Neighborhood-Based Information Costs*

Benjamin Hébert [†] Michael Woodford [‡]
Stanford University Columbia University

January 31, 2020

Abstract

We derive a new cost of information in rational inattention problems, the neighborhood-based cost functions, starting from the observation that many applications involve exogenous states with a topological structure. These cost functions summarize the results of a sequential information sampling problem (because they 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: security design, global games, modeling perceptual judgments, and linear-quadratic-Gaussian settings.

[*] The authors would like to thank 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, 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.[1] 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, as discussed, for example, in Woodford (2012), Caplin and Dean (2013), Dewan and Neligh (2017), and Caplin et al. (2018b).

We propose an alternative family of information costs, which we call neighborhood-based cost functions. These information costs have two particular properties (in addition to the standard ones described in, e.g., De Oliveira et al. (2017)) 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 the neighborhood-based cost functions avoid some of the problematic implications of the mutual-information cost function. The two properties are connected by an object we call the "information-cost matrix func-

[1] As explained in Cover and Thomas (2012), these theorems rely upon the possibility of "block coding" of a large number of independent instances of a given type of message, that can be jointly transmitted before any of the messages have to be decoded by the recipient. In our situation, an action must be taken in an individual decision problem, without waiting to learn about a large number of problems of the same form.

tion," which encodes the difficulty of distinguishing between pairs of states and summarizes the cost of a small amount of information.

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 extensive 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). Such behavior can be optimal only if it is costly to receive very different signals in similar states, but less costly to distinguish states that are dissimilar. That is, the information cost must capture some notion of "perceptual distance."

Motivated by these issues, we consider the properties that a plausible cost function should satisfy. As discussed in Fehr and Rangel (2011) and Woodford (2014), 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. More recently, some authors have proposed that data on stochastic choice and response time in economic contexts can be similarly modeled.[2] Consequently, one property we desire in a cost function is that it should summarize the results of a sequential information sampling process. In Hébert and Woodford (2018), we demonstrate that static rational inattention problems with any uniformly posterior-separable (UPS) cost functions can be derived from a sequential information sampling model.[3] Motivated by this result, and the experimental evidence of Dean and Neligh (2018) that is consistent with UPS cost functions, we restrict attention to UPS cost functions.

[2] Additional recent examples include Krajbich et al. (2014) and Clithero (2018). Shadlen and Shohamy (2016) provide a neural-process interpretation of sequential-sampling models of choice.

[3] For more on this class of cost functions, see Caplin et al. (2018b). Morris and Strack (2017) provide a related foundation for this class, in the special case in which there are only two possible states and signals are exogenous.

The key "parameter" of the dynamic model of Hébert and Woodford (2018) is a matrix-valued function that describes the local cost of information acquisition. This matrix-valued function is closely related to the Hessian of the information cost in the corresponding static rational inattention model. This matrix encodes, on its diagonal, how difficult each state is to learn about, and on its off-diagonal, how difficult it is to discriminate between states. Mutual information generates problematic predictions because its corresponding information-cost matrix function has a kind of symmetry that implies that equally likely states are equally difficult to discriminate. More generally, the comparative statics of the joint distribution of states and actions with respect to changes in payoffs are governed by the Hessian of the cost function. Intuitively, if it very costly to discriminate between some pair of states, the DM will not do so even if her payoff jumps across those states. As a result, there are UPS cost functions that are both consistent with sequential evidence accumulation and able to capture the idea of perceptual distance.

We introduce a specific family of such cost functions, the neighborhood-based cost functions. With these cost functions, information structures are more costly the greater the extent to which they allow intrinsically similar states of the world (states that share a "neighborhood") to be discriminated. 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.

We derive the neighborhood-based cost function from a set of three assumptions that connect the topology of the state space to the cost function, attempting to capture the idea that it is difficult to discriminate between nearby states. We show that, given a set of neighborhoods that cover the state space, these three assumptions plus uniform posterior separability uniquely determine a neighborhood-based cost function (up to a set of constants). We then relax one of our assumptions, introducing a generalized version of the neighborhood-based cost functions that builds on work by Dean and Neligh (2018). Dean and Neligh (2018) study these neighborhood-based cost functions in an experimental setting, and find that these costs fit observed behavior better than several other alternatives, including mutual information. We also show that these cost functions can explain the continuous variation of response frequencies in the perceptual experiments discussed previously.

We also specialize the neighborhood-based cost functions to a particularly useful case, in which the states can be ordered on a line. Throughout the paper, we use as a running example the case of a potential buyer of a security whose payoff depends on the value of some assets (an example based on Yang (2017)). In this case, it is natural to suppose that the states of the world are the asset values, and that it may be difficult for the DM to discriminate between nearby asset values even as the DM is more easily able to acquire information about whether the asset values will be very high or very low. We extend our analysis of this case to a continuum of states (in the rest of the paper, we use a discrete state space) and show that the limit of the neighborhood-based cost function for this neighborhood structure is the average Fisher information. This is the average value over the state space of a local measure of the discriminability of nearby states. Like mutual information, this measure is uniquely defined up to a scale parameter, and it can be used instead of mutual information in almost any context in which the states can be ordered on a line or a circle. We further extend this result to multi-dimensional state spaces, such as when states correspond to a vector of real numbers.

We next discuss four applications that illustrate how these cost functions are both different from and similar to mutual information in various respects. We study perceptual experiments, global games (building on Morris and Yang (2016)), security design (building on Yang (2017)), and a linear-quadratic-Gaussian setting of the kind treated in Sims (2010). In the first three of these, the neighborhood-based cost functions generate different predictions than mutual information.

In the popular 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. With a single-dimensional state space, the Fisher information leads to 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 generates a cost proportional to the log of the precision, which generates different predictions in the applications discussed by those authors.

With 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 (2019) 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 information cost function. Caplin et al. (2018b) 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. (2018) 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. (2018) 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.

In section 2, we begin by defining a general class of static rational inattention problems, that can be understood as reduced-form versions of the kind of dynamic evidence accumulation problem treated in Hébert and Woodford (2018). In section 3, we then state the additional assumptions that define the class of neighborhood-based cost functions. We apply the neighborhood-based cost functions to a series of applications in section 4. In section 5 we conclude.

2 Static Rational Inattention Problems

We begin by describing static rational inattention problems. Let $x \in X$ be the underlying state of the nature, and $a \in A$ be the action taken by the decision maker (DM).

A and X are finite sets, and the number of states is weakly larger than the number of actions, $|X| \geq |A|$. The DM's utility from taking action a in state x at time t 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 begin with prior beliefs $q_0 \in \mathscr{P}(X)$, where $\mathscr{P}(\cdot)$ denote the probability simplex on a set. The DM then chooses a "signal structure," consisting of a signal alphabet S (a finite 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 beliefs 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.

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,[4] 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. Note that the problem can be rewritten as a choice of the signal probabilities π_s and posteriors q_s, instead of the signal structure p; 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 $q_s = q_s(p, q_0)$.

Example. Suppose the DM is considering buying a security whose payoff is a function of the value of some assets. In this case, X is a set of possible values for the assets, and the actions are to either accept (L, "like") or reject (R) the offer, $A = \{L, R\}$. The utility of rejecting the offer is normalized to zero ($u_{R,x} = 0$), and

[4] 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.

the utility of accepting the offer is $u_{L,x} = s_x - K$, where s_x is the security payoff and K is the price. The stopping payoff $\hat{u}(\cdot)$ involves deciding, under the current beliefs, whether to accept or reject: $\hat{u}(q_\tau) = \max\{q_\tau^T \cdot s - K, 0\}$, where s is the vector of security payoffs.

In the classic formulation of Sims, a problem of the form of (1) 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,

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

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,

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

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

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

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,

$$D_H(q_s||q) = H(q_s) - H(q) - (q_s - q)^T H_q(q). \tag{5}$$

The Kullback-Leibler divergence, for example, is a Bregman divergence (see (3)), with a 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.

Assuming the H function is twice-differentiable, we can define a transformed version of its Hessian matrix H_{qq},

$$k(q) = Diag(q)H_{qq}(q)Diag(q), \tag{6}$$

where $Diag(q)$ is an $|X| \times |X|$ diagonal matrix with q_x on its diagonal. By the convexity and homogeneity of degree one of H, $k(q)$ is positive-definite and any vector z that is constant in the support of q satisfies $k(q) \cdot z = \vec{0}$.

In Hébert and Woodford (2018), we call this matrix the "information cost matrix function." We show that any static rational inattention problem (1) with a UPS cost function be justified from a continuous time problem in which the matrix-valued function $k(q)$ describes the cost of acquiring a small amount of information given current beliefs q. In particular, the diagonal elements of $k(q)$ determine the difficulty of learning about a particular state, whereas the off-diagonal elements determine the ease or difficult of distinguishing between particular pairs of states.

