

Constrained versus Unconstrained Rational Inattention

Yaron Azrieli

Department of Economics, The Ohio State University, Columbus, OH 43210, USA; azrieli.2@osu.edu

Abstract: The rational inattention literature is split between two versions of the model: in one, mutual information of states and signals are bounded by a hard constraint, while, in the other, it appears as an additive term in the decision maker's utility function. The resulting constrained and unconstrained maximization problems are closely related, but, nevertheless, their solutions differ in certain aspects. In particular, movements in the decision maker's prior belief and utility function lead to opposite comparative statics conclusions.

Keywords: costly information; rational inattention; Shannon entropy

1. Introduction

The Rational Inattention (RI) model was introduced to economics by Sims [1,2], and it has been widely applied since then in a variety of fields. It is based on the premise that attention is a scarce resource for decision makers, and that these decision makers optimally allocate their attention, given the environment that they face. Sims suggested that a useful way for capturing the scarcity of attention is to impose a constraint on the quantity of information that the agent can process. Specifically, the constraint is that the average reduction of entropy, from the agent's prior belief about the state of the world to her posterior, can not exceed a given threshold.¹

The follow-up literature continued for the most part to use entropy reduction in order to measure informativeness, but two different versions of the model emerged: the first, which we call the 'constrained version', continues as in Sims to study problems of the form $\max_x \{f(x)\}$ subject to $g(x) \leq c$. Here, x is the information choice of the agent,² the objective f maps each choice to the expected utility that it generates, $g(x)$ is the expected reduction of entropy induced by x , and c is the bound on the agent's capacity to process information. The second 'unconstrained version' instead analyzes maximization problems of the form $\max_x \{f(x) - \lambda g(x)\}$, where x, f, g are the same as before and λ captures the marginal cost of attention.

The purpose of this note is to point out that, while the two versions of the problem are obviously closely related, their solutions differ in several important aspects. Thus, the conclusions reached when using one of these versions do not automatically transfer to the other and tests of the validity of the RI model may reach different conclusions, depending on which of the two versions is tested.

The connection between the two versions is as follows: the Lagrangian of the constrained version is given by $f(x) - \lambda[g(x) - c]$, where λ is the multiplier of the RI 'budget constraint'.³ Therefore, the first-order conditions with respect to x are the same as in the unconstrained version and, since these programs are convex, the conditions are also sufficient. Furthermore, as long as c is not too large the budget constraint binds. Therefore, x solves the constrained version if and only if (i) x solves the unconstrained version with *some*



Citation: Azrieli, Y. Constrained versus Unconstrained Rational Inattention. *Games* **2021**, *12*, 3. <https://doi.org/10.3390/g12010003>

Received: 30 October 2020

Accepted: 29 December 2020

Published: 5 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

¹ The use of entropy for measuring quantity of information has its origins in [3] classical work on the capacity of a communication channel.

² More precisely, the agent chooses both the information structure and the action to take conditional on the realized signal. See the next section.

³ The Lagrangian contains additional terms due to constraints associated with x being a collection of probability distributions. These constraints are common to the two versions and are not relevant for the current discussion.

$\lambda > 0$ and (ii) the constraint binds at x . In the other direction, if x solves the unconstrained version, then it also solves the constrained version with parameter $c = g(x)$.

Despite this apparent equivalence, note that, in the constrained version, the Lagrange multiplier λ is determined endogenously, while, in the unconstrained version, it is part of the description of the problem. This is the reason underlying the differences between the solutions. First, for a fixed decision problem, the mapping from the parameter c in the constrained version to the corresponding multiplier λ need not be one-to-one, i.e., there may be an interval of c values mapped to the same multiplier, say λ^* . These critical values λ^* are associated with ‘regime changes’ in the unconstrained version, where the set of actions are considered by the agent shifts. When analyzing the unconstrained problem, these cases appear to be knife-edge and negligible, but, for the constrained problem, this is exactly where much of the “action” takes place. We demonstrate this phenomenon with a simple example (Section 3), and then show that it always happens in two families of decision problems (Propositions 2 and 3).

A byproduct of this observation is that some properties of the solution of the unconstrained version that have been emphasized in the literature fail to hold in the constrained version. For example, Caplin and Dean [4] show that, for the unconstrained problem, there is always a solution in which the number of posteriors (and, hence, the number of actions) chosen by the agent is, at most, the number of states. This is no longer true for the constrained problem: there may be intervals of c values at which any solution uses more actions than there are states.⁴ Another example is the dependence of the optimal set of posteriors on the parameter. It is easy to see that any two different values of λ lead to different sets of posteriors in the solution to the unconstrained version. In the constrained version, on the other hand, there may be intervals of c values where the optimal posteriors stay fixed and only the allocation of mass between them changes as c varies.

Second, as the decision problem changes the mapping from c to λ changes with it, which leads to reversal of known comparative statics results for the unconstrained version. Specifically, one important property of the unconstrained version is that changes in the prior do not affect the optimal set of posteriors, so long as the prior remains within the convex-hull of these posteriors. This property was termed “locally invariant posteriors” (LIP) by Caplin and Dean [4], and it has been experimentally tested by Dean and Neligh [6]. For the constrained version, quite the opposite is true: if the set of optimal posteriors is affinely independent, which is often the case, then changes in the prior almost always lead to changes in the optimal posteriors. See Proposition 5 for details.

On the other hand, scaling up or down the stakes of the decision problem works in the opposite way: while the solution to the unconstrained version is sensitive to such changes, for the constrained version, the solution stays the same. Indeed, scaling up the utility function has exactly the same effect as scaling down the marginal cost of attention λ in the unconstrained version. When λ changes, the solution to the unconstrained problem changes with it, as already mentioned above. However, in the constrained problem, a rescaling of utility is accompanied by a corresponding rescaling of the multiplier λ , and the two cancel each other.

These differences between the two versions have simple testable implications and can help to guide the modeling of rationally inattentive agents.⁵ For instance, Propositions 2 and 3 describe classes of decision problems, in which the two versions significantly differ in their predictions, offering a direct way to distinguish between them. Similarly, Proposition 5 on the failure of LIP in the constrained version can be used to refute the validity of this model while using experimental or empirical data.

⁴ Le Treust and Tomala [5] give an example that is similar to ours, where the number of actions considered is larger than the number of states in the constrained problem. They prove that there is always a solution with at most the number of states plus one actions. We discuss the relation to that paper below. The first draft of the current paper was written without awareness of this existing result.

⁵ See the related literature section below for a discussion of experimental works that test some of these implications.

Related Literature

This paper makes a theoretical contribution to the growing body of literature on RI, see Maćkowiak et al. [7] for a recent survey.

We work in a finite environment and make extensive use of the characterization of the solution of the unconstrained version in Matějka and McKay [8] (MM, henceforth) and Caplin et al. [9] (CDL, henceforth). For most of the analysis, we view the agent as choosing a distribution over posterior beliefs, rather than state-dependent distributions over actions, see e.g., Caplin and Dean [4] and Caplin et al. [10] for previous works using this approach.

The constrained and unconstrained versions are both extensively used in the literature. Roughly speaking, static models tend to adopt the unconstrained version, while dynamic models the constrained one, although there are exceptions in both directions.⁶ De Oliveira [13] axiomatizes the unconstrained version of the RI model and comments that, due to the Lagrangian connection, the constrained version behaves similarly for small variations in the menu of available acts. Matějka [11] points out that, in his model, the multiplier λ decreases as c increases, and Fulton ([14], Theorem 2) makes a similar observation in a continuous Gaussian framework; as we show below, in the discrete case, the relationship is monotonic, but sometimes only weakly so.

Le Treust and Tomala [5] analyze the interaction between a sender and a receiver, who communicate through a noisy channel. The receiver faces a sequence of n identical decision problems and the sender sends k messages through the channel. The main result is that, as k, n grow, the payoff of the sender converges to the value of the constrained version of the RI problem, with c being determined by the channel's capacity. They then show that the number of posteriors in the solution to the constrained problem can always be chosen to be, at most, one more than the number of states, and they give an example showing that this bound is tight.⁷ Their example is similar to the one that is given below in Section 3. Relative to that paper, the contribution of our Propositions 2 and 3 below is to show that more actions than states in the solution is a 'robust' phenomenon that holds for intervals of c values and, in general, classes of decision problems.

Our results may be relevant for the experimental tests of the RI model and for its estimation. Dean and Neligh [6] use the testable implications that were identified by Caplin and Dean [17] and by Caplin et al. [10] in order to study whether subjects' choices are consistent with models of costly information acquisition in general, and with the RI unconstrained model in particular. One of their findings is that subjects pay more attention (more likely to make the right choice) when the stakes are higher, which is consistent with the unconstrained version of the RI model, but not with the constrained version (see Proposition 6). Dewan and Neligh [18] observe a similar kind of behavior by most subjects (60%) in their experiment; however, note that many subjects were non-responsive to scaling up of the incentives, which suggests that the constrained model may better fit a significant fraction of the population.⁸

Dean and Neligh [6] also test the LIP property and find that it generally holds. In view of Proposition 5, this is another indication that the unconstrained model does a better job in explaining the data. It would be interesting to see whether LIP holds more generally in other kinds of decision problems and with other implementations of attention costs.

Finally, Cheremukhin et al. [19] use laboratory data in order to estimate a hybrid model that includes the two versions considered in this paper as special cases. The behavior of approximately 70% of their subjects is better described by an additive cost term than by a capacity constraint. This further suggests that the unconstrained model is a better fit for

⁶ For example, Matějka [11] uses the constrained version in a static application, while Steiner et al. [12] analyze a dynamic problem with the unconstrained version.

⁷ Also see [15,16] for related results.

⁸ Dewan and Neligh [18] also compare the performance of several cost functions for information in explaining their data. They find that most of the subjects fit well to one of four functional forms, including Shannon entropy; however, they do not explicitly consider the constrained model.

most decision makers, but, at the same time, that one should not dismiss the constrained version as irrelevant.

2. Two Versions of the RI Problem

For the most part, we follow the notation of CDL in order to facilitate an easy comparison. There is a finite set Ω of states, with $\omega \in \Omega$ denoting a typical state. The prior belief of the decision maker (DM) is $\mu \in \Delta(\Omega)$, where $\Delta(X)$ is the set of probability distributions over any finite set X . We assume that μ assigns positive probability to every $\omega \in \Omega$. The finite set of available actions is \mathcal{A} . For each pair $(a, \omega) \in \mathcal{A} \times \Omega$, the utility of the DM when she chooses action a and the realized state is ω is denoted by $u(a, \omega)$. A decision problem is described by the triplet (μ, \mathcal{A}, u) .

Throughout, we restrict attention to decision problems that satisfy two assumptions: first, actions are not duplicates, i.e., $u(a, \omega) \neq u(a', \omega)$ for some ω whenever $a \neq a'$. Second, different actions are optimal in different states, i.e., if $\omega \neq \omega'$, then $\arg \max_a u(a, \omega) \cap \arg \max_a u(a, \omega') = \emptyset$. The first assumption is purely for expositional reasons. As for the second, all of our results still hold without this assumption, but the upper end of the range of the cost parameter c for which they hold may decrease.

The DM chooses an information structure, i.e., a mapping from states to distributions over some set of signals, as well as which action to play after observing each signal. However, it is without loss ([8], Lemma 1, e.g.) to restrict attention to information structures with at most one signal per action in \mathcal{A} and to identify signals with the actions that they induce. Therefore, the choice variable is a mapping $P : \Omega \rightarrow \Delta(\mathcal{A})$, where $P(a|\omega)$ is the probability of action a conditional on state ω . With slight abuse of notation, P also denotes the unconditional probability of actions, which is $P(a) = \sum_{\omega} \mu(\omega) P(a|\omega)$. Following CDL, the *consideration set* of P is $B(P) = \{a \in \mathcal{A} : P(a) > 0\}$.

In order to state the problem, it is useful to introduce one more piece of notation. If $p \in \Delta(X)$ for some finite set X , then $H(p) = -\sum_{x \in X} p(x) \ln p(x)$ is the entropy of p . We will use H for distributions over \mathcal{A} as well as over Ω .

