The Cost of Information*

Luciano Pomatto[†] Philipp Strack[‡] Omer Tamuz[§]

February 4, 2019

Abstract

We develop an axiomatic theory of information acquisition that captures the idea of constant marginal costs in information production: the cost of generating two independent signals is the sum of their costs, and generating a signal with probability half costs half its original cost. Together with a monotonicity and a continuity conditions, these axioms determine the cost of a signal up to a vector of parameters. These parameters have a clear economic interpretation and determine the difficulty of distinguishing states. We argue that this cost function is a versatile modeling tool that leads to more realistic predictions than mutual information.

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

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

[†] Caltech. Email: luciano@caltech.edu.

[‡] UC Berkeley. Email: pstrack@berkeley.edu.

[§] Caltech. Email: tamuz@caltech.edu. Omer Tamuz was supported by a grant from the Simons Foundation (#419427).

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 with probability one.

As an example, the second axiom implies that the cost of collecting n independent random samples (for example by surveying n customers) 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 one.

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.

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 economic environment 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 interpretation, making it easy to judge for which applications it is appropriate.

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[1]

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

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 μ on average produces evidence that strongly favors state 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 distinguishing 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 set of states Θ in the particular information acquisition problem; this is a feature that is commonly assumed in models that exogenously restrict attention to normal signals.

Under these assumptions (Axioms a and b) Proposition 3 shows that there exists a 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

[1] 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).

choice of parameters offers a simple and tractable 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.[2] 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 LLR cost (1) of an experiment can be represented as the expected change of the function

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

from the prior q to the posterior p induced by the signal.[3] That is, the cost of an experiment equals

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

This alternative formulation makes it possible to apply techniques and insights derived for posterior-separable costs functions (Caplin and Dean, 2013; Caplin, Dean, and Leahy, 2018).

Relation to Mutual Information Cost. Following Sims' seminal work on rational inattention, cost functions based on mutual information have been commonly applied to model costly information processing (Sims, 2003, 2010). Mackowiak, Matějka, and Wiederholt (2018) review the literature on rational inattention. Mutual information costs are defined as the expected change

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

of the Shannon entropy $H(p) = -\sum_{i \in \Theta} p_i \log p_i$ 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 *generating* information. Due to this 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

[2] 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.

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

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 producing this 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 mutual information 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 $\mu = (S, (\mu_i)_{i\in\Theta})$ consists of a set S of signal realizations equipped with a sigma-algebra Σ, and, for each state $i \in \Theta$, a probability measure μ_i defined on (S, Σ). The set S represents the possible outcomes of the experiment, and each measure μ_i describes the distribution of outcomes when the true state is i.

We assume throughout that the measures (μ_i) are mutually absolutely continuous, so that each derivative (i.e. ratio between densities) $\frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}$ 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). This assumption means that no signal can ever rule out any state, and in particular can never completely reveal the true state.

Given an experiment μ, 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. The distribution of L depends on the true state generating the data. Given an experiment μ, we denote by $\bar{\mu}_i$ the distribution of L conditional on state i.[4]

We restrict our attention to signals where the induced log-likelihoods ratios (ℓ_{ij}) have finite moments. That is, experiments such that for every state i and every integral vector $\alpha \in \mathbb{N}^\Theta$ the expectation $\int_S |\prod_{k\neq i} \ell_{ik}^{\alpha_k}|\mathrm{d}\mu_i$ is finite. We denote by $\mathcal{E}$ the class of all such experiments.[5] The restriction to $\mathcal{E}$ is a technical condition that rules out experiments whose log-likelihood ratios have very heavy tails, but, to the best of our knowledge, includes all

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

[5] 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, p. 5).

(not fully revealing) experiments commonly used in applications. In particular, we do not restrict our attention to a parametric family of experiments such as normally distributed signals.

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

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

assigning to each experiment $\mu \in \mathcal{E}$ its cost $C(\mu)$. In the next section we introduce and characterize four basic properties for information cost functions.

2.1 Axioms

Our first axiom postulates that the cost of an 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.[6] Let μ and ν be two experiments, inducing the distributions over posteriors π_μ and π_ν given the uniform prior q. Then μ dominates ν in the Blackwell 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 (weakly) 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 (see, for example, Lemma 1 in the Appendix).

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, it will turn out that the cost function axiomatized in this paper will satisfy the stronger property of Blackwell monotonicity (see Proposition 1).

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. If experiments are optimally chosen by a

[6] 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.

decision maker then 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 product of the two measures.[7] 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 independent samples, one for each population, are collected.

Our second axiom states that the cost function is additive with respect to independent experiments:

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

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

An immediate implication of Axioms 1 and 2 is that a completely uninformative signal has zero cost. This follows from the fact that an uninformative experiment μ is Blackwell equivalent to the product experiment $\mu \otimes \mu$.

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 $\alpha \in [0, 1]$. We define

$$\alpha \cdot \mu = (S \cup \{o\}, (\nu_i)),$$

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:*

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

Our final assumption is a continuity condition. We first introduce a (pseudo)-metric over $\mathcal{E}$. Recall that for every experiment μ, $\bar{\mu}_i$ denotes its distribution of log-likelihood

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

ratios conditional on state i. We denote by d_{tv} the total-variation distance.[8] Given a vector $\alpha \in \mathbb{N}^{\Theta}$, let $M_i^{\mu}(\alpha) = \int_S |\prod_{k \neq i} \ell_{ik}^{\alpha_k}| d\mu_i$ be the α-moment of the vector of log-likelihood ratios $(\ell_{ik})_{k \neq i}$. 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.[9]

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.[10] 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

[8] 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 $\mathbb{R}^{\Theta \times \Theta}$.

[9] 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 clarity, we maintain the two axioms as separate throughout the paper.

[10] 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).

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. In an extended framework where the decision maker is allowed to randomize her choice of experiment, 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. Now 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.[11] 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}). The expression $\int_S \log(\mathrm{d}\mu_i/\mathrm{d}\mu_j)\mathrm{d}\mu_i$ is the Kullback-Leibler divergence $D_{\mathrm{KL}}(\mu_i \| \mu_j)$ between the two distributions, a well understood and tractable measure of informational content (Kullback and Leibler, 1951). This implies that (3) can alternatively be formulated as

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

[11] 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.