One possible $k(q)$ is the "inverse Fisher information matrix,"

$$k(q) = g^+(q) = Diag(q) - qq^T = \begin{bmatrix} q_1(1-q_1) & -q_1q_2 & \cdots & -q_1q_{|X|} \\ -q_1q_2 & q_2(1-q_2) & \cdots & -q_2q_{|X|} \\ \vdots & \vdots & \ddots & \vdots \\ -q_1q_{|X|} & -q_2q_{|X|} & \cdots & q_{|X|}(1-q_{|X|}) \end{bmatrix}. \tag{7}$$

This $k(q)$ matrix corresponds to the H function that is the negative of Shannon's entropy. In this case, the off-diagonal element $k_{xx'}(q)$ is equal to $-q(x)q(x')$ for any pair of states x, x'; thus it depends only on the prior probabilities of the two states, and is otherwise the same regardless of the states selected. Consequently, all pairs of distinct states with identical probabilities are equally easy or difficult to tell apart. While this kind of symmetry might seem appealing on a priori grounds for some applications, we view it as implausible for many cases of economic relevance.

Example. Continuing the example of a buyer considering a security, suppose the buyer's current beliefs q are uniformly distributed over the various asset values $x \in X$. If $k(q)$ is the inverse Fisher information matrix, the buyer finds it equally costly to discriminate between any pair of asset values x, x', regardless of how close or far apart those asset values are.

In many applications, we have a notion of some pairs of states x, x' being closer together than others. In the case of payoffs, quantities, or other economic variables that can be summarized by a single number, we usually think that it is harder to sharply discriminate between values that are close together than values that are far apart. Perceptual experiments, in which subjects classify stimuli that differ from one another in intensity or magnitude along a single dimension, are another example. A $k(q)$ that captures a notion of distance between states is

$$k(q) = \begin{bmatrix} \frac{q_1 q_2}{q_1+q_2} & -\frac{q_1 q_2}{q_1+q_2} & 0 & \cdots & 0 \\ -\frac{q_1 q_2}{q_1+q_2} & \frac{q_1 q_2}{q_1+q_2} + \frac{q_2 q_3}{q_2+q_3} & -\frac{q_2 q_3}{q_2+q_3} & \ddots & \vdots \\ 0 & -\frac{q_2 q_3}{q_2+q_3} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \frac{q_{|X|-1} q_{|X|-2}}{q_{|X|-2}+q_{|X|-1}} + \frac{q_{|X|} q_{|X|-1}}{q_{|X|-1}+q_{|X|}} & -\frac{q_{|X|-1} q_{|X|}}{q_{|X|}+q_{|X|-1}} \\ 0 & \cdots & 0 & -\frac{q_{|X|} q_{|X|-1}}{q_{|X|}+q_{|X|-1}} & \frac{q_{|X|-1} q_{|X|}}{q_{|X|}+q_{|X|-1}} \end{bmatrix}. \tag{8}$$

Here, the only non-zero off-diagonal elements $k_{xx'}(q)$ are negative elements when x' directly follows x in the ordering of states (or vice versa). This form of matrix $k(q)$ implies that an information structure is costly only to the extent that there are pairs of "neighboring" states x, x' for which the distributions of signals conditional on those states are different. This example information cost matrix function is closely related to the neighborhood-based cost functions we introduce in Section 3.

Example. Continuing the example of a buyer considering a security, if $k(q)$ is the function described in equation (8) above, the buyer finds discriminating between adjacent asset values costly, and the total information cost depends on how rapidly the signals the buyer receives change as a function of the asset value.

Aside from its a priori appeal, this alternative information-cost matrix function has different implications for the behavior of the DM. Intuitively, the Hessian of the cost function determines the comparative statics of how the DM responds to changing incentives. The recoverability result of Caplin et al. (2018b) also demonstrates that the cost function matters for behavior– if the cost function can be uniquely recovered from data on the likelihood of the DM's action in each state, then that likelihood must be influenced by the cost function. Our applications in section 4 provide additional examples of how information costs influence behavior.

At this point, we have defined the static rational inattention problem and the UPS cost functions. Our next section proposes a specific class of UPS cost functions, the neighborhood-based cost functions, that we will argue are superior in certain respects to the standard mutual information cost function.

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 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 (2018) and Caplin et al. (2018b), and the experimental evidence of Dean and Neligh (2018), 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.*

As the discussion in the previous section emphasized, there are many UPS cost functions, and they will 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.[5] Now suppose we are given with a point finite open cover of X (i.e. a finite set of open neighborhoods that cover X). Let us denote this collection of neighborhoods by $\mathscr{N}$, and let these neighborhoods be indexed by $i \in \mathscr{I}$. We will think of these neighborhoods as 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 the 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, meaning that the conditional distributions of signals conditional on any state except x are identical under p. Formally,

$$p_{x''} = \begin{cases} r & x'' \neq x \\ r' & x'' = x, \end{cases} \tag{9}$$

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

Define p'' as the signal structure that discriminates between (x, x') and all other

[5] The metric d will play no role in our analysis. What we really assume is a second-countable regular Hausdorff space. But since all such spaces are metrizable (Urysohn's metrization theorem), it is equivalent to just assume a metric space. We would like to thank Harald Uhlig for a helpful discussion on this point.

states, that is,

$$p''_{x''} = \begin{cases} r & x'' \notin \{x,x'\} \\ r' & x'' \in \{x,x'\}. \end{cases} \tag{10}$$

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 logic above, if x and x' share a neighborhood in $\mathscr{N}$, the structures p and p' should be costlier 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 apply this logic 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 (9) and (10), 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 contains a diagram with an example neighborhood structure that summarizes this assumption. To ease exposition, we have made this assumption is stronger than necessary for our results; we only require that it holds for values of r' close to r. Note also that distinguishing states within a neighborhood is relevant only when the neighborhood contains states that occur with positive probability under q_0. One implication of using a UPS cost function is that the conditional distributions for zero-probability (under the prior q_0) states are irrelevant.[6]

In general, a single state x will be contained in multiple neighborhoods in $\mathscr{N}$.

[6] To see this, observe that these conditional distributions change neither the unconditional signal probabilities $\pi(p,q_0)$ nor the posteriors $q_s(p,q_0)$ associated with positive probability signals.

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 (for example) the neighborhood N_1 and because it discriminates between x and a (possibly overlapping) different set of states in neighborhood N_2. Our next assumption states that this situation is equivalent to one in which x is split into two states, x_1 and x_2, with $x_1 \in N_1$ and $x_2 \in N_2$, but $x_1 \notin N_2$ and $x_2 \notin N_2$.

Let X' be the split space, $X' = (X \setminus \{x\}) \cup \{x_1, x_2\} \subset \mathscr{X}$. Define, for some distinct $r, r' \in \mathscr{P}(S)$, the signal structure p^1 on X' that discriminates between x_1 and all other states, $p^1_{x''} = r$ for all $x'' \neq x_1$ and $p^1_{x_1} = r'$, and define p^2 in similar fashion. Likewise, define p as the signal structure that discriminates between x and all other states on the original state space, (9).

Let $q \in \mathscr{P}(X)$ be some prior on X, and define $q^1 \in \mathscr{P}(X')$ by

$$q^1_{x''} = \begin{cases} q_{x''} & x'' \notin \{x_1, x_2\} \\ q_x & x'' = x_1 \\ 0 & x'' = x_2. \end{cases}$$

Define q^2 in analogous fashion.

Our assumption is that discriminating between x and all other states requires both differentiating x from all states in N_1 and all states in N_2, and therefore is as costly as doing these things separately.

Assumption 3. *Fix a prior $q \in \mathscr{P}(X)$ and distinct signals $r, r' \in \mathscr{P}(S)$. Suppose that some state $x \in X$ is contained in at least two neighborhoods in $\mathscr{N}$, with none of these neighborhoods entirely contained in another, and let x_1, x_2, X', q^1, q^2, p, p^1, and p^2 be defined as above. Then*

$$C(p, q; S, X) = C(p^1, q^1; S, X') + C(p^2, q^2; S, X').$$

Figure 2 contains a diagram with an example neighborhood structure that summarizes this assumption. The key implication of this assumption is that it is without loss of generality to suppose that the neighborhoods are disjoint. This implication allows us to invoke standard results on additive separability. Combining our first three assumptions, we derive an additive separability result. We present this result below, but first introduce some notation. For each neighborhood 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 can be written as*

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

where $H^i(\cdot; X_i) : \mathscr{P}(X_i) \to \mathbb{R}$ is a twice-differentiable convex function for all $i \in \mathscr{I}$.

Proof. See the appendix, section B.1. □

To pin down the specific H^i functions, we require an additional assumption. We will assume that the cost function is monotonically decreasing with respect to “garblings” of states that preserve the neighborhood structure. That is, if we create an random mapping from X to X'' that maps each $x \in X$ only to states in X'' that are in the exact same subset of neighborhoods in $\mathscr{N}$ as x, then this mapping only reduce the cost function, because it can only reduce the difference between the prior and posteriors within each neighborhood. Intuitively, by “merging” states, the priors and posteriors become closer in some sense. This kind of invariance is closely related to the “invariance under compression” axiom of Caplin et al. (2018b), and allows us to conclude that each H^i function is proportional to Shannon’s entropy.