In their papers, MM and CDL consider the following maximization problem, where $\lambda > 0$ is an exogenous parameter:

$$\max_P \sum_{\omega} \mu(\omega) \sum_a P(a|\omega) u(a, \omega) - \lambda \left[H(P(\cdot)) - \sum_{\omega} \mu(\omega) H(P(\cdot|\omega)) \right]. \quad (\star)$$

The first term is the expected utility that the DM obtains by conditioning her choice on the observed signal, while the second term in parentheses is the expected reduction of entropy from the marginal distribution to the state-contingent distribution of actions, which captures the cost of attention.

We compare program (\star) with the following constrained maximization problem with parameter $0 < c < H(\mu)$:⁹

$$\begin{aligned} \max_P \quad & \sum_{\omega} \mu(\omega) \sum_a P(a|\omega) u(a, \omega) \\ \text{s.t.} \quad & H(P(\cdot)) - \sum_{\omega} \mu(\omega) H(P(\cdot|\omega)) \leq c. \end{aligned} \quad (\star\star) \quad (1)$$

Here the objective only includes the benefit of receiving information, and the constraint requires that the expected reduction in entropy does not exceed the 'budget of attention' c . As mentioned above, this second formulation corresponds to the original RI problem that was introduced by Sims [2].

The relationship between programs of the form (\star) and $(\star\star)$ is well-understood in general. While λ is part of the input in the former, it is the Lagrange multiplier of the con-

⁹ If $c \geq H(\mu)$, then full information is feasible and the problem becomes trivial, similarly to the case $\lambda = 0$ in (\star) .

straint in the later, and, therefore, is endogenously determined. The following proposition formalizes this connection.

Proposition 1. *The mapping P solves program $(\star\star)$ if and only if*

- (i) *P satisfies the budget constraint (1) with equality; and,*
- (ii) *P solves (\star) with some $\lambda > 0$.*

The fact that the budget constraint necessarily binds for any $c \in (0, H(\mu))$ is a consequence of our assumption that different actions are optimal at different states; a proof is provided in Appendix A. Once this is established, and given that these are convex programs, the rest of the proof easily follows from the KKT theorem ([20], Corollary 28.3.1, e.g.), and, therefore, is omitted.

Let $Z(a, \omega, \lambda) = \exp\left(\frac{u(a, \omega)}{\lambda}\right)$. MM and CDL prove that P solves (\star) if and only if it satisfies

$$P(a|\omega) = \frac{P(a)Z(a, \omega, \lambda)}{\sum_{b \in \mathcal{A}} P(b)Z(b, \omega, \lambda)} \quad (2)$$

for every a and ω , and, in addition

$$\sum_{\omega} \mu(\omega) \left(\frac{Z(a, \omega, \lambda)}{\sum_{b \in \mathcal{A}} P(b)Z(b, \omega, \lambda)} \right) \leq 1 \quad (3)$$

for every a .

It is often more convenient to work with the distribution over posteriors that are induced by P than with P itself. Namely, instead of choosing $P : \Omega \rightarrow \Delta(\mathcal{A})$, we can equivalently think of the DM as choosing the unconditional probabilities of actions $\{P(a)\}_{a \in \mathcal{A}}$ and the posteriors $\{\gamma^a \in \Delta(\Omega)\}_{a \in B(P)}$ subject to $\sum_a P(a)\gamma^a(\omega) = \mu(\omega)$ for every $\omega \in \Omega$.¹⁰ CDL show that conditions (2) and (3) can be rewritten as

$$\frac{\gamma^a(\omega)}{Z(a, \omega, \lambda)} = \frac{\gamma^b(\omega)}{Z(b, \omega, \lambda)} \quad (4)$$

for every $a, b \in B(P)$, $\omega \in \Omega$, and

$$\sum_{\omega} \gamma^a(\omega) \frac{Z(b, \omega, \lambda)}{Z(a, \omega, \lambda)} \leq 1 \quad (5)$$

for every $a \in B(P)$ and $b \in \mathcal{A}$. Furthermore, the budget constraint (1), which must bind at the optimum, can be rewritten as

$$H(\mu) - \sum_{a \in B(P)} P(a)H(\gamma^a) \leq c. \quad (6)$$

3. An Example

We now illustrate the differences between the solutions of the two problems with an example; in the next section, we show that these differences hold more generally. Let $\Omega = \{\omega_0, \omega_1\}$. Because Ω only has two elements, we identify $\Delta(\Omega)$ with the $[0, 1]$ interval and describe its elements by the probability that the state is ω_1 . The set of actions is $\mathcal{A} = \{l, m, r\}$ standing for left, middle, and right. The following table provides the utility function:

¹⁰ Indeed, starting from $P(a|\omega)$, define $P(a) = \sum_{\omega} \mu(\omega)P(a|\omega)$ and let $\gamma^a(\omega) = \frac{\mu(\omega)P(a|\omega)}{P(a)}$ for any $a \in B(P)$. In the other direction, set $P(a|\omega) = \frac{P(a)\gamma^a(\omega)}{\mu(\omega)}$.

	l	m	r
ω_0	1	0	−2
ω_1	−2	0	1

Thus, m is a safe action with a sure payoff of zero; l and r are risky actions, where l gives a high payoff at the ‘left’ state ω_0 and r gives a high payoff at the ‘right’ state ω_1 . Note that l is optimal for beliefs $\gamma \in [0, 1/3]$, m is optimal for $\gamma \in [1/3, 2/3]$, and r is optimal for $\gamma \in [2/3, 1]$. See Figure 1.

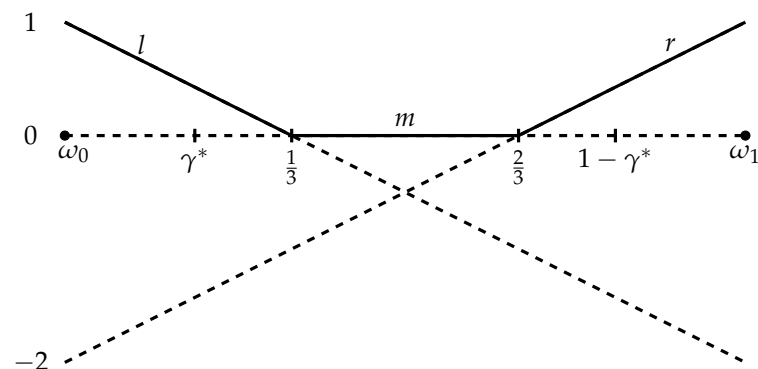


Figure 1. The dashed lines show the expected utility for each of the actions $\{l, m, r\}$ in the example as a function of the decision maker's (DM's) belief. The upper envelope of these functions is the solid black line. l is optimal for posteriors in $[0, 1/3]$, m is optimal in $[1/3, 2/3]$, and r is optimal in $[2/3, 1]$.

We use the distribution over posteriors $\{(P(l), P(m), P(r)), \{\gamma^a\}_{a \in B(P)}\}$ for the analysis, as this makes it easier to visualize the solution. We break the analysis to three different cases, depending on the location of the prior $\mu \in (0, 0.5]$, and, for each case, compare the solutions of $(\star\star)$ and (\star) .¹¹ Figures 2–4 illustrate the solutions of the three cases, while proofs of the claims can be found in the Appendix A.

In order to describe the solution, it is useful to introduce additional notation. First, let $\lambda^* = \frac{1}{\ln\left(\frac{1+\sqrt{5}}{2}\right)} \approx 2.078$ and $\gamma^* = [1 + \exp(\frac{3}{\lambda^*})]^{-1} = \frac{1}{3+\sqrt{5}} \approx 0.191$. Second, define the function $\bar{\lambda}(\mu)$ by

$$\bar{\lambda}(\mu) = \begin{cases} 3 \left[\ln\left(\frac{1-\mu}{\mu}\right) \right]^{-1} & \text{if } 0 < \mu < \gamma^*, \\ \left[\ln\left(\frac{-\mu + \sqrt{\mu(4-3\mu)}}{2\mu}\right) \right]^{-1} & \text{if } \gamma^* \leq \mu < \frac{1}{3}, \\ +\infty & \text{if } \mu = \frac{1}{3}, \\ \left[\ln\left(\frac{\mu + \sqrt{\mu(4-3\mu)}}{2(1-\mu)}\right) \right]^{-1} & \text{if } \frac{1}{3} < \mu \leq \frac{1}{2}. \end{cases}$$

Case 1: $\mu = 0.5$

For problem (\star) , the solution is as follows. If $\lambda > \lambda^*$ then $P(m) = 1$ and $\gamma^m = 0.5$. If $0 < \lambda < \lambda^*$ then $P(l) = P(r) = 0.5$ and $\gamma^l = 1 - \gamma^r = [1 + \exp(\frac{3}{\lambda})]^{-1}$. If $\lambda = \lambda^*$, then any mixture (including degenerate) of the former two solutions is optimal. Thus, if the cost parameter λ is high, then the DM chooses m for sure without seeking any information; once λ falls below the threshold level λ^* , the DM chooses sufficiently informative symmetric signals, so that she ends up choosing either l or r for sure; only at the cutoff λ^* there is the possibility that all three actions are considered, and even in this case $B(P) = \{m\}$ and $B(P) = \{l, r\}$ are still optimal. The value of λ^* is determined by the condition that the

¹¹ Because the setup is symmetric the analysis is similar for $\mu > 0.5$.

‘net utilities’ (see [9]), i.e., the difference between expected utility and cost, of the three associated posteriors $0.5, \gamma^*, 1 - \gamma^*$ are equal.

Moving onto problem $(\star\star)$, first note that it can never (for no $c > 0$) be optimal to choose no information, since, by Proposition 1, the budget constraint must bind. When c is small, specifically $0 < c < H(0.5) - H(\gamma^*)$, the solution is given by $P(l) = P(r) = \frac{c}{2[H(0.5) - H(\gamma^*)]}$, $P(m) = 1 - \frac{c}{H(0.5) - H(\gamma^*)}$, $\gamma^m = 0.5$, and $\gamma^l = 1 - \gamma^r = \gamma^*$. That is, for c in this interval, we have $B(P) = \{l, m, r\}$ and, as c increases, the posteriors stay constant while more probability is shifted from the middle $\gamma^m = 0.5$ to the extremes $\gamma^l = \gamma^*$ and $\gamma^r = 1 - \gamma^*$. In light of Proposition 1, this is possible, because, for all c in this range, the corresponding value of the lagrange multiplier is λ^* . However, in contrast to the previous paragraph, it is strictly beneficial here to the DM to move as much mass as possible to the extreme posteriors, since the cost does not directly enter the objective. For larger values of c , specifically, $H(0.5) - H(\gamma^*) \leq c < H(0.5)$, the solution is given by $P(l) = P(r) = 0.5$ with the posteriors $\gamma^l = 1 - \gamma^r$ determined by the equation $H(0.5) - H(\gamma^l) = c$.

Figure 2 illustrates the consideration sets of the two solutions. The figure also shows the mapping from c to the corresponding value of λ at the optimum of $(\star\star)$. While this mapping is weakly decreasing (a higher c implies weakly lower λ), it is neither one-to-one nor onto the entire range of λ 's.

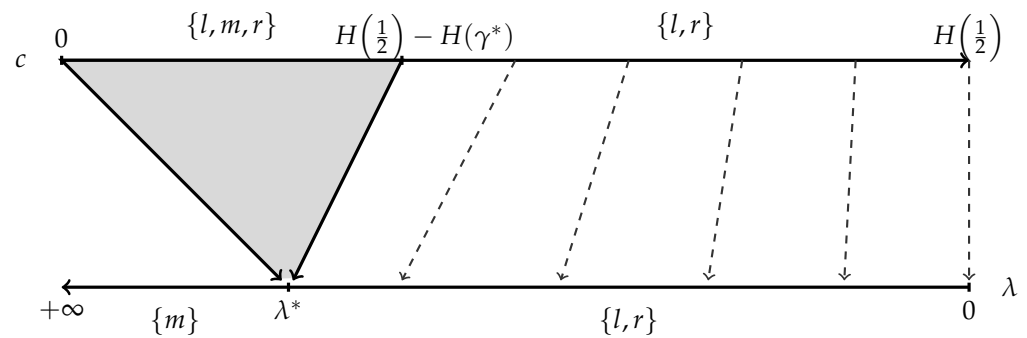


Figure 2. The solutions of problems $(\star\star)$ (upper line) and (\star) (lower line) from the example with prior $\mu = 0.5$. The arrows between the lines illustrate the mapping from c to the value of the lagrange multiplier λ at the optimum.

