The Cost of Information*

Luciano Pomatto† Philipp Strack‡ Omer Tamuz§

December 13, 2018

Abstract

We develop an axiomatic theory of costly information acquisition. Our axioms capture the idea of constant marginal costs in information production: the cost of generating two independent signals is the sum of their costs, and the cost of generating a signal with probability half equals half the cost of generating it deterministically. Together with a monotonicity and a continuity conditions, these axioms completely determine the cost of a signal up to a vector of parameters, one for each pair of states of nature. These parameters have a clear economic interpretation and determine the difficulty of distinguishing between different states. The resulting cost function, which we call *log-likelihood ratio cost*, is a linear combinations of the Kullback-Leibler divergences (i.e., the expected log-likelihood ratios) between the conditional signal distributions. We argue that this cost function is a versatile modeling tool, and that in various examples of information acquisition it leads to more realistic predictions than the approach based on Shannon entropy.

1 Introduction

> "The choice of information structures must be subject to some limits, otherwise, of course, each agent would simply observe the entire state of the world. There are costs of information, and it is an important and incompletely explored part of decision theory in general to formulate reasonable cost functions for information structures." – Arrow (1985).

Much of contemporary economic theory is built on the idea that information is scarce and valuable. A proper understanding of information as an economic commodity requires theories for its value, as well as for its production cost. While the literature on the value

*We thank Kim Border, Ben Brooks, Simone Cerreia-Vioglio, Tommaso Denti, Federico Echenique, Drew Fudenberg, Ed Green, Massimo Marinacci, Filip Matějka, Stephen Morris, and Doron Ravid for their comments. All errors and omissions are our own.

†Caltech.

‡UC Berkeley.

§Caltech. Omer Tamuz was supported by a grant from the Simons Foundation (#419427).

of information (Bohnenblust, Shapley, and Sherman, 1949; Blackwell, 1951) is by now well established, modeling the cost of information has remained an open problem. In this paper, we develop an axiomatic theory of costly information acquisition.

We characterize all cost functions over signals (i.e., Blackwell experiments or information structures) that satisfy three main axioms: First, signals that are more informative in the sense of Blackwell (1951) are more costly. Second, the cost of generating independent signals equals the sum of their individual costs. Third, the cost of generating a signal with probability half equals half the cost of generating it deterministically.

As an example, the second axiom implies that the cost of collecting n independent random samples from a population is linear in n. The third axiom implies that the cost of an experiment that produces a sample with probability α is a fraction α of the cost of acquiring the same sample with probability 1.

Interpretation of the Axioms. Our three axioms admit a straightforward economic interpretation. The first one is a simple form of monotonicity: more precise information is more costly. The second and third axioms aim to capture the idea of constant marginal costs. In the study of traditional commodities, a standard avenue for studying costs functions is by categorizing them in terms of decreasing, increasing, or constant marginal costs. The case of linear cost is, arguably, the conceptually simplest one.[1]

With this motivation in mind, the second axiom states that the cost of generating a signal is the same regardless of which additional independent signals a decision maker decides to acquire. Consider, as an example, a company surveying customers by calling them to learn about the demand for a new product. Our axiom implies that the cost of calling an additional customer is constant, i.e. calling 100 customers is 10 times more costly than calling 10. Whether this assumption is a reasonable approximation depends on the application at hand: for instance, it depends on whether or not large fixed costs are a crucial ingredient of the model under consideration.

The third axiom posits constant marginal costs with respect to the probability that an experiment is successful. To formalize this idea we study experiments that succeed with probability α, and produce no information with probability $1-\alpha$. The axiom states that for such experiments the cost is linear in α, so that the marginal cost of success is constant.

We propose the constant marginal cost assumption as a natural starting point for thinking about the cost of information acquisition. It has the advantage that it admits a clear economic motivation, making it easy to judge for which applications it is appropriate.

[1] A difficulty with modeling the production of information is that bundles of this commodity cannot be easily modeled as collections of small (or infinitesimal) identical units. Hence the traditional definition of "marginal costs" is not readily applicable. We therefore take a different approach.

Representation. The main result of this paper is a characterization theorem for cost functions over experiments. We are given a finite set Θ of states of nature. An experiment μ produces a signal realization s with probability $\mu_i(s)$ in state $i \in \Theta$. We show that for any cost function C that satisfies the above postulates, together with a continuity assumption, there exist non-negative coefficients β_{ij}, one for each ordered pair of states of nature i and j, such that[2]

$$C(\mu) = \sum_{i,j \in \Theta} \beta_{ij} \left(\sum_{s \in S} \mu_i(s) \log \frac{\mu_i(s)}{\mu_j(s)} \right). \tag{1}$$

The coefficients β_{ij} can be interpreted as capturing the difficulty of discriminating between state i and state j. To see this, note that the cost can be expressed as a linear combination

$$C(\mu) = \sum_{i,j \in \Theta} \beta_{ij} D_{\mathrm{KL}}(\mu_i \| \mu_j),$$

where the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(\mu_i \| \mu_j) = \sum_{s \in S} \mu_i(s) \log \frac{\mu_i(s)}{\mu_j(s)}$$

is the expected log-likelihood ratio between state i and state j when the state equals i. $D_{\mathrm{KL}}(\mu_i \| \mu_j)$ is thus large if the experiment on average produces evidence that strongly favors i over j, conditional on the state being i. Hence, the greater β_{ij} the more costly it is to reject the hypothesis that the state is j when it truly is i. Formally, β_{ij} is the marginal cost of increasing the expected log-likelihood ratio of an experiment with respect to states i and j, conditional on i being the true state. We refer to the cost (1) as the *log-likelihood ratio cost* (or *LLR cost*).

In many common information acquisition problems, states of the world are one dimensional. This is the case when, for instance, the unknown state is a physical quantity to be measured, or the future level of interest rates. In these examples, a signal can be seen as a noisy measurement of the unknown underlying state $i \in \mathbb{R}$. We provide a framework for choosing the coefficients β_{ij} in these contexts. Our main hypotheses are that the difficulty of distinguish between two states i and j is a function of the distance between them, and that the cost of performing a measurement with standard Gaussian noise does not depend on the particular features of the information acquisition problem.

Under these assumptions (Axioms a and b) Proposition 2 shows that there exists a

[2]Throughout the paper we assume that the set of states of nature Θ is finite. We do not assume a finite set S of signal realizations and the generalization of (1) to infinitely many signal realizations is given in (3).

constant κ such that, for every pair of states $i, j \in \Theta$,

$$\beta_{ij} = \frac{\kappa}{(i-j)^2}.$$

Thus, states that are closer are more difficult to distinguish. As we show in the paper, this choice of parameters offers a simple framework for analyzing the implications of the LLR cost.

The concept of a Blackwell experiment makes no direct reference to subjective probabilities nor to Bayesian reasoning.[3] Likewise, our axioms and characterization theorem do not presuppose the existence of a prior over the states of nature. Nevertheless, given a prior q over Θ, an experiment induces a distribution over posteriors p, making p a random variable. Under this formulation, the cost (1) of an experiment can be represented as the expected change of the function

$$H(p,q) = \sum_{i,j} \beta_{ij} \frac{p_i}{q_i} \log\left(\frac{p_i}{p_j}\right)$$

from the prior q to the posterior p induced by the signal.[4] That is, the cost of an experiment equals

$$\mathbb{E}\left[H(p,q) - H(q,q)\right].$$

This alternative formulation makes it possible to apply techniques for posterior-separable costs functions (Caplin and Dean, 2013; Caplin, Dean, and Leahy, 2016).

Relation to Entropy Cost. Following Sims' seminal work on rational inattention, cost functions based on Shannon entropy have been commonly applied to model costly information processing (Sims, 2003, 2010).[5] Entropy costs are defined as the expected change

$$\mathbb{E}\left[F(q) - F(p)\right]$$

of the Shannon entropy $F(p) = -\sum_{i\in\Theta} p_i \log p_i$ between the decision maker's prior belief q and posterior p. Equivalently, in this formulation, the cost of an experiment is given by the mutual information between the state of nature and the signal.

Compared to Sims' work—and the literature in rational inattention—our work aims at modeling a different kind of phenomenon. While Sims' goal is to model the cost of *processing* information our goal is to model the cost of *acquiring* information. Due to this

[3] Blackwell experiments have been studied both within and outside the Bayesian framework. See, for instance, Le Cam (1996) for a review of the literature on Blackwell experiments.

[4] By Bayes' rule the posterior belief p associated with the signal realization s is given by $p_i = \frac{q_i \mu_i(s)}{\sum_j q_j \mu_j(s)}$.

[5] Caplin (2016) and Mackowiak, Matějka, and Wiederholt (2018) review the literature on rational inattention.

difference in motivation, Sims' axioms postulate that signals which are harder to encode are more costly, while we assume that signals which are harder to generate are more costly. As an illustrative example of this difference consider a newspaper. Rational inattention theory models the readers' effort of processing the information contained in the newspaper. In contrast, our goal is to model the cost that the newspaper incurs in generating such information.

Given the different motivation, it is perhaps not surprising that the LLR cost leads to predictions which are profoundly different from those induced by entropy cost. We illustrate the differences by four stylized examples in §5.

2 Model

A decision maker acquires information on an unknown state of nature belonging to a finite set Θ. Elements of Θ will be denoted by i, j, k etc. Following Blackwell (1951), we model the information acquisition process by means of *signals*, or *experiments*. An experiment $\mu = (S, (\mu_i))$ consists of a set S of signal realizations equipped with a sigma-algebra Σ, and a vector of probability measures $(\mu_i)_{i\in\Theta}$ defined on a (S, Σ). Throughout we assume that the measures μ_i are mutually absolutely continuous, so that the derivative (i.e. ratio between densities) $\frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s)$ is finite almost everywhere. In the case of finite signal realizations these derivatives are simply equal to ratio between probabilities $\frac{\mu_i(s)}{\mu_j(s)}$, as in (1). Economically, this assumption means that no signal can ever rule out any state, and in particular can never completely reveal the true state.

From a decision-theoretic perspective, an experiment can be summarized by its conditional distributions of log-likelihood ratios. Given an experiment μ a realization s and pair of states i, j, we denote by

$$\ell_{ij}(s) = \log \frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s)$$

the log-likelihood ratio between states i and j upon observing the realization s. We define the vector

$$L(s) = (\ell_{ij}(s))_{i,j}$$

of log-likelihood ratios among all pairs of states, conditional on a signal realization s. The distribution of L will of course depend on what is the true state generating the data. Given an experiment μ, we denote by $\bar{\mu}_i$ the distribution of L conditional on state i.[6]

We restrict our attention to signals where the induced log-likelihoods ratios (ℓ_{ij}) have finite moments, that is for every state i and every integral vector $\alpha \in \mathbb{N}^\Theta$ the expectation $\int_S \left|\prod_{k\neq i} \ell_{ik}^{\alpha_k}\right| d\mu_i$ is finite. We denote by $\mathcal{E}$ the class of all such experiments.[7]

[6] The measure $\bar{\mu}_i$ is defined as $\bar{\mu}_i(A) = \mu_i(\{s : L(s) \in A\})$ for every measurable $A \subseteq \mathbb{R}^{\Theta\times\Theta}$.

[7] We refer to $\mathcal{E}$ as a class, rather than a set, since Blackwell experiments do not form a well-defined set. In doing so, we follow a standard convention in set theory (see, for instance, Jech, 2013).

The cost of producing information is described by an *information cost function*

$$C : \mathcal{E} \to \mathbb{R}_+$$

assigning to each experiment $\mu \in \mathcal{E}$ its cost $C(\mu)$. As in the literature on rational inattention, our model imposes virtually no restrictions over the set of feasible experiments. In particular, we do not restrict our attention to a parametric family of experiments such as the normally distributed signals.

2.1 Axioms

We introduce and characterize four basic properties for information cost functions. Our first axiom postulates that the cost of any experiment should depend only on its informational content. For instance, it should not be sensitive to the way signal realizations are labelled. In making this idea formal we follow Blackwell (1951, Section 4).

Let $q \in \mathcal{P}(\Theta)$ be the uniform prior assigning equal probability to each element of Θ.[8] Let μ and ν be two experiments, inducing the distributions over posteriors π_μ and π_ν. Then μ dominates ν in the Blacwell order if

$$\int_{\mathcal{P}(\Theta)} f(p)\,\mathrm{d}\pi_\mu(p) \geq \int_{\mathcal{P}(\Theta)} f(p)\,\mathrm{d}\pi_\nu(p)$$

for every convex function $f : \mathcal{P}(\Theta) \to \mathbb{R}$.

As is well-known, dominance with respect to the Blackwell order is equivalent to the requirement that in any decision problem, a Bayesian decision maker achieves a higher expected utility when basing her action on μ rather than ν. We say that two experiments are *Blackwell equivalent* if they dominate each other. It is a standard result that two experiments μ and ν are Blackwell equivalent if and only if for every every state i they induce the same distribution $\bar{\mu}_i = \bar{\nu}_i$ of log-likelihood ratios.

As discussed in the introduction, it is natural to require the cost of information to be increasing in the Blackwell order. For our main result, it is sufficient to require that any two experiments that are Blackwell equivalent lead to the same cost. Nevertheless, the cost function axiomatized in this paper will satisfy the stronger property of Blackwell monotonicity.

Axiom 1. *If μ and ν are Blackwell equivalent, then $C(\nu) = C(\mu)$.*

The lower envelope of a cost function assigns to each μ the minimum cost of producing an experiment that is Blackwell equivalent to μ. When experiments are optimally chosen

[8] Throughout the paper, $\mathcal{P}(\Theta)$ denotes the set of probability measures on Θ identified with their representation in $\mathbb{R}^\Theta$, so that for every $q \in \mathcal{P}(\Theta)$, q_i is the probability of the state i.

by a decision maker we can, without loss of generality, identify a cost function with its lower envelope. This results in a cost function for which Axiom 1 is automatically satisfied.

For the next axiom, we study the cost of performing multiple independent experiments. Given $\mu = (S, (\mu_i))$ and $\nu = (T, (\nu_i))$ we define the signal

$$\mu \otimes \nu = (S \times T, (\mu_i \times \nu_i))$$

where $\mu_i \times \nu_i$ denotes the independent product of the two measures.[9] Under the experiment $\mu \times \nu$, the realizations of both experiments μ and ν are observed, and the two observations are independent conditional on the state. To illustrate, suppose μ and ν consist of drawing a random sample from two possible populations. Then $\mu \otimes \nu$ is the experiment where two samples, one for each population, are collected.

Our second axiom states that the cost function is additive for independent experiments:

Axiom 2. *The cost of performing two independent experiments is the sum of their costs:*

$$C(\mu \otimes \nu) = C(\mu) + C(\nu) \text{ for all } \mu \text{ and } \nu.$$

In many settings an experiment can, with non-negligible probability, fail to produce new evidence. The next axiom states that the cost of an experiment is linear in the probability that the experiment will generate information. Given μ, we define a new experiment, which we call a *dilution* of μ and denote by $\alpha \cdot \mu$. In this new experiment, with probability α the signal μ is produced, and with probability $1 - \alpha$ a completely uninformative signal is observed. Formally, given $\mu = (S, (\mu_i))$, fix a new signal realization $o \notin S$ and $\alpha \in [0, 1]$. Then $\alpha \cdot \mu$ is defined as $\alpha \cdot \mu = (S \cup \{o\}, (\nu_i))$, where $\nu_i(E) = \alpha\mu_i(E)$ for every measurable $E \subseteq S$, and $\nu_i(\{o\}) = 1 - \alpha$. The next axiom specifies the cost of such an experiment:

Axiom 3. *The cost of a dilution $\alpha \cdot \mu$ is linear in the probability α:*

$$C(\alpha \cdot \mu) = \alpha\, C(\mu) \text{ for every } \mu \text{ and } \alpha \in [0, 1]\,.$$

An immediate implication of Axioms 3 and 1 is that a completely uninformative signal has zero cost.

Our final assumption is a continuity condition. We first introduce a distance over $\mathcal{E}$. Recall that for every experiment μ, $\bar{\mu}_i$ denotes its distribution of log-likelihood ratios conditional on state i. We denote by d_{tv} the total-variation distance.[10] Given a vector $\alpha \in \mathbb{N}^\Theta$, let $M_i^\mu(\alpha) = \int_S \left| \prod_{k \neq i} \ell_{ik}^{\alpha_k} \right| d\mu_i$ be the α-moment of the corresponding vector of

[9] When the set of signal realizations is finite, the measure $\mu_i \times \nu_i$ assigns to each realization (s, t) the probability $\mu_i(s)\nu_i(t)$.

[10] That is, $d_{tv}(\bar{\mu}_i, \bar{\nu}_i) = \sup |\bar{\mu}_i(A) - \bar{\nu}_i(A)|$, where the supremum is over all measurable subsets of $\mathbb{R}^{\Theta \times \Theta}$.

likelihood-ratios. Given an upper bound $N \geq 1$, we define the distance:

$$d_N(\mu, \nu) = \max_{i \in \Theta} d_{tv}\left(\bar{\mu}_i, \bar{\nu}_i\right) + \max_{i \in \Theta} \max_{\alpha \in \{0,\ldots,N\}^n} \left|M_i^{\mu}(\alpha) - M_i^{\nu}(\alpha)\right|.$$

According to the metric d_N, two signals μ and ν are close if, for each state i, the induced distributions of log-likelihood ratios are close in total-variation and, in addition, have similar moments (for any moment α lower or equal to $(N, \ldots, N)$).