Formally, let $X'' \subset \mathscr{X}$ be another set covered by the neighborhood covering. Define a stochastic (Markov) matrix $m_{x'',x}$, and suppose it has the following property: $m_{x'',x} > 0$ only if the subsets of $\mathscr{N}$ containing x'' and x are identical. In effect, this matrix stochastically “garbles” each state $x \in X$ into one or more $x'' \in X''$, while maintaining the neighborhood structure. We define the corresponding mapping of measures on X to measures on X'', $M : \mathscr{P}(X) \to \mathscr{P}(X'')$ by, for all $x'' \in X''$,

$$M(q)_{x''} = \sum_{x \in X} m_{x'',x} q_x. \tag{11}$$

Take as given a signal structure p defined on X, and let $q_s(p,q_0)$ be the associated posteriors. Now define a new set of posteriors, $\{q''_s \in \mathscr{P}(X'')\}_{s\in S}$ by

$$q''_s = M(q_s(p,q_0))$$

and observe that $\sum_{s\in S}\pi_s(p,q_0)q''_s = M(q_0)$, meaning that these posteriors are consistent with the prior $M(q_0)$. It follows that there exists a signal structure p'' such that $q''_s = q_s(p'',M(q_0))$ and $\pi_s(p,q_0) = \pi_s(p'',M(q_0))$. That is, there exists a signal structure that generates the garbled posteriors. Because this signal structure has posteriors and priors that are garbled within each neighborhood, and hence closer to each other, we assume that it costs weakly less than the original signal structure.

Assumption 4. *Let m be a stochastic matrix that maps measures on $X \subset \mathscr{X}$ to measures on $X'' \subset \mathscr{X}$, and suppose that $m_{x',x} > 0$ only if, for all $i \in \mathscr{I}$, $x \in N_i \iff x' \in N_i$. Then for all such m, all $q_0 \in \mathscr{P}(X)$, and all signal structures p,*

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

where M is the mapping corresponding to m defined by (11) and 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 $\pi_s(p,q_0) = \pi_s(p'',M(q_0))$.

Figure 3 contains a diagram with an example neighborhood structure that summarizes this assumption, using an invertible mapping m. In this case, equality must hold, since the inequality of this assumption applies in both directions. This form of invariance pins down, up to a scalar, a unique H^i function for each neighborhood. We call the resulting UPS cost function a *neighborhood-based cost function.*

Proposition 2. *Under Assumptions 1, 2, 3, and 4, the H function can be written as*

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

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

Proof. See the appendix, section B.2. □

These neighborhood-based cost functions are unique given the neighborhoods $\mathscr{N}$ and constants c. These neighborhoods and constants determine the difficulty of discriminating between nearby states. We consider them as 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.

Define the selection matrices E_i as the $|X_i| \times |X|$ matrices that select each of the elements of X_i from a vector of length $|X|$. The information cost matrix function associated with our neighborhood-based cost function is

$$k_N(q) = \sum_{i \in \mathscr{I}} c_i \bar{q}_i E_i^T g^+(q_i) E_i,$$

where $g^+(q_i) = Diag(q_i) - q_i q_i^T$ is the inverse Fisher information matrix defined on measures over the neighborhood X_i. We have already seen an example of this matrix, given a particular neighborhood structure, in (8).

3.1 Neighborhood with Generalized Entropy

We next introduce a generalization of the neighborhood cost function that replaces Shannon's entropy with the generalized entropy index of Shorrocks (1980). This generalization, which nests Shannon's entropy as a special case, satisfies Assumptions 1, 2, and 3, and hence Proposition 1 holds, but does not satisfy Assumption 4. The original version of this paper studied only the Shannon entropy case; our use of the generalized entropy index follows Dean and Neligh (2018), who provide experimental evidence consistent with the neighborhood-based cost function, but find that generalized entropy indices provide a better fit to their experimental data.

We start from a generalized k_N function, defined for all full-support q as

$$k_N(q;\rho) = \begin{cases} \sum_{i \in \mathscr{I}} c_i \bar{q}_i |X_i|^{1-\rho} E_i^T (g^+(q_i))^{2-\rho} E_i & \rho \neq 2, \\ \sum_{i \in \mathscr{I}} c_i \bar{q}_i |X_i|^{-1} E_i^T (I - q_i \iota^T)(I - \iota q_i^T) E_i & \rho = 2, \end{cases}$$

where $g^+(\cdot)$ is the inverse Fisher information matrix, ι is a vector of ones, ρ is a constant, and the constants c_i are strictly positive. Within each neighborhood, we have assumed that the cost of information is proportional to the inverse Fisher information to some power, nesting our axiomatically derived cost function as the $\rho = 1$ case. Using (6), we derive the corresponding entropy function, $H_N(q;\rho)$.

Lemma 1. *Let $H^{Gen}(q_i;\rho)$ be the generalized entropy index of Shorrocks (1980) on the neighborhood $i \in \mathscr{I}$, defined for any interior q_i as*

$$H^{Gen}(q_i;\rho) = \begin{cases} \frac{1}{|X_i|}\frac{1}{(\rho-2)(\rho-1)}\sum_{x\in X_i}\{(|X_i|e_x^T q_i)^{2-\rho}-1\} & \rho \notin \{1,2\} \\ -\frac{1}{|X_i|}\sum_{x\in X_i}\ln(e_x^T q_i) & \rho = 2 \\ \sum_{x\in X_i} e_x^T q_i \ln(e_x^T q_i) & \rho = 1. \end{cases}$$

The entropy function $H_N(q;\rho)$ associated with the neighborhood-based information-cost matrix function $k_N(q;\rho)$ is, for any q in the relative interior of the simplex,

$$H_N(q;\rho) = \sum_{i\in\mathscr{I}} c_i \bar{q}_i H^{Gen}(q_i;\rho),$$

and is defined on the boundary by continuity for $\rho < 2$ and as infinity for $\rho \geq 2$.

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

As mentioned above, the special case of $\rho = 1$ corresponds to (the negative of) Shannon's entropy within each neighborhood. The exponent ρ controls the curvature of the entropy function (Dean and Neligh (2018) use the following analogy: $H^{Gen}(q_i;\rho)$ is to Shannon's entropy as CRRA utility is to log utility). Using the our generalized entropy function $H_N(q;\rho)$, we can define a Bregman divergence, $D_N(q_s||q;\rho)$, as in (5), and a static rational inattention problem,[7]

$$V_N(q) = \max_{\pi\in\mathscr{P}(A),\{q_a\in\mathscr{P}(X)\}_{a\in A}} \sum_{a\in A}\pi(a)(u_a^T\cdot q_a) - \theta\sum_{a\in A}\pi(a)D_N(q_a||q;\rho). \quad (12)$$

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 the using Bayes' rule below.

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

Lemma 2. *The static rational inattention problem in (12) can be written as*

$$V_N(q) = \max_{\{p_x \in \mathscr{P}(S)\}_{x\in X}} \sum_{s\in S} \pi_s(p,q_0)\hat{u}(q_s(p,q))$$
$$-\theta \sum_{i\in\mathscr{I}} c_i |X_i|^{1-\rho} \bar{q}_i^{\rho-1} \sum_{x\in X_i} (e_x^T q)^{2-\rho} D_\rho(p_x || pE_i^T q_i),$$

where

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

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

The divergences D_ρ are known as the α-divergences (under a different parameterization) and are a transformed version of the Renyi divergences (Amari and Nagaoka (2007)). In the special case of $\rho = 1$, D_ρ is the Kullback-Leibler divergence. If $\rho = 1$ and there is only a single neighborhood, this is the standard rational inattention problem with mutual information. The relevance of alternative neighborhood structures is illustrated by the following observation.

Corollary 1. *Consider a rational inattention problem with a neighborhood-based information-cost function (Lemma 2), 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 $\rho = 1$, this result holds even if $q_x \neq q_{x'}$.*

Proof. The result follows directly from the problem in Lemma 2. □

The significance of Corollary 1 can be seen if we consider the predictions of rational inattention for a standard form of perceptual discrimination experiment. In these experiments, payments are based on correct and incorrect responses. As a result, two states in which the correct response and ex-ante likelihoods are identical will (for a single-neighborhood cost function) have the same likelihood of a correct response. Experimental evidence (intuitively) shows that in some states it is more difficult to determine the correct response than in other states.

3.2 A Specific Proposal: The Fisher Information Cost Function

Our neighborhood-based framework is flexible enough to accommodate a wide range of structures on the state space. However, in practice, there is a particular structure that is relevant for many economic models: states ordered on a line. We first discuss the case of a discrete set of states, as above, and then extend our results to allow for a continuum of states, which is useful in many applications.

