Neighborhood-Based Information Costs*

Benjamin Hébert † Michael Woodford ‡
Stanford University Columbia University

January 14, 2021

Abstract

We derive a new cost of information in rational inattention problems, the neighborhood-based cost functions, starting from the observation that many settings involve exogenous states with a topological structure. These cost functions are uniformly posterior-separable and capture notions of perceptual distance. This second property ensures that neighborhood-based costs, unlike mutual information, make accurate predictions about behavior in perceptual experiments. We compare the implications of our neighborhood-based cost functions with those of the mutual information in a series of applications: perceptual judgments; the general environment of binary choice; regime-change games; and linear-quadratic-Gaussian settings.

*The authors would like to thank Alex Bloedel, Mark Dean, Sebastian Di Tella, Mira Frick, Xavier Gabaix, Matthew Gentzkow, Emir Kamenica, Divya Kirti, Jacob Leshno, Stephen Morris, Pietro Ortoleva, José Scheinkman, Ilya Segal, Ran Shorrer, Joel Sobel, Harald Uhlig, Miguel Villas-Boas, Ming Yang, Weijie Zhong, the editor (Mikhail Golosov) and three anonymous referees, and seminar and conference participants at the Cowles Theory conference, 16th SAET Conference, Barcelona GSE Summer Conference on Stochastic Choice, Stanford GSB research lunch, the 2018 ASSA meetings, UC San Diego, and UC Berkeley for helpful discussions on this topic, and the NSF for research support. We would particularly like to thank Doron Ravid and Philipp Strack for discussing an earlier version of the paper. Portions of this paper circulated previously as the working papers "Rational Inattention with Sequential Information Sampling" and "Information Costs and Sequential Information Sampling," and appeared in Benjamin Hébert's Ph.D. dissertation at Harvard University. All remaining errors are our own.

†Hébert: Stanford University. Email: bhebert@stanford.edu.

‡Woodford: Columbia University. Email: mw2230@columbia.edu.

1 Introduction

In models of rational inattention (proposed by Christopher Sims and surveyed in Sims (2010)), a decision maker (DM) chooses her action based on a signal that provides only an imperfect indication of the true state. The information structure that generates this signal is optimal, in the sense of allowing the best possible state-contingent action choice, net of a cost of information. In Sims' theory, the cost of any information structure is proportional to the mutual information between the true state of the world and the signals generated by that information structure.

It is not obvious, though, that the theorems that justify the use of mutual information in communications engineering (Cover and Thomas (2012)) provide a warrant for using it as a cost function in a theory of attention allocation, either in the case of economic decisions or that of perceptual judgments. Moreover, the mutual-information cost function has implications that are unappealing on their face, and that seem inconsistent with evidence on the nature of sensory processing.[1]

We propose a more general family of information costs, the neighborhood-based cost functions. Cost functions in this family have two particular properties that we view as desirable. First, they can be viewed as summarizing the results of a process of sequential evidence accumulation. Second, these information costs can capture the idea that certain pairs of states are easy to distinguish, whereas others are difficult to distinguish. Our interest in both of these properties is motivated by empirical evidence about the nature of sensory processing, discussed further below. The second property, in particular, allows neighborhood-based cost functions to avoid some of the problematic implications of the mutual-information cost function.

The neighborhood-based cost functions differ from mutual information because mutual information imposes a type of symmetry across different states of nature, so that it is equally difficult to distinguish between any two states that are equally probable ex ante. This implies that under an optimal information structure, actions differ across states only to the extent that the associated payoffs differ across those states, and action probabilities jump discontinuously when payoffs jump. An ex-

[1] See, e.g., Woodford (2012), Caplin and Dean (2013), Dewan and Neligh (2020), Caplin et al. (2019), and Dean and Neligh (2019).

tensive experimental literature in psychophysics finds that subjects' probabilities of making perceptual judgments (the action) vary continuously with changes in the stimulus magnitude (the state), even when subjects are rewarded based on whether the magnitude is greater or smaller than some threshold (generating a discrete jump in payoffs).[2] The sigmoid functions that describe subjects' response frequencies in these experiments are known as "psychometric functions." We show that predictions of rational inattention with a neighborhood-based cost function (unlike mutual information) can match the key properties of these psychometric functions.

It might be thought that the continuity of the psychometric functions measured in perceptual experiments should not be relevant in many economic settings, on the ground that the available information will often be symbolic rather than perceptual. Yet there is evidence that even numerical magnitudes that are presented using number symbols are given a "semantic" representation in the brain (indicating how large the quantity is) which is imprecise, so that numbers representing similar quantities are not accurately distinguished, and that this "approximate number system" is drawn upon when judgments are made without explicit arithmetic calculations.[3] A classic experiment requires subjects to respond quickly whether a two-digit number presented on a screen is larger or smaller than a particular threshold (say, 65). Responses are slower and more mistakes are made when the number presented is closer to the threshold, rather a number that is either much smaller or much larger; and this is not simply a matter of whether the first digit of the presented number is the same as that of the threshold. Thus the accuracy with which responses can vary for different numbers seems to depend on how similar the numbers are in their meaning, and not just the similarity of their visual appearance.

Response frequencies described by psychometric functions are also observed in laboratory experiments involving regime change games, of the sort studied by Morris and Shin (1998). In a monotone equilibrium of these games (described in more detail below), the regime will change if and only if the fundamental state exceeds a threshold. As a result, from the perspective of an individual DM, the net payoff of "investing" (the action) jumps discretely at this threshold. Consistent with the

[2] See for example Figure 1 in Woodford (2020) or Figure 10.1A of Shadlen et al. (2007).

[3] See Khaw et al. (forthcoming), section 1.2, for references and further discussion.

aforementioned evidence from perceptual experiments, the observed frequency in lab experiments of investing resembles a psychometric function (see, for example, Figure 4 of Szkup and Trevino (2020)).[4] Moreover, whether response frequencies vary smoothly or discontinuously when payoffs jump turns out to have economically important ramifications in the context of regime change games.[5]

In a regime-change game, if the state can be observed with perfect precision by all agents, and this is common knowledge, the game has a large multiplicity of equilibria, so that the timing of a run is arbitrary (see e.g. Obstfeld (1986)), and the probability of a run need not even be a monotone function of the fundamental. However, a robust result that emerges from studies of bank runs (see, e.g., Calomiris and Mason (1997)) is that failure is more likely for banks with worse fundamentals.

The global-games approach of Morris and Shin (1998) instead proposes a model of imprecise perception of fundamentals that implies a unique equilibrium in the regime change game, with a decision rule described by a continuous and monotone psychometric function. But this celebrated result depends in turn on the nature of agents' imprecise perceptions of the state. When Yang (2015) models agents' information as the information structure predicted by a mutual-information cost function, he again finds a large multiplicity of equilibria, despite imprecise observation of the state and a failure of common knowledge. As stressed by Morris and Yang (2019), the critical issue is not whether agents can perfectly observe the state, but whether they will "invest" with a probability that jumps discretely when the fundamentals exceed a threshold. Thus, because the neighborhood-based cost functions predict response frequencies that resemble psychometric functions, they will also generate a unique equilibrium when used as the information cost in a regime-change game with endogenous information acquisition.

Motivated by these issues, we consider the properties that a plausible infor-

[4] Other authors, including Heinemann et al. (2009), also study regime-change games in a laboratory setting and find results consistent with choice probabilities described by psychometric curves.

[5] In theory, these differences can be observed in aggregate data. In a model of speculative attacks or bank runs, if every agent's behavior was described by a psychometric curve, the total run size would also resemble a psychometric curve. Conversely, if every agent's behavior was described be a step function at the same threshold, the total run size would also be a step function at that threshold. However, it is difficult to distinguish between these two cases in real-world data because of the difficulty of measuring the relevant fundamental state precisely.

mation cost function should satisfy. We restrict attention to uniformly posterior-separable (UPS) cost functions, motivated by the results of Hébert and Woodford (2019) (who justify UPS costs as arising from sequential information sampling),[6] the related theoretical justifications for UPS costs in Bloedel and Zhong (2020), and the experimental evidence of Dean and Neligh (2019).

We then introduce a specific family of UPS cost functions, the neighborhood-based cost functions. With these costs, information structures are more costly the greater the extent to which they discriminate between intrinsically similar states of the world (states that share a "neighborhood"). The dependence on a concept of intrinsic similarity between states (the "neighborhood structure") distinguishes these cost functions from mutual information. Neighborhood structures are closely related to the idea that the state space is equipped with a topology; that is, states of nature are not unordered sets.[7]

We derive our family of neighborhood-based cost functions from two assumptions that connect the topology of the state space to the cost function, intended to capture the idea that it is difficult to discriminate between nearby states. Given a set of neighborhoods that cover the state space, these two assumptions plus the assumption of uniform posterior separability define the class of neighborhood-based cost functions. Within this class, we describe specific neighborhood-based cost functions that differ in terms of their curvature within each neighborhood, and show that this curvature governs the elasticity of information acquisition to incentives.

We specialize the neighborhood-based cost functions to a particularly useful case, in which the states can be ordered on a line. We extend our analysis of this case to allow for a continuum of states, as in many economic models (such as the regime-change game mentioned above), and show that the limit of the neighborhood-based cost function for this limiting neighborhood structure is defined by an integral of the Fisher information over the state space. The information cost is thus the average

[6] As discussed in Fehr and Rangel (2011), a large literature in psychology and neuroscience has argued that data on both the frequency of perceptual errors and the frequency distribution of response times can be explained by models of sequential sampling; hence we view it as desirable to "micro-found" information costs using a sequential sampling process.

[7] Our definition allows however for a trivial neighborhood structure, in which all states belong to a single neighborhood. It is in this sense that the mutual information cost function remains a special case of our family of neighborhood-based cost functions.

value of a local measure of the discriminability of nearby states. This measure can be used instead of mutual information in almost any context in which the state space corresponds to a line or circle. We further extend this result to multi-dimensional state spaces, such as when states correspond to a vector of real numbers.

Lastly, we apply the Fisher information cost function in two canonical settings of rational inattention: binary choice and a multi-variate linear-quadratic-Gaussian setting. These environments cover many of the existing applications of rational inattention (for a survey, see Mackowiak et al. (2020)); our results show that the Fisher information cost can tractably replace mutual information in these settings. We also discuss perceptual experiments and regime-change games in more detail, building on our binary choice results to show that the Fisher information cost predicts psychometric function response frequencies in these settings. In Appendix Section B, we illustrate how our results on binary choice can be incorporated into more complex problems by considering security design with adverse selection (Yang, 2020).

In the linear-quadratic-Gaussian case, we find that the average Fisher information cost function shares a convenient prediction with mutual information: optimal signals will have a Gaussian structure. However, even in this case, interesting differences exist between the implications of the two cost functions. In the case of a single-dimensional state space, the Fisher information cost function implies a cost that is linear in the precision of a Gaussian signal; such costs have previously been used in the literature (e.g. Van Nieuwerburgh and Veldkamp (2010); Myatt and Wallace (2011)), and our results provide a justification for this functional form. In contrast, mutual information implies a cost proportional to the log of the precision, which generates different predictions in the applications discussed by those authors. In the case of a multi-dimensional state space, additional differences emerge. In a setting where only one dimension of the state space is payoff-relevant, we show that with mutual information, the DM receives a signal only about that dimension, whereas with Fisher information, the DM receives a signal that maximally covaries with the payoff-relevant dimension. Hébert and La'O (2020) demonstrate that this distinction leads to different predictions about efficiency and non-fundamental volatility in games with rationally inattentive agents.

Several other papers in the literature propose alternatives to the mutual infor-

mation cost function. Caplin et al. (2019) analyze the class of UPS cost functions, and direct particular attention to a class of UPS cost functions based on Tsallis entropy. These cost functions lack a notion of distance between states, but deviate from mutual information in other respects. Pomatto et al. (2020) are motivated by concerns similar to ours, and derive a different family of cost functions from axioms related to the cost of repeated experiments. These cost functions are not UPS, but are similar to our neighborhood-based cost functions in that they can also capture a notion of distance between states. The axioms of Pomatto et al. (2020) relate to the cost of performing multiple, independent experiments and to "diluted" versions of an experiment, whereas our axioms describe the relationship between the topology of the state space and information costs. Bloedel and Zhong (2020) study UPS cost functions that exhibit the "constant marginal costs" property studied by Pomatto et al. (2020), and characterize the resulting "total information" cost functions; in Appendix Section A we discuss the intersection of the neighborhood-based information costs and total information costs.

In section 2, we define a general class of static rational inattention problems, emphasizing the case of a UPS cost function. In section 3, we then state the additional assumptions that define the class of neighborhood-based cost functions, and offer a general characterization result for these functions. We then discuss a variety of more specific examples of such functions, with further attractive features, including the average Fisher information cost function for the special case when the state space is the real line. We next show how neighborhood-based cost functions can be used in a series of applications in section 4. In section 5 we conclude.

2 Static Rational Inattention Problems

We begin by defining the class of static rational inattention problems with which we are concerned. Let $x \in X$ be the underlying state of the nature, and $a \in A$ be the action taken by the decision maker (DM). Here A and X are finite sets, and the DM's utility from taking action a in state x is $u_{a,x}$.

The DM does not know the state $x \in X$, but can learn about which states are more or less likely. The DM begins with prior beliefs $q_0 \in \mathscr{P}(X)$, where $\mathscr{P}(\cdot)$

denotes the probability simplex on a set. The DM's decision is based on additional information, the nature of which is specified by a "signal structure," consisting of a signal alphabet S (a set) and a conditional probability for each state x of each signal, $p = \{p_x \in \mathscr{P}(S)\}_{x \in X}$. The signal structure p generates, under the prior q_0, an unconditional probability of each signal, $\pi_s(p, q_0)$. After receiving a signal $s \in S$, the DM will hold posterior beliefs $q_s(p, q_0)$, defined by Bayes' rule. To simplify notation, we assume S is finite, but nothing depends on this assumption.

Based on her posterior beliefs, the DM chooses an action $a \in A$. Define $\hat{u}$: $\mathscr{P}(X) \to \mathbb{R}$ as the utility when taking an optimal action given posteriors beliefs q,

$$\hat{u}(q) = \max_{a \in A} \sum_{x \in X} u_{a,x} q_x,$$

where q_x is the probability under q of state $x \in X$. In what follows, we will treat the beliefs $q \in \mathscr{P}(X)$ as vectors in $\mathbb{R}^{|X|}$.

Signal structures are costly in utility terms. Let $C(p, q_0; S) : \mathscr{P}(S)^{|X|} \times \mathscr{P}(X) \to \mathbb{R}$ be the cost of choosing a signal structure p and alphabet S, given initial prior q_0. The standard static rational inattention problem, given the signal alphabet S,[8] is

$$\max_{\{p_x \in \mathscr{P}(S)\}_{x \in X}} \sum_{s \in S} \pi_s(p, q_0) \hat{u}(q_s(p, q_0)) - \theta C(p, q_0; S), \tag{1}$$

where $\theta > 0$ parameterizes the cost of information. This endogenizes the information available to the DM. The problem can also be written as a choice over signal probabilities π_s and posteriors q_s; for any π_s and q_s such that $\sum_{s \in S} \pi_s q_s = q_0$, there is a unique signal structure p such that $\pi_s = \pi_s(p, q_0)$ and $q_s = q_s(p, q_0)$.

In the classic formulation of Sims, a problem of this kind is considered in which the cost function $C(p, q; S)$ is given by the mutual information between the signal and the state. Mutual information can be defined using Shannon's entropy measure,

$$H^{Shannon}(q) \equiv -\sum_{x \in X} q_x \ln(q_x). \tag{2}$$

[8]The full problem includes a choice over the signal alphabet S. A standard result, which will hold for all of the cost functions we study, is that $|S| = |A|$ is sufficient.

Shannon's entropy can be used to define a measure of the degree to which each posterior q_s differs from the prior q_0, the Kullback-Leibler (KL) divergence,

$$D_{KL}(q_s||q_0) \equiv H^{Shannon}(q_0) - H^{Shannon}(q_s) + (q_s - q_0)^T \cdot H_q^{Shannon}(q_0), \quad (3)$$

where $H_q^{Shannon}$ denotes the gradient of Shannon's entropy. Mutual information is the expected value of the KL divergence over possible signals,

$$C^{MI}(p, q_0; S) \equiv \sum_{s \in S} \pi_s(p, q_0) D_{KL}(q_s(p, q_0)||q_0). \quad (4)$$

Mutual information provides a measure of the degree to which the signal changes what the DM believes about the state, on average. Mutual information is not, however, the only possible measure of the informativeness of an information structure, or the only plausible cost function for a static rational inattention problem.

A more general class of cost functions, which includes mutual information, are the UPS cost functions. These cost functions can all be written as

$$C^{UPS}(p, q_0; S) \equiv \sum_{s \in S} \pi_s(p, q_0) D_H(q_s(p, q_0)||q_0),$$

where D_H is a Bregman divergence, itself defined by a convex function H,

$$D_H(q_s||q) = H(q_s) - H(q) - (q_s - q)^T \cdot H_q(q). \quad (5)$$

The Kullback-Leibler divergence, for example, is a Bregman divergence (see (3)), with an entropy function equal to the negative of Shannon's entropy.

Any differentiable convex function H defines a Bregman divergence. For notational purposes, we define H on $\mathbb{R}_+^{|X|}$ instead of $\mathscr{P}(X)$. That is, we work with non-negative vectors that may not sum to one. Given a function defined on $\mathscr{P}(X)$, we extend it to $\mathbb{R}_+^{|X|}$ by assuming that the function is homogenous of degree one.

At this point, we have defined the static rational inattention problem and the UPS cost functions. Before proceeding, we discuss the motivating examples of regime change games and perceptual experiments. These examples feature a jump in the relative utility of the two actions at a particular location in the state space.

Example 1. Suppose the DM is an agent in a "regime change game" such as the speculative currency attack game studied by Morris and Shin (1998). The state $x \in X \subseteq \mathbb{R}$ is the exogenous fundamental (e.g. the quantity of currency reserves in the speculative attack game). The DM can choose to invest or not, $A = \{invest, not\}$. If a fraction $l \geq 1 - x$ of agents invest, the regime will change (e.g. the currency peg will break in the speculative attack game). In a monotone equilibrium of this game, l is an increasing function of x and the regime will change if and only if $x \geq x^*$ for some $x^* \in \mathbb{R}$. For an individual DM, in such an equilibrium the payoff of investing depends on both a transaction cost $t \in (0,1)$ and on whether the regime changes,

$$u_{invest,x} = \begin{cases} 1-t & x \geq x^*, \\ -t & x < x^*, \end{cases}$$

while the payoff of not investing is normalized to zero, $u_{not,x} = 0$ for all $x \in X$.

Example 2. Suppose the DM is participating in an experiment intended to measure how well people can distinguish among sensory stimuli that differ with respect to a particular feature. The states $X \subseteq \mathbb{R}$ represent different values of this feature in the stimuli that may be presented to the DM, who is asked to classify the stimulus as one of two types (L or R); R is the correct answer if and only if $x \geq x^*$ for some $x^* \in \mathbb{R}$. An example is the kind of experiment discussed in Shadlen et al. (2007), in which subjects are rewarded for correctly classifying the dominant direction of motion for a field of moving dots. In this case, x is a measure of "motion strength" that varies between -1 (when all dots move to the left) and +1 (when all move to the right), with $x = 0$ corresponding to no coherent motion in either direction. The DM is rewarded for correctly classifying the stimulus, $u_{L,x} = \mathbf{1}\{x < x^*\}$ and $u_{R,x} = \mathbf{1}\{x \geq x^*\}$, where in this example $x^* = 0$.

These examples share two key features in common. First, the net utility of *invest* or R ($u_{invest,x} - u_{not,x}$ or $u_{R,x} - u_{L,x}$) is a step function that jumps from a negative value to a positive value at the threshold x^*. Second, the states (the quantity of currency reserves or the motion strength) are naturally represented by numbers on a line. In these examples, the ideal information structure p for the DM is one that

sharply discriminates between states based on whether $x < x^*$ or $x \geq x^*$. Such a signal may or may not be costly, depending on the properties of the information cost function C. With the mutual information, a signal structure in which the probability of receiving a given signal depends only on whether $x < x^*$ or $x \geq x^*$ has a finite cost, and is in fact the optimal signal structure. That is, the mutual information cost function predicts both a sharp discontinuity in response frequency at the threshold x^* and completely flat response frequencies on the domains $x < x^*$ and $x > x^*$.[9]

However, the choice frequencies observed in perceptual experiments and in experiments on regime change games do not jump discretely, and instead resemble psychometric functions. These functions typically exhibit several key properties. First, the frequency of $a = R$ is strictly increasing in x, approaching a flat asymptote only for very high or low values of x. Second, the slope of the response frequency is highest (but still finite) for intermediate values of x, typically around the threshold x^*. This second property is equivalent to observing that the frequency of $a = R$ is generally convex for low values of x and concave for high values of x. A common functional form used in the psychophysics literature (see, e.g., Wichmann and Hill (2001)) is $p_R(x) = p_L + (p_H - p_L)F(x)$, where F is the CDF of a logistic or Gaussian distribution and (p_L, p_H) are the asymptotes of the response frequency.

Our goal is to construct a family of information cost functions that predict psychometric-curve response frequencies in the binary-choice setting. We derive these cost functions, which we call the neighborhood-based cost functions, from primitive assumptions that are meant to capture the notion that it is difficult for DMs to sharply discriminate between nearby states in the state space.

3 Neighborhood-Based Cost Functions

In this section, we define the neighborhood-based cost functions. For this section only, we treat the state space X as part of the definition of the cost function, and focus on how cost functions defined on different state spaces can be related to each

[9]The aforementioned experiments of Heinemann et al. (2009) and Szkup and Trevino (2020) study games that differ slightly from our example, because $u_{invest,x}$ depends on x if $x \geq x^*$. Consequently, mutual information predicts flat probabilities only on the $x < x^*$ domain in these games.

other. That is, in this section only, we write $C(p, q_0; S, X)$ instead of $C(p, q_0; S)$.

Motivated by the theoretical results of Hébert and Woodford (2019), Bloedel and Zhong (2020), and Caplin et al. (2019), and the experimental evidence of Dean and Neligh (2019), we restrict attention to cost functions in the UPS family:

Assumption 1. *The cost function $C(p, q_0; S, X)$ is uniformly posterior-separable, and the associated H function is continuously twice-differentiable.*

There are many UPS cost functions, and they make different predictions about behavior. Our goal is to justify particular choices within the UPS family. To make progress, we begin by observing that, in many problems, the state space X has a structure. That is, some states are similar in a way that others are not.

To capture this idea, we will assume that X is a finite subset of a metric space $(\mathscr{X}, d)$, and suppose that the cardinality of $\mathscr{X}$ is at least as great at the cardinality of the real numbers.[10] Now suppose we are given a minimal point-finite open cover of X (i.e. a finite set of open neighborhoods that cover X, such that if any one neighborhood were removed, the neighborhoods would no longer cover X). Let us denote this collection of neighborhoods by $\mathscr{N}$, and let these neighborhoods be indexed by $i \in \mathscr{I}$. These neighborhoods are intended to represent regions in which it is difficult to discriminate. Each neighborhood $N_i \in \mathscr{N}$ is a subset of $\mathscr{X}$, and we will use the notation $X_i \equiv X \cap N_i$ to denote that set of states in neighborhood N_i. Except where it would cause confusion, we will also refer the sets X_i as neighborhoods.

The question is how to connect these neighborhoods with the cost function $C(\cdot)$. Intuitively, the neighborhoods define the sets of points that are difficult to distinguish. If there is no neighborhood in $\mathscr{N}$ that contains some $x, x' \in X$, it should be easy for the DM to distinguish between x and x', whereas if those states do share a neighborhood, it should be costly to distinguish them. In the context of a static rational inattention problem, the DM is distinguishing between x and x' if she receives a different distribution of signals conditional on x than conditional on x'.

To operationalize this idea, consider three different signal structures, p, p', and p''. The signal structure p discriminates between a state x and all other states,

[10] The metric d will play no role in our analysis. Our key assumption is the existence of a point finite open cover of X; an easy-to-state sufficient condition is that the space $\mathscr{X}$ be metrizable. We would like to thank Harald Uhlig for a helpful discussion on this point.

meaning that the conditional distributions of signals conditional on any state except x are identical under p. Formally,

$$p_{x''} = \begin{cases} r & x'' \neq x \\ r' & x'' = x, \end{cases} \tag{6}$$

for some $r, r' \in \mathscr{P}(S)$ with $r \neq r'$. Similarly, suppose that p' discriminates between x' and all other states. That is, let $p'_{x''} = r$ for $x'' \neq x'$ and $p'_{x'} = r'$. Let p'' be a signal structure that discriminates between $\{x, x'\}$ and all other states,

$$p''_{x''} = \begin{cases} r & x'' \notin \{x, x'\} \\ r' & x'' \in \{x, x'\}. \end{cases} \tag{7}$$

The key difference between p'' and the signal structures p and p' is that the former does not discriminate between x and x', whereas the latter structures do.

By the above logic, if x and x' share a neighborhood in $\mathscr{N}$, p and p' should be more costly than p'', because they discriminate between nearby states whereas p'' does not. Conversely, if x and x' do not share a neighborhood in $\mathscr{N}$, it is easy to distinguish between them, and p'' should be as costly as p and p'. Intuitively, what is costly is distinguishing x from its neighboring states and x' from its neighboring states, and since p'' does both these things it should be as costly as if they were done separately. We express this logic more formally in the assumption below.

Assumption 2. *Let $x, x' \in X$ be distinct states in the support of $q_0 \in \mathscr{P}(X)$, and let p, p', and p'' be defined as in equations (6) and (7), with $r \neq r'$. If there exists a neighborhood $N_i \in \mathscr{N}$ with $\{x, x'\} \subseteq N_i$, then*

$$C(p'', q_0; S, X) < C(p, q_0; S, X) + C(p', q_0; S, X).$$

If no such neighborhood exists, then

$$C(p'', q_0; S, X) = C(p, q_0; S, X) + C(p', q_0; S, X).$$

Figure 1 illustrates this assumption with an example neighborhood structure.

In general, a state x can be contained in multiple neighborhoods in $\mathscr{N}$. Suppose, for example, that a state x is contained in the neighborhoods N_1 and N_2. We interpret this situation as one in which discriminating between x and all other states is difficult both because it discriminates between x and the other states in N_1 and because it discriminates between x and a different (but possibly overlapping) set of states in neighborhood N_2. Our next assumption states that if x belongs the neighborhoods $(N_1,\ldots,N_k)$, this situation is equivalent to one in which x is split into k new distinct states, $(x_1,\ldots,x_k) \in \mathscr{X} \setminus X$, with $x_j \in N_j$ for each $j \in \{1,\ldots,k\}$, but $x_j \notin N$ for any $N \in \mathscr{N} \setminus \{N_j\}$.[11]

Let X' be the split space, $X' = (X \setminus \{x\}) \cup \{x_1,\ldots,x_k\} \subset \mathscr{X}$, and let $\phi : X' \to X$ be the surjection (many-to-one mapping) that associates each element of the split space X' with the corresponding elements of the original space X. That is, $\phi(x'') = x''$ for any $x'' \notin \{x_1,\ldots,x_k\}$ and $\phi(x'') = x$ for any $x'' \in \{x_1,\ldots,x_k\}$. Note that for any neighborhood N, the mapping ϕ establishes a one-to-one correspondence between the elements of X' belonging to N and the elements of X belonging to N.

For any prior $q \in \mathscr{P}(X)$, we can define an associated measure $q' \in \mathbb{R}_+^{|X'|}$ by $q'_{x''} = q_{\phi(x'')}$ for all x" $\in X'$. Note that the measure q' will generally not have unit mass,[12] but that the total mass assigned to the elements of any neighborhood $N \in \mathscr{N}$ is the same under the measures q and q'. Similarly, for any signal structure $p \in \mathscr{P}(S)^{|X|}$, we define $p' \in \mathscr{P}(S)^{|X'|}$ by $p'_{x''} = p_{\phi(x'')}$ for each $x'' \in X'$.

Our assumption is that the cost of discriminating between x and the other states in X is equal to the cost of discriminating between $\{x_1,\ldots,x_k\}$ and the other states in X'. This assumption captures the idea that learning about whether the state x has occurred is costly because it requires discriminating between x and the other states in each of the neighborhoods $(N_1,\ldots,N_k)$.

Assumption 3. *Suppose that some state $x \in X$ is contained in the neighborhoods $(N_1,\ldots,N_k)$ in $\mathscr{N}$, for some $k \geq 1$, and let $\{x_1,\ldots,x_k\}$, X', and $\phi : X' \to X$ be any splitting of x of the kind defined above. Then for any signal alphabet S, prior*

[11] Such a split exists by the assumption that $\mathscr{N}$ is minimal (N_j cannot be covered by $\mathscr{N} \setminus \{N_j\}$).

[12] This does not prevent us from defining the cost function for such a measure. Recall that we have adopted the convention that the $H(\cdot)$ functions are homogenous of degree one, and hence that the cost functions $C(\cdot)$ are homogenous of degree one in the prior.

$q \in \mathscr{P}(X)$, and signal structure $p \in \mathscr{P}(S)^{|X|}$,

$$C(p,q;S,X) = C(p',q';S,X'),$$

where $q' \in \mathbb{R}_+^{|X'|}$ and $p' \in \mathscr{P}(S)^{|X'|}$ are defined by $q'_{x''} = q_{\phi(x'')}$ and $p'_{x''} = p_{\phi(x'')}$ for all $x'' \in X'$.

Figure 2 illustrates this assumption an example neighborhood structure. A first key implication of this assumption is that it is without loss of generality to suppose that the X_i are disjoint. This implication, when combined with Assumption 2, allows us to invoke standard results on additive separability.

A second key implication is that the location of the states within each neighborhood is irrelevant for information costs. Suppose that we have a specification in which the X_i are disjoint (which, as just noted, is without loss of generality). Then Assumption 3 must also apply to a splitting that simply replaces some state x by a new state belonging to the same unique neighborhood as x, and with the same prior probability as x. Hence the locations of the x_j in each neighborhood are irrelevant for information costs — only the neighborhoods to which the different states belong and their prior probabilities can matter. This result ensures that the difficulty of distinguishing two states is governed entirely by the neighborhood structure, as opposed to by the nature of information costs within a neighborhood.[13]

Combining our first three assumptions, we derive an additive separability result showing that the H function can be written as the expected sum of a local information cost in each neighborhood.[14] Moreover, we show that this local information cost depends only on the cardinality of the neighborhood and probabilities of each state within the neighborhood. That is, each local information cost is a symmetric function of the probabilities of each state within the neighborhood. We present this result below, but first we introduce some additional notation. For each neighbor-

[13]Note that by partitioning a single neighborhood into a set of smaller overlapping neighborhoods we can capture more local notions of distance between states, while still satisfying this assumption.

[14]For ease of exposition, we have made Assumptions 2 and 3 stronger than necessary. Assumption 3 does not need to hold for all p, only for signal structures that discriminate between one or two states and all others (as in (6) and (7)). Both Assumptions only need hold for values of r' close to r, as opposed to more informative signal structures. Our assumptions can be weakened because, for UPS costs functions, the costs of nearly uninformative signals determine the costs of all possible signals.

hood X_i, we define the probability that some state belonging to neighborhood X_i (and N_i) occurs under beliefs $q \in \mathscr{P}(X)$, $\bar{q}_i(q) \equiv \sum_{x \in X_i} q_x$. For neighborhoods with positive probability ($\bar{q}_i(q) > 0$), we define $q_i(q) \in \mathscr{P}(X_i)$ as the conditional distribution over X_i under q, and adopt the convention that $q_i(q)$ is uniform if $\bar{q}_i(q) = 0$.

Proposition 1. *Under Assumptions 1, 2, and 3, the H function associated with the cost function C can be written as*

$$H(q; X, \mathscr{N}) = \sum_{i \in \mathscr{I}} \bar{q}_i(q) H^i(q_i(q); |X_i|),$$

with the $\{H^i(\cdot; |X_i|) : \mathbb{R}_+^{|X_i|} \to \mathbb{R}\}_{i \in \mathscr{I}}$ *symmetric, twice-differentiable, and convex.*

It is without loss of generality to assume that, for all $i \in \mathscr{I}$, H^i *is homogenous of degree one and reaches its minimum when* q_i *is uniform.*

Proof. See the Appendix, section C.1. □

We will call any information cost function that can be written in this way a "neighborhood-based cost function." To use these cost functions in applications, we must specify both the neighborhood structure of the state space (the sets $\{X_i \subseteq X\}_{i \in \mathscr{I}}$) and the local information costs (the functions $\{H^i\}_{i \in \mathscr{I}}$), but it is not necessary to explicitly specify $\mathscr{X}$ or $\mathscr{N}$. We proceed by describing several (closely related) varieties of neighborhood-based cost function.

3.1 Neighborhoods with Compression

We begin by providing an additional assumption that justifies (the negative of) Shannon's entropy as the local information cost. Under this assumption, it is without loss of generality to "compress" or "merge" states that are contained within the same set of neighborhoods, provided that the signal structure does not discriminate between these states. This leads to the conclusion that the H^i function must be proportional to Shannon's entropy. This result is closely related to the "invariance under compression" axiom of Caplin et al. (2019).

Specifically, we assume that if $x, x' \in X$ are distinct states contained in a common set of neighborhoods, and the signal structure p does not distinguish between

them ($p_x = p_{x'}$), then it is as-if they were in fact a single state. This assumption is appealing because it implies that states that are identical both in terms of their payoffs for each action and in terms of their perceptual properties can be treated as a single state when considering the DM's decision problem. Figure 3 contains a diagram with an example neighborhood structure that summarizes this assumption.

Assumption 4. *Fix X, and let $X'' \subset \mathscr{X}$ be any set covered by the neighborhood covering such that a surjection (many-to-one mapping) $m : X \to X''$ exists and satisfies, for all $i \in \mathscr{I}$ and $x \in X$, $x \in N_i \iff m(x) \in N_i$. Then for all such X'', all $q_0 \in \mathscr{P}(X)$, and all $p \in \mathscr{P}(S)^{|X|}$ such that $p_x = p_{x'}$ for all $x, x' \in X$ with $m(x) = m(x')$,*

$$C(p, q_0; S, X) = C(p'', M(q_0); S, X''),$$

where $M : \mathscr{P}(X) \to \mathscr{P}(X'')$ is defined by, for all $x'' \in X''$,

$$M(q)_{x''} = \sum_{x \in X : x'' = m(x)} q_x.$$

the signal structure $p'' : X'' \to \mathscr{P}(S)$ is the signal structure that satisfies $M(q_s(p, q_0)) = q_s(p'', M(q_0))$ and $\pi_s(p, q_0) = \pi_s(p'', M(q_0))$ for all $s \in S$.

When the mapping m is a bijection (one-to-one mapping), Assumption 4 shares with Assumption 3 the implication that the location of the states within the neighborhoods does not matter. What is being added by Assumption 4 is the implication that it is without loss of generality to merge states, provided the signal structure does not discriminate between the states being merged. If the signal structure does discriminate between the states being merged, it must be more costly than the merged signal structure (by Assumption 2). This form for monotonicity and invariance has been shown to exactly characterize Shannon's entropy (defined in (2)).[15]

Proposition 2. *If a family of neighborhood-based cost functions defined by $H(q; X, \mathscr{N})$*

[15] The result follows from known results in information geometry (Chentsov (1982)). Caplin et al. (2019) describe the behavioral axiom that characterizes the Shannon entropy cost.

satisfies Assumption 4, the H^i functions are

$$H^i(q_i;|X_i|) = c_i \sum_{j=1}^{|X_i|} q_{i,j} \ln\left(\frac{q_{i,j}}{\frac{1}{|X_i|}\sum_{k=1}^{|X_i|} q_{i,k}}\right),$$

where $\{c_i \in \mathbb{R}_+\}_{i\in I}$ are constants and $q_{i,j}$ denotes the j-th element of $q_i \in \mathbb{R}_+^{|X_i|}$.

Proof. See the Appendix, section C.2. □

The summation in Proposition 2 is written using the indices j rather than the states $x \in X_i$ to emphasize that this function depends only the cardinality of X_i. The local information costs H^i described in this proposition are an affine transformation of Shannon's entropy, extended to $\mathbb{R}_+^{|X_i|}$ by assuming homogeneity of degree one.[16]

The neighborhood structure and constants c_i determine the difficulty of discriminating between nearby states. We consider them part of the economic environment, observing that problems with similar payoffs can nevertheless differ in terms of the DM's ability to distinguish between exogenous states. This is exactly the kind of variation that occurs, for example, in perceptual experiments.

The ability to merge states that are identical in terms of their payoffs and perceptual properties is appealing, and Proposition 2 shows that this assumption alone is sufficient to determine the cost function, up to a set of constants. However, because this assumption generates such a sharp result, it also pins down the curvature of cost function, and hence the responsiveness of the DM's behavior to changing incentives. We discuss this issue in more detail next.