Axiom 4. *For some $N \geq 1$ the function C is uniformly continuous with respect to d_N.*

As is well known, convergence with respect to the total-variation distance is a demanding requirement, as compared to other topologies such as the weak topology. So, continuity with respect to d_{tv} is a relatively weak assumption. Continuity with respect to the stronger metric d_N is, therefore, an even weaker assumption.[11]

2.2 Discussion

Additivity assumptions in the spirit of Axiom 2 have appeared in multiple parametric models of information acquisition. A common assumption in Wald's classic model of sequential sampling and its variations (Wald, 1945; Arrow, Blackwell, and Girshick, 1949), is that the cost of acquiring n independent samples from a population is linear in n.[12] Likewise, in models where information is acquired by means of normally distributed experiments, a standard specification is that the cost of an experiment is inversely proportional to its variance (see, e.g. Wilson, 1975; Van Nieuwerburgh and Veldkamp, 2010). This amounts to an additivity assumption, since the product of two independent normal signals is Blackwell equivalent to a normal signal whose precision (that is, the inverse of its variance) is equal to the sum of the precisions of the two original signals.

Underlying these different models is the notion that the cost of an additional independent experiment is constant. Axiom 2 captures this idea in a non-parametric context, where no a priori restrictions are imposed over the domain of feasible experiments. As discussed in the introduction, we focus on linear cost structures as we view those as a natural starting point to reason about the cost of information, in the same way the assumption of constant marginal cost is a benchmark for the analysis of traditional commodities. Whether this assumption fits a particular application well is inevitably an empirical question.

Axiom 3 expresses the idea that the marginal cost of increasing the probability of success of an experiment is constant. The axiom admits an additional interpretation.

[11] We discuss this topology in detail in §A. Any information cost function that is continuous with respect to the metric d_N satisfies Axiom 1. For expositional simplicity, we maintain the two axioms as separate throughout the paper.

[12] A similar condition appears in the continuous-time formulation of the sequential sampling problem, where the information structure consists of observing a signal with Brownian noise over a time period of length t, under a cost that is linear in t (Dvoretzky, Kiefer, Wolfowitz, et al., 1953; Chan, Lizzeri, Suen, and Yariv, 2017; Morris and Strack, 2018).

Consider an extended framework where a decision maker can randomize her choice of experiment. Then, the property

$$C(\alpha \cdot \mu) \leq \alpha C(\mu) \tag{2}$$

ensures that the cost of the diluted experiment $\alpha \cdot \mu$ is not greater than the expected cost of performing μ with probability α and collecting no information with probability $1-\alpha$. Hence, if (2) was violated, the experiment $\alpha \cdot \mu$ could be replicated at a strictly lower cost through a simple randomization by the decision maker.

Assume Axiom 2 holds. Then, the converse inequality

$$C(\alpha \cdot \mu) \geq \alpha C(\mu)$$

ensures that the cost $C(\mu)$ of an experiment is not greater than the expected cost $(1/\alpha)C(\alpha \cdot \mu)$ of performing repeated independent copies of the diluted experiment $\alpha \cdot \mu$ until it succeeds.[13] Axiom 3 is thus automatically satisfied once one allows for dynamic and mixed strategies of information acquisition.

3 Representation

Theorem 1. *An information cost function C satisfies Axioms 1-4 if and only if there exists a collection $(\beta_{ij})_{i,j\in\Theta}$ in $\mathbb{R}_+$ such that for every experiment $\mu = (S, (\mu_i))$,*

$$C(\mu) = \sum_{i,j\in\Theta} \beta_{ij} \int_S \log \frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s)\, \mathrm{d}\mu_i(s). \tag{3}$$

Moreover, the collection $(\beta_{ij})_{i\neq j}$ is unique given C.

We refer to a cost function that satisfies Axioms 1-4 as a *log-likelihood ratio (LLR) cost*. As shown by the theorem, this class of information cost functions is uniquely determined up to the parameters (β_{ij}). A higher value of $\int_S \log(\mathrm{d}\mu_i/\mathrm{d}\mu_j)\mathrm{d}\mu_i$ describes an experiment where, conditional on state i, there is a higher probability of observing evidence in favor of state i compared to j, as represented by a higher expected value of the likelihood ratio $\mathrm{d}\mu_i/\mathrm{d}\mu_j$.

Given an experiment μ, the coefficient β_{ij} measures the cost of moving to an experiment ν where $\int_S \log(\mathrm{d}\mu_i/\mathrm{d}\mu_j)\mathrm{d}\mu_i$—the expected log-likelihood ratio between states i and j conditional on i—is increased by 1 while, all other expectations remain fixed. As we formally show in Lemma 2 this operation is always well-defined: for every strictly positive vector $(z_{ij})_{i\neq j}$ there exists an experiment $\nu \in \Theta$ such that $(z_{ij})_{i\neq j}$ equals the vector of

[13] Implicit in this interpretation is the assumption, common in the literature on rational inattention, that the decision maker's cost of an experiment is expressed in the same unit as her payoffs.

expected log-likelihood ratios induced by ν. Hence, each parameter β_{ij} is the *marginal cost* of increasing the expected log-likelihood ratio of the experiment with respect to states i and j and conditional on i.

The expression $\int_S \log(\mathrm{d}\mu_i/\mathrm{d}\mu_j)\mathrm{d}\mu_i$ is the Kullback-Leibler divergence between the two distributions $D_{\mathrm{KL}}(\mu_i\|\mu_j)$, a well understood and tractable measure of informational content. This implies that (3) can alternatively be formulated as

$$C(\mu) = \sum_{i,j\in\Theta} \beta_{ij} D_{\mathrm{KL}}(\mu_i\|\mu_j).$$

Closed form solutions for the Kullback-Leibler divergence between standard distributions, such as Normal, exponential or binomial, are readily available. This makes it immediate to compute the cost $C(\mu)$ of common parametric families of experiments.

Example: Normal Signals. Consider a Normal experiment $\mu^{m,\sigma}$ where the signal s is Normally distributed

$$s = m_i + \varepsilon$$

with mean $m_i \in \mathbb{R}$ depending on the state, and a state independent normally distributed ε with standard deviation σ. Then, by substituting (3) with the expression for the divergence between normal distributions, we obtain that the cost of such an experiment is given by

$$C(\mu^{m,\sigma}) = \sum_{i,j\in\Theta} \beta_{ij} \frac{(m_j - m_i)^2}{2\sigma^2}\,. \tag{4}$$

The cost is decreasing in the variance σ^2, as one may expect. Increasing β_{ij} increases the cost of a signal $\mu^{m,\sigma}$ by a factor that is proportional to the distance in mean between the the two states, so by the degree by which the signal is effective at distinguishing between the two states.

Example: Binary Signals. Another canonical example is the binary-binary setting in which the set of states is $\Theta = \{H, L\}$, and the signal $\nu^p = (S, (\nu_i))$ is also binary: $S = \{0, 1\}$, $\nu_H = B(p)$ and $\nu_L = B(1-p)$ for some $p > 1/2$, where $B(p)$ is the Bernoulli distribution on $\{0, 1\}$ assigning probability p to 1. In this case

$$C(\nu^p) = (\beta_{HL} + \beta_{LH}) \left[p \log \frac{p}{1-p} + (1-p) \log \frac{1-p}{p} \right]. \tag{5}$$

Hence the cost is monotone in (β_{ij}) and p.

Bayesian Representation. The framework we considered so far makes no references to subjective beliefs over the states of nature. We now show how the LLR cost function

can be embedded in a standard Bayesian framework. Consider a decision maker endowed with a prior $q \in \mathcal{P}(\Theta)$. An experiment μ induces then a distribution over posteriors π_μ. As shown by the next result, the cost of an experiment $C(\mu)$ can be reformulated in terms of the distribution π_μ :

Proposition 1. *Let C admit the representation (3) and fix a prior $q \in \mathcal{P}(\Theta)$ with full support. For every experiment μ inducing a distribution over posterior π_μ,*

$$C(\mu) = \int_{\mathcal{P}(\Theta)} F(p) - F(q)\, \mathrm{d}\pi_\mu(p) \quad \textit{where} \quad F(p) = \sum_{i,j \in \Theta} \gamma_{ij}\, p_i \log \frac{p_i}{p_j} \tag{6}$$

and $\gamma_{ij} = \beta_{ij}/q_i$ for every i, j.

Hence, in this Bayesian representation, the cost of the experiment μ can be expressed as the expected change of the function F from the prior q to the realized posterior p. In the representation each coefficient β_{ij} is normalized by the prior probability of the state q_i.[14]

Representations of the form (6) have been studied in the literature under the name of "posterior separable." An important implication of Theorem 1 is that general techniques for posterior separable costs functions, as developed by Caplin and Dean (2013), can be applied to the cost function in this paper.

4 One-Dimensional Information Acquisition Problems

Up to now we have been intentionally silent on how to specify the coefficients (β_{ij}). Each parameter β_{ij} captures how costly it is to distinguish between particular states, and thus will be highly context dependent.

A commonly encountered context is that of learning about one-dimensional characteristics where each state i is a real number.[15] In macroeconomic applications, the state may represent the future level of interest rates. In perceptual experiments in neuroscience and economics the state can correspond to the number of red/blue dots on the screen (see §5.1 below). More generally, i might represent a physical quantity to be measured.

In this section we propose a natural choice of parameters (β_{ij}) for one-dimensional information acquisition problems that extends a commonly made assumption in the literature. Given a problem where each state $i \in \Theta$ is a real number, we propose to set

[14] This may suggests that under the cost function (6) the cost of an experiment is independent of the prior, in the sense that two agents performing the same experiment but hold different beliefs will face the same expected cost. While this interpretation is consistent with the representation (6), a formal separation between the coefficients (β_{ij}) and the subjective beliefs of the decision maker generating the experiments would require a richer framework.

[15] We opt, in this section, to deviate from standard practice and use the letters i, j to refer to real numbers, in order to maintain consistency with the rest of the paper.

each coefficient β_{ij} to be equal to $\frac{\kappa}{(i-j)^2}$ for some constant $\kappa \geq 0$. So, each β_{ij} is inversely proportional to the squared distance between the corresponding states i and j. Intuitively, under this specification, two states that are closer to each other are harder to distinguish. As we show in §5, this choice of specification provides a simple framework for applying the LLR cost.

The main result of this section shows that this choice of parameters captures two main hypotheses: That the difficulty of producing a signal that allows to distinguish between state i and j is a function only of the distance $|i-j|$ between the two states, and that the cost of observing a signal that equals the state plus Gaussian noise is the same across problems. Both assumptions capture the idea that the cost of making a measurement depends only on the precision of the measurement, and not on the other details of the model (the set of states, the prior etc). For example, the cost of measuring a person's height depends only on the precision of the measurement instrument and the number of independent measurements, but not on what modeling assumption are made about the set of possible heights.

Denote by $\mathcal{T}$ the collection of finite subsets of $\mathbb{R}$ with at least two elements. Each set $\Theta \in \mathcal{T}$ is the set of states of nature in a different information acquisition problem. To simplify the language, we refer to each Θ as a *problem*.

For each Θ we are given a LLR cost function C^Θ with coefficients (β_{ij}^Θ). The next two axioms formalize the two hypotheses described above by imposing restrictions on the cost of information across problems. In particular, the marginal cost of increasing the expected LLR between two states $i, j \in \Theta$ is a function of the distance between the two and is independent of the identity of the other states.

Axiom a. *For all $\Theta, \Theta' \in \mathcal{T}$ such that $|\Theta| = |\Theta'|$, and for all $i, j \in \Theta$ and $k, l \in \Theta'$,*

$$\text{if } \ |i-j| = |k-l| \ \text{ then } \ \beta_{ij}^\Theta = \beta_{kl}^{\Theta'}.$$

Recall that β_{ij}^Θ measures the cost of increasing the expected log-likelihood ratio $D_{\mathrm{KL}}(\mu_i \| \mu_j)$ of an experiment μ, keeping all other expected log-likelihood ratios fixed. Hence Axiom a states that this marginal cost depends only on the distance between states.

For each $i \in \mathbb{R}$ denote by ν_i a Normal probability measure on the real line with mean i and variance 1. Given a problem Θ, we denote by ν^Θ the experiment $(\mathbb{R}, (\nu_i)_{i \in \Theta})$. Hence, ν^Θ is the canonical experiment consisting of a noisy measurement with standard Gaussian error. The next axiom states that the cost of such a measurement does not depend on the particular values that the state can take.

Axiom b. *For all $\Theta, \Theta' \in \mathcal{T}$, $C^\Theta(\nu^\Theta) = C^{\Theta'}(\nu^{\Theta'})$.*

Axioms a and b lead to a simple parametrization for the coefficient of the LLR cost in one-dimensional information acquisition problems:

Proposition 2. *The collection $C^\Theta, \Theta \in \mathcal{T}$, satisfies Axioms a and b if and only if there exists a constant $\kappa > 0$ such that for all $i, j \in \Theta$ and $\Theta \in \mathcal{T}$,*

$$\beta_{ij}^\Theta = \frac{1}{n(n-1)} \frac{\kappa}{(i-j)^2}$$

where n is the cardinality of Θ.

Using the cost of a Normal signal we derived in (4) we obtain that for any $\Theta \in \mathcal{T}$ the cost satisfies $C(\nu^\Theta) = \kappa$, and that, furthermore, Normal signals with mean i and variance σ^2 have cost $\kappa\sigma^{-2}$, as is often assumed in the literature (Wilson, 1975; Van Nieuwerburgh and Veldkamp, 2010). Thus, the functional form given in Proposition 2 generalizes the assumption that the cost of information is *proportional to precision* which is commonly made in models which endogenously restrict attention to normal signals.

5 Examples

5.1 Stochastic Choice

We now study the log-likelihood ratio cost in the context of stochastic choice. We consider an agent choosing an action a from a finite set A of actions. The payoff from a depends on the state of nature $i \in \Theta$ and is given by $u(a, i)$. The agent is endowed with a prior q over the set of states.

Before making her choice, the agent can acquire a signal μ at cost $C(\mu)$. As is well known, if the cost function C is monotone with respect to the Blackwell order, then it is without loss of generality to restrict attention to signals where the set of realizations S equals the set of actions A, and to assume that upon observing a signal $s = a$ the decision maker will choose the action recommended by the signal. We can then therefore identify an experiment μ with a vector of probability measures (μ_i) in $\mathcal{P}(A)$.

An optimal signal $\mu^\star = (\mu_i^\star)$ solves

$$\mu^\star \in \operatorname*{argmax}_\mu \left[\sum_{i \in \Theta} q_i \left(\sum_{a \in A} \mu_i(a) u(a, i) \right) - C(\mu) \right]. \tag{7}$$

Hence, the probability action a is chosen in state i is given by $\mu_i^\star(a)$. The LLR cost function is monotone with respect to the Blackwell order (as shown by Proposition 6 in the Appendix). In addition, the maximization problem (7) is strictly concave, provided all coefficients (β_{ij}) are strictly positive (Proposition 7).

First-Order Conditions The next result characterizes the optimal choice probabilities under the LLR cost:

Proposition 3. *The state-dependent distribution over actions $\mu = (\mu_i)_{i\in\Theta}$ solves the optimization problem (7) if and only if*

$$q_i \left[u(i,a) - u(i,a')\right] = \sum_{j\neq i} \left\{ \beta_{ij} \log\left(\frac{\mu_i(a)/\mu_i(a')}{\mu_j(a)/\mu_j(a')}\right) - \beta_{ji}\left[\frac{\mu_j(a)}{\mu_i(a)} - \frac{\mu_j(a')}{\mu_i(a')}\right]\right\} \tag{8}$$

for every state $i \in \Theta$ and every pair of actions $a, a' \in A$.

Condition (8) has an intuitive interpretation: the left-hand-side measures the expected benefit of choosing action a instead of a' in state i. The right-hand-side measures the change in information acquisition cost necessary to pick action a marginally more often and action a' marginally less often.

An Application to Perception Tasks. Consider a perception task where subjects observe 100 dots of different colors on a screen. Each dot is either red or blue. A parameter $r \in \{1, \ldots, 50\}$ is fixed. Subjects are told the value of r and that the number of blue dots i is drawn uniformly in $\Theta = \{50 - r, \ldots, 49, 51, \ldots, 50 + r\}$. The state where the number of blue and red dots is equal to 50 is ruled out to simplify the exposition.[16]

Subjects are asked to guess whether there are more blue or red dots and get rewarded if they guess correctly. So $A = \{R, B\}$ and

$$u(a,i) = \begin{cases} 1 & \text{if } a = B \text{ and } i > 50 \\ 1 & \text{if } a = R \text{ and } i < 50 \\ 0 & \text{else} \end{cases}.$$

For optimal choices according to (7), in state i an agent guesses correctly with probability

$$m(i) = \begin{cases} \mu_i(B) & \text{if } i > 50 \\ \mu_i(R) & \text{if } i < 50 \end{cases}.$$

Intuitively, it should be harder to guess whether there are more blue or red dots when the difference in the number of dots is small, i.e. when i is close to 50. Indeed, it is a well established fact in the psychology[17], neuroscience[18], economics[19] literatures that so called *psychometric functions*—the relation between the strength of a stimulus offered to a subject and the probability that the subject identifies this stimulus—are sigmoidal (or S-shaped), so that the probability that a subject chooses B transitions smoothly from values close to 0 to values close to 1.

