The Cost of Optimally Acquired Information[*]

Alexander W. Bloedel[†] Weijie Zhong[‡]

August 19, 2024

Abstract

We study the "reduced-form" *indirect cost* of information arising from flexible sequential minimization of a "primitive" *direct cost* function. Indirect cost functions are characterized by a simple recursive condition, *sequential learning-proofness (SLP)*. Under a smoothness condition, (i) SLP is equivalent to uniform posterior separability and (ii) the mapping from direct to indirect costs is tractably characterized by the cost of incrementally informative "diffusion" signals. We apply this framework to establish—and resolve—a trilemma among SLP and two other natural properties of information costs: prior invariance and constant marginal cost. Our analysis provides foundations for two new indirect cost functions: *Total Information* and the *Minimal Likelihood Ratio (MLR) cost*.

1 Introduction

Information is a valuable but costly resource. There is a unified paradigm for modeling its value based on the extent to which it facilitates decision-making (D. A. Blackwell 1951). There is less consensus on how to model its cost. In this paper, we develop a framework for modeling the cost of information based on the core tenet of production theory: that outputs are produced at minimal cost by combining inputs optimally.

Our framework features a Bayesian decision-maker (DM) who learns about an uncertain state by acquiring costly information in the form of Blackwell experiments (i.e., signals correlated with the state). The DM's "primitive" information acquisition technology is described by an arbitrary *direct cost* function over experiments. The DM produces any "target experiment" by optimizing over all sequential information acquisition strategies that generate as much information as the target. We call the DM's minimal expected cost of producing target experiments her *indirect cost* function. The indirect cost

[*] This paper partially subsumes the working paper Bloedel and Zhong (2020). We thank Nageeb Ali, Doug Bernheim, Simon Board, Andrew Caplin, Mark Dean, Henrique de Oliveira, Tommaso Denti, Ben Hébert, Ian Jewitt, Emir Kamenica, Alessandro Lizzeri, Jay Lu, Paul Milgrom, Stephen Morris, Luciano Pomatto, Doron Ravid, Ilya Segal, Philipp Strack, Balazs Szentes, Omer Tamuz, Can Urgun, Leeat Yariv, and numerous conference and seminar audiences for helpful comments. Bloedel is grateful to the B.F. Haley and E.S. Shaw Fellowship for Economics for financial support and to Caltech for its hospitality.

[†] Department of Economics, UCLA. Email: abloedel@econ.ucla.edu

[‡] Graduate School of Business, Stanford University. Email: weijie.zhong@stanford.edu

function then represents the DM's "reduced-form" cost of acquiring information in any downstream decision problem (as in the rational inattention literature, e.g., Sims 2003; Matějka and McKay 2015; Maćkowiak, Matějka, and Wiederholt 2023).

We propose this framework as a unified way of capturing two key features of real-world information acquisition. First, in many important and widely studied settings, it is both feasible and optimal for the DM to acquire information piece-by-piece in a sequential fashion. For example, in a typical *statistical sampling problem* (Wald 1945), a firm learns about the demand for a new product by sequentially sampling consumers (e.g., via surveys, A/B tests, or RCTs), subject to a direct physical/pecuniary cost that depends on the sample's size and features.[1] In a typical *encoding problem* (Shannon 1948), an online consumer chooses between products by sequentially querying their attributes (e.g., on a price-comparison website), incurring a cognitive/computational cost per query. And, in a standard *perception task* (Ratcliff 1978), a lab subject faced with a visual stimulus gradually contemplates how to classify it, paying a cognitive cost while she thinks.[2]

Second, the cost of information is highly context-specific. In the above examples, to paraphrase Sims (2010, p. 161), the physical/pecuniary costs of generating new information through statistical sampling may bear no relation to the cognitive/computational costs of processing freely available information in encoding and perception tasks.

Accommodating both of these features is challenging, which has required the literature to make significant modeling compromises. A classical approach involves studying sequential learning with specific direct costs/production technologies, as in the literature on sequential sampling in statistics (Wald 1945; Wald 1947; Arrow, D. Blackwell, and Girshick 1949), optimal encoding in information theory (Shannon 1948; Huffman 1952), and drift-diffusion models of perception in psychology/neuroscience (Ratcliff and McKoon 2008; Fehr and Rangel 2011; Fudenberg, Strack, and Strzalecki 2018). By adopting specific "units for information," such frameworks are "useful but only on very limited problems" (Arrow 1996, p. 120). On the other hand, the modern rational inattention paradigm (Sims 2003) abstracts away from the underlying production procedure and instead justifies particular reduced-form cost functions via context-specific axioms (Hébert

[1] For instance, FDA 2019 encourages the use of multi-stage RCTs in the context of clinical trials for pharmaceutical products. Sequential A/B testing is common in the tech industry (e.g., Johari et al. 2022).

[2] Other examples of sequential learning abound: scientific research and industrial R&D involve multiple adaptively designed stages of experimentation, voters learn about political issues by reading news articles one-by-one, and so on. Even in settings where information acquisition may initially appear to be one-shot, there is often some degree of sequentiality. In perception tasks, subjects' response times are nonzero but short (e.g., on the order of seconds). In statistical sampling problems, non-sequential (fixed sample size) procedures still take time to implement and can be interpreted as non-contingent sequential procedures.

and Woodford 2021; Caplin, Dean, and Leahy 2022; Denti, Marinacci, and Rustichini 2022; Pomatto, Strack, and Tamuz 2023) or their implications for the DM's choice behavior (Caplin and Dean 2015; Oliveira et al. 2017; Denti 2022; Dean and Neligh 2023). Our framework bridges these two perspectives, permitting analysis of both the *context-free* implications of sequential optimization for reduced-form (indirect) cost functions and *context-specific* cost functions arising from optimization in particular settings.

The Indirect Cost of Information. Our first contribution is to characterize the full class of indirect costs in terms of a novel recursive property that we call *sequential learning-proofness* (SLP). A cost function is SLP if the cost of acquiring any target experiment in one shot is weakly lower than the expected cost of decomposing it into two steps. We interpret this as a minimal "internal consistency" requirement for any reduced-form model of information cost: if the DM's cost function were *not* SLP, then she could optimize away some of its features using a simple two-step strategy. We show that a cost function is the indirect cost for *some* underlying direct cost *if and only if* it is SLP (Theorem 1). Thus, SLP fully characterizes the "context-free" implications of sequential optimization.

We then show (Theorem 2) that a cost function C is SLP and *Regular* (a weak notion of "local differentiability") *if and only if* it has a *uniformly posterior separable* (UPS) representation, i.e., given the set Θ of states, there is some convex "potential function" $H : \Delta(\Theta) \to (-\infty, +\infty]$ such that

$$C(\pi) = C^H_{\text{ups}}(\pi) := \mathbb{E}_\pi \left[H(q) - H(p) \right] \tag{UPS}$$

for every prior belief $p \in \Delta(\Theta)$ and distribution $\pi \in \Delta(\Delta(\Theta))$ of Bayesian posteriors $q \in \Delta(\Theta)$ induced by some experiment. The class of UPS costs (introduced by Caplin, Dean, and Leahy 2022) includes most specifications studied in the rational inattention literature, including mutual information (Sims 2003; Matějka and McKay 2015) and the more general family of neighborhood-based costs (Hébert and Woodford 2021). Theorem 2 provides a novel optimality-based rationale for using such UPS costs in applications.

The Sequential Learning Map. Our second contribution is to characterize the *sequential learning map*, Φ, that transforms each direct cost C into its corresponding indirect cost $\Phi(C)$ (see Figure 1). This map determines how properties of a given direct cost function are transformed (or preserved) under optimization. Conversely, the *pre-image of* this map determines the "primitive" economic assumptions that are implicitly imposed on the underlying direct cost when one uses a particular functional form for the indirect cost.

Central to our characterization is an object that we call the *kernel* of a cost function, which summarizes the cost of "incremental evidence," i.e., experiments that shift posterior beliefs only locally (analogous to a continuous-time "diffusion signal"). Our key observation is that the kernel of any direct cost is *invariant* under the sequential learning map, i.e., the cost of incremental evidence cannot be further optimized.

We proceed in two steps. First, we develop general lower and upper bounds on the sequential learning map. For any direct cost C, the indirect cost $\Phi(C)$ is (i) *locally* bounded below by the kernel of C and (ii) *globally* bounded above by the UPS cost obtained by integrating the kernel of C (Theorem 3). Economically, the UPS upper bound is the total expected cost associated with the *incremental learning* strategy that only acquires incremental evidence (which, in general, need not be an optimal strategy under C).

Second, we show that the upper bound is tight *if and only if* the indirect cost $\Phi(C)$ is Regular/UPS. Formally, we obtain an exact characterization of the sequential learning map for the co-domain of Regular/UPS indirect costs. Given such an indirect cost C^H_{ups}, a direct cost C satisfies $\Phi(C) = C^H_{\text{ups}}$ *if and only if* (i) the kernel of C is Hess$H$ (the Hessian of the potential function H) and (ii) C *favors learning via incremental evidence* (FLIEs), i.e., weakly exceeds the expected cost of incremental learning (Theorem 4).

Since it is natural to assume that the direct cost FLIEs is some applications but not in others, Theorem 4 helps delineate when Regular/UPS indirect costs are economically reasonable. Theorem 4 also delivers a tractable method for calculating Regular/UPS from their direct costs, and vice versa. As a proof of concept, we illustrate this method for several important classes of UPS costs from the literature (Section 5.1).

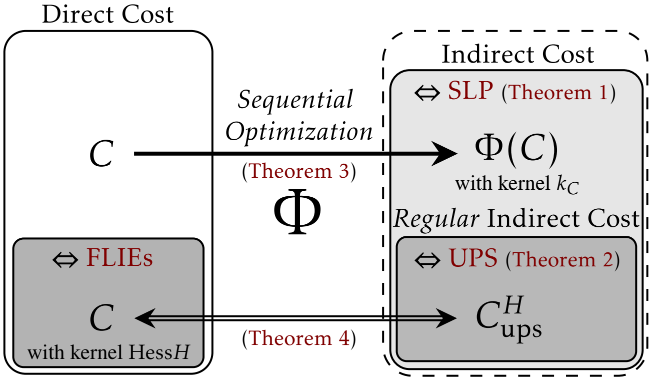

Figure 1: The sequential learning map.

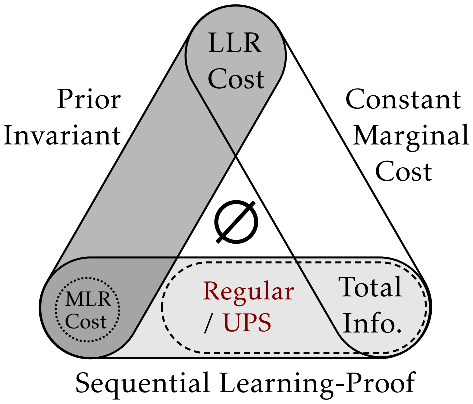

Figure 2: Information cost trilemma.

Information Cost Trilemmas. Our third contribution is to characterize the implications of sequential optimization in specific economic contexts. This exercise serves two purposes: to pinpoint specific indirect cost functions for use in applications, and to elucidate inherent modeling tradeoffs in the rational inattention literature.

To these ends, we study how our notion of indirect (or SLP) cost interacts with two axioms that the literature has advocated for imposing on reduced-form cost functions. The first axiom, Prior Invariance, requires the cost of any given Blackwell experiment to be independent of the DM's prior beliefs. This property is natural when modeling physical or pecuniary information costs (e.g., statistical sampling or R&D).[3] The second axiom, Constant Marginal Cost (CMC), posits that the cost of running two independent experiments together equals the sum of their individual costs. Pomatto, Strack, and Tamuz (2023) propose this property as a non-parametric way of modeling costs that are "linear in sample size," which is a familiar and natural assumption in statistical sampling problems.

We offer two characterization results. First, we establish an *information cost trilemma* (Theorem 5) among these three natural properties: SLP, Prior Invariance, and CMC. An information cost function can satisfy any two of these properties, but no nonzero cost function can satisfy all three of them (see Figure 2). Pomatto, Strack, and Tamuz (2023) show that the unique Prior Invariant and CMC cost function is the *Log-Likelihood Ratio (LLR) cost*.[4] We show that the unique SLP and CMC cost function is the *Total Information* cost, a novel UPS cost defined by the potential function

$$H_{\mathrm{TI}}(q) := \sum_{\theta,\theta' \in \mathrm{supp}(p)} \gamma_{\theta,\theta'} q(\theta) \log\left(\frac{q(\theta)}{q(\theta')}\right), \tag{TI}$$

where the coefficients $\gamma_{\theta,\theta'} \geq 0$, which control the cost of distinguishing between pairs of states, can be chosen by the modeler.We conclude that Total Information is the natural reduced-form cost function in applications where CMC is a desirable assumption, such as statistical sampling problems. To show that the remaining two-way intersection is nonempty, we construct the SLP and Prior Invariant *Minimal Likelihood Ratio (MLR) cost*:

$$C_{\mathrm{MLR}}(\pi) := \mathbb{E}_\pi\left[\max_{\theta \in \mathrm{supp}(p)} \left\{1 - \frac{q(\theta)}{p(\theta)}\right\}\right] \tag{MLR}$$

for every prior p and distribution $\pi \in \Delta(\Delta(\Theta))$ over posteriors q. We argue that the MLR cost is useful in certain applications, such as costly monitoring in games (see Section 6.1).

We argue that the main tension in the trilemma is between SLP and Prior Invariance. For instance, no commonly studied Prior Invariant reduced-form costs in the literature are SLP. This tension is natural: since SLP costs are derived from *expected* cost-minimization, they "should" *endogenously* depend on prior beliefs. Our framework also suggests a natural way to alleviate this tension: view Prior Invariance as a natural "prim-

[3] Many authors have advocated for this property on these and other grounds. See, e.g., Woodford (2012), Gentzkow and Kamenica (2014), Mensch (2018), and Denti, Marinacci, and Rustichini (2022).

[4] Formally, their characterization of the LLR cost requires an additional "dilution linearity" axiom, which is implied by SLP and omitted from the present discussion for simplicity.

itive" property for *direct* costs, rather than a "reduced-form" property for *indirect* costs.

Following this logic, we introduce the novel class of *Sequentially Prior Invariant* (SPI) cost functions: indirect costs that are generated by Prior Invariant direct costs. Our second characterization result uses this notion to resolve the information cost trilemma. Given a binary state space $\Theta = \{0, 1\}$ and letting $q \in [0, 1]$ denote the probability of $\theta = 1$, Morris and Strack (2019) define the *Wald cost* function as

$$C_{\text{Wald}} := C_{\text{ups}}^{H^*}, \quad \text{where } \; H^*(q) := (2q-1)\log\left(\frac{q}{1-q}\right), \tag{Wald}$$

i.e., the special case of Total Information with binary states and symmetric coefficients ($\gamma_{0,1} = \gamma_{1,0} = 1$). We establish a three-way equivalence: a cost function is SPI and CMC *if and only if* it is SPI and Regular/UPS *if and only if* it is proportional to the Wald cost (Theorem 6). From a positive perspective, the Wald cost resolves the trilemma and, equally importantly, demonstrates that the UPS model can be justified by optimization of a physical/pecuniary direct cost. However, the Wald cost is the *unique* cost function with either of these virtues and, being defined only for binary-state settings, is very special.