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. Suppose further that $c_i = 1$ for all i, implying that it is equally difficult to distinguish two neighboring states at all points in the sequence.[8] Under these assumptions, for any full-support q,

$$\begin{aligned} H_N(q;\rho,M) = \frac{1}{\rho-2}\frac{1}{\rho-1}\sum_{j=0}^{M-1}(\frac{1}{2}(e_j^T + e_{j+1}^T)q)\times \\ \{(\frac{e_j^T q}{\frac{1}{2}(e_j^T + e_{j+1}^T)q})^{2-\rho} + (\frac{e_{j+1}^T q}{\frac{1}{2}(e_j^T + e_{j+1}^T)q})^{2-\rho} - 2\}. \qquad (13) \end{aligned}$$

[8] If c_i is the same for all i, we can without loss of generality set it equal to one, as the multiplier θ can still be used to scale the overall magnitude of information costs. Note also that if we regard our discrete model as a discrete approximation to a model in which the state is actually a continuous variable, the assumption that c_i is the same for all neighborhoods requires that we choose the spacing between discrete "states" in such a way that any two adjacent states in the sequence are equally difficult to distinguish. The construction of numerical scales with which to measure physical stimuli so that equal distances along the scale 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).

Defining the function $g(x;\rho) = \frac{1}{\rho-2}\frac{1}{\rho-1}x^{2-\rho}$, we can rewrite this expression as

$$H_N(q;\rho,M) = \sum_{j=0}^{M-1}(\frac{1}{2}(e_j^T + e_{j+1}^T)q)\{g(1-\frac{\frac{1}{2}(e_{j+1}^T - e_j^T)q}{\frac{1}{2}(e_j^T + e_{j+1}^T)q};\rho)+g(1+\frac{\frac{1}{2}(e_{j+1}^T - e_j^T)q}{\frac{1}{2}(e_j^T + e_{j+1}^T)q};\rho)-2g(1;\rho)\}.$$

This function penalizes differences in the function $g(\cdot;\rho)$ between states i and $i+1$ and their average. Because the $g(\cdot;\rho)$ function is convex, any changes in probability are penalized. As a result, it will be optimal in the static rational inattention problem for the DM to smooth posterior probabilities across states of the world.

If q_i and q_{i+1} are close to each other for all i, a second-order Taylor approximation of the function $g(u;\rho)$ around $u = 1$ clarifies this point:

$$H_N(q;\rho,M) \approx \frac{1}{4}\sum_{j=0}^{M-1}\frac{((e_{j+1}^T - e_j^T)q)^2}{\frac{1}{2}(e_j^T + e_{j+1}^T)q}. \tag{14}$$

Note that this approximation is exact in the $\rho = 0$ case, and that the approximation is the same for all values of ρ. Intuitively, all of the $H^{Gen}(q_i;\rho)$ resemble each other in the neighborhood of the uniform distribution, and hence when applied to a neighborhood with two states with similar probabilities are approximately identical.

For many applications, it is convenient to work with a continuous state space. Based on this approximation result, it is tempting to suppose that, in the limit as $M \to \infty$, if the discrete distributions q_M converge to differentiable function q,

$$\lim_{M\to\infty} H_N(q_M;\rho,M) = \frac{1}{4}\int_{supp(q)}\frac{(q'(x))^2}{q(x)}dx,$$

where $supp(q)$ denotes the support of q. Based on this intuition, we can define a continuous-state rational inattention problem:

$$\begin{aligned} V_N(q) = & \sup_{\pi\in\mathscr{P}(A),\{q_a\in\mathscr{P}_{LipG}\}_{a\in A}} \sum_{a\in A}\pi(a)\int_{supp(q)} u_a(x)q_a(x)dx \\ & -\frac{\theta}{4}\sum_{a\in A}\{\pi(a)\int_{supp(q)}\frac{(q_a'(x))^2}{q_a(x)}dx\} + \frac{\theta}{4}\int_{supp(q)}\frac{(q'(x))^2}{q(x)}dx, \end{aligned} \tag{15}$$

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:

$$V_N(q) = \sup_{\{p_a\}_{a \in A} \in \mathscr{P}_{LipG}(A)} \int_{supp(q)} q(x) \sum_{a \in A} p_a(x) u_a(x) dx - \frac{\theta}{4} \int_{supp(q)} q(x) \sum_{a \in A} \frac{(p'_a(x))^2}{p_a(x)} dx, \tag{16}$$

where $\mathscr{P}_{LipG}(A)$ is the set of mappings $\{p_a : supp(q) \to \mathbb{R}_+\}_{a \in A}$ such that for each x, $\sum_{a \in A} p_a(x) = 1$, and for each action a, the function $p_a(x)$ is a differentiable function of x with a Lipschitz-continuous derivative.

This alternative 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.[9] 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 be applied in almost any context, as long as the state space is continuous.[10] Unlike the mutual-information cost function, the Fisher-information cost function depends on the topological structure of the state space.

We prove the convergence of the static problem described in section §2 to this

[9] 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.

[10] 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. This means that the formula that we give here for the cost function does not remain appropriate under arbitrary smooth re-parameterizations of the state space. The formula can easily be generalized, however, to apply to cases in which the local degree of discriminability of nearby states is a smooth nonlinear function of x, rather than constant.

problem formally in the Technical Appendix, Section C.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 of $\rho = 1$.[11] 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 quite technical, and the relevant economics are summarized by the approximation (14).

The key challenge of the proof is to demonstrate that the DM will optimally choose a signal structure such that the posteriors are in $\mathscr{P}_{LipG}$. However, if we are willing to simply assume this, it is straightforward to extend our results to the $\rho \neq 1$ case by observing that for all values of ρ, the approximation in (14) holds. We can then immediately observe that all values of ρ lead to the same continuous-state limit. Consequently, the distinctions Dean and Neligh (2018) reach with regards to the value of ρ must rely on the way the state space they study is discretized.

Moreover, provided we are willing to assume posteriors in $\mathscr{P}_{LipG}$, it is 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

$$\lim_{M\to\infty} H_N(q_M;\rho,M) = \frac{1}{4}\int_{supp(q)} \frac{|\nabla q(x)|^2}{q(x)}dx,$$

where $\nabla q(x)$ denotes the gradient. In effect, this simply adds up the one-dimensional Fisher information costs in each dimension.[12] We can also write the multi-dimensional

[11] We also assume bounded utilities. We think the result holds for other values of ρ, and without some of our regularity assumptions, but generalizing our quite technical proof is difficult.

[12] Again, the fact that these are added with equal weights on the Fisher information for the various dimensions depends on assumption that the units in which distance is measured along the various dimensions are equivalent, in the sense that a given size distance along any dimension has the same consequence for the degree of discriminability of nearby states.

problem in terms of the signal structure using the cost function

$$C_{Fisher}(p,q;A) = \frac{\theta}{4}\int_{supp(q)} q(x)\sum_{a\in A}\frac{|\nabla p_a(x)|^2}{p_a(x)}dx. \tag{17}$$

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

4 Applications of Neighborhood-Based Cost Functions

In this section, we discuss several applications of our results. Our first application considers a linear-quadratic-Gaussian environment, and the next three study two-action environments. These two environments cover a wide range of existing applications of rational inattention (for a survey, see Mackowiak et al. (2018)).

4.1 Linear-Quadratic Gaussian Environments

In this application, we consider the classic "Linear-Quadratic-Gaussian" (LQG) tracking problem, which is a major application of the standard theory of rational inattention (see, e.g., Sims (2010)). The mutual-information cost function proposed by Sims is known to be quite convenient in this case, as it leads to a very tractable solution. Here we show that our Fisher-information cost function is equally tractable, while leading to interestingly different conclusions in some cases.

For this application, we extend the continuous-state version of our model, with the multi-dimensional Fisher-information cost function, 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 the mutual-information case. In particular, we will find that the Fisher-information cost function implies information costs that are linear in precision. Our approach thus provides foundations for a cost that has already been found to be convenient in practical 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 estimate the value of the state (i.e., to "track" 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 (17). Our problem is to choose the functions $\{p_a(x)\}_{a \in \mathbb{A}} \in \mathscr{P}_{LipG}(A)$ so as to minimize

$$V(q) = \int_X q(x) \int_A [p_a(x)(a - \gamma^T x)^2 + \frac{\theta}{4} \frac{|\nabla_x p_a(x)|^2}{p_a(x)}] da dx. \tag{18}$$

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 3. *In the linear-quadratic-Gaussian tracking problem defined in (18), 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

$$\frac{\lambda}{|\lambda|} \in \arg\max_{\hat{\lambda}:|\hat{\lambda}|=1} \hat{\lambda}^T (\Sigma_0^{-1} + \sigma^{-2}\lambda\lambda^T)^{-1}\gamma. \tag{19}$$

This $p_a(x)$ is identical to 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 B.5. □

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 $p_a(x)$.