[16] This means that the prior is $q_i = \frac{1}{2r}$, for $i \in \Theta$.
[17] See, e.g., Chapter 7 in Green and Swets (1966) or Chapter 4 in Gescheider (1997).
[18] E.g., Krajbich et al. (2010); Tavares et al. (2017).
[19] See, e.g., Mosteller and Nogee (1951).

As Dean and Neligh (2017) note, under entropy (and a uniform prior, as in the experimental setup described above), the optimal signal μ^* must induce a probability of guessing correctly that is state-independent.[20]

As shown by Matějka and McKay (2015), Caplin and Dean (2013), and Steiner, Stewart, and Matějka (2017), conditional on a state i, the likelihood ratio $\mu_i^*(B)/\mu_i^*(R)$ between the two actions must equal the ratio $e^{u(i,B)}/e^{u(i,R)}$. Hence the probability that a subject chooses correctly must be the same for any two states that lead to the same utility function over actions, such as the state in which there are 51 blue dots and 49 red dots, and the state in which the composition is 99 blue dots and and 1 red dots.

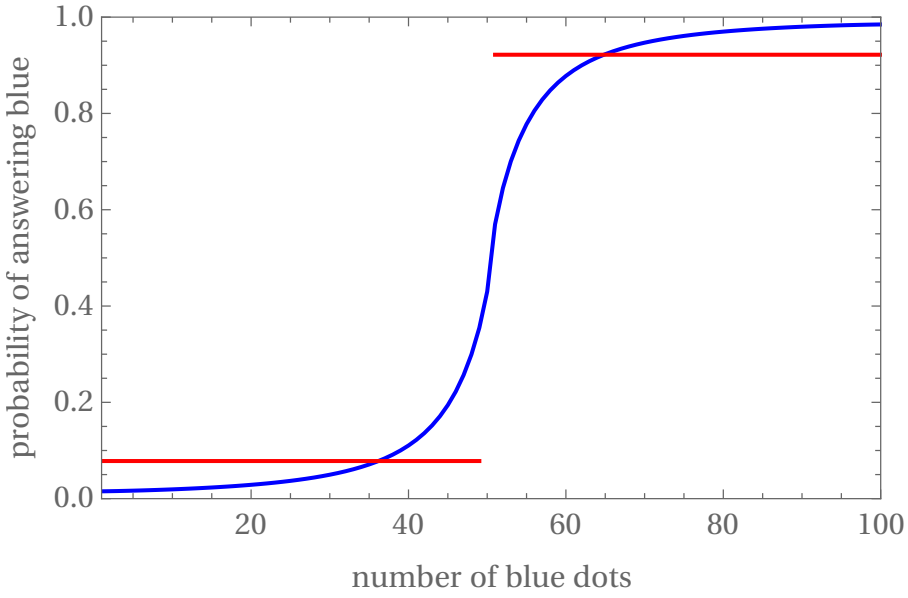

Figure 1: Predicted probability of guessing that there are more red dots as a function of the state for LLR cost with $\beta_{ij} = 1/(i-j)^2$ (in blue) and entropy cost (in red).

This unrealistic prediction is driven by the fact that in the entropy model the states are devoid of meaning and thus equally hard to distinguish. Indeed, the same conclusion holds for any cost function C in (7) that, like entropy, is invariant with respect to a permutation of the states and is convex as a function of the state-dependent action distributions (μ_i).

Our model can account for the difficulty of distinguishing different states through the coefficients β. For example, under the specification $\beta_{ij} = 1/(i-j)^2$, LLR cost predicts that the agent is less likely to chose correctly when the number of red and blue dots is close, i.e. if i is close to 50 (see Figure 1). A closely related implication is that the probability of

[20] It is well known that under entropy the physical features of the states (such as distance or similarity) do not affect the cost of information acquisition. For instance, Mackowiak, Matějka, and Wiederholt (2018) write "[..] entropy does not depend on a metric, i.e., the distance between states does not matter. With entropy, it is as difficult to distinguish the temperature of 10^oC from 20^oC, as 1^oC from 2^oC. In each case the agent needs to ask one binary question, resolve the uncertainty of one bit."

guessing R will "jump" at $\theta = 50$ for the entropy model while it will behave "smoothly" for LLR cost (see Figure 1 again).

Continuous Choice The main insight emerging from the above example is that, under LLR cost, neighboring states are harder to distinguish, hence acquiring information that can finely discriminate between them is more costly. This, in turn, implies that the choice probabilities cannot vary abruptly across nearby states.

We now extend this intuition to more general decision problems. Assume that the state space Θ is endowed with a natural distance $d : \Theta \times \Theta \to \mathbb{R}$. In the previous example, d is simply the difference $|i - j|$ in the number of blue and red dots.

We say that *nearby states are hard to distinguish* if for all $i, j \in \Theta$

$$\min\{\beta_{ij}, \beta_{ji}\} \geq \frac{1}{d(i,j)^2}\,.$$

So, the cost of acquiring information that discriminates between states i and j is high for states that are close to each other. Our next result shows that when nearby states are hard to distinguish, the optimal choice probabilities are Lipschitz continuous in the state: the agent will choose actions with similar probabilities in similar states.

Proposition 4 (Continuity of Choice)**.** *Suppose that nearby states are hard to distinguish, and let* $\|u\| = \max_{a,i} |u(a,i)|$*. Then, for every action* a*, the optimal choice probabilities* $(\mu_i^*(a))_{i\in\Theta}$ *solving* (7) *satisfy*

$$\left|\mu_i^\star(a) - \mu_j^\star(a)\right| \leq \sqrt{\|u\|}\, d(i,j) \quad \text{for all } a \in A \text{ and } i, j \in \Theta. \tag{9}$$

A crucial feature of the bound (9) is that the Lipschitz constant is given only by the maximum $\|u\|$ of the utility function, independently of the exact form of the coefficients (β_{ij}), and of the number of states.[21]

The result illustrates a stark contrast between the predictions of entropy cost and LLR cost. Entropy models predict behavior that display counter-intuitive discontinuities with respect to the true state of nature. Under the log-likelihood ratio cost, when nearby states are harder to distinguish, the change in choice probabilities across states can be bounded by the distance between states.

This difference has stark implications in coordination games. Morris and Yang (2016) show that if continuous choice holds, there is a unique equilibrium; if it fails, there are multiple equilibria. Consequently, entropy cost and LLR cost function lead to very different

[21] While we only consider finite Θ, in some applications a natural model involves a continuum of states, endowed with a distance; e.g., the unit interval. To use LLR costs for such models one can choose a fine finite subset and apply LLR costs there. Proposition 4 implies that the action probabilities will be continuous, regardless of how the continuous state space is discretized.

predictions in coordination games and their economic applications (bank-runs, currency attacks, models of regime change, etc).

5.2 Acquiring Precise Information.

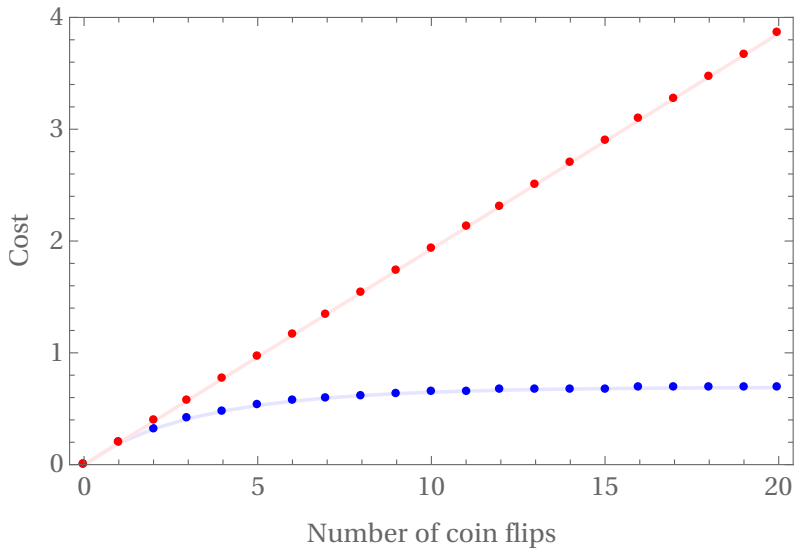

Figure 2: The LLR cost (in red) and the entropy cost (in blue) of observing multiple independent coin flips/binary signals.

In this section we use a simple example to illustrate how our additivity axiom captures a principle that is natural in settings of physical production of information, and contrast it with the sub-additivity—i.e., decreasing marginal costs—of entropy cost. Consider, for instance, the classical information acquisition problem of learning the bias of a coin by flipping it multiple times. In this context, entropy and LLR cost behave quite differently. Suppose the coin either yields heads 80% of the time or tails 80% of the time and either bias is equally likely. We are interested in how much more costly it is to observe the result of a single coin flip versus an infinite sequence of conditionally i.i.d. coin flips (or equivalently binary signals). Under LLR cost, the additivity axiom implies that the cost of observing k coin flips is linear in k, hence the cost of observing an infinite stream of coin flips is infinite.

Under entropy cost there exists $c > 0$ such that the cost of a single coin flip equals

$$\left[\{0.8\log(0.8) + 0.2\log(0.2)\} - \log\frac{1}{2}\right] c \approx 0.192745\, c.$$

Seeing an infinite sequence of coins reveals the state and thus leads to a posterior of 0 or 1.

The cost of seeing an infinite sequence of coin flips and thus learning the state is given by

$$\lim_{p\to 1}\left[\{p\log p+(1-p)\log(1-p)\}-\log\frac{1}{2}\right]c=\log(2)\,c\approx 0.693147\,c\,.$$

Thus, the cost of observing infinitely many coin flips is only approximately 3.59 times the cost of observing a single coin flip. Arguably, the low price of infinitely precise information is counter-factual in many contexts. The low cost is caused by the concavity of entropy as a cost function, which contrasts with the linearity of the log-likelihood ratio cost we propose (see Figure 2).

This difference in the marginal cost of information is not merely a mathematical difference, but could lead to substantially different predictions in economic applications. For example consider an investor who decides whether he wants to learn about the stocks in his home country or stocks in the international market. If she is already well informed about her home market entropy postulates that it will be cheaper to acquire further information about the returns of domestic stocks. Thus, under entropy the investor will focus more on domestic stocks and will be ex-ante more likely to purchase them. This rationale for the home bias has been proposed in Van Nieuwerburgh and Veldkamp (2009, 2010). We note that it crucially depends on the decreasing marginal cost of information as they mechanically bias information acquisition towards the better known home market. If the marginal cost of information is constant, acquiring a signal about the returns in the home market is equally costly as acquiring a signal about the foreign market and the investor will tend to spread his investments more evenly (this difference is explored in more detail in Van Nieuwerburgh and Veldkamp (2010) for normal states and in Morris and Strack (2018) for binary states).

5.3 Learning a Digit and Hypothesis Testing

In this section we apply the log-likelihood ratio cost to a standard hypothesis testing problem. We consider a decision maker performing an experiment with the goal of learning about an hypothesis $H\subseteq\Theta$ regarding the state of nature. We study how the cost of such an experiment depends on the structure of the hypothesis.

For concreteness, consider the case where the state is represented by a natural number i in an interval $\Theta=\{20,000,\ldots,80,000\}$, representing, for instance, the current US GDP per capita. Consider the following two questions: (1) Is i above or below 50,000? (2) Is i an odd or even number? Intuitively, under a model of costly information acquisition, answering the first question should be less expensive than answering the second (a practically impossible task for the US GDP). We show that while LLR cost can capture this difference, this is not the case under entropy cost and a uniform prior.

Given an hypothesis $H\subseteq\Theta$, we study the cost of symmetric binary signals where the set of signal realizations is $S=\{H,H^c\}$ and the output of the experiment is correct with

probability $\lambda > 1/2$, conditional on any state. Under LLR cost, the cost of such a signal is given by

$$\left(\sum_{i\in H, j\in H^c} \beta_{ij} + \beta_{ji}\right)\left(\lambda \log \frac{\lambda}{1-\lambda} + (1-\lambda)\log\frac{1-\lambda}{\lambda}\right) \tag{10}$$

The first term captures the difficulty of discerning between H and H^c in terms of the coefficients (β_{ij}). The second term is monotone in the signal precision λ and is independent of the hypothesis.

In the context of the GDP example, a simple way to capture the intuition that the cost of an experiment should depend on the structure of the hypothesis H is by setting $\beta_{ij} = 1/(i-j)^2$. Then, the difficulty $\sum_{i\in H, j\in H^c} \beta_{ij} + \beta_{ji}$ of discerning between H and H^c is higher when learning about an hypotheses that depends on the fine details of the state, such as learning whether or not i is odd or even, than when learning whether i is above or below a given threshold. It is higher in the former case because for each state is $i \in H$ there exists another state $j \notin H$ that is close to i.[22]

It is useful to compare these observations with the results that would be obtained under the entropy cost and a uniform prior on Θ. In such a model, the cost of a symmetric binary signal with precision λ is determined solely by the cardinality of H. This follows from the fact that entropy cost is invariant with respect to a relabelling of the states.

This example demonstrates that the LLR cost function can capture different phenomena from entropy cost. Rational inattention theory models the cost of paying attention to information that is freely available. In the above example, it is equally costly to read the last digit and the first digit of the per capita GDP in a newspaper. In contrast to rational inattention, we aim at modeling the cost of generating information.

5.4 Verification, Falsification, and Information Acquisition

Rejecting an hypothesis is an act of information acquisition, as is verifying it. It is common to encounter examples, both in science and everyday life, where the type of evidence that can be used to verify an hypothesis is qualitatively different from the evidence that can be used to verify it.

Consider, for instance, the classic statement "all swans are white." Say that a decision maker attaches prior probability 0.5 to the state 0 where this hypothesis is correct, and probability 0.5 to the state 1 where there exists a non-white swan. Imagine then two possible distributions over posteriors: in the first, with small probability ε it is revealed (perhaps up to a vanishing degree of doubt) that there exists at least one non-white swan, and with probability $1-\varepsilon$ the posterior probability that all swans are white remains close to the prior, increasing slightly. This distribution can be achieved by experiment I in Table

[22] Numerically, for a fixed λ the cost (10) of learning whether i is above or below 50,000 is higher than the cost learning whether i is odd or even by a factor of 6600.

	s	t
0	$1-\varepsilon^2$	ε^2
1	$1-\varepsilon$	ε

(a) Experiment I

	s	t
0	$1-\varepsilon$	ε
1	$1-\varepsilon^2$	ε^2

(b) Experiment II

Table 1: The set of states is $\Theta = \{0, 1\}$. In both experiments $S = \{s, t\}$. Under experiment I, observing the signal realization t rejects the hypothesis that the state is 0. Under experiment II, observing t verifies the same hypothesis.

1. Consider an alternative distribution in which the roles of the two states are reversed: with probability ε it is revealed that all swans are white, and with probability $1-\varepsilon$ the decision maker's belief in all swans being white decreases slightly. This distribution can be achieved by experiment II.

In practice, the first distribution over posteriors can be achieved by means of a simple experiment: as an extreme example, one may look up in the sky. There is an infinitesimal chance a black swan will be observed, in which case the posterior would shift to near certainty; if no black swan is observed the decision maker's belief would not change by much. It is not obvious what type of experiment would lead to the second distribution over posteriors, let along to conclude that the two experiment should be equally costly.

Formally, the two distributions over posteriors are equal up to a relabeling of the states: An experiment (μ_0, μ_1) that achieves the first distribution over posteriors will, by switching the roles of the two states, lead to the second one. However, the example points at an asymmetry between verification and falsification: permuting the state-dependent distributions of an experiment may, in practice, correspond to a completely different type of empirical investigation.

Hence, in order for a model of information acquisition to capture the difference between verification and falsification, the cost of an experiment should not necessarily be invariant with respect to a permutation of the states. In our model, this can be captured by assuming that the coefficients (β_{ij}) are non-symmetric, i.e. that β_{ij} and β_{ji} are are not necessarily equal. For instance, the cost of experiments I and II in Table 1 will differ whenever $\beta_{10} \neq \beta_{01}$.

Note that this is impossible under entropy and a uniform prior, since in that model the cost of an experiment is invariant with respect to a permutation of the states.

6 Related Literature

The question of how to quantify the amount of information provided by an experiment is the subject of a long-standing and interdisciplinary literature. Kullback and Leibler (1951) introduced the notion of Kullback-Leibler divergence as a measure of distance between statistical populations. Kelly (1956), Lindley (1956), Marschak (1959) and Arrow (1971) apply Shannon's entropy to the problem of ordering information structures.

