Economies and Diseconomies of Scale in the Cost of Information *

Christian Baker †

October 30, 2023

Please click here for the latest version

Abstract

We present a theory of information costs that explores the notion of economies and diseconomies of scale in information acquisition. We formulate a cost function that generalizes the constant marginal cost function in Pomatto et al. (2023). We demonstrate how to apply our cost function to decision problems. While economies of scale will push a decision maker toward a balanced learning strategy, diseconomies of scale will encourage a decision maker to focus her attention on a limited number of states.

Keywords: Information Acquisition, Marginal Costs

JEL Codes: D83

1 Introduction

A central theme in economic theory is that information is scarce. A comprehensive understanding of how this scarcity impacts decision making requires endogenizing the choice of information structure. It is important, then, to formulate reasonable cost

*I would like to thank Benjamin Brooks, Doron Ravid, Philip Reny, and Emir Kamenica for their comments and insights.

†cbaker8128@gmail.com

functions for information structures. For conventional commodities, cost functions are commonly characterized by whether or not they exhibit economies of scale. In this paper, we present a tractable reduced form cost function that can be used to model diseconomies of scale.

Recent progress in this direction has been made by Pomatto et al. (2023) who present an axiomatic theory of information acquisition with constant marginal costs. Their axioms determine the cost of an experiment up to a vector of parameters, characterized by a log-likelihood ratio (LLR) cost function. An important feature of their axioms is the assumption of constant marginal costs: costs are additive with respect to experiments that are independent conditional on the state. This implies, for instance, that if the variable is the potency of a drug, and information is gathered by testing the drug on volunteers, then the cost of recruiting an additional volunteer is constant. However, if the drug is being used to treat a particularly rare disease, then doubling the number of participants may well more than double the costs.

The question then arises of how best to capture the notion of nonconstant marginal costs in an information acquisition setting. A standard approach to studying cost functions for traditional goods is to classify them in terms of constant, increasing, or decreasing marginal costs. Hence a natural direction is to consider cost functions that are super-additive or sub-additive with respect to the number of independent experiments. We generalize the axiom of constant marginal cost in Pomatto et al. (2023) and arrive at a specific functional form for information costs that includes the LLR cost function as a special case.

When modeling the cost of information, an important question is whether we are considering the cost of acquiring a *belief*, or the cost of performing an *experiment*. For instance, following the influential work of Sims (2003) on rational inattention, cost functions based on mutual information have been widely applied. Under this approach, the cost of an experiment is given by the expected change in the Shannon entropy between prior belief and posterior belief. In contrast, we follow the experiment-based approach of Denti et al. (2022) and consider experimental cost functions, in which the cost depends solely on the experiment performed.

Our main result characterizes all cost functions over Blackwell experiments that satisfy three main conditions. Let π and ρ be two conditionally independent and identically distributed Blackwell experiments. First, the cost $C(\pi)$ of π is greater than the cost $C(\rho)$ of ρ if π is more informative than ρ in the sense of Blackwell (1951). Second, the cost of performing π with probability β and otherwise performing ρ is equal to $\beta C(\pi) + (1 - \beta)C(\rho)$. Third, the cost of performing both experiments is equal to $C(\pi) + C(\rho) + \lambda C(\pi)C(\rho)$ for some $\lambda \in \mathbb{R}$. This third axiom allows us to consider nonconstant marginal costs. When $\lambda = 0$, costs are additive and we arrive at the same LLR cost function from Pomatto et al. (2023). But when $\lambda > 0$ costs are super-additive, and when $\lambda < 0$ costs are sub-additive. This pseudo-additivity property already exists in the literature on information theory. For instance, Tsallis entropy, a generalization of Shannon entropy, satisfies this property (Tsallis, 1988).

The parameters of the cost function permit a straightforward interpretation that is amenable to applications. Given a finite state space Θ, there is one parameter for each state, $\alpha_i \in \mathbb{R}$, and a scaling parameter, λ. The parameter α_i corresponds to the difficulty of learning about state i; higher values reflect lower complexity.

An implication of our model is that diseconomies of scale can lead to unbalanced decision making between states. As an illustration, consider a pharmaceutical firm performing a clinical trial of a new drug. If the drug is effective the firm wishes to begin manufacturing it, but if it is ineffective then they will abandon the project. In the face of increasing marginal costs for recruiting new test subjects, it may be quite cheap to perform a small-scale trial but prohibitively expensive to perform a large-scale trial. Such experiments tend to have low statistical power, a high probability of a Type II error (false negative), and a low probability of a Type I error (false positive). There will then be a low chance of taking the correct action in the state of the world where the drug truly is effective, and a relatively high chance of taking the correct action when the drug is not effective.

If, on the other hand, the cost of recruiting subjects exhibits constant returns to scale, then the best experiment would be one which strikes a balance between the likelihood of a Type I error and a Type II error. The probability of taking the correct action will be balanced between the two states.

The cost function we present is not the only way to approach modeling diseconomies of scale in information acquisition. However, its relatively simple form and parameterization make it serve as a useful benchmark for this topic. For instance, we show that this cost function has a posterior-separable representation (Denti, 2022).

In Section 2 we introduce the basic model of information costs. Our main results are presented in Section 3. Section 4 provides applications of our framework to decision problems. Section 5 discusses the relationship of our results to the literature on rational inattention. Section 6 concludes.

Related Literature Caplin et al. (2022) provide an axiomatic foundation for representing information acquisition costs via a posterior-separable function. Additionally, Denti (2022) provides testable conditions for posterior separability of cost functions. The cost functions characterized in our main result would have a posterior-separable representation if equipped with a full-support prior. However, our cost functions do not have what is termed a *uniform* posterior-separable representation. In general, a uniform posterior-separable cost function cannot be represented as a prior-independent experiment cost function. Hébert and Woodford (2017) and Bloedel and Zhong (2020) show that many of the reduced form uniformly posterior-separable cost functions can be conceived of as resulting from a process of continuous experimentation and optimal stopping.

Morris and Strack (2019) consider the sampling problem of Wald (1945) and derive a cost function that exhibits constant marginal costs. They also discuss the issue of economies/diseconomies of scale in information acquisition, pointing out that entropy costs have economies of scale. We contribute by providing a cost function that exhibits such diseconomies of scale.

Mensch (2018) provides an axiomatic representation of information acquisition using Blackwell experiments as the primitive. Unlike Pomatto et al. (2023) who focus solely on cost functions that have constant marginal costs, Mensch (2018) provides an axiomatic framework that covers all posterior-separable cost functions. His main result characterizes cost functions that are additively separable over signals, and is the same as our Proposition 1. This result also comes from the statistics literature, e.g. Torgersen (1991).

This project is related to the rational inattention literature (Sims, 2003; Matějka and McKay, 2015), which also considers information frictions. However, in rational inattention models, the cost of new information is generally given by the expected reduction in entropy between the prior and the posterior beliefs of a decision maker. In such a cost function, the cost of an experiment is dependent on a prior belief. In contrast, we examine cost functions over Blackwell experiments which make no reference to a prior. Note, however, that if one is running a subsequent experiment contingent on some realization s of a first experiment, then the resulting implicit cost may depend on the prior, since the probability that s is realized may depend on the prior.

Finally, the Bayesian persuasion literature (Kamenica and Gentzkow, 2011) considers a sender who commits to an arbitrary experiment in order to influence a receiver. Gentzkow and Kamenica (2014) extends this model to the case where experiments are costly. Their model of costly information fits into our present framework.

2 Model

2.1 Experiments

There is a finite state space Θ of size N. An *experiment* is a pair (M, π), where $M \in \mathbb{N}$ denotes the size of a signal space, and $\pi : \Theta \to \Delta(\{1, \ldots, M\})$ gives the conditional probabilities of the possible signal realizations. Let Π be the class of all experiments. We will use Π_M to denote the class of all experiments with a signal space of size M. Hence for $i \in \{1, \ldots, N\}$ and $(M, \pi) \in \Pi$ we have that $\pi(i) \in \Delta(\{1, \ldots, M\})$. We will use subscripts to index signals, so that $\pi_s(i) = \Pr\{s|i\}$ and $\pi_s = (\pi_s(1), \ldots, \pi_s(N)) \in [0,1]^N$. When the size of the signal space is clear from context, we will sometimes simply refer to π as an experiment.

For example, suppose we have an experiment in which a biased coin is flipped. The coin is known to be 40-60 biased, but it is not known which face is favored. Call the state in which heads is favored H, and the state in which tails is favored T. An experiment in which the coin is flipped has two possible signal realizations, t and h. This experiment is represented as

$$\begin{aligned} \pi_h(H) &= 0.6, \quad \pi_t(H) = 0.4 \\ \pi_h(T) &= 0.4, \quad \pi_t(T) = 0.6. \end{aligned} \tag{2.1}$$

We now introduce some key definitions.

Definition 1. Given $\beta \in (0, 1)$ and experiments $\pi \in \Pi_M$ and $\rho \in \Pi_L$, the *mixed experiment* $\tau = \beta\pi \oplus (1-\beta)\rho \in \Pi_{M+L}$ is defined via

$$\begin{aligned} \tau_i &= \beta\pi_i, \text{ for } i = 1, \ldots, M, \\ \tau_i &= (1-\beta)\rho_{i-M}, \text{ for } i = M+1, \ldots, M+L. \end{aligned}$$

Hence when *mixing* two experiments, we conduct only one of the experiments, chosen randomly.

Definition 2. Given experiments $\pi \in \Pi_M$ and $\rho \in \Pi_L$, the *joint experiment* $\tau = \pi \otimes \rho \in \Pi_{ML}$ is defined via

$$\tau_{s+M(t-1)} = \pi_s \rho_t,$$

for all $s \in \{1, \ldots, M\}$ and $t \in \{1, \ldots, L\}$.

When *joining* two experiments, we conduct both experiments and see both signal realizations.

Definition 3. Let π and ρ be two experiments, and μ_0 a uniform prior over Θ. Let $\langle \pi | \mu_0 \rangle$ and $\langle \rho | \mu_0 \rangle$ be the distributions over posterior beliefs induced by π and ρ given μ_0. Then π *dominates* ρ *in the Blackwell order* if

$$\int_{\mu \in \Delta(\Theta)} g(\mu) d\langle \pi | \mu \rangle (\mu) \geq \int_{\mu \in \Delta(\Theta)} g(\mu) d\langle \rho | \mu_0 \rangle (\mu)$$

for every convex[1] function $g : \Delta(\Theta) \to \mathbb{R}$. Two experiments π and ρ are *Blackwell equivalent* if π dominates ρ in the Blackwell order and ρ dominates π in the Blackwell order.

The Blackwell ordering is commonly used to rank experiments based on their informational content.

2.2 Cost Functions

A *cost function* $C : \Pi \to \mathbb{R}_+$ assigns a (nonnegative) cost to each experiment in Π. The first two axioms that we will use to characterize cost functions are *Blackwell monotonicity* and *linearity*.

[1] A function c is convex if for all $\beta \in (0,1)$ and for all $x_1, x_2 \in (0,1)^N$

$$c(\beta x_1 + (1-\beta) x_2) \leq \beta c(x_1) + (1-\beta) c(x_2).$$

Definition 4. A cost function C is *Blackwell monotonic* if for all experiments $\pi, \rho \in \Pi$, if π dominates ρ in the Blackwell order then $C(\pi) \geq C(\rho)$.