3.2 Neighborhoods with Generalized Entropy

Suppose that the state space consists of two states, $X = \{x_1, x_2\}$, that share a neighborhood, and that the action space, $A = \{L, R\}$, consists of two actions. To simplify matters further, suppose that the DM has a uniform prior over the two states, and that the payoffs are symmetric, $u_{L,x_1} = u_{R,x_2} = \bar{u} > 0$, $u_{L,x_2} = u_{R,x_1} = 0$. In this setup,

[16] These local information costs can also be described as being proportional to the KL divergence between q_i and the uniform distribution.

we can explore how the DM's behavior changes in response to changing incentives, a question studied experimentally by Dean and Neligh (2019).[17]

Suppose that the DM solves the rational inattention problem in (1). By symmetry, we can define $p(L|x_1) = p(R|x_2) = p_c$ as the probability the DM chooses "correctly" ex-post. The associated posteriors are $q_L = [p_c, 1-p_c]$ and $q_R = [1-p_c, p_c]$. and hence by symmetry the DM solves, given the local information cost $H^i(\cdot)$,

$$\max_{p_c \in [0,1]} \bar{u} p_c - \theta H^i(\begin{bmatrix} p_c \\ 1-p_c \end{bmatrix}; 2).$$

The elasticity of the choice probability p_c to the incentives $\bar{u}$ is characterized by

$$\frac{\bar{u}}{p_c}\frac{\partial p_c}{\partial \bar{u}} = \frac{1}{p_c}\frac{H^i_q(\begin{bmatrix} p_c \\ 1-p_c \end{bmatrix}; 2) \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix}}{\begin{bmatrix} 1 & -1 \end{bmatrix} \cdot H^i_{qq}(\begin{bmatrix} p_c \\ 1-p_c \end{bmatrix}; 2) \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix}}. \tag{8}$$

By Proposition 2, Assumption 4 uniquely determines this elasticity, $\frac{\bar{u}}{p_c}\frac{\partial p_c}{\partial \bar{u}} = \frac{\ln(\frac{p_c}{1-p_c})}{1+\frac{p_c}{1-p_c}}$.

Caplin and Dean (2013) and Dean and Neligh (2019) show that the responsiveness of choice probabilities to incentives observed in their experiment is lower than the responsiveness predicted by this elasticity (and hence the Shannon's entropy cost function). Those authors proceed by considering a more general class of cost functions, defined using the generalized entropy of index of Shorrocks (1980), that nests Shannon's entropy as a special case while also including cost functions with more curvature. The curvature of these cost functions is controlled by a parameter ρ (a larger ρ leads to more curvature when $\rho > 0$), with $\rho = 1$ corresponding to the case of Shannon's entropy. Dean and Neligh (2019) use the following analogy: the

[17] Experiment #2 of Dean and Neligh (2019) considers the effects of incentives on choice probabilities, and is described by those authors as fitting this setup. The experiment involves subjects determining whether a screen contains more red balls or more blue balls. The screen always contains 49 balls of one color and 51 balls of the other, arranged randomly. Consequently, there are in fact many states, with some perceptual topology. We are justified in modeling their experiment as containing only two states only for certain neighborhood structures (which may or may not accurately reflect the perceptual structure of the experiment) and if we impose Assumption 4.

generalized entropy index is to Shannon's entropy as CRRA utility is to log utility.

Definition 1. The generalized entropy index of Shorrocks (1980), $H^G(q_i;\rho,|X_i|)$, is defined for any $\rho \in \mathbb{R}$ and interior $q_i \in \mathbb{R}_+^{|X_i|}$ with $\sum_{j=1}^{|X_i|} q_{i,j} = 1$ as

$$H^G(q_i;\rho,|X_i|) = \begin{cases} \frac{1}{|X_i|}\frac{1}{(\rho-2)(\rho-1)}\sum_{j=1}^{|X_i|}\{(|X_i|q_{i,j})^{2-\rho}-1\} & \rho \notin \{1,2\} \\ -\frac{1}{|X_i|}\sum_{j=1}^{|X_i|}\ln(q_{i,j}) & \rho = 2 \\ \sum_{j=1}^{|X_i|} q_{i,j}\ln(q_{i,j}) & \rho = 1. \end{cases}$$

Like Shannon's entropy, the generalized entropy index is a symmetric function that depends only on the cardinality of the state space, and hence satisfies the conditions of Proposition 1 for a local information cost (after being extended $\mathbb{R}_+^{|X_i|}$ by homogeneity of degree one). With this family of local information costs, the elasticity of the choice probability p_c to the incentives $\bar{u}$ is

$$\frac{\bar{u}}{p_c}\frac{\partial p_c}{\partial \bar{u}} = \frac{\frac{1}{1-\rho}[1-(\frac{1-p_c}{p_c})^{1-\rho}]}{[1+(\frac{1-p_c}{p_c})^{-\rho}]}.$$

This elasticity is decreasing in ρ for $\rho > 0$ holding fixed p_c, consistent with the intuition that (for positive ρ) a higher value of ρ generates a more curved cost function. In the set of experiments considered by Dean and Neligh (2019), a neighborhood cost function with $\rho \approx 13$ is found to best fit the experimental data under their modeling assumptions. That is, a cost function with more curvature than Shannon's entropy best fits the low responsiveness of subjects to changing incentives.

We use $H^{NG}(q;\rho,X,\mathcal{N})$ to denote the family neighborhood-based cost functions that use a generalized entropy index with parameter ρ in each neighborhood. This family of cost functions is useful because it flexibly parameterizes the elasticity of choice probabilities to incentives. The local information costs H^G are characterized by the property of being "additively decomposable" (Shorrocks, 1980), and hence could be derived via a generalization of Assumption 4. We do not find this justification intuitive in our setting, and prefer to view this class as an ad-hoc family of cost functions that can capture varying degrees of responsiveness to incentives.

Definition 2. The generalized entropy neighborhood-based cost function is the UPS cost function defined by the entropy function

$$H^{NG}(q;\rho,X,\mathscr{N}) = \sum_{i\in\mathscr{I}} c_i \bar{q}_i(q)\, H^G(q_i(q);\rho,|X_i|),$$

where $\{c_i \in \mathbb{R}_+\}_{i\in\mathscr{I}}$ are constants, for any q in the relative interior of the simplex, and is defined on the boundary by continuity for $\rho < 2$ and as infinity for $\rho \geq 2$.

Using this generalized entropy function, we can define a Bregman divergence, $D_{NG}(q_s||q;\rho,X,\mathscr{N})$, as in (5), and a static rational inattention problem,[18]

$$\begin{aligned} V_{NG}(q;\rho,X,\mathscr{N}) = &\max_{\pi\in\mathscr{P}(A),\{q_a\in\mathscr{P}(X)\}_{a\in A}} \sum_{a\in A} \pi(a)(u_a^T \cdot q_a) \\ &- \theta \sum_{a\in A} \pi(a) D_{NG}(q_a||q;\rho,X,\mathscr{N}), \end{aligned} \tag{9}$$

subject to the constraint $\sum_{a\in A} \pi(a) q_a = q$.

It is sometimes more convenient to work with cost functions defined over signals $\{p_x \in \mathscr{P}(S)\}_{x\in X}$, as opposed to posteriors q_a and unconditional probabilities π (as in (1)). We rewrite (9) using Bayes' rule below.

Lemma 1. *The static rational inattention problem in (9) can be written as*

$$\begin{aligned} V_{NG}(q;\rho,X,\mathscr{N}) = &\max_{\{p_x\in\mathscr{P}(S)\}_{x\in X}} \sum_{s\in S} \pi_s(p,q_0)\hat{u}(q_s(p,q)) \\ &- \theta \sum_{i\in\mathscr{I}} c_i |X_i|^{1-\rho} \bar{q}_i(q)^{\rho-1} \sum_{x\in X_i} (q_x)^{2-\rho} D_\rho(p_x||\pi_i), \end{aligned}$$

where $\pi_i \in \mathscr{P}(S)$ is defined by $\pi_i = \sum_{x\in X_i} p_x q_{i,x}(q)$ and

$$D_\rho(p_x||\pi) = \begin{cases} \frac{1}{(\rho-2)(\rho-1)} \sum_{s\in S:\pi_s>0} \pi_s((\frac{p_{x,s}}{\pi_s})^{2-\rho} - 1) & \rho \neq \{1,2\} \\ \sum_{s\in S:\pi_s>0} \pi_s \ln(\frac{\pi_s}{p_{x,s}}) & \rho = 2 \\ \sum_{s\in S:\pi_s>0} p_{x,s} \ln(\frac{p_{x,s}}{\pi_s}) & \rho = 1. \end{cases}$$

Proof. See the Appendix, Section C.3. □

[18] To deal with the boundaries in the $\rho \geq 2$ case, we assume q has full support in this problem.

The divergences D_ρ are known as the α-divergences (under a different parameterization) and are a transformed version of the Renyi divergences. In the $\rho = 1$ case, D_ρ is the Kullback-Leibler divergence.

The curvature of the local information cost function, which is controlled by the parameter ρ for the generalized entropy neighborhood-based cost functions, is closely related to the issue of whether the cost function exhibits what Pomatto et al. (2020) call increasing, constant, or decreasing marginal costs. We discuss this in more detail in the Appendix Section A, and in particular demonstrate that the neighborhood-based cost functions defined by $H(q;X,\mathcal{N}) = H^{NG}(q;1,X,\mathcal{N}) + H^{NG}(q;2,X,\mathcal{N})$ exhibit constant marginal costs (and thus fall into the "total information cost" family described by Bloedel and Zhong (2020)).

3.3 The Fisher Information Cost Function

Thus far, we have described an assumption that is sufficient to determine, for any neighborhood structure, the local information costs, and described a more general family of cost functions that can parameterize the elasticity of choice probabilities to incentives. We next take a different approach, and study a specific neighborhood structure, states ordered on a line, with the aim of deriving results that apply regardless of the nature of the local information costs. We first discuss the case of a discrete set of states, and then extend our results to allow for a continuum of states.

Suppose that there are $M+1$ ordered states, $X^M = \{0,1,\ldots,M\}$, and that each pair of adjacent states forms a neighborhood, $X_i = \{i, i+1\}$, for all $i \in \{0,1,\ldots,M-1\}$. Thus two states belong to a common neighborhood if and only if one comes immediately after the other in the sequence. This captures the idea that the readily available measurement technologies respond similarly in states that are "similar," in the sense of being at nearby positions in the sequence.

With this neighborhood structure ($\mathcal{N}^M$), for all full-support q^M, any neighborhood-based cost function can be written as

$$H(q^M;X^M,\mathcal{N}^M) = \sum_{j=0}^{M-1} c_j(q_j+q_{j+1})H^j(\begin{bmatrix}\frac{1}{2}-\varepsilon_j\\ \frac{1}{2}+\varepsilon_j\end{bmatrix}),$$

where

$$\varepsilon_j = \frac{1}{2}\frac{q_{j+1} - q_j}{q_j + q_{j+1}}$$

and the local information costs H^j are scaled[19] such that

$$\frac{\partial^2}{\partial \varepsilon_j^2} H^j(\begin{bmatrix} \frac{1}{2} - \varepsilon_j \\ \frac{1}{2} + \varepsilon_j \end{bmatrix})|_{\varepsilon_j = 0} = 4.$$

By Proposition 1, it is without loss of generality to assume that the H^j functions reach their minimum when $\varepsilon_j = 0$. Consequently up to second order,

$$H(q; X^M, \mathcal{N}^M) = \frac{1}{4}\sum_{j=0}^{M-1}\{c_j \frac{(q_{j+1} - q_j)^2}{\frac{1}{2}(q_j + q_{j+1})} + o(\varepsilon_j^2)\}. \tag{10}$$

This approximation is exact for $H^{NG}(q; \rho = 0, X, N)$.

Let us further assume that $c_i = 1$ for all i, implying that it is equally difficult to distinguish two neighboring states at all points in the sequence.[20] We interpret this assumption as requiring that the "perceptual distance" between adjacent states is the same for all states on the line. The construction of numerical scales measuring physical stimuli so that equal distances imply equal difficulty of discrimination is a familiar exercise in psychophysics; it often requires that the scale be a nonlinear function of measurable physical properties of the stimuli (Gescheider, 1988).

Based on this approximation, it is tempting to suppose that, in the limit as $M \to \infty$, if the discrete distributions q_M converge to differentiable function q,

$$\lim_{M\to\infty} H(q^M; X^M, \mathcal{N}^M) = \frac{1}{4}\int_{supp(q)} \frac{(q'(x))^2}{q(x)} dx,$$

where $supp(q)$ denotes the support of q. Building on this intuition, we can define a

[19] This scaling is arbitrarily chosen to match the scale of the generalized entropy indices.

[20] If c_i is the same for all i, we can without loss of generality set it equal to one, as the parameter θ can still be used to scale the overall magnitude of information costs.

continuous-state rational inattention problem:

$$V_N(q) = \sup_{\pi \in \mathscr{P}(A), \{q_a \in \mathscr{P}_{LipG}\}_{a \in A}} \sum_{a \in A} \pi(a) \int_{supp(q)} u_a(x) q_a(x) dx$$
$$- \frac{\theta}{4} \sum_{a \in A} \{\pi(a) \int_{supp(q)} \frac{(q_a'(x))^2}{q_a(x)} dx\} + \frac{\theta}{4} \int_{supp(q)} \frac{(q'(x))^2}{q(x)} dx, \qquad (11)$$

subject to the constraint that, for all x,

$$\sum_{a \in A} \pi(a) q_a(x) = q(x).$$

In this expression, the real number x is the exogenous state, $u_a(x)$ is the utility of action $a \in A$ in state x, $q(x)$ is the prior over the states, and $q_a(x)$ is the posterior belief conditional on taking action a. The notation $\mathscr{P}_{LipG}$ refers to a set of probability measures on the support of q that we describe below.

This problem can alternatively be formulated as a choice of the signal structure:

$$V_N(q) = \sup_{\{p_a\}_{a \in A} \in \mathscr{P}_{LipG}(A)} \int_{supp(q)} q(x) \sum_{a \in A} p_a(x) u_a(x) dx \qquad (12)$$
$$- \frac{\theta}{4} \int_{supp(q)} q(x) \sum_{a \in A : p_a(x) > 0} \frac{(p_a'(x))^2}{p_a(x)} dx,$$

where $\mathscr{P}_{LipG}(A)$ is the set of mappings $\{p_a : supp(q) \to [0,1]\}_{a \in A}$ such that for all x, $\sum_{a \in A} p_a(x) = 1$, and for each a, $p_a(x)$ is either everywhere zero or strictly positive and differentiable, with a Lipschitz-continuous derivative.

This formulation shows that our proposed static information-cost function is a weighted average of the Fisher information (Cover and Thomas (2012), sec. 11.10), a real number for each point in the state space that provides a measure of the local discriminability of states.[21] It is for this reason that we refer to our proposal as the "Fisher-information cost function." Like the mutual-information cost function, the Fisher-information cost function is a single-parameter cost function, and it can also

[21] The equivalence of the two formulations is shown in the Technical Appendix, section C.2, where we also provide further discussion of the connection with Fisher information.

be applied in almost any context, as long as the state space is continuous.[22]

We prove the convergence of the static problem described in section §2 to this problem formally in the Technical Appendix, Section D.1, under some regularity assumptions on the prior q (differentiability, with a Lipschitz-continuous derivative, and support on a compact set), for the specific case in which the local information costs are the negative of Shannon's entropy.[23] In the proof, we show that the limiting optimal posteriors q_a are also differentiable and have the same support as q (so the Fisher information integrals make sense) and that their derivatives are also Lipschitz-continuous (which helps prove convergence). We refer to the set of full-support, differentiable probability distribution functions with Lipschitz-continuous derivatives as $\mathscr{P}_{LipG}$. The proof is relatively technical, and the relevant economics are summarized by the approximation (10).

The key step is to demonstrate that the DM will choose a signal structure such that the posteriors are in $\mathscr{P}_{LipG}$. However, if we simply assume this, it is straightforward to extend our results to any set of local information costs using the approximation in (10). We can then immediately observe that all local information costs lead to the same continuous-state limit. We can also observe from (12) that the Fisher information cost is linear in the prior, and therefore will exhibit what Pomatto et al. (2020) call constant marginal costs (see Appendix Section A).

The argument we have outlined assumed the state variable x was structured so that the perceptual distance between each pair of adjacent states is the same. In the continuous-state limit, this led to a Fisher information integral defined above. Let us now suppose that we would like to define our state space using the alternative coordinate $y = f(x)$, where $f(\cdot)$ is a strictly monotone and differentiable function. In this case, defining $\hat{q}(y) = q(f^{-1}(x))$ and $c(y) = f'(f^{-1}(y))$, we have

$$\frac{1}{4}\int_{supp(\hat{q})} c(y)\frac{(\hat{q}'(y))^2}{\hat{q}(y)}dy = \frac{1}{4}\int_{supp(q)} \frac{(q'(x))^2}{q(x)}dx.$$

[22]The fact that we have a single free parameter depends on having chosen a coordinate x for the state space with the property that the difficulty of discriminating nearby states increases with the distance Δx between two states in a similar way at all points in the state space.

[23]We also assume bounded utilities. We believe the result holds for many other local information costs, and with weaker regularity assumptions, but generalizing our proof is not trivial.

The function $c(y)$ captures the local perceptual distance between neighboring values of y, just as the constants c_j capture this information in the discretized version of the model. In the special case in which f is linear, $c(y)$ is constant, and can be incorporated into the scalar θ. In applications, one can either transform the state variable to ensure uniform perceptual distance (use the x coordinate) or explicitly model perceptual distance as part of the cost function (using any coordinate y), whichever is more convenient. In our applications, we have often found it convenient to choose coordinates such that $c(y)\hat{q}(y)$ is constant.

If we are willing to assume posteriors in $\mathscr{P}_{LipG}$, it is also straightforward to extend our results to a multi-dimensional state space. Suppose that, instead of being ordered on a line, the state space consists of an L-dimensional grid, with each edge consisting of M states ordered on a line, and the neighborhoods are all pairs of states that are adjacent in one of the L dimensions. In this case, by arguments almost identical to those in the technical appendix, one can show that

$$\lim_{M\to\infty} H(q^M; X^M, \mathscr{N}^M) = \frac{1}{4}\int_{supp(q)} \frac{|\nabla q(x)|^2}{q(x)} dx,$$

where $\nabla q(x)$ denotes the gradient. In effect, this simply adds up the one-dimensional Fisher information costs in each dimension. Note again that this expression uses equal weights on the Fisher information for the various dimensions, implicitly assuming that the units in which distance is measured along the various dimensions are uniform and equivalent, in the sense that a given distance along any dimension has the same consequence for the degree of discriminability of nearby states. We can write the multi-dimensional problem in terms of the signal structure using

$$C_{Fisher}(p,q;A) = \frac{\theta}{4}\int_{supp(q)} q(x) \sum_{a\in A} \frac{|\nabla p_a(x)|^2}{p_a(x)} dx. \tag{13}$$

Thus, our proposed Fisher-information cost function can be readily applied to multi-dimensional settings with a continuous state space. We turn now to applications, to illustrate the effects of using our proposed alternatives instead of the standard rational inattention cost function.

4 Applications of Neighborhood-Based Cost Functions

In this section, we discuss several applications. We first discuss perceptual experiments, and then apply the Fisher information cost to the general setting of binary choice to show that the predicted choice probabilities resemble psychometric curves. We next apply these results to regime change games. Lastly, we consider the general setting of linear-quadratic-Gaussian problems.

4.1 Psychometric Functions

We begin by discussing perceptual experiments (example 2), of the sort conducted by Shadlen et al. (2007) and Dean and Neligh (2019). In some of these experiments, the state is most naturally modeled as a continuous variable; in others, the state space is finite. For expositional purposes, in this subsection we assume a discrete state space $X = \{0, 1, 2, \ldots, M\}$, where M is an odd integer. The action R is the correct response if and only if $x > x^* = M/2$. We also assume a uniform prior.

Let us first consider whether mutual information, or any other symmetric UPS cost function, can generate the kinds of psychometric curves observed in these experiments. Any symmetric UPS cost function can be thought of as a neighborhood-based cost function with the particular neighborhood structure in which all states belong to a single neighborhood. The following corollary demonstrates that for this class of costs, the likelihood of the DM choosing R in the perceptual experiment can depend only on whether the $x > M/2$ and not on how far x is from $M/2$.

Corollary 1. *Consider a rational inattention problem with a neighborhood-based cost function, and let x, x' be two states with the property that (i) $u_{a,x} = u_{a,x'}$ for all actions $a \in A$, (ii) $q_x = q_{x'}$, and (iii) the set of neighborhoods $\{X_i\}$ such that $x \in X_i$ is the same as the set such that $x' \in X_i$. Then under the optimal policy, $p^*_x = p^*_{x'}$. If the local information costs of the neighborhood-based cost function are proportional to Shannon's entropy, this result holds even if $q_x \neq q_{x'}$.*

Proof. Immediate from Lemma 1. □

Corollary 1 implies that the probability of response R must be the same for all states $x < M/2$, and also the same (but higher) for all states $x > M/2$. Changing

the scale of the information cost changes the degree to which the probability of R is higher when $x > M/2$, but the response probabilities still depend only on whether x is greater or less than $M/2$. This is illustrated in Figure 4, which plots the optimal response frequencies as a function of x, for alternative θ, under mutual information. As discussed in Section 2, this prediction of mutual information is not consistent with the psychometric functions observed in perceptual experiments.

Alternatively, consider a neighborhood-based cost function in which the neighborhoods are given by $X_i = \{i, i+1\}$ for $i = 1, 2, \ldots, M-1$, as in Section 3.3. From the approximation in (10), it is apparent that any neighborhood-based cost function with this structure, irrespective of the nature of local information costs, will penalize sharp changes in the posterior probabilities between neighboring states. Consider the optimal strategy described above, and the posterior associated with the action $a = R$. This posterior features a sharp change in the probability between the states $j = (M-1)/2$ and $j = (M+1)/2$ and therefore will be costly under any neighborhood-based cost function with this neighborhood structure.

As a result, the DM will be better off with posteriors that vary smoothly in the state, generating exactly the kinds of psychometric functions observed in experiments.[24] We illustrate this result, for the particular neighborhood-based cost function that uses Shannon's entropy as the local information cost, in Figure 5. This figure again shows the optimal response frequencies as a function of x, for alternative θ. The sigmoid functions predicted with this cost function are characteristic of measured psychometric functions in perceptual experiments of this kind.

4.2 Binary Choice with the Fisher Information Cost

We next consider the general problem of binary choice with the Fisher information cost and show that when payoffs satisfy a single-crossing property, the resulting choice probabilities resemble psychometric functions. Our results on convergence prove that (under some conditions) the optimal policies under a neighborhood-

[24] While the difference between the predictions of the mutual information and neighborhood-based cost functions is especially stark in the case of a discontinuous payoff function of the kind assumed here, there are also notable differences when the payoff function is continuous but kinked, as in the application to security design treated in appendix section B.

based cost function in the discrete state space will converge to the optimal policies with the Fisher information cost on the continuous state space. Thus, the results in this section provide formal underpinnings for the results shown in Figure 5.

We will assume that X is a compact subset of the real line, $X = [x_L, x_H]$, and that the DM has a prior $q \in \mathscr{P}_{LipG}$ with full support on X. We arbitrarily label the two actions $A = \{L, R\}$, and normalize (without loss of generality) the utility of action L to zero, $u_L(x) = 0$. We assume $u_R(x)$ is finite on X and has finitely many discontinuities, but impose no other assumptions at this point.

Taking advantage of the binary choice structure ($p_L(x) + p_R(x) = 1$ for all $x \in X$), we can rewrite the DM's problem in (12) as

$$V_N(q) = \max\{\sup_{p_R \in C^1(X,(0,1))} \int_X q(x) p_R(x) u_R(x) dx - \frac{\theta}{4} \int_X q(x) \frac{(p_R'(x))^2}{p_R(x)(1-p_R(x))} dx, \tag{14}$$

$$\int_X q(x) u_R(x) dx, 0\},$$

where $C^1(X,(0,1))$ is the set of differentiable functions from $[x_L, x_H]$ to $(0,1)$. This expression captures the key properties of the set $\mathscr{P}_{LipG}(A)$: for each action, the choice probability is either interior and differentiable or is zero everywhere. We have ignored the requirement that p_R have a Lipschitz-continuous gradient; we (implicitly) show below that this condition is satisfied.

Echoing the results of Woodford (2008) and Yang (2020) on binary choice with mutual information, there are three possible optimal strategies: always L, always R, and a strictly interior strategy, $p_R(x) \in (0,1)$ for all $x \in X$. We provide necessary and sufficient conditions for each of these three cases to be an optimal policy, and describe the differential equation that characterizes the optimal interior policy.

Proposition 3. *If*

$$\inf_{p_L \in \{p \in C^1(X,(0,\infty)) : \int_X q(x) p(x) dx = 1\}} \int_X q(x) p_L(x) u_R(x) dx + \frac{\theta}{4} \int_X q(x) \frac{(p_L'(x))^2}{p_L(x)} dx \geq 0, \tag{15}$$

then always-R is an optimal policy. If

$$\inf_{p_R \in \{p \in C^1(X,(0,\infty)):\int_X q(x)p(x)dx=1\}} -\int_X q(x)p_R(x)u_R(x)dx + \frac{\theta}{4}\int_X q(x)\frac{(p_R'(x))^2}{p_R(x)}dx \geq 0, \tag{16}$$

then always-L is an optimal policy. Any optimal policy $p_R^* \in C^1(X,(0,1))$ satisfies the following differential equation: for all $x \in X$ at which $u_R(x)$ continuous,

$$\frac{p_R^*(x)(1-p_R^*(x))}{2\theta}u_R(x) - \frac{1}{2}\frac{(p_R^{*\prime}(x))^2}{p_R^*(x)(1-p_R^*(x))}(1-2p_R^*(x)) + \frac{q'(x)}{q(x)}p_R^{*\prime}(x) + p_R^{*\prime\prime}(x) = 0.$$

with the boundary conditions

$$p_R^{*\prime}(x_L) = p_R^{*\prime}(x_H) = 0.$$

A $C^1(X,(0,1))$ solution to this differential equation satisfying the boundary conditions exists if (15) and (16) do not hold, and any such solution is an optimal policy.

Proof. See the Appendix, Section C.4. □

One particular case that is sometimes of interest in applications (such as the security design application described in Appendix Section B) is when the DM is just indifferent between always-R and gathering information. We next describe a necessary and sufficient condition for this indifference to hold.[25]

Lemma 2. *The condition*

$$\inf_{p_L \in \{p \in C^1(X,(0,\infty)):\int_X q(x)p(x)dx=1\}} \int_X q(x)p_L(x)u_R(x)dx + \frac{\theta}{4}\int_X q(x)\frac{(p_L'(x))^2}{p_L(x)}dx = 0$$

holds if and only if a function $\psi : X \to \mathbb{R}$ exists with $\psi(x_H) = 0$ and, for all $x \in X$,

$$\psi(x) = -\int_{x_L}^{x}[\frac{1}{2\theta}u_R(x') + \frac{1}{4}\psi(x')^2 + \frac{q'(x')}{q(x')}\psi(x')]dx'.$$

Proof. See the Appendix, Section C.5. □

[25] The analogous condition for always-L is identical, except that $-u_R$ replaces u_R.

These results characterize the DM's optimal choice probabilities. The next corollary demonstrates that these choice probabilities will resemble psychometric curves in a large class of decision problems.

Corollary 2. *Suppose that u_R satisfies strict single-crossing, meaning that for some $x^* \in (x_L, x_H)$, $u_R(x) < 0$ for all $x < x^*$ and $u_R(x) > 0$ for all $x > x^*$, and that the DM chooses to gather some information (does not always choose L or R). Then the optimal choice probabilities p_R^* satisfy:*

i) $p_R^*(x)$ is strictly increasing on $x \in (x_L, x_H)$ and satisfies $p_R^{*\prime}(x_L) = p_R^{*\prime}(x_H) = 0$,

ii) for some $x_1 \in (x_L, x_H)$, $p_R^*(x)$ is strictly convex on $x \in [x_L, x_1)$, and

iii) for some $x_2 \in [x_1, x_H)$ $p_R^*(x)$ is strictly concave on $(x_2, x_H]$.

Proof. See the Appendix, Section C.6. □

In the region $[x_L, x_1)$, $p_R^*(x)$ begins as a flat function and is strictly increasing and strictly convex. In the region $(x_2, x_H]$, $p_R^*(x)$ is strictly increasing and concave, and ends as a flat function at some higher value than it started, $p_R^*(x_H) > p_R^*(x_L)$.[26] Thus, under the strict single-crossing assumption (which is satisfied by the payoffs in the regime change game, in perceptual experiments, and in many other environments), the optimal choice will have a sigmoid shape reminiscent of psychometric curves.

4.3 The Regime Change Game

We next consider the implications of our results on binary choice in the regime change game (example 1 from Section 2). In this game, it is natural to model the state as a continuous variable. Building on the work of Yang (2015) and Morris and Yang (2019), we contrast the predictions of mutual information and Fisher information in this context. We assume the game consists of a continuum of agents whose signals are conditionally independent given x. We consider symmetric, monotone equilibria and invoke the law of large numbers by assuming that the fraction of

[26] This corollary leaves open the possibility $p_R^*(x)$ alternates between concave and convex regions within the interval $[x_1, x_2]$. However, all of the examples we have constructed satisfy $x_1 = x_2$.

agents investing in state x, $l(x)$, is equal to the probability of each individual investing in state x, $p^*_{invest,x}$. In a monotone equilibrium of the game, the regime will change if and only if $x \geq x^*$ for some x^*. To sustain such an equilibrium, it must be the case that $l(x)$ exceeds $1-x$ if and only if $x \geq x^*$.

As discussed in the previously, under any symmetric UPS cost function (e.g. mutual information), the DM will optimally receive a signal that sharply discriminates between states with $x \geq x^*$ and $x < x^*$, but does not discriminate within the sets of states with $x \geq x^*$ or $x < x^*$. Let us suppose that the DM's optimal signal is characterized by $p^*_{invest,x} = l_H$ for all $x \geq x^*$ and $p^*_{invest,x} = l_L$ for all $x < x_H$, with $l_H > l_L$. If all DMs follow this strategy, then their strategies will be consistent with the equilibrium cutoff x^* provided that $l_H \geq 1 - x^* > l_L$.

This argument (which follows Yang (2015)) demonstrates that there are many monotone equilibria of the regime change game.[27] In the limit as information becomes costless, $l_H \to 1$ and $l_L \to 0$, and any $x^* \in (0,1)$ characterizes a monotone equilibrium. However, this issue arises because of agents' ability to sharply discriminate between neighboring states in the vicinity of x^*. In each of these equilibria, the DMs receive a signal that abruptly changes around the threshold x^*, which is not consistent with the psychometric curves observed in experiments.

In contrast, with the Fisher information cost, a DM who chooses $p^*_{invest,x^*} = l_H$ will choose, for x less than but close to x^*, $p^*_{invest,x}$ less than but close to l_H, by Corollary 2. Consequently, in equilibrium the DM must optimally choose $p^*_{invest,x^*} = l_H = 1 - x^*$, as otherwise regime change would occur for values of x less than x^*. Under mild conditions, there is a unique solution to this fixed point, and consequently there is a unique monotone equilibrium with the Fischer information cost. This result, due to Morris and Yang (2019), holds because the Fisher information cost satisfies those authors' "infeasible perfect discrimination" condition, meaning that it assigns infinite cost to signals $p_{invest,x}$ that are not absolutely continuous in x.

[27] The optimal values of l_H and l_L depend on x^*, so proving this statement formally requires the consideration of a fixed point problem. Because the ordering of the states is irrelevant under mutual information, there are also many non-monotone equilibria. See Yang (2015) for details.

4.4 Linear-Quadratic Gaussian Environments

In our last application, we consider the classic "Linear-Quadratic-Gaussian" (LQG) tracking problem. Mutual information is known to be quite convenient in this setting, as it leads to a very tractable solution. We show that with the Fisher information cost, the solution remains equally tractable and leads to different conclusions.

For this application, we extend the continuous-state version of our model, with the multi-dimensional Fisher information cost, to the case of a continuous action space as well (though we do not formally prove convergence). An important conclusion is that, as with the mutual-information cost function, the optimal signal given a linear-quadratic payoff and a Gaussian prior will be a Gaussian signal. However, the precision of this Gaussian signal and (in the multi-dimensional case) the nature of the information it conveys will differ from mutual information. In particular, we will find that the Fisher information cost is linear in precision. Our approach thus provides foundations for a cost that has proven convenient in applications (e.g. Myatt and Wallace (2011) and Van Nieuwerburgh and Veldkamp (2010)).

Let the state space X be $\mathbb{R}^L$, with $L \geq 1$, and let the space of possible actions A be the real line $\mathbb{R}$. The DM's task is to track the variation in the state, with a reward given by $u_a(x) = -(\gamma^T x - a)^2$. In other words, the goal is to minimize the mean squared error of the DM's estimate of $\gamma^T x$, where γ is a non-zero vector that defines the payoff-relevant dimension of the state space.

We assume that the prior distribution over the state space X is a Gaussian distribution, with mean vector μ_0 and variance-covariance matrix Σ_0. Information costs are given by the multi-dimensional Fisher-information cost function, as in (13). Our problem is to choose the functions $\{p_a(x)\}_{a\in\mathbb{A}} \in \mathscr{P}_{LipG}(A)$ so as to minimize

$$V(q) = \int_X q(x) \int_A [p_a(x)(a-\gamma^T x)^2 + \frac{\theta}{4}\frac{|\nabla_x p_a(x)|^2}{p_a(x)}]dadx. \tag{17}$$

This is a problem in the calculus of variations. Our next proposition demonstrates that, if $\theta < 4|\Sigma_0\gamma|^2$, the optimal information structure is equivalent to observing a one-dimensional signal about some dimension of the state space. The fact that the optimal signal can be one-dimensional is a consequence of the usual result

that it is without loss of generality to equate signals with recommended actions.

Proposition 4. *In the linear-quadratic-Gaussian tracking problem defined in (17), if $\theta < 4|\Sigma_0\gamma|^2$, then the optimal choice of $p_a(x)$ satisfies*

$$p_a(x) = \frac{\sigma}{\sqrt{2\pi}}\exp(-\frac{\sigma^2}{2}(a-\gamma^T\mu_0-\sigma^{-2}\lambda^T(x-\mu_0))^2),$$

where $\sigma > 0$ is a constant satisfying

$$|(\Sigma_0^{-1}+\frac{4}{\theta}\sigma^{-2}I)^{-1}\gamma|^2 = \frac{\theta}{4}$$

and λ is a vector of length $|\lambda| = 2\theta^{-\frac{1}{2}}$ and direction

$$\frac{\lambda}{|\lambda|} \in \arg\max_{\hat{\lambda}:|\hat{\lambda}|=1} \hat{\lambda}^T(\Sigma_0^{-1}+\sigma^{-2}\lambda\lambda^T)^{-1}\gamma. \tag{18}$$

This $p_a(x)$ is identical to the conditional distribution of actions of a DM who observes a signal $s = \lambda^T x + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$, and then chooses her action optimally.

Proof. See the Appendix, Section C.7. □

If the DM observes the signal described in this proposition, her expectation of the payoff-relevant state $\gamma^T x$ (and hence optimal action) is

$$E[\gamma^T x|s] = \underbrace{\gamma^T\mu_0}_{\text{prior}} + \underbrace{\frac{\gamma^T\Sigma_0\lambda}{\lambda^T\Sigma_0\lambda}}_{\text{"beta" between } \gamma^T x \text{ and } \lambda^T x} \underbrace{\frac{\sigma^{-2}}{(\lambda^T\Sigma_0\lambda)^{-1}+\sigma^{-2}}(s-\lambda^T\mu_0)}_{\text{update on } \lambda^T x},$$

which given the particular optimal values of λ and σ simplifies to

$$E[\gamma^T x|s] = \gamma^T\mu_0 + \sigma^{-2}(s-\lambda^T\mu_0).$$

As a result, the action taken conditional on x is normally distributed, with mean

$$E[a|x] = \gamma^T\mu_0 + \sigma^{-2}\lambda^T(x-\mu_0),$$

the variance is $V[a|x] = \sigma^{-2}$, as implied by our solution for $p_a(x)$.

This simplification is due to the fact that $\lambda^T x$ maximally covaries with the payoff-relevant state $\gamma^T x$ under the DM's posterior after receiving the signal s. After receiving the signal s, the DM's posterior variance-covariance matrix on x is

$$V[x|s] = (\Sigma_0^{-1} + \sigma^{-2}\lambda\lambda^T)^{-1},$$

and by (18) the vector λ maximizes covariance with γ under this posterior. Letting I denote the identity matrix, an explicit formula for λ given σ is

$$\lambda = (\frac{\theta}{4}\Sigma_0^{-1} + \sigma^{-2}I)^{-1}\gamma.$$

This result is subtly different from what happens in the case of the mutual-information cost function. With a mutual-information cost function, the DM will choose to learn only about the payoff-relevant dimension of the state (λ will be a multiple of γ), and ignore all other information even when that information is correlated with the payoff-relevant state. In contrast, with the Fisher-information cost function, the DM chooses to receive a signal about a dimension of the state space that maximally covaries with the payoff-relevant dimension, and as a result will choose to receive information about dimensions of the state space that are correlated with the payoff-relevant dimension even when they are not directly payoff-relevant themselves. Hébert and La'O (2020) interpret this difference in the context of public signals, and demonstrate that this distinction leads to significantly different outcomes in coordination games ("beauty contests").

