The Cost of Information*

Luciano Pomatto† Philipp Strack‡ Omer Tamuz§

December 10, 2020

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 Blackwell monotonicity and a continuity condition, 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.

‡Yale. Email: philipp.strack@gmail.com. Philipp Strack was supported by a Sloan Fellowship.

§Caltech. Email: tamuz@caltech.edu. Omer Tamuz was supported by a grant from the Simons Foundation (#419427), a Sloan research fellowship, and a BSF award (#2018397).

We characterize all cost functions over Blackwell experiments that satisfy three main axioms: First, experiments that are more informative in the sense of Blackwell (1951) are more costly. Second, the cost of generating independent experiments equals the sum of their individual costs. Third, the cost of generating an experiment with probability half equals half the cost of generating it with probability one.

Our three axioms admit a straightforward economic interpretation. The first one is a form of monotonicity: more precise information is more costly. The second and third axioms capture the idea of linear cost. The second axiom implies that the cost of collecting n independent random samples is linear in n. For example, if information is generated by surveying customers, the axiom is satisfied if the cost of calling an additional customer is constant: i.e. calling 20 customers is twice as costly as calling 10. Similarly, the third axiom implies that the cost of producing a sample with probability α is a fraction α of the cost of acquiring the same sample with probability one. This axiom is satisfied by all posterior separable costs, which include nearly all models of information cost in the literature.

We propose these linearity assumptions as a baseline for studying cost functions over information structures. In the context of traditional commodities, a standard avenue for studying cost functions is by categorizing them in terms of decreasing, increasing, or constant marginal costs, with the latter being arguably the conceptually simplest case. In this paper we take a similar approach for studying the cost of information acquisition, and our axioms formalize the assumption of constant marginal costs for information. As in the case of traditional commodities, assuming linear costs is restrictive, and it is easy to think of decision problems where our axioms are violated. For example, surveying 20 customers might cost more than twice as much as surveying 10 if customers are hard to find, and economies of scale may result in decreasing marginal costs. Nevertheless, our axioms have the advantage of admitting a clear economic interpretation, making it possible to judge for which applications they are appropriate. We thus propose the study of linear cost functions as an obvious first step towards the wider goal of studying general information costs in terms of their economic properties.

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 $s \in S$ 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 unique non-negative coefficients (β_{ij}), one for each ordered pair of

states of nature i and j, such that[1]

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

Each coefficient β_{ij} can be interpreted as capturing the difficulty of discriminating between state i and state j, as 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. The term $D_{\mathrm{KL}}(\mu_i \| \mu_j)$ 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 larger the coefficient 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) function as the *log-likelihood ratio cost* (or *LLR cost*).

In many common information acquisition problems, states of the world are one-dimensional quantities. 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 2 shows that there exists a constant $\kappa \geq 0$ 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, this choice of

[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).

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

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

from the prior q to the posterior p induced by the signal.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 acquisition (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. The LLR cost leads to predictions which are profoundly different from those induced by mutual information cost. We illustrate the differences in four stylized examples in §5.

Examples and Applications. In §6 we apply the LLR cost function to optimal information acquisition problems and derive a number of predictions of interest. One of our main applications are binary prediction problems, where a decision maker needs to predict whether the state is above or below a given threshold. An instance of this problem is an analyst trying to predict which party will obtain the majority of votes in the election.

[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.

Another instance is a perception task, where the agent is a subject who is asked to observe a number of dots of two different colors on a screen, and must predict which color is predominant.

We show that in binary prediction problems the decision maker is strictly more likely to make the correct choice when the quantity to predicted is closer to the desired threshold, under general assumptions on the coefficients (β_{ij}). For example, it is harder for the agent to predict the winner in a close election than in an election where one of the candidates has a large lead. Moreover, we show that under the specification $\beta_{ij} = \frac{\kappa}{(i-j)^2}$, the decision maker's probability of being correct is a sigmoidal function of the state—a prediction in line with psychometric evidence on perception tasks.

This and other examples illustrate how the LLR cost function leads to optimal choice probabilities that properly take into account the difficulty of distinguishing between states. While intuitive, this property is ruled out by cost functions that treat states symmetrically, as in the case of mutual information.

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 $\frac{\mu_i(s)}{\mu_j(s)}$. 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

$$(\ell_{ij}(s))_{i,j\in\Theta}$$

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.[3]

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.[4] 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 (not fully revealing) experiments commonly used in applications.

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.[5] 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 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, as shown by Proposition 1 below.

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

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

[4] 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).

[5] 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.

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 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 two signals $\mu = (S, (\mu_i))$ and $\nu = (T, (\nu_i))$ we define the product 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.[6] 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 combining 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 it 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:*

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

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

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 ratios conditional on state i. We denote by d_{tv} the total-variation distance.[7] 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.[8] As we show in Theorem 6 in the Appendix, our characterization holds for the case of two states and bounded experiments even if one only imposes Blackwell monotonicity, Axiom 2 and 3 without requiring continuity.

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 is linear in n.[9] 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

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

[8] 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.

[9] 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).

a priori restrictions are imposed over the domain of feasible experiments. As discussed in the introduction, we focus on linear cost structures as we view those as a natural starting point to reason about the cost of information, in the same way the assumption of constant marginal cost is a benchmark for the analysis of traditional commodities. Whether this assumption fits a particular application well is inevitably an empirical question.

Axiom 3 expresses the idea that the marginal cost of increasing the probability of success of an experiment is constant. The axiom is implied by posterior separability—the standard assumption in the literature for cost functions over experiments. It is however, a strictly weaker assumption. We also note that for proving our results it suffices to restrict this axiom to $\alpha = 1/2$.[10]

The domain of our cost function rules out experiments that perfectly reveal the true state with positive probability. This domain ensures that the cost function takes finite values. For example, a perfectly revealing experiment would have have infinite cost under any nontrivial cost function that is Blackwell monotone and additive.[11] This is not special to our framework, and holds also in the Wald model, as well as in models that restrict attention to normal signals and assume convex cost in the precision (as in, e.g. , Wilson, 1975).

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 $\mu = (S, (\mu_i))$,*

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

[10] The axiom admits an additional interpretation. Suppose the decision maker is allowed to randomize her choice of experiment. Then, the property

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

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

[11] By nontrivial we mean that there exists at least one experiment μ that is not perfectly revealing and, under C, has strictly positive cost. The experiment μ will remain so regardless of the number of repetitions. This implies, by Blackwell monotonicity and additivity, that the cost of the n-times repeated experiment $\mu^{\otimes n}$ must always be below the cost of a perfectly informative experiment. Hence, a perfectly informative experiment must have infinite cost.

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). The representation (3) can be rewritten as

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

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 at hand. 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, a state of the world i represents a possible 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 differences 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 §8 we discuss environments where the coefficients might naturally assumed to be asymmetric, in the sense that $\beta_{ij} \neq \beta_{ji}$.[13] In environments where no pair of states is a priori harder to distinguish than another,[14] a simple choice is to set all the

[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] Since we do not impose symmetry axioms, it is in a sense a natural finding that the LLR cost function can capture differences in the costs of learning about different states. What we think is surprising is that the cost function has a relatively small set of parameters, of dimension $2n(n-1)$, where n is the number of states. Given the level of generality of our framework, we believe this is indeed a small set of parameters. In fact, even the fact that the set of parameters is finite is not to be taken for granted. For example, the family of cost functions based on f-divergences is a simple generalization of mutual information, and it is an infinite dimensional class of cost functions.

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

coefficients (β_{ij}) to be equal. Finally, in the next section we propose a specific functional form in the more structured case where states represent a one-dimensional quantity.

We end this section by showing that the 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 $C(\mu) \geq C(\nu)$.*

4 Learning about a One-Dimensional State

Many information acquisition problems involve learning about a one-dimensional characteristic, so that each state i is a real number.[15] In macroeconomic applications, the state may represent the future level of interest rates. In perceptual experiments in neuroscience and economics, the state can correspond to the number of red/blue dots on a screen or the number of voters voting for a given party (see §6 below). More generally, i might represent a physical quantity to be measured.

In this section we propose a 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. 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, 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 on the precision of the measurement instrument, but not on whether we restrict our attention to a group of individuals whose height is within a certain range.

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, on the cost of information.

[15] 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.

The first axiom states that β_{ij}^{Θ}, the marginal cost of increasing the expected LLR between two states i and j 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 $k, l \in \Xi$,*

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

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 $(\mathbb{R}, (\zeta_i)_{i \in \Theta})$. This is the canonical experiment consisting of a noisy measurement of the state plus standard normal error.[16] 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 2. *The collection $C^{\Theta}, \Theta \in \mathcal{T}$, satisfies Axioms a and b if and only if there exists a constant $\kappa > 0$ such that for all $i, j \in \Theta$ and $\Theta \in \mathcal{T}$,*

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

where n is the cardinality of Θ.

The result shows that under Axioms a and b each coefficient β_{ij}^{Θ} is decreasing in the distance between the states. Thus, distinguishing states that are closer to each other is more costly. Each coefficient is also divided by a factor $n(n-1)$ that normalizes the cost with respect to the number of states. This is an implication of Axiom b, which states the cost of performing a noisy measurement does not depend on the particular values the state can take.

Proposition 2 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 2 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.

[16] 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.

5 Illustrative Examples

5.1 LLR Cost for Normal and Binary Signals

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. In this example, each m_i is a feature of the information structure: choosing a signal where the distances between states $|m_i - m_j|$ are higher provides stronger information about the states.

By substituting (3) with the well-known expression for the Kullback-Leibler divergence between normal distributions, we obtain that the cost of such an experiment is given by

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

The cost is decreasing in the variance σ^2, as one may expect. Increasing β_{ij} increases the cost of a signal $\mu^{m,\sigma}$ by a factor that is proportional to the squared distance between the signal means.

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.

5.2 Hypothesis Testing

In this section we apply the log-likelihood ratio cost to a standard hypothesis testing problem. We study a decision maker performing an experiment with the goal of learning about an hypothesis, i.e. whether the state is in a subset $H \subset \Theta$.

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 let μ be the binary signal with signal realizations $S = \{H, H^c\}$ (where H^c denotes the complement of H)

$$\mu_i(s) = \begin{cases} \alpha & \text{for } i \in s \\ 1-\alpha & \text{for } i \notin s \end{cases} \tag{6}$$

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

$$\left(\sum_{i \in H, j \in H^c} \beta_{ij} + \beta_{ji} \right) \left(\alpha \log \frac{\alpha}{1-\alpha} + (1-\alpha) \log \frac{1-\alpha}{\alpha} \right) \tag{7}$$

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 sum of 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. We next illustrate with an example how this allows to capture the fact that testing two different hypotheses can lead to very different costs even if they involve the same number of states.

Learning about the GDP. For concreteness, we take a state to be a natural number i in the interval $\Theta = \{20000, \dots, 80000\}$, representing, for instance, the current US GDP per capita. We fix the following two hypotheses:[17]

(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; the same qualitative conclusion will hold as long as β_{ij} is strictly decreasing in the distance$|i-j|$. Then,

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

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

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

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.[18] 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.