Blackwell monotonicity captures the notion that more informative experiments are more costly. Additionally, it implies that information costs should not be sensitive to how signal realizations are labelled. Note that if a cost function is Blackwell monotonic then it will assign the same cost to Blackwell equivalent experiments.

Definition 5. A cost function C is *linear* if for all $\beta \in (0,1)$ and experiments $\pi \in \Pi_M$ and $\rho \in \Pi_L$,

$$C(\beta\pi \oplus (1-\beta)\rho) = \beta C(\pi) + (1-\beta)C(\rho).$$

Thus linearity requires that a cost function is linear in the mixing operation.

Definition 6. A cost function C is *continuous* if for any (M, π) and any sequence of experiments $(M, \pi^1), (M, \pi^2), (M, \pi^3), \ldots$ such that $\pi^n_s(i) \longrightarrow \pi_s(i)$ for all i and s, we have $C(\pi^n) \longrightarrow C(\pi)$.

Under a continuous cost function, if two experiments have similar conditional distributions of signals then they have similar costs. Note that in this definition we are only considering point-wise convergence of experiments whose signal spaces are the same size.

2.3 Separability

Cost functions that are linear and Blackwell monotonic have a very convenient additively separable form, and can be characterized by a generating function c. The following proposition is a known result (Mensch, 2018; Torgersen, 1991), but is included here as this representation is key to our main result.

Proposition 1. *A cost function C is Blackwell monotonic, linear, and continuous, if and only if there exists a convex, homogeneous of degree one,* [2] *and continuous*

[2] A function c is homogeneous of degree one if for all $x_1, \ldots, x_N \in (0,1)$ and $\beta \in (0, \min_i\{1/x_i\})$ we have $\beta c(x) = c(\beta x)$.

function $c : (0,1)^N \to \mathbb{R}$ *such that for all* $(M, \pi) \in \Pi$,

$$C(\pi) = \sum_{s=1}^{M} c(\pi_s). \tag{2.2}$$

A proof is given in Appendix A. The homogeneity of c relates directly to the linearity of the cost function, and the convexity of c is connected to the monotonicity of the cost function. To understand the intuition, consider comparing the two experiments of flipping a biased coin once, and flipping it twice. Each of the possible signal realizations in the latter experiment has a lower probability than any signal in the former experiment. Hence the convexity of c guarantees that it is more costly to flip the coin twice.

3 Main Results

3.1 Economies of Scale

Given some $\lambda \in \mathbb{R}$, we are interested in cost functions that satisfy

$$C(\pi \otimes \rho) = C(\pi) + C(\rho) + \lambda C(\pi)C(\rho) \tag{3.1}$$

for all experiments π and ρ, with $\lambda \neq 0$.

This third condition allows us to capture either economies or diseconomies of scale. When $\lambda > 0$ costs are super-additive, and when $\lambda < 0$ costs are sub-additive. This pseudo-additivity property already exists in the literature on information theory. For instance, Tsallis entropy, a generalization of Shannon entropy, satisfies this property (Tsallis, 1988).

Theorem 1. *A cost function* C *is monotonic, linear, continuous, nonconstant, and satisfies*

$$C(\pi \otimes \rho) = C(\pi) + C(\rho) + \lambda C(\pi)C(\rho) \tag{3.2}$$

for some $\lambda \neq 0$ *and all* $\pi, \rho \in \Pi$ *if and only if for all* $\pi \in \Pi$,

$$C(\pi) = \frac{1}{\lambda}\left(\sum_{s=1}^{M}\prod_{i=1}^{N}\pi_s(i)^{\alpha_i} - 1\right), \tag{3.3}$$

where $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$ *are parameters satisfying* $\sum_{i=1}^{N}\alpha_i = 1$ *such that*

$$c(x) = \frac{1}{\lambda}\left(\prod_{i=1}^{N}x(i)^{\alpha_i} - \frac{1}{N}\sum_{i=1}^{N}x(i)\right). \tag{3.4}$$

is convex.

The proof is given in Appendix B.1. The proof uses Proposition 1 to reduce the problem to finding a generating function c. By defining $f : (0,1)^N \to \mathbb{R}$ via $f(x) = \lambda c(x) + \frac{1}{N}\sum_{i=1}^{N}x(i)$, the condition in (3.2) reduces to

$$\sum_{s=1}^{M}\sum_{t=1}^{L}f(\pi_s\rho_t) = \sum_{s=1}^{M}\sum_{t=1}^{L}f(\pi_s)f(\rho_t).$$

Solving this functional equation gives us (3.3). Note that the -1 term in (3.3) is there to normalize the cost of an uninformative experiment to 0.

If a function is twice continuously differentiable, then its convexity can be verified by examining its (symmetric) Hessian matrix. At each $x \in (0,1)^N$, c is convex iff its Hessian matrix $\mathbf{H}$ is positive semi-definite, written as $\mathbf{H} \geq 0$. We calculate the entries of the Hessian to be

$$\mathbf{H}_{ij} = \frac{\alpha_i\alpha_j}{\lambda x(i)x(j)}\prod_{k=1}^{N}x(k)^{\alpha_k}, \quad i \neq j,$$
$$\mathbf{H}_{ii} = \frac{\alpha_i(\alpha_i - 1)}{\lambda x(i)^2}\prod_{k=1}^{N}x(k)^{\alpha_k}, \quad i = 1, \ldots, N.$$

The matrix $\mathbf{D} \coloneqq \operatorname{diag}(x(1), \ldots, x(N))$ is positive definite ($\mathbf{D} > 0$), so $\mathbf{H} \geq 0$ (and c

is convex) iff the matrix

$$\begin{aligned}\widehat{\mathbf{H}} &\coloneqq \mathbf{DHD} \\ &= \frac{1}{\lambda}\begin{bmatrix} \alpha_1(\alpha_1-1) & \alpha_1\alpha_2 & \cdots & \alpha_1\alpha_N \\ \alpha_2\alpha_1 & \alpha_2(\alpha_2-1) & \cdots & \alpha_2\alpha_N \\ \vdots & \vdots & & \vdots \\ \alpha_N\alpha_1 & \alpha_N\alpha_2 & \cdots & \alpha_N(\alpha_N-1) \end{bmatrix}\end{aligned}$$

is positive semi-definite.

For an example of how the requirement $\widehat{\mathbf{H}} \geq 0$ constrains α, consider the case $N = 2$. Then $\widehat{\mathbf{H}}$ is 2×2, so $\widehat{\mathbf{H}} \geq 0$ iff $\widehat{\mathbf{H}}_{1,1} \geq 0$ and $\det(\widehat{\mathbf{H}}) \geq 0$. We also require that $\alpha_1 + \alpha_2 = 1$, and so

$$\det(\widehat{\mathbf{H}}) = \frac{1}{\lambda^2}(\alpha_1(\alpha_1-1)\alpha_2(\alpha_2-1) - \alpha_1^2\alpha_2^2) = 0.$$

In other words, the non-negative determinant condition is always satisfied. It remains to consider the conditions under which $\widehat{\mathbf{H}}_{1,1} \geq 0$. If $\lambda > 0$, then $\widehat{\mathbf{H}}_{1,1} = \frac{1}{\lambda}\alpha_1(\alpha_1 - 1) \geq 0$ iff $\alpha_1 \leq 0$ or $\alpha_1 \geq 1$ (and hence $\alpha_2 \geq 1$ or $\alpha_2 \leq 0$ respectively). If $\lambda < 0$, then $\widehat{\mathbf{H}}_{1,1} \geq 0$ iff $0 \leq \alpha_1 \leq 1$ (and hence $0 \leq \alpha_2 < 1$).

Each α_i can be interpreted parameterizing the cost of learning about state i. The lower α_i, the more costly it is to learn about state i. The parameter λ is a scaling parameter that affects the overall cost of information. These interpretations will be explored in more detail in Section 4.

The cost function in (3.3) is closely related to the Hellinger transform (Torgersen, 1991). The Hellinger transform of an experiment π is a mapping $H_\pi : \Delta(\Theta) \to [0, 1]$ given by $H_\pi(\alpha) = \sum_{s=1}^{M} \prod_{i=1}^{N} \pi_s(i)^{\alpha_i}$. This mapping is monotonically decreasing in the statistical information of π. The primary difference between (3.3) and the Hellinger transform is that in (3.3) the parameters α_i are not restricted to be positive, and the parameter λ allows for scaling. This allows (3.3) to be monotonically increasing in the Blackwell order, and exhibit economies or diseconomies of scale depending on the parameter choices.

3.2 Relationship to Pomatto et al. (2023)

Pomatto et al. (2023) consider three main axioms for cost functions. First, C assigns the same cost to Blackwell equivalent experiments. Second, C satisfies $C(\pi \otimes \rho) = C(\pi) + C(\rho)$ for all $\pi, \rho \in \Pi$. Note that these first two axioms in turn imply that the cost function C is Blackwell monotonic. Third, they posit that the cost of an experiment is linear in the probability that the experiment will generate information. That is, if π^0 is an uninformative experiment, then for every π and $\beta \in [0, 1]$,

$$C(\beta\pi \oplus (1 - \beta)\pi^0) = \beta C(\pi).$$

This third axiom is weaker than our linearity assumption; their third axiom only requires linearity to hold when one of the experiments is uninformative. While this means that Theorem 2 below is immediate from their paper, we include it as a result because our proof technique is quite different.

Theorem 2. *A cost function C is monotonic, linear, continuous, nonconstant, and satisfies*

$$C(\pi \otimes \rho) = C(\pi) + C(\rho) \tag{3.5}$$

for all $\pi, \rho \in \Pi$ if and only if C has the form

$$C(\pi) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} \beta_{ij} \sum_{s=1}^{M} \pi_s(i) \ln\left(\frac{\pi_s(i)}{\pi_s(j)}\right), \tag{3.6}$$

with parameters $\beta_{ij} \in \mathbb{R}_+$ not all zero.

The proof is given in Appendix B.2. The approach is similar to the proof of

Theorem 1. Using Proposition 1, the problem reduces to finding a generating function c that satisfies

$$\sum_{s=1}^{M}\sum_{t=1}^{L} c(\pi_s\rho_t) = \sum_{s=1}^{M} c(\pi_s) + \sum_{t=1}^{L} c(\rho_t).$$

Solving this functional equation gives us

$$c(x) = \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \beta_{ij} x(i) \ln\left(\frac{x(i)}{x(j)}\right),$$

which yields (3.6).

A stark difference between the cost functions in (3.3) and (3.6) is the number of parameters. Each parameter β_{ij} can be interpreted as the cost of discriminating between states i and j, requiring $N(N-1)$ parameters. In contrast, (3.3) has only $N+1$ parameters.

4 Information Acquisition in Decision Problems

In this section we explore the implications of the cost function (3.3) for decision problems. The examples considered also appear in Pomatto et al. (2023). An agent chooses an action a from a finite set A. The payoff from a in state i is given by $u(a,i)$. The agent has a full-support prior $\mu_0 \in \Delta(\Theta)$. Before taking an action, the agent can choose an experiment π for a cost $C(\pi)$, where C is the cost function in (3.3). Since C is monotone with respect to the Blackwell order, it is without loss of generality to restrict attention to experiments where the set of signal realizations equals the set of actions A, and to assume that the decision maker will take the action recommended by the signal. We can thus identify an experiment π with a vector of probability measures over actions $\pi \in \Delta(A)^N$.