Thus, our application concludes with a challenge for the literature, in the form of a new *modeler's trilemma*. Every modeler desires three things: realism (SPI), tractability (Regularity), and generality (general state space). Pick any two; you can't have all three.

Roadmap. With the majority of related papers discussed already, we review additional related papers below. Section 2 presents the framework. Sections 3 and 4 characterize the class of indirect cost functions and the sequential learning map, respectively. Section 5 develops the information cost trilemma and other applications. Section 6 concludes with a discussion of extensions and open questions.

1.1 Related Literature

The earliest example of an indirect cost appears in Shannon (1948), which introduces the concept of mutual information and shows that it *approximates* the indirect cost for encoding problems where the direct cost assigns equal cost to all "bits" (i.e., binary partitions of the state space) and infinite cost to all other experiments. In Section 5.1, we characterize all direct costs that yield mutual information as their *exact* indirect cost, offering an optimization foundation for the baseline rational inattention model (Sims 2003).[5]

[5] Shannon's result is often interpreted as providing an optimization foundation for mutual information as a reduced-form cost function in rational inattention models (Sims 2003; Sims 2010). However, a well-known caveat to this interpretation is that Shannon's direct cost only generates mutual information as its *exact* indirect cost if the DM can simultaneously "block code" many i.i.d. draws of the state (Cover and Thomas 2006, Ch. 10), which may not be a compelling assumption for economic applications.

Our work builds on Morris and Strack (2019), which derives UPS costs as reduced-form cost functions arising from sequential sampling of continuous-time diffusion signals. Their key assumptions are that (i) the DM samples from an *exogenous* diffusion signal process, choosing only when to stop, and (ii) there are only two states.[6] Hébert and Woodford (2023, Proposition 7), concurrent to our work, derive a similar result, assuming that the DM has an *exogenous* preference for using only diffusion signal processes. These results could be viewed as special cases of the "sufficiency" direction of our Theorem 4 (i.e., $\Phi(C)$ is UPS *if* C FLIEs).[7] Our framework gains its generality and power by allowing the DM to optimally choose the signal process. We significantly extended these results by establishing that UPS indirect costs arise from arbitrary direct costs and general sequential learning strategies *only if* the optimal strategy is to employ diffusion signals (the "necessity" direction of Theorem 4).

There is a small literature building on our paper. Several papers impose SLP as an axiom on reduced-form cost functions in various applications (e.g., Müller-Itten, Armenter, and Stangebye 2023; Wong 2023; Li 2022). Hébert and Woodford (2023, Section 5) and Miao and Xing (2024) apply Total Information in optimal stopping and dynamic decision problems, respectively. Denti, Marinacci, and Rustichini (2022, Section 2) study the special case of our framework with Prior Invariant direct costs and develop a variant of our finding (implied by Theorem 6) that no nonzero, bounded UPS cost function is SPI.

2 Model

A decision-maker (DM) can acquire information about an unknown *state* $\theta \in \Theta$, where Θ is a finite set. The DM's *beliefs* about the state are elements $p, q \in \Delta(\Theta)$ of the simplex. By convention, we let p denote the DM's *prior* belief and q denote his *posterior* belief.

For most of our analysis, we take a belief-based approach and model information as a *random posterior* $\pi \in \mathcal{R} := \Delta(\Delta(\Theta))$ induced by some Blackwell experiment.[8] For every $\pi \in \mathcal{R}$, we let $p_\pi := \mathbb{E}_\pi[q]$ denote the prior belief that is consistent with Bayesian updating from π. Let $\mathcal{R}^\circ := \bigcup_{p \in \Delta(\Theta)} \{\delta_p\}$ denote the set of all degenerate measures (correspond-

[6] When there are more than two states, Morris and Strack (2019) show that, due to the exogeneity of their signal process, only a "small set" of target experiments can be implemented by some stopping strategy.

[7] Hébert and Woodford 2023 assume "preference for gradual learning" (PGL), which is a stronger version of FLIEs that is defined only for "posterior separable" direct costs. Formally, Theorem 4 does not nest the results of these papers because we study a discrete-time framework. In Section 6.2, we describe an extension of our framework that lets us embed the continuous-time information acquisition procedures of these papers, facilitating a more direct comparison.

[8] Formally, fixing a Polish *signal space* S, a *(statistical) experiment* σ is a measurable map $\sigma : \Theta \to \Delta(S)$. It is well-known that every experiment σ is equivalently characterized by the Bayesian random posterior that it induces. Thus, for convenience, we often refer to random posteriors and experiments interchangeably.

ing to uninformative experiments). In Section 6.1, we discuss the implications of the belief-based approach and show how our model can be extended to handle alternative formulations.

An *(information) cost function* is a map $C : \mathcal{R} \to \overline{\mathbb{R}}_+$ satisfying $C(\mathcal{R}^\circ) = \{0\}$, i.e., trivial experiments have zero cost.[9] We make no other *a priori* assumptions about the shape of C or the structure of $\mathrm{dom}(C) := \{\pi \in \mathcal{R} \mid C(\pi) < +\infty\}$, the set of feasible experiments. This generality allows us to capture a wide range of settings, including those where $\mathrm{dom}(C)$ is highly restricted.

Let $\mathcal{C}$ denote the set of all information costs. We endow $\mathcal{C}$ with addition, multiplication by positive scalars, and the pointwise order $\leq$ (i.e., $C \leq C'$ if and only if $C(\pi) \leq C'(\pi)$ for all $\pi \in \mathcal{R}$). With this structure, $\mathcal{C}$ is a convex cone, a complete lattice, and closed under pointwise limits (formal analysis relegated to Appendix I).

2.1 Sequential learning and indirect cost

Given any target random posterior $\pi \in \mathcal{R}$, the DM aims to find the cheapest information acquisition procedure that "produces" π. The information acquisition needs not be one-shot, that is, there can be many "periods" and the experiments in "later" periods are specified by contingent plans of outcome of experiments in the "earlier" periods. We begin with defining a simplest two-period contingent plan: $\Pi \in \Delta(\mathcal{R})$ is a probability measure of random posteriors with finite non-degenerate support.[10]

1. *The first-period experiment*: the experiment in the first period (π_1) is implicitly defined by the projection $\pi_1(p_{\pi_2}) := \Pi(\pi_2)$, $\forall \pi_2 \in \mathrm{supp}(\Pi)$. In words, π_1 induces the interim beliefs that each will later become the prior of some second-period experiment (p_{π_2}).

2. *The second-period experiment*: given each interim belief p_{π_2}, the corresponding second-period experiment induces random posterior π_2.

Then, at the end of period two, the posterior belief has distribution $\mathbb{E}_\Pi[\pi_2]$, i.e. $\mathbb{E}_\Pi[\pi_2]$ is "produced" by Π. We define the map Ψ that characterizes cost minimization using two-step contingent plans.

Definition 1. *The two-step sequential learning map $\Psi : \mathcal{C} \to \mathcal{C}$ is defined by:*

$$\Psi(C)(\pi) := \inf_{\mathbb{E}_\Pi[\pi_2] \geq_{mps} \pi} C(\pi_1) + \mathbb{E}_\Pi\left[C(\pi_2)\right].^{11}$$

[9] $\overline{\mathbb{R}}_+ := \mathbb{R}_+ \cup \{+\infty\}$. Unless otherwise noted, we adopt the convention that $+\infty = +\infty$ when comparing unbounded functions.

[10] Π has finite non-degenerate support means $\mathrm{supp}(\Pi) \setminus \mathcal{R}^\circ$ is finite. The finiteness restriction is purely a technical restriction for guaranteeing measurability in various places of the analysis. The restriction is without loss if an additional continuity assumption is imposed on C.

$\Psi(C)$ characterizes the minimal total expected cost of two-step contingent plans that acquire *weakly more* information than the target output C. Note that by the construction, $0 \leq \Psi(C)(\pi) \leq C(\pi)$; hence, Ψ maps into $\mathcal{C}$. By imposing the restriction $\mathbb{E}_\Pi[\pi_2] \geq_{mps} \pi$, we implicitly assume the free disposal of information. An information cost is *sequential learning-proof* if it is a fixed point of Ψ; namely, under direct cost C, acquiring information in two steps is never strictly better than doing so in one shot.

Definition 2 (SLP)**.** *$C \in \mathcal{C}$ is Sequential Learning-Proof (SLP) if $\Psi(C) = C$.*

Next, we extend the two-step contingent plans to the contingent plans with arbitrary lengths to capture fully flexible sequential information acquisition. Intuitively, by applying Ψ to $\Psi(C)$, which is already optimized in two steps, we optimize over "four-step" contingent plans. Analogously, $\Psi^n(C)$ gives the optimal total cost of "2^n-step" information acquisition, and converges to the infinite horizon limit as $n \to \infty$. In what follows, we define this limit as the outcome of *fully flexible* cost minimization.

Definition 3. *The sequential learning map $\Phi : \mathcal{C} \to \mathcal{C}$ is defined by:*

$$\Phi(C)(\pi) := \lim_{n\to\infty} \Psi^n(C)(\pi).$$

Note that the limit is always well defined as $(\Psi^n(C)(\pi))$ is a non-negative and decreasing sequence. We call $\Phi(C)$ the *indirect cost* (function) generated by C.

Definition 4 (Indirect Cost)**.** *$\Phi(C) \in \mathcal{C}$ is the Indirect Cost generated by $C \in \mathcal{C}$. The set of all indirect cost functions $\{\Phi(C) \mid C \in \mathcal{C}\}$ is denoted by $\mathcal{C}^*$.*

We note that Φ satisfies three natural properties (formal analysis relegated to Appendix I). It is (i) *isotone*: $C \preceq C'$ implies that $\Phi(C) \preceq \Phi(C')$, (ii) *positively homogenous of degree* 1: $\Phi(\alpha C) = \alpha\Phi(C)$ for all $\alpha \geq 0$, and (iii) *concave*: $\Phi(\alpha C + (1-\alpha)C') \geq \alpha\Phi(C) + (1-\alpha)\Phi(C')$ for all $C, C' \in \mathcal{C}$ and $\alpha \in [0,1]$. These three properties are familiar from producer theory, where $\Phi(C)$ corresponds to the firm's cost function for producing outputs $\pi \in \mathcal{R}$ at input prices $C \in \mathcal{C}$.

A key element of our framework is that the information acquisition process can take arbitrarily many steps and contingencies, and one might wonder how restrictive such flexibility is. Our model excludes settings where time is costly (i.e., $\mathcal{R}^\circ$ has a non-zero cost) and the settings with non-stationary restrictions on information (i.e., the primitive cost function is time-dependent). In Section 6.2, we extend our model to more general settings that allow *arbitrary* restrictions on the sequential learning process, and provide minimal sufficient conditions for our main results.

[11] If $\operatorname{supp}(\Pi) \cup \{\pi_1\} \not\subset \operatorname{dom}(C)$, $C(\pi_1) + \mathbb{E}_\Pi[C(\pi_2)] = \infty$.

2.2 Illustrative Example

We illustrate the ideas of contingent plans and indirect costs by introducing the following running example.

Example 1 (Gaussian learning)

Information and direct cost: The unknown state $\theta \in \{0,1\}$. $\forall \psi \in (0,\infty)$, let s_ψ denote a signal with additive normal noise with precision ψ:

$$s_\psi = \theta + N\left(0, \tfrac{1}{\psi}\right).$$

The direct cost of acquiring s_ψ is $f(\psi)$, which satisfies $f(0)=0$, $f'(0)>0$, $f''(\psi)>0$.

A simple contingent plan: Let $s_{\psi/2}$ and $s'_{\psi/2}$ denote two (conditionally independent) signals each with precision $\frac{\psi}{2}$. Consider a simple two-period contingent plan: in the first stage, acquire $s_{\psi/2}$ and induce the Bayesian random posterior π_1. Then, in step two, following any interim belief, acquire $s'_{\psi/2}$ and induce the Bayesian random posterior π_2.[12]

By acquiring the two signals and observing the mean of them $\frac{1}{2}(s_{\psi/2} + s'_{\psi/2}) = \theta + N\left(0, \frac{1}{\psi}\right)$, signal s_ψ is "produced" at cost $2f(\psi/2)$ which is strictly lower than $f(\psi)$, i.e., sequential learning strictly benefits the DM. Of course, this contingent plan is not necessarily optimal; hence, letting π_ψ be the random posterior induced by s_ψ given some prior,

$$\Psi(C)(\pi_\psi) \leq 2f\left(\frac{\psi}{2}\right).$$

We can apply the argument again and produce s_ψ using the same contingent plan, but this time paying the cost of $\Psi(C)$. As a result, $\Psi^2(C)(\pi_\psi) \leq 2\Psi(C)(\pi_{\psi/2}) \leq 4f(\psi/4)$. Applying the argument recursively leads to

$$\Phi(C)(\pi_\psi) = \lim_{n\to\infty} \Psi^n(C)(\pi) \leq \lim_{n\to\infty} 2^n f(2^{-n}\psi) = f'(0)\cdot\psi.$$

A sequential learning strategy: Following the previous step, one might wonder if one could produce an arbitrary random posterior via "splitting" the signal. The answer is yes. We formalize such a strategy of acquiring many asymptotically uninformative signals — a diffusion process $\langle z_t \rangle$ defined by: $\mathrm{d}z_t = \theta \mathrm{d}t + \mathrm{d}W_t$, where W_t is a Wiener process. Observe that per $\mathrm{d}t$ unit of time, $\frac{\mathrm{d}z_t}{\mathrm{d}t} \approx \theta + \frac{1}{\mathrm{d}t}N(0,\mathrm{d}t) = s_{\mathrm{d}t}$. Therefore, the flow cost of observing the signal for $\mathrm{d}t$ unit of time is $\frac{f(\mathrm{d}t)}{\mathrm{d}t} \xrightarrow{\mathrm{d}t\to 0} f'(0)$.

An important observation by Morris and Strack 2019 is that by choosing a suitable stopping time τ, z_τ produces *any* random posterior $\pi \in \Delta[0,1]$. Moreover, for any such stopping time, the total flow cost is $f'(0)\mathbb{E}[\tau] \equiv f'(0)\cdot C_{\text{Wald}}$. Therefore, the Wald cost

[12] Formally, the constructed contingent plan has infinite support, which is not permitted. However, it is inconsequential as one can "sandwich" the contingent plan with finite-support ones with converging costs.

gives the "indirect cost" of a general signal via only "incremental evidence". Since such a strategy is not necessarily optimal,

$$\Phi(C) \leq f'(0) \cdot C_{\text{Wald}}.$$

Optimal sequential learning strategy: We claim without a proof that the inequality above holds as equality: $\Phi(C) = f'(0) \cdot C_{\text{Wald}}$. The proof will be straightforward once we establish several key properties of indirect costs in Section 3. This immediately implies that learning via incremental evidence is indeed the optimal sequential learning strategy in this example. In Section 4, we utilize this observation to obtain a general characterization of the sequential learning map Ψ.