5.3 Acquiring Precise Information

We next illustrate how our additivity axiom captures constant marginal costs in information acquisition, and we contrast it with the assumption of decreasing marginal costs that is implicit in mutual information. Consider the classical problem of learning the bias of a coin by flipping it multiple times. This experiment could correspond to the act of surveying customers, who either like a product or not, in order to learn whether the product is popular. It could also represent a political party surveying voters to discover the appeal of a potential candidate.

Suppose the coin either yields heads 80% of the time or tails 80% of the time and that either bias is equally likely. We compare the cost of observing a single coin flip versus a long sequence of coin flips. 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 approximately equals $0.2\,\lambda$. The cost of observing k flips is less than $0.7\,\lambda$, for *any* k. This implies, for example, that a firm could survey a million customers at less than 4 times the cost of surveying a single customer. This low cost of acquiring perfect information is a consequence of the sub-additivity of mutual information as a cost function.

This difference in the marginal cost of information is not simply a mathematical distinction, but it 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.

6 Information Acquisition in Decision Problems

In this section we study the implications of the log-likelihood ratio cost function for decision problems. We consider a decision maker choosing an action a from a finite set A. The payoff from a depends on the state $i \in \Theta$ and is given by $u(a, i)$. The agent is endowed with a prior q over the set of states that has full support. Before making her choice, the agent can acquire a signal $\mu \in \mathcal{E}$ at cost $C(\mu)$, where C is a LLR cost function where all the coefficients (β_{ij}) are assumed to be strictly positive.

[18] This follows from the fact that the mutual information cost is invariant with respect to a relabelling of the states.

As is well known, for a cost function that is monotone with respect to the Blackwell order, 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. Throughout this section, we will therefore therefore identify an experiment μ with a vector of probability measures over actions $\mu \in \mathcal{P}(A)^n$.

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

$$\mu^\star \in \underset{\mu \in \mathcal{P}(A)^n}{\operatorname{argmax}} \left[\sum_{i \in \Theta} q_i \left(\sum_{a \in A} \mu_i(a) u(a,i) \right) - C(\mu) \right] . \tag{8}$$

Hence, the optimal action a is chosen in state i with probability $\mu_i^\star(a)$. The maximization problem (8) is well behaved: the maximand is smooth and concave (see Proposition 11 in the Appendix), and there always exists an optimal solution.[19] Thus, an optimal experiment can be found by applying standard methods in concave optimization.

It is without loss of generality to restrict the attention to choice probabilities where an action that is chosen with strictly positive probability in one state it is chosen with strictly positive probability in every state, since otherwise the experiment is not in the domain $\mathcal{E}$.

6.1 Implications for Optimal Choice Probabilities

We obtain a characterization of the decision maker's optimal choice probabilities. The characterization is based on the study of first-order conditions, and is therefore analogous to that obtained by Matějka and McKay (2015) for the case of entropy cost.

The result is based on a standard economic intuition. For choice probabilities to be optimal, the marginal benefit of choosing an action a marginally more often then a different action b must exactly offsets its marginal cost. Formally, given a vector μ of choice probabilities, we denote by $\operatorname{supp}(\mu)$ the support of μ, i.e. the set of actions which are played with strictly positive probability under μ.[20] Given two actions a and b in the support of μ, consider perturbing μ by increasing the probability $\mu_i(a)$ of taking action a in state i, while decreasing by the same amount the probability $\mu_i(b)$ of taking action b in the same state. The marginal benefit of this perturbation is denoted by $\mathrm{MB}_i(a,b)$ and is equal to

$$\mathrm{MB}_i(a,b) = q_i \left[u(a,i) - u(b,i) \right] .$$

Such a transfer of probabilities has an effect on the information cost of the experiment μ.

[19] To establish existence of an optimal solution, recall that the Kullback-Leibler divergence $D_{\mathrm{KL}} \colon \mathcal{P}(A) \times \mathcal{P}(A) \to [0, \infty]$ is a lower-semicontinuous function (Dupuis and Ellis, 2011, Lemma 1.4.3). The maximand in (8), being a sum of upper-semicontinuous functions, is therefore upper-semicontinuous. Since $\mathcal{P}(A)^n$ is compact, the problem admits a solution.

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

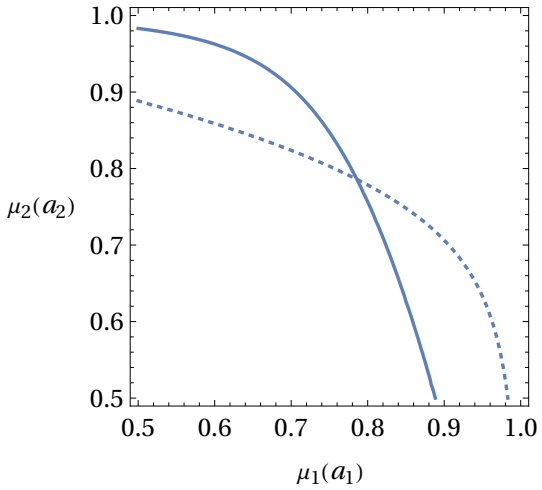

Figure 1: A decision problem where $\Theta = \{1, 2\}$ and $A = \{a_1, a_2\}$, the prior q is uniform, $\beta_{12} = \beta_{21} = 1$, and payoffs are $u_1(a_1) = u_2(a_2) = 3$ and $u_1(a_2) = u_2(a_1) = 0$. The solid line is the locus of choice probabilities such that $\mathrm{MB}_1(a_1, a_2) = \mathrm{MC}_1(a_1, a_2)$. The dotted line is the locus where $\mathrm{MB}_2(a_2, a_1) = \mathrm{MC}_2(a_2, a_1)$. The optimal vector of choice probabilities is given by the intersection of the two curves.

This is given by the expression:

$$\mathrm{MC}_i(a, b) = \sum_{j \in \Theta} \beta_{ij} \left(\log \frac{\mu_i(a)}{\mu_j(a)} - \log \frac{\mu_i(b)}{\mu_j(b)} \right) - \sum_{j \in \Theta} \beta_{ji} \left(\frac{\mu_j(a)}{\mu_i(a)} - \frac{\mu_j(b)}{\mu_i(b)} \right). \tag{9}$$

It measures the change in information acquisition cost necessary to choose action a marginally more often and action b marginally less often. For the choice probabilities μ to be chosen optimally, this change in information cost must equal the difference $q_i \left[u(i, a) - u(i, b) \right]$ in expected benefits. This is the content of the next proposition.

Proposition 3. *Assume $\beta_{ij} > 0$ for all i and j. Let $\mu = (\mu_i)_{i \in \Theta}$ be the vector of choice probabilities that solves the optimization problem (8). Then, for every state $i \in \Theta$ it holds that*

$$\mathrm{MB}_i(a, b) = \mathrm{MC}_i(a, b) \quad \text{for all } a, b \in \mathrm{supp}(\mu). \tag{10}$$

Figure 1 illustrates the result in a simple decision example with two states and two actions where the decision maker's goal is to match the state. The solid line represents the locus of choice probabilities μ_1 where the optimality equation $\mathrm{MB}_1(a_1, a_2) = \mathrm{MC}_1(a_1, a_2)$ holds, taking μ_2 as given. The dotted line represents the locus where the optimality equation holds with respect to state 2, taking now μ_1 as given. The intersection of the two determines the optimal choice probabilities.

6.2 Continuity of Choice Probabilities

A key distinguishing feature of the LLR cost is its ability to model the fact that closer states are harder to distinguish, in the sense that acquiring information that finely discriminates between them is more costly. This, in turn, suggests that choice probabilities cannot vary abruptly across nearby states.

To formalize this intuition we assume that the state space Θ is endowed with a distance $d\colon \Theta \times \Theta \to \mathbb{R}$. We say that *nearby states are hard to distinguish* if for all $i, j \in \Theta$

$$\beta_{ij} \geq \frac{1}{d(i,j)^2}\,. \tag{11}$$

Under this assumption the cost of acquiring information that discriminates between states i and j is high for states that are close to each other. Our next result shows that when nearby states are hard to distinguish, the optimal choice probabilities are Lipschitz continuous in the state: the agent will choose actions with similar probabilities in similar states. For this result, we denote by $\|u\| = \max_{i,a} |u(a,i)|$ the norm of the decision maker's utility function.

Proposition 4 (Continuity of Choice)**.** *Suppose that nearby states are hard to distinguish. Then the optimal choice probabilities $\mu^\star$ solving (8) are uniformly Lipschitz continuous with constant $\sqrt{\|u\|}$, i.e. satisfy*

$$\sum_{a\in A} \left|\mu_i^\star(a) - \mu_j^\star(a)\right| \leq \sqrt{\|u\|}\, d(i,j) \quad \text{for all } i, j \in \Theta. \tag{12}$$

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 (12) 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.[21] In addition, assumption (12) can be generalized to arbitrary ordinal transformations of the distance d. The proof of Proposition 4 shows that if the coefficients satisfy $\beta_{ij} \geq 1/f(d(i,j))^2$ for a monotone increasing function f, then the conclusion of the proposition holds with the right hand side of (12) replaced with $\sqrt{\|u\|}\, f(d(i,j))$.

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 (see §6.5 for an example). Under the log-likelihood ratio cost, when nearby states are harder to distinguish, the change in choice probabilities

[21] Proposition 4 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 (12) 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.

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.

6.3 Comparative Statics with Respect to the Coefficients (β_{ij})

While so far we have focused on the effect that the coefficients β_{ij} have on the cost of a given experiment, we now address the question of their effect on behavior. The next proposition is a comparative statics result describing how choice probabilities vary with the parameters (β_{ij}).

Proposition 5. *Consider a decision problem, and let μ and μ' be the optimal choice probabilities obtained under a LLR cost function with coefficients (β_{ij}) and (β'_{ij}), respectively. Then*

$$\sum_{i \neq j} (\beta'_{ij} - \beta_{ij}) (D_{\mathrm{KL}}(\mu'_i \| \mu'_j) - D_{\mathrm{KL}}(\mu_i \| \mu_j)) \leq 0.$$

All other things equal, increasing a coefficient β_{ij} between two states decreases the Kullback-Leibler divergence $D(\mu_i \| \mu_j)$ between the corresponding optimal choice probabilities. In words, this makes the decision maker's behavior more similar in the two states.

Proposition 5 follows from the same logic underlying the law of supply in standard microeconomic models of production. Under the LLR cost function, the decision maker solves an optimization problem that is mathematically equivalent to a profit maximization problem. Each expected log-likelihood ratio $D_{\mathrm{KL}}(\mu_i \| \mu_j)$ is an intermediate "input" which accrues to the decision maker's expected payoff. Each such input is "priced" according to a linear price β_{ij}. The comparative statics described by the result follows from such a linearity property, together with a standard revealed-preference argument.

6.4 Identifying the Cost from Observed Choices

Proposition 3 can be applied to problem of identifying and testing the model from observed choices. We illustrate this in the context of a simple example. We consider a binary choice problem where we are given two a priori equally likely states $\Theta = \{1, 2\}$. The agent can take one two actions, a_1 and a_2, and receives a payoff $v > 0$ if the action matches the state and 0 otherwise.