More recently, Hansen and Sargent (2001) and Strzalecki (2011) adopted relative entropy as a tool to model robust decision criteria under uncertainty. Cabrales, Gossner, and Serrano (2013) derive Shannon's entropy as an index of informativeness for experiments in the context of portfolio choice problems (see also Cabrales, Gossner, and Serrano, 2017). Frankel and Kamenica (2018) put forward an axiomatic framework for quantifying the value and the amount of information in an experiment.

As discussed in the introduction, our work is also motivated by the recent literature on rational inattention and models of costly information acquisition based on Shannon's entropy. A complete survey of this area is beyond the scope of this paper; we refer the interested reader to Caplin (2016) and Mackowiak, Matějka, and Wiederholt (2018) for perspectives on this growing literature.

Our axiomatic approach differs both in terms of motivation and techniques from other results in the literature. Caplin and Dean (2015) study the revealed preference implications of rational inattention models, taking as a primitive state-dependent random choice data. Within the same framework, Andrew Caplin and Leahy (2017) characterize entropy cost, Chambers, Liu, and Rehbeck (2017) study non-separable models of costly information acquisition, and Denti (2018) provides a foundation for the assumption of posterior-separability.

Decision theoretic foundations for models of information acquisition have been put forward by de Oliveira (2014), De Oliveira, Denti, Mihm, and Ozbek (2017), and Ellis (2018). Mensch (2018) provides an axiomatic characterization of posterior-separable cost functions. Hébert and Woodford (2017), Zhong (2017b), Zhong (2017a) and Morris and Strack (2018) provide dynamic microfoundations for static information cost functions.

This paper is also related to the axiomatic literature in information theory characterizing different notions of entropy and information measures. Ebanks, Sahoo, and Sander (1998) and Csiszár (2008) survey and summarize the literature in the field. In the special case where $|\Theta| = 2$ and the coefficients (β_{ij}) are set to 1, the function (1) is also known as *J-divergence*. Kannappan and Rathie (1988) provide an axiomatization of J-divergence, under axioms very different from the ones in this paper. A more general representation appears in Zanardo (2017).

Ebanks, Sahoo, and Sander (1998) characterize functions over tuples of measures with finite support. They show that a condition equivalent to our additivity axiom leads to a

functional form similar to (1). Their analysis is however quite different from ours: their starting point is an assumption which, in the notation of this paper, states the existence of a map $F : \mathbb{R}^{\Theta} \to \mathbb{R}$ such that the cost of an experiment $(S, (\mu_i))$ with finite support takes the form $C(\mu) = \sum_{s \in S} F((\mu_i(s))_{i \in \Theta})$. This assumption of additive separability does not seem to have an obvious economic interpretation, nor to be related to our motivation of capturing constant marginal costs in information production.

The results in Mattner (1999, 2004) have, perhaps, the closest connection with this paper. Mattner studies functionals over the space probability measures over $\mathbb{R}$ that are additive with respect to convolution. As we explain in the next section, additivity with respect to convolution is a property that is closely related to Axiom 2. We draw inspiration from Mattner (1999) is applying the study of cumulants to the proof of Theorem 1. However, the difference in domain makes the techniques in Mattner (1999, 2004) not applicable to this paper.

7 Proof Sketch

In this section we informally describe some of the ideas involved in the proof of Theorem 1. We consider the binary case where $\Theta = \{0, 1\}$ and so there is only one relevant log-likelihood ratio $\ell = \ell_{10}$. The proof of the general case is more involved, but conceptually similar.

Step 1. Let C satisfy Axioms 1-4. Conditional on each state i, an experiment μ induces a distribution σ_i for the ratio ℓ. Two experiments that induce the same pair of distributions (σ_0, σ_1) are equivalent in the Blackwell order. Thus, by Axiom 1, C can be identified with a map $c(\sigma_0, \sigma_1)$ defined over all pairs of distributions induced by some experiment μ.

Step 2. Axioms 2 and 3 translate into the following properties of c. The product $\mu \otimes \nu$ of two experiments induces, conditional on i, a distribution for ℓ that is the *convolution* of the distributions induced by the two experiments. Axiom 2 is equivalent to c being additive with respect to convolution, i.e.

$$c(\sigma_0 * \tau_0, \sigma_1 * \tau_1) = c(\sigma_0, \sigma_1) + c(\tau_0, \tau_1)$$

Axiom 3 is equivalent to c satisfying for all $\alpha \in [0, 1]$,

$$c(\alpha\sigma_0 + (1-\alpha)\delta_0, \alpha\sigma_1 + (1-\alpha)\delta_0) = \alpha c(\sigma_0, \sigma_1)$$

where δ_0 is the degenerate measure at 0. Axiom 4 translates into continuity of c with respect to total variation and the first N moments of σ_0 and σ_1.

Step 3. As is well known, many properties of a probability distribution can be analyzed by studying its moments. We apply this idea to the study of experiments, and show that under our axioms the cost $c(\sigma_0, \sigma_1)$ is a function of the first N moments of the two measures,

for some (arbitrarily large) N. Given an experiment μ, we consider the experiment

$$\mu^n = \frac{1}{n} \cdot (\mu \otimes \cdots \otimes \mu)$$

in which with probability $1/n$ no information is produced, and with the remaining probability the experiment μ is carried out n times. By Axioms 2 and 3, the cost of μ^n is equal to the cost of μ.[23] We show that these properties, together with the continuity axiom, imply that the cost of an experiment is a function G of the moments of (σ_0, σ_1):

$$c(\sigma_0, \sigma_1) = G\left[m_{\sigma_0}(1), \ldots, m_{\sigma_0}(N), m_{\sigma_1}(1), \ldots, m_{\sigma_1}(N)\right] \tag{11}$$

where $m_{\sigma_i}(n)$ is the n-th moment of σ_i. Each $m_{\sigma_i}(n)$ is affine in σ_i, hence Step 2 implies that G is affine with respect to mixtures with the zero vector.

Step 4. It will be useful to analyze a distribution not only through its moments but also through its cumulants. The n-th *cumulant* $\kappa_\sigma(n)$ of a probability measure σ is the n-th derivative at 0 of the logarithm of its characteristic function. By a combinatorial characterization due to Leonov and Shiryaev (1959), $\kappa_\sigma(n)$ is a polynomial function of the first n moments $m_\sigma(1), \ldots, m_\sigma(n)$. For example, the first cumulant is the expectation $\kappa_\sigma(1) = m_\sigma(1)$, the second is the variance, and the third is $\kappa_\sigma(3) = m_\sigma(3) - 2m_\sigma(2)m_\sigma(1) + 2m_\sigma(1)^3$.

Step 3 and the result by Leonov and Shiryaev (1959) imply that the cost of an experiment is a function H of the cumulants of (σ_0, σ_1):

$$c(\sigma_0, \sigma_1) = H\left[\kappa_{\sigma_0}(1), \ldots, \kappa_{\sigma_0}(N), \kappa_{\sigma_1}(1), \ldots, \kappa_{\sigma_1}(N)\right] \tag{12}$$

where $\kappa_n[\sigma_i]$ is the n-th cumulant of σ_i.

Step 5. Cumulants satisfy a crucial property: the cumulant of a sum of two independent random variables is the sum of their cumulants. So, they are additive with respect to convolution. By Step 2, this implies that H is additive. We show that H is in fact a linear funtion. This step is reminiscent of the classic Cauchy equation problem. That is, understanding under what conditions a function $\phi\colon \mathbb{R} \to \mathbb{R}$ that satisfies $\phi(x+y) = \phi(x) + \phi(y)$ must be linear. In Theorem 4 we show, very generally, that any additive function from a subset $\mathcal{K} \subset \mathbb{R}^d$ to $\mathbb{R}_+$ is linear, provided $\mathcal{K}$ is closed under addition and has a non-empty interior. We then proceed to show that both of these conditions are satisfied if $\mathcal{K}$ is taken to be the domain of H, and thus deduce that H is linear.

Step 6. In the last step we study the implications of (11) and (12). We apply the characterization by Leonov and Shiryaev (1959) and show that the affinity with respect to the origin of the map G, and the linearity of H, imply that H must be a function

[23] For n large, the experiment μ^n has a very simple structure: With high probability it is uninformative, and with probability $1/n$ is highly revealing about the states.

solely of the first cumulants $\kappa_{\sigma_0}(1)$ and $\kappa_{\sigma_1}(1)$. That is, C must be a weighted sum of the expectations of the log-likelihood ratio ℓ conditional on each state.

A Discussion of the Continuity Axiom

Our continuity axiom may seem technical, and in a sense it is. However, there are some interesting technical subtleties involved with its choice. Indeed, it seems that a more natural choice of topology would be the topology of *weak convergence* of likelihood ratios (or, equivalently, posteriors). Under that topology, two experiments would be close if they had close expected utilities for decision problems with continuous bounded utilities. The disadvantage of this topology is that *no cost* that satisfies the rest of the axioms is continuous in this topology. To see this, consider the sequence of experiments in which a coin (whose bias depends on the state) is tossed n times with probability $1/n$, and otherwise is not tossed at all. Under our axioms these experiments all have the same cost—the cost of tossing the coin once. However, in the weak topology these experiments converge to the trivial experiment that yields no information and therefore has zero cost.

In fact, even the stronger *total variation* topology suffers from the same problem, which is demonstrated using the same sequence of experiments. Therefore, one must consider a *finer* topology (which makes for a weaker continuity assumption), which we do by also requiring the first N moments to converge. Note that increasing N makes for a finer topology and therefore a weaker continuity assumption, and that our results hold for all $N > 0$. An even stronger topology (which requires the convergence of all moments) is used by Mattner (1999, 2004) to find additive linear functionals on the space of all random variables on $\mathbb{R}$.

Nevertheless, the continuity axiom is technical. We state here without proof that it is not required when there are only two states, and we conjecture that it is not required in general.

B Preliminaries

For the rest of this section, in order to simplify the notation, we let $\Theta = \{0, 1, \ldots, n\}$, so that $|\Theta| = n + 1$.

B.1 Properties of the Kullback-Leibler Divergence

In this section we summarize some well known properties of the Kullback-Leibler divergence, and derive from them straightforward properties of the LLR cost.

Given a measurable space (X, Σ) we denote by $\mathcal{P}(X, \Sigma)$ the space of probability measures on (X, Σ). If $X = \mathbb{R}^d$ for some $d \in \mathbb{N}$ then Σ is implicitly assumed to be the corresponding Borel σ-algebra and we simply write $\mathcal{P}(\mathbb{R}^d)$.

For the next result, given two measurable spaces (Ω, Σ) and (Ω', Σ'), a measurable map $F\colon \Omega \to \Omega'$, and a measure $\eta \in \mathcal{P}(\Omega, \Sigma)$, we can define the *push-forward* measure $F_*\eta \in \mathcal{P}(\Omega', \Sigma')$ by $[F_*\eta](A) = \eta(F^{-1}(A))$ for all $A \in \Sigma'$.

Proposition 5. *Let $\nu_1, \nu_2, \eta_1, \eta_2$ be measures in $\mathcal{P}(\Omega, \Sigma)$, and let μ_1, μ_2 be probability measures in $\mathcal{P}(\Omega', \Sigma')$. Assume that $D_{\mathrm{KL}}(\nu_1 \| \nu_2)$, $D_{\mathrm{KL}}(\eta_1 \| \eta_2)$ and $D_{\mathrm{KL}}(\mu_1 \| \mu_2)$ are all finite. Let $F \colon \Omega \to \Omega'$ be measurable. Then:*

1. *$D_{\mathrm{KL}}(\nu_1 \| \nu_2) \geq 0$ with equality if and only if $\nu_1 = \nu_2$.*
2. *$D_{\mathrm{KL}}(\nu_1 \times \mu_1 \| \nu_2 \times \mu_2) = D_{\mathrm{KL}}(\nu_1 \| \nu_2) + D_{\mathrm{KL}}(\mu_1 \| \mu_2)$.*
3. *For all $\alpha \in (0, 1)$,*

$$D_{\mathrm{KL}}(\alpha\nu_1 + (1-\alpha)\eta_1 \| \alpha\nu_2 + (1-\alpha)\eta_2) \leq \alpha D_{\mathrm{KL}}(\nu_1 \| \nu_2) + (1-\alpha) D_{\mathrm{KL}}(\eta_1 \| \eta_2).$$

   and this equality is strict unless $\nu_1 = \eta_1$ and $\nu_2 = \eta_2$.

4. *$D_{\mathrm{KL}}(F_* \nu_1 \| F_* \mu_1) \leq D_{\mathrm{KL}}(\nu_1 \| \mu_1)$.*

Cover and Thomas (2012) present proofs of the first three statements (see Theorems 2.6.3, 2.5.3 and 2.7.2) in the case where all measures have finite support. It is immediate to adapt their proofs to the general case. We refer the reader to Austin (2006) for a proof of the last statement (see Proposition 2.4).

Lemma 1. *Two experiments $\mu = (S, (\mu_i))$ and $\nu = (T, (\nu_i))$ that satisfy $\bar{\mu}_i = \bar{\nu}_i$ for every $i \in \Theta$ are equivalent in the Blackwell order.*

Proof. The result is standard, but we include a proof for completeness. Suppose $\bar{\mu}_i = \bar{\nu}_i$ for every $i \in \Theta$. Given the experiment μ and a uniform prior on Θ, the posterior probability of state i conditional on s is given almost surely by

$$p_i(s) = \frac{\mathrm{d}\mu_i}{\mathrm{d}\sum_{j \in \Theta} \mu_j}(s) = \frac{1}{\sum_{j \in \Theta} \frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s)} = \frac{1}{\sum_{j \in \Theta} \mathrm{e}^{\ell_{ij}}} \tag{13}$$

and the corresponding expression applies to experiment ν. By assumption, conditional on each state the two experiments induce the same distribution of log-likelihood ratios (ℓ_{ij}). Hence, by (13) they must induce the same distribution over posteriors, hence be equivalent in the Blackwell order. □

A consequence of Proposition 5 is that the LLR cost is monotone with respect to the Blackwell order.

Proposition 6. *Let $\mu = (S, (\mu_i)_{i \in \Theta})$ and $\nu = (T, (\nu_i)_{i \in \Theta})$ be experiments such that μ Blackwell dominates ν. Then any LLR cost C satisfies $C(\mu) \geq C(\nu)$.*

Proof. Let C be a LLR cost. It is immediate that if $\bar{\mu}_i = \bar{\nu}_i$ for every i then $C(\mu) = C(\nu)$. We can assume without loss of generality that $S = T = \mathcal{P}(\Theta)$, endowed with the Borel

σ-algebra. This follows from the fact that we can define a new experiments $\rho = (\mathcal{P}(\Theta), (\rho_i))$ such that $\bar{\mu}_i = \bar{\rho}_i$ for every i (see, e.g. Le Cam (1996)), and apply the same result to ν .

By Blackwell's Theorem there exists a probability space (R, λ) and a "garbling" map $G\colon S \times R \to T$ such that for each $i \in \Theta$ it holds that $\nu_i = G_*(\mu_i \times \lambda)$. Hence, by the first, second and fourth statements in Proposition 5,

$$\begin{aligned} D_{\mathrm{KL}}(\nu_i \| \nu_j) &= D_{\mathrm{KL}}(G_*(\mu_i \times \lambda) \| G_*(\mu_j \times \lambda)) \\ &\leq D_{\mathrm{KL}}(\mu_i \times \lambda \| \mu_j \times \lambda) \\ &= D_{\mathrm{KL}}(\mu_i \| \mu_j) + D_{\mathrm{KL}}(\lambda \| \lambda) \\ &= D_{\mathrm{KL}}(\mu_i \| \mu_j). \end{aligned}$$

Therefore, by Theorem 1, we have

$$C(\nu) = \sum_{i,j \in \Theta} \beta_{ij} D_{\mathrm{KL}}(\nu_i \| \nu_j) \leq \sum_{i,j \in \Theta} \beta_{ij} D_{\mathrm{KL}}(\mu_i \| \mu_j) = C(\mu).$$

□

We note that a similar argument shows that if all the coefficients β_{ij} are positive then $C(\mu) > C(\nu)$ whenever μ Blackwell dominates ν but ν does not dominate μ.

An additional direct consequence of Proposition 5 is that the LLR cost is convex:

Proposition 7. *Let $\mu = (S, (\mu_i)_{i \in \Theta})$ and $\nu = (S, (\nu_i)_{i \in \Theta})$ be experiments. Given $\alpha \in (0, 1)$, define their convex combination η by $\eta_i = \alpha \nu_i + (1 - \alpha)\mu_i$. Then any LLR cost C satisfies*

$$C(\eta) \leq \alpha C(\nu) + (1 - \alpha) C(\mu).$$

The follows immediately from the third statement in Proposition 5. As before, we note that C is strictly convex if all of its coefficients β_{ij} are positive.