An optimal experiment solves

$$\pi^* \in \underset{\pi \in \Delta(A)^N}{\arg\max} \sum_{i\in\Theta} \left(\mu_0(i) \sum_{a\in A} \pi_a(i) u(a,i) \right) - C(\pi). \tag{4.1}$$

Thus under an optimal experiment the action a is chosen in state i with probability $\pi^*_a(i)$.

4.1 Identifying the Cost Function from Observed Choices

We begin by examining the problem of identifying and testing our model using data from observed choices. Consider the simple binary choice problem with two states $\Theta = \{1, 2\}$ and a uniform prior. The agent can choose between actions a_1 and a_2. If the action matches the state she receives a payoff of $v > 0$, and otherwise she receives a payoff of 0. The agent's optimization problem is then

$$\max_{\pi_{a_1}(1), \pi_{a_2}(2) \in (0,1)} \frac{v}{2}[\pi_{a_1}(1) + \pi_{a_2}(2)] - \frac{1}{\lambda}\left[\pi_{a_1}(1)^{\alpha_1}(1-\pi_{a_2}(2))^{\alpha_2} + (1-\pi_{a_1}(1))^{\alpha_1}\pi_{a_2}(2)^{\alpha_2} - 1\right].$$

The choice probabilities $(\pi_a(i))_{i\in\Theta, a\in A}$ are observed by an analyst who wishes to determine if these probabilities are consistent with the cost function (3.3). Recalling that $\alpha_2 = 1 - \alpha_1$, this is true if there exist coefficients (λ, α_1) that satisfy the first-order conditions

$$\frac{\lambda v}{2} = \alpha_1 \left(\frac{1-\pi_{a_2}(2)}{\pi_{a_1}(1)}\right)^{1-\alpha_1} - \alpha_1 \left(\frac{\pi_{a_2}(2)}{1-\pi_{a_1}(1)}\right)^{1-\alpha_1} \tag{4.2a}$$

$$\frac{\lambda v}{2} = (1-\alpha_1) \left(\frac{1-\pi_{a_1}(1)}{\pi_{a_2}(2)}\right)^{\alpha_1} - (1-\alpha_1) \left(\frac{\pi_{a_1}(1)}{1-\pi_{a_2}(2)}\right)^{\alpha_1}. \tag{4.2b}$$

For instance, suppose that $v = 3$ and the analyst observes the agent taking the correct action 30% of the time in state 1, and 60% of the time in state 2. Then we can numerically solve the first-order conditions to find that $\lambda \approx -4.79$, $\alpha_1 \approx 4.85$, and $\alpha_2 \approx -3.85$. Since $\alpha_1 > 1$ and $\lambda < 0$, these choice probabilities are inconsistent with the cost function (3.3) as it would mean the generating function (3.4) is not convex.

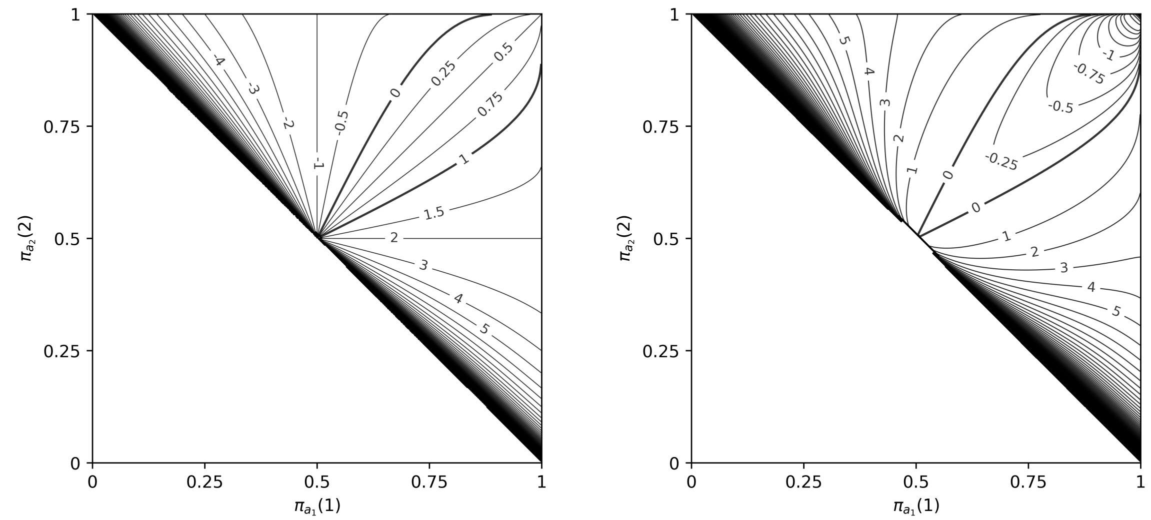
But now suppose that the agent takes the correct action 80% of the time in state 1, and 70% of the time in state 2. This corresponds to parameter values of $\lambda \approx -0.209$, $\alpha_1 \approx 0.837$, and $\alpha_2 \approx 0.163$. Hence this choice behavior can be explained by our model. Of course in general, when there are more than two states and two actions, the analyst would need data from multiple decision problems to point-identify the parameters.

Figure 1 depicts all choice probabilities that are consistent with the cost function (3.3), and the corresponding values of α_1 and λ, for $v = 3$. There are no α_1 and λ pairs that are consistent with $\pi_{a_1}(1) + \pi_{a_2}(2) \leq 1$ or $\lambda = 0$. We can think of λ as parameterizing the overall cost of learning. Since it is the inverse of λ that enters into (3.3), the amount of learning is increasing in the absolute value of λ. On the other hand, α_1 parameterizes the relative ease of learning about state 1. Higher values of α_1 will skew learning towards state 1, so that the decision maker will be more accurate in state 1.

For instance, when $\alpha_1 = \alpha_2 = 1/2$ we have that $\lambda < 0$ and $\pi_{a_1}(1) = \pi_{a_2}(2)$, i.e., the probability of guessing correctly is the same in both states. The probability of choosing correctly approaches 1/2 as λ approaches 0, and approaches 1 as $|\lambda|$ increases. If we instead hold $\lambda < 0$ constant, increasing α_1 from 1/2 leads to an increase in $\pi_{a_1}(1)$.

When $\lambda > 0$, negative values of α_1 correspond to $\pi_{a_1}(1) < \pi_{a_2}(2)$, and positive values to $\pi_{a_1}(1) > \pi_{a_2}(2)$. When $\alpha_1 = -1$, the value of $\pi_{a_1}(1) = 1/2$ is the same for any valid $\lambda > 0$, while the amount of learning about state 2, $\pi_{a_2}(2)$, is increasing in λ. In a sense, $\alpha_1 = -1$ implies that learning about state 1 is inelastic. When α_1 is held fixed between 0 and -1, then the amount of learning about state 1, $\pi_{a_1}(1)$ is decreasing in λ. If instead we fix some $\alpha_1 > -1$, then $\pi_{a_1}(1)$ is increasing in λ.

Note that when $\alpha_1 < -1$ we have $\pi_{a_1}(1) < 1/2$, so that the probability of guessing correctly in state 1 is actually *lower* than it would be in the absence of any learning. Hence if we are performing an experiment to test the hypothesis that the true state



(a) Contour plot of values of α_1.

(b) Contour plot of values of λ.

Figure 1: Values of α_1 and λ that are consistent with observed choice probabilities for $v = 3$. There are no α_1 and λ pairs that are consistent with the lower region where $\pi_{a_1}(1) + \pi_{a_2}(2) \leq 1$, or where $\lambda = 0$.

is 2, $\alpha_1 < -1$ would increase the probability of a Type II error, in the sense that the decision maker would be more likely to accept the hypothesis and choose action a_2 when the true state is 1.

The case of $\lambda > 0$ tends to lead to asymmetry in the choice probabilities. This is because in order for there to be increasing marginal costs, we must have $\alpha_i > 1$ for some i. But to ensure convexity of the generating function and hence Blackwell monotonicity, we must then have $\alpha_{-i} < 0$. This strongly skews the choice probabilities in favor of the cheaper state. Hence diseconomies of scale in information costs characterized by (3.1) differ fundamentally from traditional diseconomies of scale, as the increasing marginal costs cannot apply equally to all states.

Note the teardrop-shaped region: within this region $\lambda < 0$, and outside of it $\lambda > 0$. In fact, this teardrop-shaped region is identical to the range of choice probabilities that can be consistent with the LLR cost function for $\beta_{12}, \beta_{21} > 0$; compare Figure 2 of Pomatto et al. (2023). This can be shown simply by comparing the first-order conditions of the decision problem.

Proposition 2. *A choice behavior π can be explained by the LLR cost function with $\beta_{1,2}, \beta_{2,1} > 0$ if and only if it can be explained by the cost function* (3.3) *with $\lambda < 0$.*

Proof. Define $\ell_1 = \frac{1-\pi_{a_2}(2)}{\pi_{a_1}(1)}$ and $\ell_2 = \frac{\pi_{a_2}(2)}{1-\pi_{a_1}(1)}$. From the first order conditions for (4.1), the LLR cost function can explain a choice behavior with strictly positive β's iff

$$\ell_2 - \ell_1 + \ln(\ell_1/\ell_2) > 0 \qquad\qquad 1/\ell_1 - 1/\ell_2 + \ln(\ell_1/\ell_2) > 0. \tag{4.3}$$

The first order conditions (4.2) can be simplified to give us

$$\alpha_1(\ell_1^{1-\alpha_1} - \ell_2^{1-\alpha_1}) + (1-\alpha_1)(\ell_1^{-\alpha_1} - \ell_2^{-\alpha_1}) = 0. \tag{4.4}$$

The left-hand side of this equation as a function of α_1 resembles a negative cubic function, rising to the left and descending to the right. There are always roots at $\alpha_1 = 0$ and $\alpha_1 = 1$. The third and final root will be between 0 and 1 when the derivative of the LHS is negative at both $\alpha_1 = 0$ and $\alpha_1 = 1$. The derivative of the LHS evaluated at $\alpha_1 = 0$ is $-(\ell_2 - \ell_1 + \ln(\ell_1/\ell_2))$, and evaluated at $\alpha_1 = 1$ is $-(1/\ell_1 - 1/\ell_2 + \ln(\ell_1/\ell_2))$. These derivatives are just the negatives of the left-hand sides of (4.3). Now recall that $\lambda < 0$ is implied by $\alpha_1 \in (0, 1)$. Hence the conditions for the choice behavior to be explicable by (3.3) with $\lambda < 0$ are the same as the conditions for them to be explicable by the LLR cost function with strictly positive parameters. □

Hence the case of diseconomies of scale where $\lambda > 0$ strictly expands the set of explicable choice probabilities compared to the LLR cost function. In fact, we can obtain a special case of the LLR cost function by taking the limit of (3.3) as $\lambda \to 0$. We will use the fact that

$$\ln(z) = \lim_{a\to 0} \frac{z^a - 1}{a}.$$

Now set $\beta_{12} = 1$ and $\beta_{21} = 0$. Then

$$C_{LLR}(\pi) = \lim_{\lambda \to 0} \frac{1}{\lambda} \left(\pi_1(1)^{1+\lambda} \pi_1(2)^{-\lambda} + \pi_2(1)^{1+\lambda} \pi_2(2)^{-\lambda} - 1 \right),$$

i.e. the cost function in (3.3) with $\alpha_1 = 1 + \lambda$ and $\alpha_2 = \lambda$.