Because the optimal signal is conditionally Gaussian with a constant variance, we can rewrite the problem as a choice of the posterior covariance matrix Σ_s.

Corollary 3. *Let $\mathscr{M}_L$ be the set of $L \times L$ real symmetric positive-definite matrices. The value function $V(q)$ described in (17) can be written as*

$$V(q) = \inf_{\Sigma_s \in \mathscr{M}_L} \gamma^T \Sigma_s \gamma - \frac{\theta}{4} tr[\Sigma_s^{-1}] + \frac{\theta}{4} tr[\Sigma_0^{-1}],$$

subject to $\Sigma_s \preceq \Sigma_0$,

and the optimal policy in this problem is $\Sigma_s^* = (\Sigma_0^{-1} + \sigma^{-2}\lambda\lambda^T)^{-1}$*, where* σ *and* λ *are described as in Proposition 4.*

Proof. See the Appendix, Section C.8. □

That is, the DM chooses the variance-covariance matrix of her posterior to minimize errors subject to a cost that is proportional to the trace of the posterior precision matrix (and a "no-forgetting" constraint). This problem is a multi-dimensional analog of a problem in which costs are linear in precision. A similar result holds with mutual information, in which the trace in the above equation is replaced by the log determinant. Consequently, in the one-dimensional case, the two problems are identical up to the functional form of the precision cost. Even this difference can generate different predictions, as both Van Nieuwerburgh and Veldkamp (2010) and Myatt and Wallace (2011) discuss. In the multi-dimensional case, the two cost functions make more divergent predictions (Hébert and La'O (2020)).

Van Nieuwerburgh and Veldkamp (2010) observe that the trace-based information cost is not scale-invariant. That is, rescaling the state variable to $y = Fx$ for some invertible matrix F changes the value of the trace operator. This is a special case of the coordinate transformations discussed earlier, and should be interpreted as moving away from measuring perceptual distance in the same units across the various dimensions. In this case, the Fisher information definition must include a weight matrix, and the analog of Corollary 3 would involve $tr[F^T\Sigma_s^{-1}F]$ and $tr[F^T\Sigma_0^{-1}F]$. The optimal signals and policies, after accounting for the coordinate transformation, remain the same. Put another way, the dependence of the trace-based information cost on the scale of the variables is a reflection of the fact that the Fisher information cost incorporates a notion of perceptual distance. Conversely, the scale-invariance of mutual information is a reflection of the fact that the mutual information cost does not reflect a notion of perceptual distance.

In the solution described by Proposition 4, as θ approaches $4|\Sigma_0\gamma|^2$ from below, the optimal choice of σ diverges to infinity. That is, the DM's signal converges to something uninformative. Our next corollary shows that, as one might expect, if $\theta \geq 4|\Sigma_0\gamma|^2$, it is optimal for the information structure to be purely uninformative, and for the DM to choose an action $a = \gamma^T\mu_0$ regardless of the state.

Corollary 4. *In the linear-quadratic-Gaussian tracking problem defined in (17), if* $\theta \geq 4|\Sigma_0\gamma|^2$*, the optimal policy for the DM is to gather no information and choose* $a = \gamma^T \mu_0$ *with probability one.*

Proof. See the Appendix, Section C.9. □

The Fisher information cost function, like mutual information, allows for the possibility of a corner solution in which no attention at all is paid to some features of the environment, despite the fact that tracking them would allow the DM to achieve a higher level of welfare, and despite a finite information cost parameter θ.

5 Conclusion

In many applications of rational inattention, the space of exogenous states has a structure– for example, that of numbers ordered on a line. Imposing assumptions on the structure of the state space, and assuming a uniformly posterior-separable cost function, we have derived our neighborhood-based cost functions. These cost functions capture the idea that certain states are easier or harder to discriminate than others, and as a result are able to match experimental results on perception and on play in regime-change games. In contrast, the standard rational inattention cost function, mutual information, cannot match these results. Moreover, we have shown that the neighborhood-based cost functions and their continuous-state limit, the Fisher information cost function, remain tractable while making different predictions from those of mutual information in two leading applications of rational inattention: binary choice problems and linear-quadratic Gaussian problems.

References

Alexander Bloedel and Weijie Zhong. The cost of optimally-acquired information. *Unpublished Manuscript*, November 2020.

Charles W Calomiris and Joseph R Mason. Contagion and bank failures during the great depression: The june 1932 chicago banking panic. *The American Economic Review*, 87 (5):863–883, 1997.

Andrew Caplin and Mark Dean. The behavioral implications of rational inattention with Shannon entropy. *Unpublished manuscript*, August 2013.

Andrew Caplin, Mark Dean, and John Leahy. Rationally inattentive behavior: Characterizing and generalizing Shannon entropy. *Unpublished manuscript*, February 2019.

Nikolai Nikolaevich Chentsov. *Statistical decision rules and optimal inference*. Number 53. American Mathematical Soc., 1982.

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

Mark Dean and Nathaniel Neligh. Experimental tests of rational inattention. *Unpublished Manuscript*, June 2019.

Ambuj Dewan and Nathaniel Neligh. Estimating information cost functions in models of rational inattention. *Journal of Economic Theory*, 187:105011, 2020.

Ernst Fehr and Antonio Rangel. Neuroeconomic foundations of economic choice — recent advances. *Journal of Economic Perspectives*, 25(4):3–30, 2011.

George A Gescheider. Psychophysical scaling. *Annual review of psychology*, 39(1):169–200, 1988.

Benjamin Hébert and Jennifer La'O. Information acquisition, efficiency, and non-fundamental volatility. *Unpublished Manuscript*, February 2020.

Benjamin Hébert and Michael Woodford. Rational inattention when decisions take time. *Unpublished manuscript*, October 2019.

Frank Heinemann, Rosemarie Nagel, and Peter Ockenfels. Measuring strategic uncertainty in coordination games. *The Review of Economic Studies*, 76(1):181–221, 2009.

Mel Win Khaw, Ziang Li, and Michael Woodford. Cognitive imprecision and small-stakes risk aversion. *Review of Economic Studies*, forthcoming.

Bartosz Mackowiak, Filip Matejka, and Mirko Wiederholt. Rational inattention: A review. November 2020.

Stephen Morris and Hyun Song Shin. Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review*, pages 587–597, 1998.

Stephen Morris and Ming Yang. Coordination and continuous stochastic choice. *Available at SSRN 2889861*, 2019.

David P Myatt and Chris Wallace. Endogenous information acquisition in coordination games. *The Review of Economic Studies*, 79(1):340–374, 2011.

Maurice Obstfeld. Rational and self-fulfilling balance-of-payments crises. *The American Economic Review*, 76(1):72–81, 1986.

Luciano Pomatto, Philipp Strack, and Omer Tamuz. The cost of information. *arXiv preprint*, 1812.04211, December 2020.

Michael N. Shadlen et al. The speed and accuracy of a perceptual decision: A mathematical primer. In K. Doya et al., editors, *Bayesian Brain: Probabilistic Approaches to Neural Coding*. M.I.T. Press, 2007.

Anthony F Shorrocks. The class of additively decomposable inequality measures. *Econometrica: Journal of the Econometric Society*, pages 613–625, 1980.

Christopher A Sims. Rational inattention and monetary economics. *Handbook of Monetary Economics*, 3:155–181, 2010.

Michal Szkup and Isabel Trevino. Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior*, 124:534–553, 2020.

Stijn Van Nieuwerburgh and Laura Veldkamp. Information acquisition and underdiversification. *The Review of Economic Studies*, 77(2):779–805, 2010.

Felix A Wichmann and N Jeremy Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8):1293–1313, 2001.

Michael Woodford. Inattention as a source of randomized discrete adjustment. *Unpublished manuscript*, April 2008.

Michael Woodford. Inattentive valuation and reference-dependent choice. *Unpublished manuscript*, May 2012.

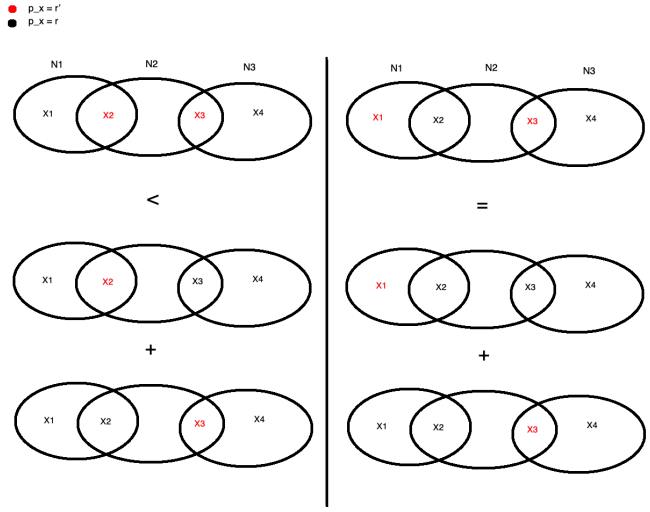
Michael Woodford. Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, 12:579–601, 2020.

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

Ming Yang. Optimality of debt under flexible information acquisition. *The Review of Economic Studies*, 87(1):487–536, 2020.

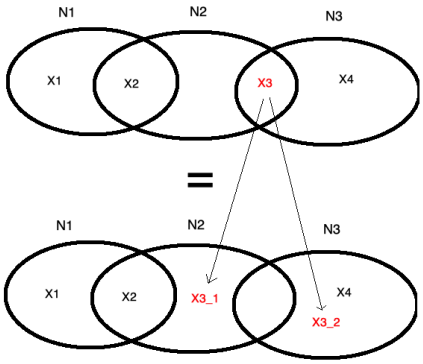
Figures

Figure 1: Diagram for Assumption 2



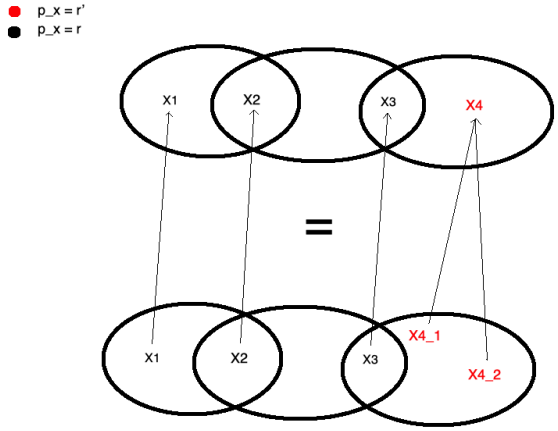
Notes: $X = \{X1, X2, X3, X4\}$ in this diagram. Each circle denotes a neighborhood, $\mathcal{N} = \{N1, N2, N3\}$. Under the signal structure p, red/gray colored states have signal distribution r', whereas black-colored states have signal distribution r. The left-hand side describes a situation in which the x and x' of Assumption 2 share a neighborhood, while the right-hand side describes a situation in x and x' do not share a neighborhood.

Figure 2: Diagram for Assumption 3



Notes: $X = \{X1, X2, X3, X4\}$ and $X' = \{X1, X2, X3\_1, X3\_2, X4\}$ in this diagram. Each circle denotes a neighborhood, $\mathcal{N} = \{N1, N2, N3\}$. The red/gray colored state is the one being "split." The arrows show how $q' \in \mathscr{P}(X')$ and $p' \in \mathscr{P}(S)^{|X'|}$ are constructed from $q \in \mathscr{P}(X)$ and $p' \in \mathscr{P}(S)^{|X|}$.

Figure 3: Diagram for Assumption 4



Notes: $X = \{X1, X2, X3, X3, X4\_1, X4\_2\}$ and $X'' = \{X1, X2, X3, X4\}$ in this diagram. Each circle denotes a neighborhood, $\mathcal{N} = \{N1, N2, N3\}$. Under the signal structure p, red/gray colored states have signal distribution r', whereas black-colored states have signal distribution r. The arrows show how $q \in \mathscr{P}(X)$ is assigned to $q'' = M(q)$ when the states $X4\_1$ and $X4\_2$ are merged.

Figure 4: Predicted response probabilities with a mutual-information cost function, for alternative values of the cost parameter θ.

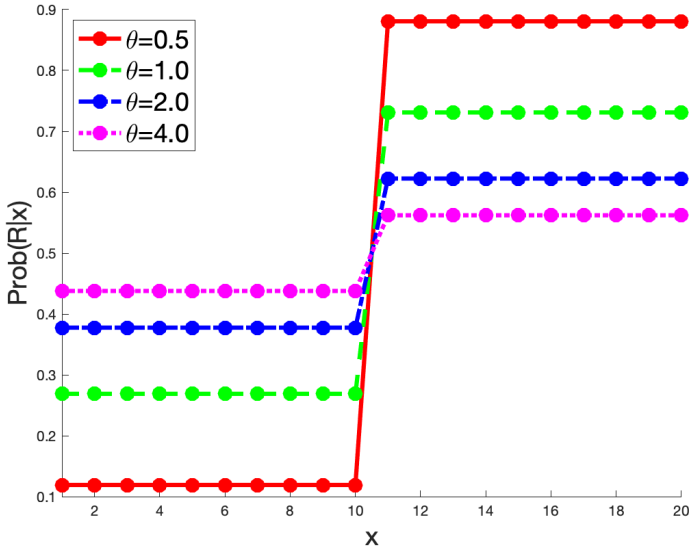
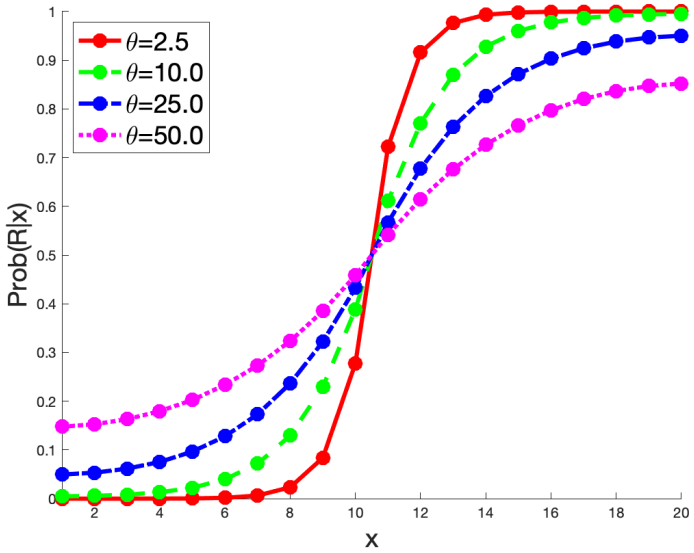


Figure 5: Predicted response probabilities with a generalized entropy neighborhood-based cost function with $\rho = 1$, in which each neighborhood consists only of two adjacent states.



Online Appendix

A Neighborhoods with Constant Marginal Costs

In this appendix section, we discuss an alternative assumption, described by Pomatto et al. (2020) as "constant marginal costs," the leads to a somewhat different local information cost, a version of the "total information" measure of Bloedel and

Zhong (2020). We first define what it means for a local information cost to exhibit constant, increasing, or decreasing marginal costs. Our definition follows Pomatto et al. (2020).

Consider two signal structures, p^1 and p^2. Define $p^1 \otimes p^2$ as the signal structure associated with receiving both signals, under the assumption that the signal realizations are independent conditional on the state $x \in X$. That is, $(p^1 \otimes p^2)_{s_1 s_2, x} = p^1_{s_1,x} p^2_{s_2,x}$. A DM who receives the signal structure $p^1 \otimes p^2$ can be thought of as observing both signals simultaneously or as sequentially receiving one signal and then the other (the equivalence of these two interpretations follows from uniform posterior separability).

We will say that a cost function exhibits increasing/constant/decreasing marginal costs if receiving both signals is more/equally/less costly than the sum of the costs of receiving the signals separately.

Definition 3. A cost function exhibits increasing marginal costs if, for all signal structures p^1, p^2 and all priors q_0,

$$C(p^1 \otimes p^2, q_0; S, X) \geq C(p^1, q_0; S, X) + C(p^2, q_0; S, X).$$

A cost function exhibits decreasing marginal costs if, for all signal structures p^1, p^2 and all priors q_0,

$$C(p^1 \otimes p^2, q_0; S, X) \leq C(p^1, q_0; S, X) + C(p^2, q_0; S, X).$$

A cost function exhibits constant marginal costs if it exhibits both increasing and decreasing marginal costs.

Note that neighborhood-based cost functions, by Assumption 2, always exhibit constant marginal costs with respect to signal structures p^1 and p^2 that provide information about states without any neighborhoods in common. When instead p^1 and p^2 both discriminate between states within some neighborhood, whether the cost function exhibits increasing, decreasing, or constant marginal costs is governed by the nature of the local information cost in that neighborhood. Note also that a cost function might not exhibit decreasing, increasing, or constant marginal costs,

if none of the above inequalities holds for all p^1, p^2, and q_0.

Intuitively, there is a connection between whether marginal costs are increasing or decreasing and the curvature of the information cost function. It is well-known that using Shannon's entropy leads to decreasing marginal costs, and in some applications of rational inattention this can lead to non-concavities and problems with equilibrium existence (Myatt and Wallace (2011)). Perhaps unsurprisingly, more curved cost functions can lead instead to increasing marginal costs. In the lemma below, we restate the familiar result for Shannon's entropy ($\rho = 1$ using the generalized entropy index), and show that with $\rho = 2$ the generalized entropy neighborhood-based cost functions exhibit increasing marginal costs.

Lemma 3. *The generalized entropy neighborhood-based cost function with* $\rho = 1$*,* $H_{NG}(q;1,X,\mathscr{N})$*, exhibits decreasing marginal costs. The generalized entropy neighborhood-based cost function with* $\rho = 2$*,* $H_{NG}(q;2,X,\mathscr{N})$*, exhibits increasing marginal costs.*

Proof. See the Appendix, Section C.10. □

Based on these results, it is tempting to speculate that $H_{NG}(q;\rho,X,\mathscr{N})$ will exhibit constant marginal costs for some $\rho \in (1,2)$. Instead, we show that it is the sum of the $\rho = 1$ and $\rho = 2$ generalized entropy neighborhood-based cost functions that exhibits constant marginal costs. This property arises because information costs in this case are linear in the prior, implying (under the UPS assumption) that the expected cost of the receiving the signals sequentially is exactly equal to the cost of receiving them simultaneously.

Proposition 5. *Suppose a neighborhood-based cost function* $H(q;X,\mathscr{N})$ *exhibits constant marginal costs. Then the local information costs are proportional to, for all* q_i *in the interior of the simplex,*

$$H^{CM}(q_i;|X_i|) = H^G(q_i;1,|X_i|) + H^G(q_i;2,|X_i|),$$

which simplifies to

$$H^{CM}(q_i;|X_i|) = \frac{1}{|X_i|}\sum_{j=1}^{|X_i|}\sum_{k=1}^{|X_i|} q_{i,j}\ln(\frac{q_{i,j}}{q_{i,k}}).$$

The static rational inattention problem can be written as

$$\begin{aligned} V_{CM}(q;X,\mathscr{N}) = &\max_{\{p_x\in\mathscr{P}(S)\}_{x\in X}} \sum_{s\in S}\pi_s(p,q_0)\hat{u}(q_s(p,q)) \\ &- \theta\sum_{i\in\mathscr{I}} c_i\bar{q}_i \sum_{x\in X_i}\sum_{x'\in X_i\setminus\{x\}} q_x D_{KL}(p_x||p_{x'}), \end{aligned}$$

where $\{c_i\in\mathbb{R}_+\}_{i\in I}$ *are positive constants.*

Proof. See the Appendix, Section C.11. The proof builds on results in Bloedel and Zhong (2020). □

Bloedel and Zhong (2020) characterize the set of UPS cost functions with constant marginal costs, which they call total information costs. This family includes the neighborhood-based cost functions with constant marginal costs (because the neighborhood based costs functions are uniformly posterior-separable). They show that any UPS cost function with constant marginal costs must satisfy

$$H^{TI}(q;X) = \sum_{x\in X}\sum_{x'\in X\setminus\{x\}} \gamma_{x,x'}q_x\ln(\frac{q_x}{q_{x'}}) \tag{19}$$

for some non-negative constants $\gamma_{x,x'}$.[28]

It is immediately apparent that any total information costs with $\gamma_{x,x'} = \gamma_{x',x}$ can be interpreted as a neighborhood-based information cost. The simplest way to do this is by defining the set of neighborhoods $\{X_i\}$ to be the set of all pairwise combinations of states in X, in which case $c_i = 2\gamma_{x,x'}$ for the states $\{x,x'\} = X_i$. These

[28]This class of cost functions is also a special case of the class of LLR cost functions defined by Pomatto et al. (2020), a larger family of cost functions that exhibit constant marginal costs. (See their Proposition 8.) Thus there is a non-empty intersection between our neighborhood-based costs and the class of LLR cost functions defined by Pomatto et al. (2020), though neither class is entirely contained in the other.

c_i constants are by definition non-negative, and hence satisfy the only restrictions required for the neighborhood-based cost functions.

Constant marginal costs is an appealing assumption if the signal structures p^1 and p^2 are interpreted as experiments (as in Pomatto et al. (2020)), because it seems natural to assume that if each experiment has a cost, doing both experiments should have a cost equal to the sum of the two costs. Under this interpretation, the constant marginal costs property can be thought of as "constant returns to scale."

However, the constant marginal costs assumption also pins down the curvature of the local information costs. An immediately corollary of Proposition 5 and (8) is that the elasticity of choice probabilities to incentives assuming constant marginal costs will be a weighted average of the elasticity with the $\rho = 2$ generalized entropy index and the Shannon entropy case ($\rho = 1$). Although this elasticity will be lower than the elasticity in the Shannon entropy case, it will be higher than the elasticity for the generalized entropy index with $\rho \approx 13$, which was found by Dean and Neligh (2019) to best describe their data. Moreover, it is not a priori obvious that marginal costs should be constant in the context of discriminating between neighboring states. It is intuitive that a DM might find it difficult to make sharp distinctions between neighboring states, and that the marginal difficulty of such distinctions might increase the more precisely the DM attempts to discriminate between the states. That is, local information costs might be characterized by increasing, and not constant, marginal costs.

B Security Design with Fisher Information

In this appendix section we consider the security design model with adverse selection in Yang (2020),[29] which builds on the results concerning binary choice described in the main text. The purpose of this application is two-fold. First, we illustrate another distinction between mutual information and Fisher information, concerning kinked payoffs rather than discrete jumps in payoffs. Second, we illustrate how our results on binary choice described previously can be incorporated into

[29] Our neighborhood cost function could also be applied in the same fashion to the model of security design with moral hazard in attention described in the appendix of Hébert (2018).

other problems.

Let $X = [0, \bar{x}]$ be the value of some assets, and let $s : X \to \mathbb{R}_+$ be a security that offers a payoff of $s(x)$ when the underlying assets value is x. Consider the problem of a risk-neutral buyer (the DM) who is presented with a take-it-or-leave-it offer of this security at a price K. The actions available to the buyer are to accept or decline this offer, $A = \{accept, decline\}$. Normalizing the utility of declining the offer to zero, the utility of accepting the security s at price K is

$$u_{accept}(x; s, K) = s(x) - K.$$

The seller and buyer share a common prior $q \in \mathscr{P}_{LipG}(X)$.

The problem facing this buyer fits exactly into the binary choice framework. As a result, we can determine whether the buyer will always decline, always accept, or pursue an interior strategy using the results of Proposition 3 (for Fisher information) and Woodford (2008)/Yang (2020) (for mutual information and variants thereof).

Let us suppose the security being offered is a debt security,

$$s(x; x^*) = \min\{x, x^*\}$$

for some $x^* \in (0, \bar{x})$. This security is continuous; by the results of Proposition 3, with the Fisher information cost the buyer will optimally choose a continuously twice-differentiable (and hence continuously differentiable) probability of acceptance $p^*_{accept,FI}(x)$, assuming the buyer chooses to gather some information. In contrast, by the results of Woodford (2008), with mutual information and again assuming the buyer gathers some information,

$$p^*_{accept,MI}(x) = \frac{\pi \exp(\theta^{-1} \min\{x, x^*\} - \theta^{-1} K)}{1 - \pi + \pi \exp(\theta^{-1} \min\{x, x^*\} - \theta^{-1} K)}$$

for some $\pi \in (0, 1)$. This probability of acceptance is kinked, and is in fact constant for all $x \geq x^*$, a property it inherits from the payoff function. The particular prediction that $p^*_{accept,MI}(x)$ is constant on $x \geq x^*$ is both testable and familiar from the regime change and perceptual experiment examples discussed in the main text, and is rejected in experimental evidence. More generally, mutual information and

Fisher information will generate different predictions in all binary choice problems with kinked but continuous payoffs– this difference is not specific to security design.

Let us now turn to the problem of a security designer/seller who wishes to design s and then offer it to the buyer described above at the price K. This problem is analyzed by Yang (2020), who shows that when the buyer's information cost is proportional to mutual information,[30] it is optimal for the seller to offer a debt security. This result holds regardless of whether it is optimal for the seller to induce the buyer to always accept the security or induce the buyer to gather information.

We will first discuss the case in which the seller wishes to avoid information gathering by the buyer. To motivate trade, we follow Yang (2020) and assume that the seller retains whatever asset value is not sold to the buyer, $x - s(x)$, and discounts these cashflows at a rate of $\beta \in (0,1)$. Let S be the class of feasible security designs; for this application, we will require that securities satisfy limited liability and be "doubly monotone," meaning that $s(x)$ and $x - s(x)$ are both non-negative, non-decreasing functions of x. Note that this assumption is common in the security design literature but is not imposed by Yang (2020).

The problem of the seller is to choose s and K to maximize her payoff, subject to the constraint that the buyer does not acquire information (characterized in Proposition 3):

$$\max_{s \in S, K \in \mathbb{R}} \int_X q(x)(K - \beta s(x))dx$$

subject to

$$\inf_{p_L \in \{p \in C^1(X,(0,\infty)) : \int_X q(x)p(x)dx = 1\}} \int_X q(x)p_L(x)(s(x) - K)dx + \frac{\theta}{4}\int_X q(x)\frac{(p_L'(x))^2}{p_L(x)}dx \geq 0.$$

It is immediately apparent that the constraint will bind; if it did not, the seller could increase the offering price K until the constraint was binding. Consequently, we can use the results of Lemma 2 to define a necessary and sufficient condition for

[30] Yang (2020) proves this result for a general class of state-separable information costs that includes mutual information but does not include Fisher information.

the buyer to choose not to acquire information.

The resulting security design problem can be analyzed using Hamiltonian methods. Because the s must be doubly monotone and $s(0) = 0$ by limited liability, we can think of $s'(x) = v(x)$ as the control variable. Using the results of Lemma 2, the state vector $(s(x), \psi(x))$ must evolve as

$$\frac{d}{dx}\left(\begin{bmatrix} s(x) \\ \psi(x) \end{bmatrix}\right) = \begin{bmatrix} v(x) \\ \frac{1}{2\theta}(s(x) - K) + \frac{1}{4}\psi(x)^2 + \frac{q'(x)}{q(x)}\psi(x) \end{bmatrix}.$$

The Hamiltonian, treating the price K as given, is

$$\begin{aligned} H(s, \psi, v, \lambda_1, \lambda_2, x; K) = q(x)(K - \beta s) + \lambda_1 v \\ + \lambda_2\left(\frac{1}{2\theta}(s - K) + \frac{1}{4}\psi^2 + \frac{q'(x)}{q(x)}\psi\right), \end{aligned}$$

noting that the choice of v is restricted to the interval $[0, 1]$. The relevant boundary conditions are $\psi(0) = \psi(\bar{x}) = 0$ (from Lemma 2), $s(0) = 0$ (from limited liability), and the free boundary condition associated with $s(\bar{x})$, $\lambda_1(\bar{x}) = 0$. The full problem must also consider the optimal price,

$$K^* \in \arg\max_{K \in \mathbb{R}} \int_X H(s^*(x;K), \psi^*(x;K), v^*(x;K), \lambda_1^*(x;K), \lambda_2^*(x;K); K)dx.$$

Given this description of the problem, it is straightforward to numerically compute the optimal security design. However, for this particular Hamiltonian system, we are able to analytically characterize the optimal security design.

Proposition 6. *Suppose the buyer's cost of information acquisition is proportional to the Fisher information cost function. The optimal doubly-monotonic, limited liability security design of seller who wishes to avoid information acquisition is a debt security, $s^*(x) = \min\{x, x^*\}$ for some $x^* \in (0, \bar{x}]$.*

Proof. See the Appendix, section C.12. □

This proposition demonstrates that the results of Yang (2020) for mutual information are robust to using the Fisher information cost under the additional restric-

tion that the security be doubly-monotonic.

The Hamiltonian approach outlined here can be readily modified to cover the case without the double-monotonicity requirement (in which $s(x)$ is a control variable instead state variable) as well as the case in which the buyer acquires information (in which $p(x)$ and $p'(x)$ replace $\psi(x)$ as state variables). The security design problem can also be analyzed on a discrete state space, using one of the neighborhood-based cost functions described in the text. In the next appendix subsection, we quantitatively analyze the some of the other cases discussed in Yang (2020) (when the seller induces information acquisition by the buyer, and without monotonicity constraints) in the discrete state case.

B.1 Security Design with Neighborhood-Based Cost Functions

In this appendix section, we numerically analyze the security design problem described above. For this section, we will use a finite state space, instead of a continuous one. The purpose of this section is to show that neighborhood-based cost functions remain tractable (at least computationally) in this application. We will briefly summarize the environment for the discrete state case.

Let X be a finite set of states. A seller offers a security $s \in \mathbb{R}_+^{|X|}$, whose payoffs are contingent on the realized value of the assets backing the security, $x \in X \subset \mathbb{R}_+$, to a buyer at a price K. The buyer's problem is to gather information about which asset values $x \in X$ are most likely and then accept ("like," L) or reject (R) this take-it-or-leave it offer. Both parties are risk-neutral, and the seller discounts the cashflows by a factor $\beta \in (0,1)$, relative to the buyer. The security is constrained by limited liability, $0 \leq s_x \leq x$. Let S_{LL} be the set of limited liability securities, and let $S \subset S_{LL}$ be the set of limited liability that are doubly monotone (as described above). The seller designs the security and offers a price,

$$\max_{s \in S_{LL}, K \in \mathbb{R}} \pi_L(s,K) q_L(s,K)^T (K\iota - \beta s),$$

where ι is a vector of ones, possibly subject to the monotonicity constraint $s \in S$. In this expression, $\pi_L(s,K)$ and $q_L(s,K)$ are the optimal policies of the buyer who

solves the rational inattention problem of (1), with $A = \{L, R\}$,

$$V(q_0; s, K) = \max_{\pi_L \in [0,1], q_L, q_R \in \mathscr{P}(X)} \pi_L q_L^T \cdot (s - K\iota) - \theta \pi_L D_H(q_L || q_0) - \theta(1 - \pi_L) D_H(q_R || q_0),$$

subject to the constraint that $\pi_L q_L + (1 - \pi_L) q_R = q_0$.

We explore, numerically, how the result of Yang (2020) on the optimality of debt with the mutual information cost function changes with alternative Bregman divergence cost functions (which are defined by the $H(\cdot)$ functions). We consider three alternatives, a generalized entropy index neighborhood-based function (Definition 2) with a pairwise neighborhood structure (as in section 3.3), a generalized entropy index cost function (i.e. a neighborhood cost function with only one neighborhood), and a "weighted" Shannon's entropy. Weighted Shannon's entropy is

$$H_w(q) = \sum_{x \in X} (e_x^T w)(e_x^T q) \ln(\frac{e_x^T q}{\iota^T q}),$$

where w is a vector of weights. Constant weights correspond to Shannon's entropy.

Summarizing our results, we replicate numerically the proof of Yang (2020) that, with mutual information, the optimal security design is always a debt. In contrast, for weighted mutual information and the generalized entropy index, the shape of the security design depends on the weights and the prior, respectively. The neighborhood cost function, on the other hand, appears to always generate the same shape irrespective of the prior.

Below, we describe our calculation procedure, and the parameters we use to generate figures 6 and 7 below, which show the optimal securities when s is not and is required to be doubly monotone, respectively. Our choice of parameters is guided by a desire to illustrate the differences between the cost functions, and to ensure that acceptance is not certain ($\pi_L < 1$). Our numerical calculation uses the

first-order approach,[31] solving

$$\max_{s\in S_{LL},K\in\mathbb{R},\pi_L\in[0,1],q_L\in\mathscr{P}(X)} \pi_L q_L^T(K\iota-\beta s)$$

subject to the buyer's first order condition and that beliefs remain in the simplex,

$$\begin{aligned} s-K\iota+\theta H_q(q_0-\pi_L q_L)&=\theta H_q(\pi_L q_L),\\ e_x^T(q_0-\pi_L q_L)&\geq 0,\ \forall x\in X. \end{aligned}$$

and the monotonicity constraints (if applicable). Combining the first-order conditions of this security design problem and the limited liability constraints,

$$\begin{aligned} (1-\beta)s^* = {}&\theta H_q(q-\pi_L^* q_L^*)-\theta H_q(\pi_L^* q_L^*)+\\ &+\theta[H_{qq}(q-\pi_L^* q_L^*)+H_{qq}(\pi_L^* q_L^*)](\beta\pi_L^* q_L^*-\lambda+\nu), \end{aligned}$$

where λ and ν are the multipliers on the limited liability constraints. This illustrates that the optimal security design is determined by the H function, subject to the caveat that $\pi_L^* q_L^*$ is endogenous.

Our numerical experiment uses an X with twenty-one states, with values of x evenly spaced from 0 to 10. We use a seller β of 0.5, and prior q that is an equal-weighted mixture of a uniform and binomial (21 outcomes of a 50-50 coin flip) distribution. We have chosen these parameters to help illustrate the differences between the cost functions.[32]

For the generalized entropy and neighborhood-based cost functions, we use $\rho=13$. This value is close to the estimated parameter of Dean and Neligh (2019) for these two cost functions, although there is no particular reason to apply parameters estimated for perceptual experiments to security design. The various cost functions

[31] We conjecture, but have not proven, that the first-order approach is valid in this context.

[32] In particular, the effects of weighted vs. standard Shannon's entropy are proportional to $\ln(\beta)$, so we choose a value of β significantly different from one. The differences between the generalized entropy index and Shannon's entropy disappear with a uniform prior, so we use the binomial part of the prior to highlight those differences. At the same time, it is helpful for numerical purposes to ensure the prior is significantly different from zero in each state, which is why we have the uniform part of the prior.

are not of the same "scale," so the same values of θ do not necessarily result in the securities of the same scale. We have chosen $\theta = \frac{1}{2}$ for Shannon's entropy, $\theta = 1$ for weighted Shannon's entropy and the neighborhood cost function, and $\theta = \frac{1}{50}$ for the generalized entropy function, which results in securities that are of the same scale but distinct in our graphs. For our weighted Shannon's entropy, we use

$$w(x) = \frac{3}{2} + \frac{x}{10}.$$

This linear weight structure assumes that it is more costly for the buyer to learn about good states than about bad states. We will see that this induces the seller to offer the buyer more in good states, and hence makes the buyer's security more equity-like. The more general point is that almost any security design could be reverse-engineered as optimal given some weight matrix. This reinforces the need to consider what kinds of information costs are reasonable.

Our numerical results are shown in figures 6 and 7. The first of these shows the optimal security designs, the second the optimal doubly monotone security designs. Our numerical calculations recover the result of Yang (2020) for the case of Shannon's entropy. They also illustrate our point that, with upward-sloping weights, the result for weighted Shannon's entropy is equity-like. The "inverse hump-shape" of the optimal security with the generalized entropy index cost function is caused by the "hump-shape" of the prior.[33] The optimal securities for mutual information and weighted mutual information are monotone, and hence do not differ between the two graphs, whereas the optimal securities for the neighborhood based cost function and (imperceptibly) the generalized entropy index are non-monotone, and hence do differ. For weighted mutual information and the generalized entropy index, monotonicity or a lack thereof is not guaranteed, as the shape of the optimal security depends on the weights and prior, respectively.

Our results for the neighborhood cost function appear, regardless of parameters, to result in the same "debt-like," but non-monotone, optimal security. This security is non-monotone and rapidly changing in one area. Rapid changes in security values would cause rapid changes in buyer behavior with Shannon's entropy, and hence be

[33] With a uniform prior, the optimal security with the generalized entropy index cost is also a debt.

sub-optimal, but this is not the case with neighborhood cost functions. As a result, it is possible for the optimal security to have rapid changes.[34] However, when we restrict the security to be monotone, the optimal security is a debt, suggesting that the result of Yang (2020) is robust to using neighborhood cost functions (but not the other two alternatives) under this additional restriction. This is consistent with our result in the previous appendix section, which shows the optimality of debt among monotone securities for the acceptance with certainty case on a continuous state space. We discuss this case with a discrete state space next.