We now study the set

$$\mathcal{D} = \{(D_{\mathrm{KL}}(\mu_i \| \mu_j))_{i \neq j} : \mu \in \mathcal{E}\} \subseteq \mathbb{R}_+^{(n+1)n}$$

of all possible pairs of expected log-likelihood ratios induced by some experiment μ. The next result shows that $\mathcal{D}$ contains the strictly positive orthant.

Lemma 2. $\mathbb{R}_{++}^{(n+1)n} \subseteq \mathcal{D}$

Proof. The set $\mathcal{D}$ is convex. To see this, let $\mu = (S, (\mu_i))$ and $\nu = (T, (\nu_i))$ be two experiments. Without loss of generality, we can suppose that $S = T$, and $S = S_1 \cup S_2$, where S_1, S_2 are disjoint, and $\mu_i(S_1) = \nu_i(S_2) = 1$ for every i.

Fix $\alpha \in (0, 1)$ and define the new experiment $\tau = (S, (\tau_i))$ where $\tau_i = \alpha \mu_i + (1 - \alpha)\nu_i$ for every i. It can be verified that τ_i-almost surely, $\frac{\mathrm{d}\tau_i}{\mathrm{d}\tau_j}$ satisfies $\frac{\mathrm{d}\tau_i}{\mathrm{d}\tau_j}(s) = \frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s)$ if $s \in S_1$

and $\frac{\mathrm{d}\tau_i}{\mathrm{d}\tau_j}(s) = \frac{\mathrm{d}\nu_i}{\mathrm{d}\nu_j}(s)$ if $s \in S_2$. It then follows that

$$D_{\mathrm{KL}}(\tau_i \| \tau_j) = \alpha D_{\mathrm{KL}}(\mu_i \| \mu_j) + (1-\alpha) D_{\mathrm{KL}}(\nu_i \| \nu_j)$$

Hence $\mathcal{D}$ is convex. We now show $\mathcal{D}$ is a convex cone. First notice that the zero vector belongs to $\mathcal{D}$, since it corresponds to the totally uninformative experiment. In addition (see §B.1),

$$D_{\mathrm{KL}}((\mu \otimes \mu)_i \| (\mu \otimes \mu)_j) = D_{\mathrm{KL}}(\mu_i \times \mu_i \| \mu_j \times \mu_j) = 2 D_{\mathrm{KL}}(\mu_i \| \mu_j)$$

Hence $\mathcal{D}$ is closed under addition. Because $\mathcal{D}$ is also convex and contains the zero vector, it follows that it is a convex cone.

Suppose, by way of contradiction, that the inclusion $\mathbb{R}_{++}^{(n+1)n} \subseteq \mathcal{D}$ does not hold. Then, we can find a vector $z \in \mathbb{R}_{+}^{(n+1)n}$ that does not belong to the closure of $\mathcal{D}$. Therefore, there exists a vector $w \in \mathbb{R}^{(n+1)n}$ and $t \in \mathbb{R}$ such that $w \cdot z > t \geq w \cdot y$ for all $y \in \mathcal{D}$. Because $\mathcal{D}$ is a cone, then $t \geq 0$ and $0 \geq w \cdot y$ for all $y \in \mathcal{D}$. Let $i_o j_o$ be a coordinate such that $w_{i_o j_o} > 0$.

Consider the following three cumulative distribution functions on $[2, \infty)$:

$$F_1(x) = 1 - \frac{2}{x}$$
$$F_2(x) = 1 - \frac{\log^2 2}{\log^2 x}$$
$$F_3(x) = 1 - \frac{\log 2}{\log x},$$

and denote by π_1, π_2, π_3 the corresponding measures. A simple calculation shows that $D_{\mathrm{KL}}(\pi_3 \| \pi_1) = \infty$, whereas $D_{\mathrm{KL}}(\pi_a \| \pi_b) < \infty$ for any other choice of $a, b \in \{1,2,3\}$.

Let $\pi_a^\varepsilon = (1-\varepsilon)\,\delta_2 + \varepsilon \pi_a$ for every $a \in \{1,2,3\}$, where δ_2 is the point mass at 2. Then still $D_{\mathrm{KL}}(\pi_3^\varepsilon \| \pi_1^\varepsilon) = \infty$, but, for any other choice of a and b in $\{1,2,3\}$, the divergence $D(\pi_a^\varepsilon \| \pi_b^\varepsilon)$ vanishes as ε goes to zero. Let $\pi_a^{\varepsilon,M}$ be the measure π_a^ε conditioned on $[2, M]$. Then $D_{\mathrm{KL}}(\pi_a^{\varepsilon,M} \| \pi_b^{\varepsilon,M})$ tends to $D_{\mathrm{KL}}(\pi_a^\varepsilon \| \pi_b^\varepsilon)$ as M tends to infinity, for any a, b. It follows that for every $N \in \mathbb{N}$ there exist ε small enough and M large enough such that $D_{\mathrm{KL}}(\pi_3^{\varepsilon,M} \| \pi_1^{\varepsilon,M}) > N$ and, for any other choice of a, b, $D_{\mathrm{KL}}(\pi_a^{\varepsilon,M} \| \pi_b^{\varepsilon,M}) < 1/N$.

Consider the experiment $\mu = (\mathbb{R}, (\mu_i))$ where $\mu_{i_0} = \pi_3^{\varepsilon,M}$, $\mu_{j_0} = \pi_1^{\varepsilon,M}$ and $\mu_k = \pi_2^{\varepsilon,M}$ for all $k \notin \{i_0, j_0\}$ and with ε and M so that the above holds for N large enough. Then $\mu \in \mathcal{E}$ since all measures have bounded support. It satisfies $D_{\mathrm{KL}}(\mu_{i_o} \| \mu_{j_o}) > N$ and $D_{\mathrm{KL}}(\mu_i \| \mu_j) < 1/N$ for every other pair ij.

Now let $y \in \mathcal{D}$ be the vector defined by μ. Then $w \cdot y > 0$ for N large enough. A contradiction. □

B.2 Experiments and Log-likelihood Ratios

It will be convenient to consider, for each experiment, the distribution over log-likelihood ratios with respect to the state $i = 0$ conditional on a state j. Given an experiment, we define $\ell_i = \ell_{i0}$ for every $i \in \Theta$. We say that a vector $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_n) \in \mathcal{P}(\mathbb{R}^n)^{n+1}$ of measures is *derived from the experiment* $(S, (\mu_i))$ if for every $i = 0, 1, \ldots, n$,

$$\sigma_i(E) = \mu_i\left(\{s : (\ell_1(s), \ldots, \ell_n(s)) \in E\}\right) \text{ for all measurable } E \subseteq \mathbb{R}^n$$

That is, σ_i is the distribution of the vector $(\ell_1, \ldots, \ell_n)$ of log-likelihood ratios (with respect to state 0) conditional on state i. There is a one-to-one relation between the vector σ and the collection $(\bar{\mu}_i)$ of distributions defined in the main text. Notice that $\ell_{ij} = \ell_{i0} - \ell_{j0}$ almost surely, hence knowing the distribution of $(\ell_{0i})_{i\in\Theta}$ is enough to recover the distribution of $(\ell_{ij})_{i,j\in\Theta}$. Nevertheless, working directly with σ (rather than $(\bar{\mu}_i)$) will simplify the notation considerably.

We call a vector $\sigma \in \mathcal{P}(\mathbb{R}^n)^{n+1}$ *admissible* if it is derived from some experiment. The next result provides a straightforward characterization of admissible vectors of measures.

Lemma 3. *A vector of measures $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_n)$ is admissible if and only if the measures are mutually absolutely continuous and, for every i, satisfy $\frac{\mathrm{d}\sigma_i}{\mathrm{d}\sigma_0}(\xi) = e^{\xi_i}$ for σ_i-almost every $\xi \in \mathbb{R}^n$.*

Proof. If $(\sigma_0, \sigma_1, \ldots, \sigma_n)$ is admissible then there exists an experiment $\mu = (S, (\mu_i))$ such that for any measurable $E \subseteq \mathbb{R}^n$

$$\begin{aligned}\int_E e^{\xi_i}\,\mathrm{d}\sigma_0(\xi) &= \int 1_E\left((\ell_1(s), \ldots \ell_n(s))\right) e^{\ell_i(s)}\,\mathrm{d}\mu_0(s) \\ &= \int 1_E\left((\ell_1(s), \ldots \ell_n(s))\right)\,\mathrm{d}\mu_i(s)\end{aligned}$$

where 1_E is the indicator function of E. So, $\int_E e^{\xi_i}\,\mathrm{d}\sigma_0(\xi) = \sigma_i(E)$ for every $E \subseteq \mathbb{R}^n$. Hence e^{ξ_i} is a version of $\frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_0}$.

Conversely, assume $\frac{\mathrm{d}\sigma_i}{\mathrm{d}\sigma_0}(\xi) = e^{\xi_i}$ for σ_i-almost every $\xi \in \mathbb{R}^n$. Define an experiment $(\mathbb{R}^{n+1}, (\mu_i))$ where $\mu_i = \sigma_i$ for every i. The experiment $(\mathbb{R}^{n+1}, (\mu_i))$ is such that $\ell_i(\xi) = \xi_i$ for every $i > 0$. Hence, for $i > 0$, $\mu_i\left(\{\xi : (\ell_1(\xi), \ldots, \ell_n(\xi)) \in E\}\right)$ is equal to

$$\int 1_E\left((\ell_1(\xi), \ldots \ell_n(\xi))\right) e^{x_i}\,\mathrm{d}\sigma_0(t) = \int 1_E(\xi) e^{x_i}\,\mathrm{d}\sigma_0 = \sigma_i(E)$$

and similarly $\mu_0\left(\{\xi : (\ell_1(\xi), \ldots, \ell_n(\xi)) \in E\}\right) = \sigma_0(E)$. So $(\sigma_0, \ldots, \sigma_n)$ is admissible. □

B.3 Properties of Cumulants

The purpose of this section is to formally describe cumulants and their relation to moments. We follow Leonov and Shiryaev (1959) and (Shiryaev, 1996, p. 289). Given a vector $\xi \in \mathbb{R}^n$ and an integral vector $\alpha \in \mathbb{N}^n$ we write $\xi^\alpha = \xi_1^{\alpha_1}\xi_2^{\alpha_2}\cdots\xi_n^{\alpha_n}$ and use the notational conventions $\alpha! = \alpha_1!\alpha_2!\cdots\alpha_n!$ and $|\alpha| = \alpha_1 + \cdots \alpha_n$.

Let $A = \{0,\ldots,N\}^n \backslash \{0,\ldots,0\}$, for some constant $N \in \mathbb{N}$ grater or equal than 1. For every probability measure $\sigma_1 \in \mathcal{P}(\mathbb{R}^n)$ and $\xi \in \mathbb{R}^n$, let $\varphi_{\sigma_1}(\xi) = \int_{\mathbb{R}^n} e^{i\langle z,\xi\rangle}\, \mathrm{d}\sigma_1(z)$ denote the characteristic function of σ_1 evaluated at ξ. We denote by $\mathcal{P}_A \subseteq \mathcal{P}(\mathbb{R}^n)$ the subset of measures σ_1 such that $\int_{\mathbb{R}^n} |\xi^\alpha|\, \mathrm{d}\sigma_1(\xi) < \infty$ for every $\alpha \in A$. Every $\sigma_1 \in \mathcal{P}_A$ is such that in a neighborhood of $\mathbf{0} \in \mathbb{R}^n$ the cumulant generating function $\log \varphi_{\sigma_1}(z)$ is well defined and the partial derivatives

$$\frac{\partial^{|\alpha|}}{\partial\xi_1^{\alpha_1}\partial\xi_2^{\alpha_2}\cdots\partial\xi_n^{\alpha_n}} \log \varphi_{\sigma_1}(\xi)$$

exists and are continuous for every $\alpha \in \mathbb{N}^n$.

For every $\sigma_1 \in \mathcal{P}_A$ and $\alpha \in A$ let $\kappa_{\sigma_1}(\alpha)$ be defined as

$$\kappa_{\sigma_1}(\alpha) = i^{-|\alpha|} \frac{\partial^{|\alpha|}}{\partial\xi_1^{\alpha_1}\partial\xi_2^{\alpha_2}\cdots\partial\xi_n^{\alpha_n}} \log \varphi_{\sigma_1}(\mathbf{0})$$

With slight abuse of terminology, we refer to $\kappa_{\sigma_1} \in \mathbb{R}^A$ as the *vector of cumulants* of σ_1. In addition, for every $\sigma_1 \in \mathcal{P}_A$ and $\alpha \in A$ we denote by $m_{\sigma_1}(\alpha) = \int_{\mathbb{R}^n} \xi^\alpha\, \mathrm{d}\sigma_1(\xi)$ the mixed moment of σ_1 of order α and refer to $m_{\sigma_1} \in \mathbb{R}^A$ as the *vector of moments* of σ_1.

Given two measures $\sigma_1, \sigma_2 \in \mathcal{P}(\mathbb{R}^n)$ we denote by $\sigma_1 * \sigma_2 \in \mathcal{P}(\mathbb{R}^n)$ the corresponding convolution.

Lemma 4. *For every $\sigma_1, \sigma_2 \in \mathcal{P}_A$, and $\alpha \in A$, $\kappa_{\sigma_1*\sigma_2}(\alpha) = \kappa_{\sigma_1}(\alpha) + \kappa_{\sigma_2}(\alpha)$.*

Proof. The result follows from the well known fact that $\varphi_{\sigma_1*\sigma_2}(\xi) = \varphi_{\sigma_1}(\xi)\varphi_{\sigma_2}(\xi)$ for every $\xi \in \mathbb{R}^n$. □

The next result, due to Leonov and Shiryaev (1959), establishes a one-to-one relation between the moments $\{m_{\sigma_1}(\alpha) : \alpha \in A\}$ and the cumulants $\{\kappa_{\sigma_1}(\alpha) : \alpha \in A\}$ of a probability measure $\sigma_1 \in \mathcal{P}_A$. Given $\alpha \in A$, let $\Lambda(\alpha)$ be the set of all ordered collections $(\lambda^1,\ldots,\lambda^q)$ of non-zero vectors in $\mathbb{N}^n$ such that $\sum_{p=1}^q \lambda^p = \alpha$.

Theorem 2. *For every $\sigma_1 \in \mathcal{P}_A$ and $\alpha \in A$,*

1. $m_{\sigma_1}(\alpha) = \sum_{(\lambda^1,\ldots,\lambda^q)\in\Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1!\cdots\lambda^q!} \prod_{p=1}^q \kappa_{\sigma_1}(\lambda^p)$
2. $\kappa_{\sigma_1}(\alpha) = \sum_{(\lambda^1,\ldots,\lambda^q)\in\Lambda(\alpha)} \frac{(-1)^{q-1}}{q} \frac{\alpha!}{\lambda^1!\cdots\lambda^q!} \prod_{p=1}^q m_{\sigma_1}(\lambda^p)$

B.4 Admissible Measures and the Cumulants Manifold

We denote by $\mathcal{A}$ the set of vectors of measures $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_n)$ that are admissible and such that $\sigma_i \in \mathcal{P}_A$ for every i. To each $\sigma \in \mathcal{A}$ we associate the vector

$$m_\sigma = (m_{\sigma_0}, m_{\sigma_1}, \ldots, m_{\sigma_n}) \in \mathbb{R}^d$$

of dimension $d = (n+1)\,|A|$. Similarly, we define

$$\kappa_\sigma = (\kappa_{\sigma_0}, \kappa_{\sigma_1}, \ldots, \kappa_{\sigma_n}) \in \mathbb{R}^d.$$

In this section we study properties of the sets $\mathcal{M} = \{m_\sigma : \sigma \in \mathcal{A}\}$ and $\mathcal{K} = \{\kappa_\sigma : \sigma \in \mathcal{A}\}$.

Lemma 5. *Let I and J be disjoint finite sets and let $(\phi_k)_{k\in I\cup J}$ be a collection of real valued functions defined on $\mathbb{R}^n$. Assume $\{\phi_k : k \in I \cup J\} \cup \{1_{\mathbb{R}^n}\}$ are linearly independent and the unit vector $(1, \ldots, 1) \in \mathbb{R}^J$ belongs to the the interior of $\{(\phi_k(\xi))_{k\in J} : \xi \in \mathbb{R}^n\}$. Then*

$$C = \left\{ \left(\int_{\mathbb{R}^n} \phi_k \, \mathrm{d}\sigma_1 \right)_{k\in I} : \sigma_1 \in \mathcal{P}(\mathbb{R}^n) \text{ has finite support and } \int_{\mathbb{R}^n} \phi_k \, \mathrm{d}\sigma_1 = 1 \text{ for all } k \in J \right\}$$

is a convex subset of $\mathbb{R}^I$ with nonempty interior.