The critical property of $\lambda^T x$ that enables this simplification is that it maximally covaries with the payoff-relevant state $\gamma^T x$ under the DM's posterior after receiving the signal s. That is, 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 (19) the vector λ maximizes covariance with γ under this posterior. An explicit formula for λ given σ is

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

where I is the identity matrix.

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 (2019) interpret this difference in the context of public signals, and demonstrate that this distinction leads to significantly different outcomes in coordination games ("beauty contests").

Armed with the knowledge that the optimal signal is conditionally Gaussian and that its variance does not depend on the state, we can rewrite the problem as a choice of the posterior variance-covariance matrix Σ_s.

Corollary 2. *Let $\mathscr{M}_L$ be the set of $L \times L$ real symmetric positive-definite matrices. The value function $V(q)$ described in (18) can be written as*

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

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 3.

Proof. See the appendix, section B.6. □

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 the 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 with the log determinant. Consequently, in the one-dimensional case, the two problems are almost 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 (2019)).

In the solution described by Proposition 3, 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 3. *In the linear-quadratic-Gaussian tracking problem defined in (18), 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 B.7. □

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.

The main features of our results are similar to but subtly different from those of the LQG tracking problem with a mutual-information cost function. We include these results to show that, if one considers the tractability of the LQG problem an appealing feature of mutual information, the problem remains tractable (and the results equally sensible) with the Fisher-information cost function. However, even though both cost functions result in Gaussian signals, Fisher-information and mutual-information cost functions imply different conclusions in interesting applications. Our next application, to psychometric functions, shows that neighborhood-based cost functions (and their Fisher-information limit) can match experimental evidence that cannot be reconciled with a mutual-information cost function.

4.2 Psychometric Functions

We next discuss neighborhood-based cost functions in the context of perceptual experiments (for example, Shadlen et al. (2007) or Dean and Neligh (2018)). Suppose that the different states $X = \{0, 1, 2, \ldots, M\}$, where M is an odd integer, represent different stimuli that may be presented to the subject, and that the subject is asked to classify the stimulus as one of two types (L or R); R is the correct answer if and only if $x > M/2$. For example, the stimuli might be visual images with different orientations relative to the vertical, with increasing values of x corresponding to increasingly clockwise orientations; the subject is asked whether the image is tilted clockwise or counter-clockwise relative to the vertical. The subject's goal is to give as many correct responses as possible; hence we suppose that $u_{x,a} = 1$ if $a = R$ and $x > M/2$ or if $a = L$ and $x < M/2$, while $u_{x,a} = 0$ in all other cases. We assume that each of the possible stimuli is presented with equal prior probability, and hence that both responses have equal ex ante probability of being correct.

Both the mutual-information cost function and generalizations of it based on a generalized entropy index represent special cases of a neighborhood-based cost function, in which all states belong to the unique neighborhood. Hence condition (iii) of Lemma 1 holds for any pair of states, and by assumption conditions (i) and (ii) hold as well. In the problem just posed, Lemma 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 severity of the information constraint changes the degree to which the probability of responding R is higher when $x > M/2$, but the response probabilities still will 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.

Alternatively, consider a neighborhood-based cost function in which the neighborhoods are given by $X_i = \{i, i+1\}$ for $i = 1, 2, \ldots, M-1$, and the constants c_i are equal to one for all neighborhoods, as in Section 3.2. Suppose further that $\rho = 1$.

With this alternative cost function, Corollary 1 no longer requires that the response frequencies be identical for any two states. Moreover, because the cost function penalizes large differences in signal frequencies (and hence in response frequencies) in the case of neighboring states, in this case an optimal policy involves a gradual increase in the probability of response R as x increases, even though the payoffs associated with the different actions jump abruptly at a particular value of x. This is illustrated in Figure 5, which again shows the optimal response frequencies as a function of x, for alternative θ, with the cost function just described. The sigmoid functions predicted with this cost function — with the property that response frequencies differ only modestly from 50 percent when the stimuli are near the threshold of being correctly classified one way or the other, and yet approach zero or one in the case of stimuli that are sufficiently extreme — are characteristic of measured "psychometric functions" in perceptual experiments of this kind.[13]

4.3 Global Games and The Fisher-Information Cost Function

The continuity of choice probabilities despite discrete changes in payoffs is also an important issue for the "global games" literature (Morris and Yang (2016)). This literature typically assumes a continuum of states, so for this application we will discuss the continuous-state limit described in (15). We will compare the implications of the Fisher-information cost function proposed in Section 3.2 with those of the more standard mutual-information cost function.

This application is motivated by the work of Yang (2015) and Morris and Yang (2016), who study global games (e.g. Morris and Shin (1998)) with endogenous information acquisition. In the well-known analysis of Morris and Shin (1998), with exogenous private information, there is a unique equilibrium despite the incentives for coordination across DMs (subject to some caveats and details that are not relevant for our discussion). In contrast, Yang (2015) demonstrates that allowing for endogenous information acquisition, with mutual information as the information cost, restores a multiplicity of equilibria.

[13] For the general concept of a psychometric function, see, for example, Gabbiani and Cox (2010), chap. 25, especially Figures 25.1 and 25.2, and discussion on p. 360; or Gold and Heekeren (2014), p. 356. For an example of an empirical psychometric function for the kind of task discussed in the text (classification of the dominant direction of motion for a field of moving dots), see Shadlen et al. (2007), Figure 10.1A. Note not only that the curve is monotonically increasing, with many data points corresponding to different response probabilities between zero and one, but also that the subject's reward function is clearly of the kind assumed in the text: only two possible reward levels (for correct vs. incorrect responses), with a discontinuous change in the reward where the sign of the "motion strength" changes from negative to positive.

The key to Yang's result is that DMs can tailor the signals they receive to sharply discriminate between nearby states of the world. As a result, they can all coordinate their decision (say, to invest or not) on a particular threshold, and there are many such thresholds that can represent equilibria if coordinated upon. But this result depends on the fact that the mutual-information cost function does not make it costly to have abrupt changes in signal probabilities as the state of the world changes continuously. Morris and Yang (2016) develop the complementary result, showing that even in the case of an endogenous information structure, if signal probabilities must vary continuously with the state, there is again a unique equilibrium.

Here we show that a neighborhood-based cost function can provide a justification for the kind of continuity condition that the result of Morris and Yang (2016) requires. Those authors study a global game with two possible actions, "invest" and "not-invest," with equilibrium behavior characterized by a probability $s(x)$ of investing when the state is x. Their equilibrium uniqueness result depends on an assumption of continuous choice, meaning that for all information costs $\theta > 0$ and all parameterizations of the relevant utility function, $s(x)$ is absolutely continuous on a compact interval for which $q(x)$ has full support.

In our continuous state problem, (15), agents always choose posteriors that are differentiable, with a Lipschitz-continuous derivative. By assumption, the prior is also differentiable with a Lipschitz-continuous derivative. Therefore the function

$$s(x) = \frac{q_{invest}(x)}{q(x)} \pi_{invest}$$

is differentiable with respect to x in the support of q. By the Lipschitz-continuity of $q'_{invest}(x)$ and $q'(x)$, and the fact that $q(x)$ has full support over the relevant compact interval, the derivative of $s(x)$ is bounded, and hence $s(x)$ is absolutely continuous.

Thus, our proposal provides a micro-foundation for the continuous choice assumption of Morris and Yang (2016), and hence for uniqueness in global games.

4.4 Security Design

Our last application considers the security design model with adverse selection in Yang (2017),[14] which builds on the buyer's decision problem that we have used as an example. The purpose of this example is to show that neighborhood-based cost functions remain tractable (at least computationally) in this application, and to demonstrate an interesting implication of contracting in the presence of a buyer who will always choose continuous choice probabilities. We will briefly summarize the environment, and encourage readers to refer to Yang (2017) for a richer exposition.

A seller offers a security $s \in \mathbb{R}_+^{|X|}$, whose payoffs are contingent on the realized value of the assets backing the security, $x \in X \subseteq \mathbb{R}_+$, to a buyer at a price K. The buyer's problem (our example earlier in the paper) 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 $\beta < 1$, relative to the buyer. The security is constrained by limited liability, $0 \leq e_x^T s \leq x$. The seller designs the security and offers a price,

$$\max_{s,K \geq 0} \pi_L(s,K) q_L(s,K)^T (K\iota - \beta s) \cdot$$

subject to the limited liability constraint, where ι is a vector of ones. 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 $A = \{L,R\}$,

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

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

Yang (2017) shows that, with the standard rational inattention cost function (D_H is the Kullback-Leibler divergence), the optimal security design is a debt contract, $s(x) = \min\{v(x), \bar{v}\}$ for some positive constant $\bar{v}$. The analysis involves two different cases, depending on whether the seller attempts to ensure acceptance with

[14] 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).

certainty ($\pi_L = 1$) or not, but the form of the optimal security is the same in both cases. To simplify our exposition, we focus on the case with some possibility of rejection ($\pi_L < 1$), and discuss acceptance with certainty in appendix section §C.2.