A higher value of $D_{\mathrm{KL}}(\mu_i \| \mu_j)$ describes an experiment which, conditional on state i, produces stronger evidence in favor of state i compared to j, as represented by a higher expected value of the log-likelihood ratio $\mathrm{d}\mu_i/\mathrm{d}\mu_j$. The coefficient β_{ij} thus measures the *marginal cost* of increasing the expected log-likelihood ratio between states i and j, conditional on i, while keeping all other expected log-likelihood ratios fixed.[12]

The specification of the parameters (β_{ij}) must of course depend on the particular application under consideration. Consider, for instance, a doctor who must choose a treatment for a patient displaying a set of symptoms, and who faces uncertainty regrading their cause. In this example, the state of the world i represents the pathology affecting the patient. In order to distinguish between two possible diseases i and j it is necessary to collect samples and run tests, whose costs will depend on factors that are specific to the two conditions, such as their similarity, or the prominence of their physical manifestations. These difference in costs can then be reflected by the coefficients β_{ij} and β_{ji}. For example, if i and j are two types of viral infections, and k is a bacterial infection, then $\beta_{ij} > \beta_{ik}$ if it is harder to tell apart the two viral infection than to tell apart a viral infection from a bacterial one. In §6 we discuss environments where the coefficients might naturally be asymmetric, in the sense that $\beta_{ij} \neq \beta_{ji}$.

In environments where no pair of states is a priori harder to distinguish than another,[13] a natural choice is to set all the coefficients (β_{ij}) to be equal. Finally, in §4 we propose a specific functional form in the more structured case where states represent a one-dimensional quantity.

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.

Normal Signals. Consider a normal experiment $\mu^{m,\sigma}$ according to which the signal s is given by

$$s = m_i + \varepsilon$$

where the mean $m_i \in \mathbb{R}$ depends on the true state i, and ε is state independent and normally distributed with standard deviation σ. By substituting (3) with the well-known expression for the Kullback-Leibler divergence between normal distributions, we obtain

[12] As we formally show in Lemma 2 in the Appendix, this operation of increasing a single expected log-likelihood ratio while keeping all other expectations fixed is well-defined: for every experiment μ and every $\varepsilon > 0$, if $D_{\mathrm{KL}}(\mu_i \| \mu_j) > 0$ then there exists a new experiment ν such that $D_{\mathrm{KL}}(\nu_i \| \nu_j) = D_{\mathrm{KL}}(\mu_i \| \mu_j) + \varepsilon$, and all other divergences are equal. Hence the difference in cost between ν and the experiment μ is given by β_{ij} times the difference ε in the expected log-likelihood ratio. The result formally justifies the interpretation of each coefficient β_{ij} as a marginal cost.

[13] An example is that of a country that faces uncertainty regarding which of its political rivals is responsible for a cyber attack.

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 squared distance between the two states.

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.

In the above examples more informative experiments are more costly. This is true for for normal signals, since the cost is decreasing in σ, and for binary signals, where the cost is increasing in p. The next result establishes that the a LLR cost function is monotone with respect to the Blackwell order:

Proposition 1. *Let μ and ν be experiments such that μ Blackwell dominates ν. Then every LLR cost C satisfies $C(\mu) \geq C(\nu)$.*

Bayesian Representation. The framework we considered so far makes no references to subjective beliefs over the states of nature. Nevertheless, a LLR cost function can be easily embedded in a standard Bayesian framework. Consider, to illustrate, a decision maker endowed with a prior $q \in \mathcal{P}(\Theta)$. Each 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 2. *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,*

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

In this representation the cost of the experiment μ is expressed as the expected change of the function F from the prior q to the realized posterior p. Each coefficient β_{ij} is normalized by the prior probability of the state q_i.

Representations of the form (6) have been studied in the literature under the name of "posterior separable" (Caplin, Dean, and Leahy, 2018, Definition 5). For example, Sims' mutual information cost has the same functional form, but where $F(p)$ is replaced by the Shannon entropy $H(p) = -\sum_i p_i \log p_i$. An important implication of Theorem 2 is that general techniques for posterior separable costs functions, as developed by Caplin and Dean (2013), can be applied to the LLR cost function.

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 necesarrily be context dependent.

A commonly encountered context is that of learning about a one-dimensional characteristic, so that each state i is a real number.[14] 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 a 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. Given a problem where each state $i \in \Theta \subset \mathbb{R}$ is a real number, we propose to set 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. Therefore, under this specification, two states that are closer to each other are harder to distinguish.

The main result of this section shows that this choice of parameters captures two main hypotheses: (a) 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 (b) the cost of a noisy measurement of the state with standard normal error is the same across information acquisition problems. Both assumptions express the idea that the cost of making a measurement depends only on its precision, and not on the other details of the model, such as the set of states Θ. For example, the cost of measuring a person's height should depend only on the precision of the measurement instrument, but not on what modeling assumptions are made about the set of possible heights.

We denote by $\mathcal{T}$ the collection of finite subsets of $\mathbb{R}$ with at least two elements. Each set $\Theta \in \mathcal{T}$ represents the set of states of nature in a different, one-dimensional, information acquisition problem. To simplify the language, we refer to each Θ as a *problem.* For each $\Theta \in \mathcal{T}$ we are given an LLR cost function C^{Θ} with coefficients (β_{ij}^{Θ}). The next two axioms formalize the two hypotheses described above by imposing restrictions, across problems,

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

on the cost of information.

The first axiom states that β^{Θ}_{ij}, 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 unaffected by changing the values of the other states.

Axiom a. *For all $\Theta, \Xi \in \mathcal{T}$ such that $|\Theta| = |\Xi|$, and for all $i, j \in \Theta$ and $k, l \in \Xi$,*

$$\textit{if} \quad |i - j| = |k - l| \quad \textit{then} \quad \beta^{\Theta}_{ij} = \beta^{\Xi}_{kl}.$$

For each $i \in \mathbb{R}$ we denote by ζ_i a normal probability measure on the real line with mean i and variance 1. Given a problem Θ, we denote by ζ^{Θ} the experiment $(\mathbb{R}, (\zeta_i)_{i \in \Theta})$. Hence, ζ^{Θ} is the canonical experiment consisting of a noisy measurement of the state plus standard normal error.[15] 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 $\Theta, \Xi \in \mathcal{T}$, $C^{\Theta}(\zeta^{\Theta}) = C^{\Xi}(\zeta^{\Xi})$.*

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

Proposition 3. *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 $\Theta \in \mathcal{T}$,*

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

where n is the cardinality of Θ.