3 The Indirect Cost of Information

In this section, we characterize the set of indirect cost functions $\mathcal{C}^*$. Evidently, every fixed point of Ψ is the indirect cost of itself, hence an indirect cost. In Section 3.1, we show that being the fixed point of Ψ (sequential learning-proof) is not only sufficient but also necessary for a cost to be indirect. In Section 3.2, we show that when restricted to Regular cost functions, a cost is indirect if and only if it has a uniformly posterior separable (UPS) representation.

3.1 Indirect Cost and Sequential Learning-Proofness

It is straightforward that an SLP cost is indirect (it is its own indirect cost), because SLP implies $\Phi(C) = \lim_{n\to\infty} \Psi^n(C) = C$. SLP is also clearly a sufficient condition for the following two weaker axioms:

Axiom 1 (Monotone)**.** *$C \in \mathcal{C}$ is Monotone if $\forall \pi \leq_{mps} \pi'$, $C(\pi) \leq C(\pi')$.*

Axiom 2 (Subadditive)**.** *$C \in \mathcal{C}$ is (Sequentially) Subadditive if $\forall$ finite support Π, $C(\mathbb{E}_\Pi[\pi_2]) \leq C(\pi_1) + \mathbb{E}_\Pi[C(\pi_2)]$.*

Monotonicity and subadditivity are weaker than SLP because the violation of either axiom directly provides a feasible two-step contingent plan that improves the direct cost. Nevertheless, Theorem 1 shows that the seemingly non-nested notions are equivalent.

Theorem 1. *For every $C \in \mathcal{C}$,*

$$C \in \mathcal{C}^* \iff C \text{ is SLP} \iff C \text{ is Monotone and Subadditive.}$$

Proof. See Appendix A.1. □

Theorem 1 establishes two equivalent characterizations of indirect cost. The first equivalence is the analogy of the principle of dynamic programming: full sequential optimality can be equivalently verified by checking single deviations (two-step contingent plans). Therefore, the SLP cost functions are justified by two distinct modeling motivations: (i) the desire to model applications where flexible sequential information acquisition is feasible (e.g. modeling a firm conducting A/B testing), and (ii) the desire to write a static information acquisition model that is internally consistent, i.e., robust to potential sequentiality that is not observable by the modeler (e.g. modeling a cognitive task with internal information processing). The second equivalence further decomposes the single deviation into two independent operations. Monotonicity captures the DM's ability to freely dispose of information. Subadditivity captures the decomposition of experiments into two steps. Monotonicity and subadditivity are simple and economically interpretable properties that lead to properties commonly studied/assumed in the literature. Theorem 1 implies that they are exactly the context-free implications of sequential optimality for reduced-form cost functions. A direct corollary of Theorem 1 is a variational characterization of Φ:

Corollary 1.1. *For every $C \in \mathcal{C}$, the indirect cost $\Phi(C) = \max\{C' \in \mathcal{C} \mid C' \leq C$ and C' is SLP$\}$.*

Proof. Theorem 1 implies that $\Phi(C)$ is SLP, and $\Phi(C) \leq C$ by definition. Moreoever, for every SLP $C' \leq C$, we have $C' = \Phi(C') \leq \Phi(C)$ because C' is SLP and Φ is isotone. □

Remark 1. *Theorem 1 has a few notable implications for the indirect cost of information, which we summarize here (details are in Appendix I):*

- *Every Subadditive $C \in \mathcal{C}$ is Convex: for any $\pi, \pi' \in \mathcal{R}$ with the same prior and $\alpha \in [0,1]$, $\alpha C(\pi) + (1-\alpha)C(\pi') \geq C(\alpha\pi + (1-\alpha)\pi')$.[13] Thus, every SLP cost function is Monotone and Convex. This is natural, as these two properties characterize optimal one-shot information acquisition (Caplin and Dean 2015; Oliveira et al. 2017).*

- *The space of indirect costs $\mathcal{C}^*$ is a convex cone and closed under suprema, i.e., for any $\mathcal{D} \subseteq \mathcal{C}^*$, the cost function $C(\pi) := \sup\{C'(\pi) \mid C' \in \mathcal{D}\}$ is also in $\mathcal{C}^*$. Thus, we can generate new indirect costs from conical combinations and suprema of existing ones.*

3.2 Foundations for Uniform Posterior Separability

A special class of SLP cost that is particularly useful in applications is the class of *uniformly posterior separable* (UPS) costs introduced by Caplin, Dean, and Leahy (2022). In this section, we show that UPS costs are fully micro-founded by sequential optimization under an additional local differentiability condition. The following definition of UPS slightly generalizes the standard one by allowing for a partial domain.

[13] This is the implication of Axiom 2 when π_1 is uninformative, and Π simply randomizes over π_2.

Definition 5 (UPS)**.** *$C \in \mathcal{C}$ is Uniformly Posterior Separable (UPS) if there exists a convex function $H : \Delta(\Theta) \to \mathbb{R} \cup \{+\infty\}$ such that $C = C^H_{ups}$, where $C^H_{ups} \in \mathcal{C}$ is defined as*

$$C^H_{ups}(\pi) := \mathbb{E}_\pi[H(q) - H(p_\pi)], \quad \forall \pi \in \Delta(\mathrm{dom}(H)).$$[14]

When $\mathrm{dom}(H) = \Delta(\Theta)$, we say that C is Strongly UPS.[15]

It has been proved that strong UPS is equivalent to the "chain-rule property", or sequential additivity (Zhong (2022, Theorem 3)). We prove in Proposition 1 that the equivalence extends to more general UPS functions as well.

Proposition 1. *For any open convex $W \subseteq \Delta(\Theta)$ and $C \in \mathcal{C}$ with $\mathrm{dom}(C) = \Delta(W) \cup \mathcal{R}^\circ$,*

C is UPS $\iff$ C is (Sequentially) Additive: $\forall\ \Pi \in \Delta(\mathcal{R})$, $C(\mathbb{E}_\Pi[\pi_2]) = C(\pi_1) + \mathbb{E}_\Pi[C(\pi_2)]$.

Proof. See Appendix D.1. □

Proposition 1 hints a close relation between UPS and sequential learning, as the additivity property is a strengthening of the subadditivity property beared by any indirect cost per Theorem 1. This relation is fully explored in Theorem 2.

We begin with an additional technical assumption that enables us to take "first derivatives" of cost functions. Two preliminary definitions are in order. First, for any $\pi \in \mathcal{R}$ and $\alpha \in [0, 1]$, the *α-dilution of π* is the random posterior $\alpha \cdot \pi := \alpha\pi + (1 - \alpha)\delta_{p_\pi}$ obtained by mixing π with the degenerate random posterior (at the prior p_π). Second, a *divergence* is a map $D : \Delta(\Theta)\times\Delta(\Theta) \to \overline{\mathbb{R}}_+$ such that $D(p \mid p) = 0$ for all $p \in \Delta(\Theta)$. If $D(\cdot \mid p)$ is differentiable at q, we denote its gradient by $\nabla_1 D(q \mid p) \in \mathbb{R}^{|\Theta|}$.

Definition 6 (Regular)**.** *$C \in \mathcal{C}$ is Regular if there is a divergence D such that*

$$\lim_{\alpha \searrow 0} \frac{C(\alpha \cdot \pi)}{\alpha} = \mathbb{E}_\pi[D(q \mid p_\pi)] \quad \forall \pi \in \mathrm{dom}(C), \tag{1}$$

and both $D(q \mid p)$ and the gradient $\nabla D_1(q \mid p)$ are (well-defined and) jointly continuous on $\mathrm{relint}(\mathrm{dom}(D))$.[16] *We call any such divergence D a derivative of C and denote $D_C := D$.*

In words, a cost function is Regular if it satisfies two conditions. First, (1) requires that C be *Gateux differentiable* at every $\delta_p \in \mathcal{R}^\circ$ in the direction of any $\pi \in \mathrm{dom}(C) \cap \mathcal{R}(p)$, with the derivative D_C representing the "marginal cost" of increasing the probability of generating π away from $\alpha = 0$. Second, the derivative D_C itself is continuously differentiable

[14] Implicitly, we define $C^H_{\text{ups}}(\pi) := 0$ for all $\pi \in \mathcal{R}^\circ$ and $C^H_{\text{ups}}(\pi) := +\infty$ for all other π.

[15] The original notion of UPS introduced by Caplin and Dean 2013 corresponds to the case where C/H has full domain (renamed to strong UPS by Caplin, Dean, and Leahy 2019). We adopt the terminologies that are consistent with Caplin, Dean, and Leahy 2019.

[16] relint denotes the relative interior.

in the posterior. We interpret Definition 6 primarily as a technical assumption made for tractability. In particular, nearly all models of information cost in the literature satisfy Definition 6 (we highlight one exception in Section 5.2).

Theorem 2. *For any open convex $W \subseteq \Delta^\circ(\Theta)$ and $C \in \mathcal{C}$ with* $\mathrm{dom}(C) = \Delta(W) \cup \mathcal{R}^\circ$,

$$C \text{ is SLP and Regular} \iff C = C^H_{ups} \text{ for some convex } H \in \mathbf{C}^1(W).$$

Proof. See Appendix A.2. □

The proof consists of three main steps. First, we show (in Lemma 2) that every SLP cost is linear in the probability of running an experiment, i.e., satisfy the following property (introduced by Pomatto, Strack, and Tamuz 2023):

Axiom 3 (Dilution Linear). *$C(\alpha \cdot \pi) = \alpha C(\pi)$ for every $\pi \in \mathrm{dom}(C)$ and $\alpha \in [0,1]$.*

Consequently, every Regular SLP cost C equals the expectation of the divergence D_C, i.e., is *Posterior Separable* (Caplin, Dean, and Leahy 2022). The second step of the proof then shows that a Posterior Separable cost is SLP if and only if the divergence that defines it satisfies an "average-case" version of the triangle inequality that defines metric distances. For the third and final step, a calculus exercise shows that the only differentiable divergences satisfying this inequality are Bregman divergences.

Theorem 2 provides an optimality-based foundation for using UPS cost functions in applications. UPS information cost is the only "tractable" (Regular) functional form a modeler can use to write a static model of information acquisition that subsumes sequential optimality or is internally consistent, independent of the specific context of the application. This foundation is orthorgonal to various axiomatic foundations for the UPS model in the literature (Denti 2022; Caplin, Dean, and Leahy 2019; Oliveira 2019; Mensch 2021), which are agnostic about where such reduced-form cost functions "come from".

4 The Sequential Learning Map

In this section, we characterize the sequential learning map. We begin by concluding Example 1 and proving that in the example, the cost of diffusion signals indeed characterizes the true indirect cost.

Example 1 (Gaussian learning–continued)

Recall that we derive the "indirect cost" of information when only diffusion signals are employed: $f'(0) \cdot C_{\text{Wald}}$. We conclude that it is an upper bound for the indirect cost

$\Phi(C) \le f'(0)\cdot C_{\text{Wald}}$. Observe that C_{Wald} is SLP/UPS and for all π_ψ, $f'(0)\cdot C_{\text{Wald}}(\pi_\psi) \le f(\psi)$ due to the convexity of f. Thus, $\Phi(C) \ge \Phi(f'(0)\cdot C_{\text{Wald}}) = f'(0)\cdot C_{\text{Wald}}$. Combining these inequalities yields $\Phi(C) = f'(0)\cdot C_{\text{Wald}}$.

Example 1 hints that the direct cost of diffusion-type experiments, namely, experiments that only shift the posterior belief locally crucially determines the indirect cost. Naturally, two questions arise for the general setting:

1. Can a general indirect cost be bounded by the cost of analogous diffusion signals?
2. Under what condition the diffusion signals are optimal and the bound is tight?

In what follows, we provide the complete answers to the two questions, which guided us toward a general characterization of the sequential learning mapping. To begin, we formally define the cost of diffusion-type signals, i.e. the cost of *incremental evidence*.

4.1 The Cost of Incremental Evidence

A piece of incremental evidence, like an infinitesimal diffusion signal, shifts the posterior belief only locally. Following this logic, we first define a modified direct cost that permits only random posteriors with local supports. For any $C \in \mathcal{C}$, let $\Delta_C := \{p_\pi \in \Delta(\Theta) \mid \exists \pi \in \text{dom}(C)\backslash\mathcal{R}^\circ\}$ denote the set of priors at which some nontrivial experiment is feasible, and let $\Omega(C)$ denote the collection of all open covers of Δ_C.[17] For every $\mathbb{O} \in \Omega(C)$, we define the modified direct cost $C|_{\mathbb{O}} \in \mathcal{C}$ as

$$C|_{\mathbb{O}}(\pi) := \begin{cases} C(\pi), & \text{if } \text{supp}(\pi) \subseteq O \in \mathbb{O} \\ 0, & \text{if } \pi \in \mathcal{R}^\circ \\ +\infty, & \text{otherwise.} \end{cases}$$

This definition modifies C by restricting its domain to experiments that move beliefs within some neighborhood $O \in \mathbb{O}$. We are interested in the limit with *infinitesimally* incremental evidence, i.e. fine coverings.

Definition 7. *The incremental evidence sequential learning map $\Phi_{IE} : \mathcal{C} \to \mathcal{C}$ is defined as*

$$\Phi_{IE}(C) := \sup_{\mathbb{O}\in\Omega(C)} \Phi(C|_{\mathbb{O}}). \tag{IE}$$

Lemma 1. *For every $C \in \mathcal{C}$, $\Phi_{IE}(C)$ is a well-defined indirect cost, i.e., $\Phi_{IE}(C) \in \mathcal{C}^*$.*

Proof. See Appendix E.1. □

[17] That is, every $\mathbb{O} \in \Omega(C)$ is a collection of open sets $O \subseteq \Delta(\Theta)$ such that $\cup_{O\in\mathbb{O}} O \supseteq \Delta_C$.

The supremum in (IE) is approached in the limit where all of the neighborhoods in $\mathbb{O}$ become vanishingly small. Therefore, we interpret this limit as approximating a continuous-time setting in which information is acquired by sampling from a diffusion signal process à la Example 1 and Morris and Strack (2019), but with full control over the signal's drift and volatility.[18]

While (IE) defines the (indirect) cost of incremental evidence in an economically natural way, it is unclear how to actually calculate and analyze $\Phi_{\mathrm{IE}}(C)$. In Example 1, we were able to tractably study the diffusion limit using differential approximations from (stochastic) calculus. By analogy, we next define the *(quadratic) kernel* of an information cost, which facilitates calculations by providing a tractable differential approximation for the cost of incremental evidence in our non-parametric framework.

Definition 8 (Locally Quadratic)**.** *For any $C \in \mathcal{C}$ and $W \subseteq \Delta(\Theta)$, we say that a (symmetric) positive semi-definite matrix-valued function $k : W \to \mathbb{R}^{|\Theta|\times|\Theta|}$ is:*

(i) An upper (quadratic) kernel of C on W if, for every $p \in W$ and $\epsilon > 0$, there exists a $\delta > 0$ such that for all $\pi \in \Delta(B_\delta(p))$,

$$C(\pi) \leq \int_{B_\delta(p)} (q - p_\pi)^\top \left(\frac{1}{2}k(p) + \epsilon I\right)(q - p_\pi)\,\mathrm{d}\pi(q).$$

(ii) A lower (quadratic) kernel of C on W if, for every $p \in W$ and $\epsilon > 0$, there exists a $\delta > 0$ such that for all $\pi \in \mathcal{R}$ with $p_\pi \in B_\delta(p)$,

$$C(\pi) \geq \int_{B_\delta(p)} (q - p_\pi)^\top \left(\frac{1}{2}k(p) - \epsilon I\right)(q - p_\pi)\,\mathrm{d}\pi(q).$$

(iii) A (quadratic) kernel of C on W if it is both a lower kernel and an upper kernel on W.