Case 2: $\gamma^* < \mu < 0.5$

Figure 3 illustrates the solution for this case. In (\star) , when $\lambda < \lambda^*$, the solution is similar to the previous case $\mu = 0.5$: the consideration set is $\{l, r\}$, the posteriors are $\gamma^l = 1 - \gamma^r = [1 + \exp(\frac{3}{\lambda})]^{-1}$, and the probabilities are set to satisfy $P(l)\gamma^l + P(r)\gamma^r = \mu$. At $\lambda = \lambda^*$, again, the solution is not unique, with $B(P) = \{l, r\}$, $B(P) = \{l, m\}$, and $B(P) = \{l, m, r\}$ all possible. When $\lambda^* < \lambda < \bar{\lambda}(\mu)$, $B(P) = \{l, m\}$, the posteriors are $\gamma^l = [1 + \exp(\frac{1}{\lambda}) + \exp(\frac{2}{\lambda})]^{-1}$, $\gamma^m = \exp(\frac{2}{\lambda})\gamma^l$, and the probabilities are adjusted, so that $P(l)\gamma^l + P(m)\gamma^m = \mu$. Finally, for $\lambda \geq \bar{\lambda}(\mu)$, it is optimal to obtain no information, and the consideration set is either $\{l\}$ or $\{r\}$ when μ is below or above $1/3$, respectively.¹²

In $(\star\star)$, when c is small, the consideration set is $\{l, m\}$. The posteriors γ^l, γ^m satisfy $\gamma^l(1 - \gamma^l)^2 = \gamma^m(1 - \gamma^m)^2$ (this follows from (4)) and, in addition, we need that $P(l)\gamma^l + P(m)\gamma^m = \mu$ and that $H(\mu) - [P(l)H(\gamma^l) + P(m)H(\gamma^m)] = c$. These three equations, combined, pin down the solution. Once c is sufficiently large, so that all three actions can be considered,¹³ the consideration set becomes $\{l, m, r\}$, the posteriors are fixed at $\gamma^m = 0.5$, $\gamma^l = 1 - \gamma^r = \gamma^*$, and the probabilities $(P(l), P(m), P(r))$ adjust, so that the

¹² Note that, for $\mu = 1/3$, we have $\bar{\lambda}(\mu) = +\infty$, so the latter case is empty.

¹³ This happens when $c > H(\mu) - [\frac{1-2\mu}{1-2\gamma^*}H(\gamma^*) + \frac{2\mu-2\gamma^*}{1-2\gamma^*}H(0.5)]$.

expected posterior is equal to μ and the cost is equal to c . For even larger c , namely $c \geq H(\mu) - H(\gamma^*)$, the consideration set is $\{l, r\}$ and the posteriors satisfy $\gamma^l = 1 - \gamma^r$.

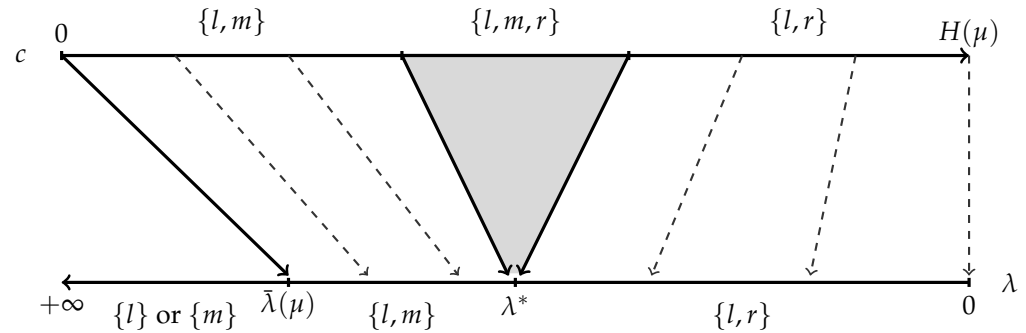


Figure 3. The consideration sets in the solutions of (*) and (**) for the case $\gamma^* < \mu < 0.5$, $\mu \neq \frac{1}{3}$. When $\mu = 1/3$, the consideration set in (*) is $\{l, m\}$ for every $\lambda > \lambda^*$.

Case 3: $0 < \mu \leq \gamma^*$

In (*), for $\lambda < \bar{\lambda}(\mu)$, we have $B(P) = \{l, r\}$, $\gamma^l = 1 - \gamma^r = [1 + \exp(\frac{3}{\lambda})]^{-1}$, and $(P(l), P(r))$ determined by $P(l)\gamma^l + P(r)\gamma^r = \mu$. For $\lambda > \bar{\lambda}(\mu)$, the solution is $B(P) = \{l\}$, i.e., no information.

In (**), for any $0 < c < H(\mu)$, the consideration set is $B(P) = \{l, r\}$ and $\gamma^l = 1 - \gamma^r$. These posteriors are determined by the equation $H(\mu) - H(\gamma^l) = c$, and the probabilities $P(l), P(r)$ are then determined by the equation $P(l)\gamma^l + P(r)\gamma^r = \mu$. See Figure 4.

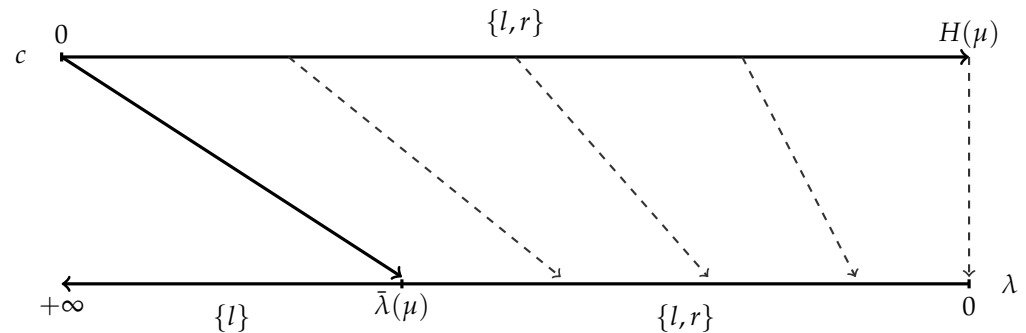


Figure 4. The solution of the example for $0 < \mu \leq \gamma^*$.

4. Model Comparison

In this section, we state several results regarding the relationship and differences between the two versions of the RI problem. These results generalize the insights that are gained from the above example. All of the proofs not in the main text can be found in the Appendix A.

We start with the following lemma that describes the correspondence between c and λ . This correspondence is key for the subsequent analysis.

Lemma 1. Fix a decision problem (μ, \mathcal{A}, u) . For every $c \in (0, H(\mu))$, there is a unique $\lambda > 0$, denoted $\lambda(c)$, such that every solution to (**) with c solves (*) only with $\lambda(c)$. The mapping $c \rightarrow \lambda(c)$ is continuous and (weakly) decreasing on $(0, H(\mu))$, and $\lim_{c \rightarrow H(\mu)-} \lambda(c) = 0$.

Lemma 1 implies that the choice consistent with optimization in problem (**) with c can be rationalized as optimal behavior in (*) only for one value of λ , namely $\lambda(c)$. The continuity and monotonicity of $\lambda(c)$ imply that its image contains the entire interval $(\lim_{c \rightarrow H(\mu)-} \lambda(c), \lim_{c \rightarrow 0+} \lambda(c)) = (0, \lim_{c \rightarrow 0+} \lambda(c))$. Therefore, for every λ in this interval, the optimal behavior in (*) can be rationalized as optimal behavior in (**) with some c .

However, as illustrated in the example, for a given λ , there may be multiple solutions that correspond to different values of c , so a solution to (\star) with $\lambda(c)$ need not be optimal (or even feasible) for $(\star\star)$ with c .

4.1. Consideration Sets and Optimal Posteriors

Perhaps the most apparent difference between the solutions of (\star) and $(\star\star)$ in the example is that there is an interval of c values for which the solution of $(\star\star)$ has all three actions $\{l, m, r\}$ considered, while, in (\star) , this can only happen at a single point λ^* , and, even at λ^* , there are other solutions that only involve subsets of actions. It is well-known [4] that this feature of the solution of (\star) is true in general: in every decision problem, there is always a solution in which the size of the consideration set is, at most, $|\Omega|$. We now show that, in several cases of interest, the solution of $(\star\star)$ behaves quite differently. Thus, the example of Section 3 is in no way special.

The first result generalizes the example to any decision problem with a binary state-space, at least three actions, and a not-too-extreme prior.

Proposition 2. Let $\Omega = \{\omega_0, \omega_1\}$ and denote, by a_i , the (unique) optimal action at ω_i ($i = 0, 1$). If neither a_0 nor a_1 are optimal at the prior μ , then there is $\lambda^* > 0$ and $0 < \underline{c} < \bar{c} < H(\mu)$, such that, for every $c \in (\underline{c}, \bar{c})$ (i) $\lambda(c) = \lambda^*$ and (ii) $|B(P)| \geq 3$ for every P that solves $(\star\star)$.

In the next proposition, we consider a class of decision problems similar to the one analyzed in (Caplin et al. [9], Section 3.1), but with an additional action corresponding to the outside option of the DM. Let $\Omega = \{\omega_1, \dots, \omega_m\}$. Consider the decision problem in which $\mathcal{A} = \{a_1, \dots, a_m, o\}$, and the utility function is given by

$$u(a_i, \omega_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

and $u(o, \omega_i) = t$ for every $1 \leq i \leq m$, where $\frac{m-1}{m} < t < 1$. Thus, if the DM correctly guesses the state, then her payoff is 1, while any wrong guess yields a payoff of 0. In addition, the safe choice o guarantees a payoff of t .¹⁴ For this class of problems, we obtain a similar result to that of Proposition 2.

Proposition 3. Consider a decision problem, as described above, and suppose that the prior μ satisfies $1 - t < \mu(\omega_i)$ for every i . Then there is $\lambda^* > 0$ and $0 < \underline{c} < \bar{c} < H(\mu)$, such that, for every $c \in (\underline{c}, \bar{c})$ (i) $\lambda(c) = \lambda^*$ and (ii) $B(P) = \mathcal{A}$ for every P that solves $(\star\star)$.

The assumption that $1 - t < \mu(\omega_i)$ for each i guarantees that the optimal action at the prior μ is the outside option o , and that μ is centrally located, in the sense that it is not in the convex-hull of any collection of $m - 1$ posteriors at which the a_i actions are optimal.

The intuition for the last two propositions is similar to the one in the example: when λ is small, the DM obtains precise information guaranteeing that one of the ‘extreme’ actions (the a_i ’s) will be selected. When λ is relatively large, the DM may seek some information, but it will often end up at a posterior in which a safer action is optimal (o in Proposition 3). The transition between these two regimes happens at λ^* . Moreover, at λ^* , mixtures of these two types of solutions are also optimal, so there is a range of c values mapped to λ^* and both types of actions are considered.

Remark 1. Another known property of (\star) is that the same set of posteriors can not be a solution with two different values of λ , while assuming that some information is obtained (this immediately follows from condition (4) above). As shown in the example, this is not true for the parameter c in $(\star\star)$: there is an interval of c values, such that the chosen posteriors are fixed, and only the allocation

¹⁴ In Caplin et al. [9], the available actions are just $\{a_1, \dots, a_m\}$. Setting the payoffs to 1 and 0 is just for convenience, the same result (with the obvious changes) applies with similar payoff structures.

of mass over these posteriors changes as c varies. The proofs of Propositions 2 and 3 make it clear that the same is also true in these families of decision problems.