Proposition 3 implies that for any $\Theta \in \mathcal{T}$, a normal signal with mean i and variance σ^2 has cost $\kappa\sigma^{-2}$ proportional to its precision; this can be seen by applying (4), the expression for the cost of normal signals. Thus, the functional form given in Proposition 3 generalizes a specification often found in the literature, where the cost of a normal signal is assumed to be proportional to its precision (Wilson, 1975; Van Nieuwerburgh and Veldkamp, 2010) to arbitrary (non-normal) information structures.

5 Examples

5.1 Information Acquisition in Decision Problems

We now study the log-likelihood ratio cost in the context of decision problems. We consider a decision maker choosing an action a from a finite set A of actions. The payoff from a

[15] Expressed differently, if $i \in \Theta$ is the true state, then the outcome of the experiment ζ^{Θ} is distributed as $s = i + \epsilon$, where ϵ is normally distributed with mean zero and variance 1 independent of the state.

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

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

Hence, action a is chosen in state i with probability $\mu_i^\star(a)$. The maximization problem (7) is strictly concave, provided all coefficients (β_{ij}) are strictly positive (Proposition 7 in the Appendix). Thus, it admits a unique solution.

First Order Conditions. Denote the support of μ of by $\operatorname{supp}(\mu)$: this is the set of actions which are played with strictly positive probability under μ.[16] The next result characterizes the optimal choice probabilities under the LLR cost:

Proposition 4. *Assume that $\beta_{ij} \neq 0$ for all $i \neq j$. Let $\mu = (\mu_i)_{i \in \Theta}$ be a state-dependent distribution over actions which solves the optimization problem (7). Then, for every state $i \in \Theta$ and every pair of actions $a_1, a_2 \in \operatorname{supp}(\mu)$ it holds that*

$$q_i \left[u(i,a_1) - u(i,a_2) \right] = \tilde{c}(i,a_1) - \tilde{c}(i,a_2) \tag{8}$$

where

$$\tilde{c}(i,a) = -\sum_{j \neq i} \left[\beta_{ij} \log \frac{\mu_j(a)}{\mu_i(a)} + \beta_{ji} \frac{\mu_j(a)}{\mu_i(a)} \right] .$$

Condition (8) can be interpreted as follows. The expression $q_i \left[u(i,a_1) - u(i,a_2) \right]$ measures the expected benefit of choosing action a_1 instead of a_2 in state i. Up to an additive constant, $\tilde{c}(i,a)$ is the informational cost of choosing action a marginally more often in state i. This marginal cost is increasing in the probability $\mu_i(a)$, due to the convexity of C. Hence the right-hand-side of (8) measures the change in information acquisition cost necessary to choose action a_1 marginally more often and action a_2 marginally less often.

An Application to Perception Tasks. Consider a perception task (see, e.g. Dean and Neligh, 2017) where subjects observe 100 dots of different colors on a screen. Each dot is

[16] $\operatorname{supp}(\mu) = \{a \in A \colon \mu_i(a) > 0 \text{ for some } i \in \Theta\}$.

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.[17]

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

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

For a tuple of distributions over actions $(\mu_i)_{i \in \Theta}$, 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[18], neuroscience[19], economics[20] 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 when the number of blue dots increases.

As Dean and Neligh (2017) note, under mutual information cost (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*.[21] 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 the state in which there are 99 blue dots.

This unrealistic prediction is driven by the fact that under mutual information the states

[17] This means that the prior is $q_i = \frac{1}{2r}$, for $i \in \Theta$.

[18] See, e.g., Chapter 7 in Green and Swets (1966) or Chapter 4 in Gescheider (1997).

[19] E.g., Krajbich et al. (2010); Tavares et al. (2017).

[20] See, e.g., Mosteller and Nogee (1951).

[21] It is well known that under mutual information costs 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."

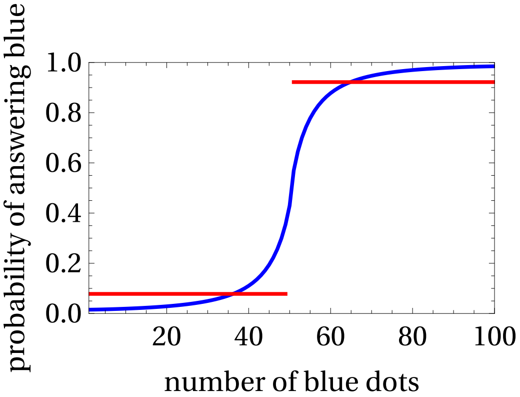

Figure 1: Predicted probability of guessing that there are more red dots as a function of the state for LLR cost with $\beta_{ij} = 1/(i-j)^2$ (in blue) and mutual information cost (in red).

are devoid of meaning and thus equally hard to distinguish. Indeed, the same conclusion holds for any cost function C in (7) that, like mutual information, 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 accounts for the difficulty of distinguishing different states through the coefficients β. As this is a one-dimensional information acquisition problem, we apply the specification $\beta_{ij} = \kappa/(i-j)^2$ of the LLR cost described in §4. As can be seen in Figure 1, the LLR cost predicts a sigmoidal relation between the state and the choice probabilities.

Continuous Choice. The main insight emerging from the above example is that under the LLR cost closer states are harder to distinguish, in the sense that acquiring information that finely discriminates 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. We assume that the state space Θ is endowed with a distance $d : \Theta \times \Theta \to \mathbb{R}$. In the previous example, d is simply the difference $|i-j|$ in the number of blue dots.

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

$$\min\{\beta_{ij}, \beta_{ji}\} \geq \frac{1}{d(i,j)^2} \,. \tag{9}$$

So, the cost of acquiring information that discriminates between states i and j is high for states that are close to each other.[22] 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. For this result, we

[22] As we show in the proof of the next proposition, the results of this section extend with minor variations to the case where the exponent in (9) is taken to be some $\gamma > 0$ rather than 2.

denote by $\|u\| = \max_{a,i} |u(a,i)|$ the norm of the decision maker's utility function.