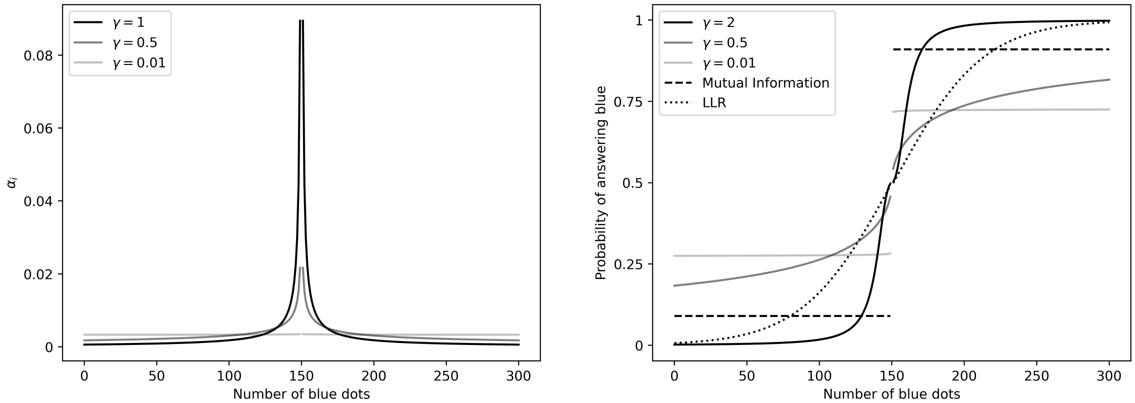
4.2 Perception Task

We will next the implications of our cost function for the classic perception task in which an agent is shown an *even* number of dots, each of which is either red or blue. The agent decides whether there are more blue or red dots, and receives a reward if correct. Intuitively, it should be harder to guess correctly when the difference in the number of dots of different colors is small. This is a canonical experiment in the literature on rational inattention (e.g. Caplin and Dean, 2013).

The total number of dots $n > 0$ is fixed and known to the agent. The number i of blue dots is drawn uniformly from the set $\Theta = \{0, \ldots, n/2 - 1, n/2 + 1, \ldots, n\}$. The action set is $A = \{B, R\}$ and the utility function is

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

For a vector of distributions over actions π, the decision maker guesses correctly in state i with probability $\pi_B(i)$ if $i > n/2$, and $\pi_R(i)$ if $i < n/2$. Under mutual information costs, the optimal experiment must induce a probability of guessing correctly that is state-independent, i.e., the probability of a correct choice must be the same for any two states that lead to the same utility function over actions (Dean and Neligh, 2017). In contrast, both the cost function in (3.3) and the LLR cost function can predict a sigmoidal relation between the state and the choice probability with an



,

(a) Values of α_i across states for different values of γ.

(b) Predicted probability of guessing that there are more blue dots as a function of the state with $\lambda = -1$.

Figure 2: Perception Task

appropriate choice of parameters. Such a sigmoidal relation is more consistent with the psychometric functions typical of such a task.

To understand the role of the parameters, consider a family of example values of $\{\alpha_i\}$ determined by a hyperparameter $\gamma > 0$, and set $\lambda = -1$. Define $\hat{\alpha}_i(\gamma) := |i - n/2|^{-\gamma}$ for $\gamma > 0$. We can then normalize to define $\alpha_i(\gamma) = \hat{\alpha}_i / \sum_j \hat{\alpha}_j$. Hence the value of α_i is decreasing in the difference between the number of blue and red dots. The parameter γ determines the magnitude of this dependence: as γ goes to zero, all states become equally costly to learn about. This is depicted in Figure 2a.

Figure 2b shows the predicted choice probabilities of optimal experiments under the cost function (3.3). When γ is high so that states close to $n/2$ are more costly to observe, the resulting function is sigmoidal. When γ is small so that all values of α_i approach the same value, the choice probabilities resemble those obtained using mutual information costs: since all states are equally costly to observe, the choice probabilities only depend on whether or not the state is above or below the critical threshold of $n/2$. Of course, this feature could be achieved using other cost functions that can account for the difficulty of distinguishing different states (Pomatto et al., 2023; Hébert and Woodford, 2021), but our cost function can do so with only n parameters.

5 Posterior Separability

Given a prior $\mu \in \Delta(\Theta)$, a distribution of posteriors is *Bayes plausible* if the expected posterior probability distribution equals the prior. We will write $\mathcal{I}(\mu) \in \Delta(\Delta(\Theta))$ for the set of Bayes plausible distributions given a prior μ. This provides an alternative to directly modeling Blackwell experiments when given a prior, as any experiment will induce a Bayes plausible distribution of posteriors. Two experiments are Blackwell equivalent if for any prior they induce the same distribution over posteriors. We will write $\langle \pi | \mu \rangle \in \mathcal{I}(\mu)$ for the distribution of posteriors induced by an experiment π given the prior μ. An *attention cost function* $K : \Delta(\Theta) \times \Delta(\Delta(\Theta)) \to \mathbb{R}$ takes as arguments a prior and a Bayes plausible distribution of posteriors.

Our definition of posterior separability follows Caplin et al. (2022):

Definition 7. Given a prior $\mu \in \Delta(\Theta)$, an attention cost function K is *posterior-separable* if there exists a strictly convex function $k_\mu : \Delta(\Theta) \to \overline{\mathbb{R}}$, real-valued if the support is the same as the prior, such that, given $p \in \mathcal{I}(\mu)$,

$$K(\mu, p) = \sum_{\nu \in \text{supp}(p)} p(\nu) k_\mu(\nu) - k_\mu(\mu). \tag{5.1}$$

Note that the function k_μ is allowed to vary with the prior. A cost function K is *uniformly posterior-separable* if the function k_μ does not depend on the prior, i.e., there is some function k such that $k_\mu = k$ for all priors μ.

Definition 8. A cost function C has a *posterior-separable representation* if there exists an attention cost function K such that for all full-support priors $\mu \in \Delta(\Theta)$ and $\pi \in \Pi$,

$$C(\pi) = K(\mu, \langle \pi | \mu \rangle).$$

We will now show that cost functions satisfying the assumptions in Proposition 1 can indeed be represented by a posterior-separable attention cost function, but this representation is not uniformly posterior-separable.

Proposition 3. *If a cost function C is Blackwell monotonic, linear, and continuous then it has a posterior-separable representation.*

For instance, consider the LLR cost function in (3.6). This cost function can be reformulated so that it has the posterior-separable form in (5.1) by defining $k_\mu : \Delta(\Theta) \to \mathbb{R}$ via

$$k_\mu(\nu) = \sum_{i \neq j} \beta_{ij} \frac{\nu(i)}{\mu(i)} \ln \frac{\nu(i)}{\nu(j)},$$

where $\mu \in \Delta(\Theta)$. Even though this k_μ depends on the prior, the cost $K(\mu, \langle \pi | \mu \rangle)$ of an experiment π is independent of the prior and depends only on the experiment.

This holds in general. While the experiment cost functions we have considered may have posterior-separable representations, they are not *uniformly* posterior-separable. On the other hand, the mutual information cost function used widely in the rational inattention literature is uniformly posterior-separable and cannot be represented with an experiment cost function in the form (2.2).

6 Conclusion

In this paper we present a theory of information acquisition that explores the notion of economies and diseconomies of scale in the cost of information production. This builds on the work on constant returns to scale in Pomatto et al. (2023). Our cost function is tractable with a straightforward interpretation and has distinct economic implications.

We propose several avenues for future research. The first is an extension of our framework beyond the setting of a finite set of states and experiments with a finite number of signals. Second, there are numerous settings which have modeled information costs using entropy reduction (Van Nieuwerburgh and Veldkamp, 2010), and it would be interesting to see how sensitive the results are to assumptions on economies of scale. Finally, it remains to explore other formulations of information costs that exhibit diseconomies of scale.

A Proof of Proposition 1

We begin by introducing two lemmas that will be used in providing conditions for the separability of a cost function. We then proceed to give a proof of Proposition 1.

As a preliminary, define $D \subset \mathbb{R}^N \times \mathbb{R}^N$ as $D = \{(x_1, x_2) | x_1, x_2, x_1 + x_2 \in (0,1)^N\}$. We will frequently use D as the domain of a function. Additionally, given $M \geq 2$, define $\Pi_M \subset \Pi$ to be the set of all experiments with a signal space of size M.

Lemma 1. *If a cost function C is linear and assigns the same cost to Blackwell equivalent experiments, then there exists a function $H : D \to \mathbb{R}$ such that for all M and $\pi \in \Pi_M$ with $\pi_1 + \pi_2 \in (0,1)^N$,*

$$C(\pi_1, \dots, \pi_M) - C(\pi_1 + \pi_2, \pi_3, \dots, \pi_M) = H(\pi_1, \pi_2).$$

Proof. Take any M and L, and choose any $\pi \in \Pi_{M+1}$ and $\pi' \in \Pi_{L+1}$ such that $\pi_1 = \pi'_1$ and $\pi_2 = \pi'_2$. Define $\rho \in \Pi_M$ via

$$\rho_1 = \pi_1 + \pi_2, \qquad \text{and} \qquad \rho_s = \pi_{s+1} \ \text{ for } \ s = 2, \dots, M.$$

Similarly, define $\rho' \in \Pi_L$ via

$$\rho'_1 = \pi'_1 + \pi'_2, \qquad \text{and} \qquad \rho'_s = \pi'_{s+1} \ \text{ for } \ s = 2, \dots, L.$$

Now define the mixed experiments

$$\tau = \frac{1}{2}\rho \oplus \frac{1}{2}\pi' \qquad \text{and} \qquad \tau' = \frac{1}{2}\rho' \oplus \frac{1}{2}\pi.$$

Given the definition of ρ and ρ', we have that

$$\begin{array}{lllll} \tau_1 = \frac{1}{2}(\pi_1 + \pi_2), & \tau_2 = \frac{1}{2}\pi_3, & \tau_3 = \frac{1}{2}\pi_4, & \ldots, & \tau_M = \frac{1}{2}\pi_{M+1}, \\ \tau_{M+1} = \frac{1}{2}\pi'_1, & \tau_{M+2} = \frac{1}{2}\pi'_2, & \tau_{M+3} = \frac{1}{2}\pi'_3, & \ldots, & \tau_{M+L+1} = \frac{1}{2}\pi'_{L+1}, \end{array}$$

and

$$\begin{array}{lllll} \tau'_1 = \frac{1}{2}(\pi'_1 + \pi'_2), & \tau'_2 = \frac{1}{2}\pi'_3 & \tau'_3 = \frac{1}{2}\pi'_4, & \ldots, & \tau'_L = \frac{1}{2}\pi'_{L+1}, \\ \tau'_{L+1} = \frac{1}{2}\pi_1, & \tau'_{L+2} = \frac{1}{2}\pi_2, & \tau'_{L+3} = \frac{1}{2}\pi_3, & \ldots, & \tau'_{M+L+1} = \frac{1}{2}\pi_{M+1}. \end{array}$$

Since $\pi_1 = \pi'_1$ and $\pi_2 = \pi'_2$, we clearly have that τ and τ' are Blackwell equivalent.