Suppose the seller designs the security to induce the buyer to accept with probability one. In other words, the buyer's "consideration set" in his rational inattention problem consists only of L, instead of both L and R. As mentioned above, we have chosen the parameters of our numerical example to ensure that, for all of the cost functions, the seller is better off inducing information acquisition ($\pi_L < 1$) than avoiding information acquisition ($\pi_L = 1$). Note that the $\pi_L = 0$ case is equivalent to trading a "nothing" security at zero price, and hence assuming $\pi_L > 0$ is without loss of generality.

We will begin by restating the acceptance with certainty problem for the discrete state case (the problem for the continuous state case is described in the text). Consider the buyer's problem,

$$\begin{aligned} V(q;s,K) = \max_{\pi_L \in [0,1], q_L, q_R \in \mathscr{P}(X)} & \pi_L q_L^T (s - K\iota) \\ & - \theta \pi_L D_H(q_L||q) - \theta(1-\pi_L) D_H(q_R||q), \end{aligned}$$

subject to the constraint that $\pi_L q_L + (1-\pi_L) q_R = q$. Rewrite the choice variables as $\hat{q}_L = \pi_L q_L$ and $\hat{q}_R = (1-\pi_L) q_R$, and use the homogeneity of the H function, so

[34] Sharp-eyed readers might notice a second feature of the optimal security for neighborhood-based cost functions: the "flat" part isn't exactly flat. This feature arises from the "tri-diagonal" nature of the information cost matrix function $k(q)$, which leads to a difference equation describing the optimal security. As the number of states increases, the "flat" part of the security becomes increasingly flat. In the continuous state case, the difference equation becomes a differential equation, and we conjecture that the flat part is truly flat.

that the problem is

$$V(q;s,K) = \max_{\hat{q}_L,\hat{q}_R \in \mathbb{R}_+^{|X|}} \hat{q}_L^T(s-K\iota) - \theta D_H(\hat{q}_L||q) - \theta D_H(\hat{q}_R||q),$$

subject to $\hat{q}_L + \hat{q}_R = q$. Observe that the objective is concave and the constraints linear, so it suffices to consider local perturbations.

Suppose that it is optimal to set $\pi_L = 1$, implying $\hat{q}_L = q$. Consider a perturbation to $\hat{q}_L = q - \varepsilon q_R$, $\hat{q}_R = \varepsilon q_R$, for any arbitrary $q_R \in \mathscr{P}(X)$. For such a perturbation to reduce utility, we must have

$$-\varepsilon q_R^T(s-K\iota) - \theta D_H(q-\varepsilon q_R||q) - \theta\varepsilon D_H(q_R||q) \leq 0.$$

Taking the limit as $\varepsilon \to 0^+$, we must have, for all q_R, and hence for the minimizer,

$$\min_{q_R \in \mathscr{P}(X)} q_R^T(s-K\iota) + \theta D_H(q_R||q) \geq 0.$$

Note that this condition closely resembles the problem for the continuous state case above (Proposition 3).

If this condition is satisfied, it is at least weakly optimal for the buyer to choose $\pi_L = 1$ and gather no information. Consequently, the Lagrangian version of the optimal security design problem, subject to the constraint of inducing no information acquisition, is

$$\max_{\{s\in\mathbb{R}_+^{|X|},K\geq 0\}} \min_{\{\lambda\geq 0, q_R\in\mathscr{P}(X), \omega\in\mathbb{R}_+^{|X|}\}} q^T(K\iota-\beta s) + \lambda(q_R^T(s-K\iota) + \theta D_H(q_R||q)) + \omega^T(v-s),$$

where λ is the multiplier on the no-information-gathering constraint, $v \in \mathbb{R}^{|X|}$ is a vector with $v_x = x$, and ω is the multiplier on the upper-bound of the limited liability requirement.

Defining $\tilde{q}_R = \lambda q_R$, the dual of this problem is

$$\min_{\tilde{q}_R \in \mathbb{R}_+^{|X|}, \omega \in \mathbb{R}_+^{|X|}} \max_{s \in \mathbb{R}_+^{|X|}, K \geq 0} q^T(K\iota - \beta s) + \tilde{q}_R^T(s - K\iota) + \theta D_H(\tilde{q}_R||q) + \omega^T(v - s),$$

which can be understood as

$$\min_{\tilde{q}_R \in \mathbb{R}_+^{|X|}, \omega \in \mathbb{R}_+^{|X|}} \theta D_H(\tilde{q}_R||q) + \omega^T v,$$

subject to

$$\tilde{q}_R - \beta q - \omega \leq 0,$$

$$1 - q_R^T \iota \leq 0.$$

The multipliers of this convex minimization problem are the optimal security design and price. After solving the problem for $\tilde{q}_R$ and ω, we can use the first-order condition to recover the security design:

$$s - K\iota = H_q(q) - H_q(\tilde{q}_R).$$

We use the convention that in the lowest state, the asset value is zero, and therefore $s_0 = 0$, and hence

$$s_x = (e_x - e_0)^T (H_q(q) - H_q(\tilde{q}_R)),$$

where e_x and e_0 are basis vectors associated with the states $x \in X$ and $0 \in X$.

To implement the problem with the additional requirement of monotonicity for the security design, write the monotonicity requirement as $Ms \gg 0$, where M is an $|X| - 1 \times |X|$ matrix. The dual problem is

$$\min_{\tilde{q}_R \in \mathbb{R}_+^{|X|}, \omega \in \mathbb{R}_+^{|X|}, \rho \in \mathbb{R}_+^{|X|}} \theta D_H(\tilde{q}_R||q) + \omega^T v,$$

subject to

$$\tilde{q}_R - \beta q - \omega + M^T \rho \leq 0,$$

$$1 - q_R^T \iota \leq 0.$$

As mentioned above, under our parameters it is not optimal for the seller to avoid information acquisition. We first present the optimal securities that induce information acquisition. We then present the optimal securities that avoid information acquisition below. Note the shapes of these securities are very similar in these two cases, although the level is often quite different.

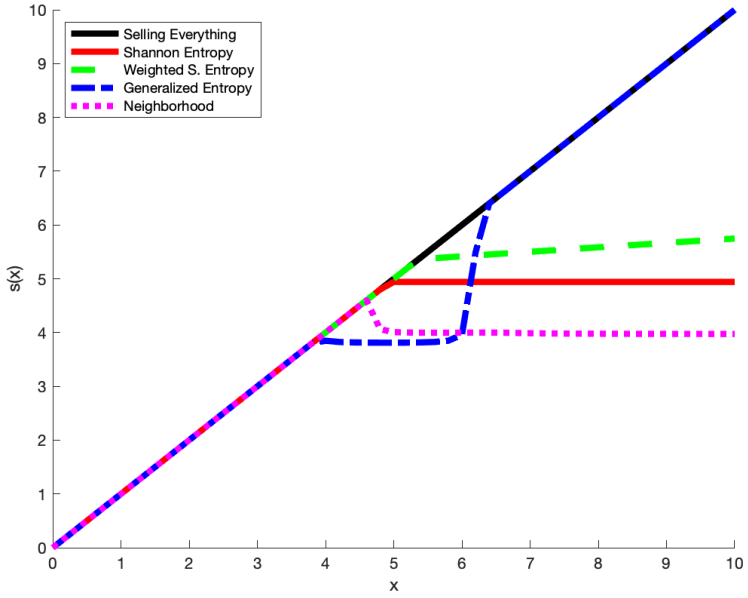


Figure 6: Optimal Security Designs

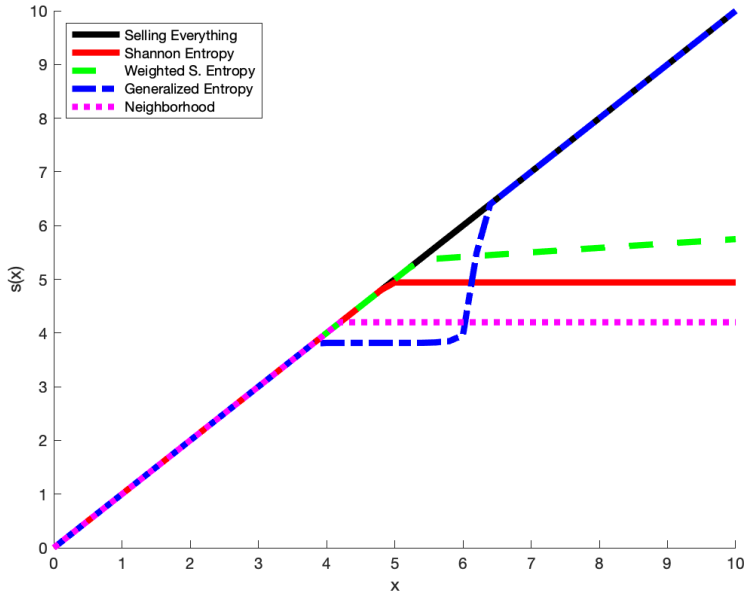


Figure 7: Optimal Monotone Security Designs

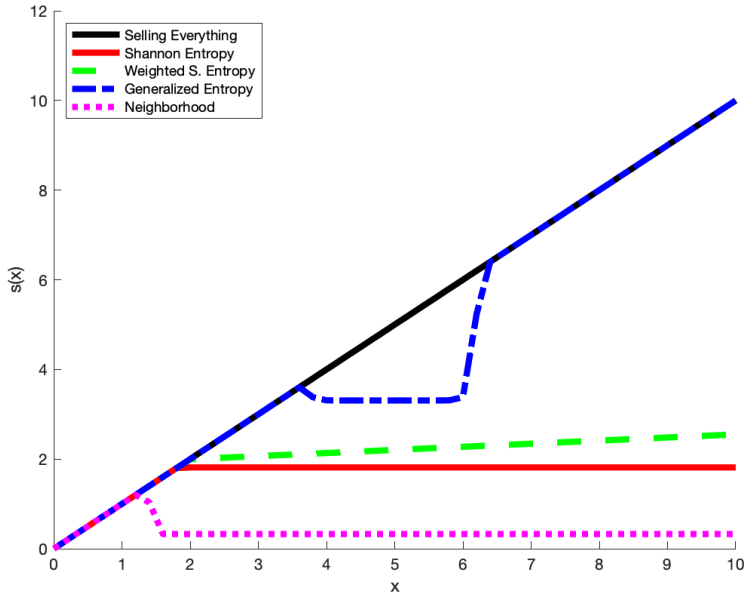


Figure 8: Optimal Security Designs that Avoid Info. Acquisition

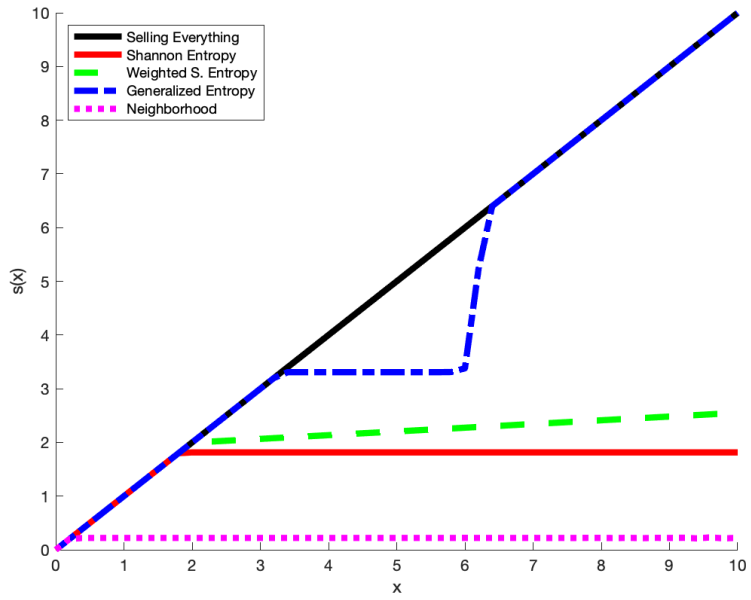


Figure 9: Optimal Monotone Security Designs that Avoid Info Acquisition

C Proofs

C.1 Proof of Proposition 1

We begin with the following lemma, which restates results in Hébert and Woodford (2019). For completeness, we include a proof of this lemma in the technical appendix.

Lemma 4. *Let $C(p,q_0;S,X)$ be any cost function satisfying Assumption 1 (i.e. any continuously twice-differentiable UPS cost function). Suppose that, for all $x \in X$,*

$$p_x = r + \varepsilon v_x$$

for some $\varepsilon > 0$, $r \in \mathscr{P}(S)$ with full support on S, and $v_x \in \mathbb{R}^{|S|}$, and that q_0 has full support on X. Then for the matrix-valued function

$$k(q) = Diag(q) \cdot H_{qq}(q;X) \cdot Diag(q),$$

where $Diag(q)$ is the diagonal matrix with q on its diagonal and $H_{qq}(q;X)$ is the Hessian of the H function associated with C,

$$C(p,q_0;S) = \frac{1}{2}\varepsilon^2 \sum_{x\in X, x'\in X} k_{x,x'}(q_0)\nu_x^T \cdot Diag(r)^{-1} \cdot \nu_{x'} + o(\varepsilon^2),$$

where $Diag(r)$ is a diagonal matrix with r on the diagonal and ι is a vector of ones.

Proof. See Hébert and Woodford (2019) or the Technical Appendix, section D.2.1. □

Consider now the signal structures p, p', and p'' defined in (6) and (7). Applying this lemma to those particular signal structures, with

$$r' = r + \varepsilon v,$$

we have

$$C(p,q_0;S,X) = \frac{1}{2}\varepsilon^2 k_{x,x}(q_0)\nu^T \cdot Diag(r)^{-1} \cdot \nu + o(\varepsilon^2),$$
$$C(p',q_0;S,X) = \frac{1}{2}\varepsilon^2 k_{x',x'}(q_0)\nu^T \cdot Diag(r)^{-1} \cdot \nu + o(\varepsilon^2),$$

and

$$\begin{aligned} C(p'',q_0;S,X) = &\frac{1}{2}\varepsilon^2 k_{x,x}(q_0)\nu^T \cdot Diag(r)^{-1} \cdot \nu + o(\varepsilon^2) \\ &+ \frac{1}{2}\varepsilon^2 k_{x',x'}(q_0)\nu^T \cdot Diag(r)^{-1} \cdot \nu + o(\varepsilon^2) \\ &+ \varepsilon^2 k_{x,x'}(q_0)\nu^T \cdot Diag(r)^{-1} \cdot \nu + o(\varepsilon^2). \end{aligned}$$

Consequently, by Assumption 2, if x and x' do not share a neighborhood in $\mathcal{N}$, $k_{x,x'}(q_0) = 0$, and if they do, $k_{x,x'}(q_0) < 0$. By definition, this property also applies to the Hessian matrix of H. That is, if x and x' share a neighborhood in $\mathcal{N}$, then

$$\frac{\partial^2}{\partial q_x \partial q_{x'}} H(q;X,\mathcal{N}) < 0,$$

and otherwise

$$\frac{\partial^2}{\partial q_x \partial q_{x'}} H(q;X,\mathscr{N}) = 0.$$

It follows that if x and x' do not share a neighborhood, then

$$\frac{\partial}{\partial q_x} H(q;X,\mathscr{N}) = \frac{\partial}{\partial q_x} H(q';X,\mathscr{N}) \tag{20}$$

for all measures q, q' that differ only in the mass on x'.

Let us now suppose we are given some $\bar{X}$ such that the $\bar{X}_i = \bar{X} \cap N_i$ are disjoint. Observe that it is without loss of generality to suppose H is strictly positive on $q \in \mathscr{P}(X)$ (shifting H by a constant does not change the cost function C). By the strict positivity and homogeneity of degree one of H, at least one partial derivative must be positive, and note that this must continue to hold even if we assume instead that H is only weakly positive. Consequently, by the General Theorem on Functional Dependence (see Leontief (1947) and Gorman (1968)), separability holds:

$$H(q;\bar{X},\mathscr{N}) = f(\hat{H}^1(q_1(q),\bar{q}_1(q)),\hat{H}^2(q_2(q),\bar{q}_2(q)),...),$$

where the $\hat{H}^i$ are continuously differentiable functions only of the values of q_x within the neighborhood $\bar{X}_i$ (and hence of q_i and $\bar{q}_i$), and f is a continuously differentiable function.

By the condition

$$\frac{\partial^2}{\partial q_x \partial q_{x'}} H(q;\bar{X},\mathscr{N}) = 0$$

for x, x' that do not share a neighborhood, the function f must be linear in its arguments. The constant term in f is irrelevant for cost function under Assumption 1, and hence without loss of generality we assume it is zero. We have concluded that $f(x) = \alpha x$ for some constant α, and without loss of generality to rescale the $\hat{H}^i$ functions and assume $\alpha = 1$. Therefore, we can write

$$H(q;\bar{X},\mathscr{N}) = \sum_{i \in \mathscr{I}} \hat{H}^i(q_i(q),\bar{q}_i(q);\bar{X}_i).$$

Under Assumption 1, the level of the cost functions $\hat{H}^i(q_i,\bar{q}_i;\bar{X}_i)$ has no impact

on the cost functions. We can therefore assume without loss of generality that $\hat{H}^i(q_i, 0; X_i) = 0$, consistent with the assumption of homogeneity of degree one for $H(q; X, \mathscr{N})$. Considering distributions that place all support within a single neighborhood, it follows that the $\hat{H}^i$ are homogenous of degree one in $\bar{q}_i$ and twice-differentiable in q_i. We can therefore write

$$H(q; \bar{X}, \mathscr{N}) = \sum_{i \in \mathscr{I}} \bar{q}_i(q) \hat{H}^i(q_i(q), 1; \bar{X}_i).$$

Let us now applying Assumption 3 in the $k = 1$ case. In this case,

$$C(p, q; S, \bar{X}) = C(p', q'; S, \bar{X}') = C(p, q; S, \bar{X}')$$

for any X', and hence the $\hat{H}^i$ function can depend on $\bar{X}_i$, holding fixed $|\bar{X}_i|$, only in ways that are irrelevant for information costs. It follows that it is without loss of generality to assume $\hat{H}^i$ depends only on the cardinality of $\bar{X}_i$.

As with standard UPS cost functions, any strictly increasing affine transformation of the $\hat{H}^i$ functions generates an equivalent cost function. It is therefore without loss of generality to assume they reach their minima at the uniform distribution, and also without loss of generality to extend $\hat{H}^i$ to the set of measures by assuming homogeneity of degree one.

This completes the proof for the case in which the X_i are disjoint. Let us now suppose we are given some X such that the X_i are not disjoint. Repeatedly applying Assumption 3, for any S, $p \in \mathscr{P}(S)^{|X|}$, and $q \in \mathscr{P}(X)$, there exists a $\bar{X}$, surjection $m : \bar{X} \to X$, $\bar{p} \in \mathscr{P}(S)^{|\bar{X}|}$, and $\bar{q} \in \mathscr{P}(\bar{X})$ such that the $\bar{X}_i$ are disjoint and

$$C(p, q; S, X) = C(\bar{p}, \bar{q}; S, \bar{X}),$$

where for all $\bar{x} \in \bar{X}$,

$$\bar{q}_{\bar{x}} = q_{m(\bar{x})},$$
$$\bar{p}_{\bar{x}} = p_{m(\bar{x})}.$$

Moreover, $x \in X_i$ if and only if there is exactly one $\bar{x} \in \bar{X}$ such that $\bar{x} \in \bar{X}_i$ and

$m(\bar{x}) = x$, and therefore $|X_i| = |\bar{X}_i|$.

Define

$$\hat{q}_s(p,q) = \pi(p,q)q_s(p,q)$$

and observe by Bayes' rule that

$$\hat{q}_{s,x}(p,q) = p_{s,x}q_x.$$

By Assumption 1 and homogeneity of degree one, and using the fact that the $\bar{X}_i$ are disjoint,

$$\begin{aligned} C(\bar{p},\bar{q};S,\bar{X}) &= -H(\bar{q};\bar{X},\mathscr{N}) + \sum_{s\in S} H(\hat{q}_s(\bar{p},\bar{q});\bar{X},\mathscr{N}) \\ &= \sum_{i\in\mathscr{I}} \{-\bar{q}_i(\bar{q})\hat{H}^i(q_i(\bar{q}),1;|\bar{X}_i|) + \sum_{s\in S} \bar{q}_i(\hat{q}_s(\bar{p},\bar{q}))\hat{H}^i(q_i(\hat{q}_s(\bar{p},\bar{q})),1;|\bar{X}_i|)\}. \end{aligned}$$

By definition,

$$\bar{q}_i(\bar{q}) = \sum_{\bar{x}\in\bar{X}_i} \bar{q}_{\bar{x}} = \sum_{\bar{x}\in\bar{X}_i} q_{m(\bar{x})} = \sum_{x\in X_i} q_x = \bar{q}_i(q),$$

and (assuming $\bar{q}_i(q) > 0$)

$$q_{i,\bar{x}}(\bar{q}) = \frac{\bar{q}_{\bar{x}}}{\bar{q}_i(\bar{q})} = \frac{q_{m(\bar{x})}}{\bar{q}_i(q)} = q_{i,m(\bar{x})}(q).$$

Hence it follows that

$$\bar{q}_i(\bar{q})\hat{H}^i(q_i(\bar{q}),1;|\bar{X}_i|) = \bar{q}_i(q)\hat{H}^i(q_i(q),1;|X_i|).$$

By a similar argument,

$$\bar{q}_i(\hat{q}_s(\bar{p},\bar{q})) = \sum_{\bar{x}\in\bar{X}_i} \bar{p}_{s,x}\bar{q}_x = \sum_{\bar{x}\in\bar{X}_i} p_{s,m(\bar{x})}q_{m(\bar{x})} = \sum_{x\in X_i} p_{s,m(\bar{x})}q_{m(\bar{x})} = \bar{q}_i(\hat{q}_s(p,q))$$

and

$$q_{i,\bar{x}}(\hat{q}_s(\bar{p},\bar{q})) = \frac{\bar{p}_{s,\bar{x}}\bar{q}_{\bar{x}}}{\bar{q}_i(\bar{q})} = \frac{p_{s,m(\bar{x})}q_{m(\bar{x})}}{\bar{q}_i(q)} = q_{i,m(\bar{x})}(\hat{q}_s(p,q)),$$

and therefore

$$C(p,q;S,X) = \sum_{i\in\mathscr{I}} \{-\bar{q}_i(q)\hat{H}^i(q_i(q),1;|X_i|) + \sum_{s\in S} \bar{q}_i(\hat{q}_s(p,q))\hat{H}^i(q_i(\hat{q}_s(p,q)),1;|X_i|)\}.$$

Consequently, it is without loss of generality to suppose that

$$H(q;X,\mathscr{N}) = \sum_{i\in\mathscr{I}} \bar{q}_i(q)\hat{H}^i(q_i(q),1;|X_i|),$$

concluding the proof.

C.2 Proof of Proposition 2

As argued in the proof of section C.1, it is without loss of generality to suppose that the neighborhoods are disjoint. It follows immediately by Assumption 4 that the Hessian matrix of H^i is invariant to all embeddings in the sense of Chentsov (1982) (see also Amari and Nagaoka (2007) or Hébert and Woodford (2019) for a discussion of this invariance). Consequently, by Theorem 11.1 in Chentsov (1982), the Hessian matrix is proportional to the Fisher matrix. Let c_i denote the constant of proportionality, and note by the convexity of H^i that it is weakly positive. Integrating the Hessian of H^i, it follows that H^i must be proportional to the negative of Shannon's entropy.

C.3 Proof of Lemma 1

First, note that if $\rho \geq 2$ and q_s does not have full support, then p_x will not have full support for the state x such that $e_x^T q_s = 0$, and we will have $D_\rho(p_x||pE_i^T q_i) = \infty$ for any i with $x \in X_i$, as required. For $\rho < 2$, continuity holds, and therefore both boundary cases are satisfied, provided the result holds for interior q_s.

To prove this claim, it is sufficient to show that, if all q_s are interior,

$$\sum_{i\in\mathscr{I}} c_i|X_i|^{1-\rho}\bar{q}_i(q)^{\rho-1}\sum_{x\in X_i}(q_x)^{2-\rho}D_\rho(p_x||\pi_i) = -H_N(q) + \sum_{s\in S}\pi_s(p,q)H_N(q_s(p,q)).$$

By definition,

$$\sum_{s\in S}\pi_s H_N(q_s)=\sum_{s\in S:\pi_s>0}\pi_s\sum_{i\in\mathscr{I}}c_i\bar{q}_i(q_s)\frac{1}{|X_i|}\frac{1}{(\rho-2)(\rho-1)}\sum_{x\in X_i}\{(\frac{q_{s,x}}{\frac{1}{|X_i|}\bar{q}_i(q_s)})^{2-\rho}-1\}.$$

Using Bayes' rule, $\pi_s\bar{q}_i(q_s)=\sum_{x\in X_i}p_{s,x}q_x=\bar{q}_i(q)\pi_{i,s}$, and therefore

$$\begin{aligned}\sum_{s\in S}\pi_s H_N(q_s)=&\sum_{i\in\mathscr{I}}c_i|X_i|^{1-\rho}\bar{q}_i(q)^{\rho-1}\frac{1}{(\rho-2)(\rho-1)}\sum_{x\in X_i}(q_x)^{2-\rho}\sum_{s\in S:\pi_{i,s}>0}\pi_{i,s}(\frac{p_{s,x}}{\pi_{i,s}})^{2-\rho}\\&-\sum_{i\in\mathscr{I}}c_i\bar{q}_i(q)\frac{1}{(\rho-2)(\rho-1)}.\end{aligned}$$

Therefore,

$$-H_N(q)+\sum_{s\in S}\pi_s H_N(q_s)=\sum_{i\in\mathscr{I}}c_i|X_i|^{1-\rho}\bar{q}_i(q)^{\rho-1}\sum_{x\in X_i}(q_s)^{2-\rho}D_\rho(p_x||\pi_i),$$

as required. The proof is essentially identical in the $\rho=1$ and $\rho=2$ cases.

C.4 Proof of Proposition 3

It is convenient to work with the transformed variable

$$y=G(x)=\int_{x_L}^{x}\frac{dx}{q(x)},$$

which is well-defined by the compactness of X and the full-support property of $q(x)$.

Using this change-of-variable,

$$\begin{aligned}V_N(q)=\max\{&\sup_{p_R\in C^1([0,\bar{y}],(0,1))}\int_0^{\bar{y}}g(y)p_R(y)u_R(y)dy-\frac{\theta}{4}\int_0^{\bar{y}}\frac{(p_R'(y))^2}{p_R(y)(1-p_R(y))}dy,\\&\int_0^{\bar{y}}g(y)u_R(y)dy,0\},\end{aligned}$$

where $\bar{y}=G(x_H)$ and $g(y)=q(G^{-1}(y))^2$.

A necessary condition for always-L to be the optimal strategy is that

$$\int_0^{\bar{y}} g(y)\hat{p}_R(y)u_R(y)dy - \frac{\theta}{4}\int_0^{\bar{y}} \frac{(\hat{p}'_R(y))^2}{\hat{p}_R(y)(1-\hat{p}_R(y))}dy \leq 0$$

where

$$\hat{p}_R(y) = \varepsilon p_R(y)$$

for some $\varepsilon > 0$ and all $p_R(y) \in C^1([0,\bar{y}],(0,1))$. Considering the limit as $\varepsilon \to 0^+$, we must have

$$\int_0^{\bar{y}} g(y)p_R(y)u_R(y)dy - \frac{\theta}{4}\int_0^{\bar{y}} \frac{(p'_R(y))^2}{p_R(y)}dy \leq 0.$$

Using a change of variable back to the x variable, this is

$$\int_X q(x)p_R(x)u_R(x)dx - \frac{\theta}{4}\int_X q(x)\frac{(p'_R(x))^2}{p_R(x)}dx \leq 0. \tag{21}$$

Now suppose this condition holds for all $p_R(x) \in C^1(X,(0,1))$. It must hold for p_R constant, and hence $\int_X q(x)u_R(x)dx \leq 0$, meaning that always-$L$ is preferred to always-R. Define the functional $J : C^1([0,\bar{y}],(0,1)) \to \mathbb{R}$ by

$$J[p] = \int_0^{\bar{y}} g(y)p(y)u_R(y)dy - \frac{\theta}{4}\int_0^{\bar{y}} \frac{(p'(y))^2}{p(y)(1-p(y))}dy \tag{22}$$

The following lemma demonstrates that this functional is concave on its domain.

Lemma 5. *The functional $J : C^1([0,\bar{y}],(0,1)) \to \mathbb{R}$ defined in (22) is concave on $C^1([0,\bar{y}],(0,1))$.*

Proof. See the Technical Appendix, Section D.3.2. □

Consequently, for all $p \in C^1([0,\bar{y}],(0,1))$, and $\varepsilon > 0$,

$$J[p] \leq J[\varepsilon p] + \delta J[\varepsilon p, p - \varepsilon p],$$

where $\delta J[\varepsilon p, p-\varepsilon p]$ is the first variation from εp in the direction $p-\varepsilon p$,

$$\begin{aligned}\delta J[\varepsilon p, p-\varepsilon p] = &\int_0^{\bar{y}} g(y)(1-\varepsilon)p(y)u_R(y)dy \\ &+\frac{\theta}{4}(1-\varepsilon)\int_0^{\bar{y}} \frac{(\varepsilon p'(y))^2}{(\varepsilon p(y)(1-\varepsilon p(y)))^2}(1-2\varepsilon p(y))p(y)dy \\ &-\frac{\theta}{2}(1-\varepsilon)\int_0^{\bar{y}} \frac{(\varepsilon p'(y))}{(\varepsilon p(y)(1-\varepsilon p(y)))}p'(y)dy.\end{aligned}$$

In the limit as $\varepsilon \to 0^+$,

$$J[p] \leq \int_0^{\bar{y}} g(y)p(y)u_R(y)dy - \frac{\theta}{4}\int_0^{\bar{y}} \frac{(p'(y))^2}{p(y)}dy$$

and consequently $J[p] \leq 0$. It follows that if (21) holds for all $p_R(x) \in C^1(X,(0,1)$, the optimal policy is always-L. Consequently, the condition is both necessary and sufficient. Moreover, observe that this condition will hold if and only if it holds for all $p_R \in C^1(X,(0,\infty))$ such that $\int_X q(x)p(x)dx = 1$, by the homogeneity of degree of the functional J.

By symmetry, the analogous condition for always-R is, for all $p_L \in \{p \in C^1(X,(0,\infty)) : \int_X q(x)p(x)dx = 1\}$,

$$-\int_0^{\bar{x}} q(x)p_L(x)u_R(x)dx - \frac{\theta}{4}\int_0^{\bar{y}} q(x)\frac{(p_L'(x))^2}{p_L(x)}dx \leq 0$$

If neither of these conditions hold, then it must be the case that

$$\sup_{p_R \in C^1([0,\bar{y}],(0,1))} J[p_R] > \max\{\int_0^{\bar{y}} g(y)u_R(y)dy, 0\}.$$

The space $C^1([0,\bar{y}],(0,1))$ is not compact, so the existence of a maximizer does not follow immediately from concavity. However, the following lemma demonstrates that a maximizer does in fact exist.

Lemma 6. *If*

$$\sup_{p_R \in C^1([0,\bar{y}],(0,1))} J[p_R] > \max\{\int_0^{\bar{y}} g(y)u_R(y)dy, 0\},$$

then there exists an extremal $p_R^* \in C^1([0,\bar{y}],(0,1))$ *that is a maximizer and is continuously twice-differentiable except at the discontinuities of* $u_R(y)$.

Proof. See the Technical Appendix, Section D.3.1. □

Anywhere it is twice-differentiable, p_R^* must satisfy the Euler-Lagrange equation,

$$q(x)u_R(x) + \frac{\theta}{4}q(x)\frac{(p_R^{*\prime}(x))^2}{(p_R^*(x)(1-p_R^*(x)))^2}(1-2p_R^*(x)) = -\frac{\theta}{2}\frac{d}{dx}[q(x)\frac{p_R^{*\prime}(x)}{p_R^*(x)(1-p_R^*(x))}],$$

along with the natural boundary conditions

$$q(0)\frac{p_R^{*\prime}(0)}{p_R^*(0)(1-p_R(0))} = q(\bar{x})\frac{p_R^{*\prime}(\bar{x})}{p_R^*(\bar{x})(1-p_R^*(\bar{x}))} = 0.$$

The Euler-Lagrange equation can be rewritten as

$$q(x)u_R(x) - \frac{\theta}{4}q(x)\frac{(p_R^{*\prime}(x))^2}{(p_R^*(x)(1-p_R^*(x)))^2}(1-2p_R^*(x)) + \frac{\theta}{2}q'(x)\frac{p_R^{*\prime}(x)}{p_R^*(x)(1-p_R^*(x))} = -\frac{\theta}{2}q(x)\frac{p_R^{*\prime\prime}(x)}{p_R^*(x)(1-p_R^*(x))}$$

and further simplified to

$$\frac{p_R^*(x)(1-p_R^*(x))}{2\theta}u_R(x) - \frac{1}{2}\frac{(p_R^{*\prime}(x))^2}{p_R^*(x)(1-p_R^*(x))}(1-2p_R^*(x)) + \frac{q'(x)}{q(x)}p_R^{*\prime}(x) = -p_R^{*\prime\prime}(x).$$

By the concavity of J, any extremal satisfying these conditions is a maximizer.

C.5 Proof of Lemma 2

We first prove the "if": suppose a function

$$q(x)\psi(x) = \int_{x_L}^{x} q(x')[\frac{2}{\theta}u_R(x') - \frac{1}{2}\psi(x')^2]dx'$$

satisfying $\psi(x_H) = 0$ exists. Observe that this function is continuous. Defining the functional

$$J[p] = \int_{x_L}^{x_H} F(x, p(x), p'(x))dx, \tag{23}$$

$$F(x, p, v) = q(x)u_R(x)p + \frac{\theta}{4}q(x)\frac{v^2}{p},$$

we will prove that the existence of ψ implies that

$$\inf_{p_L \in \{p \in C^1(X,(0,\infty)): \int_X q(x)p(x)dx = 1\}} J[p_L] = 0.$$

The integrated Euler-Lagrange equation associated with this functional is, for some constant c,

$$\frac{\theta}{2}q(x)\frac{p'(x)}{p(x)} = c + \int_{x_L}^{x} q(x')[u_R(x') - \frac{\theta}{4}(\frac{p'(x')}{p(x')})^2]dx'$$

and the natural boundary conditions are $\frac{\theta}{2}q(x_L)\frac{p'(x_L)}{p(x_L)} = \frac{\theta}{2}q(x_H)\frac{p'(x_H)}{p(x_H)} = 0$. Defining

$$\psi(x) = \frac{p'(x)}{p(x)}$$

demonstrates that if the function ψ exists, an extremal of the functional $J[p]$ on $p \in C^1(X,(0,\infty))$ exists,

$$p^*(x) = A\exp(\int_{x_L}^{x} \psi(x')dx')$$

for any constant $A > 0$.

We next invoke the following lemma to show that the functional $J[p]$ is convex

on $C^1(X,(0,\infty))$.

Lemma 7. *The functional $J : C^1(X,(0,\infty)) \to \mathbb{R}$ defined in (23) is convex on $C^1(X,(0,\infty))$.*

Proof. See the Technical Appendix, Section D.3.2. □

Consequently, the $p^*(x)$ are minimizers, and must achieve the same value of the functional for all values of A, which by the homogeneity of degree one of $J[p]$ must be zero. Hence, for the particular value of A satisfying

$$A^{-1} = \int_{x_L}^{x_H} q(x) \exp\left(\int_{x_L}^{x} \psi(x')dx'\right)dx,$$

the associated p^* must minimize $J[\cdot]$ on $\{p \in C^1(X,(0,\infty)) : \int_X q(x)p(x)dx = 1\}$ and have $J[p^*] = 0$.

We next prove the "only if", a proof that largely follows the arguments of the proof of Proposition 3. We will show that if

$$\inf_{p_L \in \{p \in C^1(X,(0,\infty)) : \int_X q(x)p(x)dx = 1\}} J[p_L] = 0, \tag{24}$$

then the function ψ must exist.

It is convenient to work with the transformed variable

$$y = G(x) = \int_{x_L}^{x} \frac{dx}{q(x)},$$

which is well-defined by the compactness of X and the full-support property of $q(x)$. Define $y_L = G(x_L)$ and $y_H = G(x_H)$ as the boundary points, and define $g(y) = q(G^{-1}(y))^2$.