An analyst observes the agent's choice probabilities $(\mu_i(a))_{i \in \Theta, a \in A}$, and is interested in testing if such probabilities are consistent with LLR cost. This is true if there exist

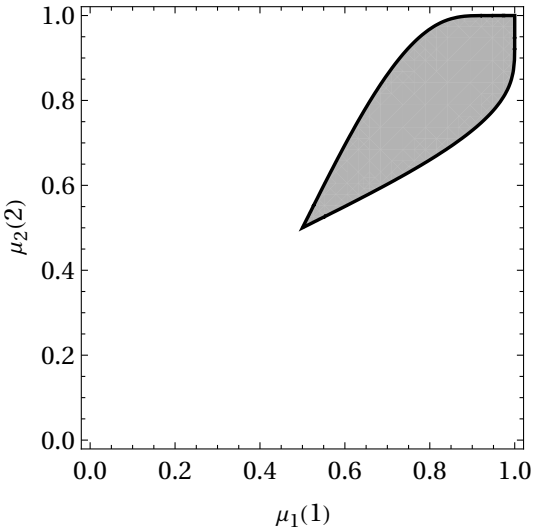

Figure 2: The probabilities of choosing correctly in state 1 and state 2 that are consistent with LLR cost.

coefficients (β_{12}, β_{21}) that satisfy equation (10). The equation simplifies to

$$\begin{aligned} \frac{v}{2} &= -\left[\beta_{12}\left(\log l_1 - \log l_2\right) + \beta_{21}\left(l_1 - l_2\right)\right] \\ -\frac{v}{2} &= -\left[\beta_{21}\left(-\log l_1 + \log l_2\right) + \beta_{12}\left(1/l_1 - 1/l_2\right)\right] . \end{aligned} \tag{13}$$

where $l_1 = \frac{\mu_2(1)}{\mu_1(1)}$ and $l_2 = \frac{\mu_2(2)}{\mu_1(2)}$. Rearranging the above conditions yields that one can infer the information cost parameters (β_{ij}) from her choice probabilities μ as

$$\beta_{12} = \frac{v}{2}\frac{l_2 - l_1 + \log\frac{l_1}{l_2}}{\frac{(l_1-l_2)^2}{l_1 l_2} - (\log\frac{l_1}{l_2})^2} \qquad \beta_{21} = \frac{v}{2}\frac{\frac{l_2-l_1}{l_1 l_2} + \log\frac{l_1}{l_2}}{\frac{(l_1-l_2)^2}{l_1 l_2} - (\log\frac{l_1}{l_2})^2} . \tag{14}$$

For example, if the agent takes the correct action 80% of the time in state 1 and 60% of the time in state 2, we have that $(\mu_1(1), \mu_1(2), \mu_2(1), \mu_2(2)) = (0.8, 0.2, 0.4, 0.6)$ and the above formula yields that $(\beta_{12}, \beta_{21}) = (0.37v, -0.07v)$. As the implied β_{21} is negative these choice probabilities are inconsistent with any LLR cost function and this type of choice behavior would reject our model. In contrast, if the agent takes the correct action 80% of the time in state 1 and 70% of the time in state 2, we have that $(\mu_1(1), \mu_1(2), \mu_2(1), \mu_2(2)) = (0.8, 0.2, 0.3, 0.7)$ which implies that $(\beta_{12}, \beta_{21}) = (0.18v, 0.03v)$, and thus that this choice behavior can be explained by a LLR cost. Figure 2 more generally depicts all probabilities of choosing correctly in state 1 and state 2 that are consistent with LLR cost.

This examples illustrates how an analyst could use choice data to either reject LLR cost or to identify the information cost parameter β. In general, when there are more than

two states and actions the analyst might need data from multiple decision problems to point identify β. For a general decision problem with $|A|$ actions and $|\Theta|$ the model admits $|\Theta|\Big(|\Theta|-1\Big)$ degrees of freedom and (10) imposes $|\Theta| \times \frac{1}{2}|A| \times (|A|-1)$ linear equations on β which suggests that to identify the analyst needs to observe behaviour in

$$\frac{|\Theta|-1}{\frac{1}{2}|A| \times (|A|-1)}$$

decision problems. While there is in general no closed-form solution for the inferred coefficients β akin to (14) identifying them from data is easy numerically as the corresponding system of equation is linear.[22]

6.5 Perception Tasks

In this section we study the implications of the LLR cost function for perception tasks, a well known and long studied family of decision problems. In a perception task there is an even number n of dots and each dot is either red or blue. The agent guesses 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[23]

$$u(a,i) = \begin{cases} 1 & \text{if } a = B \text{ and } i > n/2 \\ 1 & \text{if } a = R \text{ and } i < n/2 \\ 0 & \text{otherwise.} \end{cases}$$

Such perception tasks can be used to model many applied learning problems. For example, each dot could correspond to a voter whose color indicates whether they vote for the red or blue party and the agent is an analyst trying to predict which party will obtain the majority of votes in the election. Perception tasks have also been studied experimentally, both in neuroscience as well as in economics and psychology. In a typical experiment subjects observe 100 dots each of which is either red or blue on a screen (see, e.g. Caplin and Dean, 2013; Dean and Neligh, 2017) and are asked whether there are more red or blue dots on the screen.

We fix some $r \in \{1, \ldots, n/2\}$ and assume that the prior q over states is uniform over the set $\Theta = \{n/2 - r, \ldots, n/2 - 1, n/2 + 1, \ldots, n/2 + r\}$, so that $q_i = \frac{1}{2r}$ for all $i \in \Theta$. The state where the number of blue and red dots is exactly equal to n is ruled out to simplify the exposition. As in the case of binary decision problems, it is without loss of generality to assume that $\mu_i(B)$ is strictly between 0 and 1 in every state. For a vector of distributions

[22] Due to the linear structure of the implied restrictions, one could also construct finite sample test for the LLR model using standard econometric methods, but this is beyond the scope of this paper.

[23] None of our results depend on the particular value chosen for the payoff from a correct guess.

over actions (μ_i), the decision maker guesses correctly in state i with probability

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

Intuitively, it should be harder to guess correctly when the difference in the number of dots of different colors is small, i.e. when i is close to $n/2$. For example, it should be harder to predict the winner in a close election than in an election where one of the candidates has a large lead. Also in single agent experiments, it is a well established fact in the psychology[24], neuroscience[25], economics[26] 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 (i.e. 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 $\mu^\star$ must induce a probability of guessing correctly that is state-independent.[27] As shown by Matějka and McKay (2015), Caplin and Dean (2013), and Steiner, Stewart, and Matějka (2017), conditional on a state i, the log-likelihood ratio $\log(\mu_i(B)/\mu_i(R))$ between the two actions must equal the difference in payoffs $u(B,i) - u(R,i)$, up to a constant. Hence, the probability of a correct choice 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 out of 100 and the state in which there are 99 blue dots. In the context of the election example, this means that, under mutual information cost, is equally hard for the analyst to predict an election where one candidate wins with 80% of the votes as it is to predict an election where one candidate wins with 50.1% of the votes.

This unrealistic prediction is driven by the fact that under mutual information the states are devoid of meaning and thus equally hard to distinguish. The same conclusion holds for any cost function C that, like mutual information, is invariant with respect to a permutation of the states and is convex as a function of the choice probabilities (μ_i).

In contrast, the LLR cost function can account for the difficulty of distinguishing different states through the coefficients β. As this is a one-dimensional information acquisition problem, it is natural to apply the specification $\beta_{ij} = \kappa/(i-j)^2$ of the LLR cost

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

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

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

[27] 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."

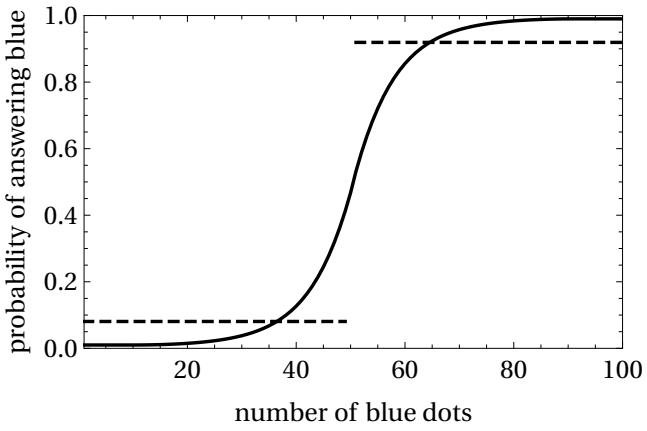

Figure 3: Predicted probability of guessing that there are more red dots as a function of the state, for the LLR cost with $\beta_{ij} = 1/(i-j)^2$ (solid line) and for mutual information cost (dashed line).

described in §4. As can be seen in Figure 3, this LLR cost predicts a sigmoidal relation between the state and the choice probability. For example when the election is almost split and one candidate wins with 50.1% of the votes LLR costs predict that the analyst will often make the wrong prediction while an election where one candidate receives 80% of the votes will rarely be called wrongly by the analyst. Thus, the model matches the qualitative features of choice probabilities commonly observed in practice.

To gain additional insight, we now consider a more basic assumption on the cost function. Rather than assuming a particular specification, we assume that the coefficients (β_{ij}) are *strictly decreasing in the distance between states*: there exists a positive and strictly decreasing function f such that $\beta_{ij} = f(|i-j|)$ for all pairs of states. The condition captures the idea that states which are closer to each other are harder to distinguish.

Even under this general non-parametric assumption, the LLR cost function leads to the intuitive prediction that the decision maker will guess correctly with strictly higher probability when the difference in the number of dots of different colors is smaller:

Proposition 6. *Consider the perception task above. Let C be a LLR cost function where the parameters (β_{ij}) are strictly increasing in the distance between states. Then, the resulting optimal probabilities (m_i) of guessing correctly satisfy $m_i > m_j$ whenever $|i - \frac{n}{2}| > |j - \frac{n}{2}|$.*

Thus, the decision maker is strictly more likely to make the correct choice in states where the composition of dots is farther away from fifty-fifty.

6.6 The Effect of Greater Incentives

We first apply the characterization of Proposition 3 to study more in detail the classic problem of predicting the probability of choosing between two options as a function of

their relative values. In its simplest implementation, it consists of a task where there are two equally likely states, two actions a_1 and a_2, and each action yields a payoff $v \in \mathbb{R}$ when chosen in the corresponding state, and 0 otherwise. We term this the *binary choice problem*. Compared to §6.4, we focus here on the question of how the decision maker's behavior varies as a function of v.

The payoff v may represent money, under the assumption that the decision maker is risk neutral, or utils, after appropriately correcting for risk aversion. Alternatively, the quantity v can represent the probability of receiving a fixed prize.

The next result derives the optimal choice probabilities in a binary choice problem under a symmetric LLR cost function. Without loss of generality we restrict our attention to choice probabilities where both actions are chosen with strictly positive probability in every state. The result follows by rearranging the optimality conditions of Proposition 3.

Proposition 7. *In a binary choice problem, let $\mu_i[v]$ denote the optimal choice probability of choosing action a_i, in state i, as a function of the reward v, under an LLR cost function. Assume the cost function satisfies $\beta_{12} = \beta_{21} = \beta$. Then $\mu_1[v] = \mu_2[v] = m[v]$, where*

$$m[v] = \frac{\mathrm{e}^{\eta\left(\frac{v}{2\beta}\right)}}{1 + \mathrm{e}^{\eta\left(\frac{v}{2\beta}\right)}}$$

and $\eta\colon \mathbb{R} \to \mathbb{R}$ is the inverse of the function $x \mapsto 2x + e^x - e^{-x}$.