Proposition 5 (Continuity of Choice)**.** *Suppose that nearby states are hard to distinguish. Then the optimal choice probabilities $\mu^\star$ solving (7) are uniformly Lipschitz continuous with constant $\sqrt{\|u\|}$, i.e. satisfy*

$$\left|\mu_i^\star(a) - \mu_j^\star(a)\right| \le \sqrt{\|u\|}\, d(i,j) \quad \textit{for all } a \in A \textit{ and } i,j \in \Theta. \tag{10}$$

Lipschitz continuity is a standard notion of continuity in discrete settings, such as the one of this paper, where the relevant variable i takes finitely many values. A crucial feature of the bound (10) is that the Lipschitz constant depends only on the norm $\|u\|$ of the utility function, independently of the exact form of the coefficients (β_{ij}), and of the number of states.[23]

This result highlights a contrast between the predictions of mutual information cost and LLR cost. Mutual information predicts behavior that displays counter-intuitive discontinuities with respect to the state. 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 them.

This difference has stark implications in coordination games. Morris and Yang (2016) study information acquisition in coordination problems. In their model, continuity of the choice probabilities with respect to the state leads to a unique equilibrium; if continuity fails, then there are multiple equilibria. This suggests that mutual information and LLR costs lead to very different predictions in coordination games and their economic applications (bank-runs, currency attacks, models of regime change, etc).

5.2 Acquiring Precise Information

In this section we use a simple example to illustrate how our additivity axiom captures constant marginal costs, 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 mutual information.

Consider, for instance, the classical problem of learning the bias of a coin by flipping it multiple times. In this context, mutual information 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 comparing the cost of observing a single coin flip versus a long sequence of coin flips.

[23] Proposition 5 suggests that the analysis of choices probabilities might be extended to the case where the set of states Θ is an interval in $\mathbb{R}$, or, more generally, a metric space. Given a (possibly infinite) state space Θ endowed with a metric, and a sequence of finite discretizations (Θ_n) converging to Θ, the bound (10) implies that if the corresponding sequence of choice probabilities converges, then it must converge to a collection of choice probabilities that are continuous, and moreover Lipschitz.

Under LLR cost, the additivity axiom implies that the cost of observing k coin flips is linear in k. Hence the cost of observing a sequence of k flips goes to infinity with k. Under mutual information cost with constant $\lambda > 0$ the cost of a single coin flip equals

$$\left[\{0.8 \log(0.8) + 0.2 \log(0.2)\} - \log \frac{1}{2}\right] \lambda \approx 0.19\, \lambda\,.$$

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

$$\lim_{p \to 1} \left[\{p \log p + (1-p) \log(1-p)\} - \log \frac{1}{2}\right] \lambda = \log(2)\, \lambda \approx 0.69\, \lambda\,.$$

Thus, the cost of observing infinitely many coin flips is only approximately 3.6 times the cost of observing a single coin flip. The low—and arguably in many applications unrealistic—cost of acquiring perfect information is caused by the sub-additivity of mutual information as a cost function, which contrasts with the additivity of the log-likelihood ratio cost we propose (see Figure 2).

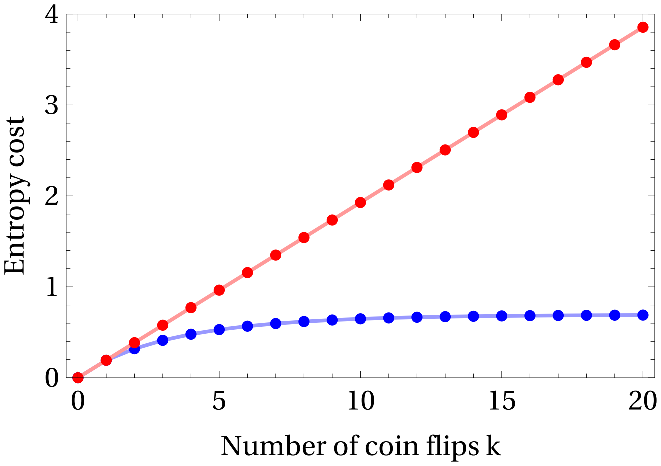

Figure 2: The LLR cost (in red) and the mutual information cost (in blue) of observing multiple independent coin flips/binary signals.

These simple calculations suggest that using Sims' mutual information cost as a model of information production rather than information processing (as originally intended by Sims) may lead to counterintuitive predictions.

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, it might lead to different predictions about whether investors tend to learn and ultimately invest in domestic or foreign stocks, as shown in Section 2.5 of Van Nieuwerburgh and Veldkamp (2010), for the case where signals are exogenously restricted to be normal.

5.3 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, i.e. whether the state is in a subset[24]

$$H \subset \Theta\,.$$

We consider an experiment that reveals with some probability whether the hypothesis is true or not, and study how its cost depends on the structure of H. For a given hypothesis H and a precision α consider the binary signal μ, with signal realizations $S = \{H, H^c\}$

$$\mu_i(s) = \begin{cases} \alpha & \text{for } i \in s \\ 1-\alpha & \text{for } i \notin s \end{cases} \tag{11}$$

Conditional on each state i, this experiment yields a correct signal with probability α. Under LLR cost, the cost of such a signal is given by

$$\left(\sum_{i \in H, j \in H^c} \beta_{ij} + \beta_{ji}\right)\left(\alpha \log \frac{\alpha}{1-\alpha} + (1-\alpha)\log\frac{1-\alpha}{\alpha}\right) \tag{12}$$

The first term captures the difficulty of discerning between H and H^c. The harder the states in H and H^c are to distinguish, the larger the coefficients β_{ij} and β_{ji} will be, and the more costly it will thus be to learn whether the hypothesis H is true. The second term is monotone in the signal precision α and is independent of the hypothesis.

Learning about the GDP. For concreteness, consider the case where the state is represented by a natural number i in the interval $\Theta = \{20000, \ldots, 80000\}$, representing, for instance, the current US GDP per capita. Consider the following two different hypotheses:[25]

(H1) The GDP is above 50000.

(H2) The GDP is an even number.