Proof. To ease the notation, let $Y = \mathbb{R}^n$ and denote by $\mathcal{P}_o$ be the set of probability measures on Y with finite support. Consider $F = \{\phi_k : k \in I \cup J\} \cup \{1_{\mathbb{R}^d}\}$ as a subset of the vector space $\mathbb{R}^Y$, where the latter is endowed with the topology of pointwise convergence. The topological dual of $\mathbb{R}^Y$ is the vector space of signed measures on Y with finite support. Let

$$D = \left\{ \left(\int_{\mathbb{R}^n} \phi_k \, \mathrm{d}\sigma_1 \right)_{k\in I\cup J} : \sigma_1 \in \mathcal{P}_o \right\} \subseteq \mathbb{R}^{I\cup J}.$$

Fix $k \in I \cup J$. Since ϕ_k does not belong to the linear space V generated by $\{\phi \in F : \phi \neq \phi_k\}$, then there exists a signed measure

$$\rho = \alpha\sigma_1 - \beta\sigma_2$$

where $\alpha, \beta \geq 0$, $\alpha + \beta > 0$ and $\sigma_1, \sigma_2 \in \mathcal{P}_o$, such that ρ satisfies $\int \phi_k \, \mathrm{d}\rho > 0 \geq \int \phi \, \mathrm{d}\rho$ for every $\phi \in V$.

This implies $\int \phi \, \mathrm{d}\rho = 0$ for every $\phi \in V$. By taking $\phi = 1_{\mathbb{R}^n}$, we obtain $\rho(\mathbb{R}^n) = 0$. Hence, $\alpha = \beta$. Therefore, $\int \phi_k \, \mathrm{d}\sigma_1 > \int \phi_k \, \mathrm{d}\sigma_2$ and $\int \phi_m \, \mathrm{d}\sigma_1 = \int \phi_m \, \mathrm{d}\sigma_2$ for every ϕ_m in F that is distinct from ϕ_k. Because k is arbitrary, it follows that the linear space generated by D equals $\mathbb{R}^{I\cup J}$. Because D is convex and spans $\mathbb{R}^{I\cup J}$, then D has nonempty interior.

Now consider the hyperplane

$$H = \{z \in \mathbb{R}^{I \cup J} : z_k = 1 \text{ for all } k \in J\}$$

Let D^o be the interior of D. It remains to show that the hyperplane H satisfies $H \cap D^o \neq \emptyset$. This will imply that the projection of $H \cap D$ on $\mathbb{R}^I$, which equals C, has non-empty interior.

Let $w \in D^o$. By assumption, $(1, \ldots, 1) \in \mathbb{R}^J$ is in the interior of $\{(\phi_k(\xi))_{k \in J} : \xi \in Y\}$. Hence, there exists $\alpha \in (0, 1)$ small enough and $\xi \in Y$ such that $\phi_k(\xi) = \frac{1}{1-\alpha} - \frac{\alpha}{1-\alpha} w_k$ for every $k \in J$. Define $z = \alpha w + (1 - \alpha)(\phi_k(\xi))_{k \in I \cup J} \in D$. Then $z_k = 1$ for every $k \in J$. In addition, because $w \in D^o$ then $z \in D^o$ as well. Hence $z \in H \cap D^o$. □

Lemma 6. *The set $\mathcal{M} = \{m_\sigma : \sigma \in \mathcal{A}\}$ has nonempty interior.*

Proof. For eevery $\alpha \in A$ define the functions $(\phi_{i,\alpha})_{i \in \Theta}$ as

$$\phi_{0,\alpha}(\xi) = \xi^\alpha \text{ and } \phi_{i,\alpha}(\xi) = \xi^\alpha e^{\xi_i} \text{ for all } i > 0.$$

Define $\psi_0 = 1_{\mathbb{R}^n}$ and $\psi_i(\xi) = e^{\xi_i}$ for all $i > 0$. It is immediate to verify that

$$\{\phi_{i,\alpha} : i \in \Theta, \alpha \in A\} \cup \{\psi_i : i \in \Theta\}$$

is a linearly independent set of functions. In addition, $(1, \ldots, 1) \in \mathbb{R}^n$ is in the interior of $\{(e^{\xi_1}, \ldots, e^{\xi_n}) : \xi \in \mathbb{R}^n\}$. Lemma 5 implies that the set

$$C = \left\{ \left(\int_{\mathbb{R}^n} \phi_{i,\alpha} \, \mathrm{d}\sigma_0 \right)_{\substack{i \in \Theta \\ \alpha \in A}} : \sigma_0 \in \mathcal{P}(\mathbb{R}^n) \text{ has finite support and } \int_{\mathbb{R}^n} e^{\xi_i} \, \mathrm{d}\sigma_0(\xi) = 1 \text{ for all } i \right\}$$

has nonempty interior.

Given σ_0 as in the definition of C, construct a vector $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_n)$ where for each $i > 0$ the measure σ_i is defined so that $(\mathrm{d}\sigma_i / \mathrm{d}\sigma_0)(\xi) = \mathrm{e}^{\xi_i}$, σ_0-almost surely. Then, Lemma 3 implies σ is admissible. Because each σ_i has finite support then $\sigma \in \mathcal{A}$. In addition,

$$m_\sigma = \left(\int_{\mathbb{R}^n} \phi_{i,\alpha} \, \mathrm{d}\sigma_0 \right)_{\substack{i \in \Theta \\ \alpha \in A}}$$

hence $C \subseteq \mathcal{M}$. Thus, $\mathcal{M}$ has nonempty interior. □

Theorem 3. *The set $\mathcal{K} = \{\kappa_\sigma : \sigma \in \mathcal{A}\}$ has nonempty interior.*

Proof. Theorem 2 establishes the existence of a continuous one-to-one map $m_{\sigma_0} \mapsto \kappa_{\sigma_0}$, $\sigma_0 \in \mathcal{P}_A$. Therefore, we can define a one-to-one function $H : \mathcal{M} \to \mathbb{R}^d$ such that $H(m_\sigma) = \kappa_\sigma$ for every $\sigma \in \mathcal{A}$. Lemma 6 shows there exists an open set $U \subseteq \mathbb{R}^d$ included in $\mathcal{M}$. Let H_U be the restriction of H on U. Then H_U satisfies all the assumptions of

Brouwer's Invariance of Domain Theorem,[24] which implies that $H_U(U)$ is an open subset of $\mathbb{R}^d$. Since $H(\mathcal{M}) \subseteq \mathcal{K}$, it follows that $\mathcal{K}$ has nonempty interior. □

B.5 Automatic continuity in the Cauchy problem for subsemigroups of $\mathbb{R}^d$.

A *subsemigroup* of $\mathbb{R}^d$ is a subset $\mathcal{S} \subseteq \mathbb{R}^d$ that is closed under addition, so that $x + y \in \mathcal{S}$ for all $x, y \in \mathcal{S}$. We say that a map $F\colon \mathcal{S} \to \mathbb{R}_+$ is *additive* if $F(x+y) = F(x) + F(y)$ for all $x, y, x+y \in \mathcal{S}$. We say that F is *linear* if there exists $(a_1, \dots, a_d) \in \mathbb{R}^d$ such that $F(x) = F(x_1, \dots, x_d) = a_1 x_1 + \cdots + a_d x_d$ for all $x \in \mathcal{S}$.

We can now state the main result of this section:

Theorem 4. *Let $\mathcal{S}$ be a subsemigroup of $\mathbb{R}^d$ with a nonempty interior. Then every additive function $F\colon \mathcal{S} \to \mathbb{R}_+$ is linear.*

Before proving the theorem we will establish a number of claims.

Claim 1. Let $\mathcal{S}$ be a subsemigroup of $\mathbb{R}^d$ with a nonempty interior. Then there exists an open ball $B \subset \mathbb{R}^d$ such that $aB \subset \mathcal{S}$ for all real $a \geq 1$.

Proof. Let B_0 be an open ball contained in $\mathcal{S}$, with center x_0 and radius r. Given a positive integer k, note that kB_0 is the ball of radius kr centered at kr_0, and that it is contained in $\mathcal{S}$, since $\mathcal{S}$ is a semigroup. Choose a positive integer $M \geq 4$ such that $\frac{2}{3}Mr > \|x_0\|$, and let B be the open ball with center at Mx_0 and radius r (see Figure 3). Fix any $a \geq 1$, and write $a = \frac{1}{M}(n+\gamma)$ for some integer $n \geq M$ and $\gamma \in [0,1)$. Then $\frac{n}{M}B$ is the ball of radius $\frac{n}{M}r$ centered at nx_0, which is contained in nB_0, since nB_0 also has center nx_0, but has a larger radius nr. So $\frac{n}{M}B \subset nB_0$. We claim that furthermore $\frac{n+1}{M}B$ is also contained in nB_0. To see this, observe that the center of $\frac{n+1}{M}B$ is $(n+1)x_0$ and its radius is $\frac{n+1}{M}r$. Hence the center of $\frac{n+1}{M}B$ is at distance $\|x_0\|$ from the center of nB_0, and so the furthest point in $\frac{n+1}{M}B$ is at distance $\|x_0\| + \frac{n+1}{M}r$ from the center of nB_0. But the radius of nB_0 is

$$nr = \frac{2}{3}nr + \frac{1}{3}nr \geq \frac{2}{3}Mr + \frac{1}{3}nr > \|x_0\| + \frac{n+1}{M}r,$$

where the first inequality follows since $n \geq M$, and the second since $\frac{2}{3}Mr > \|x_0\|$ and $M \geq 4$. So nB_0 indeed contains both $\frac{n}{M}B$ and $\frac{n+1}{M}B$. Thus it also contains aB, and so $\mathcal{S}$ contains aB. □

[24] Theorem 2 at https://terrytao.wordpress.com/2011/06/13/brouwers-fixed-point-and-invariance-of-domain-theorems-and-hilberts-fifth-problem/

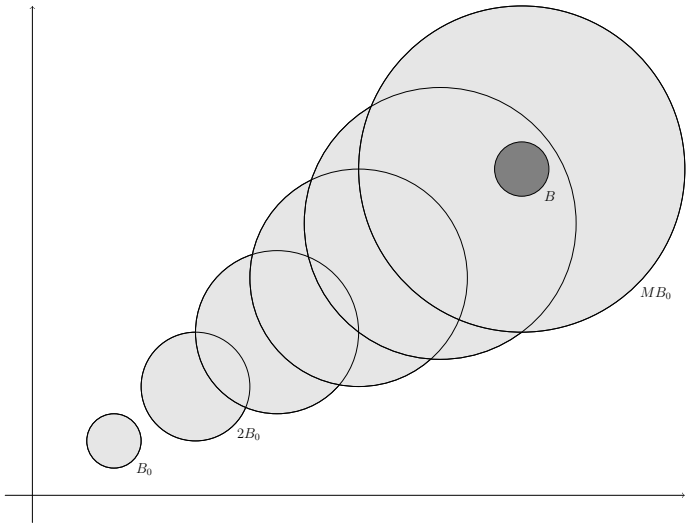

Figure 3: Illustration of the proof of Claim 1. The dark ball B is contained in the light ones, and it is apparent from this image that so is any multiple of B by $a \geq 1$.

Claim 2. Let $\mathcal{S}$ be a subsemigroup of $\mathbb{R}^d$ with a nonempty interior. Let $F\colon \mathcal{S} \to \mathbb{R}_+$ be additive and satisfy $F(ay) = aF(y)$ for every $y \in \mathcal{S}$ and $a \in \mathbb{R}_+$ such that $ay \in \mathcal{S}$. Then F is linear.

Proof. If $\mathcal{S}$ does not include zero, then without loss of generality we add zero to it and set $F(0) = 0$. Let B be an open ball such that $aB \subset \mathcal{S}$ for all $a \geq 1$; the existence of such a ball is guaranteed by Claim 1. Choose a basis $\{b^1, \ldots, b^d\}$ of $\mathbb{R}^d$ that is a subset of B, and let $x = \beta_1 b^1 + \cdots + \beta_d b^d$ be an arbitrary element of $\mathcal{S}$. Let $b = \max\{1/|\beta_i| : \beta_i \neq 0\}$, and let $a = \max\{1, b\}$. Then

$$F(ax) = F(a\beta_1 b^1 + \cdots + a\beta_d b^d).$$

Assume without loss of generality that for some $0 \leq k \leq d$ it holds that the first k coefficients β_i are non-negative, and the rest are negative. Then for $i \leq k$ it holds that $a\beta_i b^i \in \mathcal{S}$ and for $i > k$ it holds that $-a\beta_i b^i \in \mathcal{S}$; this follows from the defining property of the ball B, since each b^i is in B, and since $|a\beta_i| \geq 1$. Hence we can add $F(-a\beta_{k+1} b^{k+1} - \cdots - a\beta_d b^d)$ to both sides of the above displayed equation, and then by additivity,

$$\begin{aligned}
&F(ax) + F(-a\beta_{k+1} b^{k+1} - \cdots - a\beta_d b^d) \\
&= F(a\beta_1 b^1 + \cdots + a\beta_d b^d) + F(-a\beta_{k+1} b^{k+1} - \cdots - a\beta_d b^d) \\
&= F(a\beta_1 b^1 + \cdots + a\beta_k b^k).
\end{aligned}$$

Using additivity again yields

$$F(ax) + F(-a\beta_{k+1}b^{k+1}) + \cdots + F(-a\beta_d b^d) = F(a\beta_1 b^1) + \cdots + F(a\beta_k b^k).$$

Applying now the claim hypothesis that $F(ay) = aF(y)$ whenever $y, ay \in \mathcal{S}$ yields

$$aF(x) + (-a\beta_{k+1})F(b^{k+1}) + \cdots + (-a\beta_d)F(b^d) = a\beta_1 F(b^1) + \cdots + a\beta_k F(b^k).$$

Rearranging and dividing by a, we arrive at

$$F(x) = \beta_1 F(b^1) + \cdots + \beta_d F(b^d).$$

We can therefore extend F to a function that satisfies this on all of $\mathbb{R}^d$, which is then clearly linear. □

Claim 3. Let B be an open ball in $\mathbb{R}^d$, and let $\mathcal{B}$ be the semigroup given by $\cup_{a\geq 1} aB$. Then every additive $F\colon \mathcal{B} \to \mathbb{R}_+$ is linear.

Proof. Fix any $x \in \mathcal{B}$, and assume $ax \in \mathcal{B}$ for some $a \in \mathbb{R}_+$. Since $\mathcal{B}$ is open, by Claim 2 it suffices to show that $F(ax) = aF(x)$. The defining property of $\mathcal{B}$ implies that the intersection of $\mathcal{B}$ and the ray $\{bx : b \geq 0\}$ is of the form $\{bx : b > a_0\}$ for some $a_0 \geq 0$. By the additive property of F, we have that $F(qx) = qF(x)$ for every rational $q > a_0$. Furthermore, if $b > b' > a_0$ then $n(b-b')x \in \mathcal{S}$ for n large enough. Hence

$$\begin{aligned} F(bx) &= \frac{1}{n}F(nbx) \\ &= \frac{1}{n}F\left(nb'x + (n(b-b')x)\right) \\ &= \frac{1}{n}F\left(nb'x\right) + \frac{1}{n}F\left(n(b-b')x\right) \\ &= F(b'x) + \frac{1}{n}F\left(n(b-b')x\right) \\ &\geq F(b'x). \end{aligned}$$

Thus the map $f\colon (a_0, \infty) \to \mathbb{R}^+$ given by $f(b) = F(bx)$ is monotone increasing, and its restriction to the rationals is linear. So f must be linear, and hence $F(ax) = aF(x)$. □

Given these claims, we are ready to prove our theorem.

Proof of Theorem 4. Fix any $x \in \mathcal{S}$, and assume $ax \in \mathcal{S}$ for some $a \in \mathbb{R}_+$. By Claim 2 it suffices to show that $F(ax) = aF(x)$. Let B be a ball with the property described in Claim 1, and denote its center by x_0 and its radius by r. As in Claim 3, let $\mathcal{B}$ be the semigroup given by $\cup_{a\geq 1} aB$; note that $\mathcal{B} \subseteq \mathcal{S}$. Then there is some y such that $x+y, a(x+y), y, ay \in \mathcal{B}$;

in fact, we can take $y = bx_0$ for $b = \max\{a, 1/a, |x|/r\}$ (see Figure 4). Then, on the one hand, by additivity,

$$F(ax + ay) = F(ax) + F(ay).$$

On the other hand, since $x + y, a(x + y), y, ay \in \mathcal{B}$,and since, by Claim 3, the restriction of F to $\mathcal{B}$ is linear, we have that

$$F(ax + ay) = F(a(x + y)) = aF(x + y) = aF(x) + aF(y) = aF(x) + F(ay),$$

thus

$$F(ax) + F(ay) = aF(x) + F(ay)$$

and so $F(ax) = aF(x)$. □

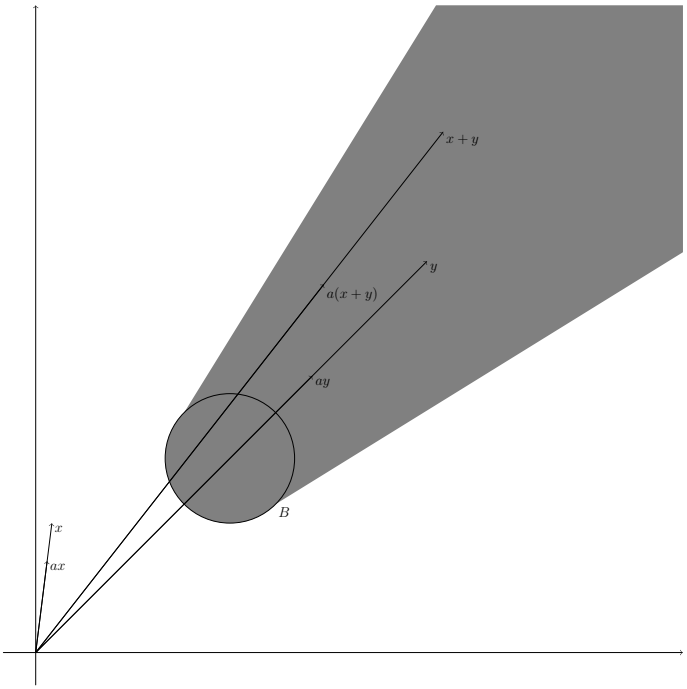

Figure 4: An illustration of the proof of Theorem 4.