4.1.1. An ‘Anything Goes’ Result

We end this section by arguing that the testable implications of the constrained model are limited if the analyst does not know the DM’s utility function and prior. Namely, any finite set of posteriors that is not convex independent can arise as the solution to problem (★★) for an interval of c values in *some* decision problem. Note that the set of posteriors can be arbitrarily large.

Proposition 4. Fix Ω , let $n \geq 3$ be an arbitrary integer, and consider a collection $\Gamma = \{\tilde{\gamma}^i\}_{i=1}^n$ of distinct elements in the relative interior of $\Delta(\Omega)$. If Γ is not convex independent (i.e., if there is $\tilde{\gamma}^i$ in the convex-hull of $\Gamma \setminus \{\tilde{\gamma}^i\}$), then there is a decision problem (μ, \mathcal{A}, u) and $0 < \underline{c} < \bar{c} < H(\mu)$ such that for every $c \in (\underline{c}, \bar{c})$ (i) there is a solution of (★★), in which the set of posteriors is Γ ; and (ii) if $\gamma \notin \Gamma$, then γ is not part of any solution of (★★).

A couple of comments are in order. First, property (ii) of the proposition guarantees that optimal posteriors must be in Γ ; in particular, the decision problem is non-trivial in the sense that not every P is optimal. Second, while we know that the set Γ is also a solution to the unconstrained problem for some λ , Remark 1 above implies that it is not robustly so in the sense that arbitrarily small changes in λ would change the optimal set of posteriors. This is in contrast to the constrained version, in which Γ remains optimal for all $c \in (\underline{c}, \bar{c})$.

4.2. Comparative Statics

4.2.1. Locally Variant Posteriors

The “locally invariant posteriors” (LIP) property [4] states that changes in the prior do not affect the optimal set of posteriors for (★) whenever this set is still feasible. Dean and Neligh [6] experimentally test this property and find that it is generally satisfied.

In problem (★★), on the other hand, arbitrarily small changes in the prior typically induce different sets of optimal posteriors. Consider the example of Section 3 with some given prior μ , and suppose that c is large enough, so that the optimal consideration set is $\{l, r\}$. The optimal posteriors satisfy $\gamma^l = 1 - \gamma^r$ in this case. Therefore, by the symmetry of H , the cost of an optimal P is given by $H(\mu) - [P(l)H(\gamma^l) + P(r)H(\gamma^r)] = H(\mu) - H(\gamma^l)$. Because the budget constraint binds, we must have $H(\mu) - H(\gamma^l) = c$. Therefore, if c is fixed and the prior μ changes, then the optimal posteriors γ^l, γ^r must also change.

The reason for the failure of LIP is clear: suppose that we fix a set of affinely independent posteriors $\{\gamma^a\}$. For μ in the convex-hull of this set, say $\mu = \sum_a P(a)\gamma^a$, the cost associated with the choice of these posteriors is $H(\mu) - \sum_a P(a)H(\gamma^a)$. The first term is strictly concave in μ , while the second is linear in μ (since the vector $\{P(a)\}$ changes linearly with μ). Therefore, with a fixed set of posteriors, the cost is a strictly concave function of the prior μ , which implies that changes in the prior typically lead to variation in the cost. Because the budget constraint binds, the set of posteriors associated with an optimal solution must adjust to keep the cost unchanged. Therefore, we have the following.

Proposition 5. Consider a decision problem (μ, \mathcal{A}, u) and parameter c , such that (★★) has a unique solution P with posteriors $\{\gamma^a\}_{a \in B(P)}$ that are affinely independent. Then the set of priors μ' for which the same set of posteriors $\{\gamma^a\}_{a \in B(P)}$ is optimal for decision problem (μ', \mathcal{A}, u) with parameter c is a nowhere dense subset of $\Delta(\Omega)$.¹⁵

¹⁵ We view $\Delta(\Omega)$ as a subset of the Euclidean space \mathbb{R}^Ω and endow it with the topology that it inherits from that space. A set in a topological space is nowhere dense if its closure has an empty interior. The proposition immediately follows from the fact that the level sets of a strictly concave and continuous function are closed and contain no open set.

Another way to think about the failure of LIP is that, under the conditions of the proposition, $\lambda(c)$ for decision problem (μ, \mathcal{A}, u) is usually different than $\lambda(c)$ for (μ', \mathcal{A}, u) . When λ changes, the optimal set of posteriors in (\star) changes with it, as mentioned in the previous subsection.

We note that the assumption of affinely independent posteriors can not be dispensed with. Indeed, going back to the example, consider the case $\mu = 0.5$ and c not too large, so that the solution to $(\star\star)$ has $B(P) = \{l, m, r\}$. For μ' close to μ and the same c , the solution to $(\star\star)$ still has $B(P) = \{l, m, r\}$. Additionally, when all three actions are considered, the posteriors must be $\gamma^l = \gamma^*, \gamma^m = 0.5$, and $\gamma^r = 1 - \gamma^*$.

Finally, the difference between the unconstrained and constrained versions of the RI model that are exhibited in Proposition 5 is much more general than in the case where cost is measured by a reduction of entropy. Indeed, both the LIP property in the unconstrained version and its failure in the constrained version continue to hold so long as cost is measured by the expected reduction in the value of a strictly concave function of posteriors.

4.2.2. Utility Scaling

While the solution of problem $(\star\star)$ is sensitive to changes in the prior, it does not change when the stakes of the decision problem are scaled up or down.

Proposition 6. Consider a decision problem (μ, \mathcal{A}, u) and suppose that P solves $(\star\star)$ with parameter c . Then P is also a solution of $(\star\star)$ with parameter c in the decision problem $(\mu, \mathcal{A}, \rho \cdot u)$ for every $\rho > 0$.

The argument is straightforward: since P is optimal in (μ, \mathcal{A}, u) , it follows from Proposition 1 that P solves (\star) for this decision problem with some λ . Therefore, P also solves (\star) for the scaled decision problem $(\mu, \mathcal{A}, \rho \cdot u)$ with $\lambda' = \rho \cdot \lambda$. Because the cost of P does not change with the problem, Proposition 1 implies that P also solves $(\star\star)$ for $(\mu, \mathcal{A}, \rho \cdot u)$ with the same c .

Notice that stakes do matter in the unconstrained version: scaling up the utility has the exact same effect as scaling down the marginal cost of information λ , which, as already discussed, necessarily changes the solution.

4.3. Optimality of ‘No Information’

For large values of λ , the solution of (\star) often involves the DM choosing not to be informed at all, as demonstrated by the example. Indeed, that was the case for any prior μ , except for $\mu = \frac{1}{3}$. On the other hand, since the budget constraint must bind, choosing no information can not be optimal in $(\star\star)$ for any $c > 0$.

In light of Lemma 1, this gap between the two versions occurs if and only if the limit $\bar{\lambda} := \lim_{c \rightarrow 0+} \lambda(c)$ is finite. Indeed, there is a zero-cost (i.e., uninformative) solution to (\star) with λ if and only if $\lambda \geq \bar{\lambda}$. In the next proposition, we characterize those decision problems for which $\bar{\lambda}$ is finite and show that this is the ‘typical’ case.

Formally, we say that P is *uninformative* if $P(\cdot|\omega)$ is the same for all $\omega \in \Omega$, or, equivalently, if the posterior is equal to the prior with probability 1. Additionally, given (\mathcal{A}, u) , say that the prior μ is an *indifference point* if there are two different actions $a, a' \in \mathcal{A}$, such that both a and a' are optimal at belief μ , i.e., $\{a, a'\} \subseteq \arg \max_b \sum_{\omega} \mu(\omega) u(b, \omega)$. Note that, for a given (\mathcal{A}, u) , the set of indifference points is “small” in $\Delta(\Omega)$, e.g., it is nowhere dense and has Lebesgue measure zero in $\Delta(\Omega)$ viewed as a subset of $\mathbb{R}^{|\Omega|} - 1$.

Proposition 7. Suppose that μ is not an indifference point of (\mathcal{A}, u) and let a^* be the (unique) optimal action given belief μ . Then the limit $\bar{\lambda} = \lim_{c \rightarrow 0+} \lambda(c)$ is finite and, for every $\lambda > \bar{\lambda}$, the unique solution to (\star) is given by $P(a^*|\omega) = 1$ for all ω . If μ is an indifference point, then $\bar{\lambda} = +\infty$ and uninformative P ’s are never optimal.

Intuitively, when λ is large condition (4) that characterizes the solution to (\star) requires that the posteriors are close to each other and, therefore, close to the prior μ . Thus, if μ is

not an indifference point, then, for large enough λ , the same action that is optimal at μ is also optimal at all posteriors. But then obtaining no information yields the same expected utility at a lower cost. Conversely, when μ is an indifference point, the marginal value of little information is positive, since it allows the DM to learn which of the a-priori optimal actions is better. The marginal cost of little information is zero due to the smoothness of the entropy function. Therefore, it is not optimal to obtain no information.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proofs

Appendix A.1. Notation

The following notation and remarks will be used in several of the proofs below. We use P to denote the state-dependent stochastic choice of the DM, i.e., $P : \Omega \rightarrow \Delta(\mathcal{A})$. The distribution over posteriors induced by P is denoted by $\Gamma(P)$. Recall that the support of $\Gamma(P)$ is the collection $\{\gamma^a\}_{a \in B(P)}$, where $\gamma^a(\omega) = \frac{\mu(\omega)P(a|\omega)}{P(a)}$, and $P(a) = \sum_{\omega} \mu(\omega)P(a|\omega)$ is the probability assigned to posterior γ^a . Given P, P' and $\alpha \in [0, 1]$, the mixture of the induced distributions over posteriors $\alpha\Gamma(P) + (1 - \alpha)\Gamma(P')$ is defined as usual: The support of the mixture is the union of the supports of $\Gamma(P)$ and $\Gamma(P')$, and the probability of each γ in the support is the corresponding average of probabilities of γ in the two distributions.

Remark A1. Suppose that P solves either (\star) or $(\star\star)$. If $a \in B(P)$, then choosing a must be optimal given belief γ^a , that is $a \in \arg \max_b \sum_{\omega} \gamma^a(\omega)u(b, \omega)$. Indeed, if that was not the case then choosing the same distribution over posteriors but playing some action from the $\arg \max$ would increase the utility of the DM without changing the cost of information. Furthermore, a must be the unique optimal action given belief γ^a . Indeed, if there are multiple optimal actions at some induced posterior, then this is inconsistent with optimality, as shown in the proof of Proposition 7.

Remark A2. Given P and P' , if $a \in B(P) \cap B(P')$ and the posteriors γ^a induced by P and γ'^a induced by P' are not equal, then the mixture $\alpha\Gamma(P) + (1 - \alpha)\Gamma(P')$ has two different posteriors associated with the same action a . In particular, if P and P' are both optimal in (\star) , and if $a \in B(P) \cap B(P')$, then $\gamma^a = \gamma'^a$. Indeed, if P and P' are optimal, then so is any mixture of $\Gamma(P)$ and $\Gamma(P')$.

Given a decision problem (μ, \mathcal{A}, u) , we write

$$V(P) = \sum_{\omega} \mu(\omega) \sum_a P(a|\omega) u(a, \omega)$$

for the expected utility that the DM obtains with choice P . Notice that this can also be expressed using $\Gamma(P)$ as

$$V(P) = \sum_a P(a) \sum_{\omega} \gamma^a(\omega) u(a, \omega).$$

Similarly, we write

$$C(P) = H(P(\cdot)) - \sum_{\omega} \mu(\omega) H(P(\cdot|\omega)) = H(\mu) - \sum_a P(a) H(\gamma^a)$$

for the cost of P . Thus, problem (\star) can be written as $\max_P \{V(P) - \lambda C(P)\}$, and problem $(\star\star)$ as $\max_P \{V(P)\}$ subject to $C(P) \leq c$.

Appendix A.2. Proofs of Propositions

Proof of Proposition 1. Fix a decision problem (μ, \mathcal{A}, u) and $c \in (0, H(\mu))$. We prove here that if P is optimal for $(\star\star)$ then the budget constraint (1) binds. As explained in the text, the rest of the argument is a standard application of the KKT theorem.