Intuitively, producing enough information to answer with high accuracy whether (H1) is true should be less expensive than producing enough information to answer whether (H2) is true, a practically impossible task. Our model captures this intuition. As the state is one-dimensional we set $\beta_{ij} = \kappa/(i-j)^2$, following §4. Then,

$$\sum_{i \in H1, j \in H1^c} \beta_{ij} + \beta_{ji} \approx 22\,\kappa \qquad\qquad \sum_{i \in H2, j \in H2^c} \beta_{ij} + \beta_{ji} \approx 148033\,\kappa.$$

[24] We denote the complement of H by $H^c = \Theta \setminus H$.

[25] Formally, $H1 = \{i \in \Theta \colon i > 50000\}$ and $H2 = \{i \in \Theta \colon i \text{ even}\}$

That is, learning whether the GDP is even or odd is by an order of magnitude more costly than learning whether the GDP is above or below 50000.

It is useful to compare these observations with the results that would be obtained under mutual information and a uniform prior on Θ. In such a model, the cost of a symmetric binary signal with precision α is determined solely by the cardinality of H. In particular, under mutual information learning whether the GDP is above or below 50000 is *equally* costly as learning whether it is even or odd. This follows from the fact that the mutual information cost is invariant with respect to a relabelling of the states.

This example demonstrates that the LLR cost function can capture different phenomena from mutual information 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, and capture the intuitive fact that measuring the most significant digit of the GDP is much easier than measuring the least significant one.

6 Verification and Falsification

It is well understood that verification and falsification are fundamentally different forms of empirical research. This can be seen most clearly through Karl Popper's famous example of the statement "all swans are white." Regardless of how many white swans are observed, no amount of evidence can imply that the next one will be white. However, observing a single black swan is enough to prove the statement false.

Popper's argument highlights a crucial asymmetry between verification and falsification. A given experiment, such as the observation of swans, can make it feasible to reject an hypothesis, yet have no power to prove that the same hypothesis is true.

This principle extends from science to everyday life. In a legal case, the type of evidence necessary to prove that a person is guilty can be quite different from the the type of evidence necessary to demonstrate that a person is innocent. In a similar way, corroborating the claim "Ann has a sibling" might require empirical evidence (such as the outcome of a DNA test) that is distinct from the sort of evidence necessary to prove that she has no siblings. These examples lead to the question of how to capture Popper's distinction between verification and falsification in a formal model of information acquisition.

In this section we show that the asymmetry between verification and falsification can be captured by the LLR cost. As an example, we consider a state space $\Theta = \{a, e\}$ that consists of two hypotheses. For simplicity, let a corresponds to the hypothesis "all swans are white" and e to the event "there exists a nonwhite swan." Imagine a decision maker who attaches equal probability to the each state, and consider the experiments described

	s_1	s_2
a	$1-\varepsilon^2$	ε^2
e	$1-\varepsilon$	ε

(a) Experiment I

	s_1	s_2
a	$1-\varepsilon$	ε
e	$1-\varepsilon^2$	ε^2

(b) Experiment II

Table 1: The set of states is $\Theta = \{a, e\}$. In both experiments $S = \{s, t\}$. Under experiment I, observing the signal realization t rejects the hypothesis that the state is a (up to a small probability of error ε^2). Under experiment II, observing t verifies the same hypothesis.

in Table 1:[26]

- In experiment I, regardless of the state, an uninformative signal realization s_1 occurs with probability greater than $1-\varepsilon$, where ε is positive and small. If a nonwhite swan exists, then one is observed with probability ε. Formally, this corresponds to observing the signal realization s_2. If all swans are white, then signal s_1 is observed, up to an infinitesimal probability of error ε^2. Hence, conditional on observing s_2, the decision maker's belief in state a approaches zero, while conditional on observing s_1 the decision maker's belief remains close to the prior. So, the experiment can reject the hypothesis that the state is a, but cannot verify it.[27]

- In experiment II the roles of the two states are reversed: if all swans are white, then this fact is revealed to the decision maker with probability ε. If there is a non-white swan, then the uninformative signal s_1 is observed (up to an infinitesimal probability of error ε^2). Conditional on observing s_2, the decision maker's belief in state a approaches one, and conditional on observing s_1 the decision maker's belief is essentially unchanged. Thus, the experiment can verify the hypothesis that the state is a, but cannot reject it.

As shown by the example, permuting the state-dependent distributions of an experiment may affect its power to verify or falsify an hypothesis. However, permuting the role of the states may, in reality, correspond to a completely different type of empirical investigation. For instance, experiment I can be easily implemented in practice: as an extreme example,

[26] Popper (1959) intended verification and falsifications as deterministic procedures, which exclude even small probabilities of error. In our informal discussion we do not distinguish between events that are deemed extremely unlikely (such as thinking of having observed a black swan in world where all swans are white) and events that have zero probability. We refer the reader to (Popper, 1959, chapter 8) and Olszewski and Sandroni (2011) for a discussion of falsifiability and small probability events.

[27] The error term ε^2 can be interpreted as small noise in the observation. Its role is simply to ensure that log-likelihood ratios are finite for each observation.

the decision maker may look up in the sky. There is a small chance a nonwhite swan will be observed; if not, the decision maker's belief will not change by much. It is not obvious exactly what tests or samples would be necessary to implement experiment II, let along to conclude that the two experiments should be equally costly to perform.

We conclude that 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 the coefficients of the LLR cost satisfy $\beta_{ae} \neq \beta_{ea}$. For example, if we set $\beta_{ae} = \kappa$ and $\beta_{ea} = 0$, and if we consider small ε, then the cost of experiment I is $\kappa\varepsilon$, to first order in ε. In comparison, the cost of experiment II is—again to first order—a factor of $\log(1/\varepsilon)$ higher. Hence the ratio between the costs of these experiments is arbitrarily high for small ε.

We note that a difference between the costs of these experiments is impossible under mutual information and a uniform prior, since in that model the cost of an experiment is invariant with respect to a permutation of the states.

7 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 KL-divergence as a tool to model robust decision criteria under uncertainty. Cabrales, Gossner, and Serrano (2013) derive Shannon 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.