This result prescribes a precise relation between higher incentives and the probability of a correct choice. As shown in Figure 4, and can be easily proven analytically, the optimal choice probabilities $\mu[v]$ are a sigmoidal function of the payoff v. The prediction is in line with a vast empirical literature.

It is useful to compare the results with those obtained under the entropy cost function. As shown by Matějka and McKay (2015), under entropy the optimal choice probabilities follow a logistic relation, where the probability of matching the state, as a function of v, is given by

$$\frac{e^{\frac{v}{2\lambda}}}{1 + e^{\frac{v}{2\lambda}}},$$

and $\lambda > 0$ is the parameter controlling the cost of information acquisition. The two functional forms are similar, with the only difference being the transformation η. The function is strictly increasing and S-shaped, onto, and satisfies $\eta(x) = \eta(-x)$ (in particular, $\eta(0) = 0$).

While both the LLR and models lead to choice probabilities that are sigmoidal, the two theories lead to significantly different predictions on how the probabilities of errors scale with the payoff v. Figure 4 displays the implied probabilities with which a decision maker takes a correct choice as a function of v, under the two theories. To make the comparison

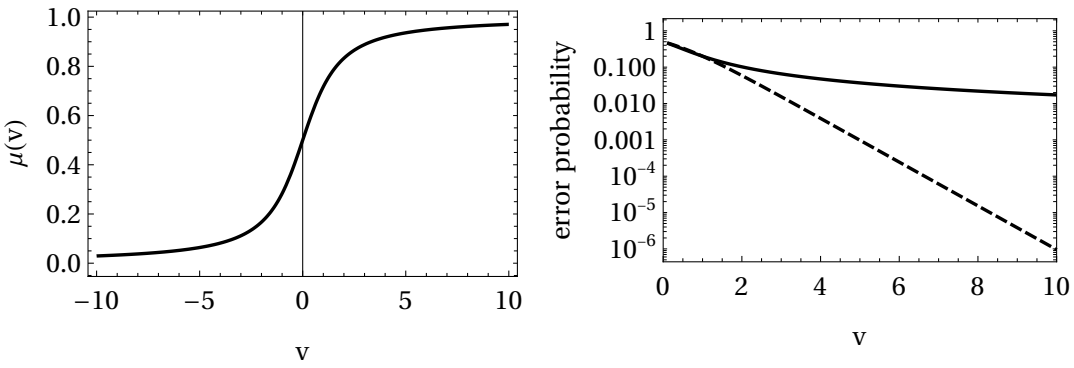

Figure 4: On the left: The optimal choice probabilities in a binary decision problem for a LLR cost function. On the right: The implied probabilities of choosing incorrectly at different levels of incentives v if the agent chooses correctly with 80% probability for $v = 1$ for the LLR cost (solid line) and mutual information cost (dashed line) on a log-scale.

meaningful, the parameters β and λ are chosen so that in both models the agent chooses incorrectly with probability 20% when the payoff is $v = 1$.

As one can see in the figure, the probability of choosing correctly reacts more strongly to incentives under mutual information cost. For example, suppose that the payoff v is measured in dollars. A simple calculation shows that under mutual information cost, if the decision maker chooses incorrectly with probability 20% when $v = \$1$, then she must choose incorrectly with probability less than one in million if $v = \$10$. LLR costs imply that this probability is about 1/60.[28] These are starkly different predictions about behavior which can be tested experimentally.

The finding is not special to this example. Under logistic choice, the probability of making a mistake decays quickly, as v grows, at the exponential rate e^{-v}. We find that as suggested by the examples above, this is at odds with casual introspection and everyday experience about the probability of mistakes. Under the LLR cost function the same probability decreases at the much slower rate $1/v$. This follows from Proposition 7, together with the fact that as v increases, the transformation η approximates the logarithm.

[28]For an alternative interpretation, suppose the decision maker is payed in chance rather than money, so that the payoff v denotes the probability of receiving a prize conditional on making a correct choice. Suppose that when v is 0.05%, the decision maker makes a mistake with probability 20%. Then, if the probability v is increased to 0.5%, the prediction under mutual information is that the decision maker must make a mistake with a probability that is less than one in a million. Under LLR the probability is about 1/60.

7 Bayesian LLR Cost

Given a prior q and an LLR cost function C, one can express the cost of an experiment μ in terms of the distribution π_μ of the posterior belief $p \in \Delta(\Theta)$ that it induces, via

$$C(\mu) = \int F(p) - F(q) \mathrm{d}\pi_\mu(p) \tag{15}$$

where

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

This follows from the definition of the LLR cost, together with Bayes' law, which states that given a prior q and a signal s, the posterior p is given by $\log \frac{p_i}{p_j} = \log \frac{q_i}{q_j} + \log \frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s)$. This reformulation shows that the LLR cost is posterior separable (Caplin and Dean, 2013).

A stronger property studied in the literature is *uniform* posterior separability, where the function F is independent of the prior q and convex. In addition to being standard, this assumption ensures, for instance, that in a dynamic environment an agent is indifferent between performing two experiments—with the choice of the second one perhaps depending on the outcome of the first—and carrying out the Blackwell equivalent one-shot experiment.

As we now show, this assumption can be accommodated in our framework by allowing the cost $C(\mu, q)$ of an experiment μ to be a function the prior, where for each prior the cost function $C(\cdot, q)$ belongs to the LLR family, and the resulting coefficients $(\beta_{ij}(q))$ depend on the prior. While any functional relation between the prior and the coefficients is consistent with LLR cost, there is a unique choice that makes the Bayesian LLR cost function uniformly posterior-separable, as the next proposition shows. An analogous result was derived independently by Bloedel and Zhong (2020).

Proposition 8. *A Bayesian LLR cost function C given by*

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

is uniform posterior separable if and only if there exist positive constants $(b_{ij})_{i,j \in \Theta}$ such that for all priors $q \in \mathcal{P}(\Theta)$ with full support, $\beta_{ij}(q) = b_{ij} q_i$.

Letting

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

and substituting this into (15), we see that Bayesian LLR cost of an experiment can be represented as the expected change of F from the prior q to the posterior p induced by the

signal, for a fixed choice of (b_{ij}). That is, the cost of the experiment equals

$$C(\mu, q) = \int [F(p) - F(q)] \, \mathrm{d}\pi(p), \tag{16}$$

and in particular is uniformly posterior-separable. For a given, fixed prior, this cost is the LLR cost with $\beta_{ij} = b_{ij} q_i$, so that, in terms of the distributions (μ_i), this cost is

$$C(\mu, q) = \sum_{i,j \in \Theta} b_{ij} q_i \int_S \log \frac{\mathrm{d}\mu_i}{\mathrm{d}\mu_j}(s) \, \mathrm{d}\mu_i(s) = \sum_{i,j \in \Theta} b_{ij} q_i D_{\mathrm{KL}}(\mu_i \| \mu_j). \tag{17}$$

Prior Dependence of Bayesian LLR Cost. As we prove in Proposition 8, the only uniformly posterior separable LLR cost potentially assigns different cost to the same experiment at different prior beliefs. We next explore which experiments have prior dependent cost, through a simple example of binary experiments. Consider the standard setting of a binary state space $\Theta = \{1, 2\}$, and an experiment μ with a binary signal which equals the state with some probability $1/2 < r < 1$. For concreteness, imagine a coin whose probability of heads depends on the state and is either r or $1 - r$, and the experiment μ consists of tossing the coin. Consider a Bayesian LLR cost, with $b_{12} = b_{21} = b$. In this case, even though the effective (β_{ij})'s depend on the prior, a simple calculation shows that the cost of the experiment does not, and equals

$$C(\mu, q) = b(2r - 1) \log \frac{r}{1 - r}$$

for every prior q.[29]

Consider now the experiment ν in which the coin is tossed until a "heads" outcome. Under Bayesian LLR costs, the cost can be calculated to be

$$C(\nu, q) = \left(\frac{q_1}{r} + \frac{q_2}{1 - r} \right) C(\mu, q).$$

This cost does depend on the prior: as the above display shows, it is equal to the cost of one toss of the coin, times the expected number of times that it is to be tossed. The latter quantity depends on the prior, in the obvious way. This cost is thus consistent with our additivity axiom, in the sense that this one-shot experiment ν—which is equivalent to a dynamic experiment in which μ is carried out a random number of times—has a cost that equals the expected number of repetition of μ, times the cost of each independent realization of μ.

We generalize the example of a biased coin toss to any experiment μ for which $D_{\mathrm{KL}}(\mu_1 \| \mu_2) = D_{\mathrm{KL}}(\mu_2 \| \mu_1)$. As the next proposition shows, this condition exactly captures

[29] This contrasts with mutual information, where the prior affects the cost of this experiment: the cost is highest for the uniform prior, and vanishes as the prior tends towards certainty.

prior independence of Bayesian LLR costs, in the symmetric case in which $b_{12} = b_{21}$.

Proposition 9. *Let $\Theta = \{1, 2\}$. Let C be a Bayesian LLR cost specified by $b_{12} = b_{21} = b > 0$. Let μ be a Blackwell experiment. Then the following are equivalent.*

(i) $D_{\mathrm{KL}}(\mu_1 \| \mu_2) = D_{\mathrm{KL}}(\mu_2 \| \mu_1)$.

(ii) $C(\mu, q)$ *is independent of the prior* q.

8 Verification and Falsification

All the specifications that we have discussed in the previous sections have the property that $\beta_{ij} = \beta_{ji}$. In this section we explain why some information costs are best modeled by specifications that break this symmetry.

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 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 in Table 1:[30]

[30] 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.

	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 $S = \{s_1, s_2\}$. Under experiment I, observing the signal realization s_2 rejects the hypothesis that the state is a (up to a small probability of error ε^2). Under experiment II, observing s_2 verifies the same hypothesis.

- 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 a minuscule 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.[31]

- 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, 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.

[31] 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.

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, set $\beta_{ae} = \kappa$ and $\beta_{ea} = 0$, and consider small ε. Then, to first order in ε, the cost of experiment I is $\kappa\varepsilon$, while the cost of experiment II is 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.

9 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 mutual information 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 mutual information. 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 2. 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 2 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 costs are flexible enough to accommodate optimal response probabilities that are S-shaped, similarly to our analysis in §6. 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.

10 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.[32] 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):

$$c(\sigma_0, \sigma_1) = G\left[m_{\sigma_0}(1), \ldots, m_{\sigma_0}(N), m_{\sigma_1}(1), \ldots, m_{\sigma_1}(N)\right] \tag{18}$$

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.

[32] 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.

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):

$$c(\sigma_0, \sigma_1) = H\left[\kappa_{\sigma_0}(1), \ldots, \kappa_{\sigma_0}(N), \kappa_{\sigma_1}(1), \ldots, \kappa_{\sigma_1}(N)\right] \tag{19}$$

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 function. 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 (18) and (19). 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.

11 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.