For each $\omega \in \Omega$ let $A^*(\omega) = \arg \max_a u(a, \omega)$. Denote $\bar{V} = \sum_{\omega} \mu(\omega) u(a^*(\omega), \omega)$, where $a^*(\omega) \in A^*(\omega)$. Then a choice P gives the DM expected utility of \bar{V} if and only if $\sum_{a \in A^*(\omega)} P(a|\omega) = 1$ for every ω . Our assumption that different actions are optimal at different states means that $A^*(\omega) \cap A^*(\omega') = \emptyset$ whenever $\omega \neq \omega'$. It follows that if P achieves \bar{V} then P reveals the realized state with probability one, and therefore that $C(P) = H(\mu)$.

Now, suppose that P is such that the constraint is slack, $C(P) < c$. Because $c < H(\mu)$ it follows from the previous paragraph that $V(P) < \bar{V}$. For $\epsilon > 0$ define $P(\epsilon) = \epsilon \bar{P} + (1 - \epsilon)P$, where \bar{P} satisfies $V(\bar{P}) = \bar{V}$. Then $V(P(\epsilon)) = \epsilon \bar{V} + (1 - \epsilon)V(P) > V(P)$, and since C is continuous in P we have that $C(P(\epsilon)) < c$ for ϵ small enough, so $P(\epsilon)$ is feasible. This shows that P is not a solution for $(\star\star)$. \square

Proof of Lemma 1. Fix c and let P be a solution to $(\star\star)$. By Proposition 1 the budget constraint binds, implying that P must be informative. In particular, there must be two actions $a \neq b$ such that $a, b \in B(P)$. Moreover, Proposition 1 implies that there is $\lambda > 0$, such that conditions (4) and (5) are satisfied. From (4) we have that for every ω

$$\frac{\gamma^a(\omega)}{Z(a, \omega, \lambda)} = \frac{\gamma^b(\omega)}{Z(b, \omega, \lambda)},$$

or, after rearranging,

$$\frac{\gamma^a(\omega)}{\gamma^b(\omega)} = \frac{Z(a, \omega, \lambda)}{Z(b, \omega, \lambda)} = \exp\left(\frac{u(a, \omega) - u(b, \omega)}{\lambda}\right).$$

Because we assumed that actions are not duplicates, there exists some ω at which $u(a, \omega) - u(b, \omega) \neq 0$, which implies that the right-hand side of the last equation is strictly monotone in λ . It follows that λ is pinned down uniquely by P . Denote this λ by $\lambda(c)$.

Now, suppose that P' also solves $(\star\star)$ with the same c . By Proposition 1 P' solves (\star) with some $\lambda' > 0$. We claim that it must be the case that $\lambda' = \lambda(c)$. Indeed $V(P') = V(P)$ since they are both optimal in $(\star\star)$, and, since the budget constraint binds, $C(P') = c = C(P)$. This implies that P is also optimal for (\star) with λ' , so by the previous paragraph $\lambda' = \lambda(c)$.

Next, we prove that $\lambda(c)$ is (weakly) decreasing. Let $c > c'$ and suppose per absurdum that $\lambda(c) > \lambda(c')$. Let P be optimal for $(\star\star)$ with c and P' optimal for $(\star\star)$ with c' . Then

$$V(P) - V(P') \geq \lambda(c)[C(P) - C(P')] = \lambda(c)[c - c'] > \lambda(c')[c - c'] = \lambda(c')[C(P) - C(P')],$$

where the first inequality follows from P being optimal for (\star) with $\lambda(c)$, the next equality holds since the budget constraint in $(\star\star)$ binds, the strict inequality is by the assumptions that $c > c'$ and $\lambda(c) > \lambda(c')$, and the last equality is again by the binding budget constraint. Rearranging gives

$$V(P) - \lambda(c')C(P) > V(P') - \lambda(c')C(P'),$$

contradicting the optimality of P' for (\star) with $\lambda(c')$.

Finally, we show that the image of $\lambda(c)$ contains the entire open interval $(0, \lim_{c \rightarrow 0+} \lambda(c))$. Combining this with the monotonicity proved above implies both continuity and $\lim_{c \rightarrow H(\mu)-} \lambda(c) = 0$. Suppose that $\tilde{\lambda}$ is in this interval. Let P be optimal for (\star) with $\tilde{\lambda}$. Notice that $C(P) = 0$ is impossible, since that would imply that any $\lambda > \tilde{\lambda}$ also has an uninformative solution, contradicting the assumption that $\tilde{\lambda} < \lim_{c \rightarrow 0+} \lambda(c)$. It also can not be that $C(P) = H(\mu)$ since that would require the posteriors to be at the vertices of the simplex, contradicting (4). Thus, $0 < C(P) < H(\mu)$ and it follows from Proposition 1 that P is

also optimal for $(\star\star)$ with $c = C(P)$. Therefore, $\tilde{\lambda} = \lambda(C(P))$, i.e., $\tilde{\lambda}$ is in the image. This completes the proof. \square

Proof of Proposition 2. Because there are only two states, we identify distributions over Ω with the probability they assign to ω_1 . Define Λ to be the set of all $\lambda > 0$, such that there exists a solution P to (\star) satisfying $B(P) = \{a_0, a_1\}$. We break the proof into several claims.

Claim 1. *The set Λ is non-empty and bounded from above.*

Proof. If P is optimal and $a, b \in B(P)$ then by (4) we have $\frac{\gamma^a}{\gamma^b} = \exp\left(\frac{u(a, \omega_1) - u(b, \omega_1)}{\lambda}\right)$ and also $\frac{1 - \gamma^a}{1 - \gamma^b} = \exp\left(\frac{u(a, \omega_0) - u(b, \omega_0)}{\lambda}\right)$. As $\lambda \rightarrow 0$ we either get $\gamma^b \rightarrow 0$ and $\gamma^a \rightarrow 1$, or vice versa.¹⁶ From Remark A1, a and b must be optimal given beliefs γ^a and γ^b , respectively. This is only possible if $\{a, b\} = \{a_0, a_1\}$ when λ is sufficiently small. It is also clear that choosing ‘no information’ is not optimal for λ small enough, since it is not optimal at $\lambda = 0$. Thus, every sufficiently small $\lambda > 0$ is in Λ .

On the other hand, as $\lambda \rightarrow +\infty$ the ratio $\frac{\gamma^a}{\gamma^b}$ converges to 1. Because μ must be in the convex-hull of the induced posteriors, all the induced posteriors necessarily converge to μ as $\lambda \rightarrow +\infty$ (see the proof of Proposition 7 for details). Additionally, since we assumed that neither a_0 nor a_1 are optimal at μ , this implies that these actions are not considered for large enough λ . Thus, Λ is bounded from above. \square

Claim 2. *Let $\lambda^* = \sup \Lambda$. Then in problem (\star) with λ^* there are two solutions, P^* and P^{**} , such that $B(P^*) = \{a_0, a_1\}$ and $B(P^{**}) \neq \{a_0, a_1\}$.*

Proof. First, consider a sequence $\{\lambda_n\}$ converging to λ^* from below, such that for each n there is a solution P_n to (\star) with λ_n satisfying $B(P_n) = \{a_0, a_1\}$. Such a sequence exists by the definition of λ^* . By taking a subsequence if needed we may assume that P_n converges. By the theorem of the maximum the limit P^* is optimal at λ^* . In addition, we must have $B(P^*) = \{a_0, a_1\}$: For every $a \neq a_0, a_1$ we have $P_n(a) = 0$ for all n , implying $P^*(a) = 0$. Additionally, it is impossible that $B(P^*) = \{a_0\}$ or $B(P^*) = \{a_1\}$ since these actions are not optimal at μ .

Second, let $\{\lambda'_n\}$ be a sequence converging to λ^* , but this time from above. Let P'_n be a corresponding solution sequence such that $|B(P'_n)| \leq 2$ for each n .¹⁷ Subsequently, the limit $P^{**} = \lim_n P'_n$ is optimal at λ^* , and by the definition of λ^* we have $B(P'_n) \neq \{a_0, a_1\}$ for every n , also implying that $B(P^{**}) \neq \{a_0, a_1\}$. \square

Claim 3. *For P^* and P^{**} constructed in Claim 2, $C(P^*) > C(P^{**})$.*

Proof. Because H is strictly concave over $\Delta(\Omega)$, it is sufficient to show that $\Gamma(P^*)$ is a mean-preserving spread of $\Gamma(P^{**})$. We consider two different cases. Suppose first that $B(P^{**}) \cap B(P^*) = \emptyset$, i.e., neither of the two extreme actions is in the consideration set of P^{**} . Subsequently, by Remark A1, every posterior induced by P^{**} is in-between the two posteriors induced by P^* . This implies that $\Gamma(P^*)$ is a mean-preserving spread of $\Gamma(P^{**})$ as needed.

The other case is when $B(P^{**}) \cap B(P^*) \neq \emptyset$. Since $B(P^{**})$ has at most two elements, the intersection contains exactly one element, say $B(P^{**}) \cap B(P^*) = \{a_0\}$. Denote the support of $\Gamma(P^*)$ by $\{\gamma_*^{a_0}, \gamma_*^{a_1}\}$ and the support $\Gamma(P^{**})$ by $\{\gamma_{**}^{a_0}, \gamma_{**}^{a_1}\}$. Remark A1 implies that $\gamma_{**}^{a_1} < \gamma_*^{a_1}$. In addition, since both P^* and P^{**} are optimal, Remark A2 implies that $\gamma_*^{a_0} = \gamma_{**}^{a_0}$. Therefore, we again get that $\Gamma(P^*)$ is a mean-preserving spread of $\Gamma(P^{**})$ and the claim is proved. If $B(P^{**}) \cap B(P^*) = \{a_1\}$ the argument is symmetric. \square

¹⁶ If one of the actions (weakly) dominates the other then the latter can not be part of an optimal solution.

¹⁷ Recall that there is always a solution in which the size of the consideration set is at most the number of states [4].

So far, we have shown that $\lambda(C(P^*)) = \lambda(C(P^{**})) = \lambda^*$, and that $C(P^*) > C(P^{**})$. By monotonicity proved in Lemma 1 this implies that $\lambda(c) = \lambda^*$ for every $c \in (C(P^{**}), C(P^*))$. The next claim completes the proof of the proposition.

Claim 4. Denote $\bar{c} = C(P^*)$. There is $\underline{c} < \bar{c}$ such that if P is optimal for $(\star\star)$ with $c \in (\underline{c}, \bar{c})$ then $|B(P)| \geq 3$.

Proof. Assume, contrary to the claim, that there is a sequence c_n converging to \bar{c} from below and a corresponding sequence P_n , such that P_n solves $(\star\star)$ with c_n and $|B(P_n)| \leq 2$ for every n . Because of the budget constraint binds, we have $C(P_n) = c_n$ for all n , so for n large enough P_n also solves (\star) with λ^* . Denote $B(P_n) = \{a(n), b(n)\}$ and the corresponding posteriors by $\gamma^{a(n)}, \gamma^{b(n)}$ (it can not be that $|B(P_n)| = 1$ since then $C(P_n) = 0$). Additionally, recall that $\gamma_*^{a_0}, \gamma_*^{a_1}$ are the posteriors induced by P^* . There are three cases to consider:

First, it can not be that $\{a(n), b(n)\} = \{a_0, a_1\}$, since by Remark A2 this would imply $\gamma^{a(n)} = \gamma_*^{a_0}$ and $\gamma^{b(n)} = \gamma_*^{a_1}$, contradicting the assumption that $c_n < \bar{c}$.

Second, suppose that $\{a(n), b(n)\} = \{a_0, a\}$ for some action $a \neq a_1$. Then again by Remark A2 this would imply $\gamma^{a(n)} = \gamma_*^{a_0}$. Additionally by Remark A1 the other posterior γ^a is smaller and bounded away from $\gamma_*^{a_1}$, contradicting the assumption that $c_n \rightarrow \bar{c}$. The argument for $\{a(n), b(n)\} = \{a_1, a\}$ is analogous.