Rational Inattention. 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 instead refer the interested reader to Caplin (2016) and Mackowiak, Matějka, and Wiederholt (2018) for perspectives on this growing literature.

Decision Theory. 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, Caplin, Dean, and Leahy (2018) characterize mutual information cost, Chambers, Liu, and Rehbeck (2017) study non-separable models of costly information acquisition, and Denti (2018) provides a revealed preference 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.

The Wald Model of Sequential Sampling. The notion of constant marginal costs over independent experiments goes back to Wald's (1945) classic sequential sampling model; our axioms extend some of Wald's ideas to a model of flexible information acquisition. In its most general form, Wald's model considers a decision maker who acquires information by collecting multiple independent copies of a fixed experiment, and incurs a cost equal to number of repetitions. In this model, every stopping strategy corresponds to an experiment, and so every such model defines a cost over some family of experiments. It is easy to see that such a cost satisfies our axioms.

Morris and Strack (2018) consider a continuous-time version where the decision maker observes a one-dimensional diffusion process whose drift depends on the state, and incurs a cost proportional to the expected time spent observing. This cost is again easily seen to satisfy our axioms, and indeed, for the experiments that can be generated using this sampling process, they show that the expected cost of a given distribution over posteriors is of the form obtained in Proposition 3. Outside of the binary state case, only a restricted family of distributions over posteriors can be implemented by means of a sampling strategy. This has to be expected, since in Wald's model the decision maker has in each period a single, exogenously fixed, signal at their disposal.

One could imagine modifying the exercise in their paper by considering families of processes other than one-dimensional diffusion processes; for example, one could take Poisson processes with rates depending on the state. One of the contributions of our paper is to abstract away from such parametric assumptions, and show that a few simple axioms which capture the most basic intuition behind Wald's model suffice to pin down a specific family of cost functions over experiments. Nevertheless, one may view the result in Morris and Strack (2018) as complementary evidence that the cost function obtained in Proposition 3 is a natural choice for one-dimensional information acquisition problems.

Dynamic Information Acquisition Models. Hébert and Woodford (2018), Zhong (2017, 2019), and Morris and Strack (2018) relate cost functions over experiments and sequential models of costly information acquisition. In these papers, the cost $C(\mu)$ is the minimum expected cost of generating the experiment μ by means of a dynamic sequential sampling strategy.

Hébert and Woodford (2018) analyze a continuous-time model where the decision maker's beliefs follow a diffusion process and the decision maker can acquire information by varying its volatility. They propose and characterize a family of "neighborhood-based" cost functions that generalize mutual information, and allow for the cost of learning about states to be affected by their proximity. In a perception task, these cost are flexible enough to accommodate optimal response probabilities that are S-shaped, similarly to our analysis in §5.1. The LLR cost does not generalize mutual information, but has a structure similar to a neighborhood-based cost where the neighboring structure consists of all pairs of states.

Zhong (2017) provides general conditions for a cost function over experiments to be induced by some dynamic model of information acquisition. Zhong (2019) studies a dynamic model of non-parametric information acquisition, where a decision maker can choose any dynamic signal process as an information source, and pays a flow cost that is a function of the informativeness of the process. A key assumption is discounting of delayed payoffs. The paper shows that the optimal strategy corresponds to a Poisson signal.

Information Theory. 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.

Probability Theory. 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) in 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.

8 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 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.[28] 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):

$$c(\sigma_0, \sigma_1) = G\left[m_{\sigma_0}(1), \ldots, m_{\sigma_0}(N), m_{\sigma_1}(1), \ldots, m_{\sigma_1}(N)\right] \tag{13}$$

[28] 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.

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):

$$c(\sigma_0, \sigma_1) = H\left[\kappa_{\sigma_0}(1), \ldots, \kappa_{\sigma_0}(N), \kappa_{\sigma_1}(1), \ldots, \kappa_{\sigma_1}(N)\right] \tag{14}$$

where $\kappa_{\sigma_i}(n)$ 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 (13) and (14). 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 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.

9 Conclusions

In this paper we put forward an axiomatic approach to modeling the cost of information acquisition, characterizing a family of cost functions that capture a notion of constant marginal returns in the production of information. We study the predictions implied by our assumptions in various settings, and compare them to the predictions of mutual information costs.

We propose a number of possible avenues for future research, all of which would require the solution of some non-trivial technical challenges: The first is an extension of our framework beyond the setting of a finite set of states to a continuum of states.

In particular, this is natural in the context of one-dimensional problems we study in §4. Second, one could consider a generalization of the study of one-dimensional problems in §4 to multidimensional problems in which Θ is a subset of $\mathbb{R}^d$. This would constitute a rather general, widely applicable setting. Third, there are a number of important additional settings which have been modeled using mutual information cost, where it may be of interest to understand the sensitivity of the conclusions to this assumption, and how it may change if we assume constant marginal costs (see, e.g., Van Nieuwerburgh and Veldkamp, 2010).

Finally, if one accepts our axioms (and hence LLR costs) as capturing constant marginal costs, a natural definition for convex cost is a cost that given by the supremum over a family of LLR costs. Likewise, concave costs would be infima over LLR costs. It may be interesting to understand if such costs are characterized by simple axioms (e.g., by substituting the appropriate inequalities in our axioms) and whether they admit a simple functional form.

References

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

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

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

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

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

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

Brouwer, L. (1911). Beweis der invarianz des n-dimensionalen gebiets. *Mathematische Annalen 71*(3), 305–313.

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

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

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

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

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

Caplin, A., M. Dean, and J. Leahy (2018). Rational inattentive behavior: Characterizing and generalizing shannon entropy. Technical report, National Bureau of Economic Research.

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

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

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

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

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

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

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

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

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

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

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

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

Gescheider, G. A. (1997). *Psychophysics: the fundamentals* (3 ed.). Psychology Press.

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

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

Hébert, B. and M. Woodford (2018). Information costs and sequential information sampling.

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

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

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

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

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

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

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

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

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

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

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

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

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

Mensch, J. (2018). Cardinal representations of information.

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

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

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

Olszewski, W. and A. Sandroni (2011). Falsifiability. *American Economic Review 101*(2), 788–818.

Popper, K. (1959). *The logic of scientific discovery.* Routledge.