Thus since C assigns the same cost to Blackwell equivalent experiments, we have that

$$C\left(\frac{1}{2}\rho \oplus \frac{1}{2}\pi'\right) = C\left(\frac{1}{2}\rho' \oplus \frac{1}{2}\pi\right).$$

Now since C is linear, it follows that

$$\frac{1}{2}C(\rho) + \frac{1}{2}C(\pi') = \frac{1}{2}C(\rho') + \frac{1}{2}C(\pi),$$

and hence

$$C(\pi) - C(\rho) = C(\pi') - C(\rho').$$

Therefore, for any $M, L \geq 2$ and any $\pi \in \Pi_M$ and $\pi' \in \Pi_L$, if $\pi_1 = \pi'_1$ and $\pi_2 = \pi'_2$ then

$$C(\pi_1, \ldots, \pi_M) - C(\pi_1 + \pi_2, \ldots, \pi_M) = C(\pi'_1, \ldots, \pi'_L) - C(\pi'_1 + \pi'_2, \ldots, \pi'_L).$$

Hence there exists a function $H : D \to \mathbb{R}$ such that for all $\pi \in \Pi$,

$$C(\pi_1, \ldots, \pi_M) - C(\pi_1 + \pi_2, \pi_3, \ldots, \pi_M) = H(\pi_1, \pi_2).$$

□

The following lemma is Theorem 2.2.4 in Ebanks et al. (1998).

Lemma 2. *A function $H : D \to \mathbb{R}$ satisfies*

$$H(x_1, x_2) = H(x_2, x_1),$$

for all $x_1, x_2 \in (0,1)^N$ with $x_1 + x_2 \in (0,1)^N$, and

$$H(x_1, x_2) + H(x_1 + x_2, x_3) = H(x_1, x_2 + x_3) + H(x_2, x_3),$$

for all $x_1, x_2, x_3 \in (0,1)^N$ with $x_1 + x_2 + x_3 \in (0,1)^N$, if and only if there exists a function $h : (0,1)^N \to \mathbb{R}$ such that

$$H(x_1, x_2) = h(x_1) + h(x_2) - h(x_1 + x_2) \tag{A.1}$$

for all $x_1, x_2 \in (0,1)^N$ with $x_1 + x_2 \in (0,1)^N$.

Proof of Proposition 1. ($\Longrightarrow$) First, suppose a cost function C is Blackwell monotonic, linear and continuous. We will show that there is a convex, homogeneous, and continuous function c such that (2.2) holds for all π.

Separability We first show that there exists a function $c : (0,1) \to \mathbb{R}$ such that (2.2) holds for all π.

By Lemma 1, there exists a function $H : D \to \mathbb{R}$ such that for all M and $\pi \in \Pi_M$,

$$C(\pi_1, \pi_2, \pi_3, \ldots, \pi_M) - C(\pi_1 + \pi_2, \pi_3, \ldots, \pi_M) = H(\pi_1, \pi_2). \tag{A.2}$$

Since Blackwell equivalent experiments have the same cost, the cost function C is invariant to permutations of the signals. Hence

$$\begin{aligned} H(\pi_1, \pi_2) &= C(\pi_1, \pi_2, \pi_3, \ldots, \pi_M) - C(\pi_1 + \pi_2, \pi_3, \ldots, \pi_M) \\ &= C(\pi_2, \pi_1, \pi_3, \ldots, \pi_M) - C(\pi_2 + \pi_1, \pi_3, \ldots, \pi_M) \\ &= H(\pi_2, \pi_1), \end{aligned} \tag{A.3}$$

so that H is symmetric. We also have that

$$\begin{aligned} C(\pi_1, \ldots, \pi_M) &= C(\pi_1 + \pi_2, \ldots, \pi_M) + H(\pi_1, \pi_2) \\ &= C(\pi_1 + \pi_2 + \pi_3, \ldots, \pi_M) + H(\pi_1 + \pi_2, \pi_3) + H(\pi_1, \pi_2) \\ &\vdots \\ &= C(\pi_1 + \cdots + \pi_t, \ldots, \pi_M) + \sum_{s=2}^{t} H(\pi_1 + \cdots + \pi_{s-1}, \pi_s) \\ &\vdots \\ &= C(\mathbf{1} - \pi_M, \pi_M) + \sum_{s=2}^{M-1} H(\pi_1 + \cdots + \pi_{s-1}, \pi_s). \end{aligned} \tag{A.4}$$

Thus, using (A.4) and Blackwell monotonicity, we have that for all $M \geq 3$ and $\pi \in \Pi_M$

$$\begin{aligned} &H(\pi_1 + \pi_2, \pi_3) + H(\pi_1, \pi_2) \\ &\qquad = \big[C(\pi_1 + \pi_2, \ldots, \pi_M) - C(\pi_1 + \pi_2 + \pi_3, \ldots, \pi_M)\big] \\ &\qquad\quad + \big[C(\pi_1, \ldots, \pi_M) - C(\pi_1 + \pi_2, \ldots, \pi_M)\big] \\ &\qquad = C(\pi_1, \ldots, \pi_M) - C(\pi_1 + \pi_2 + \pi_3, \ldots, \pi_M) \\ &\qquad = C(\pi_1, \pi_3, \pi_2 \ldots, \pi_M) - C(\pi_1 + \pi_2 + \pi_3, \ldots, \pi_M) \\ &\qquad = \big[C(\pi_1 + \pi_3, \pi_2, \ldots, \pi_M) - C(\pi_1 + \pi_2 + \pi_3, \ldots, \pi_M)\big] \\ &\qquad\quad + \big[C(\pi_1, \pi_3, \pi_2, \ldots, \pi_M) - C(\pi_1 + \pi_3, \pi_2, \ldots, \pi_M)\big] \\ &\qquad = H(\pi_1 + \pi_3, \pi_2) + H(\pi_1, \pi_3). \end{aligned} \tag{A.5}$$

By equations (A.3) and (A.5) and Lemma 2, we have that there exists a function $h : (0,1)^N \to \mathbb{R}$ such that (A.1) holds. Hence, using (A.4), we have that for all $M \geq 3$ and $\pi \in \Pi_M$

$$\begin{aligned}
&C(\pi_1, \ldots, \pi_M) \\
&\qquad = C(\mathbf{1} - \pi_M, \pi_M) + \sum_{s=2}^{M-1} \left[h(\pi_1 + \cdots + \pi_{s-1}) + h(\pi_s) - h(\pi_1 + \cdots + \pi_s) \right] \\
&\qquad = C(\mathbf{1} - \pi_M, \pi_M) + \sum_{s=2}^{M-1} h(\pi_s) - h(\pi_1 + \cdots + \pi_{M-1}) \\
&\qquad = C(\mathbf{1} - \pi_M, \pi_M) + \sum_{s=2}^{M-1} h(\pi_s) - h(\mathbf{1} - \pi_{M-1}).
\end{aligned}$$

Now define $g : (0,1)^N \to \mathbb{R}$ via

$$g(x) = C(\mathbf{1} - x, x) - h(\mathbf{1} - x),$$

so that for all $M \geq 3$ and $\pi \in \Pi_M$,

$$C(\pi_1, \ldots, \pi_M) = \sum_{s=1}^{M-1} h(\pi_s) + g(\pi_M). \tag{A.6}$$

Now again using the assumption of Blackwell monotonicity we have that for all $M \geq 3$ and $\pi \in \Pi_M$

$$\begin{aligned}
\sum_{s=1}^{M-1} h(\pi_s) + g(\pi_M) &= C(\pi_1, \ldots, \pi_{M-1}, \pi_M) \\
&= C(\pi_1, \ldots, \pi_M, \pi_{M-1}) \\
&= \sum_{s=1}^{M-2} h(\pi_s) + h(\pi_M) + g(\pi_{M-1}),
\end{aligned}$$

which reduces to

$$g(\pi_M) + h(\pi_{M-1}) = g(\pi_{M-1}) + h(\pi_M),$$

so that for all $x_1, x_2 \in (0,1)^N$ with $x_1 + x_2 \in (0,1)^N$

$$g(x_1) - h(x_1) = g(x_2) - h(x_2). \tag{A.7}$$

Now given $x_1, x_2 \in (0,1)^N$ with $x_1 + x_2 \in (0,1)^N$, choose $x_3 \in (0,1)^N$ such that $x_1 + x_3 \in (0,1)^N$ and $x_2 + x_3 \in (0,1)^N$. Two applications of (A.7) gives us that

$$g(x_1) - h(x_1) = g(x_3) - h(x_3) = g(x_2) - h(x_2).$$

Thus, $g(x) - h(x)$ is constant for all $x \in (0,1)^N$, i.e., $g(x) = h(x) + a$ for some $a \in \mathbb{R}$. Putting this into (A.6) we have that for all $M \geq 3$ and $\pi \in \Pi_M$,

$$\begin{aligned} C(\pi_1, \ldots, \pi_M) &= \sum_{s=1}^{M} h(\pi_s) + a \\ &= \sum_{s=1}^{M} \big[h(\pi_s) + a\pi_s(1)\big]. \end{aligned}$$

Now define $c : (0,1)^N \to \mathbb{R}$ via

$$c(x) = h(x) + ax(1), \tag{A.8}$$

where $x(1)$ is the first component of $x \in (0,1)^N$. Thus, we have (2.2) for all $M \geq 3$ and $\pi \in \Pi_M$.

For the case of $M = 2$, note that from (A.1) and (A.8) we have that for all $x_1, x_2 \in (0,1)^N$ with $x_1 + x_2 \in (0,1)^N$,

$$H(x_1, x_2) = c(x_1) + c(x_2) - c(x_1 + x_2). \tag{A.9}$$

Now choose some $\pi \in \Pi_2$ and $x_1, x_2 \in (0,1)^N$ such that $x_1 + x_2 = \pi_1$. It follows from

(A.9), (2.2) for $M = 3$, and (A.2) for $M = 2$ that

$$\begin{aligned} C(\pi_1, \pi_2) &= C(x_1 + x_2, \pi_2) \\ &= C(x_1, x_2, \pi_2) - H(x_1, x_2) \\ &= c(x_1) + c(x_2) + c(\pi_2) - \big[c(x_1) + c(x_2) - c(x_1 + x_2)\big] \\ &= c(\pi_1) + c(\pi_2). \end{aligned}$$

Thus we have (2.2) for all $\pi \in \Pi$.

Homogeneity (rational β) We will show that for all $x \in (0,1)^N$ and for all $\beta \in \mathbb{Q}$ with $0 < \beta < \max\{1/x(1), \ldots, 1/x(N)\}$,

$$\beta c(x(1), \ldots, x(N)) = c(\beta x(1), \ldots, \beta x(N)), \tag{A.10}$$

where $x(i)$ denotes the i^{th} element of x. We will cover the case of irrational β after proving convexity of c.

Choose $p_1 \ldots, p_N, m, r \in \mathbb{Q}$ such that $m \geq 2$, $1 \leq r \leq m$, and $0 < p_i < \frac{m}{r}$, $i = 1, \ldots, N$. Letting $M = 2$ and $L = r + 1$, define the experiments $\pi \in \Pi_M$ and $\rho \in \Pi_L$ via

$$\begin{aligned} \pi(i) &= \left(p_i \frac{r}{m}, 1 - p_i \frac{r}{m}\right), \quad \text{and} \\ \rho(i) &= \left(p_i \frac{1}{m}, \ldots, p_i \frac{1}{m}, 1 - p_i \frac{r}{m}\right). \end{aligned}$$

Note that π and ρ are Blackwell equivalent—the first $r - 1$ signals in ρ are redundant.