We explore, numerically, how the result of Yang (2017) changes with alternative $H(\cdot)$ functions. We consider three alternatives, our neighborhood-based function H_N with our pairwise neighborhood structure (equation (13)), a generalized entropy index cost function (the 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 (2017) 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, a result we speculate could be proven analytically in the continuous-state version of the model.

Below, we describe our calculation procedure, and the parameters we use to generate figures 6 and 7. 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," solving

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

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 limited liability constraints.[15] Combining the first-order conditions of this security design problem and the 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 entropy function, and hence the information cost matrix 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.[16]

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 (2018) 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 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

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

[16] 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.

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 monotone (in x) security designs. Our numerical calculations recover the result of Yang (2017) 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.[17] 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 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. However, when we restrict the security to be monotone, the optimal security is a debt, suggesting that the result of Yang (2017) is robust to using neighborhood cost functions (but not the other two alternatives) under this additional restriction. We conjecture that it is possible to prove the optimality of debt among monotone securities with a Fisher information cost, in the continuous state case.[18]

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

[18] 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

5 Conclusion

In many applications of rational inattention, the space of exogenous states has a structure– for example, that of numbers ordered on line. Imposing assumptions on the structure of the state space, and assuming a uniformly posterior separable cost function, we have derived the 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 the results of perceptual experiments. In contrast, the standard rational inattention cost function, mutual information, cannot match the results of these experiments.

Moreover, we have shown that the neighborhood-based cost functions and their continuous state limit, the Fisher information cost function, make predictions that differ from those of mutual information in important settings: linear-quadratic Gaussian problems, global games, and security design. The Fisher information cost function, in particular, is a one-free-parameter family of cost functions that can be used in place of mutual information in any application in which states are continuous and described by a vector of real numbers.

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

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*, 2018b.

John A Clithero. Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, 148:344–375, 2018.

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

Henrique De Oliveira, Tommaso Denti, Maximilian Mihm, and Kemal Ozbek. Rationally

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.

inattentive preferences and hidden information costs. *Theoretical Economics*, 12:621–624, 2017.

Mark Dean and Nathaniel Neligh. Experimental tests of rational inattention. Technical report, Working Paper, Columbia University, 2018.

Ambuj Dewan and Nate Neligh. Estimating information cost functions in models of rational inattention. *Unpublished manuscript*, January 2017.

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

Fabrizio Gabbiani and Steven J. Cox. *Mathematics for Neuroscientists*. Academic Press, 2010.

Joshua I. Gold and Hauke R. Heekeren. Neural mechanisms for perceptual decision making. In Paul W. Glimcher and Ernst Fehr, editors, *Neuroeconomics: Decision Making and the Brain, 2d ed.* Academic Press, 2014.

Benjamin Hébert. Moral hazard and the optimality of debt. *The Review of Economic Studies*, 85(4):2214–2252, 2018.

Benjamin Hébert and Jennifer La'O. Information acquisition, efficiency, and non-fundamental volatility. 2019.

Benjamin Hébert and Michael Woodford. Rational inattention in continuous time. *Unpublished manuscript*, September 2018.

Ian Krajbich, Bastiaan Oud, and Ernst Fehr. Benefits of neuroeconomics modeling: New policy interventions and predictors of preference. *American Economic Review*, 104(5): 501–506, 2014.

Bartosz Mackowiak, Filip Matejka, and Mirko Wiederholt. Rational inattention: A disciplined behavioral model. 2018.

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 Philipp Strack. The Wald problem and the equivalence of sequential sampling and static information costs. *Unpublished manuscript*, June 2017.

Stephen Morris and Ming Yang. Coordination and the relative cost of distinguishing nearby states. *Unpublished manuscript*, 2016.

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 arXiv:1812.04211*, 2018.

Michael Shadlen and Daphna Shohamy. Decision making and sequential sampling from memory. *Neuron*, 90(5):927–939, 2016.

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.

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

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

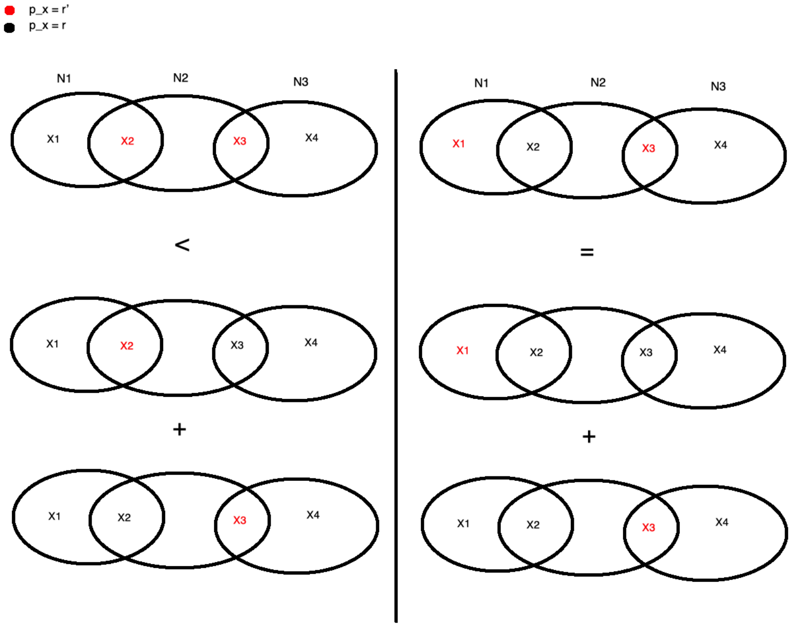
Michael Woodford. An optimizing neuroeconomic model of discrete choice. Technical report, National Bureau of Economic Research, February 2014.

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

Ming Yang. Optimality of debt under flexible information acquisition. 2017.

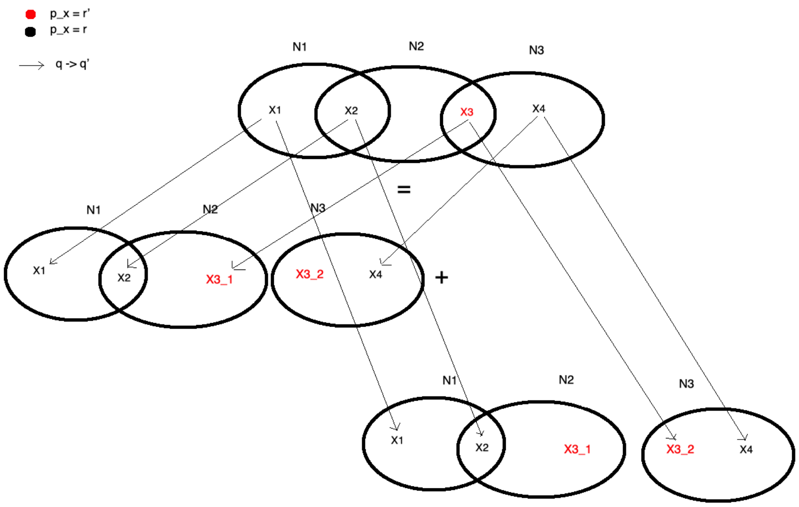
A Figures

Figure 1: Diagram for Assumption 2



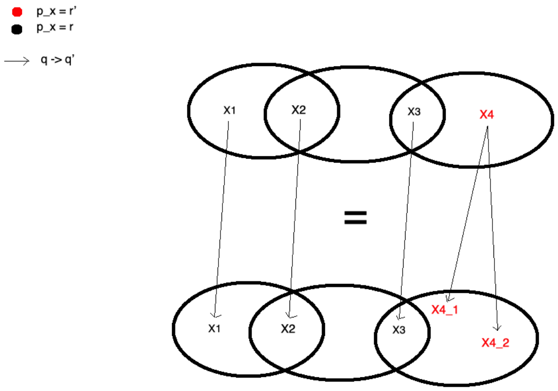
Notes: $X = \{X1, X2, X3, X4\}$ in this diagram. Each circle denotes a neighborhood, $\mathscr{N} = \{N1, N2, N3\}$. Under the signal structure p, red/gray colored states have signal distribution r', whereas black-colored states have signal distribution r. The left-hand side describes a situation in 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, $\mathscr{N} = \{N1, N2, N3\}$. Under the signal structure p, red/gray colored states have signal distribution r', whereas black-colored states have signal distribution r. The arrows describe how the probability of $q \in \mathscr{P}(X)$ is assigned to $q^1, q^2 \in \mathscr{P}(X')$.