If C admits a kernel k on W, we say that C is Locally Quadratic on W and denote $k_C := k$. In each case above, we omit the qualifier "on W" when $\Delta^\circ(\Theta) \subseteq W$.

The kernel $k_C(p)$ represents the "local second derivative" of C at the degenerate random posterior $\delta_p \in \mathcal{R}^\circ$, in the direction of any $\pi \in \mathcal{R}$ whose support is contained in an infinitesimal neighborhood of p. Economically, k_C provides a local quadratic approximation for the cost of a piece of incremental evidence, generalizing the "Itô expansion" for the cost of a diffusion signal in Example 1. This approximation represents the "weighted variance" of a random posterior, with k_C determining the relative cost of moving beliefs in

[18] We do not formally prove convergence to the continuous-time limit. However, in Section 6.2 and Appendix H, we show that our main results extend naturally to a continuous-time framework that includes those studied in Morris and Strack (2019), Zhong (2022), and Hébert and Woodford (2023).

different directions. We define the lower and upper kernels separately because, although not every $C \in \mathcal{C}$ is Locally Quadratic, the lower/upper kernels exists very generally.

Intuitively, since the kernel k_C measures the cost of incremental evidence under C, the indirect cost of incremental evidence $\Phi_{\mathrm{IE}}(C)$ should correspond to "integrating" k_C. We establish below in Proposition 2 that, under certain regularity conditions, this intuition can be made precise: for any convex function $H : \Delta(\Theta) \to \mathbb{R}$,

$$k_C = \mathrm{Hess} H \iff \Phi_{\mathrm{IE}}(C) = C^H_{\mathrm{ups}}.$$

We say that k_C is *integrable* if it is the Hessian of some convex H. Economically, integrability corresponds to the "path-independence" property that all incremental learning strategies are equally costly.[19] When there are two states, *every* kernel is integrable (e.g., Morris and Strack 2019). When there are more than two states, this is no longer true and calculating $\Phi_{\mathrm{IE}}(C)$ for non-integrable k_C is much more involved. In what follows, we primarily focus on direct costs with integrable kernels; in such instances, we will often use $\Phi_{\mathrm{IE}}(C)$ and C^H_{ups} as synonyms.

Remark 2. *Without loss of generality, we henceforth normalize all upper/lower kernels $k(p)$ to $\overline{k}(p) := (I - \mathbf{1}p^\top)k(p)(I - p\mathbf{1}^\top)$, so that $\overline{k}(p) \cdot p = \mathbf{0}$. We also normalize the gradient and Hessian of any function $f : \Delta(\Theta) \to \overline{\mathbb{R}}$ to satisfy $\nabla f(p) \cdot p = f(p)$ and $\mathrm{Hess} f(p) \cdot p = \mathbf{0}$. This normalization amounts to extending the quadratic form defined by k and function f from the simplex $\Delta(\Theta)$ to $\mathbb{R}^{|\Theta|}_+$ by homogeneity of degree 1 (HD1) and defining derivatives in the usual way.*

4.2 Bounding the Sequential Learning Map

In this section, we provide the complete answer to Question 1: learning via incremental evidence characterizes a *global* upper bound for $\Phi(C)$ and a *local* lower bound for $\Phi(C)$. For technical reasons, we will occasionally impose the following condition:

Definition 9 (Strongly Positive)**.** *$C \in \mathcal{C}$ is Strongly Positive if there exists an $m > 0$ such that $C(\pi) \geq m \cdot Var(\pi)$ for all $\pi \in \mathcal{R}$, where $Var(\pi) := \mathbb{E}_\pi\left[\|q - p_\pi\|^2\right]$ is the variance of π.*

Theorem 3. *For any $C \in \mathcal{C}$ and $W \subseteq \Delta(\Theta)$, the following holds:*

(i) If W is open and convex, $H \in \mathbf{C}^2(W)$, and $\mathrm{Hess} H$ is an upper kernel of C on W, then

$$\Phi(C)(\pi) \leq C^H_{ups}(\pi) \quad \text{for all } \pi \in \Delta(W).$$

(ii) If C is Strongly Positive and $k \gg_{psd} \mathbf{0}$ is a lower kernel of C on W, then k is also a lower kernel of $\Phi(C)$ on W.[20]

[19] Formally, integrability of k_C implies (via Proposition 1) that $\Phi_{\mathrm{IE}}(C)$ is Additive.

[20] Here, $k \gg_{psd} \mathbf{0}$ denotes that there exists an $m > 0$ such that $q^\top k(p) q \geq m$ for every $p \in W, q \in \Delta(\Theta)$, It is easy to show that every Strongly Positive $C \in \mathcal{C}$ has lower kernels with this property.

Proof. See Appendix B.1. □

Theorem 3(i) shows that (integrable) upper kernels of the direct cost are sufficient to characterize a *global* upper bound for the corresponding indirect cost.[21] This upper bound is powerful because it applies even when non-incremental experiments are extremely costly (or even infeasbile) under the direct cost. To prove this result, we effectively generalize the one-dimensional diffusion signals from Example 1 by explicitly constructing incremental learning strategies that implement any $\pi \in \Delta(W)$. Given such strategies, we can then integrate the kernel to obtain the upper bound.

Next, Theorem 3(ii) shows that $\Phi(C)$ and C have the same lower kernels, despite the fact that $\Phi(C) \preceq C$. In other words, sequential optimization *cannot* reduce the cost of incremental evidence; the lower kernel of the direct cost yields a *local* lower bound for the indirect cost. The intuition is simple if we temporarily disable "free disposal" of information. In that case, since information only accumulates over time, any piece of incremental evidence can only be decomposed into less informative pieces of incremental evidence; since each of these pieces has a direct cost bounded below by the lower kernel, the *indirect* cost of the original piece must also bounded below by the same lower kernel. The formal proof deals with the more involved "free disposal" case.

Notably, the lower kernel of the direct cost is *not* sufficient to obtain a *global* lower bound for the indirect cost, because the direct cost of non-incremental experiments might be arbitrarily low. However, it does yield a global lower bound for the indirect cost if we restrict attention to incremental learning strategies, i.e., consider $\Phi_{\mathrm{IE}}(C)$ rather than $\Phi(C)$. In fact, the kernel of C fully characterizes $\Phi_{\mathrm{IE}}(C)$.

Proposition 2. *For any open convex set $W \subseteq \Delta(\Theta)$, strongly convex $H \in \mathbf{C}^2(W)$, and direct cost $C \in \mathcal{C}$ with $\mathrm{dom}(C) \subseteq \Delta(W) \cup \mathcal{R}^\circ$, the following properties hold:*

(i) If $\mathrm{Hess}H$ is an upper kernel of C on W, then $\Phi_{IE}(C) \preceq C^H_{ups}$.

(ii) If $\mathrm{Hess}H$ is a lower kernel of C on W, then $\Phi_{IE}(C) \succeq C^H_{ups}$.

(iii) If C is Locally Quadratic on W, then $k_C = \mathrm{Hess}H \iff \Phi_{IE}(C) = C^H_{ups}$.

Proof. See Appendix E.2. □

We note that Proposition 2(iii) can be viewed as an extension of Morris and Strack 2019 to settings with a general state space, general direct cost of incremental evidence, and optimization over incremental learning strategies.

[21] Absent integrability, we have the analogous (but trivial) upper bound $\Phi(C) \preceq \Phi_{\mathrm{IE}}(C)$. Moreover, since $\Phi(C) \preceq C$, the upper kernels of C must also be upper kernels of $\Phi(C)$, so the upper kernels of C always provide a "local" upper bound for $\Phi(C)$.

4.3 Determining the Sequential Learning Map

In this section, we provide the complete answer to Question 2: the bounds we obtain from Theorem 3 are tight if and only if the direct cost satisfies the following property.

Axiom 4 (FLIEs)**.** *$C \in \mathcal{C}$ favors learning via incremental evidence (FLIEs) if $C \geq \Phi_{IE}(C)$.*

A cost function FLIEs if the total cost of producing an experiment through incremental learning is weakly lower than the cost of acquiring the experiment directly.

Theorem 4. *For any open convex set $W \subseteq \Delta(\Theta)$, strongly convex $H \in \mathbf{C}^2(W)$, and direct cost $C \in \mathcal{C}$ that is Locally Quadratic on W and satisfies $\mathrm{dom}(C) \subseteq \Delta(W) \cup \mathcal{R}^\circ$,*

$$C \text{ FLIEs and } k_C = \mathrm{Hess}H \iff \Phi(C) = C^H_{ups}.$$

Proof. See Appendix B.2. □

We interpret Theorem 4 as offering two characterizations. First, it characterizes the domain of direct costs whose corresponding indirect costs are Regular: a Locally Quadratic direct cost generates a Regular indirect cost *if and only if* the former FLIEs and has an integrable kernel. Second, it fully determines the map Φ for the codomain of Regular indirect costs, which are pinned down by the kernels of their direct costs.

The first characterization suggests a novel economic foundation for the UPS model. An immediate implication of Theorem 4 is that the indirect cost of information is UPS *if and only if* incremental learning is globally optimal and all incremental learning strategies are equally costly. [22] This can help delineate the set of applications in which the UPS model is economically reasonable. The following examples illustrate:

- *Cognitive costs of attention:* Following Sims (2003), UPS costs are often interpreted as describing humans' cognitive costs of processing available information (e.g., reading a newspaper).[23] In psychology and neuroscience, a leading theory of human attention is the *drift-diffusion model (DDM)*, which models the cognitive process as the sequential sampling of diffusion signals (Ratcliff 1978; Ratcliff and McKoon 2008; Fudenberg, Strack, and Strzalecki 2018). Theorem 4 suggests a bridge between these literatures: the indirect cost of attention is UPS if (and only if) DDM-style sampling is the optimal cognitive process (see also Hébert and Woodford 2023).

[22] If C FLIEs, then $\Phi(C) \geq [\Phi \circ \Phi_{\mathrm{IE}}](C) \geq \Phi_{\mathrm{IE}}(C)$ because Φ is isotone and $\Phi_{\mathrm{IE}}(C) \in \mathcal{C}^*$. If $\Phi(C) \geq \Phi_{\mathrm{IE}}(C)$, then C FLIEs because $C \geq \Phi(C)$. Since $\Phi(C) \leq \Phi_{\mathrm{IE}}(C)$ by construction, the claimed equivalence follows.

[23] See, e.g., Sims (2010), Matějka and McKay (2015), Hébert and Woodford (2021), and Caplin, Dean, and Leahy (2022). Experimental evidence on perception tasks provides support for this interpretation (Dean and Neligh 2023; Dewan and Neligh 2020; Denti 2022).

- *Statistical sampling:* Consider the problem of testing a hypothesis by drawing samples from a large population, e.g., political polling, market research, or clinical trials (Wald 1945; Arrow, D. Blackwell, and Girshick 1949). These applications almost perfectly fit the setting of Example 1: FLIEs is satisfied and the indirect cost is the Wald cost.

- *Research & development:* If C FLIEs, it is never strictly optimal to generate discrete belief jumps. This property may be inappropriate for modeling industrial R&D or the process of scientific research, as in these applications, learning often occurs through infrequent but discrete "breakthroughs" that can be modeled as jumps of a Poisson signal process (Che and Mierendorff 2019; Zhong 2022).

The second characterization suggests a methodological tool for analyzing the sequential learning map. Specifically, Theorem 4 allows one to calculate both (i) the indirect cost function generated by a given direct cost and (ii) the set of direct costs that generate a given indirect cost. We depict this methodology in Figure 3, where the solid arrows represent the calculation of Φ and the dotted arrows represent the calculation of Φ^{-1}. We explore several applications of this tool in Section 5 below.

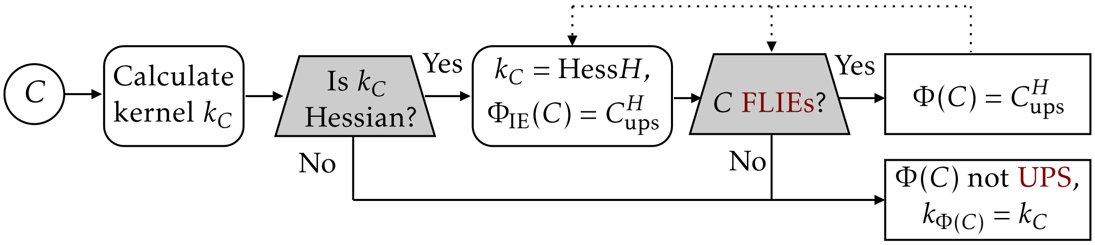
C
Calculate kernel k_C
Is k_C Hessian?
Yes
No
$k_C = \mathrm{Hess}H$, $\Phi_{\mathrm{IE}}(C) = C^H_{\mathrm{ups}}$
C FLIEs?
Yes
No
$\Phi(C) = C^H_{\mathrm{ups}}$
$\Phi(C)$ not UPS, $k_{\Phi(C)} = k_C$

Figure 3: Flow diagram for calculating direct/indirect cost.

Remark 3. *The hypothesis that C is Locally Quadratic is "nearly" without loss of generality for Theorem 4 in two respects: (i) minor variants of both directions hold without it, and (ii) any $C \in \mathcal{C}$ for which $\Phi(C) = C^H_{ups}$ can be "locally smoothed" into a Locally Quadratic $C' \in \mathcal{C}$ for which both directions hold exactly. See Corollary 6.1 in Appendix E.4 for formal details.*

5 Applications: Reduced-Form Information Costs

In this section, we operationalize our framework by studying specific (classes of) "reduced-form" information costs through the lens of optimization. Section 5.1 begins with a few illustrative examples. In Section 5.2, we introduce two new indirect cost functions that satisfy two properties that are important in applications: prior invariance and constant marginal cost. In Sections 5.3 and 5.4, we fully characterize the relationship between indirect cost and these properties, obtaining and resolving a trilemma.

5.1 Illustrative Examples

We begin by demonstrating how our characterization of Φ^{-1} can clarify the implicit economic assumptions that are imposed on the direct cost when one adopts a particular functional form for the indirect cost.

Mutual Information. The benchmark rational inattention model (Sims 2003) is based on the *Mutual Information* cost function (Shannon 1948), which we denote by C_{MI}. In our terminology, C_{MI} is the Strong UPS cost function with potential and kernel

$$H_{\mathrm{MI}}(p) = \sum_{\theta \in \Theta} p(\theta)\log(p(\theta)) \quad \text{and} \quad k_{\mathrm{MI}}(p) = \mathrm{Diag}(p)^{-1} - \mathbf{1}\mathbf{1}^{\top}, \tag{MI}$$

where H_{MI} is *(negative) Shannon's entropy* and k_{MI} is the *Fisher information matrix*. By Theorem 4, a direct cost $C \in \mathcal{C}$ generates the indirect cost $\Phi(C) = C_{\mathrm{MI}}$ if and only if $C \geq C_{\mathrm{MI}}$ and $k_C = k_{\mathrm{MI}}$.