Hence by the separability of C we have that

$$
\begin{aligned}
0 &= C(\pi) - C(\rho) \\
&= \left[c\left(p_1\frac{r}{m}, \ldots, p_N\frac{r}{m}\right) + c\left(1 - p_1\frac{r}{m}, \ldots, 1 - p_N\frac{r}{m}\right)\right] \\
&\quad - \left[\sum_{n=1}^{r} c\left(p_1\frac{1}{m}, \ldots, p_N\frac{1}{m}\right) + c\left(1 - p_1\frac{r}{m}, \ldots, 1 - p_N\frac{r}{m}\right)\right] \\
&= c\left(p_1\frac{r}{m}, \ldots, p_N\frac{r}{m}\right) - rc\left(p_1\frac{1}{m}, \ldots, p_N\frac{1}{m}\right),
\end{aligned}
$$

and thus

$$rc\left(p_1\frac{1}{m}, \ldots, p_N\frac{1}{m}\right) = c\left(p_1\frac{r}{m}, \ldots, p_N\frac{r}{m}\right).$$

In the case $r = m$ we have

$$c\left(p_1\frac{1}{m}, \ldots, p_N\frac{1}{m}\right) = \frac{1}{m}c\left(p_1, \ldots, p_N\right),$$

and so in general

$$c\left(p_1\frac{r}{m}, \ldots, p_N\frac{r}{m}\right) = \frac{r}{m}c\left(p_1, \ldots, p_N\right).$$

Choosing $\beta = r/m$ gives us (A.10) for $\beta \in \mathbb{Q}$.

Note that we have only proved (A.10) for rational β—we will show that it also holds for all real β below, after proving the convexity and continuity of c.

Convexity Take any $\beta \in (0, 1)$ and $x_1, x_2 \in (0, 1)^N$. We will prove that

$$c(\beta x_1 + (1 - \beta)x_2) \leq \beta c(x_1) + (1 - \beta)c(x_2). \tag{A.11}$$

Choose any $\pi \in \Pi_M$ such that $\pi_1 = \beta x_1 + (1 - \beta)x_2$. As a first step, we will show that

$$c(\beta x_1) + c((1-\beta)x_2) = \beta c(x_1) + (1-\beta)c(x_2). \tag{A.12}$$

If $\beta \in \mathbb{Q}$, then (A.12) is a direct result of (A.10). So consider the case where β is irrational. Then we can construct a sequence $(\beta_i)_{i=1}^{\infty}$, with $\beta_i \in \mathbb{Q}$ and $\beta_i \in (0,1)$ for all i, such that $\beta_i \longrightarrow \beta$. By (A.10), we have that for all i and $x \in (0,1)^N$, $c(\beta_i x) = \beta_i c(x)$. It follows that for all $x \in (0,1)^N$, $c(\beta_i x) \longrightarrow \beta c(x)$. Thus

$$\begin{aligned} C(\beta_i x_1, (1-\beta_i)x_2, \pi_2, \ldots, \pi_M) &= c(\beta_i x_1) + c((1-\beta_i)x_2) + \sum_{s=2}^{M} c(\pi_s) \\ &\longrightarrow \beta c(x_1) + (1-\beta)c(x_2) + \sum_{s=2}^{M} c(\pi_s). \end{aligned}$$

And since C is continuous, we have also have that

$$\begin{aligned} C(\beta_i x_1, (1-\beta_i)x_2, \pi_2, \ldots, \pi_M) &\longrightarrow C(\beta x_1, (1-\beta)x_2, \pi_2, \ldots, \pi_M) \\ &= c(\beta x_1) + c((1-\beta)x_2) + \sum_{s=2}^{M} c(\pi_s). \end{aligned}$$

This gives us (A.12).

Next, note that by Blackwell monotonicity of C we have

$$C(\beta x_1 + (1-\beta)x_2, \pi_2, \ldots, \pi_M) \leq C(\beta x_1, (1-\beta)x_2, \pi_2, \ldots, \pi_M).$$

It follows that

$$\begin{aligned}
c(\beta x_1 + (1-\beta)x_2) + \sum_{s=2}^{M} c(\pi_s) &= C(\beta x_1 + (1-\beta)x_2, \ldots, \pi_M) \\
&\leq C(\beta x_1, (1-\beta)x_2, \pi_2, \ldots, \pi_M) \\
&= c(\beta x_1) + c((1-\beta)x_2) + \sum_{s=2}^{M} c(\pi_s) \\
&= \beta c(x_1) + (1-\beta)c(x_2) + \sum_{s=2}^{M} c(\pi_s),
\end{aligned}$$

which implies (A.11).

Continuity Since $c : (0,1)^N \to \mathbb{R}$ is convex with an open domain, it is continuous.

Homogeneity (irrational β) Finally, it remains to show that for all $x \in (0,1)^N$, (A.10) holds for all irrational $\beta \in (0, \max\{1/x(1), \ldots, 1/x(N)\})$. Take such a β. Similar to what we have done before, construct a sequence $(\beta_i)_{i=1}^{\infty}$, with $\beta_i \in \mathbb{Q}$ and $\beta_i \in (0, \max\{1/x(1), \ldots, 1/x(N)\})$ for all i, such that $\beta_i \longrightarrow \beta$. By (A.10), we have that for all i and $x \in (0,1)^N$, $c(\beta_i x) = \beta_i c(x)$. It follows that $c(\beta_i x) \longrightarrow \beta c(x)$. And since c is continuous, we know that $c(\beta_i x) \longrightarrow c(\beta x)$. We then have that $c(\beta x) = \beta c(x)$.

Therefore, if a cost function C is Blackwell monotonic, linear and continuous, then there is a convex, homogeneous, and continuous function c such that (2.2) holds for all π.

($\Longleftarrow$) For the converse, take a cost function C and suppose that there exists a continuous function $c : (0,1)^N \to \mathbb{R}$ that is convex and homogeneous of degree one such that (2.2) holds for all $\pi \in \Pi$. We will show that C is Blackwell monotonic, linear, and continuous.

Linearity Take any $\beta \in (0,1)$ and experiments $\pi \in \Pi_M$ and $\rho \in \Pi_L$. Then by (2.2) and the homogeneity of c we have

$$\begin{aligned} C(\beta\pi \oplus (1-\beta)\rho) &= \sum_{s=1}^{M} c(\beta\pi_s) + \sum_{t=1}^{L} c((1-\beta)\rho_t) \\ &= \sum_{s=1}^{M} \beta c(\pi_s) + \sum_{t=1}^{L} (1-\beta)c(\rho_t) \\ &= \beta C(\pi) + (1-\beta)C(\rho). \end{aligned}$$

Thus C is linear.

Blackwell Monotonicity Given $\pi \in \Pi$, take $\beta \in (0,1)$ and $x_1, x_2 \in (0,1)^N$ such that $\beta x_1 + (1-\beta)x_2 = \pi_1$. Then by the convexity and homogeneity of c we have that

$$\begin{aligned} C(\beta x_1, (1-\beta)x_2, \pi_2, \ldots, \pi_M) &= c(\beta x_1) + c((1-\beta)x_2) + \sum_{s=2}^{M} c(\pi_s) \\ &= \beta c(x_1) + (1-\beta)c(x_2) + \sum_{s=2}^{M} c(\pi_s) \\ &\geq c(\beta x_1 + (1-\beta)x_2) + \sum_{s=2}^{M} c(\pi_s) \\ &= C(\pi_1, \ldots, \pi_M). \end{aligned}$$

It follows that for all $\pi \in \Pi$ such that $\pi_1 + \pi_2 \in (0,1)^N$,

$$C(\pi_1, \ldots, \pi_M) \geq C(\pi_1 + \pi_2, \pi_3, \ldots, \pi_M).$$

Similarly, we have that for any $s \in \{1, \ldots, M-1\}$,

$$C(\pi_1, \ldots, \pi_M) \geq C(\pi_1 \ldots, \pi_s + \pi_{s+1}, \ldots, \pi_M). \tag{A.13}$$

Now take $\pi \in \Pi_M$ and $\rho \in \Pi_L$ such that π strictly dominates ρ in the Blackwell ordering. Then ρ is a garbling of π, i.e., there exists some $g : S_M \times S_L \to [0,1]$ such that

$$\sum_{t=1}^{L} g(s,t) = 1 \ \forall s \in S_M, \quad \text{and}$$
$$\rho_t(i) = \sum_{s=1}^{M} g(s,t)\pi_s(i) \ \forall i \in \Theta, t \in S_L.$$

Construct the experiment $\sigma \in \Pi_{M\times L}$, with $\sigma_{s,t}(i) = g(s,t)\pi_s(i)$.[3] Note that π and σ are Blackwell equivalent—we have just added some redundant signals. By the homogeneity of c we have that

$$\begin{aligned} C(\pi) &= \sum_{s=1}^{M} c(\pi_s) \\ &= \sum_{s=1}^{M}\sum_{t=1}^{L} g(s,t)c(\pi_s) \\ &= \sum_{s=1}^{M}\sum_{t=1}^{L} c(g(s,t)\pi_s) \\ &= C(\sigma). \end{aligned}$$

[3]Indexing σ by writing $\sigma_{s,t}$ is a slight abuse of notation, but it simplifies the exposition.

By repeated application of (A.13) it follows that

$$
\begin{aligned}
C(\pi) &= C(\sigma) \\
&= C(\sigma_{1,1}, \sigma_{2,1}, \sigma_{3,1}, \ldots, \sigma_{M,L-1}, \sigma_{M,L}) \\
&\geq C(\sigma_{1,1} + \sigma_{2,1}, \sigma_{3,1}, \ldots, \sigma_{M,L-1}, \sigma_{M,L}) \\
&\vdots \\
&\geq C(\sum_{s=1}^{M} \sigma_{s,1}, \sigma_{1,2}, \sigma_{2,2}, \ldots, \sigma_{M,L-1}, \sigma_{M,L}) \\
&\vdots \\
&\geq C\left(\sum_{s=1}^{M} \sigma_{s,1}, \sum_{s=1}^{M} \sigma_{s,2}, \ldots, \sum_{s=1}^{M} \sigma_{s,L}\right) \\
&= C(\rho).
\end{aligned}
$$

Hence C is Blackwell monotonic.

Continuity Take any experiment (M, π) and any sequence of experiments $(M, \pi^1), (M, \pi^2), (M, \pi^3), \ldots$ such that for all i we have $\pi^n(i) \to \pi(i)$. This implies that $\pi^n_s \to \pi_s$ for $s = 1, \ldots, M$. Since c is continuous, we have that $c(\pi^n_s) \to c(\pi_s)$. Using (2.2), we find that $C(\pi^n) \to C(\pi)$. Hence C is continuous.

Therefore, C is Blackwell monotonic, linear, and continuous.

□

B Proofs for Section 3

B.1 Proof of Theorem 1

Before proceeding, we introduce the following definitions.

Definition 9. A function $A : \mathbb{R}^N \to \mathbb{R}$ is *additive* if for all $x, y \in \mathbb{R}^N$,

$$A(x + y) = A(x) + A(y).$$

Definition 10. A function $D : \mathbb{R}^N \to \mathbb{R}$ is *multiplicative* if for all $x, y \in \mathbb{R}^N$,

$$D(xy) = D(x)D(y),$$

where the multiplication xy is performed component-wise.

The following lemma is from Theorem 8.3.5 in Ebanks et al. (1998).