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

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

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

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

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

Tao, T. (2011). Brouwer's fixed point and invariance of domain theorems, and Hilbert's fifth problem. https://terrytao.wordpress.com/2011/06/13/brouwers-fixed-point-and-invariance-of-domain-theorems-and-hilberts-fifth-problem.

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

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

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

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

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

Zhong, W. (2017). Indirect information measure and dynamic learning.

Zhong, W. (2019). Optimal dynamic information acquisition.

Appendix 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. 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.

Appendix 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 6. *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. *$D_{\mathrm{KL}}(F_*\nu_1 \| F_*\mu_1) \leq D_{\mathrm{KL}}(\nu_1 \| \mu_1)$.*

It is well known that KL-divergence satisfies the first three properties in the statement of the proposition. We refer the reader to (Austin, 2006, Proposition 2.4) for a proof of the last property.

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

$$p_i(s) = \frac{\mathrm{d}\mu_i}{\mathrm{d}\sum_{j \in \Theta} \mu_j}(s) = \frac{1}{\sum_{j \in \Theta} \frac{\mathrm{d}\mu_j}{\mathrm{d}\mu_i}(s)} = \frac{1}{\sum_{j \in \Theta} \mathrm{e}^{\ell_{ji}}} \tag{15}$$

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 (15) they must induce the same distribution over posteriors, hence be equivalent in the Blackwell order. □

A consequence of Proposition 6 is that the LLR cost is monotone with respect to the Blackwell order.

Proof of Proposition 1. 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 experiment $\rho = (\mathcal{P}(\Theta), (\rho_i))$ 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 6,

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

Therefore, by Theorem 1, we have

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

□

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 6 is that the LLR cost is convex:

Proposition 7. *Let $\mu = (S, (\mu_i))$ and $\nu = (S, (\nu_i))$ be experiments in $\mathcal{E}$. Given $\alpha \in (0, 1)$, define the experiment $\eta = (S, (\nu_i))$ as $\eta_i = \alpha\nu_i + (1 - \alpha)\mu_i$ for each 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 6. We note that if ν and μ are not Blackwell equivalent, and if all the coefficients β_{ij} are positive, then the inequality above is strict.

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. This implies 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 nonzero 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 $[2, \infty)$:

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

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 $A = \{0, \dots, N\}^n \backslash \{0, \dots, 0\}$, for some constant $N \in \mathbb{N}$ greater 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 $(\lambda^1, \dots, \lambda^q)$ 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. $m_{\sigma_1}(\alpha) = \sum_{(\lambda^1, \dots, \lambda^q) \in \Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} \prod_{p=1}^q \kappa_{\sigma_1}(\lambda^p)$

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

B.4 Admissible Measures and the Cumulants Manifold

We denote by $\mathcal{A}$ the set of vectors of measures $\sigma = (\sigma_0, \sigma_1, \dots, \sigma_n)$ 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

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

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 every $\alpha \in A$ 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, $(1, \dots, 1) \in \mathbb{R}^n$ is in the interior of $\{(e^{\xi_1}, \dots, e^{\xi_n}) : \xi \in \mathbb{R}^n\}$. 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 $\sigma = (\sigma_0, \sigma_1, \dots, \sigma_n)$ 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,[29] 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. □

[29] Brouwer (1911). See also (Tao, 2011, Theorem 2).

Appendix C 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 $(a_1, \ldots, a_d) \in \mathbb{R}^d$ such that $F(x) = F(x_1, \ldots, x_d) = a_1 x_1 + \cdots + a_d x_d$ 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. □

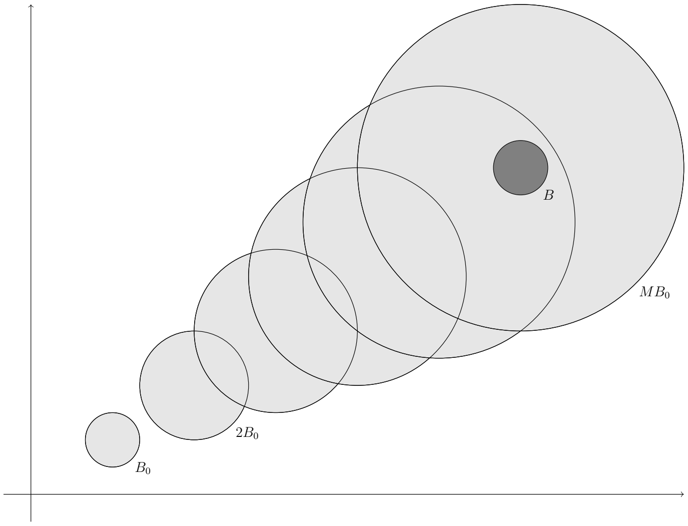

Figure 3: Illustration of the proof of Claim 1. The dark ball B is contained in the light ones, and it is apparent from this image that so is any multiple of B by $a \geq 1$.

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 $F(ax) = aF(x)$. □

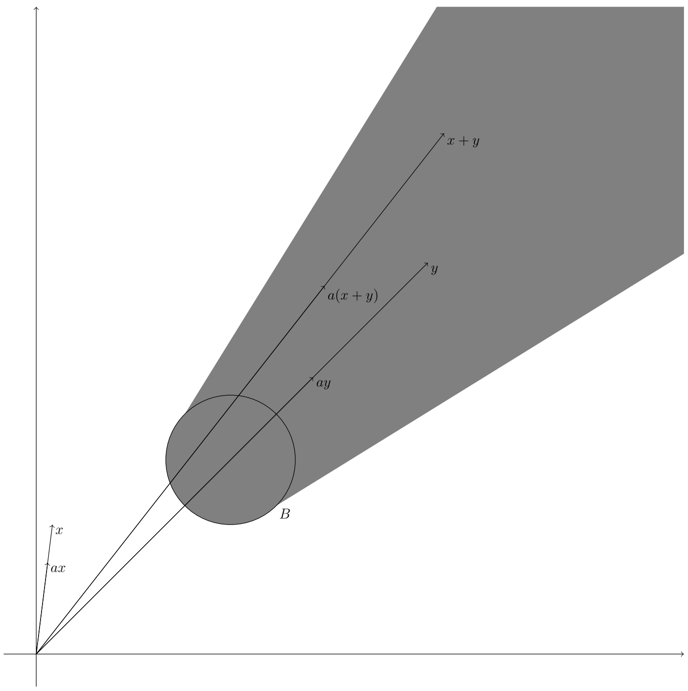

Figure 4: An illustration of the proof of Theorem 4.