Figure 3: Diagram for Assumption 4



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

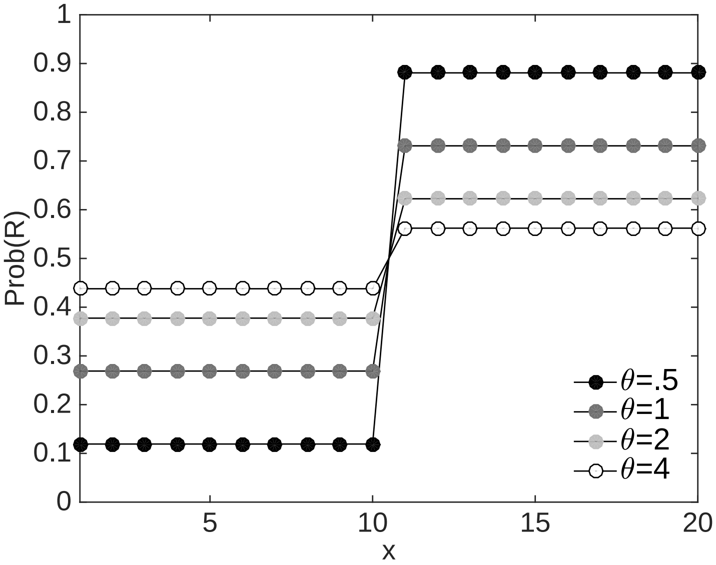


Figure 4: Predicted response probabilities with a mutual-information cost function, for alternative values of the cost parameter θ.

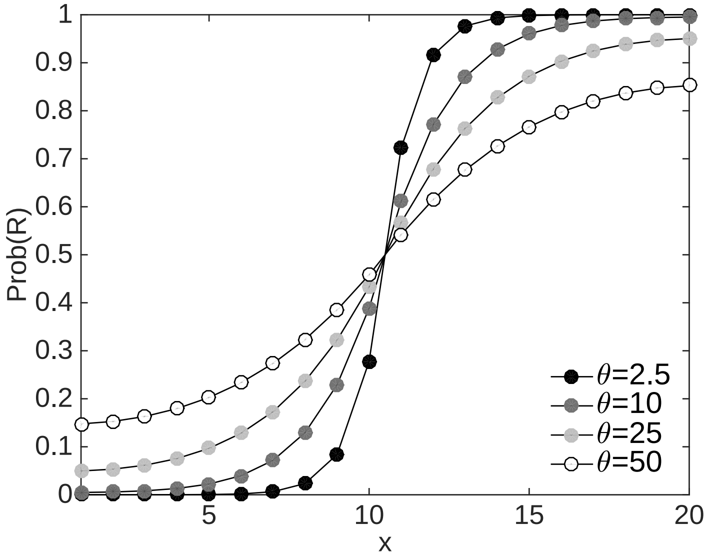


Figure 5: Predicted response probabilities with a neighborhood-based cost function, in which each neighborhood consists only of two adjacent states.

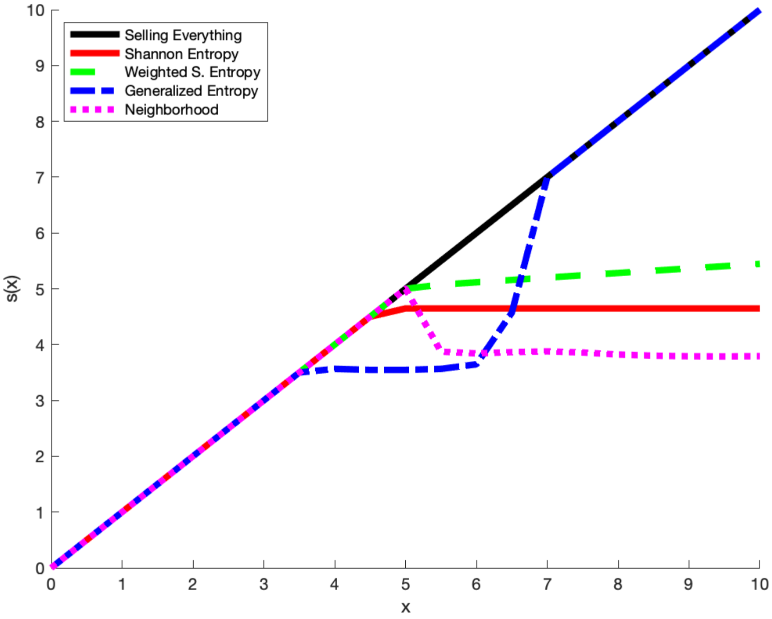


Figure 6: Optimal Security Designs by Entropy Function

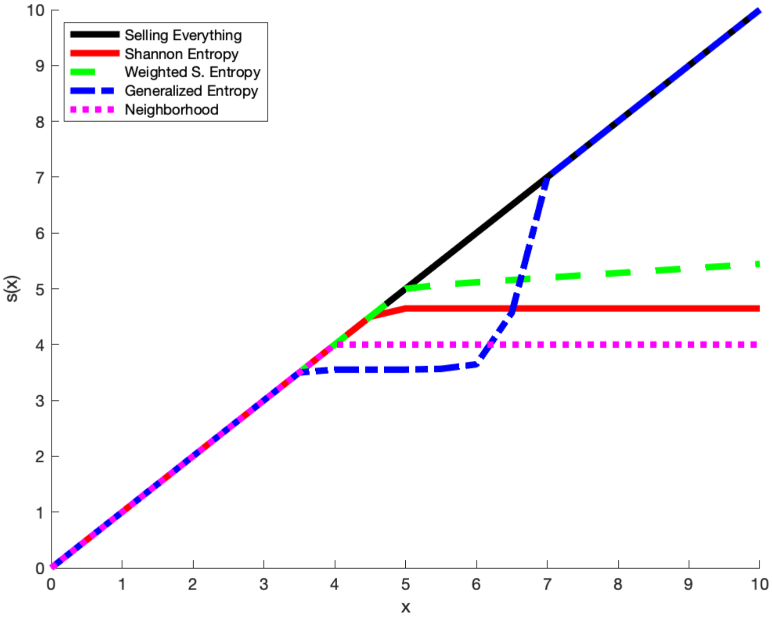


Figure 7: Optimal Monotone Security Designs by Entropy Function

Online Appendix

B Proofs

B.1 Proof of Proposition 1

This proof will refer to a few results in Hébert and Woodford (2018).

Observe that all continuously twice-differentiable UPS cost functions satisfy conditions 1-4 of Hébert and Woodford (2018) (see Lemma 8 and its proof). Suppose that

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

for some $\varepsilon > 0$ and non-zero $v \in \mathbb{R}^{|S|}$, and that r has full support on S. Then by Lemma 11 of Hébert and Woodford (2018),

$$C(p, q_0; S) = \frac{1}{2}\varepsilon^2 k_{x,x}(q_0) v^T \cdot g(r) \cdot v + o(\varepsilon^2)$$

where $g(r) = Diag(r)^{-1} - \iota\iota^T$ and $k(\cdot)$ is the information cost matrix function. Similarly,

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

and

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

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, $k_{x,x'}(q_0) < 0$.

Observe also that if $q_{0,x} = 0$, we must have $k_{xx}(q_0) = 0$ by the fact that p is costless under these circumstances, which follows from Assumption 1 as discussed in the text.

By the definition of the information cost matrix (equation (6)), this property also applies to the Hessian matrix of H. That is, if x and x' share a neighborhood in $\mathscr{N}$, then

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

and otherwise

$$\frac{\partial^2}{\partial q_x \partial q_{x'}} H(q) = 0.$$

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

$$\frac{\partial}{\partial q_x} H(q) = \frac{\partial}{\partial q_x} H(q')$$

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

We next take a detour to argue that it is without loss of generality to suppose neighborhoods are disjoint. Observe first that if one neighborhood is entirely contained in another, the existence of the smaller neighborhood imposes no additional restrictions under our assumptions. Consequently, it is without loss of generality to assume no neighborhood is contained in another neighborhood.

Under these circumstances, if there is an $x \in X$ contained in multiple neighborhoods, we can split it by Assumption 3. In the context of that assumption, define $q' \in \mathbb{R}_+^{|X'|}$ by

$$q'_{x''} = \begin{cases} q_{x''} & x'' \notin \{x_1, x_2\} \\ q_x & x'' \in \{x_1, x_2\}. \end{cases}$$

It follows immediately that, because q' differs from q^1 only on the mass x_2,

$$C(p^1, q^1; S, X') = C(p^1, q'; S, X'),$$

and likewise $C(p^2, q^2; S, X') = C(p^2, q'; S, X')$. Defining the signal structure p' by $p'_{x''} = r\mathbf{1}(x'' \notin \{x_1, x_2\}) + r'\mathbf{1}(x'' \in \{x_1, x_2\})$, by Assumption 2,

$$\begin{aligned} C(p', q'; S, X') &= C(p^1, q'; S, X') + C(p^2, q'; S, X') \\ &= C(p, q; S, X). \end{aligned}$$

Consequently, it is without loss of generality to write the problem on the split space, and repeating this argument implies it is without loss of generality to assume neighborhoods are disjoint on X, provided that we do not impose the requirement that $q \in \mathscr{P}(X)$. However, the costs when $q \notin \mathscr{P}(X)$ are determined by Assumption 1 and the homogeneity of degree one of H.