Bloedel, A. W. and W. Zhong (2020). The cost of optimally-acquired information. Technical report, Working paper, Stanford University.

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

Borwein, J. M. and J. D. Vanderwerff (2010). *Convex functions: constructions, characterizations and counterexamples*, Volume 109. Cambridge University Press Cambridge.

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.

Dupuis, P. and R. S. Ellis (2011). *A weak convergence approach to the theory of large deviations*, Volume 902. John Wiley & Sons.

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.

Mu, X., L. Pomatto, P. Strack, and O. Tamuz (2020). From blackwell dominance in large samples to rényi divergences and back again. Forthcoming in Econometrica.

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 characterize all continunous additive linear functionals on the space of all random variables on $\mathbb{R}$.

Nevertheless, the continuity axiom is technical. As we show in Theorem 5 it is not required when there are only two states, and we conjecture that it is not required in general.

Appendix B Preliminaries

In order to simplify the notation, throughout the Appendix we set $\Theta = \{0, 1, \ldots, n\}$.

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 10. *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

$$p_i(s) = \frac{\mathrm{d}\mu_i}{\mathrm{d}\sum_{j \in \Theta} \mu_j}(s) = \frac{1}{\sum_{j \in \Theta} \frac{\mathrm{d}\mu_j}{\mathrm{d}\mu_i}(s)} = \frac{1}{\sum_{j \in \Theta} \mathrm{e}^{\ell_{ji}(s)}} \tag{20}$$

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

A consequence of Proposition 10 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 10,

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

Therefore, by Theorem 1, we have

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

□

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

An additional direct consequence of Proposition 10 is that the LLR cost is convex:

Proposition 11. *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 result follows immediately from the third statement in Proposition 10.

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),

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

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 $y \in \mathcal{D}$. Hence, there must exist a coordinate $i_o j_o$ such that $w_{i_o j_o} > 0$. We now show this leads to a contradiction.

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 $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_n) \in \mathcal{P}(\mathbb{R}^n)^{n+1}$ 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 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

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

and similarly $\mu_0\left(\{\xi : (\ell_1(\xi), \ldots, \ell_n(\xi)) \in E\}\right) = \sigma_0(E)$. So $(\sigma_0, \sigma_1, \ldots, \sigma_n)$ is admissible. □

B.3 Properties of Cumulants

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

Let $A = \{0, \ldots, N\}^n \backslash \{0, \ldots, 0\}$, for some constant $N \in \mathbb{N}$ 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 $\log \varphi_{\sigma_1}$ is well defined and the partial derivatives

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

exist and are continuous for every $\alpha \in A$.

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) (see also Shiryaev, 1996, p. 290) establishes a one-to-one relation between the vector of moments m_{σ_1} and vector of cumulants κ_{σ_1} of a probability measure $\sigma_1 \in \mathcal{P}_A$. Given $\alpha \in A$, let $\Lambda(\alpha)$ be the set of all ordered collections $(\lambda^1, \ldots, \lambda^q)$ of non-zero vectors in $\mathbb{N}^n$ such that $\sum_{p=1}^q \lambda^p = \alpha$.

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

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

The result yields the following implication. Let $M_A = \{m_{\sigma_1} : \sigma_1 \in \mathcal{P}_A\} \subseteq \mathbb{R}^A$ and $K_A = \{\kappa_{\sigma_1} : \sigma_1 \in \mathcal{P}_A\} \subseteq \mathbb{R}^A$. Statement 2 in Theorem 2 shows the existence of a continuous function $h : M_A \to K_A$ such that $\kappa_{\sigma_1} = h(m_{\sigma_1})$ for every $\sigma_1 \in \mathcal{P}_A$. Moreover, statement 1 implies h is one-to-one.

B.4 Cumulants and Admissible Measures

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

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

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

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

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

Lemma 5. *Let I and J be disjoint finite sets and let $(\phi_k)_{k\in I\cup J}$ be a collection of real valued functions defined on $\mathbb{R}^n$. Assume $\{\phi_k : k \in I \cup J\} \cup \{1_{\mathbb{R}^n}\}$ are linearly independent and the unit vector $(1, \ldots, 1) \in \mathbb{R}^J$ belongs to the 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 $F\backslash\{\phi_k\}$, then a standard application of the hyperplane separation theorem implies the existence of 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 $\int \phi_l \, \mathrm{d}\sigma_1 = \int \phi_l \, \mathrm{d}\sigma_2$ for every $l \neq k$. To summarize, we have shown that for every $k \in I \cup J$ there exist vectors $w^k, z^k \in D$ such that $w^k_k > z^k_k$ and $w^k_l = z^k_l$ for $l \neq k$.

Now let $\mathrm{aff}(D)$ be the affine hull of D. As is well known, for every $d \in D$ we have the identity $\mathrm{aff}(D) = d + \mathrm{span}(D - d)$, where $\mathrm{span}(D - d)$ is the vector space generated by $D - d$. Moreover, $\mathrm{span}(D - d)$ is independent of the choice of $d \in D$ (see, for example, Borwein and Vanderwerff, 2010, Lemma 2.4.5).

Let $k \in I \cup J$ and let $1_k \in \mathbb{R}^{I \cup J}$ be the corresponding unit vector. By taking $d = z^k$ we obtain that $w^k - z^k \in \operatorname{span}(D - z^k)$. Thus, $1_k \in \operatorname{span}(D - d)$ for every k. Hence $\operatorname{span}(D - d) = \mathbb{R}^{I \cup J}$. Therefore $\operatorname{aff}(D) = \mathbb{R}^{I \cup J}$. Since D is convex, it has nonempty relative interior as a subset of $\operatorname{aff}(D)$. We conclude that 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 $(\phi_{i,\alpha})_{i \in \Theta}$ on $\mathbb{R}^n$ as

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

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

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

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

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

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

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

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

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

Proof. Let $h : M_A \to K_A$ be the function defined in the discussion following Theorem 2, mapping vectors of moments to vectors of cumulants. Define $H : \mathcal{M} \to \mathcal{K}$ as

$$H(m_\sigma) = (h(m_{\sigma_0}), h(m_{\sigma_1}), \ldots, h(m_{\sigma_n})) = \kappa_\sigma$$

for every $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_n) \in \mathcal{A}$. Since h is continuous and one-to-one then so is H. 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,[33] 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. □

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 $F(x) = F(x_1, \ldots, x_d) = a_1x_1 + \cdots + a_dx_d$ 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 5). 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

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

[33] Brouwer (1911). See also Theorem 2 in Tao (2011).

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. □

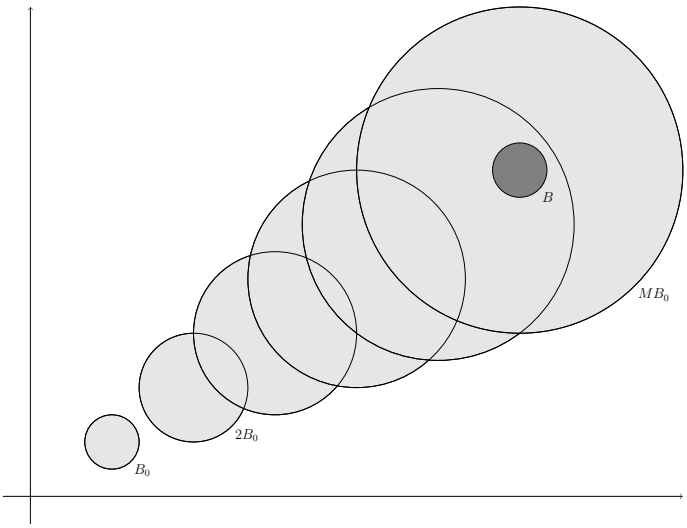

Figure 5: Illustration of the proof of Claim 1. The dark ball B is contained in the light ones, and it is apparent from this image that so is any multiple of B by $a \geq 1$.

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 6). 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 $F(ax) = aF(x)$. □

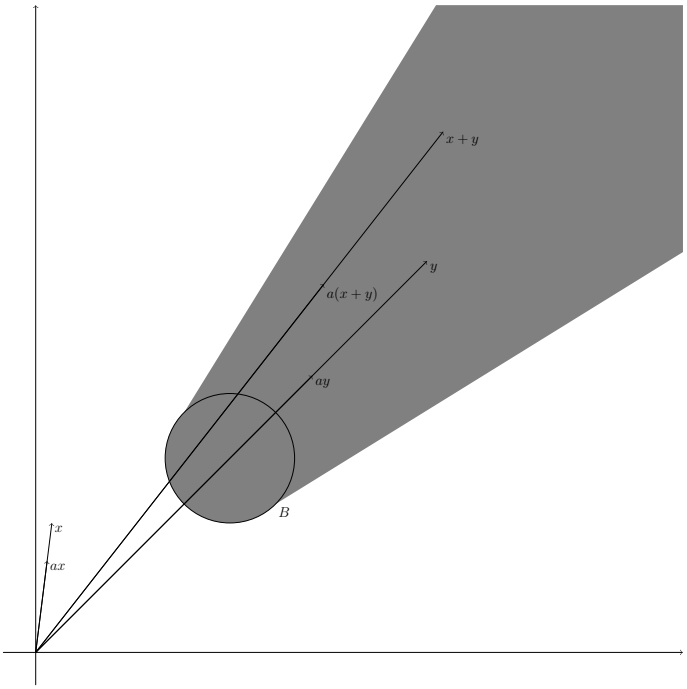

Figure 6: An illustration of the proof of Theorem 4.

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 $A = \{0, \dots, N\}^n \backslash \{0, \dots, 0\}$ 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 $(\ell_{i0})_{i \in \Theta}$. 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 $(\sigma_0 * \tau_0, \dots, \sigma_n * \tau_n) \in \mathcal{A}$;*
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,

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

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

Proof. (1) Let $\mu \in \mathcal{E}$ induce σ and let $\nu \in \mathcal{E}$ induce τ. Then $C(\mu) = c(\sigma), C(\nu) = c(\tau)$ and, by Axiom 2 and Lemma 8, $c(\sigma_0 * \tau_0, \ldots, \sigma_n * \tau_n) = C(\mu \otimes \nu) = c(\sigma) + c(\tau)$. Claim (2) follows directly from Axiom 3 and Lemma 8. □

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

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

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

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

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

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

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

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

For every $\alpha \in A$ we have

$$M_i^{\mu[r]}(\alpha) = \int \ell_{10}^{\alpha_1} \ldots \ell_{n0}^{\alpha_n} \, \mathrm{d}\mu[r]_i = \int_{\mathbb{R}^n} \xi_1^{\alpha_1} \cdots \xi_n^{\alpha_n} \, \mathrm{d}\sigma[r]_i(\xi) = m_{\sigma[r]_i}(\alpha) \tag{21}$$

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

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

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

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

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

Equation (21) therefore implies that $M_i^{\mu[r]}(\alpha) = M_i^{\nu[r]}(\alpha)$. Thus

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

Hence $d_N(\mu[r], \nu[r])$ converges to 0. Since C is uniformly continuous, then $C(\mu[r]) - C(\nu[r]) = 0$ must converge to 0 as well. This implies $C(\mu) = C(\nu)$.