Employing the change of variable

$$\phi(y) = \sqrt{p(y)},$$

we use the domain $[y_L - \varepsilon, y_H + \varepsilon]$ for some $\varepsilon > 0$, and define

$$\hat{J}[\phi] = \int_{y_L-\varepsilon}^{y_H+\varepsilon} F(y, \phi(y), \phi'(y))dy, \tag{25}$$

$$F(x, \phi, v) = \begin{cases} g(y)u_R(y)\phi^2 + \theta v^2 & y \in [y_L, y_H] \\ \theta v^2 & y \notin [y_L, y_H]. \end{cases}$$

If (24) holds, there must exist a sequence $\{\phi_n \in \{\phi \in C^1([y_L - \varepsilon, y_H + \varepsilon], (0,\infty)) : \int_{y_L}^{y_H} g(y)\phi(y)^2 dy = 1\}\}_{n=1}^{\infty}$ satisfying

$$\lim_{n\to\infty} \hat{J}[\phi_n] = 0.$$

The functions ϕ_n are elements of the Sobolev space $W^{1,2}([y_L - \varepsilon, y_H + \varepsilon], \mathbb{R})$. By definition, for any $\delta > 0$, there exists an n_0 such that for all $n > n_0$, $|\hat{J}[\phi_n]| < \delta$. Consequently,

$$\theta \int_{y_L-\varepsilon}^{y_H+\varepsilon} \phi_n'(y)^2 dy < \delta,$$

and

$$B \int_{y_L-\varepsilon}^{y_H+\varepsilon} \phi_n(y)dy < \delta,$$

where $B = \max_{y\in[y_L,y_H]} |g(y)u_R(y)|$. The sequence $\{\phi_n\}_{n=n_0}^{\infty}$ is therefore bounded in the $W^{1,2}$ norm, and hence converges weakly to some $\phi^* \in W^{1,2}([y_L - \varepsilon, y_H + \varepsilon], \mathbb{R})$, immediately implying that

$$\hat{J}[\phi^*] = 0$$

and

$$\int_{y_L}^{y_H} g(y)\phi^*(y)^2 dy = 1.$$

By the homogeneity of degree two of $\hat{J}$ and the observation that $F(y, \phi, v) = F(y, -\phi, v)$, ϕ^* must be a minimizer of $\hat{J}$ on $W^{1,2}([y_L - \varepsilon, y_H + \varepsilon], \mathbb{R})$, and it is without loss of generality to assume $\phi^*(y) \geq 0$ for all $y \in [y_L - \varepsilon, y_H + \varepsilon]$.

We invoke the following regularity result, proven in the technical appendix, to show that ϕ^* is continuously differentiable, and continuous twice differentiable on

any interval on which u_R is continuous.

Lemma 8. *If $\phi^* \in W^{1,2}([y_L - \varepsilon, y_H + \varepsilon], \mathbb{R})$ is a minimizer of the functional $\hat{J}$ defined in (25), then $\phi^* \in C^1([y_L - \varepsilon, y_H + \varepsilon], \mathbb{R})$, and ϕ^* is continuously twice-differentiable on any interval on which u_R is continuous.*

Proof. See the Technical Appendix, Section D.3.3. □

Let $y_1, \ldots y_{k-1}$ be the (possibly empty) set of points of discontinuity for u_R, and let $y_0 = y_L$ and $y_k = y_H$. This regularity result implies that the Euler-Lagrange equation,

$$2\theta\phi^{*\prime\prime}(y) = 2g(y)\phi^*(y)u_R(y)$$

must hold on all $y \in (y_{i-1}, y_i)$.

Suppose that for some $y' \in [y_L, y_H]$, $\phi^*(y') = 0$. By the fact that $\phi^*(y)$ is continuously differentiable and it is without loss of generality to assume $\phi^*(y) \geq 0$, it must be the case that $\phi^{*\prime}(y') = 0$. In this case, $\phi^*(y)$ constant on $y \in [y_L, y_H]$ satisfies the Euler-Lagrange equations. The system

$$\frac{d}{dy}\begin{bmatrix}\phi^{*\prime}(y)\\ \phi^*(y)\end{bmatrix} = \begin{bmatrix}\theta^{-1}g(y)u_R(y)\phi^*(y)\\ \phi^{*\prime}(y)\end{bmatrix}$$

is uniformly Lipschitz-continuous in $(\phi^*(y), \phi^{*\prime}(y))$ and continuous in y on all intervals (y_{i-1}, y_i), and hence by the Picard-Lindelof theorem, a unique solution to any initial value problem on any interval $[y_{i-1}, y_i]$ exists. Consequently, if $\phi^*(y') = 0$ for any $y' \in [y_L, y_H]$, $\phi^*(y) = 0$ for all $y \in [y_L, y_H]$.

But by the result that

$$\int_{y_L}^{y_H} q(y)\phi^*(y)^2 dy = 1,$$

$\phi^*(y)$ cannot be zero everywhere. It follows that $\phi^*(y) > 0$.

Defining $\hat{\psi} : [y_L, y_H] \to \mathbb{R}$ by

$$\hat{\psi}(y) = \frac{2\phi^{*\prime}(y)}{\phi^*(y)},$$

the Euler-Lagrange equation implies that everywhere u_R is continuous,

$$\hat{\psi}'(y) = 2\theta^{-1} g(y) u_R(y) - \frac{1}{2} g(y) \hat{\psi}(y)^2.$$

In the x variable, this is, for $\psi(x) = \hat{\psi}(G(x))$,

$$\frac{d}{dx}[\psi'(x)q(x)] = 2\theta^{-1} q(x) u_R(x) - \frac{1}{2} q(x) \psi(x)^2$$

Integrating and applying $\phi^{*\prime}(y_L) = \phi^{*\prime}(y_H) = 0$ proves the result.

C.6 Proof of Corollary 2

We first prove that $p_R^*(x)$ is strictly increasing on (x_L, x_H).

It is convenient to work with the transformed variable

$$y = G(x) = \int_{x_L}^{x} \frac{dx}{q(x)},$$

which is well-defined by the compactness of X and the full-support property of $q(x)$. We have

$$p_R^{*\prime}(y) = p_R^{*\prime}(x(y)) \frac{dx}{dy} = q(x(y)) p_R^{*\prime}(x(y)).$$

The Euler-Lagrange equation rewritten with this change of variable is

$$g(y)u_R(y) + \frac{\theta}{4} \frac{(p_R^{*\prime}(y))^2}{(p_R^*(y)(1-p_R^*(y)))^2}(1 - 2p_R^*(y)) = -\frac{\theta}{2} \frac{d}{dy}[\frac{p_R^{*\prime}(y)}{p_R^*(y)(1-p_R^*(y))}].$$

It also also convenient to work with the transformed function

$$\phi(y) = \cos^{-1}(\sqrt{p_R^*(y)}),$$

which satisfies

$$\phi'(y) = -\frac{1}{2} \frac{p_R^{*\prime}(y)}{\sqrt{p_R^*(y)(1-p_R^*(y))}}.$$

We assume (without loss of generality) that $\phi(y) \in [0, \frac{\pi}{2}]$. The corresponding Euler-

Lagrange equation is

$$g(y)u_R(y) + \frac{\theta}{4}\frac{(\phi'(y))^2}{p_R^*(y)(1-p_R^*(y))}(1-2p_R^*(y)) = \frac{\theta}{2}\frac{d}{dy}[\frac{\phi'(y)}{\sqrt{p_R^*(y)(1-p_R^*(y))}}]$$

which further simplifies to

$$\sin(2\phi(y))g(y)u_R(y) = \frac{\theta}{2}\phi''(y).$$

By Proposition 3, this equation is satisfies everywhere $u_R(y)$ is continuous, and ϕ is continuously differentiable everywhere. The boundary conditions are $\phi'(0) = \phi'(\bar{y}) = 0$, where $\bar{y} = G(x_H)$.

Define $y^* = G(x^*)$. By the single-crossing property, $u_R(y) < 0$ for all $y < y^*$, and hence ϕ is strictly concave on $y \in (0, y^*)$. It follows by $\phi'(0) = 0$ that $\phi'(y) < 0$ for all $y \in (0, y^*)$. Similarly, $u_R(y) > 0$ for all $y > y^*$. It follows that ϕ strictly convex on $(y^*, \bar{y})$, and by $\phi'(\bar{y}) = 0$ we must have $\phi'(y) < 0$ for all $y \in (y^*, \bar{y})$. By the continuity of $\phi'(y)$, $\phi'(y) < 0$ for all $y \in (0, \bar{y})$. It follows immediately that $p_R^{*\prime}(x) > 0$ for all $x \in (x_L, x_H)$.

We next prove that for some $x_1 \in (x_L, x_H)$, $p_R^*(x)$ is strictly convex on $x \in [x_L, x_1)$. Define x_1' as the lesser of x^* and the smallest point of discontinuity for $u_R(x)$ on (x_L, x_H). On the interval (x_L, x_1'), $u_R(x)$ is continuous and strictly negative (by single crossing). The Euler-Lagrange equation from Proposition 3 can be written as

$$\frac{1}{2\theta}u_R(x) - \frac{1}{2}\frac{(p_R^{*\prime}(x))^2}{(p_R^*(x)(1-p_R^*(x)))^2}(1-2p_R^*(x)) + \frac{q'(x)}{q(x)}\frac{p_R^{*\prime}(x)}{p_R^*(x)(1-p_R^*(x))} = -\frac{p_R^{*\prime\prime}(x)}{p_R^*(x)(1-p_R^*(x))}.$$

By the continuity of $p_R^{*\prime}(x)$ and $u_R(x)$, the fact that $p_R^*(x)$ has a strictly positive minimum on X, the boundary condition $p_R^{*\prime}(x_L) = 0$, and $q \in \mathscr{P}_{LipG}(x)$ (implying $\frac{q'(x)}{q(x)}$ bounded), we must have

$$\lim_{x \to x_L} p_R^{*\prime\prime}(x) > 0,$$

and by the continuous twice-differentiability of p_R^* on (x_L, x_1'), this must hold on some interval (x_L, x_1) with $x_1 \in (x_L, x_1']$, implying that $p_R^*(x)$ is convex on $[x_L, x_1)$. The argument that for some $x_2 \in (x_L, x_H)$, $p_R^*(x)$ is strictly concave on $x \in (x_2, x_H]$ is symmetric, and $x_2 \geq x^* \geq x_1$ proves the result.

C.7 Proof of Proposition 4

Here we solve the multi-variate problem in the calculus of variations stated in Section 4.4,

$$\inf_{\{p_a(x)\}_{a\in\mathbb{A}}\in\mathscr{P}_{LipG}(A)} \int_X q(x) \int_A [p_a(x)(a-\gamma^T x)^2 + \frac{\theta}{4}\frac{|\nabla_x p_a(x)|^2}{p_a(x)}] da dx$$

where under the prior $q(x)$ $x \sim N(\mu_0, \Sigma_0)$, $X = \mathbb{R}^L$, and $A = \mathbb{R}$.

We can write this as

$$\int_X q(x) \int_A F(a, p_a(x), \nabla_x p_a(x); x)\, da\, dx,$$

where for each pair (x, a), the function

$$F(a, f, g; x) \;\equiv\; f \cdot (a - \gamma^T x)^2 + \frac{\theta}{4}\frac{|g|^2}{f}$$

is a convex function of the arguments (f, g) everywhere on its domain (the half-plane on which $f > 0$). To prove convexity, observe that

$$\begin{bmatrix} F_{gg} & F_{fg} \\ F_{gf} & F_{ff} \end{bmatrix} = \frac{\theta}{4}\begin{bmatrix} \frac{1}{f}I & -\frac{g}{f^2} \\ -\frac{g^T}{f^2} & 2\frac{g^T g}{f^3} \end{bmatrix}.$$

The upper left block is positive definite, and the determinant of the matrix is strictly positive, and consequently the matrix is strictly positive-definite.

Given the convexity of the objective, the first-order conditions are both necessary and sufficient for an optimum. The relevant first-order conditions are further-

more the same as those for minimization of the Lagrangian

$$\int_X q(x) \int_A \mathscr{L}(a, p_a(x), \nabla p_a(x); x)\, da\, dx,$$

where

$$\mathscr{L}(a, f, g; x) \;=\; F(a, f, g; x) + \varphi(x) f + \psi_a(x) f. \tag{26}$$

Here $\varphi(x)$ is the Lagrange multiplier associated with the constraint

$$\int_A p_a(x) da \;=\; 1 \tag{27}$$

for each $x \in X$, as is required in order for $p_a(x)$ to be a probability density function, and $\psi_a(x)$ is the multiplier on the constraint that $p_a(x)$ be weakly positive.

For given Lagrange multipliers, the problem of minimizing the Lagrangian can further be expressed as a separate minimization problem for each possible action a. Then if we can find a function $\varphi(x)$ and a function $p_a(x)$ for each $a \in A$, with $p_a(x) > 0$ for all x, such that (i) for each $a \in A$, the function $p_a(x)$ minimizes

$$\int_X q(x) \mathscr{L}(a, p_a(x), \nabla_x p_a(x); x)\, dx, \tag{28}$$

and (ii) condition (27) holds for all $x \in X$, then we will have derived an optimal information structure.

For the problem of choosing a function $p_a(x)$ to minimize (28), the first-order conditions are given by the Euler-Lagrange equations

$$q(x) \frac{\partial \mathscr{L}}{\partial f}(a, p_a(x), \nabla_x p_a(x); x) \;= \sum_{k=1}^{L} \frac{d}{dx^k} \left[q(x) \frac{\partial L}{\partial g^k}(a, p_a(x), \nabla_x p_a(x); x) \right],$$

or equivalently,

$$\frac{\partial \mathscr{L}}{\partial f}(a, p_a(x), \nabla_x p_a(x); x) \;=\; \nabla_g \mathscr{L}(a, p_a(x), \nabla_x p_a(x); x) \cdot \nabla_x [\log q(x)] + \nabla_x \cdot \left[\nabla_g \mathscr{L}(a, p_a(x), p_a'(x); x) \right].$$

In the case of the objective function (26), we have

$$\frac{\partial \mathscr{L}}{\partial f} \;=\; (a-\gamma^T x)^2 - \frac{\theta}{4}\,|\nabla_x v_a(x)|^2 + \varphi(x) + \psi_a(x),$$

$$\nabla_g \mathscr{L} \;=\; \frac{\theta}{2}\nabla_x v_a(x),$$

where $v_a(x) \equiv \log p_a(x)$. Under our assumption of a Gaussian prior, we also have

$$\nabla_x[\log q(x)] \;=\; \Sigma_0^{-1}(\mu_0 - x).$$

Substituting these expressions, the Euler-Lagrange equations take the form

$$(a-\gamma^T x)^2 + \varphi(x) + \psi_a(x) - \frac{\theta}{4}|\nabla_x v_a(x)|^2 = \frac{\theta}{2}(\mu_0 - x)^T \Sigma_0^{-1} \nabla_x v_a(x) + \frac{\theta}{2}\nabla_x \cdot \nabla_x v_a(x)$$

for all x and a.

In the case that $4|\Sigma_0 \gamma|^2 > \theta$, we conjecture and verify that these equations have a solution given by

$$\psi_a(x) = 0,$$

$$\nabla_x v_a(x) \;= \lambda[a - \gamma^T \mu_0 - \sigma^{-2}\lambda^T(x-\mu_0)], \tag{29}$$

for some values of $\sigma \in \mathbb{R}, \lambda \in \mathbb{R}^L$ and some $\phi(x)$. Note that this conjecture can be integrated, with

$$\exp(v_a(x)) = p_a(x) = -\frac{\sigma}{\sqrt{2\pi}}\exp(-\frac{\sigma^2}{2}(a - \gamma^T\mu - \sigma^{-2}\lambda^T(x-\mu))^2).$$

Plugging in this conjecture,

$$\begin{aligned}
\varphi(x) = &-(a-\gamma^T x)^2 + \frac{\theta}{4}\lambda^T\lambda(a - \gamma^T x + (\gamma - \sigma^{-2}\lambda)^T(x-\mu_0))^2 \\
&+ \frac{\theta}{2}(\mu_0 - x)^T\Sigma_0^{-1}\lambda(a-\gamma^T x) \\
&+ \frac{\theta}{2}(\mu_0 - x)^T\Sigma_0^{-1}\lambda(\gamma - \sigma^{-2}\lambda)^T(x-\mu_0) \\
&+ \frac{\theta}{2}\sigma^{-2}\lambda^T\lambda.
\end{aligned}$$

By variation of parameters in a, we must have (as in the proposition)

$$\lambda^T \lambda = \frac{4}{\theta}$$

and, for all x,

$$(x-\mu_0)^T \Sigma_0^{-1} \lambda = \lambda^T \lambda (x-\mu_0)^T (\gamma - \sigma^{-2}\lambda).$$

Hence we require

$$\frac{\theta}{4}\Sigma_0^{-1}\lambda = \gamma - \sigma^{-2}\lambda,$$

which implies (as stated in the text) that

$$\lambda = (\frac{\theta}{4}\Sigma_0^{-1} + \sigma^{-2} I)^{-1}\gamma, \tag{30}$$

and

$$|(\frac{\theta}{4}\Sigma_0^{-1} + \sigma^{-2} I)^{-1}\gamma|^2 = \frac{4}{\theta}, \tag{31}$$

which is feasible for $\sigma > 0$ under the assumption that $|\Sigma_0 \gamma|^2 > \frac{\theta}{4}$. Note that this formula is a rescaled version of the one stated in the proposition.

Observe that we can rewrite this equations as

$$\Sigma_0^{-1}\lambda = \frac{4}{\theta}\gamma - \sigma^{-2}\lambda\lambda^T\lambda,$$

and hence that

$$\lambda = \frac{4}{\theta}(\Sigma_0^{-1} + \sigma^{-2}\lambda\lambda^T)\gamma. \tag{32}$$

Now suppose the DM receives a Gaussian signal $s = \lambda^T x + \varepsilon$, where the "observation error" ε is normally distributed, with mean zero and a variance σ^2, and independent of the value of x. Here, σ and λ are the solutions to (30) and (31) above.

With such a signal, and given the Gaussian prior beliefs, the DM's posterior beliefs are Gaussian. The posterior precision of the DM's belief about $\lambda^T x$ is

$$(\lambda^T \Sigma_0 \lambda)^{-1} + \sigma^{-2},$$

and the posterior mean is

$$((\lambda^T\Sigma_0\lambda)^{-1}+\sigma^{-2})^{-1}((\lambda^T\Sigma_0\lambda)^{-1}\lambda^T\mu_0+\sigma^{-2}s),$$

while the posterior mean and precision about any z^Tx with $z^T\Sigma_0\lambda=0$ is unchanged. An orthogonal basis of these z vectors and λ form an orthogonal basis, and let

$$\gamma = b_0\lambda + b_1 z_1 + \ldots,$$

observing that

$$b_0 = \frac{\gamma^T\Sigma_0\lambda}{\lambda^T\Sigma_0\lambda}.$$

The posterior variance-covariance matrix is

$$\Sigma_s = \Sigma_0 + \frac{\Sigma_0\lambda\lambda^T\Sigma_0}{(\lambda^T\Sigma_0\lambda)^2}(\frac{1}{(\lambda^T\Sigma_0\lambda)^{-1}+\sigma^{-2}} - \lambda^T\Sigma_0\lambda),$$

which simplifies to

$$\begin{aligned}\Sigma_s &= \Sigma_0 + \frac{\Sigma_0\lambda\lambda^T\Sigma_0}{(\lambda^T\Sigma_0\lambda)}(\frac{1}{1+\sigma^{-2}\lambda^T\Sigma_0\lambda} - 1)\\ &= \Sigma_0 - \Sigma_0\lambda\lambda^T\Sigma_0\frac{\sigma^{-2}}{1+\sigma^{-2}\lambda^T\Sigma_0\lambda},\end{aligned}$$

and therefore by the Sherman-Morrison lemma,

$$\Sigma_s^{-1} = \Sigma_0^{-1} + \sigma^{-2}\lambda\lambda^T.$$

The posterior mean of γ^Tx (and hence optimal action $a(s)$) is

$$\begin{aligned}E[\gamma^Tx|s] = \frac{\gamma^T\Sigma_0\lambda}{\lambda^T\Sigma_0\lambda}&[((\lambda^T\Sigma_0\lambda)^{-1}+\sigma^{-2})^{-1}((\lambda^T\Sigma_0\lambda)^{-1}\lambda^T\mu_0+\sigma^{-2}s) - \lambda^T\mu_0]\\ &+\gamma^T\mu_0,\end{aligned}$$

which simplifies to (as given in the text)

$$E[\gamma^T x|s] = \gamma^T \mu_0 + \frac{\gamma^T \Sigma_0 \lambda}{\lambda^T \Sigma_0 \lambda} \frac{\sigma^{-2}}{(\lambda^T \Sigma_0 \lambda)^{-1} + \sigma^{-2}} (s - \lambda^T \mu).$$

Observe by the definitions of λ and σ that

$$1 = \lambda^T \Sigma_0 \gamma - \sigma^{-2} \lambda^T \Sigma_0 \lambda$$

and therefore (as stated in the text)

$$E[\gamma^T x|s] = \gamma^T \mu_0 + \sigma^{-2}(s - \lambda^T \mu_0).$$

Consequently, a is normally distributed conditional on x, with conditional mean

$$E[a(s)|x] = \gamma^T \mu_0 + \sigma^{-2} \lambda^T (x - \mu_0)$$

and conditional variance

$$Var[a(s)|x] = \sigma^{-2}.$$

That is,

$$p_a(x) = \frac{\sigma}{\sqrt{2\pi}} \exp(-\frac{\sigma^2}{2}(a - \gamma^T \mu_0 - \sigma^{-2} \lambda^T (x - \mu_0))^2,$$

and

$$\nabla_x \ln(p_a(x)) = \lambda (a - \gamma^T \mu_0 - \sigma^{-2} \lambda^T (x - \mu_0)),$$

which is the conjectured and verified functional form in (29).

Now consider the problem

$$z^* \in \arg \min_{z:|z|^2=1} z^T (\Sigma_0^{-1} + \sigma^{-2} \lambda \lambda^T)^{-1} \gamma.$$

The first-order condition is

$$\Sigma_s \gamma - \psi z^* = 0,$$

where ψ is the multiplier on $z^T z = 1$, and therefore by (32)

$$z^* \propto \lambda,$$

concluding the proof.

C.8 Proof of Corollary 3

In this corollary we rewrite the problem in terms of a choice of a normally distributed signal $s \in \mathbb{R}^L$ with conditional mean μ_x and positive-semidefinite variance matrix Ω. Given such a signal, the posterior is normally distributed with mean μ_s and posterior variance

$$\Sigma_s = (\Sigma_0^{-1} + \Omega^{-1})^{-1}.$$

Observe by Proposition 4 that the optimal signal structure falls into this class.

Now consider the original problem in posterior form (as in the multi-dimensional generalization of equation (11)). Because the posteriors of this problem are normally distributed, we have

$$\int_{\mathbb{R}^k} \frac{|\nabla_x q_s(x)|^2}{q_s(x)} dx = E[|\Sigma_s^{-1}(x-\mu_s)|^2 | s]$$

and therefore

$$\begin{aligned}\int_{R^k} \pi(s) \int_X \frac{|\nabla_x q_s(x)|^2}{q_s(x)} dx ds &= E[tr[\Sigma_s^{-1}(x-\mu_s)(x-\mu_s)^T \Sigma_s^{-1}]] \\ &= tr[\Sigma_s^{-1}].\end{aligned}$$

By the same argument, for the prior q,

$$\int_X \frac{|\nabla_x q(x)|^2}{q(x)} dx = tr[\Sigma_0^{-1}].$$

Given such a signal structure, the optimal action is

$$a^*(s) = \gamma^T \mu_s,$$

and therefore

$$\begin{aligned}\int_X q(x) \int_{\mathbb{R}^k} p_s(x)(a^*(s) - \gamma^T x)^2 ds dx &= E[Var[\gamma^T x|s]] \\ &= \gamma^T \Sigma_s \gamma.\end{aligned}$$

Let $\mathscr{M}_k$ be the set of $k \times k$ real symmetric positive-definite matrices. We can write the posterior-based problem as

$$\inf_{\Sigma_s \in \mathscr{M}_k} \gamma^T \Sigma_s \gamma - \frac{\theta}{4} tr[\Sigma_s^{-1}] + \frac{\theta}{4} tr[\Sigma_0^{-1}]$$

subject to the constraint

$$\Sigma_s^{-1} \succeq \Sigma_0^{-1},$$

which equivalent to $\Sigma_s \preceq \Sigma_0$. By Proposition 4, the optimal solution to this problem is

$$\Sigma_s^* = (\Sigma_0^{-1} + \sigma^{-2} \lambda \lambda^T)^{-1}.$$

C.9 Proof of Corollary 4

In the case that $\theta \geq 4|\Sigma_0 \gamma|^2$, instead, there is no solution to the Euler-Lagrange equations from the proof of Proposition 4, and we can show that there is no interior solution to the optimization problem. Instead it is optimal to choose a completely uninformative information structure, and to choose the estimate $a = \mu$ at all times. This is because in this case, one can show that any information structure and estimation rule implies that

$$V \equiv \mathrm{E}[(a - \gamma^T x)^2] + \frac{\theta}{4} \mathrm{E}[I(x)] \geq \mathrm{E}[(\gamma^T (x - \mu))^T] = \gamma^T \Sigma_0 \gamma,$$

where $I(x)$ is the Fisher information, with the lower bound achieved only in the case that $a = \mu$ with probability 1.

Consider some hypothetical policy $p_a(x)$. We begin by observing that the Cramér-Rao bound for a biased estimator[35] implies that

$$\mathrm{E}^p[(a-\gamma^T x)^2|x] \geq (\nabla_x \bar{a}(x))^T \cdot I(x;p)^{-1} \cdot \nabla_x \bar{a}(x) + (\bar{a}(x) - \gamma^T x)^2.$$

where $\bar{a}(x) \equiv \mathrm{E}^p[a|x]$ under the measure $p_a(x)$, and $I(x;p)$ is the Fisher information of x under $p_a(x)$.

Thus,

$$\begin{aligned}\mathrm{E}^p[(a-\gamma^T x)^2|x] + \frac{\theta}{4} tr[I(x)] &\geq (\nabla_x \bar{a}(x))^T \cdot I(x;p)^{-1} \cdot \nabla_x \bar{a}(x) + \frac{\theta}{4} tr[I(x;p)] + (\bar{a}(x) - \gamma^T x)^2 \\ &\geq \inf_I \{\nabla_x \bar{a}(x))^T \cdot I^{-1} \cdot \nabla_x \bar{a}(x) + \frac{\theta}{4} tr[I]\} + (\bar{a}(x) - \gamma^T x)^2\end{aligned}$$

where the minimization is taken over the set of positive-definite matrices.

In the technical appendix, we prove the following lemma:

Lemma 9. *Let Λ_0 be a $k \times k$ real symmetric positive-semidefinite matrix, let $\mathscr{M}_k$ be the set of $k \times k$ real symmetric positive-definite matrices, and let $v \in \mathbb{R}^k$ be a vector. Then*

$$2|v| = \inf_{\Lambda \in \mathscr{M}_k} v^T \Lambda^{-1} v + tr[\Lambda]$$

Proof. See the Technical Appendix, D.2.2. □

By this lemma,

$$\inf_I \{\frac{4}{\theta} \nabla_x \bar{a}(x))^T \cdot I^{-1} \cdot \nabla_x \bar{a}(x) + tr[I]\} = 4\theta^{-\frac{1}{2}} |\nabla_x \bar{a}(x)|.$$

[35] See Cover and Thomas (2006), p. 396.

Therefore,

$$\begin{aligned} \mathrm{E}^p[(a-\gamma^T x)^2|x] + \frac{\theta}{4} tr[I(x)] &\geq \theta^{1/2} \left|\nabla_x \bar{a}(x)\right| + (\bar{a}(x)-\gamma^T x)^2 \\ &\geq 2|\Sigma_0 \gamma| \left|\nabla_x \bar{a}(x)\right| + (\bar{a}(x)-\gamma^T x)^2 \\ &\geq 2\gamma^T \Sigma_0 \nabla_x \bar{a}(x) + (\bar{a}(x)-\gamma^T x)^2, \end{aligned}$$

where the next-to-last inequality follows from the assumption that $\theta \geq 4|\Sigma_0\gamma|^2$ and the last from the Cauchy-Schwarz inequality. Taking the expected value under the prior $q(x)$, it then follows that

$$V \geq \int_X q(x) \left[2\gamma^T \Sigma_0 \nabla_x \bar{a}(x) + (\bar{a}(x)-\gamma^T x)^2\right] dx. \tag{33}$$

We wish to obtain a lower bound for the integral on the right-hand side of (33). To do this, we solve for the function $\bar{a}(x)$ that minimizes this integral, using the calculus of variations. Once again, we note that the integrand is a convex function of $\bar{a}$ and $\nabla_x \bar{a}$, so that the first-order conditions are both necessary and sufficient for a minimum. The first-order conditions are given by the Euler-Lagrange equations

$$\begin{aligned} 2q(x)(\bar{a}(x)-\gamma^T x) &= 2\gamma^T \Sigma_0 \nabla_x q(x) \\ &= 2q(x)\gamma^T (x-\mu_0) \end{aligned}$$

which have a unique solution $\bar{a}(x) = \gamma^T \mu_0$ for all x.

Substituting this solution into the integral (33), we obtain the tighter lower bound

$$V \geq \int_X q(x)(\gamma^T (x-\mu_0))^2 dx = \gamma^T \Sigma_0 \gamma. \tag{34}$$

But this lower bound is achievable by choosing $a = \gamma^T \mu_0$ with probability 1, regardless of the value of x (the optimal estimate in the case of a perfectly uninformative information structure). Hence a perfectly uninformative information structure is optimal for all $\theta \geq 4|\Sigma_0\gamma|^2$.

This solution is not only *one* way of achieving the lower bound, it is the only way. It follows from the reasoning used to derive the lower bound for V that the

lower bound can be achieved only if each of the weak inequalities holds as an equality. But the bound in (34) is equal to the bound in (33) only if $\bar{a}(x) = \gamma^T \mu_0$ almost surely; thus optimality requires this. And the restriction that $\mathrm{E}[a|x] = \gamma^T \mu_0$ for a set of x with full measure implies that we must have

$$\mathrm{E}[(a - \gamma^T x)^2|x] = (\gamma^T (x - \mu_0))^2 + Var[a|x].$$

This in turn implies that

$$\mathrm{E}[(a - \gamma^T x)^2] \; = \; \mathrm{E}[(\gamma^T (x - \mu_0))^2] + \mathrm{E}[Var[a|x]] \; = \gamma^T \Sigma_0 \gamma + \mathrm{E}[Var[a|x]].$$

Hence the lower bound can be achieved only if $\mathrm{E}[Var[a|x]] = 0$.

Given that the variance is necessarily non-negative, this requires that $Var[a|x] = 0$ almost surely. This together with the requirement that $\mathrm{E}[a|x] = \gamma^T \mu_0$ almost surely implies that $a = \gamma^T \mu_0$ almost surely. Hence optimality requires that $a = \gamma^T \mu_0$ with probability 1, whenever $\theta \geq 4|\Sigma\gamma|^2$.

C.10 Proof of Lemma 3

Applying Lemma 1, for the $\rho = 1$ case and any $p^1 \in \mathscr{P}(S)^{|X|}$,

$$C_{NG}(p^1, q_0; S, X; \rho = 1) = \sum_{i \in \mathscr{I}} c_i \sum_{x \in X_i} q_x \sum_{s \in S: \pi_{i,s} > 0} p^1_{x,s} \ln(\frac{p^1_{x,s}}{\pi^1_{i,s}})$$

where

$$\pi^1_i = \sum_{x \in X_i} p^1_x q_{i,x}(q).$$

Therefore, defining $p^{12} = p^1 \otimes p^2$ as in Definition 3,

$$C_{NG}(p^{12}, q_0; S, X; \rho = 1) = \sum_{i \in \mathscr{I}} c_i \sum_{x \in X_i} q_x \sum_{(s_1, s_2) \in S \times S: \pi^{12}_{i,s_1,s_2} > 0} p^1_{x,s_1} p^2_{x,s_2} \ln(\frac{p^1_{x,s_1} p^2_{x,s_2}}{\pi^{12}_{i,s_1,s_2}}),$$

where

$$\pi^1_{i,s_1,s_2} = \sum_{x \in X_i} p^1_{x,s_1} p^2_{x,s_2} q_{i,x}(q).$$

It follows that

$$C_{NG}(p^{12},q_0;S,X;\rho=1)-C_{NG}(p^1,q_0;S,X;\rho=1)-C_{NG}(p^2,q_0;S,X;\rho=1)=$$
$$\sum_{i\in\mathscr{I}}c_i\sum_{x\in X_i}q_x\sum_{(s_1,s_2)\in S\times S:\pi^{12}_{i,s_1,s_2}>0}p^1_{x,s_1}p^2_{x,s_2}\ln(\frac{\pi^1_{i,s_1}\pi^1_{i,s_2}}{\pi^{12}_{i,s_1,s_2}})=$$
$$-\sum_{i\in\mathscr{I}}c_i\bar{q}_i(q)\sum_{(s_1,s_2)\in S\times S:\pi^{12}_{i,s_1,s_2}>0}\pi^{12}_{i,s_1,s_2}\ln(\frac{\pi^{12}_{i,s_1,s_2}}{\pi^1_{i,s_1}\pi^1_{i,s_2}}).$$

This quantity is the negative of the conditional (on being in $i\in\mathscr{I}$) mutual information between s_1 and s_2, and hence is negative, strictly so if the signals are not independent. Therefore, the $\rho=1$ case exhibits decreasing marginal costs.

Next consider the $\rho=2$ case:

$$C_{NG}(p^1,q_0;S,X;\rho=2)=\sum_{i\in\mathscr{I}}c_i|X_i|^{-1}\bar{q}_i(q)\sum_{x\in X_i}\sum_{s\in S:\pi_{i,s}>0}\pi^1_{i,s}\ln(\frac{\pi^1_{i,s}}{p^1_{x,s}}),$$

and therefore

$$C_{NG}(p^{12},q_0;S,X;\rho=2)=\sum_{i\in\mathscr{I}}c_i|X_i|^{-1}\bar{q}_i(q)\sum_{x\in X_i}\sum_{(s_1,s_2)\in S\times S:\pi^{12}_{i,s_1,s_2}>0}\pi^{12}_{i,s_1,s_2}\ln(\frac{\pi^{12}_{i,s_1,s_2}}{p^1_{x,s_1}p^2_{x,s_2}}).$$

It follows that

$$C_{NG}(p^{12},q_0;S,X;\rho=2)-C_{NG}(p^1,q_0;S,X;\rho=2)-C_{NG}(p^2,q_0;S,X;\rho=2)=$$
$$\sum_{i\in\mathscr{I}}c_i\bar{q}_i(q)\sum_{(s_1,s_2)\in S\times S:\pi^{12}_{i,s_1,s_2}>0}\pi^{12}_{i,s_1,s_2}\ln(\pi^{12}_{i,s_1,s_2})-$$
$$\sum_{i\in\mathscr{I}}c_i\bar{q}_i(q)\sum_{(s_1,s_2)\in S\times S:\pi^{12}_{i,s_1,s_2}>0}\pi^1_{i,s_1}\pi^1_{i,s_2}\ln(\pi^1_{i,s_1}\pi^1_{i,s_2}).$$

This quantity is also the conditional (on being in $i\in\mathscr{I}$) mutual information between s_1 and s_2, and hence is positive, strictly so if the signals are not independent. Therefore, the $\rho=2$ case exhibits increasing marginal costs.

C.11 Proof of Proposition 5

By Definition 1,

$$H^G(q_i;1,|X_i|)+H^G(q_i;2,|X_i|)=(\sum_{j=1}^{|X_i|}q_{i,j}\ln(q_{i,j}))-\frac{1}{|X_i|}\sum_{j=1}^{|X_i|}\ln(q_{i,j})$$

A little algebra shows that

$$\begin{aligned}H^G(q_i;1,|X_i|)+H^G(q_i;2,|X_i|)&=\frac{1}{|X_i|}\sum_{j=1}^{|X_i|}\sum_{k=1}^{|X_i|}q_{i,j}\ln(\frac{q_{i,j}}{q_{i,k}})\\&=H^{CM}(q_i(\bar{q});|X_i|).\end{aligned}$$

By the results of Bloedel and Zhong (2020), if the neighborhood-based cost function $H(q;X,\mathscr{N})$ has constant marginal costs, it must satisfy the functional form in (19).

By the arguments used to prove Proposition 1 (invoking Assumption 3), it is without loss of generality to suppose that a set $\bar{X}$, measure $\bar{q}\in\mathbb{R}_+^{|\bar{X}|}$, and surjection $m:\bar{X}\to X$ exists such that the $\bar{X}_i=\bar{X}\cap N_i$ are disjoint and

$$H(q;X,\mathscr{N})=H(\bar{q};\bar{X},\mathscr{N}),$$

where $\bar{q}_{\bar{x}}=q_{m(x)}$ for all $x\in\bar{X}$. By Assumption 2, we must have $\gamma_{x,x'}=0$ for any x,x' that do not share a neighborhood. By Proposition 1, within each disjoint neighborhood i we must (due to the symmetry of the local information cost) have $\gamma_{x,x'}=\gamma_{x',x}=\gamma_{x,x''}$ for all $x,x',x''\in\bar{X}_i$. Consequently, defining $c_i=|X_i|\gamma_{x,x'}$ for some (any) $x,x'\in\bar{X}_i$,

$$\begin{aligned}H(\bar{q};\bar{X},\mathscr{N})&=\sum_{i\in\mathscr{I}}\frac{c_i}{|X_i|}\sum_{x\in\bar{X}_i}\sum_{x'\in\bar{X}_i\setminus\{x\}}\bar{q}_x\ln(\frac{\bar{q}_x}{\bar{q}_{x'}})\\&=\sum_{i\in\mathscr{I}}\frac{c_i}{|X_i|}\bar{q}_i(\bar{q})\sum_{x\in\bar{X}_i}\sum_{x'\in\bar{X}_i\setminus\{x\}}q_{i,x}(\bar{q})\ln(\frac{q_{i,x}(\bar{q})}{q_{i,x'}(\bar{q})})\\&=\sum_{i\in\mathscr{I}}c_i\bar{q}_i(\bar{q})H^{CM}(q_i(\bar{q});|X_i|).\end{aligned}$$

It follows immediately that for all X, regardless of whether the X_i are disjoint,

$$H(q;X,\mathscr{N}) = \sum_{i\in\mathscr{I}} c_i \bar{q}_i(q) H^{CM}(q_i(q);|X_i|).$$

The representation for V_{CM} with the KL divergence follows immediately (or see Bloedel and Zhong (2020)).