Lemma 3. *For fixed $M \geq 3$, $L \geq 2$ or $M \geq 2$, $L \geq 3$ let $f : (0,1)^N \to \mathbb{R}$ satisfy*

$$\sum_{s=1}^{M} \sum_{t=1}^{L} f(\pi_s \rho_t) = \sum_{s=1}^{M} \sum_{t=1}^{L} f(\pi_s) f(\rho_t).$$

for all $\pi \in \Pi_M$ and $\rho \in \Pi_L$. Then one of the following holds:

1. *There exists a constant $a \in \mathbb{R}$ and an additive function $A : \mathbb{R}^N \to \mathbb{R}$ such that*

$$\begin{aligned} f(x) &= A(x) + a, \quad \text{and} \\ A(\mathbf{1}) + MLa &= (A(\mathbf{1}) + Ma)(A(\mathbf{1}) + La), \end{aligned} \tag{B.1}$$

   for all $x \in (0,1)^N$.

2. *There exists a multiplicative function $D : \mathbb{R}_+^N \to \mathbb{R}$ and an additive function $B : \mathbb{R}^N \to \mathbb{R}$ such that*

$$\begin{aligned} f(x) &= D(x) + B(x), \quad \text{and} \\ B(\mathbf{1}) &= 0. \end{aligned} \tag{B.2}$$

Proof of Theorem 1. Suppose that a cost function C is monotonic, linear, continuous, nonconstant, and satisfies (3.2) for all $\pi, \rho \in \Pi$. Then by Proposition 1 there exists a function $c : (0,1)^N$ such that $C(\pi) = \sum_{s=1}^{M} c(\pi_s)$ for all $\pi \in \Pi$. It follows that for all $\pi \in \Pi_M$ and $\rho \in \Pi_L$,

$$C(\pi \otimes \rho) = \sum_{s=1}^{M} \sum_{t=1}^{L} c(\pi_s \rho_t).$$

We can then rewrite (3.2) as

$$\sum_{s=1}^{M} \sum_{t=1}^{L} c(\pi_s \rho_t) = \sum_{s=1}^{M} c(\pi_s) + \sum_{t=1}^{L} c(\rho_t) + \lambda \sum_{s=1}^{M} c(\pi_s) \sum_{t=1}^{L} c(\rho_t). \tag{B.3}$$

Now define $f : (0,1)^N \to \mathbb{R}$ via

$$f(x) = \lambda c(x) + \frac{1}{N} \sum_{i=1}^{N} x(i) \tag{B.4}$$

for $x \in (0,1)^N$, where $x(i)$ denotes the i^{th} component of x. Using the definition of f and (B.3) we have that for all $\pi \in \Pi_M$ and $\rho \in \Pi_L$

$$\begin{aligned}
\sum_{s=1}^{M} f(\pi_s) \sum_{t=1}^{L} f(\rho_t) &= \sum_{s=1}^{M} \left(\lambda c(\pi_s) + \frac{1}{N} \sum_{i=1}^{N} \pi_s(i) \right) \sum_{t=1}^{L} \left(\lambda c(\rho_t) + \frac{1}{N} \sum_{i=1}^{N} \rho_t(i) \right) \\
&= \left(\lambda \sum_{s=1}^{M} c(\pi_s) + 1 \right) \left(\lambda \sum_{t=1}^{L} c(\rho_t) + 1 \right) \\
&= \lambda \left(\sum_{s=1}^{M} c(\pi_s) + \sum_{t=1}^{L} c(\rho_t) + \lambda \sum_{s=1}^{M} c(\pi_s) \sum_{t=1}^{L} c(\rho_t) \right) + 1 \\
&= \lambda \sum_{s=1}^{M} \sum_{t=1}^{L} c(\pi_s \rho_t) + 1 \\
&= \sum_{s=1}^{M} \sum_{t=1}^{L} \left(\lambda c(\pi_s \rho_t) + \frac{1}{N} \sum_{i=1}^{N} \pi_s \rho_t(i) \right) \\
&= \sum_{s=1}^{M} \sum_{t=1}^{L} f(\pi_s \rho_t).
\end{aligned} \tag{B.5}$$

Thus, C satisfies (3.2) if and only if f satisfies

$$\sum_{s=1}^{M}\sum_{t=1}^{L} f(\pi_s \rho_t) = \sum_{s=1}^{M}\sum_{t=1}^{L} f(\pi_s) f(\rho_t). \tag{B.6}$$

It follows from Lemma 3 that one of the following holds:

1. There exists a constant $a \in \mathbb{R}$ and an additive $A : \mathbb{R}^N \to \mathbb{R}$ such that for all $x \in (0,1)^N$,

$$c(x) = \frac{1}{\lambda}\left(A(x) + a - \frac{1}{N}\sum_{i=1}^{N} x(i)\right), \tag{B.7}$$

   where a and A satisfy

$$A(\mathbf{1}) + MLa = (A(\mathbf{1}) + Ma)(A(\mathbf{1}) + La) \tag{B.8}$$

   for all $M \geq 3$, $L \geq 2$ or $M \geq 2$, $L \geq 3$.

2. There exists a multiplicative $D : \mathbb{R}_+^N \to \mathbb{R}$ and an additive $B : \mathbb{R}^N \to \mathbb{R}$ with $B(\mathbf{1}) = 0$ such that for all $x \in (0,1)^N$,

$$c(x) = \frac{1}{\lambda}\left(D(x) + B(x) - \frac{1}{N}\sum_{i=1}^{N} x(i)\right). \tag{B.9}$$

We will consider the two cases separately.

Case 1 Suppose that c has the form in (B.7). First we will split A and D by dimension to define the additive function $A_i : \mathbb{R} \to \mathbb{R}$ and the multiplicative function $D_i : \mathbb{R}_+ \to \mathbb{R}$. When looking at the function A for the vector which has $x(i)$ in entry i and 0's elsewhere, we write $A_i(x(i))$, so that $A_i(x(i)) = A(0, \ldots, 0, x(i), 0, \ldots, 0)$. Similarly for D when taking the vector which has $x(i)$ in entry i and 1's elsewhere, so that $D_i(x(i)) = D(1,\ldots,1,x(i),\ldots,1)$. Hence we can write A and D in the forms

$$A(x(1),\ldots,x(N)) = \sum_{i=1}^{N} A_i(x(i)),$$
$$D(x(1),\ldots,x(N)) = \prod_{i=1}^{N} D_i(x(i)).$$

We then we have that for all $\pi \in \Pi$,

$$\begin{aligned}
\lambda C(\pi) &= \lambda \sum_{s=1}^{M} c(\pi_s) \\
&= \sum_{s=1}^{M} \left(A(\pi_s) + a - \frac{1}{N} \sum_{i=1}^{N} \pi_s(i)) \right) \\
&= \sum_{s=1}^{M} \sum_{i=1}^{N} A_i(\pi_s(i)) + Ma - \frac{1}{N} \sum_{s=1}^{M} \sum_{i=1}^{N} \pi_s(i) \\
&= \sum_{i=1}^{N} A_i \left(\sum_{s=1}^{M} \pi_s(i) \right) + Ma - \frac{N}{N} \\
&= \sum_{i=1}^{N} A_i(1) + Ma - 1 \\
&= A(\mathbf{1}) + Ma - 1,
\end{aligned}$$

so that $C(\pi) = \frac{1}{\lambda}(Ma + A(\mathbf{1}) - 1)$.

Now take any $M, M', M'', L, L', L''' \geq 3$. Using (B.8) we can solve the system of equations

$$A(\mathbf{1}) + MLa = (A(\mathbf{1}) + Ma)(A(\mathbf{1}) + La)$$
$$A(\mathbf{1}) + M'L'a = (A(\mathbf{1}) + M'a)(A(\mathbf{1}) + L'a)$$
$$A(\mathbf{1}) + M''L''a = (A(\mathbf{1}) + M''a)(A(\mathbf{1}) + L''a),$$

and find that there are three possible solutions:

$$
\begin{aligned}
&A(\mathbf{1}) = 0 \text{ and } a = 0, \text{ or} \\
&A(\mathbf{1}) = 1 \text{ and } a = 0, \text{ or} \\
&A(\mathbf{1}) = 0 \text{ and } a = 1.
\end{aligned}
$$

The forms of C corresponding to these solutions are

$$C(\pi) = -\frac{1}{\lambda}, \tag{B.10}$$

$$C(\pi) = 0, \tag{B.11}$$

$$C(\pi) = \frac{1}{\lambda}(M-1). \tag{B.12}$$

The assumption that C is nonconstant rules out (B.10) and (B.11). The case in (B.12) gives us (3.3) with $\alpha_1 = \cdots = \alpha_N = 0$.

Case 2 If c is of the form (B.9), then for some additive $B : \mathbb{R}^N \to \mathbb{R}$ we have that for all $\pi \in \Pi$,

$$
\begin{aligned}
\lambda C(\pi) &= \lambda \sum_{s=1}^{M} c(\pi_s) \\
&= \sum_{s=1}^{M} \left(D(\pi_s) + B(\pi_s) - \frac{1}{N} \sum_{i=1}^{N} \pi_s(i) \right) \\
&= \sum_{s=1}^{M} D(\pi_s) + B(\mathbf{1}) - 1 \\
&= \sum_{s=1}^{M} D(\pi_s) - 1,
\end{aligned}
$$

so that $C(\pi) = \frac{1}{\lambda}\left(\sum_{s=1}^{M} D(\pi_s) - 1\right)$. Since C is continuous $D : \mathbb{R}_+^N \to \mathbb{R}$ is measurable. The only form for a measurable multiplicative function is

$$D(x(1), \dots, x(N)) = \prod_{i=1}^{N} x(i)^{\alpha_i}$$

for some $\alpha_1, \dots, \alpha_N \in \mathbb{R}$ (Ebanks et al., 1998, Example 1.3.5). So C is of the form in (3.3).

Therefore, C is of the form in (3.3) for some for some $\alpha_1, \dots, \alpha_N \in \mathbb{R}$. Note that the generating function is

$$c(x) = \frac{1}{\lambda}\left(\prod_{i=1}^{N} x(i)^{\alpha_i} - \frac{1}{N}\sum_{i=1}^{N} x(i)\right). \tag{B.13}$$

Since C assigns the same cost to Blackwell equivalent experiments, by Proposition 1, we require c in (B.13) to be homogeneous. For $\beta \in (0,1)$ and $x \in (0,1)^N$ we have that

$$\begin{aligned}
c(\beta x) &= \frac{1}{\lambda}\left(\prod_{i=1}^{N} \beta^{\alpha_i} x(i)^{\alpha_i} - \frac{1}{N}\sum_{i=1}^{N} \beta x(i)\right) \\
&= \frac{1}{\lambda}\left(\prod_{i=1}^{N} \beta^{\alpha_i} \prod_{i=1}^{N} x(i)^{\alpha_i} - \beta\frac{1}{N}\sum_{i=1}^{N} x(i)\right) \\
&= \frac{1}{\lambda}\left(\beta^{\alpha_1+\dots+\alpha_N} \prod_{i=1}^{N} x(i)^{\alpha_i} - \beta\frac{1}{N}\sum_{i=1}^{N} x(i)\right) \\
&= \beta\frac{1}{\lambda}\left(\beta^{\alpha_1+\dots+\alpha_N-1} \prod_{i=1}^{N} x(i)^{\alpha_i} - \frac{1}{N}\sum_{i=1}^{N} x(i)\right).
\end{aligned}$$

Since the nonconstancy of C implies that we cannot have $\alpha_i = 0$ for all i, it follows that that $c(\beta x) = \beta c(x)$ if and only $\sum_{i=1}^{N} \alpha_i = 1$.