Observe, by the strict positivity and homogeneity of degree one of H, that at least one partial derivative must be positive. Consequently, by the General Theorem on Functional Dependence (see Leontief (1947) and Gorman (1968)), separability holds:

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

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

By the condition

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

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

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

Under Assumption 1, the level of the cost functions $\hat{H}^i(q_i,\bar{q}_i;X_i)$ 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

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

which is the result.

B.2 Proof of Proposition 2

As argued in the proof of section B.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 (2018) 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.

B.3 Proof of Lemma 1

We have, for any interior q,

$$\begin{aligned} H_N(q;\rho) &= -\sum_{i\in\mathscr{I}} c_i\bar{q}_i H^{Gen}(q_i;\rho) \\ &= \sum_{i\in\mathscr{I}} c_i\bar{q}_i \frac{1}{|X_i|}\frac{1}{(\rho-2)(\rho-1)}\sum_{x\in X_i}\{(\frac{e_x^T q}{\frac{1}{|X_i|}\bar{q}_i})^{2-\rho}-1\}. \end{aligned}$$

Differentiating,

$$\begin{aligned} \frac{\partial H_N(q;\rho)}{\partial q_{x'}} = &-\sum_{i\in\mathscr{I}:x'\in X_i} c_i|X_i|^{1-\rho}\frac{1}{\rho-1}\bar{q}_i^{\rho-1}(e_{x'}^T q)^{1-\rho} \\ &+\sum_{i\in\mathscr{I}:x'\in X_i} c_i|X_i|^{1-\rho}\frac{1}{\rho-2}\bar{q}_i^{\rho-2}\sum_{x''\in X_i}(e_{x''}^T q)^{2-\rho} \\ &-\sum_{i\in\mathscr{I}:x'\in X_i} c_i\frac{1}{|X_i|}\frac{1}{(\rho-2)(\rho-1)}. \end{aligned}$$

Differentiating again,

$$\begin{aligned}\frac{\partial^2 H_N(q;\rho)}{\partial q_{x'}\partial q_{x''}} &= \delta_{x',x''}\sum_{i\in\mathscr{I}:x'\in X_i} c_i|X_i|^{1-\rho}\bar{q}_i^{\rho-1}q_{x'}^{-\rho} \\ &- \sum_{i\in\mathscr{I}:x',x''\in X_i} c_i|X_i|^{1-\rho}\bar{q}_i^{\rho-2}q_{x'}^{1-\rho} \\ &- \sum_{i\in\mathscr{I}:x',x''\in X_i} c_i|X_i|^{1-\rho}\bar{q}_i^{\rho-2}q_{x''}^{1-\rho} \\ &+ \sum_{i\in\mathscr{I}:x',x''\in X_i} c_i|X_i|^{1-\rho}\bar{q}_i^{\rho-3}\sum_{x'''\in X_i} q_{x'''}^{2-\rho}.\end{aligned}$$

Thus,

$$\begin{aligned}q_{x'}(\frac{\partial^2 H_N(q;\rho)}{\partial q_{x'}\partial q_{x''}})q_{x''} = &\sum_{i\in\mathscr{I}:x',x''\in X_i} c_i|X_i|^{1-\rho}\bar{q}_i\{\delta_{x',x''}(\frac{q_{x'}}{\bar{q}_i})^{2-\rho} \\ &-(\frac{q_{x'}}{\bar{q}_i})^{2-\rho}(\frac{q_{x''}}{\bar{q}_i})-(\frac{q_{x''}}{\bar{q}_i})^{2-\rho}(\frac{q_{x'}}{\bar{q}_i})+(\frac{q_{x''}}{\bar{q}_i})(\frac{q_{x'}}{\bar{q}_i})(\sum_{x'''\in X_i}(\frac{q_{x'''}}{\bar{q}_i})^{2-\rho})\}.\end{aligned}$$

Note that this equation also holds in the $\rho = 2$ and $\rho = 1$ cases. We can write this as

$$q_{x'}(\frac{\partial^2 H_N(q;\rho)}{\partial q_{x'}\partial q_{x''}})q_{x''} = \sum_{i\in\mathscr{I}} c_i|X_i|^{1-\rho}\bar{q}_i e_{x'}^T E_i^T m(q_i)E_i e_{x''},$$

where

$$\begin{aligned}m(q_i) &= Diag(q_i)^{2-\rho} - Diag(q_i)^{2-\rho}\imath q_i^T - q_i\imath^T Diag(q_i)^{2-\rho} + q_i\imath^T Diag(q_i)^{2-\rho}\imath q_i^T \\ &= (I-\imath q_i^T)^T Diag(q_i)^{2-\rho}(I-\imath q_i^T).\end{aligned}$$

The result immediately follows in the $\rho = 2$ case. For any $\rho \neq 2$,

$$\begin{aligned}m(q_i)^{\frac{1}{2-\rho}} &= (I-q_i\imath^T)Diag(q_i)(I-\imath q_i^T) \\ &= Diag(q) - q_iq_i^T - q_iq_i^T + q_iq_i^T \\ &= g^+(q_i).\end{aligned}$$

If $\rho < 2$, $H_N(q;\rho)$ is a bounded convex function on the relative interior of the simplex, and hence by theorem 10.3 of Rockafellar (1970) there is a unique extension to the simplex.

B.4 Proof of Lemma 2

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,

$$\sum_{i \in \mathscr{I}} c_i |X_i|^{1-\rho} \bar{q}_i^{\rho-1} \sum_{x \in X_i} (e_x^T q)^{2-\rho} D_\rho(pe_x || pE_i^T q_i) = -H_N(q) + \sum_{s \in S} (e_s^T pq) H_N(e_s^T pDiag(q)).$$

Using Lemma 1,

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

Using Bayes' rule, $\pi_s \bar{q}_{i,s} = \bar{q}_i \bar{p}_{i,s}$, where $\bar{p}_{i,s} = pE_i^T q_i$, and therefore

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

Therefore,

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

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

B.5 Proof of Proposition 3

Here we solve the multi-variate problem in the calculus of variations stated in Section 4.1,

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

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 furthermore 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

$$\mathscr{L}(a,f,g;x) \;=\; F(a,f,g;x) + \varphi(x)f + \psi_a(x)f. \tag{20}$$

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

$$\int_A p_a(x)da \;=\; 1 \tag{21}$$

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

$$\int_X q(x)\mathscr{L}(a, p_a(x), \nabla_x p_a(x); x)\, dx, \tag{22}$$

and (ii) condition (21) 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 (22), 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 (20), 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

$$\begin{aligned} \psi_a(x) &= 0, \\ \nabla_x v_a(x) \ &= \lambda[a - \gamma^T \mu_0 - \sigma^{-2} \lambda^T (x - \mu_0)], \end{aligned} \tag{23}$$

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

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

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

$$\lambda = (\frac{\theta}{4}\Sigma_0^{-1}+\sigma^{-2}I)^{-1}\gamma, \tag{24}$$

and

$$|(\frac{\theta}{4}\Sigma_0^{-1}+\sigma^{-2}I)^{-1}\gamma|^2 = \frac{4}{\theta}, \tag{25}$$

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

$$\lambda = \frac{4}{\theta}(\Sigma_0^{-1}+\sigma^{-2}\lambda\lambda^T)\gamma. \tag{26}$$

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 (24) and (25) 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 $z^T x$ with $z^T \Sigma_0 \lambda = 0$ is unchanged. An orthogonal basis of these z vectors and λ form an orthogonal basis, and let

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

observing that

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

The posterior variance-covariance matrix is

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

which simplifies to

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

and therefore by the Sherman-Morrison lemma,

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

The posterior mean of $\gamma^T x$ (and hence optimal action $a(s)$) is

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

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 (23).

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 (26)

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

concluding the proof.

B.6 Proof of Corollary 2

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 3 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 (15)). 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

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

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

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

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 3, the optimal solution to this problem is

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

B.7 Proof of Corollary 3

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 3, 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 Cramï¿œer-Rao bound for a biased estimator[19] 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 $p_a(x)$.

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

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 3. *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, C.6. □

By this lemma,

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

Therefore,

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

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

$$V\;\geq\;\int_X q(x)\,[2\gamma^T\Sigma_0\nabla_x\bar{a}(x)+\,(\bar{a}(x)-\gamma^T x)^2]\,dx. \tag{27}$$

We wish to obtain a lower bound for the integral on the right-hand side of (27). 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 (27), we obtain the tighter lower bound

$$V \;\geq\; \int_X q(x)(\gamma^T(x - \mu_0))^2 dx \;=\; \gamma^T \Sigma_0 \gamma. \tag{28}$$

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 (28) is equal to the bound in (27) 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 $\theta \geq 4|\Sigma\gamma|^2$.