C.12 Proof of Proposition 6

It is convenient to work with the transformed variable

$$y = G(x) = \int_{x_L}^{x} \frac{dx}{q(x)},$$

which is well-defined by the compactness of X and the full-support property of $q(x)$. Define $\bar{y} = G(\bar{x})$ and $g(y) = q(G^{-1}(y))^2$.

The associated problem is

$$\max_{s\in S, K\in\mathbb{R}} \int_0^{\bar{y}} g(y)(K - \beta s(y))dy$$

subject to

$$\inf_{p_L\in\{p\in C^1([0,\bar{y}],(0,\infty)):\int_X g(y)p(y)dy=1\}} \int_0^{\bar{y}} g(y)p_L(y)(s(y)-K)dy + \frac{\theta}{4}\int_0^{\bar{y}} \frac{(p_L'(y))^2}{p_L(y)}dy \geq 0.$$

Note that fixing any K, maximizer on S exists (provided the problem is feasible), by the Lipschitz property of S (which ensures S is compact).

It is without loss of generality to restrict K to the set $[0,\bar{x}]$, because $K<0$ will always be dominated by $K=0$ and $K>\bar{x}$ will never satisfy the constraint (due to the limited liability of $s\in S$). Therefore, a maximizing (s,K) exists.

We proceed by taking $K\geq 0$ as given, and determining the optimal security s, and then consider the optimal choice of K. If $K=0$, the optimal security is $s(y)=0$ for all $y\in[0,\bar{y}]$.

Suppose $K>0$. As argued in the text, it is without loss of generality to assume

the constraint binds, and hence that the results of Lemma 2 apply. This can only occur if $s(\bar{y}) > K$, as otherwise rejection must be a strictly dominating action.

Defining $q(y) = \bar{y}^{-1}$ and $u_R(y) = g(y)(s(y) - K)$, there exists a function $\psi : [0, \bar{y}] \to \mathbb{R}$ satisfying $\psi(0) = \psi(\bar{y}) = 0$ and

$$\frac{1}{\bar{y}}\psi(y) = \int_0^{\bar{y}} \frac{1}{\bar{y}}[\frac{2}{\theta}g(y')(s(y') - K) - \frac{1}{2}\psi(y')^2]dy.$$

We begin by proving that $\psi(y) < 0$. Define the function

$$\phi(y) = \exp(\frac{1}{2}\int_0^y \psi(y')dy),$$

and observe that

$$\frac{2\phi'(y)}{\phi(y)} = \psi(y).$$

Plugging $\phi(y)$ into the Euler-Lagrange equation for ψ,

$$\frac{d}{dy}[\frac{2\phi'(y)}{\phi(y)}] = 2\theta^{-1}(s(y) - K)g(y) - \frac{1}{2}(\frac{2\phi'(y)}{\phi(y)})^2,$$

which simplifies to

$$\phi''(y) = 2\theta^{-1}(s(y) - K)g(y)\phi(y).$$

By definition, $\phi'(0) = \phi'(\bar{y}) = 0$. By the single-crossing property of $g(y)(s(y) - K)\phi(y)$, $\phi(y)$ must be strictly concave wherever $s(y) < K$ and convex wherever $s(y) > K$. A single crossing must exist, by the continuity of s.

Consequently, if $K > 0$, we must have $\phi'(y) < 0$ and hence that $\psi(y) < 0$. In the case of $K = 0$, $s(y) = 0$, Lemma 2 holds with $\psi(y) = 0$ for all $y \in [0, \bar{y}]$.

The Hamiltonian can be written as

$$H(s, \psi, v, \lambda_1, \lambda_2, y; K) = g(y)(K - \beta s) + \lambda_1 v + \lambda_2(\frac{2}{\theta}(s - K)g(y) - \frac{1}{2}\psi(y)^2).$$

The constraints on v are $v \geq 0$ and $v \leq \sqrt{g(y)}$, ensuring that $s'(y)y'(x) \in [0, 1]$.

The associated necessary conditions are

$$\lambda_1(y) + \rho_0(y) - \rho_1(y) = 0,$$

where $\rho_0(y)$ and $\rho_1(y)$ are the multipliers on the constraints $v \geq 0$ and $v \leq \sqrt{g(y)}$, respectively, and

$$\begin{aligned} -\lambda_1'(y) &= g(y)(\frac{2}{\theta}\lambda_2(y) - \beta), \\ -\lambda_2'(y) &= -\lambda_2(y)\psi(y). \end{aligned} \tag{35}$$

The associated boundary conditions are $\psi(0) = \psi(\bar{y}) = 0$, $s(0) = 0$, and $\lambda_1(\bar{y}) = 0$.

If $\lambda_2(0) \leq 0$, by (35) we will have $\lambda_2(y) \leq 0$ for all $y \in [0, \bar{y}]$, implying $\lambda_1'(y) > 0$ for all $y \in [0, \bar{y}]$. By the boundary condition $\lambda_1(\bar{y}) = 0$, this requires $\lambda_1(y) < 0$ for all $y \in [0, \bar{y})$, and hence $s'(y) = 0$ for all such y. It follows in this case that $s(y) = 0$ for all $y \in [0, \bar{y}]$. This can occur only if $K = 0$.

If $K > 0$, we must have $\lambda_2(0) > 0$, and hence that $\lambda_2(y) > 0$ for all y. In this case, $\lambda_2(y)$ must be strictly decreasing (by $\psi(y) < 0$), and hence crosses β at most once. It follows that for some $\hat{y} \in [0, \bar{y}]$, $\lambda_1'(y) < 0$ on some interval $[0, \hat{y})$, if such an interval exists, and $\lambda_1'(y) > 0$ on the interval $[\hat{y}, \bar{y}]$, if such an interval exists. By the boundary condition $\lambda_1(\bar{y}) = 0$, three outcomes are possible: (1) $\lambda_1(y) < 0$ on $y \in [0, \bar{y}]$, or (2) $\lambda_1(y) > 0$ on $[0, \bar{y}]$, or (3) $\lambda_1(y) > 0$ for all $y \in [0, y^*)$ and $\lambda_1(y) < 0$ for all $y \in [y^*, \bar{y}]$. The first of these, however, is ruled out by $s(\bar{y}) > K > 0$, as above.

The optimal securities in cases (2) and (3) are described by $s(y) = \int_0^{\min\{y, y^*\}} g(y)^{\frac{1}{2}} dy$, for some $y^* \in (0, \bar{y}]$. This integral can be written as

$$\begin{aligned} s(y) &= \int_0^{\min\{G^{-1}(y), G^{-1}(y^*)\}} dx \\ &= \min\{G^{-1}(y), G^{-1}(y^*)\}, \end{aligned}$$

and therefore

$$s(x) = \min\{x, x^*\}$$

for $x^* = G^{-1}(y^*) \in (0, \bar{x}]$. By $s(\bar{x}) > K$, we must have $x^* > K > 0$.

Hence, the optimal security design is either a debt with a strictly positive price, or the zero contract with zero price.

Suppose $K = 0$ was optimal. We would require $\lambda_1^*(y) \leq 0$ for all $y \in [0, \bar{y}]$, and therefore

$$\int_0^{\bar{y}} g(y)(\frac{2}{\theta}\lambda_2(y) - \beta)dy \leq 0.$$

The first order condition for K requires that

$$\frac{\partial}{\partial K}\int_0^{\bar{y}} H(s^*(y), \psi^*(y), v^*(y), \lambda_1^*(y), \lambda_2^*(y), y; K)dy < 0,$$

and therefore

$$\int_0^{\bar{y}} g(y)(1 - \lambda_2^*(y)\frac{2}{\theta})dy = 0.$$

By $\beta < 1$, this is a contradiction, and therefore $K > 0$ and $s(\bar{x}) > K$.

D Technical Appendix

D.1 Convergence to the Continuous State Model

For each of a sequence of values for the integer M, we assume a neighborhood structure of the kind discussed in section 3.3 with $M+1$ states. The set of states is ordered, $X^M = \{0, 1, \ldots, M\}$, and each pair of adjacent states forms a neighborhood, $X_i = \{i, i+1\}$, for all $i \in \{0, 1, \ldots, M-1\}$. We will also assume that there is an $M+1$st neighborhood containing all of the states. Note that M indexes both the number of states and the number of neighborhoods. We consider the limit as $M \to \infty$.

To study this limit, we need to define how the prior beliefs, q_M, and the magnitude of the information costs vary with M. For the initial beliefs, we shall assume that there is a differentiable probability density function $q : [0,1] \to \mathbb{R}^+$, with full support on the unit interval and with a derivative that is Lipschitz continuous.

For this section and its proofs, we will use the notation e_i to indicate a basis vector equal to one for the i-th element of X^M and zero otherwise. Using the q function, we define, for any $i \in X^M$, the prior $q_M \in \mathscr{P}(X^M)$ by

$$e_i^T q_M = \int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} q(x)dx.$$

That is, for each value of M, the prior q_M is assumed to be a discrete approximation to the p.d.f. $q(x)$, which becomes increasingly accurate as $M \to \infty$.

For our neighborhood structures, we assume that that the constants associated with the cost of each neighborhood, c_j, are equal to M^2 for all $j < M$, and M^{-1} for $j = M$. In this particular example, the scaling ensures that the DM is neither able to determine the state with certainty, nor prevented from gathering any useful information, even as M is made arbitrarily large; moreover, the scaling ensures that the neighborhood containing all states plays no role in the limiting behavior, so that in the limit all information costs are local. We also scale the entire cost function by a constant, $\theta > 0$.

We also need to define the set of actions, and the utility from those actions. We

will assume the set of actions, A, remains fixed as N grows, and define the utility from a particular action, in a particular state, as

$$e_i^T u_{a,M} = \frac{\int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} q(x)u_a(x)dx}{e_i^T q_M}.$$

Here, the utility $u_a : [0,1] \to \mathbb{R}$ is a bounded measurable function for each action $a \in A$.[36] In other words, as M grows large, the prior converges to $q(x)$ and the utilities converge to the functions $u_a(x)$.

We consider only the case of generalized entropy index neighborhood cost functions with $\rho = 1$ (see Definition 2). Under these assumptions, the static model of section §2 can be written as

$$\begin{aligned} V_N(q_M;M) = \max_{\pi_M \in \mathscr{P}(A), \{q_{a,M} \in \mathscr{P}(X^M)\}_{a\in A}} & \sum_{a\in A} \pi_M(a)(u_{a,M}^T \cdot q_{a,M}) \\ & - \theta \sum_{a\in A} \pi_M(a) D_{NG}(q_{a,M}||q_M;\rho = 1, X^M, \mathscr{N}^M), \end{aligned} \tag{36}$$

subject to the constraint that

$$\sum_{a\in A} \pi_N(a) q_{a,M} = q_M.$$

The following theorem shows that the solution to this problem, both in terms of the value function and the optimal policies, converges to the solution of a static rational inattention problem with a continuous state space.

Proposition 7. *Consider the sequence of finite-state-space static rational inattention problems (36), with progressively larger state spaces indexed by the natural numbers M. There exists a sub-sequence of integers $n \in \mathbb{N}$ for which the solutions to the sub-sequence of problems converge, in the sense that, for some $\pi^* \in \mathscr{P}(A)$ and $\{q_a^* \in \mathscr{P}([0,1])\}_{a\in A}$,*

[36] Note that we do not require the payoff resulting from an action to be a continuous function of x at all points, though it will be continuous almost everywhere. This allows for the possibility that a DM's payoffs change discontinuously when the state x crosses some threshold, as in some of our applications.

i) $\lim_{n\to\infty} V_N(q_n;n) = V_N(q)$*;*

ii) $\lim_{n\to\infty} \pi_n^* = \pi^*$*; and*

iii) for all $a \in A$ *and all* $x \in [0,1]$, $\lim_{n\to\infty} \sum_{i=0}^{\lfloor xn \rfloor} e_i^T q_{a,n}^* = \int_0^x q_a^*(y)dy$.

Moreover, the limiting value function $V_N(q)$ *is the value function for the following continuous-state-space static rational inattention problem:*

$$V_N(q) = \sup_{\pi\in\mathscr{P}(A),\{q_a\in\mathscr{P}_{LipG}([0,1])\}_{a\in A}} \sum_{a\in A} \pi(a) \int_{supp(q)} u_a(x) q_a(x) dx$$
$$- \frac{\theta}{4} \sum_{a\in A} \{\pi(a) \int_0^1 \frac{(q_a'(x))^2}{q_a(x)} dx\} + \frac{\theta}{4} \int_0^1 \frac{(q'(x))^2}{q(x)} dx,$$

subject to the constraint that, for all $x \in [0,1]$,

$$\sum_{a\in A} \pi(a) q_a(x) = q(x), \tag{37}$$

and where $\mathscr{P}_{LipG}([0,1])$ *denotes the set of differentiable probability density functions with full support on* $[0,1]$*, whose derivatives are Lipschitz-continuous. Furthermore, the limiting action probabilities* $\pi^*(a)$ *and posteriors* q_a^* *are the optimal policies for this continuous-state-space problem.*

Proof. See the technical appendix, section D.4.1. □

This theorem demonstrates that the value function, choice probabilities, and posterior beliefs of the discrete state problem converge to the value function, choice probabilities, and posterior beliefs associated with a continuous state problem. The continuous state problem uses a particular cost function, the expected value of the Fisher information $I^{Fisher}(x;p)$, defined locally for each element of the continuum of possible states x, with the expectation taken with respect to the prior over possible states. The posterior beliefs in the continuous state problem, $q_a(x)$, are required to be differentiable, with a Lipschitz-continuous derivative, on their support. This is a result; the limiting posterior beliefs of the discrete state problem will have these properties. This restriction also ensures that the Fisher information is finite, so that the optimization associated with the continuous state problem is well-behaved.

The static rational inattention problem for the limiting case of a continuous state space can be given an alternative, equivalent formulation, in which the objects of choice are the conditional probabilities of taking different actions in the different possible states, rather than the posteriors associated with different actions. This is essentially the continuous state analog of Lemma 1.

Lemma 10. *Consider the alternative continuous-state-space static rational inattention problem:*

$$\bar{V}_N(q) \;=\; \sup_{p\in\mathscr{P}_{LipG}(A)} \int_0^1 q(x)\sum_{a\in A} p_a(x)u_a(x)dx - \frac{\theta}{4}\int_0^1 q(x)I^{Fisher}(x;p)dx,$$

where $\mathscr{P}_{LipG}(A)$ is the set of mappings $\{p_a : [0,1]\to[0,1]\}_{a\in A}$ such that for each action a, the function $p_a(x)$[37] is either everywhere zero or a strictly positive differentiable function of x with a Lipschitz-continuous derivative, and for any information structure $p\in\mathscr{P}_{LipG}(A)$, the Fisher information *at state $x\in X$ is defined as*

$$I^{Fisher}(x;p) \;\equiv\; \sum_{a\in A: p_a(x)>0} \frac{(p_a'(x))^2}{p_a(x)}.$$

This problem is equivalent to the one defined in Theorem 7, in the sense that the information structure p^* that is the limiting optimal policy of this problem defines action probabilities and posteriors

$$\pi^*(a) \;=\; \int_0^1 q(x)p_a^*(x), \qquad q_a^*(x) \;=\; \frac{q(x)p_a^*(x)}{\pi^*(a)} \tag{38}$$

that solve the problem in Theorem 7, and conversely, the action probabilities and posteriors $\{\pi^*(a), q_a^*\}$ that solve the problem stated in the theorem define state-contingent action probabilities

$$p_a^*(x) \;=\; \frac{\pi^*(a)q_a^*(x)}{q(x)} \tag{39}$$

that are the limiting optimal policies in the problem stated here. Moreover, the

[37] Here for any $x\in[0,1]$, we use the notation $p_a(x)$ to indicate the probability of action a implied by the probability distribution $p(x)\in\mathscr{P}(A)$.

maximum achievable value is the same for both problems: $\bar{V}_N(q) = V_N(q)$.

Proof. See the appendix, section D.4.2. □

D.2 Additional Technical Lemmas for Proposition 1 and Corollary 4

D.2.1 Proof of Lemma 4

We first state the lemma.

Lemma. *Let* $C(p, q_0; S, X)$ *be any cost function satisfying Assumption 1 (i.e. any continuously twice-differentiable UPS cost function). Suppose that, for all* $x \in X$,

$$p_x = r + \varepsilon v_x$$

for some $\varepsilon > 0$, $r \in \mathscr{P}(S)$ *with full support on S, and* $v_x \in \mathbb{R}^{|S|}$, *and that* q_0 *has full support on* X. *Then for the matrix-valued function*

$$k(q) = Diag(q) \cdot H_{qq}(q; X, \mathscr{N}) \cdot Diag(q),$$

where $Diag(q)$ *is the diagonal matrix with* q *on its diagonal and* $H_{qq}(q; X, \mathscr{N})$ *is the Hessian of the H function associated with C,*

$$C(p, q_0; S) = \frac{1}{2}\varepsilon^2 \sum_{x \in X, x' \in X} k_{x,x'}(q_0) v_x^T \cdot Diag(r)^{-1} \cdot v_{x'} + o(\varepsilon^2),$$

where $Diag(r)$ *is a diagonal matrix with* r *on the diagonal and* ι *is a vector of ones.*

Under the stated assumptions,

$$p_{s,x} = r_s + \varepsilon v_{s,x} + o(\varepsilon).$$

By Bayes' rule, for any $s \in S$ such that $\pi_s(p, q) > 0$, and any $x \in X$,

$$q_{s,x}(p, q) = \frac{p_{s,x} q_x}{\pi_s(p, q)},$$

where

$$\pi_s(p,q) = r_s + \varepsilon \sum_{x' \in X} v_{s,x'} q_{x'}.$$

It follows immediately that

$$\lim_{\varepsilon \to 0^+} q_{s,x}(p,q) = \frac{q_x r_s}{r_s} = q_x.$$

We also have

$$\varepsilon^{-1}(q_{s,x}(p,q) - q_x) = \frac{q_x(v_{s,x} - \sum_{x' \in X} v_{s,x'} q_{x'})}{\pi_s(p,q)},$$

and therefore for any s,

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}(q_{s,x}(p,q) - q_x) = \frac{q_x(v_{s,x} - \sum_{x' \in X} v_{s,x'} q_{x'})}{r_s}$$

and hence

$$q_{s,x}(p,q) - q_x = q_x \frac{v_{s,x} - \sum_{x' \in X} v_{s,x'} q_{x'}}{r_s} + o(\varepsilon).$$

In matrix form, where $v_s \in \mathbb{R}^{|X|}$ is the vector of $\{v_{s,x}\}_{x \in X}$,

$$(q_s(p,q) - q) = \frac{1}{r_s} Diag(q) \cdot (v_s - \iota q^T v_s) + o(\varepsilon).$$

By Assumption 1,

$$C(p,q_0;S) = \sum_{s \in S} \pi_s(p,q_0) D_H(q_s(p,q_0) || q_0).$$

Taylor-expanding up to second-order,

$$C(p,q_0;S) = \frac{1}{2} \sum_{s \in S} \frac{r_s}{r_s^2} (v_s - \iota q^T v_s)^T \cdot Diag(q_0) \cdot H_{qq}(q_0) \cdot Diag(q_0) \cdot (v_s - \iota q^T v_s) + o(\varepsilon^2).$$

Recalling that H is homogenous of degree one, we must have

$$\iota^T Diag(q_0) H_{qq}(q_0) = q_0^T H_{qq}(q_0) = \vec{0},$$

and consequently this expression is

$$C(p, q_0; S) = \frac{1}{2} \sum_{s \in S} \sum_{x \in X, x' \in X} k_{x.x'}(q_0) \frac{1}{r_s} v_{s,x} v_{s,x'} + o(\varepsilon^2),$$

which is the result.

D.2.2 Proof of Lemma 9

We first state the lemma.

Lemma. *Let Λ_0 be a $k \times k$ real symmetric positive-semidefinite matrix, let $\mathscr{M}_k$ be the set of $k \times k$ real symmetric positive-definite matrices, and let $v \in \mathbb{R}^k$ be a vector. Then*

$$2|v| = \inf_{\Lambda \in \mathscr{M}_k} v^T \Lambda^{-1} v + tr[\Lambda]$$

Proof. Let $\frac{v}{|v|} = z_1, z_2, \ldots, z_k$ be an orthonormal basis, and let V be the associated orthonormal matrix ($V^T V = I$) whose columns are the basis vectors. Suppose there is a minimizer, Λ^*, with

$$\Lambda^* = VMV^T$$

for some positive-definite, real symmetric M.

Consider a perturbation

$$\Lambda(\varepsilon) = \Lambda^* + \varepsilon VMzz^T MV^T$$

for some arbitrary vector z. Such a perturbation is always feasible for $\varepsilon > 0$, and is feasible for $\varepsilon < 0$ if

$$z^T MV^T \Lambda^* VMz > 0.$$

We have

$$\frac{d}{d\varepsilon}(\Lambda(\varepsilon))^{-1}|_{\varepsilon=0} = -(\Lambda^*)^{-1}VMzz^TMV^T(\Lambda^*)^{-1}.$$

Observing that

$$(\Lambda^*)^{-1} = VM^{-1}V^T$$

and using the orthonormality of V,

$$\frac{d}{d\varepsilon}(\Lambda(\varepsilon))^{-1}|_{\varepsilon=0} = -Vzz^TV^T.$$

It follows that optimality requires

$$-v^TVzz^TV^Tv + tr[VMzz^TMV^T] \geq 0,$$

with equality if the perturbation is feasible in both directions.

Because v is a basis vector of the orthonormal basis that defines V,

$$v^TV = \frac{v^Tv}{|v|}e_1^T,$$

where e_1 is a basis vector with one in index 1 and zero otherwise. Again using orthonormality to insert $V^TV = I$, we must have

$$-|v|^2e_1^Tzz^Te_1 + tr[VMV^TVzz^TV^TVMV^T] \geq 0,$$

which simplifies to

$$|v|^2e_1^Tzz^Te_1 \leq tr[\Lambda^*Vzz^TV^T\Lambda^*],$$

which is

$$z^T(V^T\Lambda^*\Lambda^*V - |v|^2e_1e_1^T)z \geq 0.$$

It follows that for all z with $e_1^Tz = 0$, we must have

$$z^TV^T\Lambda^*\Lambda^*\Lambda^*Vz = 0,$$

which requires

$$z_j^T \Lambda^* \Lambda^* \Lambda^* z_j = 0$$

for all $j \neq 1$. It follows immediately that the nullity of Λ^* is at least $k-1$, and hence that the rank is at most one. Conjecture therefore that

$$\Lambda^* = xx^T$$

for some vector x. The objective is

$$\lim_{\varepsilon \to 0^+} v^T (\varepsilon I + xx^T) v + x^T x,$$

which by the Sherman-Morrison lemma is

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1} v^T v - \frac{\varepsilon^{-2} v^T x x^T v}{1 + \varepsilon^{-1} x^T x} + x^T x.$$

By Cauchy-Schwarz,

$$\varepsilon^{-1} v^T v - \frac{\varepsilon^{-2} v^T x x^T v}{1 + \varepsilon^{-1} x^T x} \geq \frac{\varepsilon^{-1} v^T v}{1 + \varepsilon^{-1} x^T x},$$

and therefore holding fixed $|x|$ is optimal to set

$$\frac{x}{|x|} = v,$$

and the problem solves

$$\inf_{|x|^2 \geq 0} \frac{|v|^2}{|x|^2} + |x|^2,$$

and hence

$$|x|^2 = |v|.$$

It follows that

$$\inf_{\Lambda \in \mathscr{M}_k} v^T \Lambda^{-1} v + tr[\Lambda] = 2|v|.$$

□

D.3 Additional Technical Lemmas for Binary Choice

D.3.1 Proof of Lemma 6

We first repeat the lemma to be proven (from the proof of Proposition 3).

Lemma. *If*

$$\sup_{p_R \in C^1([0,\bar{y}],(0,1))} J[p_R] > \max\{\int_0^{\bar{y}} g(y)u_R(y)dy, 0\},$$

then there exists an extremal $p_R^* \in C^1([0,\bar{y}],(0,1))$ that is a maximizer and is continuously twice-differentiable except at the discontinuities of $u_R(y)$.

We begin by proving the existence of a maximizer of $J[p_R]$ on $C^1([0,\bar{y}],(0,1))$. Observe first by the concavity of J (Lemma 5) that any extremal must a maximizer.

Let us define a transformed domain for the problem, $\phi \in C^1([0,\bar{y}],\mathbb{R})$, by

$$\phi(y) = \cos^{-1}(\sqrt{p_R(y)}),$$

which satisfies

$$\phi'(y) = -\frac{p_R'(y)}{2\sqrt{p_R(y)(1-p_R(y))}}.$$

The corresponding functional is

$$\hat{J}[\phi] = \int_0^{\bar{y}} g(y)\cos(\phi(y))^2 u_R(y)dy - \theta\int_0^{\bar{y}} \phi'(y)^2 dy. \tag{40}$$

Consider the relaxed problem, for some $y_H > \bar{y} > 0 > y_L$,

$$\inf_{\phi \in C^1([0,\bar{y}+\varepsilon],\mathbb{R})} \int_{y_L}^{y_H} F(y,\phi(y),\phi'(y))dy$$

where

$$F(y,\phi,v) = \begin{cases} \theta v^2 - g(y)\cos(\phi)^2 u_R(y) & y \in [0,\bar{y}], \\ \theta v^2 & y \notin [0,\bar{y}]. \end{cases}$$

Note that this problem does not restrict the range of ϕ, but it is without loss of generality to assume that $\phi(y) \in [0, \frac{1}{2}\pi]$ for all $y \in [y_L, y_H]$. The problem is relaxed

by the possibility that $\phi(y) = 0$ or $\phi(y) = \frac{1}{2}\pi$ (which corresponds to $p(y) = 0$ or $p(y) = 1$) and extended to the domain $[y_L, y_H]$.

Because it is without loss of generality to assume bounded $\phi(y)$, and always optimal to satisfy

$$\int_{y_L}^{y_H} \phi'(y)^2 dy < \infty,$$

it is without loss of generality to assume $\phi \in W^{1,2}([y_L, y_H], \mathbb{R})$ (the Sobolev space with square-integrable weak first derivatives).

Observing that $F(y, \phi, v)$ is convex in v and satisfies, for $B = \max_{y \in [0, \bar{y}]} |g(y)u_R(y)|$,

$$F(y, \phi, v) \geq \theta v^2 - B.$$

By theorem 4.1 of Dacorogna (2007), for any given values ϕ_L and ϕ_H, the problem

$$\inf_{\phi \in \{W^{1,2}([y_L, y_H], \mathbb{R}) : \phi(y_L) = \phi_L, \phi(y_H) = \phi_H\}} \int_{y_L}^{y_H} F(y, \phi(y), \phi'(y)) dy$$

has a minimizer (where ϕ' is understood as a weak derivative). Minimizing over the compact set $(\phi_L, \phi_H) \in [0, \frac{1}{2}\pi]^2$ demonstrates that a minimizer exists for $W^{1,2}([y_L, y_H], \mathbb{R})$.

We next invoke the following lemma to show that the minimizer ϕ^* is in fact continuously differentiable, and continuously twice-differentiable everywhere $u_R(y)$ is continuous.

Lemma 11. *If $\phi^* \in W^{1,2}([-\varepsilon, \bar{y} + \varepsilon], \mathbb{R})$ is a minimizer of the functional $\hat{J}$ defined above, then $\phi^* \in C^1([-\varepsilon, \bar{y} + \varepsilon], \mathbb{R})$, and ϕ^* is continuously twice-differentiable on any interval on which u_R is continuous.*

Proof. See the Technical Appendix, Section D.3.3, defining (in the context of that proof) $u(y) = g(y)u_R(y)$. □

Let $y_1, \ldots y_{k-1}$ be the (possibly empty) set of points of discontinuity for u_R, and let $y_0 = y_L$ and $y_k = y_H$. This regularity result implies that the Euler-Lagrange equation,

$$\phi^{*\prime\prime}(y) = g(y) \sin(2\phi^*(y)) u_R(y)$$

must hold on all $y \in (y_{i-1}, y_i)$.

Suppose that for some $y \in [0,\bar{y}]$, $\phi^*(y) \in \{0, \frac{\pi}{2}\}$. By the fact that $\phi^*(y)$ is continuously differentiable and it is without loss of generality to assume $\phi^*(y) \in [0, \frac{\pi}{2}]$, it must be the case that $\phi^{*\prime}(y) = 0$ if $\phi^*(y) \in \{0, \frac{\pi}{2}\}$. In this case, $\phi^*(y)$ constant on $y \in [y_L, y_H]$ satisfies the Euler-Lagrange equations. The system

$$\frac{d}{dy}\begin{bmatrix}\phi^{*\prime}(y)\\ \phi^*(y)\end{bmatrix} = \begin{bmatrix}g(y)\sin(2\phi^*(y))u_R(y)\\ \phi^{*\prime}(y)\end{bmatrix}$$

is uniformly Lipschitz-continuous in $(\phi^*, \phi^{*\prime})$ and continuous in y on all intervals (y_{i-1}, y_i), and hence by the Picard-Lindelof theorem, a unique solution to the initial value problem on any interval $[y_{i-1}, y_i]$ exists. Consequently, if $\phi^*(y) \in \{0, \frac{\pi}{2}\}$ for any $y \in [0,\bar{y}]$, $\phi^*(y) \in \{0, \frac{\pi}{2}\}$ for all $y \in [0,\bar{y}]$.

But by the assumption that

$$\sup_{p_R \in C^1([0,\bar{y}],(0,1))} J[p_R] > \max\{\int_0^{\bar{y}} g(y)u_R(y)dy, 0\},$$

a constant solution cannot be a optimal. Therefore, $\phi^*(y) \in (0, \frac{\pi}{2})$ for all $y \in [y_L, y_H]$. Consequently, the function $p^* \in C^1([0,\bar{y}],(0,1))$ defined by

$$p^*(y) = \cos(\phi^*(y))^2$$

for $y \in [0,\bar{y}]$ is a maximizer of $J[\cdot]$ and is continuously twice-differentiable everywhere $u_R(y)$ is continuous.

D.3.2 Proof of Lemmas 5 and 7

We prove that the functionals $J_0 : C^1([0,\bar{y}],(0,1)) \to \mathbb{R}$ and $J_1 : C^1([x_L, x_H],(0,\infty)) \to \mathbb{R}$, defined by

$$J_0[p] = \int_0^{\bar{y}} g(y)p(y)u_R(y)dy - \frac{\theta}{4}\int_0^{\bar{y}} \frac{(p'(y))^2}{p(y)(1-p(y))}dy,$$
$$J_1[p] = -\int_{x_L}^{x_H} q(x)p(x)u_R(x)dx - \frac{\theta}{4}\int_{x_L}^{x_H} q(x)\frac{(p'(x))^2}{p(x)}dx,$$

are concave.

Proof. Per chapter 4, section 2.2 of Giaquinta and Hildebrandt (1996), a sufficient condition for the concavity of the functional

$$J[p] = \int_0^{\bar{y}} F(y, p(y), p'(y))dy$$

is that the Hessian

$$\begin{bmatrix} F_{22} & F_{23} \\ F_{32} & F_{33} \end{bmatrix}$$

be negative semi-definite for all p on the relevant domain. In this context,

$$F_0(y, p, v) = g(y)u_R(y)p - \frac{\theta}{4}\frac{v^2}{p(1-p)},$$

and therefore

$$\begin{bmatrix} F_{0,22}(y,p,v) & F_{0,23}(y,p,v) \\ F_{0,32}(y,p,v) & F_{0,33}(y,p,v) \end{bmatrix} = -\frac{\theta}{4}\begin{bmatrix} \frac{2}{p(1-p)} & -\frac{2v(1-2p)}{(p(1-p))^2} \\ -\frac{2v(1-2p)}{(p(1-p))^2} & \frac{2v^2(1-2p)^2}{(p(1-p))^3} + \frac{2v^2}{(p(1-p))^2} \end{bmatrix}.$$

The trace (and hence sum of the eigenvalues) is strictly negative for all $p \in (0,1)$, and the determinant (and hence product of the eigenvalues) is positive (strictly so if $v^2 > 0$), implying that all eigenvalues are weakly negative and hence that the matrix is negative semi-definite.

Similarly,

$$\begin{bmatrix} F_{1,22}(x,p,v) & F_{1,23}(x,p,v) \\ F_{1,32}(x,p,v) & F_{1,33}(x,p,v) \end{bmatrix} = -\frac{\theta}{2}q(x)\begin{bmatrix} \frac{v^2}{p^3} & -\frac{v}{p^2} \\ -\frac{v}{p^2} & \frac{1}{p} \end{bmatrix}.$$

On $p > 0$, the determinant is zero and trace negative, and hence one eigenvalue is negative and the other is zero, implying this matrix is negative semi-definite. □

D.3.3 Proof of Lemmas 11 and 8

Let $u : [y_L, y_H] \to \mathbb{R}$ be a bounded function with finitely many discontinuities. If for some $\varepsilon > 0$, $\theta > 0$, $\phi^* \in W^{1,2}([y_L - \varepsilon, y_H + \varepsilon], \mathbb{R})$ is a minimizer of

$$J[p] = \int_{y_L - \varepsilon}^{y_H + \varepsilon} F(y, p(y), p'(y)) dy,$$

where either $F = F_1$ or $F = F_0$,

$$F_0(y, \phi, v) = \begin{cases} \theta v^2 - u(y)\cos(\phi)^2 & y \in [y_L, y_H] \\ \theta v^2 & y \notin [y_L, y_H] \end{cases}$$

$$F_1(y, \phi, v) = \begin{cases} \theta v^2 + u(y)\phi^2 & y \in [y_L, y_H] \\ \theta v^2 & y \notin [y_L, y_H] \end{cases}$$

then $\phi^* \in C^1([y_L - \varepsilon, y_L + \varepsilon], \mathbb{R})$, and ϕ^* is continuously twice-differentiable on any interval on which u_R is continuous.

Proof. The functional $F(y, \phi, v)$ satisfies the growth conditions of theorem 4.12 of Dacorogna (2007). Define, for any $R > 0$,

$$\alpha_1(y) = 2\max\{R^2, 1\}|u(y)|,$$

For all $|\phi| \leq R$,

$$\begin{aligned} |F(y, \phi, v)| &\leq \alpha_1(y) + 2\theta v^2, \\ |F_\phi(y, \phi, v)| &\leq \alpha_1(y) + 2\theta v^2, \\ |F_v(y, \phi, v)|| &\leq 2\theta |v|. \end{aligned}$$

Consequently, by theorem 4.12 of Dacorogna (2007), for all $\omega \in W_0^{1,2}([y_L - \varepsilon, y_H + \varepsilon], \mathbb{R})$ (the set of $W^{1,2}$ functions with $\omega(y_L - \varepsilon) = \omega(y_H + \varepsilon) = 0$), the integrated

Euler-Lagrange equation holds:

$$\int_{y_L-\varepsilon}^{y_H+\varepsilon} [F_\phi(y,\phi^*(y),\phi^{*\prime}(y))\omega(y) + 2\theta\phi^{*\prime}(y)\omega'(y)]dy = 0.$$

Consider the particular test function defined by some $y_L \le y < y' \le y_H$,

$$\omega'(x) = \begin{cases} 0 & y \in [y_L-\varepsilon, y), \\ 1 & y \in [y, y'), \\ 0 & y \in [y', y_H], \\ -\frac{y'-y}{\varepsilon} & y \in (y_H, y_H+\varepsilon]. \end{cases}$$

It is immediate from the definition of $F(y,\phi,v)$ that if ϕ^* is a minimizer it must satisfy $\phi^{*\prime}(y) = 0$ for all $y \notin [y_L, y_H]$. Consequently, for this test function,

$$\begin{aligned} -\int_{y_L}^{y_H} F_\phi(y,\phi^*(y),\phi^{*\prime}(y))\omega(y)dy &= 2\theta\int_x^{x'} \phi^{*\prime}(x)dx \\ &= 2\theta(\phi^*(x') - \phi^*(x)). \end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} |\int_{y_L}^{y_H} F_\phi(y,\phi^*(y),\phi^{*\prime}(y))\omega(y)dy| &\le (\int_{y_L}^{y_H} F_\phi(y,\phi^*(y),\phi^{*\prime}(y))^2 dy)^{\frac{1}{2}} (\int_{y_L}^{y_H} \omega(y)^2 dy)^{\frac{1}{2}} \\ &\le (\int_{y_L}^{y_H} F_\phi(y,\phi^*(y),\phi^{*\prime}(y))^2 dy)^{\frac{1}{2}} (y'-y). \end{aligned}$$

Define $B = \max_{y\in[y_L,y_H]} |u(y)|$. For $F = F_0$, $|F_\phi(y,\phi,v)| = |u(y)\sin(2\phi)| \le B$, and consequently

$$(\int_{y_L}^{y_H} F_\phi(y,\phi^*(y),\phi^{*\prime}(y))^2 dy)^{\frac{1}{2}} \le B(y_H - y_L).$$

For $F = F_1$, $|F_\phi(y,\phi,v)| = |2u(y)\phi|$, and

$$\int_{y_L}^{y_H} F_\phi(y,\phi^*(y),\phi^{*\prime}(y))^2 dy \le 4B^2 \int_{y_L}^{y_H} \phi^*(y)^2 dy,$$

and consequently

$$(\int_{y_L}^{y_H} F_\phi(y,\phi^*(y),\phi^{*\prime}(y))^2 dy)^{\frac{1}{2}} < K$$

for some constant $K > 0$ by the square integrability of ϕ^*. We conclude that ϕ^* is Lipschitz-continuous.