□

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 $\sigma \in \mathcal{A}$. We can use Theorem 2 to define a bijection $H : \mathcal{M} \to \mathcal{K}$ such that $H(m_\sigma) = \kappa_\sigma$. Hence $F = G \circ H^{-1}$ satisfies $c(\sigma) = F(\kappa_\sigma)$ for every σ. For every $\sigma, \tau \in \mathcal{A}$, Lemmas 8 and 9 imply

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

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

Lemma 12. *There exist $(\lambda_{i,\alpha})_{i \in \Theta, \alpha \in A}$ 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 $(\lambda_{i,\alpha})_{i \in \Theta, \alpha \in A}$ 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 $t \in (0,1)$, Lemma 12 and Theorem 2 imply

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

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

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

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

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

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

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

□

Lemma 14. *Let $(\lambda_{i,\alpha})_{i \in \Theta, \alpha \in A}$ 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

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

For all $r \in \mathbb{N}$, 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 $\kappa_{\sigma_i^{*r}} = r \kappa_{\sigma_i}$, we obtain

$$c(\sigma^{*r}) = \sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} r^q \prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) \right) \tag{22}$$

By the additivity of c, $c(\sigma^{*r}) = r c(\sigma)$. Hence, because $\gamma > 1$, $c(\sigma^{*r})/r^\gamma \to 0$ as $r \to \infty$. Therefore, diving (22) by r^γ implies

$$\sum_{i \in \Theta} \sum_{\alpha \in A} \lambda_{i,\alpha} \left(\sum_{(\lambda^1, \ldots, \lambda^q) \in \Lambda(\alpha)} \frac{1}{q!} \frac{\alpha!}{\lambda^1! \cdots \lambda^q!} r^{q-\gamma} \prod_{p=1}^{q} \kappa_{\sigma_i}(\lambda^p) \right) \to 0 \text{ as } r \to \infty. \tag{23}$$

We now show that (23) 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 (23) we have $r^{q-\gamma} \to 0$ as

$r \to \infty$ whenever $q < \gamma$. Hence in order for (23) to hold it must be that

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

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[34]

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

and

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

where $L(\alpha)$ is the cardinality of the set of $(\lambda^1,\ldots,\lambda^q)\in\Lambda(\alpha)$ such that $q=|\alpha|$. We obtain that

$$\sum_{i\in\Theta}\sum_{\alpha\in A:|\alpha|=\gamma}L(\alpha)\lambda_{i,\alpha}\left(\int_{\mathbb{R}^n}\xi_1\,\mathrm{d}\sigma_i\left(\xi\right)\right)^{\alpha_1}\cdots\left(\int_{\mathbb{R}^n}\xi_n\,\mathrm{d}\sigma_i\left(\xi\right)\right)^{\alpha_n}=0. \tag{24}$$

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

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

Then (24) implies that f equals 0 on U. Hence, for every $z\in U$,$i\in\Theta$ and $\alpha\in A$ such that $|\alpha|=\gamma$,

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

hence $\lambda_{i,\alpha}=0$. This contradicts the assumption that $\gamma>1$ and concludes the proof. □

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

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

$(S, (\mu_i))$, the function c satisfies

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

Hence, using the fact that $\frac{\mathrm{d}\mu_j}{\mathrm{d}\mu_0}\frac{\mathrm{d}\mu_0}{\mathrm{d}\mu_i} = \frac{\mathrm{d}\mu_j}{\mathrm{d}\mu_i}$, we obtain

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

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

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

Appendix E Proofs of the Results of Section 6

Proof of Proposition 3. Let $\mu^\star \in \mathcal{P}(A)^n$ be an optimal experiment. Let $A^\star = \mathrm{supp}(\mu^\star)$ be the set of actions played in $\mu^\star$. It solves

$$\max_{\mu\in\mathbb{R}_+^{|\Theta|\times|A^\star|}} \left[\sum_{i\in\Theta} q_i\left(\sum_{a\in A}\mu_i(a)u(a,i)\right) - \sum_{i,j\in\Theta}\beta_{ij}\sum_{a\in A^\star}\mu_i(a)\log\frac{\mu_i(a)}{\mu_j(a)}\right] \tag{25}$$

$$\text{subject to} \quad \sum_{a\in A^\star}\mu_i(a) = 1 \text{ for all } i\in\Theta. \tag{26}$$

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 distributions to pairs of vectors in $\mathbb{R}_+^{|A^\star|}$. Moreover, expected utility is linear in the choice probabilities. It then follows that the objective function in (25) is concave over $\mathbb{R}_+^{|\Theta|\times|A^\star|}$.

As (25) 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 solution to (25), it follows from the Karush-Kuhn-Tucker theorem 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 $a \in A^\star$

$$0 = q_i u_i(a) - \lambda_i - \sum_{j\neq i} \left\{\beta_{ij}\left[\log\left(\frac{\mu_i^\star(a)}{\mu_j^\star(a)}\right) - 1\right] - \beta_{ji}\frac{\mu_j^\star(a)}{\mu_i^\star(a)}\right\}\,. \tag{27}$$

Subtracting (27) evaluated at a' from (27) evaluated at a yields the desired necessary conditions for the optimality of $\mu^\star$. □

Proof of Proposition 4. We prove a slightly more general result. Assume the coefficients satisfy $\beta_{ij} \geq 1/f(d(i,j))^2$, where f is a strictly positive and increasing function f.

The cost of the optimal experiment $\mu^\star$ must satisfy $\|u\| \geq C(\mu^\star)$, otherwise the decision maker would be made better off by acquiring no information. Pinsker's inequality (see Borwein and Vanderwerff, 2010, p. 13) implies

$$C(\mu^\star) \geq \min\{\beta_{ij},\beta_{ji}\}(D_{\mathrm{KL}}(\mu_i^\star\|\mu_j^\star) + D_{\mathrm{KL}}(\mu_j^\star\|\mu_i^\star)) \geq \min\{\beta_{ij},\beta_{ji}\}\|\mu_i^\star - \mu_j^\star\|_1^2.$$

where $\|\mu_i^\star - \mu_j^\star\|_1 = \sum_{a\in A} |\mu_i^\star(a) - \mu_j^\star(a)|$ denotes the total-variation norm between the two distributions. We then obtain

$$\|\mu_i^\star - \mu_j^\star\|_1 \leq \sqrt{\|u\|\frac{1}{\min\{\beta_{ij},\beta_{ji}\}}} \leq \sqrt{\|u\|}f(d(i,j))\,.$$

In particular, if f is the identity function then $\|\mu_i^\star - \mu_j^\star \leq \sqrt{\|u\|}d(i,j)$. □

Proof of Proposition 6. Given a vector $\mu \in \mathcal{P}(\{B,R\})^2$, we simply denote by μ_i the probability $\mu_i(B)$ of guessing B in state i. For every μ, let

$$U(\mu) = \frac{1}{2r}\left(\sum_{i<n/2}(1-\mu_i) + \sum_{i>n/2}\mu_i\right) - C(\mu) \tag{28}$$

be the net expected payoff provided by μ, where C is a LLR cost function such that

$\beta_{ij} = f(|i-j|)$ for some positive and strictly decreasing function f.

Let $\mathcal{P}_+$ be the set of probabilities μ such that $\text{supp}(\mu) = \{B, R\}$. Let $\mu^\star$ be a solution to the problem $\max_{\mu \in \mathcal{P}_+} U(\mu)$. Such a solution exists and is unique. Indeed, the problem $\max_{\mu \in \mathcal{P}(\{B,R\})^2} U(\mu)$ has a solution. Now, if $\mu^\star$ is optimal and $\mu^\star \notin \mathcal{P}_+$, then either $\mu_i^\star = 0$ for every i or $\mu_i^\star = 1$ for every i. In either case $U(\mu^\star) = U(\mu)$, where $\mu \in \mathcal{P}_+$ is defined as $\mu_i = 1/2$ for every i. It follows that the problem $\max_{\mu \in \mathcal{P}_+} U(\mu)$ admits a solution $\mu^\star$. Over $\mathcal{P}_+$ the function C is strictly convex, and thus U is strictly concave. Thus, the solution is unique.

We claim $\mu^\star$ satisfies $\mu^\star_{n/2+r} = 1 - \mu^\star_{n/2-r}$ for every r. To see this, define $\mu \in \mathcal{P}_+$ as $\mu_{n/2+r} = 1 - \mu^\star_{n/2-r}$ for every r. Because $U(\mu^\star) = U(\mu)$ and U is strictly concave on $\mathcal{P}_+$, we conclude that $\mu = \mu^\star$, completing the proof of the claim.

Let $I \subseteq \mathcal{P}(\{B, R\})^2$ be the set of vectors μ that are increasing, i.e. satisfy $\mu_i \leq \mu_{i+1}$ for every $i \in \Theta \backslash \{n/2 + r\}$, and consider the optimization problem

$$\max_{\mu \in I \cap \mathcal{P}_+} U(\mu).$$

The set I is closed and U is upper semi-continuous. Thus, the problem $\max_{\mu \in I} U(\mu)$ has a solution. The same argument applied in the previous paragraph implies $\max_{\mu \in I \cap \mathcal{P}_+} U(\mu)$ admits a solution as well, and that such a solution is unique. We denote it by $\hat{\mu}$.

As we show in the next paragraph, the vector $\hat{\mu}$ is strictly increasing: it satisfies $\hat{\mu}_i < \hat{\mu}_{i+1}$ for every i. This implies $\mu^\star = \hat{\mu}$. Indeed, by definition we have $U(\mu^\star) \geq U(\hat{\mu})$. If $U(\mu^\star) > U(\hat{\mu})$ the concavity of U implies $U(\alpha\mu^\star + (1-\alpha)\hat{\mu}) > U(\hat{\mu})$ for all $\alpha \in [0, 1]$. Because $\hat{\mu}$ is strictly increasing, then for α small enough the vector $\alpha\mu^\star + (1-\alpha)\hat{\mu}$ belongs to I, contradicting the optimality of $\hat{\mu}$. It follows that $U(\mu^\star) = U(\hat{\mu})$, and hence $\mu^\star = \hat{\mu}$, since the problem $\max_{\mu \in \mathcal{P}_+} U(\mu)$ has a unique solution.

We now show $\hat{\mu}$ is strictly increasing. Given $\nu, \rho \in (0, 1)$ we denote by $D_1(\nu \| \rho)$ and $D_2(\nu \| \rho)$ the partial derivatives of the Kullback-Leibler divergence D_{KL} with respect to its the first and second arguments:

$$D_1(\rho \| \nu) = \log \frac{\rho}{\nu} - \log \frac{1-\rho}{1-\nu}$$
$$D_2(\rho \| \nu) = -\frac{\rho}{\nu} + \frac{1-\rho}{1-\nu}.$$

Both derivatives are equal to zero if and only if $\nu = \rho$.