C Proof of Theorem 1

Throughout this section we maintain the notation and terminology introduced in §B. It follows from the results in §B.1 that a LLR cost satisfies Axioms 1-4. For the rest of this section, we denote by C a cost function that satisfies the axioms. Let N be such that C is uniformly continuous with respect to the distance d_N. We use the same N to define the set $A = \{0, \ldots, N\}^n \backslash \{0, \ldots, 0\}$ introduced in §B.3.

Lemma 7. *Let μ and ν be two experiments that induce the same vector $\sigma \in \mathcal{A}$. Then $C(\mu) = C(\nu)$.*

Proof. Conditional on each $k \in \Theta$, the two experiments induce the same distribution for $(\ell_{0i})_{i\in\Theta}$. Because $\ell_{ij} = \ell_{i0} - \ell_{j0}$ almost surely, it follows that conditional on each state the two experiments induce the same distribution over the vector of all log-likelihood ratios $(\ell_{ij})_{i,j\in\Theta}$. Hence, $\bar{\mu}_i = \bar{\nu}_i$ for every i. Hence, by Lemma 1 the two experiments are equivalent in the Blackwell order. The result now follows directly from Axiom 1. □

Lemma 7 implies we can define a function $c : \mathcal{A} \to \mathbb{R}_+$ as $c(\sigma) = C(\mu)$ where μ is an experiment inducing σ.

Lemma 8. *Consider two experiments $\mu = (S, (\mu_i))$ and $\nu = (T, (\nu_i))$ inducing σ and τ in $\mathcal{A}$, respectively. Then*

1. *The experiment $\mu \otimes \nu$ induces the vector $(\sigma_0 * \tau_0, \ldots, \sigma_n * \tau_n) \in \mathcal{A}$;*
2. *The experiment $\alpha \cdot \mu$ induces the measure $\alpha\sigma + (1-\alpha)\delta_{\mathbf{0}}$.*

Proof. (1) For every $E \subseteq \mathbb{R}^n$ and every state i,

$$\begin{aligned} & (\mu_i \times \nu_i)\left(\{(s,t) : (\ell_1(s,t), \ldots \ell_n(s,t)) \in E\}\right) \\ = \; & (\mu_i \times \nu_i)\left(\left\{(s,t) : \left(\log \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_0}(s) + \log \frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_0}(t), \ldots, \log \frac{\mathrm{d}\mu_n}{\mathrm{d}\mu_0}(s) + \log \frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_n}(t)\right) \in E\right\}\right) \\ = \; & (\sigma_i * \tau_i)(E) \end{aligned}$$

where the last equality follows from the definition of σ_i and τ_i. This concludes the proof of the claim.

(2) Immediate from the definition of $\alpha \cdot \mu$. □

Lemma 9. *The function $c : \mathcal{A} \to \mathbb{R}$ satisfies, for all $\sigma, \tau \in \mathcal{A}$ and $\alpha \in [0,1]$:*

1. *$c(\sigma_0 * \tau_0, \ldots, \sigma_n * \tau_n) = c(\sigma) + c(\tau)$;*
2. *$c(\alpha\sigma + (1-\alpha)\delta_{\mathbf{0}}) = \alpha c(\sigma)$.*

Proof. (1) Suppose μ induces σ and ν induces τ. Then $C(\mu) = c(\sigma), C(\nu) = c(\tau)$ and, by Axiom 2 and Lemma 8, $c(\sigma_0 * \tau_0, \ldots, \sigma_n * \tau_n) = C(\mu \otimes \nu) = c(\sigma) + c(\tau)$. Claim (2) follows directly from Axiom 3 and Lemma 8. □

Lemma 10. *If $\sigma, \tau \in \mathcal{A}$ satisfy $m_\sigma = m_\tau$ then $c(\sigma) = c(\tau)$.*

Proof. Let μ be and ν be two experiments inducing σ and τ, respectively. Let $\mu^{\otimes r} = \mu \otimes \ldots \otimes \mu$ be the experiment obtained as the r-th fold independent product of μ. Axioms 2 and 3 imply

$$C((1/r) \cdot \mu^{\otimes r}) = C(\mu) \quad \text{and} \quad C((1/r) \cdot \nu^{\otimes r}) = C(\nu)$$

In order to show that $C(\mu) = C(\nu)$ we now prove that $C((1/r)\cdot\mu^{\otimes r}) - C((1/r)\cdot\nu^{\otimes r}) \to 0$ as $r \to \infty$. To simplify the notation let, for every $r \in \mathbb{N}$,

$$\mu[r] = (1/r)\cdot\mu^{\otimes r} \quad \text{and} \quad \nu[r] = (1/r)\cdot\nu^{\otimes r}$$

Let $\sigma[r] = (\sigma[r]_0, \ldots, \sigma[r]_n)$ and $\tau[r] = (\tau[r]_0, \ldots, \tau[r]_n)$ in $\mathcal{A}$ be the vectors of measures induced by $\mu[r]$ and $\nu[r]$.

We claim that $d_N(\mu[r], \nu[r]) \to 0$ as $r \to \infty$. First, notice that $\overline{\mu[r]}_i$ and $\overline{\nu[r]}_i$ assign probability $(r-1)/r$ to the zero vector $\mathbf{0} \in \mathbb{R}^{(n+1)^2}$. Hence

$$d_{tv}(\overline{\mu[r]}_i, \overline{\nu[r]}_i) = \sup_E \frac{1}{r}\left|\overline{\mu^{\otimes r}}_i(E) - \overline{\nu^{\otimes r}}_i(E)\right| \leq \frac{1}{r}.$$

For every $\alpha \in A$ we have

$$M_i^{\mu[r]}(\alpha) = \int \ell_{10}^{\alpha_1} \ldots \ell_{n0}^{\alpha_n}\, \mathrm{d}\mu[r]_i = \int_{\mathbb{R}^n} \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n}\, \mathrm{d}\sigma[r]_i(\xi) = m_{\sigma[r]_i}(\alpha) \tag{14}$$

We claim that $m_{\sigma[r]} = m_{\tau[r]}$. Theorem 2 shows the existence of a bijection $H : \mathcal{M} \to \mathcal{K}$ such that $H(m_\upsilon) = \kappa_\upsilon$ for every $\upsilon \in \mathcal{A}$. The experiment $\mu^{\otimes r}$ induces the vector $(\sigma_0^{*r}, \ldots, \sigma_n^{*r}) \in \mathcal{A}$, where σ_i^{*r} denotes the r-th fold convolution of σ_i with itself. Denote such a vector as σ^{*r}. Let $\tau^{*r} \in \mathcal{A}$ be the corresponding vector induced by $\nu^{\otimes r}$. Thus we have $\kappa_\sigma = H(m_\sigma) = H(m_\tau) = \kappa_\tau$, and

$$H(m_{\mu^{*r}}) = \kappa_{\sigma^{*r}} = (\kappa_{\sigma_0}^{*r}, \ldots, \kappa_{\sigma_n}^{*r}) = (r\kappa_{\sigma_0}, \ldots, r\kappa_{\sigma_n}) = r\kappa_\sigma = r\kappa_\tau = \kappa_{\tau^{*r}} = H(m_{\tau^{*r}})$$

Hence $m_{\sigma^{*r}} = m_{\tau^{*r}}$. It now follows from

$$m_{\sigma[r]_i}(\alpha) = \frac{1}{r} m_{\sigma_i^{*r}}(\alpha) + \frac{r-1}{r} 0$$

that $m_{\sigma[r]} = m_{\tau[r]}$, concluding the proof of the claim.

Equation (14) therefore implies that $M_i^{\mu[r]}(\alpha) = M_i^{\nu[r]}(\alpha)$. Thus

$$d_N(\mu[r], \nu[r]) = \max_i d_{tv}(\overline{\mu[r]}_i, \overline{\nu[r]}_i) \leq \frac{1}{r}.$$

Hence $d_N(\mu[r], \nu[r])$ converges to 0. Since C is uniformly continuous, then $C(\mu[r]) - C(\nu[r]) = 0$. So, $C(\mu) = C(\nu)$.

□

Lemma 11. *There exists an additive function $F : \mathcal{K} \to \mathbb{R}$ such that $c(\sigma) = F(\kappa_\sigma)$.*

Proof. It follows from Lemma 10 that we can define a map $G : \mathcal{M} \to \mathbb{R}$ such that $c(\sigma) = G(m_\sigma)$ for every $\mu \in \mathcal{A}$. We can use Theorem 2 to define a bijection $H : \mathcal{M} \to \mathcal{K}$

such that $H(m_\sigma) = \kappa_\sigma$. Hence $F = G \circ H^{-1}$ satisfies $c(\sigma) = F(\kappa_\sigma)$ for every σ. For every $\sigma, \tau \in \mathcal{A}$, Lemmas 8 and 9 imply

$$F(\kappa_\sigma) + F(\kappa_\tau) = c(\sigma) + c(\tau) = c(\sigma_0 * \tau_0, \ldots, \sigma_n * \tau_n) = F(\kappa_{\sigma_0 * \tau_0}, \ldots, \kappa_{\sigma_n * \tau_n}) = F(\kappa_\sigma + \kappa_\tau)$$

where the last equality follows from the additivity of the cumulants with respect to convolution. □

Lemma 12. *There exist* $(\lambda_{i,\alpha})_{i \in \Theta \setminus \{0\}, \alpha \in A}$ *in* $\mathbb{R}$ *such that*

$$c(\sigma) = \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \kappa_{\sigma_i}(\alpha) \quad \textit{for every} \quad \sigma \in \mathcal{A}.$$

Proof. As implied by Theorem 3, the set $\mathcal{K} \subseteq \mathbb{R}^d$ has nonempty interior. It is closed under addition, i.e. a subsemigroup. We can therefore apply Theorem 4 and conclude that the function F in Lemma 11 is linear. □

Lemma 13. *Let* $(\lambda_{i,\alpha})_{i \in \Theta \setminus \{0\}, \alpha \in A}$ *be as in Lemma 12. Then*

$$c(\sigma) = \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} m_{\sigma_i}(\alpha) \quad \textit{for every} \quad \sigma \in \mathcal{A}$$

Proof. Fix $\sigma \in \mathcal{A}$. Given $t \in (0,1)$, the Leonov-Shirayev identity implies

$$\begin{aligned} c(t\sigma + (1-t)\delta_{\mathbf{0}}) &= \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{(-1)^{q-1}}{q} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} \prod_{p=1}^{q} m_{t\sigma_i + (1-t)\delta_0}(\lambda^p) \right) \\ &= \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{(-1)^{q-1}}{q} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} t^q \prod_{p=1}^{q} m_{\sigma_i}(\lambda^p) \right) \\ &= \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{\lambda = (\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \rho(\lambda) t^q \prod_{p=1}^{q} m_{\sigma_i}(\lambda^p) \right) \end{aligned}$$

where for every tuple $\lambda = (\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)$ we let

$$\rho(\lambda) = \frac{(-1)^{q-1}}{q} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!}$$

Lemma 9 implies $c(\sigma) = \frac{1}{t} c(t\mu + (1-t)\delta_0)$ for every t. Hence

$$c(\sigma) = \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{\lambda = (\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \rho(\lambda) t^{q-1} \prod_{p=1}^{q} m_{\sigma_i}(\lambda^p) \right) \text{ for all } t \in (0,1).$$

By considering the limit $t \downarrow 0$, we have $t^{q-1} \to 0$ whenever $q \neq 1$. Therefore

$$c(\sigma) = \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} m_{\sigma_i}(\alpha) \quad \text{for all } \sigma \in \mathcal{A}.$$

□

Lemma 14. *Let $(\lambda_{i,\alpha})_{i \in \Theta \setminus \{0\}, \alpha \in A}$ be as in Lemmas 12 and 13. Then, for every i, if $|\alpha| > 1$ then $\lambda_{i,\alpha} = 0$.*

Proof. Let $\gamma = \max \{|\alpha| : \lambda_{i,\alpha} \neq 0 \text{ for some } i\}$. Assume, as a way of contradiction, that $\gamma > 1$. Fix $\sigma \in \mathcal{A}$. Theorem 2 implies

$$\begin{aligned} c(\sigma) &= \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} m_{\sigma_i}(\alpha) \\ &= \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} \prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) \right) \end{aligned}$$

Let $\sigma^{*r} = (\sigma_0^{*r}, \ldots, \sigma_0^{*r})$, where each σ_i^{*r} is the r-th fold convolution of σ_i with itself. Hence, using the fact that $\kappa_{\sigma_i^{*r}} = r \kappa_{\sigma_i}$ for all $r \in \mathbb{N}$,

$$c(\sigma^{*r}) = \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} r^q \prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) \right) \tag{15}$$

By the additivity of c, $c(\sigma^{*r}) = r c(\sigma)$. Hence, because $\gamma > 1$, $c(\sigma^{*r})/r^\gamma \to 0$ as $r \to \infty$. Therefore, diving (15) by r^γ we obtain

$$\sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} r^{q-\gamma} \prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) \right) \to 0 \text{ as } r \to \infty. \tag{16}$$

We now show that (16) leads to a contradiction. By construction, if $(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)$ then $q \leq |\alpha|$. Hence $q \leq \gamma$ whenever $\lambda_{i,\alpha} \neq 0$. So, in equation (16) we have $r^{q-\gamma} \to 0$ for all $q < \gamma$. Hence (16) implies

$$\sum_{i \in \Theta} \sum_{\alpha \in A : |\alpha| = \gamma} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha), q = |\alpha|} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} \prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) \right) = 0.$$

If $q = \gamma$ and $\lambda_{i,\alpha} > 0$ then $\gamma = |\alpha|$. In this case, in order for $\lambda = (\lambda^1, \ldots, \lambda^q)$ to satisfy

$\sum_{p=1}^{q} \lambda^p = \alpha$, it must be that each λ^p is a unit vector. Every such λ satisfies[25]

$$\prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) = \left(\int_{\mathbb{R}^n} \xi_1 \,\mathrm{d}\sigma_i(\xi)\right)^{\alpha_1} \cdots \left(\int_{\mathbb{R}^n} \xi_n \,\mathrm{d}\sigma_i(\xi)\right)^{\alpha_n}$$

and

$$\sum_{(\lambda^1,\ldots,\lambda^q)\in\Lambda(\alpha),q=|\alpha|} \frac{1}{q!}\frac{\alpha!}{\lambda^1!\cdots\lambda^q!} = \sum_{(\lambda^1,\ldots,\lambda^q)\in\Lambda(\alpha),q=|\alpha|} \frac{\alpha!}{|\alpha|!} = 1$$

so we obtain that

$$\sum_{i\in\Theta}\sum_{\alpha\in A:|\alpha|=\gamma} \lambda_{i,\alpha}\left(\int_{\mathbb{R}^n} \xi_1 \,\mathrm{d}\sigma_i(\xi)\right)^{\alpha_1} \cdots \left(\int_{\mathbb{R}^n} \xi_n \,\mathrm{d}\sigma_i(\xi)\right)^{\alpha_n} = 0. \tag{17}$$

By replicating the argument in the proof of Lemma 6 we obtain that the set

$$\left\{\left(\int_{\mathbb{R}^n} \xi_j \,\mathrm{d}\sigma_i(\xi)\right)_{i,j\in\Theta,j>0} : \sigma \in \mathcal{A}\right\} \subseteq \mathbb{R}^{(n+1)n}$$

contains an open set U. Consider now the function $f : \mathbb{R}^{(n+1)n} \to \mathbb{R}$ defined as

$$f(z) = \sum_{i\in\Theta}\sum_{\alpha\in A:|\alpha|=\gamma} \lambda_{i,\alpha} z_{i,1}^{\alpha_1}\cdots z_{i,n}^{\alpha_n}, \quad z \in \mathbb{R}^{(n+1)n}$$

Then (17) implies that f equals 0 on U. Hence, for every $z \in U$,$i \in \Theta$ and $\alpha \in A$ such that $|\alpha| = \gamma$,

$$\lambda_{i,\alpha} = \frac{\partial^\gamma}{\partial^{\alpha_1} z_{i,1}\cdots\partial^{\alpha_n} z_{i,n}} f(z) = 0$$

This contradicts the assumption that $\gamma > 1$ and concludes the proof. □

For every $j \in \{1,\ldots,n\}$ let $1_j \in A$ be the corresponding unit vector. We write λ_{ij} for $\lambda_{i,j}$. Lemma 14 implies that for every distribution $\sigma \in \mathcal{A}$ induced by an experiment $(S,(\mu_i))$, the function c satisfies

$$\begin{aligned}
c(\sigma) &= \sum_{i\in\Theta}\sum_{j\in\{1,\ldots,n\}} \lambda_{ij}\int_{\mathbb{R}^n} \xi_j \,\mathrm{d}\sigma_i(\xi) \\
&= \sum_{i\in\Theta}\sum_{j\in\{1,\ldots,n\}} \lambda_{ij}\int_S \log\frac{\mathrm{d}\mu_j}{\mathrm{d}\mu_0}(s)\,\mathrm{d}\mu_i(s) \\
&= \sum_{i\in\Theta}\sum_{j\in\{1,\ldots,n\}} \lambda_{ij}\int_S \log\frac{\mathrm{d}\mu_j}{\mathrm{d}\mu_0}(s) + \log\frac{\mathrm{d}\mu_0}{\mathrm{d}\mu_i}(s) - \log\frac{\mathrm{d}\mu_0}{\mathrm{d}\mu_i}(s)\,\mathrm{d}\mu_i(s)
\end{aligned}$$

[25] It follows from the definition of cumulant that for every unit vector $1_j \in \mathbb{R}^n$, $\kappa_{\sigma_i}(1_j) = \int_{\mathbb{R}^n} \xi_j \,\mathrm{d}\sigma_i(\xi)$.