The following example demonstrates how our characterization of Φ can produce new indirect costs (and provide new foundations for extant ones).

Aggregating Technologies. We posit a simple scheme to model acquiring information from multiple "sources" that are informative about different "aspects" of the state. This is the case, for instance, when sampling from multiple heterogeneous subpopulations, obtaining news from differentially biased media outlets, or learning about the value of a financial portfolio via both fundamental research about the assets and technical analysis of market prices.[24]

Consider a finite collection of information costs $\{C^i\} \in \mathcal{C}$, each representing the cost of learning from one source. The direct cost of information is given by $C(\pi) := f(C^1(\pi), \ldots, C^I(\pi))$, where $f : \mathbb{R}^I_+ \to \mathbb{R}_+$ is non-decreasing, satisfies $f(\mathbf{0}) = 0$, and is continuously differentiable. We interpret f as a "production function" that aggregates the source-specific costs C^i, and which can encode general complementarities and substitutabilities among the sources. We assume that each C^i is Locally Quadratic with kernel $k_{C^i} = \mathrm{Hess}H^i$ for some strongly convex $H^i \in \mathbf{C}^2(\Delta(\Theta))$. It is easy to verify that $k_C = \sum \nabla_i f(\mathbf{0})\mathrm{Hess}H^i$. Then, Theorems 3 and 4 immediately implies:

Corollary 4.1. *Let $H := \sum_{i \in I} \nabla_i f(\mathbf{0}) H^i$. Then,*

$$\Phi(C) \preceq \Phi_{IE}(C) = C^H_{ups} \quad \text{and} \quad k_{\Phi(C)} = \mathrm{Hess}H = \sum_{i \in I} \nabla_i f(\mathbf{0})\,\mathrm{Hess}H^i.$$

[24] For examples in the literature, see Myatt and Wallace (2012), Liang, Mu, and Syrgkanis (2022), Angeletos and Sastry (2024), and Hébert and La'O (2023).

Moreover, if each C^i FLIEs and f is subdifferentiable at $\mathbf{0}$,[25] *then C FLIEs and $\Phi(C) = C^H_{ups}$.*

Corollary 4.1 provides simple conditions under which the indirect cost $\Phi(C)$ is Regular and characterizes its functional form. Notably, optimization "smooths away" all nonlinearities in the production function, so $\Phi(C)$ is simply a weighted sum of the source-specific indirect costs $\Phi(C^i)$. Two special cases are of particular interest:

Example 2 (Neighborhood-Based Costs)

Each $i \in I$ represents a *neighborhood* of states $N_i \subseteq \Theta$, where $\{N_i\}_{i\in I}$ covers Θ. For each $i \in I$, let $H^i(q) := q(N_i)G^i(q(\cdot \mid N_i))$ for some strongly convex $G^i : \Delta(N_i) \to \mathbb{R}$. Then the indirect cost C^H_{ups} in Corollary 4.1 is the *neighborhood-based cost* of Hébert and Woodford (2021), where $f_i'(\mathbf{0})$ is the marginal cost of learning within neighborhood N_i.

Example 3 (Pairwise Separable Costs)

Let $I = \Theta \times \Theta$, so that each $i = (\theta, \theta')$ is an ordered pair of states. For each such pair, let $H^{(\theta,\theta')}(p) := p(\theta)\phi\left(\frac{p(\theta')}{p(\theta)}\right)$ for some (strongly) convex, $\mathbf{C}^2$-smooth, $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\phi(1) = 0$. Letting $\gamma_{\theta,\theta'} := f'_{(\theta,\theta')}(\mathbf{0})$, the indirect cost in Corollary 4.1 is C^H_{ups}, where

$$H(p) = \sum_\theta p(\theta) \sum_{\theta'} \gamma_{\theta,\theta'} \phi\left(\frac{p(\theta')}{p(\theta)}\right)$$

represents the expectation over true states θ of the cost of distinguishing between other states $\theta' \neq \theta$. These indirect costs are finite-state variants of the *pairwise-separable costs* of Morris and Yang (2019). The case where $\phi(t) = -\log(t)$ is of particular interest, as it yields the Total Information cost function introduced in Section 5.2.2 below.

5.2 Normative Axioms

In this section, we introduce several novel SLP cost functions that satisfy context-specific axioms that have been extensively studied in the literature. For this purpose, it will be convenient to move between the formulations of information as Blackwell experiments $\sigma : \Theta \to \Delta(S)$ and random posteriors $\pi \in \mathcal{R}$. To this end, let $h_B : (\sigma, p) \mapsto \pi$ denote the Bayesian map that takes experiment-prior pairs to their induced random posteriors.

5.2.1 Prior Invariance

For certain applications, it is natural to assume that the cost of information does not vary with the DM's prior beliefs. This restriction is particularly appropriate when modeling the *physical* cost of conducting experiments (e.g., sampling from a population, conducting R&D) or the *pecuniary* cost of purchasing information (e.g., in a market for data). We formalize the prior invariance restriction as follows:

[25] Formally, f is *subdifferentiable at* $\mathbf{0}$ if $f(\boldsymbol{x}) \geq \nabla f(\mathbf{0}) \cdot \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^I_+$. This property is implied by, but much weaker than, the assumption that f is convex.

Axiom 5 (Prior Invariant). *$C \in \mathcal{C}$ is (Weakly) Prior Invariant if $C(h_B(\sigma,p)) = C(h_B(\sigma,p'))$ for any experiment σ and priors $p,p' \in \Delta(\Theta)$ with the same support,* $\mathrm{supp}(p) = \mathrm{supp}(p')$.[26]

Despite its intuitive appeal, prior invariance is violated by many functional forms that are commonly used in applications of flexible information acquisition. For instance, it is known that any cost function in the Strong UPS class (which includes mutual information) that is Prior Invariant must be identically zero. Nevertheless, prior invariance is consistent with SLP once we relax Regularity.

Definition 10 (MLR). *The Minimal Likelihood Ratio (MLR) cost function C_{MLR} is defined as follows. $\forall \pi \in \mathcal{R}$,*

$$C_{MLR}(\pi) = \mathbb{E}_\pi\left[D_{MLR}(q \mid p_\pi)\right] \quad \textit{for} \quad D_{MLR}(q \mid p) := 1 - \min_{\theta \in \mathrm{supp}(p)} \frac{q(\theta)}{p(\theta)}.$$

Equivalently, for all $\sigma \in \mathcal{E}$ and $p \in \Delta(\Theta)$,

$$C_{MLR}(h_B(\sigma,p)) = 1 - \int_S \bigwedge_{\theta \in \mathrm{supp}(p)} \sigma(\mathrm{d}s \mid \theta).^{27}$$

By construction, the MLR cost is Prior Invariant as it only depends on σ and the support of p. It is also SLP because D_{MLR} is (i) convex in q and (ii) a quasi-metric and, in particular, satisfies the triangle inequality. The former property implies that C is Monotone, while the latter implies that C is Subadditive. [28]

Example 4 (EAD)

In the special case of binary states, $\Theta = \{0,1\}$ and full support prior , the MLR cost reduces to what we call the *Expected Absolute Difference (EAD)* cost function:

$$C_{\mathrm{EAD}}(\pi) := \frac{1}{2p_\pi(1-p_\pi)} \mathbb{E}_\pi\left[|q - p_\pi|\right],$$

where $q, p_\pi \in [0,1]$ are probabilities that $\theta = 1$. To see why they are equivalent, observe that $D_{MLR}(q|p)$ differs from $\frac{|q(1)-p(1)|}{2p(1)p(0)}$ by $\frac{0.5-q(1)}{p(1)p(0)}$, an affine function of q. Equivalently, the EAD cost is the *total variation distance* between the state-contingent signal distributions

[26] This definition is slightly weaker than requiring complete prior-independence. This extra degree of flexibility is a technical artifact of our belief-based formulation. In Section 6.1 below, we illustrate that full prior-independence is easily incorporated into an enrichment of our framework.

[27] The meet (minimum) of two Radon measures $\mu \wedge \nu$ is $\mu - (\mu - \nu)^+$ per the Hahn decomposition theorem. The finite meet is defined accordingly. When S is finite, $C_{\mathrm{MLR}}(h_B(\sigma,p)) = 1 - \sum_{s \in S} \min_{\theta \in \mathrm{supp}(p)} \sigma(s|\theta)$.

[28] A *quasi-metric* is a map $d : \Delta(\Theta) \times \Delta(\Theta) \to \mathbb{R}_+$ that satisfies all the properties of a metric except for symmetry: (i) $d(q \mid p) = 0$ iff $p = q$ and (ii) $d(q \mid p) + d(q' \mid q) \geq d(q' \mid p)$ for all $p, q, q' \in \Delta(\Theta)$. It can be verified that any expected quasi-metric divergence is SLP.

$\sigma(\cdot \mid \theta) \in \Delta(S)$, i.e.,

$$C_{\text{EAD}}(h_B(\sigma, p)) = \frac{1}{2}\|\sigma(\cdot \mid 1) - \sigma(\cdot \mid 0)\|_{\text{TV}} := \sup_{\text{Borel } E \subseteq S} |\sigma(E \mid 1) - \sigma(E \mid 0)|.^{29}$$

The EAD cost has been used in Che and Mierendorff (2019, Section V.A) and Zhong (2022, Proposition 2) to model the Prior Invariant flow cost of Poisson signals in optimal stopping problems with continuous time and discounting.

5.2.2 Constant Marginal Cost

For some applications, it is natural to assume that the cost of drawing (conditionally) independent samples from a population is linear in sample size. The following axiom, first introduced by Pomatto, Strack, and Tamuz (2023), formalizes this intuitive notion in a non-parametric way.

Axiom 6 (CMC). *$C \in \mathcal{C}$ exhibits Constant Marginal Cost (CMC) if for arbitrary σ_1, σ_2,*

$$C(h_B(\sigma_1 \otimes \sigma_2, p)) = C(h_B(\sigma_1, p)) + C(h_B(\sigma_2, p)) \quad \forall p \in \Delta(\Theta),$$

where $[\sigma_1 \otimes \sigma_2](E_1 \times E_2 \mid \theta) := \sigma_1(E_1 \mid \theta)\sigma_2(E_2 \mid \theta)$ for all (Borel) $E_1, E_2 \subseteq S$ and $\theta \in \Theta$.

CMC specifies that the cost of running any two experiments σ_1 and σ_2 together equals the total cost of running them separately and *simultaneously*, i.e., (i) under the *same prior* and (ii) *without conditioning* the choice of σ_2 on the signal s_1 generated by σ_1. Therefore, we interpret CMC as a "static" additivity, in contrast to the sequential additivity property that characterizes UPS costs (Proposition 1). It is natural to ask whether these distinct notions of additivity are compatible. We introduce a new class of cost functions that provide an affirmative answer:

Definition 11 (Total Information). *We call $C_{TI} \in \mathcal{C}$ a Total Information cost function if there exist non-negative coefficients $(\gamma_{\theta,\theta'})_{\theta,\theta' \in \Theta}$ such that, for all $\pi \in \Delta(\Delta^\circ(\Theta))$,*

$$C_{TI}(\pi) = C_{ups}^{H_{TI}}(\pi) \quad \text{for} \quad H_{TI}(p) := \sum_{\theta,\theta' \in \Theta} \gamma_{\theta,\theta'} p(\theta) \log\left(\frac{p(\theta)}{p(\theta')}\right).$$

Equivalently, for arbitrary σ and $p \in \Delta^\circ(\Theta)$ such that $h_B(\sigma, p) \in \Delta(\Delta^\circ(\Theta))$,

$$C_{TI}(h_B(\sigma, p)) = \sum_{\theta \in \Theta} p(\theta) \sum_{\theta' \in \Theta} \gamma_{\theta,\theta'} D_{KL}(\sigma(\cdot \mid \theta) \mid \sigma(\cdot \mid \theta')).$$

Total Information is of special interest for three reasons. First, it both is UPS and exhibits CMC, where the latter can be seen from either (i) the additivity of KL divergence for independent random variables or (ii) the fact that C_{TI} is linear in the prior for each fixed experiment. We interpret the conjunction of these properties as a strong form of

[29] When S is finite, $C_{\text{EAD}}(h_B(\sigma, p)) = \frac{1}{2}\sum_{s \in S} |\sigma(s \mid 1) - \sigma(s \mid 0)|$.

"process invariance," whereby the overall expected cost of replicating a target experiment is invariant to the way in which the acquired information is decomposed within or across rounds. In other words, costs depend only on the *total amount of information* that is generated, not on the process through which it is acquired.

Second, two limiting cases of Total Information encompass important alternatives to the Mutual Information cost that have been proposed in the literature:

Example 5 (Wald)

In the special case of binary states, $\Theta = \{0,1\}$, Total Information with symmetric coefficients ($\gamma_{0,1} = \gamma_{1,0}$) reduces to the *Wald cost* from Morris and Strack (2019), which we have already seen in Example 1. Total Information can be viewed as the natural generalization of the Wald cost to multiple states and asymmetric coefficients.

Example 6 (Fisher Information)

Hébert and Woodford (2021) introduce the *Fisher Information cost* function as a way to model "perceptual distance" in continuous-state settings, where the state space $\widehat{\Theta} = (\underline{\theta}, \overline{\theta}) \subseteq \mathbb{R}$ is an interval and the DM can acquire experiments $\widehat{\sigma} : \widehat{\Theta} \to \Delta(S)$ satisfying certain technical conditions. The Fisher Information cost of $\widehat{\sigma}$ is given by the expectation (under the DM's prior) of the function $\widehat{\theta} \mapsto \mathcal{I}(\widehat{\sigma} \mid \widehat{\theta})$, where $\mathcal{I}(\widehat{\sigma} \mid \widehat{\theta})$ is the Fisher Information of $\widehat{\sigma}$ in state $\widehat{\theta}$ (see, e.g., Cover and Thomas (2006) for a textbook treatment).

Total Information can be viewed as a finite-state generalization of the Fisher Information cost, for two reasons. First, CMC is a defining property of the Fisher Information cost. Second, using the standard fact that $D_{\mathrm{KL}}\big(\widehat{\sigma}(\cdot \mid \widehat{\theta}) \mid \widehat{\sigma}(\cdot \mid \widehat{\theta}')\big) = \frac{(\widehat{\theta}-\widehat{\theta}')^2}{2}\mathcal{I}(\widehat{\sigma} \mid \widehat{\theta}) + o((\widehat{\theta} - \widehat{\theta}')^2$, we can approximate the Fisher Information cost using Total Information on a discretized state space $\Theta = \{\theta_1, \ldots, \theta_{|\Theta|}\}$ with $\underline{\theta} = \theta_1 < \theta_2 < \cdots < \theta_n = \overline{\theta}$ and the specific coefficients $\gamma_{\theta_i,\theta_j} = \mathbf{1}(j = i+1)/(\theta_i - \theta_j)^2$, which are only nonzero for "adjacent" states.