Third, if $\{a(n), b(n)\} \cap \{a_0, a_1\} = \emptyset$, then both $\gamma^{a(n)}$ and $\gamma^{b(n)}$ are between and bounded away from $\gamma_*^{a_0}, \gamma_*^{a_1}$, so $C(P_n)$ can not converge to \bar{c} . \square

Proof of Proposition 3. For the most part the proof follows the footsteps of the proof of Proposition 2. We only provide details when a different argument is needed.

Claim 5. For any $\lambda > 0$, if P solves (\star) then either $B(P) = \{a_1, \dots, a_m\}$ or $o \in B(P)$.

Proof. Suppose by contradiction that P is optimal and $B(P) \subsetneq \{a_1, \dots, a_m\}$. From Remark A1, if $a_j \in B(P)$ then the associated posterior γ^{a_j} must satisfy $\gamma^{a_j}(\omega_j) > t$, implying that $\gamma^{a_j}(\omega_i) < 1 - t$ for any $i \neq j$. Let i be such that $a_i \notin B(P)$. Then

$$\mu(\omega_i) = \sum_{a_j \in B(P)} P(a_j) \gamma^{a_j}(\omega_i) < \sum_{a_j \in B(P)} P(a_j) (1 - t) = (1 - t),$$

contradicting the assumption in the proposition. \square

Define Λ to be the set of all $\lambda > 0$, such that there exists a solution P to (\star) , satisfying $B(P) = \{a_1, \dots, a_m\}$.

Claim 6. The set Λ is non-empty and bounded from above.

Proof. We first prove that, if λ is small enough and P solves (\star) , then $o \notin B(P)$. By Claim 5, this implies that $B(P) = \{a_1, \dots, a_m\}$, so $\Lambda \neq \emptyset$.

For o to be considered, Condition (5) requires that for every $1 \leq i \leq m$

$$1 \geq \sum_{j=1}^m \gamma^0(\omega_j) \exp\left(\frac{u(a_i, \omega_j) - t}{\lambda}\right) = \gamma^0(\omega_i) \exp\left(\frac{1-t}{\lambda}\right) + (1 - \gamma^0(\omega_i)) \exp\left(\frac{-t}{\lambda}\right) \geq \gamma^0(\omega_i) \exp\left(\frac{1-t}{\lambda}\right).$$

Summing up over all i gives

$$m \geq \sum_{i=1}^m \gamma^0(\omega_i) \exp\left(\frac{1-t}{\lambda}\right) = \exp\left(\frac{1-t}{\lambda}\right).$$

Since $t < 1$ this inequality clearly can not hold for $\lambda > 0$ small enough.

To show that Λ is bounded, note that the assumption of the proposition implies that o is the unique optimal action at μ . By Proposition 7, this implies that obtaining no information is the unique optimal choice for all λ large enough. \square

Claim 7. Let $\lambda^* = \sup \Lambda$. Then in problem (\star) with λ^* there are two solutions, P^* and P^{**} , such that $B(P^*) = \{a_1, \dots, a_m\}$ and $B(P^{**}) \neq \{a_1, \dots, a_m\}$.

Proof. The proof is identical to that in Claim 2, except that to argue that $B(P^*) = \{a_1, \dots, a_m\}$ we need to use Claim 5. \square

Claim 8. For P^* and P^{**} constructed in the previous claim, $C(P^*) > C(P^{**})$.

Proof. We denote by γ_*^i the posterior corresponding to a_i induced by P^* and by γ_{**}^i the one induced by P^{**} (if $a_i \in B(P^{**})$).

Because $o \notin B(P^*)$, P^* must coincide with the solution given in ([9], Theorem 1) when all of the a_i 's are considered. In particular, the induced posteriors are symmetric in the sense that $\gamma_*^i(\omega_i)$ is the same for all i . It follows that $V(P^*) = \gamma_*^1(\omega_1)$. On the other hand, by Remark A2 we must have $\gamma_*^i = \gamma_{**}^i$ whenever $a_i \in B(P^{**})$, implying that

$$\begin{aligned} V(P^{**}) &= P^{**}(o)t + \sum_{a_i \in B(P^{**})} P^{**}(a_i)\gamma_{**}^i(\omega_i) = P^{**}(o)t + \sum_{a_i \in B(P^{**})} P^{**}(a_i)\gamma_*^i(\omega_i) = \\ &P^{**}(o)t + (1 - P^{**}(o))\gamma_*^1(\omega_1) < \gamma_*^1(\omega_1) = V(P^*), \end{aligned}$$

where the strict inequality follows from $o \in B(P^{**})$ (recall Claim 5) and $t < \gamma_*^1(\omega_1)$. Therefore, we got that $V(P^{**}) < V(P^*)$, and since both are optimal for (\star) with λ^* it must be that $C(P^*) > C(P^{**})$. \square

Claim 9. Denote $\bar{c} = C(P^*)$. There is $\underline{c} < \bar{c}$ such that if P is optimal for $(\star\star)$ with $c \in (\underline{c}, \bar{c})$ then $B(P) = \mathcal{A}$.

Proof. Suppose by contradiction that $c_n \uparrow \bar{c}$, and that for each n P_n solves $(\star\star)$ with c_n but $B(P_n) \neq \mathcal{A}$. Then $C(P_n) = c_n$, so for n large enough P_n also solves (\star) with λ^* . Since $c_n < \bar{c}$ we can not have $B(P_n) = \{a_1, \dots, a_m\}$, so by Claim 5 $o \in B(P_n)$. The exact same argument as in the previous claim gives $V(P_n) = P_n(o)t + (1 - P_n(o))\gamma_*^1(\omega_1)$. Moreover, $P_n(o)$ is bounded away from zero, since μ is not in the convex-hull of any strict subset of $\{\gamma_*^i\}_{i=1}^m$. This implies that there is $\delta > 0$ such that $V(P^*) - V(P_n) > \delta$ for every n . However, P^* and P_n are both optimal for (\star) with λ^* , so $V(P^*) - \lambda^*C(P^*) = V(P_n) - \lambda^*C(P_n)$. This contradicts the convergence of $c_n = C(P_n)$ to $\bar{c} = C(P^*)$. \square

Proof of Proposition 4. Fix a collection Γ as in the proposition. Define $f : \Delta(\Omega) \rightarrow \mathbb{R}$ by $f(\gamma) = H(\mu) - H(\gamma)$. Because f is strictly convex, for each $1 \leq i \leq n$ there is an affine function $f_i : \Delta(\Omega) \rightarrow \mathbb{R}$ such that $f_i(\tilde{\gamma}^i) = f(\tilde{\gamma}^i)$ and $f_i(\gamma) < f(\gamma)$ for every $\gamma \in \Delta(\Omega)$. Let the set of actions be $\mathcal{A} = \{1, \dots, n\}$ and the utility function be $u(i, \omega) = f_i(\gamma_\omega)$, where $\gamma_\omega \in \Delta(\Omega)$ is the Dirac measure on state ω .

To complete the description of the decision problem we need to choose the prior μ . By assumption, one of the elements of Γ , say $\tilde{\gamma}^1$, is in the convex hull of the others. Define $\mu = (1 - \alpha)\tilde{\gamma}^1 + \frac{\alpha}{n} \sum_{i=2}^n \tilde{\gamma}^i$ for some $0 < \alpha < 1$.

Consider the set

$$C = \left\{ c = \sum_i p_i H(\tilde{\gamma}^i) : \sum_i p_i \tilde{\gamma}^i = \mu, p_i > 0 \forall i \right\}.$$

We claim that this set contains a non-degenerate interval of c values. Indeed, one element of this set is $[(1 - \alpha) + \frac{\alpha}{n}]H(\tilde{\gamma}^1) + \frac{\alpha}{n} \sum_{i=2}^n H(\tilde{\gamma}^i)$. Additionally, we have $\tilde{\gamma}^1 = \sum_{i=2}^n \tilde{p}_i \tilde{\gamma}^i$ for some probability vector $\tilde{p} = \{\tilde{p}_i\}_{i=2}^n$, so $\frac{\alpha}{n} H(\tilde{\gamma}^1) + \sum_{i=2}^n [(1 - \alpha)\tilde{p}_i + \frac{\alpha}{n}]H(\tilde{\gamma}^i)$ is also in this set. The strict concavity of H implies that the former is strictly larger than the latter. By taking convex combinations of these two representations of μ we can get any c in between. Denote $\underline{c} = H(\mu) - \sup C$ and $\bar{c} = H(\mu) - \inf C$. Note that $0 < \underline{c} < \bar{c} < H(\mu)$.

Finally, consider problem $(\star\star)$ with $c \in (\underline{c}, \bar{c})$. Subsequently, by the definition of \underline{c} and \bar{c} , there are strictly positive $\{\tilde{P}(i)\}_{i=1}^n$ satisfying $\sum_i \tilde{P}(i)\tilde{\gamma}^i = \mu$ and $H(\mu) - \sum_i \tilde{P}(i)H(\tilde{\gamma}^i) = c$.

We claim that this distribution over posteriors is optimal. Indeed, this gives an expected utility of

$$\sum_i \tilde{P}(i) \max_{1 \leq j \leq n} \sum_{\omega} \tilde{\gamma}^i(\omega) u(j, \omega) = \sum_i \tilde{P}(i) \max_{1 \leq j \leq n} f_j(\tilde{\gamma}^i) \geq \sum_i \tilde{P}(i) f_i(\tilde{\gamma}^i) = H(\mu) - \sum_i \tilde{P}(i) H(\tilde{\gamma}^i) = c,$$

where the first equality is by the definition of u and the affinity of f_j , the inequality is obvious, the next equality follows from $f_i(\tilde{\gamma}^i) = f(\tilde{\gamma}^i)$, and the last equality is by construction of $\{\tilde{P}(i)\}$. On the other hand, for any feasible distribution over posteriors $\{P(i), \gamma^i\}_{i=1}^I$ we have

$$\sum_i P(i) \max_{1 \leq j \leq n} \sum_{\omega} \gamma^i(\omega) u(j, \omega) = \sum_i P(i) \max_{1 \leq j \leq n} f_j(\gamma^i) \leq \sum_i P(i) f(\gamma^i) = H(\mu) - \sum_i P(i) H(\gamma^i) \leq c,$$

where the first equality is again by the definition of u and the affinity of f_j , the inequality follows from $f_j \leq f$ for all j , and the last inequality follows from feasibility (the reduction of entropy must be at most c). Moreover, if $\gamma_i \notin \Gamma$ for some i , then $f_j(\gamma^i) < f(\gamma^i)$ for every j , so the first inequality is strict. This completes the proof. \square

Proof of Proposition 7. From Proposition 1 and Lemma 1 it immediately follows that $\bar{\lambda}$ is equal to the infimum of the set of λ 's for which an uninformative P is optimal in problem (\star) . Thus, to prove the proposition, it is enough to show that μ is an indifference point if and only if the set of such λ 's is empty.

We start by showing that if μ is not an indifference point then for λ large enough the optimal solution to (\star) is uninformative. For this proof, we view $\Delta(\Omega)$ as the unit simplex of \mathbb{R}^Ω endowed with the metric $d(\gamma, \gamma') = \max_{\omega} |\gamma(\omega) - \gamma'(\omega)|$. Let

$$K = \left\{ \gamma \in \Delta(\Omega) : \{a^*\} = \arg \max_a \sum_{\omega} \gamma(\omega) u(a, \omega) \right\}$$

be the set of beliefs at which a^* is the unique optimal action. Because K is relatively open in $\Delta(\Omega)$ and $\mu \in K$, there is $\delta > 0$ such that $\gamma \in K$ whenever $d(\gamma, \mu) < \delta$.