Hence

$$\begin{aligned} c(\sigma) &= \sum_{i\in\Theta}\sum_{j\in\{1,\dots,n\}} \lambda_{ij}\int_S \log\frac{\mathrm{d}\mu_j}{\mathrm{d}\mu_i}\,\mathrm{d}\mu_i(s) + \sum_{i\in\Theta}\left(-\sum_{j\in\{1,\dots,n\}}\lambda_{ij}\right)\int_S \log\frac{\mathrm{d}\mu_0}{\mathrm{d}\mu_i}(s)\,\mathrm{d}\mu_i(s) \\ &= \sum_{i,j\in\Theta}\beta_{ij}\int_S \log\frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s)\,\mathrm{d}\mu_i(s) \end{aligned}$$

where in the last step, for every i, we set $\beta_{ij} = -\lambda_{ij}$ if $j \neq 0$ and $\beta_{i0} = \sum_{j\neq 0}\lambda_{ij}$.

It remains to show that the coefficients (β_{ij}) are positive and unique. Because C takes positive values, Lemma 2 immediately implies $\beta_{ij} \geq 0$ for all i, j. The same Lemma easily implies that the coefficients are unique given C.

D Proofs of other Results

Proof of Proposition 1. Consider a signal $(S, (\mu_i))$. Recall that by $\ell_i = \frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_0}$. The posterior probability of state i given a signal realizations s is, almost surely,

$$p_i(s) = \frac{q_i \mathrm{d}\mu_i}{\mathrm{d}\sum_{j\in\Theta}\mu_j}(s) = \frac{q_i\ell_i(s)}{\sum_{j\in\Theta} q_j\ell_j(s)}.$$

Thus $\frac{p_i(s)}{p_j(s)} = \frac{q_i\ell_i(s)}{q_j\ell_j(s)}$. We denote by $\bar{\mu} = \sum_{i\in\Theta} q_i\mu_i$ the unconditional distribution over S. Letting $\gamma_{ij} = \beta_{ij}/q_i$ we have

$$\begin{aligned} C(\mu) &= \sum_{i,j\in\Theta}\gamma_{ij}\,q_i\int_S \log\frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s)\,\mathrm{d}\mu_i(s) \\ &= \int_S \sum_{i,j\in\Theta}\gamma_{ij}\,\log\frac{\ell_i(s)}{\ell_j(s)} q_i\,\ell_i(s)\,\mathrm{d}\mu_0(s) \\ &= \int_S \sum_{i,j\in\Theta}\gamma_{ij}\,\log\left(\frac{p_i(s)q_j}{p_j(s)q_i}\right)\frac{q_i\,\ell_i(s)}{\sum_k q_k\,\ell_k(s)}\,\mathrm{d}\bar{\mu}(s) \end{aligned}$$

which equals

$$\begin{aligned} &\int_S \sum_{i,j\in\Theta}\gamma_{ij}\left[\log\frac{p_i(s)}{p_j(s)} - \log\frac{q_i}{q_j}\right]\underbrace{\frac{q_i\,\ell_i(s)}{\sum_k q_k\,\ell_k(s)}}_{p_i(s)}\,\mathrm{d}\bar{\mu}(s) \\ &= \int_S \sum_{i,j\in\Theta}\gamma_{ij}\,p_i(s)\log\frac{p_i(s)}{p_j(s)}\,\mathrm{d}\bar{\mu}(s) - \int_S \sum_{i,j\in\Theta}\gamma_{ij}\,p_i(s)\log\frac{q_i}{q_j}\,\mathrm{d}\bar{\mu}(s) \\ &= \int_S \sum_{i,j\in\Theta}\gamma_{ij}\,p_i\log\frac{p_i}{p_j}\,\mathrm{d}\pi_\mu(p) - \sum_{i,j\in\Theta}\beta_{ij}q_i\log\frac{q_i}{q_j}. \end{aligned}$$

The proof is then concluded by applying the definition of F. □

Proof of Proposition 4. We prove a slightly stronger result: Suppose $\min\{\beta_{ij}, \beta_{ji}\} \geq \frac{1}{d(i,j)^\gamma}$ for any $i, j \in \Theta$. Then for every action a, and every pair of states i, j,

$$\left|\mu_i^\star(a) - \mu_j^\star(a)\right| \leq \sqrt{\|u\|}\, d(i,j)^{\gamma/2}\,.$$

Clearly, the cost of the optimal experiment $C(\mu^\star)$ cannot exceed $\|u\|_\infty$. Thus for any action $\hat{a} \in A$ and any pair of states k, m

$$\begin{aligned}
\|u\| \geq C(\mu^\star) &= \sum_{i,j} \beta_{ij} \sum_{a\in A} \mu_i(a) \log \frac{\mu_i(a)}{\mu_j(a)} \\
&\geq \sum_{i,j} \min\{\beta_{ij}, \beta_{ji}\} \sum_{a\in A} \left(\mu_i(a) \log \frac{\mu_i(a)}{\mu_j(a)} + \mu_j(a) \log \frac{\mu_j(a)}{\mu_i(a)}\right) \\
&= \sum_{i,j} \min\{\beta_{ij}, \beta_{ji}\} \sum_{a\in A} |\mu_i(a) - \mu_j(a)| \times \left|\log\left(\frac{\mu_i(a)}{\mu_j(a)}\right)\right| \\
&\geq \sum_{i,j} \min\{\beta_{ij}, \beta_{ji}\} \, |\mu_i(\hat{a}) - \mu_j(\hat{a})| \times |\log \mu_i(\hat{a}) - \log \mu_j(\hat{a})|
\end{aligned}$$

Thus

$$\begin{aligned}
\|u\| &\geq \min\{\beta_{km}, \beta_{mk}\} \, |\mu_k(\hat{a}) - \mu_m(\hat{a})| \times |\log \mu_k(\hat{a}) - \log \mu_m(\hat{a})| \\
&\geq \min\{\beta_{km}, \beta_{km}\} \, |\mu_k(\hat{a}) - \mu_m(\hat{a})|^2 \\
&\geq \frac{1}{d(km)^\gamma} \, |\mu_k(\hat{a}) - \mu_m(\hat{a})|^2
\end{aligned}$$

□

Proof of Proposition 2. Let $|\Theta| = n$. By Axiom a there exists a function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\beta_{ij}^\Theta = f(|i-j|)$. Let $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ be given by $g(t) = f(t)t^2$. The Kullback-Leibler divergence between two normal distributions with unit variance and expectations i and j is $(i-j)^2/2$. Hence, by Axiom b there exists a constant $\kappa \geq 0$, independent of n, so that for each $\Theta \in \mathcal{T}$

$$\kappa = C^\Theta(\nu^\Theta) = \sum_{i\neq j\in\Theta} \beta_{ij}^\Theta \frac{(i-j)^2}{2} = \sum_{i\neq j\in\Theta} g(|i-j|). \tag{18}$$

We show that g must be constant, which will complete the proof. The case $n = 2$ is immediate, since then $\Theta = \{i, j\}$ and so (18) reduces to

$$\kappa = g(|i-j|).$$

For $n > 2$, let $\Theta = \{i_1, i_2, \ldots, i_{n-1}, x\}$ with $i_1 < i_2 < \cdots < i_{n-1} < x$. Then (18) implies

$$\kappa = \sum_{\ell=1}^{n-1} g(x - i_\ell) + \sum_{k=1}^{n-1} \sum_{\ell=1}^{k-1} g(i_k - i_\ell).$$

Taking the difference between this equation and the analogous one corresponding to $\Theta' = \{i_1, i_2, \ldots, i_{n-1}, y\}$ (with $y > i_{n-1}$) yields

$$0 = \sum_{\ell=1}^{n-1} g(x - i_\ell) - g(y - i_\ell).$$

Denoting $i_1 = -z$, we can write this as

$$0 = g(x + z) - g(y + z) + \sum_{\ell=2}^{n-1} g(x - i_\ell) - g(y - i_\ell).$$

Again taking a difference—this time of this equation with the analogous one obtained by setting $i_1 = -w$—we get

$$g(x + w) - g(y + w) = g(x + z) - g(y + z),$$

which by construction holds for all $x, y > -z, -w$. Consider in particular the case that $x, y > 0$, $w = 0$ and $z > 0$. Then

$$g(x) - g(y) = g(x + z) - g(y + z) \qquad \text{for all } x, y, z > 0. \tag{19}$$

Since g is non-negative, it follows from (18) that g is bounded by κ. Let

$$A = \sup_{t>0} g(t) \leq \kappa$$

and

$$B = \inf_{t>0} g(t) \geq 0.$$

For every $\varepsilon > 0$, there are some $x, y > 0$ such that $g(x) \geq A - \varepsilon/2$ and $g(y) \leq B + \varepsilon/2$, and so $g(x) - g(y) \geq A - B - \varepsilon$. By (19) it holds for all $z > 0$ that $g(x+z) - g(y+z) \geq A - B - \varepsilon$. For this to hold, since A and B are, respectively, the supremum and infimum of g, it must be that $g(x + z) \geq A - \varepsilon$ and that $g(y + z) \leq B - \varepsilon$ for every $z > 0$. By choosing z appropriately, it follows that $A - \varepsilon \leq g(\max\{x, y\} + 1) \geq B - \varepsilon$. Since this holds for any $\varepsilon > 0$, we have shown that $A = B$ and so g is constant. □

Proof of Proposition 3. As C is convex (Proposition 7) and the expected utility is linear in the choice probabilities we have that the objective function in (7) is concave. Thus, the solution to the optimization problem given (7) is characterized by a first order condition. Taking the first order condition from (7) yields that there exists Lagrange multipliers $\lambda \in \mathbb{R}_+^{|\Theta|}$ such that for every state i and every action a

$$0 = q_i u_i(a) - \lambda_i - \sum_{j \neq i} \left\{ \beta_{ij} \left[\log \left(\frac{\mu_i(a)}{\mu_j(a)} \right) - 1 \right] - \beta_{ji} \frac{\mu_j(a)}{\mu_i(a)} \right\} . \tag{20}$$

Subtracting (20) evaluated at a' from (20) evaluated at a yields that (8) is a necessary condition for the optimality of μ. □

References

M. D. Andrew Caplin and J. Leahy. Rational inattentive behavior: Characterizing and generalizing shannon entropy. Technical report, National Bureau of Economic Research, 2017.

K. J. Arrow. The value of and demand for information. *Decision and organization*, 2: 131–139, 1971.

K. J. Arrow. Informational structure of the firm. *The American Economic Review*, 75(2): 303–307, 1985.

K. J. Arrow, D. Blackwell, and M. A. Girshick. Bayes and minimax solutions of sequential decision problems. *Econometrica, Journal of the Econometric Society*, pages 213–244, 1949.

T. D. Austin. Entropy and Sinai theorem. *mimeo*, 2006.

D. Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. The Regents of the University of California, 1951.

H. F. Bohnenblust, L. S. Shapley, and S. Sherman. Reconnaissance in game theory. 1949.

A. Cabrales, O. Gossner, and R. Serrano. Entropy and the value of information for investors. *American Economic Review*, 103(1):360–77, 2013.

A. Cabrales, O. Gossner, and R. Serrano. A normalized value for information purchases. *Journal of Economic Theory*, 170:266–288, 2017.

A. Caplin. Measuring and modeling attention. *Annual Review of Economics*, 8:379–403, 2016.

A. Caplin and M. Dean. Behavioral implications of rational inattention with shannon entropy. Technical report, National Bureau of Economic Research, 2013.

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

A. Caplin, M. Dean, and J. Leahy. Rational inattention, optimal consideration sets and stochastic choice. Technical report, Working paper, 2016.

C. P. Chambers, C. Liu, and J. Rehbeck. Nonseparable costly information acquisition and revealed preference, 2017.

J. Chan, A. Lizzeri, W. Suen, and L. Yariv. Deliberating collective decisions. *The Review of Economic Studies*, 85(2):929–963, 2017.

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

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

H. de Oliveira. Axiomatic foundations for entropic costs of attention. Technical report, Mimeo, 2014.

H. De Oliveira, T. Denti, M. Mihm, and K. Ozbek. Rationally inattentive preferences and hidden information costs. *Theoretical Economics*, 12(2):621–654, 2017.

M. Dean and N. Neligh. Experimental tests of rational inattention, 2017.

T. Denti. Posterior-separable cost of information, 2018.

A. Dvoretzky, J. Kiefer, J. Wolfowitz, et al. Sequential decision problems for processes with continuous time parameter. testing hypotheses. *The Annals of Mathematical Statistics*, 24(2):254–264, 1953.

B. Ebanks, P. Sahoo, and W. Sander. *Characterizations of information measures*. World Scientific, 1998.

A. Ellis. Foundations for optimal inattention. *Journal of Economic Theory*, 173:56–94, 2018.

A. Frankel and E. Kamenica. Quantifying information and uncertainty. Technical report, Working paper, 2018.

G. A. Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 3 edition, 1997.

D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*. New York : Wiley, 1966. ISBN 0882751395. Includes indexes. Bibliography: p. 437-486.

L. Hansen and T. J. Sargent. Robust control and model uncertainty. *American Economic Review*, 91(2):60–66, 2001.

B. Hébert and M. Woodford. Rational inattention with sequential information sampling, 2017.

T. Jech. *Set theory*. Springer Science & Business Media, 2013.

P. Kannappan and P. Rathie. An axiomatic characterization of j-divergence. In *Transactions of the Tenth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 29–36. Springer, 1988.

J. Kelly. A new interpretation of information rate. *bell system technical journal*, 1956.

I. Krajbich, C. Armel, and A. Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, 13(10):1292, 2010.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

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

V. Leonov and A. N. Shiryaev. On a method of calculation of semi-invariants. *Theory of Probability & its applications*, 4(3):319–329, 1959.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

B. Mackowiak, F. Matějka, and M. Wiederholt. Rational inattention: A disciplined behavioral model, 2018.

J. Marschak. Remarks on the economics of information. Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.

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

L. Mattner. What are cumulants? *Documenta Mathematica*, 4:601–622, 1999.

L. Mattner. Cumulants are universal homomorphisms into hausdorff groups. *Probability theory and related fields*, 130(2):151–166, 2004.

J. Mensch. Cardinal representations of information, 2018.

S. Morris and P. Strack. The wald problem and the relation of sequential sampling and static information costs, 2018.

S. Morris and M. Yang. Coordination and continuous choice, 2016.

F. Mosteller and P. Nogee. An experimental measurement of utility. *Journal of Political Economy*, 59(5):371–404, 1951.

A. N. Shiryaev. *Probability*. Springer, 1996.

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

C. A. Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3): 665–690, 2003.

J. Steiner, C. Stewart, and F. Matějka. Rational inattention dynamics: Inertia and delay in decision-making. *Econometrica*, 85(2):521–553, 2017.

T. Strzalecki. Axiomatic foundations of multiplier preferences. *Econometrica*, 79(1):47–73, 2011.

G. Tavares, P. Perona, and A. Rangel. The attentional drift diffusion model of simple perceptual decision-making. *Frontiers in neuroscience*, 11:468, 2017.

S. Van Nieuwerburgh and L. Veldkamp. Information immobility and the home bias puzzle. *The Journal of Finance*, 64(3):1187–1215, 2009.

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

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

R. Wilson. Informational economies of scale. *The Bell Journal of Economics*, pages 184–195, 1975.

E. Zanardo. How to measure disagreement. Technical report, 2017.

W. Zhong. Optimal dynamic information acquisition, 2017a.

W. Zhong. Optimal information acquisition with linear waiting cost, 2017b.