For the converse, suppose that a cost function C takes the form in (3.6) for all $\pi \in \Pi$, with $\sum_{i=1}^{N} \alpha_i = 1$ and c in (B.13) convex. Then the continuity and homogeneity of c in (B.13), along with Proposition 1, implies that C is monotonic, continuous, and linear. Moreover, since $\sum_{i=1}^{N} \alpha_i = 1$, it follows that C is nonconstant. Finally, by plugging (B.13) into (B.3) we can verify that C satisfies (3.5).

□

B.2 Proof of Theorem 2

The following lemma is from Theorem 7.2.1 in Ebanks et al. (1998)

Lemma 4. *Let $f : (0,1)^N \to \mathbb{R}$ be measurable in each variable. Then f satisfies*

$$\sum_{s=1}^{M}\sum_{t=1}^{L} f(\pi_s \rho_t) = \sum_{s=1}^{M} f(\pi_s) + \sum_{t=1}^{L} f(\rho_t)$$

for all $\pi \in \Pi_M$ and $\rho \in \Pi_L$, for a fixed pair of integers $M \geq 2$, $L \geq 3$ or $M \geq 3$, $L \geq 2$, if and only if there exist constants $a_1, \ldots, a_N \in \mathbb{R}$, $b \in \mathbb{R}$, and δ_{ij}, $i, j = 1, \ldots, N$, such that for all $x \in (0,1)^N$

$$f(x) = \sum_{i=1}^{N}\sum_{j=1}^{N} x(i)\delta_{ij} \ln x(j) + \sum_{i=1}^{N} a_i x(i) + b,$$

with

$$\sum_{i=1}^{N} a_i = (ML - M - L)b,$$

where $x(i)$ denotes the i^{th} component of x.

Proof of Theorem 2. Suppose that a cost function C is monotonic, linear, continuous, nonconstant, and satisfies (3.5) for all $\pi, \rho \in \Pi$. Then by Proposition 1 there exists a continuous, homogeneous function $c : (0,1)^N$ such that $C(\pi) = \sum_{s=1}^{M} c(\pi_s)$ for all $\pi \in \Pi$. It follows that for all $\pi \in \Pi_M$ and $\rho \in \Pi_L$,

$$C(\pi \otimes \rho) = \sum_{s=1}^{M}\sum_{t=1}^{L} c(\pi_s \rho_t).$$

Since C satisfies (3.5), it follows that c satisfies

$$\sum_{s=1}^{M}\sum_{t=1}^{L} c(\pi_s \rho_t) = \sum_{s=1}^{M} c(\pi_s) + \sum_{t=1}^{L} c(\rho_t)$$

for all M and L. So by Lemma 4, there exist constants $a_1, \ldots, a_N \in \mathbb{R}$ and δ_{ij}, $i, j = 1, \ldots, N$ such that for all $x \in (0,1)^N$,

$$c(x) = \sum_{i=1}^{N}\sum_{j=1}^{N} x(i)\delta_{ij} \ln x(j) + \sum_{i=1}^{N} a_i x(i),$$

with $\sum_{i=1}^{N} a_i = 0$. It follows that

$$\begin{aligned} C(\pi) &= \sum_{s=1}^{M}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} \pi_s(i)\delta_{ij} \ln \pi_s(j) + \sum_{i=1}^{N} a_i \pi_s(i)\right] \\ &= \sum_{s=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{N} \pi_s(i)\delta_{ij} \ln \pi_s(j). \end{aligned} \tag{B.14}$$

Since c is homogeneous of degree one, for any $\beta \in (0,1)$ it must hold that $c(\beta x) = \beta c(x)$ for all $x \in (0,1)^N$. Now note that for $\beta \in (0,1)$ and $x \in (0,1)^N$,

$$\begin{aligned} c(\beta x) &= \sum_{i=1}^{N}\sum_{j=1}^{N} \beta x(i)\delta_{ij} \ln(\beta x(j)) + \sum_{i=1}^{N} a_i \beta x(i) \\ &= \beta \sum_{i=1}^{N}\sum_{j=1}^{N} x(i)\delta_{ij}(\ln \beta + \ln x(j)) + \beta \sum_{i=1}^{N} a_i x(i) \\ &= \beta c(x) + \sum_{i=1}^{N}\sum_{j=1}^{N} x(i)\delta_{ij} \ln \beta. \end{aligned}$$

It must follow that

$$\sum_{i=1}^{N}\sum_{j=1}^{N} x(i)\delta_{ij} = 0$$

for all $x \in (0,1)^N$. This in turn implies that $\sum_{j=1}^N \delta_{ij} = 0$ for all i, or equivalently

$$\delta_{ii} = -\sum_{\substack{j=1 \\ j \neq i}}^{N} \delta_{ij}, \tag{B.15}$$

for all i. Defining $\beta_{ij} = -\delta_{ij}$ for $i \neq j$, we can combine (B.14) and (B.15) to get

$$C(\pi) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \beta_{ij} \sum_{s=1}^{M} \pi_s(i) \ln\left(\frac{\pi_s(i)}{\pi_s(j)}\right).$$

Since C is nonconstant, the parameters β_{ij} are not all zero.

For the converse, suppose that a cost function has the form in (3.6) with parameters $\beta_{ij} \in \mathbb{R}_+$ not all zero. Then we can write C as $C(\pi) = \sum_{s=1}^M c(\pi_s)$, where for $x \in (0,1)^N$,

$$c(x) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \beta_{ij} x(i) \ln\left(\frac{x(i)}{x(j)}\right).$$

Since c is homogeneous of degree one, convex, and continuous, it follows from Proposition 1 that C assigns the same cost to Blackwell equivalent experiments and is continuous and linear. We also have that for all $\pi, \rho \in \Pi$,

$$\begin{aligned}
\sum_{s=1}^{M} \sum_{t=1}^{L} c(\pi_s \rho_t) &= \sum_{s=1}^{M} \sum_{t=1}^{L} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \beta_{ij} \pi_s(i) \rho_t(i) \ln\left(\frac{\pi_s(i)\rho_t(i)}{\pi_s(j)\rho_t(j)}\right) \\
&= \sum_{s=1}^{M} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \beta_{ij} \pi_s(i) \ln\left(\frac{\pi_s(i)}{\pi_s(j)}\right) + \sum_{t=1}^{L} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \beta_{ij} \rho_t(i) \ln\left(\frac{\rho_t(i)}{\rho_t(j)}\right) \\
&= \sum_{s=1}^{M} c(\pi_s) + \sum_{t=1}^{L} c(\rho_t),
\end{aligned}$$

so that C satisfies (3.5). Since the parameters β_{ij} are not all zero, C is nonconstant.

□

C Proof of Proposition 3

Proof. We will first construct an attention cost function K and then show that it represents C.

By Proposition 1, the cost function C can be written as in (2.2) with some generating function $c : (0,1)^N \to \mathbb{R}$ that is convex and homogeneous of degree one. Let $\mu \in \Delta(\Theta)$ be an interior belief. Given a distribution $\nu \in \Delta(\Theta)$, define $i_\nu \in \Theta$ as a state that maximizes the ratio $\nu(i)/\mu(i)$:

$$i_\nu \in \arg\max_{i \in \Theta} \frac{\nu(i)}{\mu(i)}.$$

Now define the measure of uncertainty $H_\mu : \Delta(\Theta) \to \mathbb{R}$ as

$$H_\mu(\nu) = \frac{\nu(i_\nu)}{\mu(i_\nu)} c\left(\frac{\nu(1)}{\mu(1)}\frac{\mu(i_\nu)}{\nu(i_\nu)}, \ldots, \frac{\nu(N)}{\mu(N)}\frac{\mu(i_\nu)}{\nu(i_\nu)}\right). \tag{C.1}$$

Since c is convex and homogeneous of degree one, H_μ must also be convex and homogeneous of degree one. Thus we can consider the attention cost function

$$K(\mu, \tau) = \sum_{\nu \in \operatorname{supp}(\tau)} \tau(\nu) H_\mu(\nu) - H_\mu(\mu).$$

Note that $H_\mu(\mu) = c(1, \ldots, 1) = 0$.

Next we will show that for all experiments $\pi \in \Pi$

$$C(\pi) = K(\mu, \langle \pi | \mu \rangle).$$

Take an experiment (M, π) and for all $s = 1, \ldots, M$ define $\nu_s \in \Delta(\Theta)$ to be the posterior distribution over states after having observed the signal realization s from π:

$$\nu_s(i) = \frac{\pi_s(i)\mu(i)}{\sum_{j \in \Theta} \pi_s(j)\mu(j)}.$$

Let $\tau = \langle \pi | \mu \rangle \in \Delta\Delta(\Theta)$ be the distribution over posteriors induced by π, so that for all $\nu \in \text{supp}(\tau)$

$$\tau(\nu) = \sum_{s|\nu_s=\nu} \sum_{i \in \Theta} \mu(i)\pi_s(i).$$

Using the definition in (C.1) and the homogeneity of H_μ we have that

$$\begin{aligned}
C(\pi) &= \sum_{s=1}^{M} c(\pi_s) \\
&= \sum_{s=1}^{M} H_\mu(\pi_s \mu) \\
&= \sum_{\nu \in \text{supp}(\tau)} \sum_{s|\nu_s=\nu} H_\mu(\pi_s \mu) \\
&= \sum_{\nu \in \text{supp}(\tau)} \sum_{s|\nu_s=\nu} H_\mu\left(\nu_s \sum_{i \in \Theta} \pi_s(i)\mu(i)\right) \\
&= \sum_{\nu \in \text{supp}(\tau)} \sum_{s|\nu_s=\nu} \left(\sum_{i \in \Theta} \pi_s(i)\mu(i)\right) H_\mu\left(\nu\right) \\
&= \sum_{\nu \in \text{supp}(\tau)} \tau(\nu) H_\mu(\nu) \\
&= \sum_{\nu \in \text{supp}(\tau)} \tau(\nu) H_\mu(\nu) - H_\mu(\mu) \\
&= K(\mu, \tau)
\end{aligned}$$

where multiplication of vectors is performed component-wise.

Therefore, C has a posterior-separable representation. □

References

Blackwell, D. (1951). Comparison of experiments, proceedings of the second berkeley symposium on mathematical statistics and probability.

Bloedel, A. W. and W. Zhong (2020). The cost of optimally-acquired information.

Unpublished Manuscript, November.

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

Caplin, A., M. Dean, and J. Leahy (2022). Rationally inattentive behavior: Characterizing and generalizing shannon entropy. *Journal of Political Economy 130*(6), 1676–1715.

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

Denti, T. (2022). Posterior separable cost of information. *American Economic Review 112*(10), 3215–3259.

Denti, T., M. Marinacci, and A. Rustichini (2022). Experimental cost of information. *American Economic Review 112*(9), 3106–3123.

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

Gentzkow, M. and E. Kamenica (2014). Costly persuasion. *American Economic Review 104*(5), 457–62.

Hébert, B. and M. Woodford (2017). Rational inattention and sequential information sampling. Technical report, National Bureau of Economic Research.

Hébert, B. and M. Woodford (2021). Neighborhood-based information costs. *American Economic Review 111*(10), 3225–3255.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

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.

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

Morris, S. and P. Strack (2019). The wald problem and the relation of sequential sampling and ex-ante information costs. *Available at SSRN 2991567*.

Pomatto, L., P. Strack, and O. Tamuz (2023). The cost of information: The case of constant marginal costs. *American Economic Review*.

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

Torgersen, E. (1991). *Comparison of statistical experiments*, Volume 36. Cambridge University Press.

Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics 52*(1-2), 479–487.

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

Wald, A. (1945). Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 265–280.