Let $y_1, \ldots y_{k-1}$ be the (possibly empty) set of points of discontinuity for u, and let $y_0 = y_L$ and $y_k = y_H$. On the intervals (y_{i-1}, y_i) for $i \in \{1, \ldots, k\}$, the derivative F_ϕ is continuous. Following the arguments of propositions 1-3 in section 3.1, chapter 1 of Giaquinta and Hildebrandt (1996) proves that ϕ^* is continuously twice-differentiable on (y_{i-1}, y_i) for all $i \in \{1, \ldots, k\}$.[38] Moreover, the Euler-Lagrange equation

$$2\theta\phi^{*\prime\prime}(y) = F_\phi(y,\phi^*(y),\phi^{*\prime}(y))$$

must hold on all $y \in (y_{i-1}, y_i)$.

By the Weierstrauss-Erdmann corner conditions (or see also proposition 1 in section 3.1, chapter 1 of Giaquinta and Hildebrandt (1996)), at a hypothetical corner at y_i, we would have

$$F_v(y_i,\phi^*(y_i),v_i^-) = F_v(y_i,\phi^*(y_i),v_i^+),$$

where $v_i^- = \lim_{y\uparrow y_i} \phi^{*\prime}(y)$ and $v_i^+ = \lim_{y\downarrow y_i} \phi^{*\prime}(y)$. It follows immediately no corners exist, and hence that $\phi^{*\prime}(y)$ is continuous. □

D.4 Additional Definition and Lemmas for Convergence

Definition 4. Let X^M be a sequence of state spaces, as described in section 4.3. A sequence of policies $\{p_M \in \mathscr{P}(X^M)\}_{M\in\mathbb{N}}$ satisfies the "convergence condition" if:

i) The sequence satisfies, for some constants $c_H > c_L > 0$, all M, and all $i \in X^M$,

$$\frac{c_H}{M+1} \geq e_i^T p_M \geq \frac{c_L}{M+1}.$$

[38] In the aforementioned section of Giaquinta and Hildebrandt (1996), it is assumed that $F(y,\phi,v)$ is continuously differentiable. However, the proofs given in that section require only that F_ϕ and F_v be continuous, and not that $F(y,\phi,v)$ be differentiable in y.

ii) The sequence satisfies, for some constant $K_1 > 0$, all M, and all $i \in X^M \setminus \{0, M\}$,

$$M^3 |\frac{1}{2}(e_{i+1}^T + e_{i-1}^T - 2e_i^T) p_M| \leq K_1,$$

and

$$M^2 |\frac{1}{2}(e_M^T - e_{M-1}^T) p_M| \leq K_1$$

and

$$M^2 |\frac{1}{2}(e_1^T - e_0^T) p_M| \leq K_1.$$

Definition 5. Let $\{p_M \in \mathscr{P}(X^M)\}_{M \in \mathbb{N}}$ be a sequence of probability distributions over the state spaces associated with Theorem 7. The interpolating functions $\{\hat{p}_M \in \mathscr{P}([0,1])\}_{M \in \mathbb{N}}$ are, for $x \in [\frac{1}{2(M+1)}, 1 - \frac{1}{2(M+1)})$,

$$\begin{aligned}\hat{p}_M(x) = (M+1)((M+1)x + \frac{1}{2} - \lfloor (M+1)x + \frac{1}{2} \rfloor) e_{\lfloor (M+1)x + \frac{1}{2} \rfloor}^T p_M + \\ + (M+1)(\frac{1}{2} - (M+1)x + \lfloor (M+1)x + \frac{1}{2} \rfloor) e_{\lfloor (M+1)x + \frac{1}{2} \rfloor - 1}^T p_M,\end{aligned}$$

and, for $x \in [0, \frac{1}{2(M+1)})$,

$$\hat{p}_M(x) = (M+1) e_0^T q_M,$$

and. for $x \in [1 - \frac{1}{2(M+1)}, 1]$,

$$\hat{p}_M(x) = (M+1) e_M^T q_M.$$

Lemma 12. *Given a function $p \in \mathscr{P}([0,1])$, define the sequence $\{p_M \in \mathscr{P}(X^M)\}_{M \in \mathbb{N}}$,*

$$e_i^T p_M = \int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} p(x) dx,$$

where X^M is the state space described in section 4.3. If the function p is strictly greater than zero for all $x \in [0,1]$, differentiable, and its derivative is Lipschitz continuous, then the sequence $\{p_M \in \mathscr{P}(X^M)\}_{N \in \mathbb{N}}$ satisfies the convergence condition,

and satisfies, for some constant $K > 0$, all M, and all $i \in X^N \setminus \{0, M\}$,

$$M^2 |\ln(\frac{1}{2}(e_{i+1}^T + e_i^T)q_M) + \ln(\frac{1}{2}(e_{i-1}^T + e_i^T)q_M) - 2\ln(e_i^T q_M)| \leq K,$$

and

$$M|\ln(\frac{1}{2}(e_1^T + e_0^T)q_M) - \ln(e_0^T q_M))| < K$$

and

$$M|\ln(\frac{1}{2}(e_M^T + e_{M-1}^T)q_M) - \ln(e_M^T q_M))| < K.$$

Proof. See the technical appendix, D.4.3. □

Lemma 13. *Let $\{p_M \in \mathscr{P}(X^M)\}_{M\in\mathbb{N}}$ be a sequence of probability distributions over the state spaces associated with Theorem 7. If the sequence $\{p_M \in \mathscr{P}(X^M)\}_{M\in\mathbb{N}}$ satisfies the convergence condition (Definition 4), then there exists a sub-sequence, whose elements we denote by n, such that:*

i) The interpolating functions (5) $\hat{p}_n(x)$ converge point-wise to a differentiable function $p(x) \in \mathscr{P}([0,1])$, whose derivative is Lipschitz-continuous, with $p(x) > 0$ for all $x \in [0,1]$,

ii) the following sum converges:

$$\lim_{n\to\infty} n^2 \sum_{i\in X^n\setminus\{n\}} \{g(e_i^T p_n) + g(e_{i+1}^T p_n) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T)p_n)\} = \frac{1}{4}\int_0^1 \frac{(p'(x))^2}{p(x)}dx,$$

where $g(x) = x\ln(x)$,

iii) for all $a \in A$, $\lim_{n\to\infty} u_{a,n}^T p_n = \int_0^1 u_a(x)p(x)dx$,

iv) and, if the sequence $\{p_M \in \mathscr{P}(X^M)\}_{M\in\mathbb{N}}$ is constructed from some function $\tilde{p}(x)$, as in Lemma 12, then $p(x) = \tilde{p}(x)$ for all $x \in [0,1]$.

Proof. See the technical appendix, section D.4.4. □

Lemma 14. *Let $\pi_M(a) \in \mathscr{P}(A)$ and $\{q_{a,M} \in \mathscr{P}(X^M)\}_{a\in A}$ denote optimal policies in the discrete state setting described in section 4.3. For each $a \in A$, the sequence $\{q_{a,N}\}$ satisfies the convergence condition (Definition 4).*

Proof. See the technical appendix, section D.4.5. □

D.4.1 Proof of Theorem 7

By the boundedness of $\mathscr{P}(A)$, there exists a convergent sub-sequence of the optimal policy $\pi_n(a)$, which we also denote by n. Define

$$\pi(a) = \lim_{n\to\infty} \pi_n(a).$$

By Lemma 14, for all $a \in A$, each sequence of optimal policies $\{q_{a,n}\}$ satisfies the convergence condition (Definition 4). Therefore, by Lemma 13, each sequence of interpolating functions (5), $\{\hat{q}_{a,n}(x)\}$, has a convergent sub-sequence that converges to a differentiable function $q_a(x)$, whose derivative is Lipschitz continuous. We can construct a sub-sequence in which $\pi_n(a)$ and all $\{\hat{q}_{a,n}(x)\}$ converge by iteratively applying this argument. Pass to this subsequence.

We can write the discrete value function, defining $g(x) = x\ln x$, as

$$\begin{aligned}
V_N(q_n;n) = &\max_{\{p_{x,n}\in\mathscr{P}(A)\}_{i\in X}} \sum_{a\in A} e_a^T p_n Diag(q) u_n e_a \\
&- \theta n^2 \sum_{a\in A} (e_a^T p_n q_n) \sum_{i=0}^{n-1} [g(\frac{e_i^T q_{a,n}}{\bar{q}_{i,a,n}}) + g(\frac{e_{i+1}^T q_{a,n}}{\bar{q}_{i,a,n}})] \\
&+ \theta n^2 \sum_{i=0}^{n-1} [g(\frac{e_i^T q_N}{\bar{q}_{i,a,N}}) + g(\frac{e_{i+1}^T q_N}{\bar{q}_{i,a,N}})] \\
&- \theta n^{-1} \sum_{i=0}^{n-1} (e_i^T q_n) D_{KL}(p_n e_i || p_n q_n).
\end{aligned}$$

We can re-arrange this to

$$\begin{aligned}
V_N(q_n;n) = &\max_{\{p_{x,n}\in\mathscr{P}(A)\}_{i\in X}} \sum_{a\in A} e_a^T p_n Diag(q) u_n e_a \\
&- \theta n^2 \sum_{a\in A} (e_a^T pq) \sum_{i=0}^{n-1} [g(e_i^T q_{a,n}) + g(e_{i+1}^T q_{a,n}) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T) q_{a,n})] \\
&+ \theta n^2 \sum_{i=0}^{N-1} [g(e_i^T q_n) + g(e_{i+1}^T q_n) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T) q_n)] \\
&- \theta n^{-1} \sum_{i=0}^{N-1} (e_i^T q_N) D_{KL}(p_{i,n} || p_n q_n).
\end{aligned}$$

By Lemma 13 and the boundedness of the KL divergence,

$$\lim_{n\to\infty} V_N(q_n;n) = \sum_{a\in A}\pi(a)\int_0^1 u_a(x)q_a(x)dx \\ -\frac{\theta}{4}\sum_{a\in A}\{\pi(a)\int_0^1\frac{(q_a'(x))^2}{q_a(x)}dx\} + \frac{\theta}{4}\int_0^1\frac{(q'(x))^2}{q(x)}dx.$$

Suppose that $\pi(a)$ and the $q_a(x)$ functions do not maximize this expression (subject to the constraints stated in Theorem 7). Let $\pi^*(a)$ and $q_a^*(x)$ be some superior policy. Define, for all n,

$$\tilde{\pi}_n(a) = \pi^*(a),$$

$$e_i^T\tilde{q}_{a,n} = \int_{\frac{i}{n+1}}^{\frac{i+1}{n+1}} q_a^*(x)dx.$$

Note that, by construction, $\tilde{q}_{a,n}\in\mathscr{P}(X^n)$ and $\sum_{a\in A}\tilde{\pi}_N(a)\tilde{q}_{a,n} = q_n$. That is, the constraints of the discrete-state problem are satisfied for all n. Denote the value function under these policies as $\tilde{V}_N(q_n;n)$.

Because of the constraints stated in Theorem 7, each q_a^* satisfies the conditions of Lemma 12, and therefore the sequence $\tilde{q}_{a,n}$ satisfies the convergence condition for all $a\in A$. It follows by Lemma 13 that this sequence of policies delivers, in the limit, the value function $V_N(q)$. If this function is strictly larger than $\lim_{n\to\infty}V_N(q_n;n)$, there must exist some $\bar{n}$ such that

$$\tilde{V}_N(q_{\bar{n}};\bar{n}) > V_N(q_{\bar{n}};\bar{n}),$$

contradicting optimality. Therefore, the functions $q_a(x)$ and $\pi(a)$ are maximizers.

It remains to show that

$$\lim_{n\to\infty}\sum_{i=0}^{\lfloor xn\rfloor} e_i^T q_{a,n} = \int_0^x q_a(y)dy.$$

Note that

$$e_i^T q_{a,n} = (n+1)\int_{\frac{i}{n+1}}^{\frac{i+1}{n+1}}\hat{q}_{a,n}(\frac{2i+1}{2(n+1)})dy,$$

where $\hat{q}_{a,n}$ is the function defined in Lemma 13. Therefore, the sum is equal to

$$\sum_{i=0}^{\lfloor xn \rfloor} e_i^T q_{a,n} = \int_0^{\frac{\lfloor xn \rfloor + 1}{n+1}} \hat{q}_{a,n}\left(\frac{\lfloor (n+1)y + \frac{1}{2} \rfloor + \frac{1}{2}}{(n+1)}\right) dy.$$

By the boundedness of $\hat{q}_{a,n}$ (which follows from the convergence condition) and the dominated convergence theorem,

$$\lim_{n\to\infty} \int_0^{\frac{\lfloor xn \rfloor + 1}{n+1}} \hat{q}\left(\frac{\lfloor (n+1)y + \frac{1}{2} \rfloor + \frac{1}{2}}{(n+1)}\right) dy = \int_0^x q_a(y) dy,$$

as required.

D.4.2 Proof of Lemma 10

We begin by observing that any information structure $p \in \mathscr{P}_{LipG}(A)$ defines unconditional action frequencies $\pi \in \mathscr{P}(A)$ and posteriors $q_a \in \mathscr{P}_{LipG}([0,1])$ satisfying (37), using definitions (38). And conversely, any unconditional action frequencies and posteriors satisfying (37) define an information structure, using definitions (39). Hence the set of candidate structures is the same in both problems, and the problems are equivalent if the two objective functions are equivalent as well. It is also easily seen that in each problem, the first term of the objective function is the expected value of the DM's reward $u(x,a)$, integrating over the joint distribution for (x,a). Hence it remains only to establish that the remaining terms of the objective function are equivalent as well.

Consider any information structure $p \in \mathscr{P}_{LipG}(A)$ and the corresponding unconditional action frequencies and posteriors, and let x be any point at which $q(x) > 0$, and at which $p_a(x)$ is twice differentiable for all a (and as a consequence, $q_a(x)$ is twice differentiable for all a as well). (We note that, given the Lipschitz continuity of the first derivatives, the set of x for which this is true must be of full measure.) Then the fact that $\sum_{a\in A} p_a(x) = 1$ for all x implies that

$$\sum_{a\in A} p_a''(x) = 0, \tag{41}$$

and similarly, constraint (37) implies that

$$\sum_{a\in A} \pi(a) q_a''(x) = q''(x). \tag{42}$$

At any such point, the definition of the Fisher information implies that

$$\begin{aligned}
I^{Fisher}(x) &\equiv \sum_{a\in A} \frac{(p_a'(x))^2}{p_a(x)} \\
&= \sum_a p_a''(x) - \sum_{a\in A} p_a(x) \frac{\partial^2 \log p_a(x)}{\partial x^2} \\
&= -\frac{\pi(a) q_a(x)}{q(x)} \frac{\partial^2}{\partial x^2} [\log \pi(a) + \log q_a(x) - \log q(x)] \\
&= \frac{1}{q(x)} \left[\sum_{a\in A} \pi(a) \frac{(q_a'(x))^2}{q_a(x)} - \sum_{a\in A} \pi(a) q_a''(x) - \frac{(q'(x))^2}{q(x)} + q''(x) \right] \\
&= \frac{1}{q(x)} \left[\sum_{a\in A} \pi(a) \frac{(q_a'(x))^2}{q_a(x)} - \frac{(q'(x))^2}{q(x)} \right].
\end{aligned}$$

Here the first line is the definition of the Fisher information (given in the lemma), and the second line follows from twice differentiating the function $\log p_a(x)$ with respect to x. In the third line, the first term from the second line vanishes because of (41); the remaining term from the second line is rewritten using (39). The fourth line follows from the third line by twice differentiating each of the terms inside the square brackets with respect to x. The fifth line then follows from (42).

Since this result holds for a set of x of full measure, we obtain expression

$$\int_0^1 q(x) I^{Fisher}(x) dx = \sum_{a\in A} \pi(a) \int_0^1 \frac{(q_a'(x))^2}{q_a(x)} dx - \int_0^1 \frac{(q'(x))^2}{q(x)} dx$$

for the mean Fisher information. This shows that the information-cost terms in both objective functions are equivalent, and hence the two problems are equivalent, and have equivalent solutions.

D.4.3 Proof of Lemma 12

Proof. The function p is strictly greater than zero, and continuous, and therefore attains a maximum and minimum on $[0,1]$, which we denote with c_H and c_L, respectively. By construction,

$$e_i^T p_M \geq \frac{c_L}{M+1}$$

and likewise for c_H, satisfying the bounds.

For all $i \in X^M \setminus \{M\}$,

$$\begin{aligned}(e_{i+1}^T - e_i^T)p_M &= \int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} (p(x+\frac{1}{M+1}) - p(x))dx \\ &= \int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} \int_0^{\frac{1}{M+1}} p'(x+y)dydx\end{aligned}$$

and therefore, letting K_2 be the maximum of the absolute value of p' on $[0,1]$ (which exists by the continuity of p'), we have

$$|(e_{i+1}^T - e_i^T)p_M| \leq \frac{1}{(M+1)^2}K_2, \tag{43}$$

satisfying the convergence condition for the endpoints.

For all $i \in X^M \setminus \{0, M\}$,

$$\begin{aligned}(e_{i+1}^T + e_{i-1}^T - 2e_i^T)p_M &= \int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} (p(x+\frac{1}{M+1}) + p(x-\frac{1}{M+1}) - 2p(x))dx \\ &= \int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} \int_0^{\frac{1}{M+1}} (p'(x+y) - p'(x-y))dydx.\end{aligned}$$

Let K_3 denote the Lipschitz constant associated with p'. It follows that

$$|(e_{i+1}^T + e_{i-1}^T - 2e_i^T)p_M| \leq \frac{2K_3}{(M+1)^3}.$$

Therefore, the convergence condition is satisfied for $K_1 = \max(\frac{1}{2}K_2, K_3)$.

By the concavity of the log function, and the inequality $\ln(x) \leq x-1$,

$$\ln(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}{e_i^T p_M}) + \ln(\frac{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M}{e_i^T p_M}) \leq 2\ln(\frac{\frac{1}{4}(e_{i+1}^T + e_{i-1} + 2e_i^T)p_M}{e_i^T p_M}) \leq \frac{\frac{1}{2}(e_{i+1}^T + e_{i-1} - 2e_i^T)p_M}{e_i^T p_M}.$$

Therefore, by the convergence condition we have established,

$$\ln(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}{e_i^T p_M}) + \ln(\frac{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M}{e_i^T p_M}) \leq \frac{(M+1)K_1}{M^3 c_L} \leq \frac{2K_1}{M^2 c_L}.$$

By the inequality $-\ln(\frac{1}{x}) \leq x-1$,

$$\ln(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}{e_i^T p_M}) + \ln(\frac{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M}{e_i^T p_M}) \geq \frac{\frac{1}{2}(e_{i+1}^T - e_i^T)p_M}{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M} + \frac{\frac{1}{2}(e_{i-1}^T - e_i^T)p_M}{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M}.$$

We can rewrite this as

$$\ln(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}{e_i^T p_M}) + \ln(\frac{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M}{e_i^T p_M}) \geq (\frac{\frac{1}{2}(e_{i+1}^T + e_{i-1}^T - 2e_i^T)p_M}{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M} + \frac{\frac{1}{2}(e_{i-1}^T - e_i^T)p_M}{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M} - 1)).$$

By the bounds above,

$$\frac{\frac{1}{2}(e_{i+1}^T + e_{i-1}^T - 2e_i^T)p_M}{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M} \geq -\frac{2K_1}{M^2 c_L}$$

and, using equation (43),

$$\begin{aligned}\frac{\frac{1}{2}(e_{i-1}^T - e_i^T)p_M}{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}\left(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M} - 1\right) &= \frac{\frac{1}{2}(e_{i-1}^T - e_i^T)p_M}{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}\left(\frac{\frac{1}{2}(e_{i+1}^T - e_{i-1}^T)p_M}{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M}\right) \\ &\geq -\frac{M^2}{c_L^2}\frac{1}{(M+1)^4}(K_2)^2 \\ &\geq -\left(\frac{K_2}{2Mc_L}\right)^2.\end{aligned}$$

Therefore,

$$M^2\left|\ln\left(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T)p_M}{e_i^T p_M}\right) + \ln\left(\frac{\frac{1}{2}(e_{i-1}^T + e_i^T)p_M}{e_i^T p_M}\right)\right| \leq \frac{2K_1}{c_L} + \left(\frac{K_2}{2c_L}\right)^2.$$

For the end-points,

$$\frac{\frac{1}{2}(e_1^T - e_0^T)q_M}{\frac{1}{2}(e_1^T + e_0^T)q_M} \leq \ln\left(\frac{\frac{1}{2}(e_1^T + e_0^T)q_M}{e_0^T q_M}\right) \leq \frac{\frac{1}{2}(e_1^T - e_0^T)q_M}{e_0^T q_M}$$

and therefore

$$\left|\ln\left(\frac{\frac{1}{2}(e_1^T + e_0^T)q_M}{e_0^T q_M}\right)\right| \leq \frac{K_2}{Mc_L}.$$

A similar property holds for the other endpoint, and therefore the claim holds for $K = \max\left(\frac{K_2}{c_L}, \frac{2K_1}{c_L} + \left(\frac{K_2}{2c_L}\right)^2\right)$. □

D.4.4 Proof of Lemma 13

Proof. We begin by noting that the functions $\hat{p}_M(x)$ are absolutely continuous. Almost everywhere in $\left[\frac{1}{2(M+1)}, 1 - \frac{1}{2(M+1)}\right]$,

$$\hat{p}'_M(x) = (M+1)^2\left(e^T_{\lfloor (M+1)x + \frac{1}{2}\rfloor} - e^T_{\lfloor (M+1)x + \frac{1}{2}\rfloor - 1}\right)p_M,$$

and outside this region, $\hat{p}'_M(x) = 0$. Let $\tilde{p}'_M(x)$ denote the right-continuous Lebesgue-integrable function on $[0,1]$ such that

$$\hat{p}_M(x) = \hat{p}_M(0) + \int_0^x \tilde{p}'_M(y)dy,$$

which is equal to $\hat{p}'_M(x)$ anywhere the latter exists.

The total variation of $\tilde{p}'_M(x)$ is equal to

$$TV(\tilde{p}'_M) = \sum_{i=1}^{M-1} (M+1)^2 |(e_{i+1}^T + e_{i-1}^T - 2e_i^T)p_M)| + \\ + (M+1)^2 |(e_M^T - e_{M-1}^T)p_M| + (M+1)^2 |(e_1^T - e_0^T)p_M|.$$

By the convergence condition,

$$TV(\tilde{p}'_M) \leq \frac{(M+1)^3}{M^3} 2K_1,$$

and therefore the sequence of functions $\tilde{p}'_M(x)$ has uniformly bounded variation.

For any $1 - \frac{1}{2(M+1)} > x > y \geq \frac{1}{2(M+1)}$, the quantity

$$|\tilde{p}'_M(x) - \tilde{p}'_M(y)| = (M+1)^2 | \sum_{i=\lfloor (M+1)y + \frac{1}{2} \rfloor}^{\lfloor (M+1)x + \frac{1}{2} \rfloor} (e_{i+1}^T + e_{i-1}^T - 2e_i^T)p_M| \\ \leq \frac{(M+1)^2((M+1)(x-y)+2)}{M^3} 2K_1.$$

At the end points, for all $x \in [0, \frac{1}{2(M+1)})$,

$$|\tilde{p}'_M(\frac{1}{2(M+1)}) - \tilde{p}'_M(x)| \leq \frac{2K_1}{M+1},$$

and for all $x \in [1 - \frac{1}{2(M+1)}, 1]$,

$$|\tilde{p}'_M(x) - \lim_{y \uparrow 1 - \frac{1}{2(M+1)}} \tilde{p}'_M(y)| \leq \frac{2K_1}{M+1}.$$

By $\tilde{p}'_M(0) = 0$, we have, for all $x \in [0,1]$,

$$|\tilde{p}'_M(x)| \leq (\frac{(M+1)^2((M+1)(1 - \frac{1}{2(M+1)}) + 2)}{M^3} + \frac{1}{M+1}) 2K_1,$$

proving that $\tilde{p}'_M(x)$ is bounded uniformly in M for all $x \in [0,1]$.

Therefore Helly's selection theorem applies. That is, there exists a sub-sequence, which we denote by n, such that $\tilde{p}'_n(x)$ converges point-wise to some $p'(x)$. Moreover, by the point-wise convergence of $\tilde{p}'_M$ to p', for all $x > y$,

$$|p'(x) - p'(y)| \leq 2K_1(x-y),$$

meaning that p' is Lipschitz-continuous. By the fact that $p'(0) = 0$, this implies that $|p'(x)| \leq 2K_1$ for all $x \in [0,1]$.

By the convergence condition, $c_L \leq \hat{p}_N(0) \leq c_H$. Therefore, there exists a convergent sub-sequence. We now use n to denote the sub-sequence for which $\lim_{n\to\infty} \hat{p}_n(0) = p(0)$ and for which $\tilde{p}'_n(x)$ converges point-wise to $p'(x)$. By the dominated convergence theorem, for all $x \in [0,1]$,

$$\lim_{n\to\infty} \hat{p}_n(x) = \lim_{n\to\infty} \{\hat{p}_n(0) + \int_0^x \tilde{p}'_n(y)dy\} = p(0) + \int_0^x p'(y)dy.$$

Define the function $p(x) = p(0) + \int_0^x p'(y)dy$ for all $x \in [0,1]$. By the convergence conditions, this function is bounded, $0 < c_L \leq p(x) \leq c_H$, by construction it is differentiable, and its derivative is Lipschitz continuous. Moreover,

$$\int_0^1 p(x)dx = 1,$$

and therefore $p \in \mathscr{P}([0,1])$.

Next, consider the limiting cost function. We have, using the function $g(x) = x\ln x$ and Taylor-expanding,

$$g(y) = g(x) + g'(x)(y-x) + \frac{1}{2}g''(cy + (1-c)x)(y-x)^2$$

for some $c \in (0,1)$. Therefore,

$$g(e_i^T p_M) + g(e_{i+1}^T p_M) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T)p_M) =$$
$$\frac{1}{8}g''(c_1 e_i^T p_M + (1-c_1)\frac{1}{2}(e_i^T + e_{i+1}^T)p_M)((e_{i+1}^T - e_i^T)p_M)^2$$
$$+ \frac{1}{8}g''(c_2 e_i^T p_M + (1-c_2)\frac{1}{2}(e_i^T + e_{i+1}^T)p_M)((e_{i+1}^T - e_i^T)p_M)^2$$

for constants $c_1, c_2 \in (0,1)$. Note that, by the boundedness $\hat{p}_M(x)$ from below, $e_i^T p_M \geq (M+1)^{-1} c_L$ for all $i \in X^M$. It follows that

$$g''(c_1 e_i^T p_M + (1-c_1)\frac{1}{2}(e_i^T + e_{i+1}^T)p_M) = \frac{1}{c_1 e_i^T p_M + (1-c_1)\frac{1}{2}(e_i^T + e_{i+1}^T)p_M} \leq (M+1)c_L^{-1}.$$

Therefore,

$$0 \leq g(e_i^T p_M) + g(e_{i+1}^T p_M) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T)p_M) \leq \frac{(M+1)c_L^{-1}}{4}((e_{i+1}^T - e_i^T)p_M)^2.$$

By construction,

$$e_i^T p_M = \frac{1}{(M+1)}\hat{p}_M(\frac{2i+1}{2(M+1)}).$$

Therefore,

$$(M+1)(g(e_i^T p_M) + g(e_{i+1}^T p_M) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T)p_M)) =$$
$$g(\hat{p}_M(\frac{2i+1}{2(M+1)})) + g(\hat{p}_M(\frac{2i+3}{2(M+1)})) - 2g(\hat{p}_M(\frac{2i+2}{2(M+1)})).$$

and

$$g(e_i^T p_M) + g(e_{i+1}^T p_M) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T)p_M) \leq \frac{c_L^{-1}}{4(M+1)}(\hat{p}(\frac{2i+3}{2(M+1)}) - \hat{p}(\frac{2i+1}{2(M+1)}))^2.$$

By the boundedness of $\tilde{p}'_M(x)$,

$$g(\hat{p}(\frac{2i+1}{2(M+1)})) + g(\hat{p}(\frac{2i+3}{2(M+1)})) - 2g(\hat{p}(\frac{2i+2}{2(M+1)})) \leq \frac{B}{(M+1)^2}$$

for some finite bound B.

Writing the limiting cost as an integral, and switching to the sub-sequence n defined above,

$$n^2 \sum_{i \in X^n \setminus \{n\}} \{g(e_i^T p_n) + g(e_{i+1}^T p_n) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T)p_n)\} =$$
$$\frac{n^3}{n+1} \int_0^1 \{g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 1}{2(n+1)})) + g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 3}{2(n+1)})) - 2g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 2}{2(n+1)}))\} dx.$$

By the bound above,

$$\frac{n^3}{n+1} \int_0^1 \{g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 1}{2(n+1)})) + g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 3}{2(n+1)})) - 2g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 2}{2(n+1)}))\} dx \leq$$
$$\frac{n^3}{(n+1)^3} \int_0^1 B dx.$$

Applying the dominated convergence theorem,

$$\lim_{n \to \infty} n^2 \sum_{i \in X^n \setminus \{n\}} \{g(e_i^T p_n) + g(e_{i+1}^T p_n) - 2g(\frac{1}{2}(e_i^T + e_{i+1}^T)p_n)\} =$$
$$\int_0^1 \lim_{n \to \infty} \frac{n^3}{n+1} \{g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 1}{2(n+1)})) + g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 3}{2(n+1)})) - 2g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 2}{2(n+1)}))\} dx.$$

By the Taylor expansion above,

$$\lim_{n \to \infty} \frac{n^3}{n+1} \{g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 1}{2(n+1)})) + g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 3}{2(n+1)})) - 2g(\hat{p}_n(\frac{2\lfloor nx \rfloor + 2}{2(n+1)}))\} =$$
$$\lim_{n \to \infty} \frac{1}{8} \frac{n^3}{n+1} \{g''(\cdot) + g''(\cdot)\} (\hat{p}_n(\frac{2\lfloor nx \rfloor + 3}{2(n+1)}) - \hat{p}_n(\frac{2\lfloor nx \rfloor + 1}{2(n+1)}))^2.$$

By definition,

$$(n+1)(\hat{p}_n(\frac{2\lfloor nx \rfloor + 3}{2(n+1)}) - \hat{p}_n(\frac{2\lfloor nx \rfloor + 1}{2(n+1)})) = \tilde{p}_n'(\frac{2\lfloor nx \rfloor + 2}{2(n+1)})$$

and

$$\lim_{n\to\infty} g''(\hat{p}_n(\frac{2\lfloor nx\rfloor+2}{2(n+1)})+c_n(\hat{p}_n(\frac{2\lfloor nx\rfloor+3}{2(n+1)})-\hat{p}_n(\frac{2\lfloor nx\rfloor+2}{2(n+1)})))=\frac{1}{p(x)},$$

with $c_n \in (0,1)$ for all n, and therefore

$$\lim_{n\to\infty}\frac{n^3}{n+1}\{g(\hat{p}_n(\frac{2\lfloor nx\rfloor+1}{2(n+1)}))+g(\hat{p}_n(\frac{2\lfloor nx\rfloor+3}{2(n+1)}))-2g(\hat{p}_n(\frac{2\lfloor nx\rfloor+2}{2(n+1)}))\}=$$
$$\lim_{n\to\infty}\frac{1}{4}\frac{(p'(x))^2}{p(x)},$$

proving the second claim.

Turning to the third claim, recall that, by definition,

$$e_i^T u_{a,M}=\frac{\int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} u_a(x)q(x)dx}{\int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} q(x)dx.}$$

We define the function, for $x\in[0,1)$, as

$$u_{a,M}(x)=e^T_{\lfloor(M+1)x\rfloor}u_{a,M},$$

and let $u_{a,M}(1)=e_M^T u_{a,M}$. We also define the function

$$\tilde{x}_M(x)=\frac{2\lfloor(M+1)x\rfloor+1}{2(M+1)}.$$

By construction, $\hat{p}_M(\tilde{x}_M(x))=(M+1)e^T_{\lfloor(M+1)x\rfloor}p_{a,M}$ for all $x\in[0,1)$, and equals $e_M^T p_{a,M}$ for $x=1$. Therefore,

$$\begin{aligned}u_{a,M}^T p_M&=\sum_{i\in X^M}(e_i^T u_{a,M})(e_i^T p_M)\\&=\int_0^1 \hat{p}_M(\tilde{x}_M(x))u_{a,M}(x)dx.\end{aligned}$$

By the measurability of $u_a(x)$,

$$\lim_{M\to\infty} u_{a,M}(x) = u_a(x).$$

Therefore, by the boundedness of utilities and the dominated convergence theorem,

$$\lim_{n\to\infty} u_{a,n}^T p_n = \int_0^1 p(x)u_a(x)dx.$$

Finally, suppose that, for all M

$$e_i^T p_{a,M} = \int_{\frac{i}{M+1}}^{\frac{i+1}{M+1}} \tilde{p}(x)dx.$$

It follows that $\lim_{n\to\infty} \hat{p}_{a,n}(x) = \tilde{p}(x)$ for all $x \in [0,1]$, and therefore $\tilde{p}(x) = p(x)$. □

D.4.5 Proof of Lemma 14

Proof. We begin by noting that the conditions given for the function $q(x)$ satisfy the conditions of Lemma 12, and therefore the sequence q_M satisfies the convergence condition. We will use the constants c_H and c_L to refer to its bounds,

$$\frac{c_H}{M+1} \geq e_i^T q_M \geq \frac{c_L}{M+1},$$

and the constants K_1 and K to refer to the constants described by convergence condition and Lemma 12 for the sequence q_M. By the convention that $q_{a,M} = q_M$ if $\pi_M(a) = 0$, $q_{a,M}$ also satisfies the convergence condition whenever $\pi_M(a) = 0$.

The problem of size M is

$$V_N(q_M;M) = \max_{\pi_M\in\mathscr{P}(A),\{q_{a,M}\in\mathscr{P}(X^M)\}_{a\in A}} \sum_{a\in A} \pi_M(a)(u_{a,M}^T \cdot q_{a,M}) - \theta \sum_{a\in A} \pi_M(a) D_N(q_{a,M}||q_M;M)$$

subject to

$$\sum_{a\in A} \pi_M(a) q_{a,M} = q_M,$$

where

$$D_N(q_{a,M}||q_M;\rho,M) = M^2(H_N(q_{a,M};1,M) - H_N(q_M;1,M)) + M^{-1}(H^S(q_{a,M};M) - H^S(q_M;M)$$

and

$$H_N(q;1,M) = -\sum_{i=0}^{M-1} \bar{q}_i H^S(q_i).$$

Let u_M denote that $|X^M| \times |A|$ matrix whose columns are $u_{a,M}$. Using Lemma 1, we can rewrite the problem as

$$\begin{aligned} V_N(q_M;M) = & \max_{\{p_{i,M} \in \mathscr{P}(A)\}_{i \in X^M}} \sum_{a \in A} e_a^T p_M Diag(q) u_M e_a \\ & - \theta M^2 \sum_{i=0}^{M-1} (e_i^T q_M) D_{KL}(p_{i,M}||\frac{p_{i,M}(e_i^T q_M) + p_{i+1,M}(e_{i+1}^T q_M)}{(e_i^T + e_{i+1}^T) q_M}) \\ & - \theta M^2 \sum_{i=1}^{M} (e_i^T q_M) D_{KL}(p_{i,M}||\frac{p_{i,M}(e_i^T q_N) + p_{i-1,M}(e_{i-1}^T q_M)}{(e_i^T + e_{i-1}^T) q_M}) \\ & - \theta M^{-1} \sum_{i=0}^{M-1} (e_i^T q_M) D_{KL}(p_{i,M}||p_M q_M). \end{aligned}$$

The FOC for this problem is, for all $i \in [1, M-1]$ and $a \in A$ such that $e_a^T p_{i,M} > 0$,

$$\begin{aligned} & e_i^T u_{a,M} - \theta M^2 \ln(\frac{e_a^T p_{i,M}(e_i^T + e_{i+1}^T) q_M}{e_a^T (p_{i,M}(e_i^T q_M) + p_{i+1,M}(e_{i+1}^T q_M))}) \\ -\theta M^2 \ln(\frac{e_a^T p_{i,M}(e_i^T + e_{i-1}^T) q_M}{e_a^T (p_{i,M}(e_i^T q_M) + p_{i-1,N}(e_{i-1}^T q_M))}) & - \theta M^{-1} \ln(\frac{e_a^T p_{i,M}}{e_a^T p_M q_M}) - e_i^T \kappa_M = 0, \end{aligned}$$

where $\kappa_M \in \mathbb{R}^{M+1}$ are the multipliers (scaled by $e_i^T q_M$) on the constraints that $\sum_{a \in A} e_a^T p_{i,M} = 1$ for all $i \in X$. Defining $e_{i-1}^T q_M = e_{M+1}^T q_M = 0$, and defining $p_{-1,M}$ and $p_{M+1,M}$ in arbitrary fashion, we can recover this FOC for all $i \in X$.