Let $M = \max_{a,b} \max_{\omega} \{u(a, \omega) - u(b, \omega)\} > 0$. Suppose that λ is large enough, so that $\exp\left(\frac{M}{\lambda}\right) < 1 + \delta$. If P is optimal for (\star) with λ , then for every $a, b \in B(P)$ and every ω , we have

$$\frac{\gamma^a(\omega)}{\gamma^b(\omega)} = \exp\left(\frac{u(a, \omega) - u(b, \omega)}{\lambda}\right) \leq \exp\left(\frac{M}{\lambda}\right) < 1 + \delta, \quad (\text{A1})$$

where the first equality is by (4), and the next inequality is by the definition of M . We thus get

$$d(\gamma^a, \gamma^b) = \max_{\omega} \left\{ \max \left\{ \gamma^b(\omega) \left(\frac{\gamma^a(\omega)}{\gamma^b(\omega)} - 1 \right), \gamma^a(\omega) \left(\frac{\gamma^b(\omega)}{\gamma^a(\omega)} - 1 \right) \right\} \right\} \leq \max_{\omega} \left\{ \max \left\{ \gamma^b(\omega) \delta, \gamma^a(\omega) \delta \right\} \right\} < \delta, \quad (\text{A2})$$

where in the first equality we used the fact that optimal posteriors are never on the boundary of $\Delta(\Omega)$ (implied by condition (4)), and the following inequality is by (A1).

Now, since $\mu = \sum_{b \in B(P)} P(b) \gamma^b$, we also have for every $a \in B(P)$ and every ω

$$|\mu(\omega) - \gamma^a(\omega)| \leq \sum_{b \in B(P)} P(b) |\gamma^b(\omega) - \gamma^a(\omega)|.$$

Taking the maximum over ω gives

$$d(\mu, \gamma^a) \leq \sum_{b \in B(P)} P(b) d(\gamma^a, \gamma^b) < \delta,$$

where the last inequality is from (A2). By the construction of δ this implies that $\gamma^a \in K$ for every $a \in B(P)$. Thus, by Remark A1, P induces a unique posterior and is therefore uninformative.

In the other direction, we now claim that, if μ is an indifference point, then an uninformative P can not be a solution of (\star) with any λ . Indeed, let a, a' be two optimal actions given belief μ . Consider \tilde{P} , as given by $\tilde{P}(a|\omega) = \tilde{P}(a'|\omega) = 0.5$ for every ω . Note that \tilde{P} is optimal among the set of uninformative P 's. However, we have $B(\tilde{P}) = \{a, a'\}$ and $\gamma^a = \gamma^{a'} = \mu$. Because a, a' are not duplicates, condition (4) can not hold for any $\lambda > 0$, implying that \tilde{P} is not optimal.¹⁸ \square

Appendix A.3. Proofs of Claims in the Example

We start with the following lemma.

Lemma A1. Fix a consideration set $B(P) \subseteq \{l, m, r\}$. The posteriors $\{\gamma^a\}_{a \in B(P)}$ satisfy (4) with some $\lambda > 0$ if and only if:

- (i) For $B(P) = \{l, r\}$: $\gamma^l = 1 - \gamma^r = \frac{1}{1 + \exp(\frac{3}{\lambda})}$.
- (ii) For $B(P) = \{l, m\}$: $\gamma^l = \frac{1}{1 + \exp(\frac{1}{\lambda}) + \exp(\frac{2}{\lambda})}$ and $\gamma^m = \frac{\exp(\frac{2}{\lambda})}{1 + \exp(\frac{1}{\lambda}) + \exp(\frac{2}{\lambda})}$.
- (iii) For $B(P) = \{l, m, r\}$: $\gamma^m = 0.5$, $\gamma^l = 1 - \gamma^r = \gamma^*$ and $\lambda = \lambda^*$.

Proof.

- (i) Condition (4) with $a = l$ and $b = r$ requires

$$\frac{\gamma^l}{\exp(\frac{-2}{\lambda})} = \frac{\gamma^r}{\exp(\frac{1}{\lambda})}$$

in state ω_1 and

$$\frac{1 - \gamma^l}{\exp(\frac{1}{\lambda})} = \frac{1 - \gamma^r}{\exp(\frac{-2}{\lambda})}$$

in ω_0 . It is immediate to verify that these two equations are equivalent to $\gamma^l = 1 - \gamma^r = \frac{1}{1 + \exp(\frac{3}{\lambda})}$.

- (ii) Similarly, condition (4) with $a = l$ and $b = m$ requires

$$\frac{\gamma^l}{\exp(\frac{-2}{\lambda})} = \frac{\gamma^m}{1}$$

in ω_1 and

$$\frac{1 - \gamma^l}{\exp(\frac{1}{\lambda})} = \frac{1 - \gamma^m}{1}$$

in ω_0 , and these two equations are satisfied if and only if the posteriors are as in the lemma.

- (iii) When all three actions are considered condition (4) is equivalent to the equations in both previous cases holding simultaneously. In particular, it implies that both $\gamma^l = \frac{1}{1 + \exp(\frac{3}{\lambda})}$ and $\gamma^l = \frac{1}{1 + \exp(\frac{1}{\lambda}) + \exp(\frac{2}{\lambda})}$. This pins down λ at λ^* , and consequentially the posteriors $\gamma^l, \gamma^m, \gamma^r$, as above.

\square

Case 1: $\mu = 0.5$

¹⁸ The same argument also implies the stronger result that if P is optimal for (\star) with some $\lambda > 0$, then none of the posteriors $\{\gamma^a\}_{a \in B(P)}$ is an indifference point.

Claim 10. *Solution of (\star): If $0 < \lambda < \lambda^*$ then $P(l) = P(r) = 0.5$ and $\gamma^l = 1 - \gamma^r = \frac{1}{1+\exp(\frac{3}{\lambda})}$. If $\lambda > \lambda^*$ then $P(m) = 1$ and $\gamma^m = 0.5$. If $\lambda = \lambda^*$ then any mixture (including degenerate) of the former two solutions is optimal.*

Proof. Suppose $\lambda < \lambda^*$, and we will show that $B(P) = \{l, r\}$ is optimal. From Lemma A1 (i), condition (4) is equivalent to $\gamma^l = 1 - \gamma^r = \frac{1}{1+\exp(\frac{3}{\lambda})}$. Condition (5) with $a = l$ and $b = m$ requires

$$\frac{\gamma^l}{\exp(\frac{-2}{\lambda})} * 1 + \frac{1 - \gamma^l}{\exp(\frac{1}{\lambda})} * 1 \leq 1.$$

Plugging in $\gamma^l = \frac{1}{1+\exp(\frac{3}{\lambda})}$ and rearranging, this condition becomes $\lambda \leq \lambda^*$, which holds by assumption. Condition (5), with $a = r$ and $b = m$, is identical, given that $\gamma^l = 1 - \gamma^r$.

Suppose now that $\lambda > \lambda^*$. We need to show that obtaining no information is optimal. Condition (4) is trivially satisfied, so we only need to check (5) for $a = m$ and $b = l, r$. For $b = l$, this gives

$$\frac{0.5}{1} \exp\left(\frac{1}{\lambda}\right) + \frac{0.5}{1} \exp\left(\frac{-2}{\lambda}\right) \leq 1,$$

which is equivalent to $\lambda \geq \lambda^*$. For $b = r$, we get the exact same condition.

Finally, when $\lambda = \lambda^*$ it follows from the last two paragraphs that both $B(P) = \{m\}$ and $B(P) = \{l, r\}$ are still optimal. Because the set of optimal distributions over posteriors is convex, it follows that any mixture of these solutions is optimal as well. Note that at $\lambda = \lambda^*$ the posteriors of the risky actions in the solution are given by $\gamma^l = 1 - \gamma^r = \gamma^*$. \square

Claim 11. *Solution of ($\star\star$): If $0 < c < H(0.5) - H(\gamma^*)$ then the solution is given by $P(l) = P(r) = \frac{c}{2(H(0.5) - H(\gamma^*))}$, $P(m) = 1 - \frac{c}{(H(0.5) - H(\gamma^*))}$, $\gamma^m = 0.5$, and $\gamma^l = 1 - \gamma^r = \gamma^*$. If $H(0.5) - H(\gamma^*) \leq c < H(0.5)$ then $P(l) = P(r) = 0.5$ and $\gamma^l = 1 - \gamma^r$ is determined by the equation $H(0.5) - H(\gamma^l) = c$.*

Proof. Suppose $0 < c < H(0.5) - H(\gamma^*)$. We show that $B(P) = \{l, m, r\}$ is optimal. From Lemma A1 (iii) it must be that $\gamma^m = 0.5$, $\gamma^l = 1 - \gamma^r = \gamma^*$, and that the value of the lagrange multiplier is $\lambda = \lambda^*$. Setting $P(l) = P(r) = \frac{c}{2(H(0.5) - H(\gamma^*))}$ and $P(m) = 1 - \frac{c}{(H(0.5) - H(\gamma^*))}$, we get that $P(l)\gamma^l + P(m)\gamma^m + P(r)\gamma^r = 0.5$. In addition, the expected reduction of entropy from the prior to the posteriors is given by

$$H(0.5) - \left[\frac{c}{2(H(0.5) - H(\gamma^*))} H(\gamma^*) + \frac{c}{2(H(0.5) - H(\gamma^*))} H(1 - \gamma^*) + \left(1 - \frac{c}{(H(0.5) - H(\gamma^*))}\right) H(0.5) \right] = c,$$

, so the budget constraint (6) binds. By Proposition 1, we are done.

Once $H(0.5) - H(\gamma^*) \leq c$ it is no longer possible that all three actions are considered, since, given the location of the posteriors, the budget constraint can not bind. We show that $B(P) = \{l, r\}$ is optimal. Define γ^l by the equation $H(0.5) - H(\gamma^l) = c$, and set $\gamma^r = 1 - \gamma^l$. Let $P(l) = P(r) = 0.5$. Subsequently, $P(l)\gamma^l + P(r)\gamma^r = \mu$ and the budget constraint binds. Set λ to solve $\gamma^l = \frac{1}{1+\exp(\frac{3}{\lambda})}$. Afterwards, condition (4) is satisfied by Lemma A1 (i). Furthermore, since $H(0.5) - H(\gamma^l) = c \geq H(0.5) - H(\gamma^*)$ we know that $\gamma^* \geq \gamma^l$, which implies that $\lambda \leq \lambda^*$. Thus, (5) holds for $a = l, r$ and $b = m$ in the same way as in the case $\lambda < \lambda^*$ of the previous claim. \square

Case 2: $\gamma^* < \mu < 0.5$

Claim 12. *Solution of (\star): When $\lambda < \lambda^*$ we have $B(P) = \{l, r\}$, $\gamma^l = 1 - \gamma^r = \frac{1}{1+\exp(\frac{3}{\lambda})}$, and the probabilities are set to satisfy $P(l)\gamma^l + P(r)\gamma^r = \mu$. At $\lambda = \lambda^*$ the solution is not unique, with $B(P) = \{l, r\}$, $B(P) = \{l, m\}$, and $B(P) = \{l, m, r\}$ all possible. When $\lambda^* < \lambda < \bar{\lambda}(\mu)$,*

$B(P) = \{l, m\}$, the posteriors are $\gamma^l = \frac{1}{1+\exp(\frac{1}{\lambda})+\exp(\frac{2}{\lambda})}$, $\gamma^m = \frac{\exp(\frac{2}{\lambda})}{1+\exp(\frac{1}{\lambda})+\exp(\frac{2}{\lambda})}$, and the probabilities are adjusted, so that $P(l)\gamma^l + P(m)\gamma^m = \mu$. Finally, for $\lambda \geq \bar{\lambda}(\mu)$, it is optimal to obtain no information, and the consideration set is either $\{l\}$ or $\{m\}$ when μ is below or above $1/3$, respectively.

Proof. The proof for $\lambda < \lambda^*$ is identical to the $\mu = 0.5$ case. Suppose that $\lambda^* < \lambda < \bar{\lambda}(\mu)$ and we prove that $B(P) = \{l, m\}$ is optimal. By Lemma A1 (ii), condition (4) is equivalent to $\gamma^l = \frac{1}{1+\exp(\frac{1}{\lambda})+\exp(\frac{2}{\lambda})}$ and $\gamma^m = \frac{\exp(\frac{2}{\lambda})}{1+\exp(\frac{1}{\lambda})+\exp(\frac{2}{\lambda})}$. We need to check condition (5) with $a = l, m$ and $b = r$. For $a = l$, this requires

$$\frac{\gamma^l}{\exp(\frac{-2}{\lambda})} \exp\left(\frac{1}{\lambda}\right) + \frac{1-\gamma^l}{\exp(\frac{1}{\lambda})} \exp\left(\frac{-2}{\lambda}\right) \leq 1,$$

, which boils down to $\lambda^* \leq \lambda$. For $a = m$, the condition is

$$\frac{\gamma^m}{1} \exp\left(\frac{1}{\lambda}\right) + \frac{1-\gamma^m}{1} \exp\left(\frac{-2}{\lambda}\right) \leq 1,$$

, which, again, is equivalent to $\lambda^* \leq \lambda$. Finally, we need that μ is in the interior of the convex hull of γ^l and γ^m . It is not hard to check that this is equivalent to $\lambda < \bar{\lambda}(\mu)$.