Third, Total Information is closely related to the *Log-Likelihood Ratio (LLR)* cost functions of Pomatto, Strack, and Tamuz (2023). In our notation, $C_{\mathrm{LLR}} \in \mathcal{C}$ is an LLR cost if there exist non-negative coefficients $(\beta_{\theta,\theta'})_{\theta,\theta' \in \Theta}$ such that, for all $\sigma \in \mathcal{E}$ and $p \in \Delta(\Theta)$,

$$C_{\mathrm{LLR}}(h_B(\sigma,p)) = \sum_{\theta,\theta' \in \mathrm{supp}(p)} \beta_{\theta,\theta'} D_{\mathrm{KL}}(\sigma(\cdot \mid \theta) \mid \sigma(\cdot \mid \theta')). \tag{LLR}$$

Total Information costs can be viewed as expectations over collections of LLR costs, one for each possible state.[30] Moreover, for any *fixed* prior $p \in \Delta(\Theta)$, the Total Information cost with coefficients $\gamma_{\theta,\theta'}$ equals the LLR cost with coefficients $\beta_{\theta,\theta'} \equiv p(\theta)\gamma_{\theta,\theta'}$.

[30] Formally, the Total Information cost with coefficients $\gamma_{\theta,\theta'}$ can be written as $C_{\mathrm{TI}}(h_B(\sigma,p)) \equiv \sum_{\theta\in\Theta} p(\theta) C^{(\theta)}_{\mathrm{LLR}}(h_B(\sigma,p))$, where each $C^{(\theta)}_{\mathrm{LLR}}$ is the LLR cost with coefficients $\beta^{(\theta)}_{\tau,\theta'} \equiv \mathbf{1}(\tau = \theta)\gamma_{\theta,\theta'}$.

5.3 An Information Cost Trilemma

We have seen that these three important properties SLP, prior invariance, and CMC have nontrivial intersections. This suggests modeling tradeoffs. In this section, we fully characterize these tradeoffs by establishing a trilemma among the three properties.

For technical reasons, we first introduce two assumptions. First, we restrict attention to $C \in \mathcal{C}$ that has *rich domain*: $\text{dom}(C) \supseteq \Delta(\Delta^{\circ}(\Theta))$. Second, we follow Pomatto, Strack, and Tamuz (2023, Axiom 4) by augmenting CMC with a mild but complex continuity condition, which we refer to as *PST-continuity*. For simplicity, we embed it in the following definition and relegate the formal details to Appendix I.[31]

Definition 12 (CMC$^{\copyright}$)**.** *$C \in \mathcal{C}$ is CMC$^{\copyright}$ if it is CMC and PST-continuous (Appendix I).*

We can now formally state the trilemma, which is depicted in Figure 2. Point (ii) restates Theorem 1 in Pomatto, Strack, and Tamuz (2023) and is included here only for completeness.

Theorem 5. *For any non-zero $C \in \mathcal{C}$ with rich domain, the following properties hold:*

(i) C is SLP and CMC$^{\copyright}$ $\iff$ C is a Total Information cost.

(ii) C is Prior Invariant, CMC$^{\copyright}$, and Dilution Linear $\iff$ C is an LLR cost.

(iii) C is SLP and Prior Invariant $\begin{array}{l} \longrightarrow C \text{ is neither UPS nor CMC.} \\ \longleftarrow C \text{ is an MLR cost.} \end{array}$

Consequently, no such C is SLP, Prior Invariant, and CMC.

Proof. See Appendix I. □

We interpret Theorem 5(i) primarily as providing strong support for Total Information in applications where both SLP and CMC are natural assumptions. A secondary lesson is that CMC is typically not preserved under optimization. We further show in Appendix I that essentially *no* prior-dependent ("Bayesian") LLR direct cost—aside from Total Information itself—yields an indirect cost satisfying CMC$^{\copyright}$.[32] This suggests that, when modeling the "reduced form" (SLP) cost of information, CMC may be best understood as an *emergent* property of the optimization process (as in Example 1).

Next, Theorem 5(iii) identifies a strong tension between SLP and Prior Invariance. First, the most prominent subclass of SLP costs—the UPS costs—does not intersect the

[31] Informally, Axiom 4 in Pomatto, Strack, and Tamuz (2023) imposes a very permissive form of continuity on $C(h_B(\cdot, p)) : \mathcal{E} \to \overline{\mathbb{R}}_+$ for each fixed $p \in \Delta(\Theta)$. What we call PST-continuity combines this with continuity of $C(h_B(\sigma, \cdot)) : \Delta(\Theta) \to \overline{\mathbb{R}}_+$ along sequences of priors with common support, for each fixed $\sigma \in \mathcal{E}$.

[32] Such costs are Strongly Positive whenever the coefficients $\beta_{\theta,\theta'}(p)$ are bounded away from 0.

class of Prior Invariant costs. By Theorem 2, this implies that any Prior Invariant SLP cost must be non-Regular (e.g. the MLR cost). Second, a prominent sublcass of Prior Invariant cost—those exhibiting CMC, which includes LLR costs and the more general Renyi divergence costs from Mu et al. (2021)—does not intersect the class of SLP costs. This implies that no cost function satisfies all three properties in the trilemma.

It is perhaps surprising, then, that the tension between SLP and Prior Invariance can be reconciled at all. We illustrate this possibility using the MLR cost because of its convenient functional form.

5.4 A Resolution: Sequential Prior Invariance

We conclude from Theorem 5 that Prior Invariance is typically an overly restrictive assumption for modeling "reduced form" (SLP) information costs. In particular, because SLP costs arise from sequential *expected* cost-minimization, they "should" *endogenously* depend on prior beliefs—even if the underlying direct cost function is Prior Invariant (as in Example 1). This motivates the following novel class of cost functions:

Definition 13 (SPI)**.** *$C \in \mathcal{C}$ is Sequentially Prior Invariant (SPI) if $C = \Phi(C')$ for some Prior Invariant $C' \in \mathcal{C}$.*

We view SPI cost functions as the natural modeling tool in most applications where the literature has advocated for using Prior Invariant costs. Indeed, many real-world settings in which information costs are physical or pecuniary (e.g., clinical trials) feature at least some degree of flexible sequential learning (e.g., the design of multi-stage trials).

The class of SPI costs clearly includes all Prior Invariant SLP costs (e.g., the MLR cost). It also includes the Wald cost, which is UPS and CMC$^{©}$ but not Prior Invariant (Example 1). Thus, relaxing Prior Invariance to SPI delivers at least one resolution to the information cost trilemma. In fact, the Wald cost provides the *only* such resolution:

Theorem 6. *For any Strongly Positive $C \in \mathcal{C}$ with rich domain,*

$$C \textit{ is SPI and CMC}^{©} \iff C \textit{ is SPI, UPS, and Locally Quadratic} \iff |\Theta| = 2 \textit{ and } C = \gamma \cdot C_{Wald}.$$

Proof. See Appendix C.1. □

Theorem 6 offers two characterizations (see Figure 4). First, it *uniquely* resolves the information cost trilemma: the Wald cost is the *only* SPI and CMC$^{©}$ cost function. Second, it *uniquely* resolves the tension between Prior Invariance and UPS: the Wald cost, again, is the *only* SPI and (smooth) UPS cost function.

It is instructive to sketch the proof of Theorem 6. First, as is suggested by Example 1, the Wald cost is SPI. [33] The more subtle direction is to show that C_{Wald} is the *only* SPI and (smooth) UPS cost, which hinges on a novel "local" implication of (Sequential) Prior Invariance: we show in Lemma 6 that for any Prior Invariant cost C, the "statistic kernel" (the kernel represented in the space of statistical experiments)

$$\kappa_C(p) := \text{diag}(p)k_C(p)\text{diag}(p)$$

is a constant matrix. Then, so does every SPI cost function due to kernel invariance (Theorem 3). Basic calculus implies that the Wald kernel is the only kernel that is both consistent with a constant κ_C and integrable, which directly implies Theorem 6.

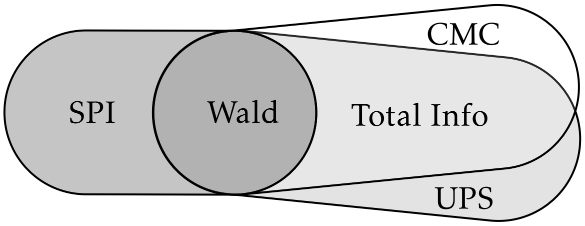

Figure 4: Resolution of the trilemma

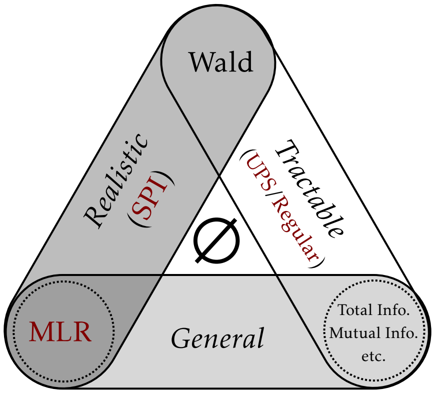

Figure 5: Modeler's trilemma

We conclude by discussing the implications of Theorem 6. The resolution of the information cost trilemma gives rise to a new "modeler's trilemma" (Figure 5): there is an inherent modeling tradeoff among realism (SPI), tractability (UPS/Regularity) and applicability (general state space). The tension between realism and tractability can be reconciled by adopting the Wald cost, but only when restricted to specific applications in binary state settings, e.g., hypothesis testing. To obtain greater applicability, one necessarily needs to sacrifice either tractability (e.g., adopting the irregular MLR cost) or some of the realism (e.g., adopting the non-SPI Total Information or Mutual Information costs).

6 Extensions and Discussion

6.1 Beyond the Belief-Based Framework

For convenience, we have developed our framework using the standard *belief-based approach* in which information structures are modeled as random posteriors. This approach makes two implicit assumptions on the cost of information: (i) all experiments

[33] The formal proof identifies a different prior invariant direct cost for C_{Wald} that is $\max\{D_{\text{KL}}(\sigma_0 \mid \sigma_1), D_{\text{KL}}(\sigma_1 \mid \sigma_0)\}$. In fact, this exercise characterizes the full set of Prior Invariant (and Locally Quadratic) direct costs that generate the Wald cost.

that generate the same random posterior are assigned the same cost, and (ii) experiments acquired in different rounds of a sequential procedure generate signals with conditionally independent noise. In this section, we discuss the extent to which each assumption is economically restrictive, how each assumption can be removed or relaxed, and the nuances that arise from doing so. Formal details are in Appendix G.

From Beliefs to Experiments. The first assumption is irrelevant if the DM only cares about the random posterior generated by the target experiment—e.g., if he uses the information to solve a canonical decision problem—because the optimization process implicitly picks the cheapest experiment to induce each random posterior. However, it may be consequential if the DM cares about the whole "informational content" of the target experiment *and* his prior belief has partial support. For example, when the prior $p = \delta_\theta$ is concentrated on a single state θ, *all* experiments induce the trivial random posterior $\delta_p \in \mathcal{R}^\circ$ and thus are assigned *zero cost*, even if they are very informative about the other states $\theta' \neq \theta$. This feature of the belief-based approach can create subtleties in applications to *costly monitoring* in games, where the state represents another agent's action, the prior represents that agent's mixed strategy, and the DM needs to monitor for off-path deviations (see, e.g., Ravid (2020) and Denti, Marinacci, and Rustichini (2022)).[34]

In Appendix G.1, we address this limitation by developing a richer *experiment-based framework* in which (i) cost functions are defined directly on pairs of statistical experiments and prior beliefs, and (ii) the MPS constraint in the definition of Ψ is replaced by a Blackwell dominance constraint (which is more restrictive at partial-support priors). We construct a scheme for mapping between the belief- and experiment-based frameworks, which reveals that they are *equivalent*—and hence our results directly extend—whenever the DM's *initial* prior belief has full support. While the experiment-based framework permits *strictly richer* behavior of the sequential learning map under *partial-support* priors, we show that Theorem 1 extends in the natural way. This provides a foundation for particular experiment-based cost functions, such as:

- **(Experiment-based) Total Information**: For all experiments $\sigma : \Theta \to \Delta(S)$ and priors $p \in \Delta(\Theta)$,
$$K_{\text{TI}}(\sigma, p) := \sum_{\theta,\theta' \in \Theta} p(\theta)\gamma_{\theta,\theta'} D_{\text{KL}}(\sigma(\cdot \mid \theta) \mid \sigma(\cdot \mid \theta')).$$

Notably, K_{TI} is Additive (hence, SLP) with respect to experiments (see Appendix G.1).

[34] More broadly, in the spirit of classical statistics, it may be conceptually desirable to decouple the "objective information" generated by an experiment from the "subjective uncertainty" embodied by the DM's prior (e.g., D. A. Blackwell (1951) does not even presuppose the existence of prior beliefs).

- **(Experiment-based) MLR Cost:** For all experiments $\sigma : \Theta \to \Delta(S)$ and priors $p \in \Delta(\Theta)$,
$$K_{\text{MLR}}(\sigma, p) := 1 - \int_S \bigwedge_{\theta \in \Theta} \sigma(\mathrm{d}s \mid \theta).$$
Notably, K_{MLR} satisfies a stronger notion of *Strong* Prior Invariance: $K(\cdot, p) \equiv K(\cdot, p')$ for *all* priors $p, p' \in \Delta(\Theta)$.

The functional forms K_{TI} and K_{MLR} can be used as reduced-form cost functions in applications to costly monitoring. For instance, Wong (2023) and Georgiadis and Szentes 2020 apply the experiment-based SLP cost functions, and K_{TI} in particular, in this context.

Correlated Signals. The second assumption is more substantial, as conditionally correlated noises may lead to the DM update his beliefs differently when observing statistical experiments that induce the same random posterior.[35] As a result, a given sequence of random posteriors does not uniquely pin down the information that is acquired. Nevertheless, conditionally dependent noises are readily nested by our framework via an enrichment of the state space: one may extend the state space to include the potential common components in noises just as a different source of information. Then, the cost can be defined correspondingly to price those correlations properly. In Appendix G.2, we develop a formal model that accommodates correlated noises via extension of the state space.

6.2 Beyond Flexible Sequential Learning

Our baseline framework endows the DM with *full flexibility* to optimize over all conceivable sequential learning strategies. While this is a useful benchmark, real-world DMs may face additional constraints/frictions like delay cost or fatigue. In this section, we posit a generalized optimization framework that can incorporate such constraints, bringing our model closer to reality and helping to isolate the key forces driving our results. We begin by defining a generalized notion of indirect cost.