As a way of contradiction, suppose $\hat{\mu}$ is not strictly increasing. Let $[i, k]$ be a maximal interval of states over which $\hat{\mu}$ is constant. Let μ^ϵ be the vector obtained from $\hat{\mu}$ by increasing $\hat{\mu}_k$ by ϵ and decreasing $\hat{\mu}_i$ by ϵ (since $\hat{\mu} \in \mathcal{P}_+$, both operations are feasible). The

function $\epsilon \mapsto U(\mu^\epsilon)$ is differentiable. Its derivative at $\epsilon = 0$ is equal to

$$\frac{\operatorname{sgn}(k-n/2)}{2r} - \sum_{j\neq k} \beta_{jk}(D_2(\hat{\mu}_j\|\hat{\mu}_k) + D_1(\hat{\mu}_k\|\hat{\mu}_j)) - \frac{\operatorname{sgn}(i-n/2)}{2r} + \sum_{j\neq i} \beta_{ij}(D_2(\hat{\mu}_j\|\hat{\mu}_i) + D_1(\hat{\mu}_k\|\hat{\mu}_i))\,. \tag{29}$$

Since $\hat{\mu}$ is constant in the interval $[i,k]$, then $D_1(\hat{\mu}_j\|\hat{\mu}_m) = D_2(\hat{\mu}_j\|\hat{\mu}_m)$ whenever $i \leq j \leq m \leq k$. We can therefore rewrite (29) as

$$\begin{aligned}&\frac{\operatorname{sgn}(k-n/2)}{2r} - \sum_{j>k} \beta_{jk}(D_2(\hat{\mu}_j\|\hat{\mu}_k) + D_1(\hat{\mu}_k\|\hat{\mu}_j)) - \sum_{j<i} \beta_{jk}(D_2(\hat{\mu}_j\|\hat{\mu}_k) + D_1(\hat{\mu}_k\|\hat{\mu}_j))\\ &-\frac{\operatorname{sgn}(i-n/2)}{2r} + \sum_{j>k} \beta_{ij}(D_2(\hat{\mu}_j\|\hat{\mu}_i) + D_1(\hat{\mu}_k\|\hat{\mu}_i)) + \sum_{j<i} \beta_{ij}(D_2(\hat{\mu}_j\|\hat{\mu}_i) + D_1(\hat{\mu}_i\|\hat{\mu}_j))\,.\end{aligned} \tag{30}$$

The derivative (30) is strictly positive. Indeed, because $k \geq i$ then $\operatorname{sgn}(k-n/2) - \operatorname{sgn}(i-n/2) \geq 0$. Whenever $j > k$, since $\hat{\mu}_j > \hat{\mu}_k = \hat{\mu}_i$ and D is strictly convex over $\mathcal{P}_+$, we have

$$D_2(\hat{\mu}_j\|\hat{\mu}_k) = D_2(\hat{\mu}_j\|\hat{\mu}_i) < 0 \text{ and } D_1(\hat{\mu}_k\|\hat{\mu}_j) = D_1(\hat{\mu}_i\|\hat{\mu}_j) < 0$$

Moreover $\beta_{jk} > \beta_{ji}$ since $|j-k| < |i-k|$. It follows that

$$-\sum_{j>k} \beta_{jk}(D_2(\hat{\mu}_j\|\hat{\mu}_k) + D_1(\hat{\mu}_k\|\hat{\mu}_j)) + \sum_{j>k} \beta_{ij}(D_2(\hat{\mu}_j\|\hat{\mu}_i) + D_1(\hat{\mu}_i\|\hat{\mu}_j))$$

is strictly positive if $k < n/2 + r$, and equal to 0 if $k = n/2 + r$. An analogous argument shows that

$$-\sum_{j<i} \beta_{jk}(D_2(\hat{\mu}_j\|\hat{\mu}_k) + D_1(\hat{\mu}_k\|\hat{\mu}_j)) + \sum_{j<i} \beta_{ij}(D_2(\hat{\mu}_j\|\hat{\mu}_i) + D_1(\hat{\mu}_i\|\hat{\mu}_j))$$

is strictly positive if $i > n/2 - r$, and equal to 0 if $i = n/2 - r$. Because $\hat{\mu} \in \mathcal{P}_+$, then either $k < n/2 + r$, $i > n/2 - r$, or both. This implies that (30) is strictly positive. Hence, for small enough ϵ, the vector μ^ϵ satisfies $U(\mu^\epsilon) > U(\hat{\mu})$, contradicting the hypothesis that $\hat{\mu}$ is optimal. We therefore conclude that $\hat{\mu}$ is strictly increasing, and thus $\mu^\star$ is strictly increasing as well.

Because $\mu^\star$ satisfies $\mu^\star_{n/2+r} = \mu^\star_{n/2-r}$ for every r, and $\mu^\star$ is stricly increasing, it follows that $m_i > m_j$ for every pair of states such that $|i - n/2| > |j - n/2|$. □

Proof of Proposition 7. Denote by $\mathcal{P}_+$ be the set of probabilities $\mu' \in \mathcal{P}(\{a_1, a_2\})^2$ such that $\operatorname{supp}(\mu) = \{a_1, a_2\}$, and let $\mu \in \mathcal{P}_+$ be an optimal experiment. We first show that μ satisfies $\mu_1(a_1) = \mu_2(a_2)$. To see this, define μ' as $\mu'_1(a_1) = \mu_2(a_2)$ and $\mu'_2(a_2) = \mu_1(a_1)$.

Let $\mu'' = \frac{1}{2}\mu + \frac{1}{2}\mu'$. By the symmetry of the payoffs functions and of the prior, we have

$$\sum_{i\in\Theta} q_i \left(\sum_{a\in A} \mu_i(a)u(a,i)\right) = \sum_{i\in\Theta} q_i \left(\sum_{a\in A} \mu'_i(a)u(a,i)\right) = \sum_{i\in\Theta} q_i \left(\sum_{a\in A} \mu''_i(a)u(a,i)\right).$$

Moreover, $C(\mu'') \leq \frac{1}{2}C(\mu) + \frac{1}{2}C(\mu')$ if $\mu \neq \mu'$, as C is strictly convex on $\mathcal{P}$. Since μ is optimal, it must be that $\mu = \mu'$.

The optimality equation $\mathrm{MB}_1(a_1, a_2) = \mathrm{MC}_1(a_1, a_2)$ can now be rewritten as

$$\frac{1}{2}v = \beta\left[\xi\left(\log\left(\frac{\mu_1(a_1)}{\mu_2(a_1)}\right)\right) - \xi\left(\log\left(\frac{\mu_1(a_2)}{\mu_2(a_2)}\right)\right)\right].$$

Simple calculations show the expression is in turn equal to

$$\frac{v}{2\beta} = \xi\left(\log\left(\frac{\mu[v]}{1-\mu[v]}\right)\right) - \xi\left(\log\left(-\frac{\mu[v]}{1-\mu[v]}\right)\right) = \zeta\left(\log\left(\frac{\mu[v]}{1-\mu[v]}\right)\right)$$

where $\zeta(x) = 2x + e^x - e^{-x}$. The result now follows by defining $\eta = \zeta^{-1}$. □

Proof of Proposition 5. Consider a decision problem described by a payoff function u and a prior q. let μ and μ' be the optimal choice probabilities obtained under the coefficients (β_{ij}) and (β'_{ij}). The optimality of μ and μ' implies

$$\sum_{i,a} q_i u(i,a)\mu_i(a) - \sum_{i,j} \beta_{ij} D(\mu_i \| \mu_j) \geq \sum_{i,a} q_i u(i,a)\mu'_i(a) - \sum_{i,j} \beta_{ij} D(\mu'_i \| \mu'_j)$$
$$\sum_{i,a} q_i u(i,a)\mu'_i(a) - \sum_{i,j} \beta'_{ij} D(\mu'_i \| \mu'_j) \geq \sum_{i,a} q_i u(i,a)\mu_i(a) - \sum_{i,j} \beta'_{ij} D(\mu_i \| \mu_j)$$

Rearranging the two inequalities leads to

$$\sum_{i,j} \beta_{ij}(D(\mu'_i\|\mu'_j) - D(\mu_i\|\mu_j)) \geq \sum_{i,a} q_i u(i,a)(\mu'_i(a) - \mu_i(a)) \geq \sum_{i,j} \beta'_{ij}(D(\mu'_i\|\mu'_j) - D(\mu_i\|\mu_j)).$$

The result now follows. □

Appendix F Additional Proofs

Proof of Proposition 2. Let $|\Theta| = n$. By Axiom a there exists a function $f\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\beta^{\Theta}_{ij} = f(|i-j|)$. Let $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ be given by $g(t) = \frac{1}{2}f(t)t^2$. 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 $\Theta \in \mathcal{T}$

$$\kappa = C^{\Theta}(\nu^{\Theta}) = \sum_{i \neq j \in \Theta} \beta_{ij}^{\Theta} \frac{(i-j)^2}{2} = \sum_{i \neq j \in \Theta} g(|i-j|). \tag{31}$$

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 (31) 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 (31) 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

$$g(x) - g(y) = g(x+z) - g(y+z) \qquad \text{for all } x, y, z > 0. \tag{32}$$

Since g is non-negative, it follows from (31) 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 (32) 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. □

Appendix G The cost of bounded experiments with binary state

In this section we restrict ourselves to the case of a binary state space $\Theta = \{0, 1\}$, and the class of *bounded* experiments $\mathcal{B}$: an experiment is said to be bounded if the beliefs that it induces are bounded away from 0 and 1. In terms of log-likelihood ratios, it is bounded if there is some M such that $\ell_{01}(s)$ is μ_0- and μ_1-almost surely in $[-M, M]$. The class of bounded experiments is contained in the class $\mathcal{E}$ of experiments considered in the rest of the paper. The bounded experiments contain all the experiments that have a finite set of possible realizations, and in which not state is ever conclusively excluded.

As we discuss above, a strengthening of Axiom 1 is Blackwell monotonicity: C is said to be Blackwell monotone if $C(\mu) \geq C(\nu)$ whenever If μ Blackwell dominates ν.

For the class of bounded experiments, we show that 2 and 3 are sufficient for proving that a Blackwell monotone cost is an LLR cost: the continuity axiom 4 is not needed. This proof heavily relies on a recent result of Mu, Pomatto, Strack, and Tamuz (2020), which characterizes the monotone and additive functions on the class of bounded Blackwell experiments with binary state. An extension of this result to large state spaces is currently out of reach, and so we do not have a more general proof. Nevertheless, we conjecture that the continuity axiom is generally redundant.

Theorem 5. *Let $\Theta = \{0, 1\}$. A Blackwell monotone information cost function $C \colon \mathcal{B} \to \mathbb{R}_+$ satisfies Axioms 2 and 3 if and only if there exist $\beta_{01}, \beta_{10} \geq 0$ such that for every experiment $\mu \in \mathcal{B}$,*

$$C(\mu) = \beta_{01} D_{\mathrm{KL}}(\mu_0 \| \mu_1) + \beta_{10} D_{\mathrm{KL}}(\mu_1 \| \mu_0).$$

Before proving Theorem 5, we will introduce some definitions and results from Mu et al. (2020).