Note that at $\lambda = \lambda^*$ both solutions with $B(P) = \{l, r\}$ and with $B(P) = \{l, m\}$ are optimal, and therefore any mixture is optimal as well. This implies that $B(P) = \{l, m, r\}$ is also optimal.

Suppose now that $\lambda \geq \bar{\lambda}(\mu)$. We consider the case where $\mu > \frac{1}{3}$, so obtaining no information implies that $B(P) = \{m\}$ and $\gamma^m = \mu$ (the case $\mu < \frac{1}{3}$ in which $B(P) = \{l\}$ is similar). We need to check (5) with $a = m$ and $b = l, r$. These conditions are given by

$$\frac{\mu}{1} \exp\left(\frac{-2}{\lambda}\right) + \frac{1-\mu}{1} \exp\left(\frac{1}{\lambda}\right) \leq 1$$

and

$$\frac{\mu}{1} \exp\left(\frac{1}{\lambda}\right) + \frac{1-\mu}{1} \exp\left(\frac{-2}{\lambda}\right) \leq 1,$$

, respectively. For every $\mu > \frac{1}{3}$, the second inequality implies the first, and it is satisfied if and only if $\lambda \geq \bar{\lambda}(\mu)$. This completes the proof. \square

Claim 13. *Solution of (★): For $0 < c \leq H(\mu) - \left[\frac{1-2\mu}{1-2\gamma^*} H(\gamma^*) + \frac{2\mu-2\gamma^*}{1-2\gamma^*} H(0.5) \right]$ the consideration set is $B(P) = \{l, m\}$, and the distribution over posteriors is determined by the equations $\gamma^l(1-\gamma^l)^2 = \gamma^m(1-\gamma^m)^2$, $P(l)\gamma^l + P(m)\gamma^m = \mu$ and $H(\mu) - \left[P(l)H(\gamma^l) + P(m)H(\gamma^m) \right] = c$. For $H(\mu) - \left[\frac{1-2\mu}{1-2\gamma^*} H(\gamma^*) + \frac{2\mu-2\gamma^*}{1-2\gamma^*} H(0.5) \right] < c < H(\mu) - H(\gamma^*)$ the consideration set is $B(P) = \{l, m, r\}$, the posteriors are $\gamma^m = 0.5$, $\gamma^l = 1 - \gamma^r = \gamma^*$, and the probabilities $(P(l), P(m), P(r))$ adjust so that the expected posterior is equal to μ and the cost is equal to c . For $H(\mu) - H(\gamma^*) \leq c < H(\mu)$, the consideration set is $B(P) = \{l, r\}$, the posteriors are determined by $H(\mu) - H(\gamma^l) = c$ and $\gamma^l = 1 - \gamma^r$, and their probabilities by $P(l)\gamma^l + P(r)\gamma^r = \mu$.*

Proof. Fix c in the first range. The function $f(x) = x(1-x)^2$ is strictly increasing on $\left[\gamma^*, \frac{1}{3}\right)$, strictly decreasing on $\left(\frac{1}{3}, 0.5\right]$, and is equal to $\frac{1}{8}$ at both γ^* and 0.5 . It follows that, for each $\gamma^l \in \left[\gamma^*, \frac{1}{3}\right)$, there is a unique $\gamma^m \in \left(\frac{1}{3}, 0.5\right]$, such that $\gamma^l(1-\gamma^l)^2 = \gamma^m(1-\gamma^m)^2$, and that as γ^l increases the corresponding γ^m decreases. Therefore, given μ , there is a unique way to choose the posteriors γ^l, γ^m and the probabilities $P(l), P(m) = 1 - P(l)$, such that (1) $\gamma^l(1-\gamma^l)^2 = \gamma^m(1-\gamma^m)^2$; (2) $P(l)\gamma^l + P(m)\gamma^m = \mu$; and, (3)

$H(\mu) - [P(l)H(\gamma^l) + P(m)H(\gamma^m)] = c$.¹⁹ Now, define λ as the (unique) solution to $\gamma^l = \frac{1}{1+\exp(\frac{1}{\lambda})+\exp(\frac{2}{\lambda})}$, and note that this implies $\gamma^m = \frac{\exp(\frac{2}{\lambda})}{1+\exp(\frac{1}{\lambda})+\exp(\frac{2}{\lambda})}$. By Lemma A1 (ii) condition (4) holds. Finally, $\gamma^l \geq \gamma^*$ implies that $\lambda \geq \lambda^*$, so condition (4) holds in the same way as in the previous claim.

Moving on to $H(\mu) - [\frac{1-2\mu}{1-2\gamma^*}H(\gamma^*) + \frac{2\mu-2\gamma^*}{1-2\gamma^*}H(0.5)] < c < H(\mu) - H(\gamma^*)$, the distribution in the claim is optimal by Proposition 1 with $\lambda = \lambda^*$. The lower end of the range of c is obtained by the distribution with support γ^* and 0.5 only, while the upper end by the distribution with support γ^* and $1 - \gamma^*$ only.

When $H(\mu) - H(\gamma^*) \leq c < H(\mu)$ the proof is as in the previous claims. \square

Case 3: $0 < \mu \leq \gamma^*$

Claim 14. The solution of (\star): for $\lambda < \bar{\lambda}(\mu)$ we have $B(P) = \{l, r\}$, $\gamma^l = 1 - \gamma^r = \frac{1}{1+\exp(\frac{3}{\lambda})}$, and $(P(l), P(r))$ determined by $P(l)\gamma^l + P(r)\gamma^r = \mu$. For $\lambda \geq \bar{\lambda}(\mu)$, the solution is $B(P) = \{l\}$, i.e., no information.

Proof. Fix $\lambda < \bar{\lambda}(\mu)$ and set the posteriors as in the claim. Subsequently, (4) holds by Lemma A1 (i). As in the previous cases above, condition (5) with $a = l, r$ and $b = m$ is equivalent to $\lambda \leq \lambda^*$; for μ in the current range it is easy to check that $\bar{\lambda}(\mu) \leq \lambda^*$, so the condition holds by assumption. Finally, we need that $\mu \in (\gamma^l, \gamma^r)$, which is equivalent to $\lambda < \bar{\lambda}(\mu)$.

Assume now $\lambda \geq \bar{\lambda}(\mu)$. Set $B(P) = \{l\}$ and $\gamma^l = \mu$. Condition (5) with $a = l$ and $b = m, r$ requires

$$\frac{\mu}{\exp(\frac{-2}{\lambda})} * 1 + \frac{1-\mu}{\exp(\frac{1}{\lambda})} * 1 \leq 1$$

and

$$\frac{\mu}{\exp(\frac{-2}{\lambda})} \exp\left(\frac{1}{\lambda}\right) + \frac{1-\mu}{\exp(\frac{1}{\lambda})} \exp\left(\frac{-2}{\lambda}\right) \leq 1,$$

respectively. For any μ in the range, the second inequality implies the first and it is satisfied if and only if $\lambda \geq \bar{\lambda}(\mu)$. \square

Claim 15. The solution of ($\star\star$): For any $0 < c < H(\mu)$ the consideration set is $B(P) = \{l, r\}$ and $\gamma^l = 1 - \gamma^r$. The posterior γ^l is determined by the equation $H(\mu) - H(\gamma^l) = c$, and the probability $P(l)$ is then determined by the equation $P(l)\gamma^l + P(r)\gamma^r = \mu$.

Proof. Fix c and define γ^l by $H(\mu) - H(\gamma^l) = c$. Let $\gamma^r = 1 - \gamma^l$ and set λ to solve $\gamma^l = \frac{1}{1+\exp(\frac{3}{\lambda})}$. The fact that $H(\mu) - H(\gamma^l) = c > 0$ implies that $\gamma^l < \mu$, which in turn implies that $\mu \in (\gamma^l, \gamma^r)$ as well as that $\lambda < \bar{\lambda}(\mu)$. The optimality of these posteriors now follows in the same way as in the previous claim. \square

References

1. Sims, C.A. Stickiness. *Carnegie Rochester Conf. Ser. Public Policy* **1998**, *49*, 317–356. [\[CrossRef\]](#)
2. Sims, C.A. Implications of rational inattention. *J. Monet. Econ.* **2003**, *50*, 665–690. [\[CrossRef\]](#)
3. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [\[CrossRef\]](#)
4. Caplin, A.; Dean, M. *Behavioral Implications of Rational Inattention with Shannon Entropy*; National Bureau of Economic Research: Cambridge, MA, USA, 2013.
5. Le Treust, M.; Tomala, T. Persuasion with limited communication capacity. *J. Econ. Theory* **2019**, *184*, 104940. [\[CrossRef\]](#)

¹⁹ Note that the bound on c is obtained by setting $\gamma^l = \gamma^*$ and $\gamma^m = 0.5$.

6. Dean, M.; Neligh, N. Experimental Tests of Rational Inattention. Working Paper. 2019. Available online: http://www.columbia.edu/~md3405/Working_Paper_21.pdf (accessed on 29 December 2020).
7. Maćkowiak, B.; Matějka, F.; Wiederholt, M. Rational Inattention: A Review. Working Paper. 2020. Available online: <http://home.cerge-ei.cz/matejka/RIsurvey.pdf> (accessed on 29 December 2020).
8. Matějka, F.; McKay, A. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *Am. Econ. Rev.* **2015**, *105*, 272–298. [[CrossRef](#)]
9. Caplin, A.; Dean, M.; Leahy, J. Rational inattention, optimal consideration sets, and stochastic choice. *Rev. Econ. Stud.* **2019**, *86*, 1061–1094. [[CrossRef](#)]
10. Caplin, A.; Dean, M.; Leahy, J. *Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy*; National Bureau of Economic Research: Cambridge, MA, USA, 2017.
11. Matějka, F. Rigid pricing and rationally inattentive consumer. *J. Econ. Theory* **2015**, *158*, 656–678. [[CrossRef](#)]
12. Steiner, J.; Stewart, C.; Matějka, F. Rational inattention dynamics: Inertia and delay in decision-making. *Econometrica* **2017**, *85*, 521–553. [[CrossRef](#)]
13. De Oliveira, H. Axiomatic Foundations for Entropic Costs of Attention. Working Paper. 2014. Available online: https://6964c30c-a4a4-4464-a435-9021be4b7ccd.filesusr.com/ugd/21e9a6_dab3d6bc526d40fa86894de99e3a48ba.pdf (accessed on 29 December 2020).
14. Fulton, C. Mechanics of Linear Quadratic Gaussian Rational Inattention Tracking Problems. Working Paper. 2017. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3065488# (accessed on 29 December 2020).
15. Doval, L.; Skreta, V. Constrained Information Design: Toolkit. Working Paper. 2018. Available online: <https://arxiv.org/abs/1811.03588> (accessed on 29 December 2020).
16. Boleslavsky, R.; Kim, K. Bayesian Persuasion and Moral Hazard. Working Paper. 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2913669 (accessed on 29 December 2020).
17. Caplin, A.; Dean, M. Revealed preference, rational inattention, and costly information acquisition. *Am. Econ. Rev.* **2015**, *105*, 2183–2203. [[CrossRef](#)]
18. Dewan, A.; Neligh, N. Estimating information cost functions in models of rational inattention. *J. Econ. Theory* **2020**, *187*, 105011. [[CrossRef](#)]
19. Cheremukhin, A.; Popova, A.; Tutino, A. A theory of discrete choice with information costs. *J. Econ. Behav. Organ.* **2015**, *113*, 34–50. [[CrossRef](#)]
20. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 1970.