Definition 14 (GSLM). *$\widehat{\Phi} : \mathcal{C} \to \mathcal{C}$ is a generalized sequential learning map (GSLM) if it (i) is isotone and (ii) satisfies $\widehat{\Phi}(C^H_{ups}) = C^H_{ups}$ for all convex $H \in \mathbf{C}^2(\Delta(\Theta))$. For any $C \in \mathcal{C}$, we call $\widehat{\Phi}(C)$ the generalized indirect cost of C and say that C is generalized SLP if $\widehat{\Phi}(C) = C$.*

We interpret GSLMs as modeling "some optimization procedure," in which the DM may have less (or more) flexibility than in our baseline model. Definition 14 requires

[35] A simple illustrating example is $s_1 = \theta + \epsilon_1$ and $s_2 = \epsilon_2$, where ϵ_i are state independent noises. Then, the second experiment induces a degenerate random posterior. However, conditional on observing s_1 already, s_2 may be very informative about θ if $\text{corr}(\epsilon_1, \epsilon_2)$ is high (e.g. $\epsilon_1 = \epsilon_2$). Then, θ can be perfectly learned with arbitrarily small cost (when ϵ_i are extremely noisy). We thank Ian Jewitt for raising this paradox.

this procedure to satisfy two properties. The first, isotonicity, is a minimal consistency condition on the strategy space and objective function. Second, we require all smooth, strong UPS cost functions to be fixed points of $\widehat{\Phi}$. [36]

This abstract notion of GSLM allows us to study a broad range of optimization procedures without explicitly modeling them. In Appendix H, we identify (minimal) sufficient conditions on the GSLM under which each of main results generalizes. Table 1 summarizes these properties (which are self-explanatory) and generalized results; in Appendix H, we also explain how each sub-part of Theorems 5 and 6 generalizes. [37]

Properties of GSLMs	*Definitions*
Allows Direct Learning (ADL)	$\widehat{\Phi}(C) \leq C$
Allows Incremental Evidence (AIE)	$\widehat{\Phi}(C) \leq \Phi_{\text{IE}}(C)$
Exhausts Optimization (EO)	$\widehat{\Phi} = \widehat{\Phi} \circ \widehat{\Phi}$
Generates Subadditivity (GS)	$\widehat{\Phi}(C)$ is Subadditive

Results	*Hold under*
Corollary 1.1	ADL & EO
Theorem 2	GS
Theorem 3	AIE
Theorem 4	ADL & AIE

Table 1: Properties of GSLMs (left) and extensions of main results (right).

Applications. In Appendix H, we develop three applications of the GSLM framework.

Restricted free disposal: In the first application, we consider two model variants with restrictions on "free disposal" of information. First, we disallow free disposal altogether (via the stronger constraint $\mathbb{E}_{\Pi}[\pi_2] = \pi$). The resulting GSLM satisfies all of the properties in Table 1, but the indirect cost *need not* be Monotone—*even if* the underlying direct cost is. Second, we study a variant of Φ that permits free disposal, *but only in the final round* (after all information has been acquired). While this ensures that the indirect cost is Monotone, the corresponding GSLM may *violate* EO and GS. We discuss the role of free disposal in dynamic vs. static information acquisition problems (cf. Caplin and Dean 2015; Oliveira et al. 2017).

Continuous-time model: In the second application, we use GSLMs to explicitly model a continuous-time information acquisition procedure and show that (a suitable version of) Theorem 4 holds. In doing so, we unify and derive novel converses to the results from several related continuous-time papers (Morris and Strack 2019; Zhong 2022; Hébert and Woodford 2023).

[36] A sufficient (but not necessary) condition for this property to hold is that the map $\widehat{\Phi} : \mathcal{C} \to \mathcal{C}$ satisfies $\Phi(C) \leq \widehat{\Phi}(C) \leq \max\{C, \Phi_{\text{IE}}(C)\}$ for all $C \in \mathcal{C}$, i.e., represents an optimization procedure that is more constrained than our baseline model but is flexible enough to permit *either* direct learning *or* incremental learning. We thus view it as an innocuous requirement.

[37] Theorem 1 does not appear in Table 1 because the characterization of SLP as "Monotonicity + Subadditivity" relies on the specific structure of our baseline Φ map. To extend the "lower SLP envelope" characterization from Corollary 1.1, we use the notion of *generalized SLP* from Definition 14.

History-dependent direct cost: We have assumed that the direct cost function does not change over time. While this assumption is ubiquitous in models of dynamic information acquisition,[38] in reality the direct cost function may "increase" or "decrease" over time as the DM develops "fatigue" or "expertise," respectively (Dillenberger, Krishna, and Sadowski 2023). In the third application, we use GSLMs to embed into our framework direct cost functions that can depend arbitrarily on the *full history* of acquired experiments and realized signals. Under a mild form of "expertise," the resulting GSLM satisfies GS, so Theorem 2 holds. Under a mild form of "fatigue," the resulting GSLM satisfies ADL and AIE, so Theorems 3 and 4 hold.

6.3 Beyond Regularity

Theorem 4 fully characterizes the sequential learning map Φ under two smoothness conditions: it restricts attention to (i) the *domain* of Locally Quadratic direct costs and (ii) the *co-domain* of Regular/UPS indirect costs. The domain restriction is mild and made only for technical convenience (see Remark 3). Thus, the main question left open by our analysis is a characterization of Φ for the *full co-domain* of indirect costs. That is, how might we extend Theorem 4 to the case in which $\Phi(C)$ is not Regular/UPS?[39]

While Theorem 2 suggests that the Regular case is most amenable to applications, Theorems 5 and 6 indicate that it can be economically restrictive. In particular, "most" SPI indirect costs—including the indirect LLR cost—are non-Regular. Consequently, to the extent that Prior Invariance and CMC are natural properties for the *direct* cost of information, it is economically important to look beyond Regular *indirect* costs.

By analogy to Theorem 4, it is natural to conjecture that the key to characterizing Φ in general is finding a suitable *local* property of the direct cost C that is *invariant* under Φ. Theorem 3(ii) shows that the lower kernels of C (which always exist) provide one such invariant property. However, it is currently unclear how to leverage this fact into a *global* characterization of $\Phi(C)$ when C violates FLIEs and/or its lower kernels are not integrable. It therefore seems likely that tackling the non-Regular case requires new techniques, the development of which is an exciting task for future research.

[38] See, e.g., Wald (1945), Arrow, D. Blackwell, and Girshick (1949), Moscarini and Smith (2001), Fudenberg, Strack, and Strzalecki (2018), Che and Mierendorff (2019), Morris and Strack (2019), Liang, Mu, and Syrgkanis (2022), Zhong (2022), and Hébert and Woodford (2023).

[39] We note that Corollary 1.1 offers a fully general (variational) characterization of Φ, but not a practical method to actually calculate $\Phi(C)$ or $\Phi^{-1}(C^*)$ for general $C \in \mathcal{C}$ and $C^* \in \mathcal{C}^*$. A proper generalization of Theorem 4 would both fully characterize Φ and deliver a tractable method to calculate these objects.

Appendix

A Main Proofs for Section 3

A.1 Proof of Theorem 1

Proof. To prove the first equivalence, it suffices to show that $\Phi(C)$ is SLP. $\forall \pi \in \mathcal{R}$, $\forall \Pi$ s.t. $\mathbb{E}_{\Pi}[\pi_2] \geq_{mps} \pi$, by the definition of Φ, $\forall \epsilon > 0$, there exists n s.t.

$$\Psi^n(C)(\pi') \leq \Phi(C)(\pi') + \epsilon$$

for all $\pi' \in \{\pi_1\} \cup \mathrm{Supp}(\Pi) \setminus \mathcal{R}^{\circ}$ (note that such n exists because there are only finitely many non-degenerate π'. For degenerate random posterior, the inequality holds trivially for any n.). Therefore,

$$\begin{aligned}
&\Phi(C)(\pi_1) + \mathbb{E}_{\Pi(\pi')}[\Phi(C)(\pi')] \\
\geq &\Psi^n(C)(\pi_1) + \mathbb{E}_{\Pi(\pi')}[\Psi^n(C)(\pi')] - 2\epsilon \\
\geq &\Psi^{n+1}(C)(\pi) - 2\epsilon \\
\geq &\Phi(C)(\pi) - 2\epsilon
\end{aligned}$$

Since ϵ can be chosen arbitrarily, $\Psi(\Phi(C)(\pi)) \geq \Phi(C)(\pi)$. Therefore, $\Phi(C)$ is SLP.

Next, we prove the second equivalence. For any contingent plan Π such that $\mathbb{E}_{\Pi}[\pi_2] \geq_{mps} \pi$, Subadditive implies that $C(\pi_1) + \mathbb{E}_{\Pi}[C(\pi_2)] \geq C(\mathbb{E}_{\Pi}[\pi_2])$, which is greater than $C(\pi)$ by Monotone. Therefore, $C \leq \Psi(C) \leq \Phi(C)$; hence, C is SLP. If C is SLP, then $\pi' \geq_{mps} \pi$ defines a contingent plan for π ($\pi_1 = \pi'$ and π_2's are trivial). Therefore, SLP implies that $C(\pi') \geq \Phi(C)(\pi) = C(\pi)$. Any Π is a contingent plan for $\mathbb{E}_{\Pi}[\pi_2]$. Therefore, SLP implies that $C(\pi_2) + \mathbb{E}_{\pi}[C(\pi_2)] \geq \Phi(C)(\mathbb{E}_{\Pi}[\pi_2]) = C(\mathbb{E}_{\Pi}[\pi_2])$. □

A.2 Proof of Theorem 2

We begin with stating a few key lemmas for the proof of Theorem 2.

Lemma 2. *If $C \in \mathcal{C}$ is Subadditive, then it is Convex (as in Remark 1) and Dilution Linear.*

Proof. See Appendix D.2. □

Recall that a cost function $C \in \mathcal{C}$ is *Posterior Separable* if there exists a divergence D such that $C(\pi) = \mathbb{E}_{\pi}[D(q \mid p_{\pi})]$ for all $\pi \in \mathcal{R}$ (Caplin, Dean, and Leahy 2022).

Lemma 3. *For any open convex $W \subseteq \Delta(\Theta)$ and Posterior $C \in \mathcal{C}$ with* $\mathrm{dom}(C) = \Delta(W) \cup \mathcal{R}^{\circ}$ *and divergence D, C is Subadditive if and only if*

$$\mathbb{E}_{\pi}[D(q \mid p)] \leq D(p_{\pi} \mid p) + \mathbb{E}_{\pi}[D(q \mid p_{\pi})] \quad \forall \pi \in \Delta(W) \textit{ and } p \in W \textit{ s.t. } p_{\pi} \ll p. \tag{2}$$

Proof. See Appendix D.2. □

Lemma 4. *Let $W \subseteq \Delta^\circ(\Theta)$ be open and convex, and let $p \in W$ be given. If $f : W \to \mathbb{R}^{|\Theta|}$ satisfies $f(p) = \mathbf{0}$ and $\mathbb{E}_\pi[f(q)] = \mathbf{0}$ for all finite-support $\pi \in \mathcal{R}(p)$, then there exists a matrix $A \in \mathbb{R}^{|\Theta| \times |\Theta|}$ such that $Ap = \mathbf{0}$ and $f(q) \equiv -Aq$ on W.*

Proof. See Appendix D.2. □

Lemma 5. *Let $W \subseteq \Delta^\circ(\Theta)$ be open and convex. Let $C \in \mathcal{C}$ satisfy $\operatorname{dom}(C) = \Delta(W) \cup \mathcal{R}^\circ$, be Subadditive, and be Posterior Separable with divergence D. If D satisfies*

$$(q, p) \mapsto \nabla_2 D(q \mid p) \text{ is well-defined and continuous on } \operatorname{dom}(D) = W \times W, \tag{3}$$

then $C = C^H_{ups}$ for some convex $H \in \mathbf{C}^1(W)$.

Proof. We prove the lemma in four steps.

Step 1: Linear prior-gradient. Since C is Subadditive and Posterior Separable, Lemma 3 implies that, for every $\pi \in \Delta(W)$,

$$0 \leq f^\pi(p) := D(p_\pi \mid p) + \mathbb{E}_\pi[D(q \mid p_\pi) - D(q \mid p)] \quad \forall p \in W.$$

Note that the functions $f^\pi : W \to \mathbb{R}_+$ and $D(p_\pi \mid \cdot) : W \to \mathbb{R}_+$ are both minimized at $p = p_\pi$ (where they both equal 0). Furthermore, if $|\operatorname{supp}(\pi)| < \infty$, then f^π is differentiable and its gradient is given by

$$\nabla f^\pi(p) = \nabla_2 D(p_\pi \mid p) - \mathbb{E}_\pi[\nabla_2 D(q \mid p)],$$

where (3) ensures that $\nabla_2 D(\cdot \mid p)$ is well-defined on W, and $|\operatorname{supp}(\pi)| < \infty$ ensures that we can interchange the order of differentiation and integration in the second term. Thus, the necessary FOCs for minimization of f^π and $D(p_\pi \mid \cdot)$ at $p = p_\pi$ yield, respectively, $\nabla f^\pi(p_\pi) = \nabla_2 D(p_\pi \mid p_\pi) - \mathbb{E}_\pi[\nabla_2 D(q \mid p_\pi)] = \mathbf{0}$ and $\nabla_2 D(p_\pi \mid p_\pi) = \mathbf{0}$. Hence,

$$\mathbb{E}_\pi[\nabla_2 D(q \mid p_\pi)] = \mathbf{0} \quad \forall \text{ finite-support } \pi \in \Delta(W).$$

Thus, for each $p_\pi \in W$, applying Lemma 4 to the map $\nabla_2 D(\cdot \mid p_\pi) : W \to \mathbb{R}^{|\Theta|}$ delivers

$$\nabla_2 D(q \mid p_\pi) = -A(p_\pi) q \quad \forall q \in W \tag{4}$$

for some matrix $A(p_\pi) \in \mathbb{R}^{|\Theta| \times |\Theta|}$ satisfying $A(p_\pi) p_\pi = \mathbf{0}$. Let $A : W \to \mathbb{R}^{|\Theta| \times |\Theta|}$ denote the corresponding matrix-valued function.

Step 2: Directional posterior-derivatives. For any $p, q \in W$, the Gradient Theorem and (4) deliver

$$D(q \mid p) = \int_a^b \nabla_2 D(q \mid r(x)) \cdot r'(x) \, \mathrm{d}x = -\int_a^b A(r(x)) q \cdot r'(x) \, \mathrm{d}x \tag{5}$$

for all $a, b \in \mathbb{R}$ and $\mathbf{C}^1$-smooth curves $r : [a, b] \to W$ such that $r(a) = q$ and $r(b) = p$, where $r'(x) \in \mathcal{T}(\Delta)$ for all $x \in [a, b]$.