Rewriting the FOC in terms of the posteriors, and again defining $e_{i-1}^T q_{a,M} =$

$e_{M+1}^T q_{a,M} = 0$, for any a such that $\pi_M(a) > 0$,

$$\begin{aligned}
e_i^T(u_{a,M} - \kappa_M) &= \theta M^2 \ln(\frac{(e_i^T q_{a,M})(1 + \frac{e_{i+1}^T q_M}{e_i^T q_M})}{(e_{i+1} + e_i)^T q_{a,M}}) + \theta M^2 \ln(\frac{(e_i^T q_{a,N})(1 + \frac{e_{i-1}^T q_N}{e_i^T q_N})}{(e_{i-1} + e_i)^T q_{a,N}}) \\
&\quad + \theta M^{-1} \ln(\frac{e_a^T p_{i,M}}{e_a^T p_M q_M}) \\
&= -\theta M^2 \ln(1 + \frac{e_{i+1}^T q_{a,M}}{e_i^T q_{a,M}}) + \theta M^2 \ln(1 + \frac{e_{i+1}^T q_M}{e_i^T q_M}) - \theta M^2 \ln(1 + \frac{e_{i-1}^T q_{a,M}}{e_i^T q_{a,M}}) \\
&\quad + \theta M^2 \ln(1 + \frac{e_{i-1}^T q_M}{e_i^T q_M}) + \theta M^{-1} \ln(\frac{e_i^T q_{a,M}}{e_i^T q_M}),
\end{aligned}$$

which can be rewritten as

$$\begin{aligned}
e_i^T(u_{a,M} - \kappa_M) &= -\theta M^2(\ln(\frac{1}{2}(e_{i+1}^T + e_i^T)q_{a,M}) + \ln(\frac{1}{2}(e_{i-1}^T + e_i^T)q_{a,M}) - (2 + M^{-3})\ln(e_i^T q_{a,M})) \\
&\quad + \theta M^2(\ln(\frac{1}{2}(e_{i+1}^T + e_i^T)q_M) + \ln(\frac{1}{2}(e_{i-1}^T + e_i^T)q_M) - (2 + M^{-3})\ln(e_i^T q_M)).
\end{aligned} \tag{44}$$

Our analysis proceeds by analyzing this first-order condition.

We next describe a series of lemmas that use this first-order condition to establish various bounds, which will ultimately be used to establish the bounds required by the convergence condition. As part of the proof, we find it useful to consider the interpolating functions $\hat{q}_{a,M}(x)$ (5) constructed from $q_{a,M}$. We define from these interpolating functions the function

$$l_{a,N}(x) = (M+1)(\ln(\hat{q}_{a,M}(x)) - \ln(\hat{q}_{a,M}(x - \frac{1}{2(M+1)})))$$

on $x \in [\frac{1}{2(M+1)}, 1]$, observing that, for any $i \in X^M \setminus \{0\}$,

$$l_{a,M}(\frac{2i+1}{2(M+1)}) = (M+1)\ln(\frac{(M+1)e_i^T q_{a,M}}{\frac{1}{2}(M+1)(e_i^T + e_{i-1}^T)q_{a,M}}),$$

and for any $i \in X^M \setminus \{M\}$,

$$l_{a,M}(\frac{2i+2}{2(M+1)}) = (M+1)\ln(\frac{\frac{1}{2}(M+1)(e_i^T + e_{i+1}^T)q_{a,M}}{(M+1)e_i^T q_{a,M}}).$$

□

Lemma 15. *For all $M \in \mathbb{N}$ and $i \in X^M \setminus \{0,M\}$, $e_i^T \kappa_M \leq B_\kappa$ for some positive constant B_κ.*

Proof. See the technical appendix, section D.4.6. □

Lemma 16. *For all $M \in \mathbb{N}$ and $i \in \{0,M\}$, $|e_i^T \kappa_M| \leq B_0$ for some positive constant B_0, and*

$$\ln(\frac{\frac{1}{2}(e_0^T + e_1^T)q_{a,M}}{e_0^T q_{a,M}}) \leq M^{-1}B_1$$

and

$$\ln(\frac{e_M^T q_{a,M}}{\frac{1}{2}(e_M^T + e_{M-1}^T)q_{a,M}}) \geq -M^{-1}B_1$$

for some positive constant B_1.

Proof. See the technical appendix, section D.4.7. □

Lemma 17. *For all $M \in \mathbb{N}$ and $j \in \{2,3,\ldots,2M+1\}$, and some positive constant B_l,*

$$|l_{a,N}(\frac{j}{2(M+1)})| \leq B_l.$$

Proof. See the technical appendix, section D.4.8. The proof uses the previous two lemmas. □

Armed with these lemmas, we prove that the convergence condition (Definition 4) is satisfied.

Proof that $\frac{c_H}{M+1} \geq e_i^T q_{a,M} \geq \frac{c_L}{M+1}$ We next apply the above lemmas to prove that the first part of the convergence condition is satisfied. Begin by observing that there must exist some $\tilde{i}_{a,M} \in X^M$ such that $e_{\tilde{i}_{a,M}}^T q_{a,M} \geq \frac{1}{N+1}$, implying that

$$\ln((M+1)e_{\tilde{i}_{a,M}}^T q_{a,M}) \geq 0.$$

By the definition of $l_{a,M}$, for any $i \in X^M \setminus \{0\}$,

$$l_{a,M}(\frac{2i+1}{2(M+1)}) + l_{a,M}(\frac{2i}{2(M+1)}) = (M+1)\ln(\frac{(M+1)e_i^T q_{a,M}}{(M+1)e_{i-1}^T q_{a,M}}).$$

For any $i > \tilde{i}_{a,M}$, using Lemma 17,

$$\begin{aligned}
\ln((M+1)e_i^T q_{a,M}) &= \ln((M+1)e_{\tilde{i}_{a,M}}^T q_{a,M}) + \sum_{j=\tilde{i}_{a,M}+1}^{i} \ln(\frac{(M+1)e_j^T q_{a,M}}{(M+1)e_{j-1}^T q_{a,M}}) \\
&= \ln((M+1)e_{\tilde{i}_{a,M}}^T q_{a,M}) + \frac{1}{M+1} \sum_{j=\tilde{i}_{a,M}+1}^{i} l_{a,M}(\frac{2j+1}{2(M+1)}) + l_{a,N}(\frac{2j}{2(M+1)}) \\
&\geq -\frac{1}{M+1} \sum_{j=\tilde{i}_{a,M}+1}^{i} 2B_l \\
&\geq -2B_l.
\end{aligned}$$

Similarly, for any $i < \tilde{i}_{a,M}$,

$$\ln((M+1)e_{\tilde{i}_{a,M}}^T q_{a,M}) = \ln((M+1)e_i^T q_{a,M}) + \sum_{j=i+1}^{\tilde{i}_{a,M}} \ln(\frac{(N+1)e_j^T q_{a,N}}{(N+1)e_{j-1}^T q_{a,N}}).$$

Therefore, for any $i < \tilde{i}_{a,M}$,

$$\ln((M+1)e_i^T q_{a,M}) \geq -\sum_{j=i+1}^{\tilde{i}_{a,M}} \ln(\frac{(M+1)e_j^T q_{a,M}}{(M+1)e_{j-1}^T q_{a,M}}),$$

and thus, using Lemma 17, for all $i \in X^M$,

$$\ln((M+1)e_i^T q_{a,M}) \geq -2B_l.$$

Repeating this argument, there must be some $\hat{i}_{a,M}$ such that $e_{\hat{i}_{a,M}}^T q_{a,M} \leq M^{-1}$, and using the bounds on $l_{a,M}$ in similar fashion yields

$$\ln((M+1)e_i^T q_{a,M}) \leq 2B_l.$$

It follows that, for all M, $a \in A$ such that $\pi_M(a) > 0$, and $i \in X^M$,

$$\frac{exp(2B_l)}{(M+1)} \geq e_i^T q_{a,M} \geq \frac{\exp(-2B_l)}{M+1}, \tag{45}$$

demonstrating that $q_{a,N}$ satisfies the first part of the convergence condition.

Proof that $M^3|\frac{1}{2}(e_{i+1}^T + e_{i-1}^T - 2e_i^T)q_{a,M}| \leq K_1$ We start by proving a bound on $(M+1)^2|\frac{1}{2}(e_{i+1}^T - e_i^T)q_{a,M}|$.

Using Lemma 17, and a Taylor expansion of $\ln(1+x)$, for some $c \in (0,1)$, for any $i \in X^M \setminus \{M\}$,

$$\begin{aligned}
|l_{a,M}(\frac{2i+2}{2(M+1)})| &= |(M+1)\ln(\frac{\frac{1}{2}(M+1)(e_i^T + e_{i+1}^T)q_{a,M}}{(M+1)e_i^T q_{a,M}})| \\
&= \frac{(M+1)|\frac{1}{2}(e_{i+1}^T - e_i^T)q_{a,M}|}{e_i^T q_{a,M} + \frac{c}{2}(e_{i+1}^T - e_i^T)q_{a,M}} \\
&\leq B_l,
\end{aligned}$$

and therefore, by the bound on $e_i^T q_{a,M}$, for any $i \in X^M \setminus \{M\}$,

$$(M+1)^2|\frac{1}{2}(e_{i+1}^T - e_i^T)q_{a,M}| \leq B_l \exp(-2B_l). \tag{46}$$

Returning to the first-order condition, for $i \in X^N \setminus \{0,N\}$, and using the bounds on utility and on the terms involving q_M,

$$\begin{aligned}
e_i^T \kappa_M \geq &-\bar{u} - \theta K + \theta M^{-1}\ln(\frac{e_i^T q_M}{e_i^T q_{a,M}}) \\
&+ \theta M^2(\ln(\frac{1}{2}(e_{i+1}^T + e_i^T)q_{a,M}) + \ln(\frac{1}{2}(e_{i-1}^T + e_i^T)q_{a,M}) - 2\ln(e_i^T q_{a,M})).
\end{aligned}$$

We have

$$M^{-1}\ln(\frac{e_i^T q_M}{e_i^T q_{a,M}}) \geq M^{-1}\ln(\frac{c_L}{\exp(2B_l)}),$$

and therefore

$$e_i^T \kappa_M \geq -\bar{u} - \theta K + M^{-1} \ln(\frac{c_L}{\exp(2B_l)}) + \theta M^2 (\ln(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T) q_{a,M}}{e_i^T q_{a,M}}) + \ln(\frac{\frac{1}{2}(e_{i-1}^T + e_i^T) q_{a,M}}{e_i^T q_{a,M}})).$$

Using the mean-value theorem, for some $c_1 \in (0,1)$,

$$\begin{aligned}\ln(\frac{\frac{1}{2}(e_{i+1}^T + e_i^T) q_{a,M}}{e_i^T q_{a,M}}) &= \ln(1 + \frac{\frac{1}{2}(e_{i+1}^T - e_i^T) q_{a,M}}{e_i^T q_{a,M}}) \\ &= \frac{e_i^T q_{a,M}}{e_i^T q_{a,M} + c_1 \frac{1}{2}(e_{i+1}^T - e_i^T) q_{a,M}} \frac{\frac{1}{2}(e_{i+1}^T - e_i^T) q_{a,M}}{e_i^T q_{a,M}},\end{aligned}$$

and likewise

$$\ln(\frac{\frac{1}{2}(e_{i-1}^T + e_i^T) q_{a,M}}{e_i^T q_{a,M}}) = \frac{\frac{1}{2}(e_{i-1}^T - e_i^T) q_{a,M}}{(1 - \frac{1}{2}c_2) e_i^T q_{a,M} + \frac{1}{2} c_1 e_{i-1}^T q_{a,M}}$$

for some $c_2 \in (0,1)$. Therefore,

$$\begin{aligned}e_i^T \kappa_M \geq & -\bar{u} - \theta K + M^{-1} \ln(\frac{c_L}{\exp(2B_l)}) \\ & + \theta M^2 (\frac{\frac{1}{2}(e_{i+1}^T - e_i^T) q_{a,M}}{(1 - \frac{1}{2}c_1) e_i^T q_{a,M} + \frac{1}{2} c_1 e_{i+1}^T q_{a,M}} + \frac{\frac{1}{2}(e_{i-1}^T - e_i^T) q_{a,M}}{(1 - \frac{1}{2}c_2) e_i^T q_{a,M} + \frac{1}{2} c_2 e_{i-1}^T q_{a,M}}).\end{aligned}$$

Multiplying through,

$$\begin{aligned}& [(1 - \frac{1}{2}c_1) e_i^T q_{a,M} + \frac{1}{2} c_1 e_{i+1}^T q_{a,M}](e_i^T \kappa_M + \bar{u} + \theta K - M^{-1} \ln(\frac{c_L}{\exp(2B_l)})) \\ & \geq \theta M^2 (\frac{1}{2}(e_{i+1}^T - e_i^T) q_{a,M} + \frac{1}{2}(e_{i-1}^T - e_i^T) q_{a,M} \frac{(1 - \frac{1}{2}c_1) e_i^T q_{a,M} + \frac{1}{2} c_1 e_{i+1}^T q_{a,M}}{(1 - \frac{1}{2}c_2) e_i^T q_{a,M} + \frac{1}{2} c_2 e_{i-1}^T q_{a,M}}). \\ & \geq \theta M^2 (\frac{1}{2}(e_{i+1}^T + e_{i-1}^T - 2e_i^T) q_{a,M} + \frac{1}{2}(e_{i-1}^T - e_i^T) q_{a,M} (\frac{\frac{1}{2}c_1 (e_{i+1}^T - e_i^T) q_{a,M} - \frac{1}{2} c_2 (e_i^T - e_{i-1}^T) q_{a,M}}{(1 - \frac{1}{2}c_2) e_i^T q_{a,M} + \frac{1}{2} c_2 e_{i-1}^T q_{a,M}})).\end{aligned}$$

Using equations (45) and (46),

$$
\begin{aligned}
&[(1-\frac{1}{2}c_1)e_i^T q_{a,M}+\frac{1}{2}c_1 e_{i+1}^T q_{a,M}](e_i^T \kappa_M+\bar{u}+\theta K-M^{-1}\ln(\frac{c_L}{\exp(2B_l)}))\\
&\geq \theta M^2(\frac{1}{2}(e_{i+1}^T+e_{i-1}^T-2e_i^T)q_{a,M}-\frac{B_l\exp(2B_l)}{(M+1)^2}(\frac{\frac{2B_l\exp(2B_l)}{(M+1)^2}}{\frac{\exp(-2B_l)}{M+1}}))\\
&\geq \theta M^2\frac{1}{2}(e_{i+1}^T+e_{i-1}^T-2e_i^T)q_{a,M}-\theta\frac{2B_l^2M^2\exp(6B_l)}{(M+1)^3}.
\end{aligned}
$$

Summing over a, weighted by $\pi_N(a)$, and applying Lemma 12,

$$
\begin{aligned}
(e_i^T \kappa_M+\bar{u}+\theta K-M^{-1}\ln(\frac{c_L}{\exp(2B_l)}))&\geq -\theta\frac{\frac{K_1}{M}+\frac{2B_l^2M^2\exp(6B_l)}{(M+1)^3}}{\frac{c_L}{(M+1)}}\\
&\geq -\theta c_L^{-1}(2K_1+2B_l^2\exp(6B_l)).
\end{aligned}
$$

Therefore, $|e_i^T \kappa_N|$ is bounded below by some $B_\kappa^+>0$ for all $i\in X^N$ (recalling that this was shown for $i\in\{0,N\}$ in Lemma 16 and in the other direction in Lemma 15).

It also follows, using equation (45), that

$$
\begin{aligned}
\theta M^2(M+1)\frac{1}{2}(e_{i+1}^T+e_{i-1}^T-2e_i^T)q_{a,M}\leq exp(2B_l)(B_\kappa^+&+\bar{u}+\theta K-M^{-1}\ln(\frac{c_L}{\exp(2B_l)})\\
&+\theta\frac{2B_l^2M^2\exp(6B_l)}{(M+1)^2},
\end{aligned}
$$

which establishes one side of the bound on $|\frac{1}{2}(e_{i+1}^T+e_{i-1}^T-2e_i^T)q_{a,M}|$.

Rewriting the FOC (equation (44)) and using Lemma 12 and the boundedness of the utility and the bound on $|e_i^T \kappa_N|$,

$$
\begin{aligned}
&-B_\kappa^+-\bar{u}-\theta K-\theta M^{-1}\ln(\frac{e_i^T q_M}{e_i^T q_{a,M}})\\
\leq&\theta M^2(\ln(\frac{1}{2}(e_{i+1}^T+e_i^T)q_{a,M})+\ln(\frac{1}{2}(e_{i-1}^T+e_i^T)q_{a,M})-2\ln(e_i^T q_{a,M})).
\end{aligned}
$$

By equation (45),

$$M^{-1}\ln(\frac{e_i^T q_M}{e_i^T q_{a,M}}) \leq M^{-1}\ln(\frac{c_H}{\exp(-2B_l)}),$$

and therefore, by the concavity of the log function,

$$-B_\kappa^+ - \bar{u} - \theta K - \theta M^{-1}\ln(\frac{c_H}{\exp(-2B_l)}) \leq 2\theta M^2 \ln(\frac{\frac{1}{4}(e_{i+1}^T + e_{i-1}^T + 2e_i^T)q_{a,M}}{e_i^T q_{a,M}}).$$

By the inequality $\ln(x) \leq x-1$,

$$-B_\kappa^+ - \bar{u} - \theta K - \theta M^{-1}\ln(\frac{c_H}{\exp(-2B_l)}) \leq 2\theta M^2 (\frac{\frac{1}{4}(e_{i+1}^T + e_{i-1}^T - 2e_i^T)q_{a,M}}{e_i^T q_{a,M}}),$$

and therefore, using the lower bound on $e_i^T q_{a,M}$ (equation (45)),

$$-B_\kappa^+ - \bar{u} - \theta K - \theta M^{-1}\ln(\frac{c_H}{\exp(-2B_l)}) \leq \theta M^2 (M+1)\frac{1}{2}(e_{i+1}^T + e_{i-1}^T - 2e_i^T)q_{a,M},$$

which proves the other side of the bound.

Proof that $M^2|\frac{1}{2}(e_1^T - e_0^T)q_{a,M}| \leq K_1$ By Lemma 17,

$$-B_l \leq (M+1)\ln(\frac{\frac{1}{2}(e_0^T + e_1^T)q_{a,M}}{e_0^T q_{a,M}}) \leq B_l.$$

Using the mean-value theorem, for some $c \in (0,1)$,

$$\ln(\frac{\frac{1}{2}(e_0^T + e_1^T)q_{a,M}}{e_0^T q_{a,M}}) = \frac{\frac{1}{2}(e_1^T - e_0^T)q_{a,M}}{(1-\frac{1}{2}c)e_0^T q_{a,M} + \frac{1}{2}c e_i^T q_{a,M}}.$$

Therefore, by equation (45),

$$\frac{\exp(2B_l)}{(M+1)^2}B_l \geq \frac{1}{2}(e_1^T - e_0^T)q_{a,M} \geq -\frac{\exp(2B_l)}{(M+1)^2}B_l,$$

proving the bound. The proof for the other endpoint is identical.

D.4.6 Proof of Lemma 15

First, using Lemma 12, for all $i \in X^M \setminus \{0, M\}$, observe that

$$M^2 |\ln(\frac{1}{2}(e_{i+1}^T + e_i^T)q_M) + \ln(\frac{1}{2}(e_{i-1}^T + e_i^T)q_M) - 2\ln(e_i^T q_M)| \leq K.$$

Rewriting the FOC (equation (44)) and using this bound,

$$\begin{aligned} e_i^T \kappa_M \leq\ & e_i^T u_{a,M} + \theta K + \theta M^{-1} \ln(e_i^T q_M) \\ & + \theta M^2 (\ln(\frac{1}{2}(e_{i+1}^T + e_i^T)q_{a,M}) + \ln(\frac{1}{2}(e_{i-1}^T + e_i^T)q_{a,M}) - (2 + M^{-3})\ln(e_i^T q_{a,M})). \end{aligned}$$

By the boundedness of the utility function, this can be rewritten as

$$e_i^T \kappa_M \leq \bar{u} + \theta K - \theta M^2 (\ln(\frac{e_i^T q_{a,M}}{\frac{1}{2}(e_{i+1}^T + e_i^T)q_{a,M}}) + \ln(\frac{e_i^T q_{a,M}}{\frac{1}{2}(e_{i-1}^T + e_i^T)q_{a,M}})) - \theta M^{-1} \ln(\frac{e_i^T q_{a,M}}{e_i^T q_M}).$$

By the concavity of the log function,

$$\begin{aligned} & \ln(\frac{1}{2}(e_{i+1}^T + e_i^T)q_{a,M}) + \ln(\frac{1}{2}(e_{i-1}^T + e_i^T)q_{a,M}) + M^{-3}\ln(e_i^T q_M) \leq \\ & \qquad (2 + M^{-3}) \ln(\frac{1}{2(2+M^{-3})}(e_{i+1}^T + e_{i-1}^T + 2e_i^T)q_{a,M} + \frac{M^{-3}}{2+M^{-3}} e_i^T q_M), \end{aligned}$$

It follows that

$$e_i^T \kappa_N \leq \bar{u} + \theta K + (2 + M^{-3})\theta M^2 \ln(\frac{\frac{1}{2(2+M^{-3})}(e_{i+1}^T + e_{i-1}^T + 2e_i^T)q_{a,M} + \frac{N^{-3}}{2+N^{-3}} e_i^T q_M}{e_i^T q_{a,M}}).$$

Exponentiating,

$$\begin{aligned} & (e_i^T q_{a,M}) \exp(-\frac{1}{2+M^{-3}} \theta^{-1} M^{-2} (\bar{u} + \bar{\theta} K - e_i^T \kappa_M)) \leq \\ & \qquad \frac{1}{2(2+M^{-3})}(e_{i+1}^T + e_{i-1}^T + 2e_i^T)q_{a,M} + \frac{M^{-3}}{2+M^{-3}} e_i^T q_M. \end{aligned}$$

Summing over a, weighted by $\pi_N(a)$,

$$(e_i^T q_M)\exp(-\frac{1}{2+M^{-3}}\theta^{-1}M^{-2}(\bar{u}+\bar{\theta}K-e_i^T\kappa_M)) \leq \\ \frac{1}{2(2+M^{-3})}(e_{i+1}^T+e_{i-1}^T+2e_i^T)q_M+\frac{M^{-3}}{2+M^{-3}}e_i^T q_M.$$

Taking logs,

$$-\frac{1}{2+M^{-3}}\theta^{-1}M^{-2}(\bar{u}+\bar{\theta}K-e_i^T\kappa_M) \leq \ln(\frac{\frac{1}{2(2+M^{-3})}(e_{i+1}^T+e_{i-1}^T+2e_i^T)q_M+\frac{M^{-3}}{2+M^{-3}}e_i^T q_M}{(e_i^T q_M)}) \\ \leq \ln(1+\frac{M^{-3}}{2+M^{-3}}+\frac{1}{2+M^{-3}}\frac{K_1M^{-3}}{c_LM^{-1}}),$$

where the last step follows by Lemma 12, recalling that c_L is the lower bound on $q(x)$. We have

$$\begin{aligned} e_i^T\kappa_N &\leq 3\theta M^2\ln(1+\frac{M^{-3}}{2+M^{-3}}+\frac{1}{2+M^{-3}}\frac{K_1}{c_L}M^{-2})+\bar{u}+\bar{\theta}K \\ &\leq \bar{u}+\theta K+\frac{3\theta M^{-1}}{2+M^{-3}}+\frac{3\theta}{2+M^{-3}}\frac{K_1}{c_L} \\ &\leq \bar{u}+\theta K+\frac{3\theta}{2}+\frac{3\theta}{2}\frac{K_1}{c_L}. \end{aligned}$$

where the second step follows by the inequality $\ln(1+x)<x$ for $x>0$.

D.4.7 Proof of Lemma 16

For the lower end point, the FOC (equation (44)) can be simplified to

$$e_0^T(u_{a,M}-\kappa_M) = -\theta M^2(\ln(\frac{1}{2}(e_1^T+e_0^T)q_{a,M})+\ln(\frac{1}{2})-(1+M^{-3})\ln(e_0^T q_{a,M})) \\ +\theta M^2(\ln(\frac{1}{2}(e_1^T+e_0^T)q_M)+\ln(\frac{1}{2})-(1+M^{-3})\ln(e_0^T q_M)).$$

Rearranging this,

$$\theta^{-1}M^{-2}e_0^T(u_{a,M}-\kappa_M)+\ln(\frac{1}{2}(e_1^T+e_0^T)q_{a,M}) =$$
$$(1+M^{-3})\ln(\frac{e_0^T q_{a,M}}{e_0^T q_M})+\ln(\frac{1}{2}(e_1^T+e_0^T)q_M).$$

Exponentiating,

$$\frac{1}{2}(e_1^T+e_0^T)q_{a,M}\exp(\theta^{-1}M^{-2}e_0^T(u_{a,M}-\kappa_M)) = (\frac{e_0^T q_{a,M}}{e_0^T q_M})^{1+M^{-3}}\frac{1}{2}(e_1^T+e_0^T)q_M.$$

By the boundedness of the utility function,

$$\frac{1}{2}(e_1^T+e_0^T)q_{a,M}\exp(\theta^{-1}M^{-2}(\bar{u}-e_0^T\kappa_M)) \geq (\frac{e_0^T q_{a,M}}{e_0^T q_M})^{1+M^{-3}}\frac{1}{2}(e_1^T+e_0^T)q_M.$$

Taking a sum over a, weighted by $\pi(a)$, and applying Jensen's inequality,

$$\frac{1}{2}(e_1^T+e_0^T)q_M\exp(\theta^{-1}M^{-2}(\bar{u}-e_0^T\kappa_M)) \geq \frac{1}{2}(e_1^T+e_0^T)q_M,$$

and therefore

$$e_0^T\kappa_M \leq \bar{u}.$$

Observing that

$$M^{-1}\ln(\frac{e_0^T q_{a,M}}{e_0^T q_M}) \leq M^{-1}\ln(\frac{M}{c_L}) \leq M^{-1}(\frac{M}{c_L}-1) \leq c_L^{-1}, \tag{47}$$

we have

$$\theta^{-1}M^{-2}e_0^T(u_{a,M}-\kappa_M)+\ln(\frac{1}{2}(e_1^T+e_0^T)q_{a,M}) \leq M^{-2}c_L^{-1}+\ln(\frac{e_0^T q_{a,M}}{e_0^T q_M})+\ln(\frac{1}{2}(e_1^T+e_0^T)q_M).$$

Exponentiating,

$$\frac{1}{2}(e_1^T+e_0^T)q_{a,M}\exp(\theta^{-1}M^{-2}(-\theta c_L^{-1}+e_0^T(u_{a,M}-\kappa_M))) \leq (\frac{e_0^T q_{a,M}}{e_0^T q_M})\frac{1}{2}(e_1^T+e_0^T)q_M$$

Using the boundedness of the utility function, then taking a sum over a, weighted by $\pi(a)$,

$$\frac{1}{2}(e_1^T + e_0^T)q_{a,M}\exp(\theta^{-1}M^{-2}(-\theta c_L^{-1} - \bar{u} - e_0^T\kappa_M)) \leq \frac{1}{2}(e_1^T + e_0^T)q_M.$$

Therefore,

$$e_0^T\kappa_M \geq -\bar{u} - \theta c_L^{-1},$$

and thus

$$|e_0^T\kappa_M| \leq B_0$$

for $B_0 = \bar{u} + \theta c_L^{-1}$. A similar argument applies to the other end-point ($e_M^T\kappa_M$).

Using the bound on utility and equation (47), the FOC requires that

$$\ln(\frac{\frac{1}{2}(e_1^T + e_0^T)q_{a,M}}{e_0^T q_{a,M}}) \leq \theta^{-1}M^{-2}(\bar{u} + B_0 + \theta c_L^{-1}) + \ln(\frac{\frac{1}{2}(e_1^T + e_0^T)q_M}{e_0^T q_M}).$$

By Lemma 12, it follows that

$$\ln(\frac{\frac{1}{2}(e_1^T + e_0^T)q_{a,M}}{e_0^T q_{a,M}}) \leq \theta^{-1}M^{-2}(\bar{u} + B_0 + \theta c_L^{-1}) + M^{-1}K,$$

and therefore the constraint with $B_1 = K + \theta^{-1}(\bar{u} + B_0 + \theta c_L^{-1})$ is satisfied.

Similarly, the FOC for the highest state is

$$\theta^{-1}M^{-2}e_M^T(u_{a,M} - \kappa_M) + \ln(\frac{\frac{1}{2}(e_M^T + e_{M-1}^T)q_{a,M}}{e_M^T q_{a,M}}) =$$
$$(1 + M^{-3})\ln(\frac{e_M^T q_{a,M}}{e_M^T q_M}) + \ln(\frac{1}{2}(e_M^T + e_{M-1}^T)q_M),$$

and therefore

$$\ln(\frac{\frac{1}{2}(e_M^T + e_{M-1}^T)q_{a,M}}{e_M^T q_{a,M}}) \leq \theta^{-1}M^{-2}(\bar{u} + B_0 + \theta c_L^{-1}) + \ln(\frac{\frac{1}{2}(e_M^T + e_{M-1}^T)q_M}{e_M^T q_M}),$$

implying that

$$\ln(\frac{\frac{1}{2}(e_M^T + e_{M-1}^T)q_{a,M}}{e_M^T q_{a,M}}) \leq \theta^{-1}M^{-2}(\bar{u} + B_0 + \theta c_L^{-1}) + M^{-1}K,$$

and therefore

$$\ln(\frac{e_M^T q_{a,M}}{\frac{1}{2}(e_M^T + e_{M-1}^T)q_{a,M}}) \geq -M^{-1}B_1.$$

D.4.8 Proof of Lemma 17

The first-order condition is, for any $i \in X^M \setminus \{0, M\}$ can be re-written using the function $l_{a,M}$ (and the function l_M, defined from $\hat{q}_M$ along the same lines) as

$$\begin{aligned} e_i^T(\kappa_M - u_{a,M}) + \theta M^{-1}\ln(\frac{e_i^T q_{a,M}}{e_i^T q_M}) = \theta\frac{M^2}{(M+1)}(l_{a,M}(\frac{2i+2}{2(M+1)}) - l_{a,M}(\frac{2i+1}{2(M+1)})) \\ - \theta\frac{M^2}{(M+1)}(l_M(\frac{2i+2}{2(M+1)}) - l_M(\frac{2i+1}{2(M+1)})). \end{aligned}$$

Note that

$$\theta M^{-1}\ln(\frac{e_i^T q_{a,M}}{e_i^T q_M}) \leq \theta M^{-1}\ln(\frac{1}{c_L M^{-1}}) \leq \theta M^{-1}(\frac{M}{c_L} - 1) \leq \theta c_L^{-1}.$$

By Lemma 12 and Lemma 15 and the bound on utility,

$$\theta\frac{M^2}{(M+1)}(l_{a,M}(\frac{2i+2}{2(M+1)}) - l_{a,M}(\frac{2i+1}{2(M+1)}) \leq B_\kappa + \bar{u} + \theta K + \theta c_L^{-1}.$$

We also have, for all $i \in X^M \setminus \{M\}$

$$\begin{aligned} &\frac{M^2}{M+1}(l_{a,M}(\frac{2i+3}{2(M+1)}) - l_{a,M}(\frac{2i+2}{2(M+1)})) \\ &= M^2(\ln(\frac{(M+1)e_{i+1}^T q_{a,M}}{\frac{1}{2}(M+1)(e_{i+1}^T + e_i^T)q_{a,M}}) - \ln(\frac{\frac{1}{2}(M+1)(e_i^T + e_{i+1}^T)q_{a,M}}{(M+1)e_i^T q_{a,M}})) \\ &\leq 0, \end{aligned}$$

by the concavity of the log function. Observe also that, by Lemma 16,

$$l_{a,M}(\frac{2}{2(M+1)}) = (M+1)\ln(\frac{\frac{1}{2}(e_0^T+e_1^T)q_{a,M}}{e_0^T q_{a,M}}) \leq \frac{M+1}{M}B_1.$$

It follows that, for all $j \in \{2,3,\ldots,2M+1\}$,

$$\begin{aligned} l_{a,M}(\frac{j}{2(M+1)}) &= l_{a,M}(\frac{2}{2(M+1)}) + \sum_{k=2}^{j-1}(l_{a,M}(\frac{k+1}{2(N+1)}) - l_{a,M}(\frac{k}{2(M+1)})) \\ &\leq \theta^{-1}(B_\kappa + \bar{u} + \theta K + \theta c_L^{-1})\frac{M+1}{M^2}(j-2) + \frac{M+1}{M}B_1. \end{aligned}$$

Similarly, for all $j \in \{2,3,\ldots,2M+1\}$,

$$l_{a,M}(\frac{2M+1}{2(M+1)}) = l_{a,M}(\frac{j}{2(M+1)}) + \sum_{k=j}^{2M}(l_{a,M}(\frac{k+1}{2(M+1)}) - l_{a,M}(\frac{k}{2(M+1)})).$$

Observing that

$$-l_{a,M}(\frac{2M+1}{2(M+1)}) = -\ln(\frac{(M+1)e_M^T q_{a,M}}{\frac{1}{2}(M+1)(e_M^T+e_{M-1}^T)q_{a,M}}) \leq \frac{M+1}{M}B_1,$$

using Lemma 16,

$$-l_{a,M}(\frac{j}{2(M+1)}) \leq \theta^{-1}(B_\kappa + \bar{u} + \theta K + \theta c_L^{-1})\frac{M+1}{M^2}(2M-j+1) + \frac{M+1}{M}B_1.$$

It follows that, for all $j \in \{2,3,\ldots,2M+1\}$,

$$\begin{aligned} |l_{a,N}(\frac{j}{2(N+1)})| &\leq \theta^{-1}(B_\kappa + \bar{u} + \theta K + \theta c_L^{-1})\frac{M+1}{M^2}(2M-1) + \frac{M+1}{M}B_1 \\ &\leq 4\theta^{-1}(B_\kappa + \bar{u} + \theta K + \theta c_L^{-1}) + 2B_1. \end{aligned}$$

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

Alexander Bloedel and Weijie Zhong. The cost of optimally-acquired information. *Unpublished Manuscript*, November 2020.

Nikolai Nikolaevich Chentsov. *Statistical decision rules and optimal inference*. Number 53. American Mathematical Soc., 1982.

Bernard Dacorogna. *Direct methods in the calculus of variations*, volume 78. Springer Science & Business Media, 2007.

Mark Dean and Nathaniel Neligh. Experimental tests of rational inattention. *Unpublished Manuscript*, June 2019.

M Giaquinta and S Hildebrandt. Calculus of variations, vol. i. number 310 in a series of comprehensive studies in mathematics, 1996.

WM Gorman. Conditions for additive separability. *Econometrica: Journal of the Econometric Society*, pages 605–609, 1968.

Benjamin Hébert. Moral hazard and the optimality of debt. *The Review of Economic Studies*, 85(4):2214–2252, 2018.

Benjamin Hébert and Michael Woodford. Rational inattention when decisions take time. *Unpublished manuscript*, October 2019.

Wassily Leontief. A note on the interrelation of subsets of independent variables of a continuous function with continuous first derivatives. *Bulletin of the American mathematical Society*, 53(4):343–350, 1947.

David P Myatt and Chris Wallace. Endogenous information acquisition in coordination games. *The Review of Economic Studies*, 79(1):340–374, 2011.

Luciano Pomatto, Philipp Strack, and Omer Tamuz. The cost of information. *arXiv preprint*, 1812.04211, December 2020.

Michael Woodford. Inattention as a source of randomized discrete adjustment. *Unpublished manuscript*, April 2008.

Ming Yang. Optimality of debt under flexible information acquisition. *The Review of Economic Studies*, 87(1):487–536, 2020.