Appendix D 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

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

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

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

For every $\alpha \in A$ we have

$$M_i^{\mu[r]}(\alpha) = \int \ell_{10}^{\alpha_1} \ldots \ell_{n0}^{\alpha_n} \, \mathrm{d}\mu[r]_i = \int_{\mathbb{R}^n} \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n} \, \mathrm{d}\sigma[r]_i(\xi) = m_{\sigma[r]_i}(\alpha) \tag{16}$$

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 $H(m_v) = \kappa_v$ for every $v \in \mathcal{A}$. 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 (16) therefore implies that $M_i^{\mu[r]}(\alpha) = M_i^{\nu[r]}(\alpha)$. Thus

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

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

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

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

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

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

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

Lemma 9 implies $c(\sigma) = \frac{1}{t} c(t\mu + (1-t)\,\delta_{\mathbf{0}})$ for every t. Hence

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

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 $r \in \mathbb{N}$,

$$c(\sigma^{*r}) = \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} r^q \prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) \right) \tag{17}$$

By the additivity of c, $c(\sigma^{*r}) = rc(\sigma)$. Hence, because $\gamma > 1$, $c(\sigma^{*r})/r^\gamma \to 0$ as $r \to \infty$. Therefore, diving (17) by r^γ we obtain

$$\sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} r^{q-\gamma} \prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) \right) \to 0 \text{ as } r \to \infty. \tag{18}$$

We now show that (18) 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 (18) we have $r^{q-\gamma} \to 0$ as $r \to \infty$ whenever $q < \gamma$. Hence (18) implies

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

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[30]

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

and

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

so we obtain that

$$\sum_{i\in\Theta}\sum_{\alpha\in A:|\alpha|=\gamma} \lambda_{i,\alpha} \left(\int_{\mathbb{R}^n} \xi_1 \, \mathrm{d}\sigma_i\,(\xi)\right)^{\alpha_1} \cdots \left(\int_{\mathbb{R}^n} \xi_n \, \mathrm{d}\sigma_i\,(\xi)\right)^{\alpha_n} = 0. \tag{19}$$

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 (19) 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 $j \in \{1,\dots,n\}$ 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

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

[30] 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

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

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.

Appendix E Additional Proofs

Proof of Proposition 2. 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 5. 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

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

□

Proof of Proposition 3. 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 $\Theta \in \mathcal{T}$

$$\kappa = C^\Theta(\nu^\Theta) = \sum_{i \neq j \in \Theta} \beta_{ij}^\Theta \frac{(i-j)^2}{2} = \sum_{i \neq j \in \Theta} g(|i-j|). \tag{20}$$

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 (20) 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 (20) 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

$$g(x) - g(y) = g(x + z) - g(y + z) \qquad \text{for all } x, y, z > 0. \tag{21}$$

Since g is non-negative, it follows from (20) 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 (21) 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 4. Let $\mu^\star$ be an optimal experiment. As argued in the text, $\mu^\star$ is such that $S = A$, so that it reveals to the decision maker what actions to play. Let $A^\star = \operatorname{supp}(\mu^\star)$ be the set of actions played in $\mu^\star$. It solves

$$\max_{\mu \in \mathbb{R}_+^{|\Theta| \times |A^\star|}} \left[\sum_{i \in \Theta} q_i \left(\sum_{a \in A} \mu_i(a) u(a,i) \right) - \sum_{i,j \in \Theta} \beta_{ij} \sum_{a \in A^\star} \mu_i(a) \log \frac{\mu_i(a)}{\mu_j(a)} \right] \tag{22}$$

$$\text{subject to} \qquad \sum_{a \in A^\star} \mu_i(a) = 1 \text{ for all } i \in \Theta. \tag{23}$$

Reasoning as in (Cover and Thomas, 2012, Theorem 2.7.2) the Log-sum inequality implies that the function D_{KL} is convex when its domain is extended from pairs of probability distrubutions to pairs of positive measures. Moreover, expected utility is linear in the choice probabilities. It then follows that the objective function in (22) is concave over $\mathbb{R}_+^{|\Theta| \times |A^\star|}$.

As (22) equals $-\infty$ whenever $\mu_i(a) = 0$ for some i and $\mu_j(a) > 0$ for some $j \neq i$ we have that $\mu_i^\star(a) > 0$ for all $i \in \Theta, a \in A^\star$. For every $\lambda \in \mathbb{R}^{|\Theta|}$ we define the Lagrangian $L_\lambda(\mu)$ as

$$L_\lambda(\mu) = \left[\sum_{i \in \Theta} q_i \left(\sum_{a \in A} \mu_i(a) u(a,i) \right) - \sum_{i,j \in \Theta} \beta_{ij} \sum_{a \in A} \mu_i(a) \log \frac{\mu_i(a)}{\mu_j(a)} \right] - \sum_{i \in \Theta} \lambda_i \sum_{a \in A} \mu_i(a)\,.$$

As $\mu^\star$ is an interior maximizer it follows from the Karush-Kuhn-Tucker conditions that there exists Lagrange multipliers $\lambda \in \mathbb{R}^{|\Theta|}$ such that $\mu^\star$ maximizes $L_\lambda(\cdot)$ over $\mathbb{R}_+^{|\Theta| \times |A^\star|}$. As $\mu^\star$ is interior it satisfies the first order condition

$$\nabla L_\lambda(\mu^\star) = 0\,.$$

We thus have that for every state $i \in \Theta$ and every action $a \in A^\star$

$$0 = q_i u_i(a) - \lambda_i - \sum_{j \neq i} \left\{ \beta_{ij} \left[\log \left(\frac{\mu_i^\star(a)}{\mu_j^\star(a)} \right) - 1 \right] - \beta_{ji} \frac{\mu_j^\star(a)}{\mu_i^\star(a)} \right\}\,. \tag{24}$$

Subtracting (24) evaluated at a' from (24) evaluated at a yields that (8) is a necessary condition for the optimality of $\mu^\star$. □