For $t \in (0, \infty]$, we denote by $R_t(\mu_0 \| \mu_1)$ the Rényi t-divergence between two probability μ_0, μ_1 defined on the same measurable space S. For $t \neq 1$, $t \neq \infty$,

$$R_t(\mu_0 \| \nu_1) = \frac{1}{t-1} \log \int_S \left(\frac{\mathrm{d}\mu_0}{\mathrm{d}\mu_1}(s) \right)^{t-1} \mathrm{d}\mu_0(s).$$

For $t = 1$

$$R_1(\mu_0 \| \mu_1) = \int_S \log \frac{\mathrm{d}\mu_0}{\mathrm{d}\mu_1}(s)\, \mathrm{d}\mu_0(s) = D_{\mathrm{KL}}(\mu_0 \| \mu_1).$$

For $t = \infty$, $R_\infty(\mu_0 \| \mu_1)$ is the essential maximum of the log-likelihood ratio $\log \frac{\mathrm{d}\mu_0}{\mathrm{d}\mu_1}$. Note that $R_t(\mu_0 \| \mu_1)$ is always non-negative, and positive whenever $\mu_0 \neq \mu_1$.

The following result is a reformulation of Theorem 2 in Mu et al. (2020) (see also Lemmas 5 and 6).[35]

Theorem 6 (Mu et al. 2020)**.** *An information cost function $C\colon \mathcal{B} \to \mathbb{R}_+$ satisfies Axioms 1 and 2 if and only if there exist two finite Borel measures m_0, m_1 on $[1/2, \infty]$ such that for every bounded experiment $\mu = (S, \mu_0, \mu_1)$ it holds that*

$$C(\mu) = \int_{[1/2,\infty]} R_t(\mu_0 \| \mu_1)\, \mathrm{d}m_0(t) + \int_{[1/2,\infty]} R_t(\mu_1 \| \mu_0)\, \mathrm{d}m_1(t).$$

Using this result, we can now prove Theorem 5.

Proof of Theorem 5. The argument that this representation satisfies the axioms is identical to the same argument in the proof of Theorem 1. It thus remains to be shown that the representation is implied by the axioms.

By Theorem 6,

$$\begin{aligned} C(\mu) = {} & \beta_{01} D_{\mathrm{KL}}(\mu_0 \| \mu_1) + \beta_{10} D_{\mathrm{KL}}(\mu_1 \| \mu_0) \\ & + \int_{[1/2,1)} R_t(\mu_0 \| \mu_1)\, \mathrm{d}m_0(t) + \int_{[1/2,1)} R_t(\mu_1 \| \mu_0)\, \mathrm{d}m_1(t) \\ & + \int_{(1,\infty]} R_t(\mu_0 \| \mu_1)\, \mathrm{d}m_0(t) + \int_{(1,\infty]} R_t(\mu_1 \| \mu_0)\, \mathrm{d}m_1(t). \end{aligned} \tag{33}$$

for some $\beta_{01}, \beta_{10} \geq 0$ and m_0, m_1 finite Borel measures on $[1/2, \infty]$ that assign measure 0 to the singleton $\{1\}$. To prove the claim, we show that m_0 and m_1 are the zero measures.

Let $\mu = (S, \mu_0, \mu_1)$ be a non-trivial bounded experiment, and let $\nu = (1/r) \cdot \mu^{\otimes r}$ for some r. It follows from the definition of Rényi t-divergences that for $t \neq 1$, $t \neq \infty$

$$R_t(\nu_0 \| \nu_1) = \frac{1}{t-1} \log \left(\frac{r-1}{r} + \frac{1}{r} \left(\int_S \left(\frac{\mathrm{d}\mu_0}{\mathrm{d}\mu_1}(s) \right)^{t-1} \mathrm{d}\mu_0(s) \right)^r \right).$$

Now, for $x > 1$,

$$\lim_{r \to \infty} \log \left(\frac{r-1}{r} + \frac{1}{r} x^r \right) = \infty,$$

[35] The *data processing inequality* in that paper is monotonicity with respect to deterministic garblings, which is implied by Blackwell monotonicity. The additivity there translates immediately to additivity in the sense of Axiom 2.

and for $x < 1$ this same limit is 0. It thus follows that for $t > 1$ (including, trivially, $t = \infty$)

$$\lim_{r\to\infty} R_t(\nu_0\|\nu_1) = \infty, \tag{34}$$

since R_t is positive for non-trivial experiments, and so the integral in the expression for R_t is strictly greater than 1. For $t < 0$

$$\lim_{r\to\infty} R_t(\nu_0\|\nu_1) = 0, \tag{35}$$

since, again by the positivity of R_t, the integral in the expression for R_t is strictly less than 1.

It follows from (34) that both m_0 and m_1 must assign no mass to $(1,\infty]$, i.e. $m_0((1,\infty]) = m_1((1,\infty]) = 0$, since otherwise the integral $\int_{(1,\infty]} R_t(\mu_0\|\mu_1)\,\mathrm{d}m_0(t)$ or $\int_{(1,\infty]} R_t(\mu_0\|\mu_1)\,\mathrm{d}m_1(t)$ would diverge and by (33) the cost of the experiment $(1/r)\cdot\mu^{\otimes r}$ would diverge

$$\lim_{r\to\infty} C((1/r)\cdot\mu^{\otimes r}) = \infty\,.$$

This would contradict the axioms which imply that $C((1/r)\cdot\mu^{\otimes r}) = C(\mu)$. It then follows from (35) that $m_0((1/2,1)) = m_1((1/2,1)) = 0$, since otherwise

$$\lim_{r\to\infty} C((1/r)\cdot\mu^{\otimes r}) < C(\mu).\ \square$$

Appendix H Uniform Separable Bayesian LLR Cost

Proof of Proposition 8. It is straightforward to verify that if the parameters satisfy $\beta_{ij}(q) = \gamma_{ij}q_i$, then C is uniformly posterior separable. We now prove the opposite implication.

Fix a prior q with full support, and consider an experiment μ where the set of signal realizations is a product $S_1 \times S_2$, with S_1 a finite set, and each μ_i satisfies $\mu_i(\{s\}\times S_2) > 0$ for every $s \in S_1$. We denote by μ_i^1 the marginal of μ_i on S_1, and by $\mu_i(\cdot|s)$ the measure on S_2 obtained by conditioning μ_i on $s \in S_1$.

The chain rule for the KL-divergence implies that the cost of such an experiment can be written as

$$C(\mu,q) = \sum_{ij}\beta_{ij}(q)\left[D_{\mathrm{KL}}(\mu_i^1\|\mu_j^1) + \sum_{s_1\in S_1}\mu_i^1(s_1)D_{\mathrm{KL}}(\mu_i(\cdot|s_1)\|\mu_j(\cdot|s_1))\right]. \tag{36}$$

Now assume C is uniformly posterior separable with respect to a function G. The cost of the experiment μ can then be written as follows. It will be convenient to denote posterior beliefs as random variables defined over the probability space $(\Theta\times S_1\times S_2,\mathbb{P})$ where $\mathbb{P}$ is obtained from q and μ in the obvious way. Let p^2 be the posterior belief over Θ obtained by conditioning q on a realization (s_1,s_2), and let p^1 be the posterior belief

obtained by conditioning q on a realization s_1. Then

$$\begin{aligned} C(\mu, q) &= \mathbb{E}\left[G(p^2) - G(p^1) - G(p^1) - G(q)\right] \\ &= \mathbb{E}\left[G(p^1) - G(q)\right] + \sum_{s_1 \in S_1} \mathbb{P}(s_1)\mathbb{E}\left[G(p^2) - G(p^1)|p^1 = q(\cdot|s_1)\right]. \end{aligned}$$

Now consider the experiment $((\mu_i^1), S)$ which consists of observing the first realization s_1 but not the second. By uniform posterior separability, its cost, at the prior q, is given by

$$\mathbb{E}\left[G(p^1) - G(q)\right] = \sum_{ij} \beta_{ij}(q) D_{\mathrm{KL}}(\mu_i^1 \| \mu_j^1).$$

Given a realization $s_1 \in S$, consider the experiment $((\mu_i(\cdot|s_1)), S_2)$. By considering now $p^1 = q(\cdot|s_1)$ as a prior, uniform separability implies that the cost of the experiment $((\mu_i(\cdot|s_1)), S_2)$ is equal to

$$\mathbb{E}\left[G(p^2) - G(p^1)|p^1 = q(\cdot|s^1)\right] = \sum_{ij} \beta_{ij}(q(\cdot|s_1)) D_{\mathrm{KL}}(\mu_i(\cdot|s_1) \| \mu_j(\cdot|s_1)).$$

The last two equations imply that the cost $C(\mu, q)$ can be rewritten as

$$\sum_{ij} \beta_{ij}(q) D_{\mathrm{KL}}(\mu_i^1 \| \mu_j^1) + \sum_{s_1 \in S_1} \mathbb{P}(s_1) \left(\sum_{ij} \beta_{ij}(q(\cdot|s_1)) D_{\mathrm{KL}}(\mu_i(\cdot|s_1) \| \mu_j(\cdot|s_1)) \right). \tag{37}$$

This equation can be interpreted as saying that the cost of running the experiment μ is equal to the cost of running the first experiment $((\mu_i^1), S_1)$ plus the expected cost of running the second experiment $((\mu_i(\cdot|s_1)), S_2)$, conditional on the signal realization s_1 from the first experiment. By equating (36) and (37) we obtain that

$$\sum_{s_1 \in S_1} \sum_{ij} \left[\beta_{ij}(q) \mu_i^1(s_1) - \mathbb{P}(s_1) \beta_{ij}(q(\cdot|s_1)) \right] D_{\mathrm{KL}}(\mu_i(\cdot|s_1) \| \mu_j(\cdot|s_1)) = 0. \tag{38}$$

Given a particular realization $s_1 \in S_1$, we are free to choose μ such that all the conditional experiments $((\mu_i(\cdot|s_1')), S_2)$, $s_1' \neq s_1$, are completely uninformative, and hence have cost 0. Thus, it must hold that for every $s_1 \in S_1$,

$$\sum_{ij} \left[\beta_{ij}(q) \mu_i^1(s_1) - \mathbb{P}(s_1) \beta_{ij}(q(\cdot|s_1)) \right] D_{\mathrm{KL}}(\mu_i(\cdot|s_1) \| \mu_j(\cdot|s_1)) = 0.$$

By Lemma 2, the latter can hold only if

$$\beta_{ij}(q) \mu_i^1(s_1) = \mathbb{P}(s_1) \beta_{ij}(q(\cdot|s_1)).$$

By dividing and multiplying the left-hand side by q_i and then applying Bayes' rule we obtain that

$$\frac{\beta_{ij}(q)}{q_i} = \frac{\beta_{ij}(q(\cdot|s_1))}{q(\cdot|s_1)}.$$

Given any $q' \in \mathcal{P}(\Theta)$ with full support, we can choose μ such that $q(\cdot|s_1) = q'$ for some s_1. The conclusion now follows by defining $\gamma_{ij} = \beta_{ij}(q)/q_i$. □

Proof of Proposition 9. Under the assumption that $b_{12} = b_{21} = b > 0$, the cost of an experiment μ at prior q is

$$C(\mu, q) = b\left[q_1 D_{\mathrm{KL}}(\mu_1 \| \mu_2) + q_2 D_{\mathrm{KL}}(\mu_2 \| \mu_1)\right].$$

Clearly, this quantity depends on q if and only if $D_{\mathrm{KL}}(\mu_1 \| \mu_2) \neq D_{\mathrm{KL}}(\mu_2 \| \mu_1)$. □