Let $q, p \in W$ and $y \in \mathcal{T}(\Delta)$ be given. Let $\delta \in (0, 1/2)$ be given and sufficiently small that $q + \eta y \in W$ for all $\eta \in [-\delta, \delta]$, and consider any $\mathbf{C}^1$-smooth curve $r : [0, 1] \to W$ for which

(i) $r(x) = q + (x - \delta)y$ for all $x \in [0, 2\delta]$ and (ii) $r(1) = p$.[40] Note that $r(\delta) = q$ and $r'(x) = y$ for all $x \in [0, 2\delta]$. Thus, for any $\epsilon' \in (-\delta, \delta)$ and corresponding $\zeta := \delta + \epsilon'$, the (two-sided) directional derivative of $D(\cdot \mid p)$ at $r(\zeta) = q + \epsilon' y$ in direction y is given by

$$\begin{aligned}\frac{\partial}{\partial \epsilon} D(q + \epsilon' y + \epsilon y \mid p)\Big|_{\epsilon=0} &= \frac{\mathrm{d}}{\mathrm{d}t} D(r(t) \mid p)\Big|_{t=\zeta} \\ &= -\frac{\mathrm{d}}{\mathrm{d}t}\left[\int_t^1 A(r(x))\, r(t) \cdot r'(x)\, \mathrm{d}x\right]\Big|_{t=\zeta} \\ &= A(r(\zeta))r(\zeta) \cdot r'(\zeta) - \int_\zeta^1 A(r(x))r'(\zeta) \cdot r'(x)\, \mathrm{d}x \\ &= -\int_{\delta+\epsilon'}^1 A(r(x))y \cdot r'(x)\, \mathrm{d}x,\end{aligned}$$

where the first two lines follow from the definition of the curve r and the identity (5), the third line follows from the standard Leibniz rule,[41] and the final line follows from the definition of ζ and the facts that $A(q')q' \equiv \mathbf{0}$ on W and $r'(\zeta) = y$. Then the second-order (two-sided) directional derivative of $D(\cdot \mid p)$ at q in direction y as

$$\begin{aligned}\frac{\partial^2}{\partial \epsilon' \partial \epsilon} D(q + \epsilon' y + \epsilon y \mid p)\Big|_{\epsilon=\epsilon'=0} &= -\frac{\mathrm{d}}{\mathrm{d}t}\left[\int_t^1 A(r(x))y \cdot r'(x)\, \mathrm{d}x\right]\Big|_{t=\delta} \\ &= A(r(\delta))y \cdot r'(\delta) \\ &= y^\top A(q) y \qquad (6)\end{aligned}$$

where the first line follows from the preceding display, the second line follows from the Leibniz rule, and the final line follows from the definitions of the curve r (viz., $r(\delta) = q$ and $r'(\delta) = y$) and the dot product.

Step 3: UPS Representation. We now show that C has a UPS representation. Let $p^* \in W$ be given and define the map $H : W \to \mathbb{R}_+$ as $H(q) := D(q \mid p^*)$. For all $q, p \in W$, define $L(q, p) := D(q \mid p) - D(q \mid p^*)$, so that $D(q \mid p) = H(q) + L(q, p)$.

We claim that, for every $p \in W$, the map $L(\cdot, p) : W \to \mathbb{R}$ is affine, i.e., $L(\alpha q_1 + (1-\alpha)q_0) = \alpha L(q_1, p) + (1 - \alpha)L(q_0, p)$ for all $q_0, q_1 \in W$ and $\alpha \in [0, 1]$. Let $p, q_0, q_1 \in W$ be given; the $q_0 = q_1$ case is trivial, so let $q_0 \neq q_1$. Define $y := q_1 - q_0 \in \mathcal{T}(\Delta)$ and the map $f : [0, 1] \to \mathbb{R}$ as $f(t) := L(q_0 + ty, p)$. It follows from Step 2 that f is continuous on $[0, 1]$ and twice

[40] Such $\delta > 0$ and curves r exist because $W \subseteq \Delta^\circ(\Theta)$ and W is open and convex.

[41] The Leibniz rule applies because the function $x \mapsto \frac{\mathrm{d}}{\mathrm{d}t} A(r(x))r(t) \cdot r'(x)\big|_{t=\zeta} = A(r(x))y \cdot r'(x) \in \mathbb{R}$ is continuous on $[0, 1]$. In particular, for every $x \in [0, 1]$ and $\eta \in [-\delta, \delta]$, (4) implies that $\nabla_2 D(q \mid r(x)) = -A(r(x))q$ and $\nabla_2 D(q + \eta y \mid r(x)) = -A(r(x))(q + \eta y)$. For any given $\eta \in [-\delta, \delta] \backslash \{0\}$, this implies that $-A(r(x))y = \frac{1}{\eta}(\nabla_2 D(q + \eta y \mid r(x)) - \nabla_2 D(q \mid r(x)))$. Thus, (3) implies that $x \mapsto -A(r(x))y \in \mathbb{R}^{|\Theta|}$ is continuous on $[0, 1]$. Continuity of $r' : [0, 1] \to \mathcal{T}(\Delta)$ concludes the argument.

differentiable on $(0,1)$. Moreover, since (6) implies that

$$\forall t \in [0,1], \quad p \mapsto \frac{\partial^2}{\partial\epsilon'\partial\epsilon} D(q_0 + ty + \epsilon y + \epsilon' y \mid p)\Big|_{\epsilon=\epsilon'=0} \text{ is constant on } W,$$

it follows from the definition of $L(\cdot, p)$ that

$$f''(t) = \frac{\partial^2}{\partial\epsilon'\partial\epsilon} L(q_0 + ty + \epsilon y + \epsilon' y, p)\Big|_{\epsilon=\epsilon'=0} = 0 \qquad \forall t \in [0,1].$$

Thus, $f(t) = tf(0) + (1-t)f(1)$ for all $t \in [0,1]$ (e.g., Lemma 13 in Royden and Fitzpatrick (2010, Ch. 6)). We conclude that every $L(\cdot, p)$ is affine (hence, continuous) on W.

It follows that, for every $\pi \in \Delta(W)$, $\mathbb{E}_\pi[L(q, p_\pi)] = L(p_\pi, p_\pi)$ and therefore $C(\pi) = \mathbb{E}_\pi[H(q) + L(p_\pi, p_\pi)]$. Since $C(\delta_p) = 0$ for all $p \in W$, $H(p) \equiv -L(p,p)$ on W. Since $C(\pi) \geq 0$ for all $\pi \in \Delta(W)$, H is convex. We conclude that $C = C^H_{ups}$.

Step 4: Smooth Potential. It remains to show that $H \in \mathbf{C}^1(W)$. To this end, note that we have shown in Step 2 that $H(\cdot) = D(\cdot \mid p^*)$ has directional derivatives at every $q \in W$ and every direction $y \in \mathcal{T}(\Delta)$. Being that H is convex and $W \subseteq \Delta^\circ(\Theta)$ is open, Theorem 25.2 and Corollary 2.5.5.1 in Rockafellar (1970) imply that $H \in \mathbf{C}^1(W)$, as desired. □

With Lemma 5 in hand, we are in a position to prove Theorem 2 itself. To do so, we adapt a mollification argument from Banerjee, Guo, and Wang (2005). Recall that a map $\xi : \mathcal{T}(\Delta) \to \mathbb{R}_+$ is a *(positive) mollifier* if it satisfies the following conditions: (i) $\xi \in \mathbf{C}^\infty(\mathcal{T}(\Delta))$, (ii) $\operatorname{supp}(\xi)$ is compact, (iii) $\int_{\mathcal{T}(\Delta)} \xi(y)\,\mathrm{d}y = 1$, and (iv) defining for every $\epsilon > 0$ the function $\xi_\epsilon(\cdot) := \epsilon^{-|\Theta|}\xi(\cdot/\epsilon)$, $\lim_{\epsilon\to 0} \xi_\epsilon(y) = \delta(y)$ for all $y \in \mathcal{T}(\Delta)$, where $\delta(y)$ denotes the Dirac delta function on y.

Proof of Theorem 2. **($\Longleftarrow$ direction)** Let $C = C^H_{ups}$ with $H \in \mathbf{C}^1(W)$. Let $D_H(q \mid p) := H(q) - H(p) - \nabla H(p) \cdot (q - p)$ denote the Bregman divergence associated with H. Then $C(\pi) \equiv \mathbb{E}_\pi[D_H(q \mid p_\pi)]$ and $\nabla_1 D_H(q \mid p) = \nabla H(q) - \nabla H(p)$, which is jointly continuous on W. Thus, C is Regular with derivative D_H.

($\Longrightarrow$ direction) We begin with some preliminaries. For each $\epsilon > 0$, let $B^{\mathcal{T}}_a(\mathbf{0}) := \{y \in \mathcal{T} \mid \|y\| < \epsilon\}$ denote the ball in $\mathcal{T}$ of radius ϵ centered at $\mathbf{0}$. Let a mollifier $\xi : \mathcal{T}(\Delta) \to \mathbb{R}_+$ with $\operatorname{supp}(\xi) \subseteq B^{\mathcal{T}}_1(\mathbf{0})$ be given. Then, for every $\epsilon > 0$, $\operatorname{supp}(\xi_\epsilon) \subseteq B^{\mathcal{T}}_\epsilon(\mathbf{0})$. Let $p^* \in \operatorname{relint}(W)$ be given; by definition, there exists a $\delta > 0$ such that $B_\delta(p^*) \subseteq \operatorname{relint}(W)$. For every $\eta \in (0,1)$, define $W_\eta := \{(1-\eta)p + \eta p^* \mid p \in W\}$. It is easy to see that W_η is a convex subset of $\operatorname{relint}(W)$; in particular, for every $p \in W_\eta$, $B_{\eta\delta}(p) \subset \operatorname{relint}(W)$. Thus, for every $p \in W_\eta$ and $\epsilon \in (0, \eta\delta)$, $p + y \in \operatorname{ri}(W)$ for all $y \in \operatorname{supp}(\xi_\epsilon)$.

Now, let $C \in \mathcal{C}$ be Regular with derivative D. Let $\eta \in (0,1)$ and $\epsilon \in (0, \eta\delta)$ be given.

Define the map $D_\epsilon : W_\eta \times W_\eta \to \mathbb{R}_+$ as

$$D_\epsilon(q \mid p) := \int_{\mathcal{T}} D(q+y \mid p+y)\xi_\epsilon(y)\,\mathrm{d}y,$$

which is well-defined and continuous by the preceding paragraph and continuity of D. Changing variables from y to $r := p+y$, we have

$$D_\epsilon(q \mid p) = \int_{\Delta(\Theta)} D(q-p+r \mid r)\xi_\epsilon(r-p)\,\mathrm{d}r. \tag{7}$$

We claim that D_ϵ satisfies the hypotheses of Lemma 5 on W_η. First, we show that it satisfies (3) on this restricted domain. Denote the (continuous) integrand in (3) by $f(q,p,r) := D(q-p+r \mid r)\xi_\epsilon(r-p)$. Since $\nabla_1 D : W \times W \to \mathbb{R}^{|\Theta|}$ is continuous by hypothesis, we have $(q,p,r) \mapsto \nabla_2 f(q,p,r)$ is well-defined and continuous. Thus, by the fundamental theorem of calculus, $\nabla_2 D_\epsilon : W_\eta \times W_\eta \to \mathbb{R}^{|\Theta|}$ is well-defined and continuous, as desired. Next, we show that it satisfies (2) for all $\pi \in \Delta(W_\eta)$ and $p \in W_\eta$. So, let $\pi \in \Delta(W_\eta)$ and $p \in W_\eta$ be given. For each $y \in \operatorname{supp}(\xi_\epsilon)$, define $\pi_y \in \Delta(W)$ as $\pi_y(\{q+y \mid q \in E\}) := \pi(E)$ for all Borel $E \subseteq W_\eta$. Note that $p_{\pi_y} = p_\pi + y \in W$. Then we have

$$\begin{aligned}
\mathbb{E}_\pi\left[D_\epsilon(q \mid p)\right] &= \int_{\mathcal{T}} \mathbb{E}_\pi\left[D(q+y \mid p+y)\right]\xi_\epsilon(y)\,\mathrm{d}y \\
&= \int_{\mathcal{T}} \mathbb{E}_{\pi_y}\left[D(q \mid p+y)\right]\xi_\epsilon(y)\,\mathrm{d}y \\
&\le \int_{\mathcal{T}} \Big[D(p_\pi+y \mid p+y) + \mathbb{E}_{\pi_y}\left[D(q \mid p_\pi+y\right]\Big]\xi_\epsilon(y)\,\mathrm{d}y \\
&= \int_{\mathcal{T}} \Big[D(p_\pi+y \mid p+y) + \mathbb{E}_\pi\left[D(q+y \mid p_\pi+y\right]\Big]\xi_\epsilon(y)\,\mathrm{d}y \\
&= D_\epsilon(p_\pi \mid p) + \mathbb{E}_\pi\left[D_\epsilon(q \mid p_\pi\right],
\end{aligned}$$

where the first line is by Fubini, the second line is by definition of π_y, the third line is because D satisfies (2) on W, the fourth line is by definition of π_y, and the final line is by Fubini. Thus, D_ϵ satisfies (2) on W_η, as desired.

Thus, we can apply Lemma 5 to D_ϵ on W_η to establish that $H_{\eta,\epsilon}(\cdot) := D_\epsilon(\cdot \mid p^*) \in \mathbf{C}^1(W_\eta)$ is convex and satisfies $\mathbb{E}_\pi\left[D_\epsilon(q \mid p_\pi)\right] = C_{ups}^{H_{\eta,\epsilon}}(\pi)$ for all $\pi \in \Delta(W_\eta)$. Fixing η and sending $\epsilon \to 0$, we have $D_\epsilon(q \mid p) \to D(q \mid p)$ for all $q,p \in W_\eta$ by construction of the mollifier. Thus, for each $\eta \in (0,1)$, $H_\eta(q) := \lim_{\epsilon\to 0} H_{\eta,\epsilon}(q) = \lim_{\epsilon\to 0} D(q \mid p^*)$ defines a convex function $H_\eta : W_\eta \to \mathbb{R}_+$. Moreover, $H(q) := \inf\{H_{1/n}(q) : n \in \mathbb{N}\}$ defines a convex function $H : W \to \mathbb{R}_+$ such that $H(q) = H_\eta(q)$ for any $\eta \in (0,1)$ such that $q \in W_\eta$. Thus, for any finite-support $\pi \in \Delta(W)$, since $\operatorname{supp}(\pi) \subseteq W_{1/n}$ for some $n \in \mathbb{N}$ (being that $W = \operatorname{relint}(W)$), we have $C(\pi) = \mathbb{E}_\pi[D(q \mid p_\pi)] = C_{ups}^{H_\eta}(\pi) = C_{ups}^{H}(\pi)$. Since D is continuous and hence C is weak* continuous on each $\Delta(W_\eta)$, a standard approximation argument implies that this representation extends to all $\pi \in \Delta(W)$. □

B Main Proofs for Section 4

B.1 Proof of Theorem 3

Proof. **The statement about upper quadratic kernels.**

Step 1. Prove the statement for $\pi \in \Delta(W)$ with binary support. Let $p = p_\pi$ and $\text{Supp}(\pi) = \{q_1, q_2\}$. Since $W \subset \Delta(\Theta)^\circ$ is open, there exists interval $[q_1', q_2']$ in $\Delta(\Theta)^\circ$ such that q_1, q_2 are interior points of the interval. Let $\widehat{\pi}$ denote the (unique) random posterior with prior p and support $\left\{q_1', q_2'\right\}$.

$\forall \epsilon > 0$, by Definition 8, $\forall p_0$ in $[q_1', q_2']$, there exists a corresponding $\delta > 0$ defining $\text{Hess}H(p_0)$ as an upper quadratic kernel. Then, $B_{\delta/2}(p_0)^\circ$'s constitute a collection of open covers of the compact interval, and therefore, a finite cover exists. Let δ' be the radius of the smallest ball. WLOG, δ' can be chosen smaller than the continuity parameter of $\text{Hess}H$ on $[q_1', q_2']$. Now, construct a finite grid of size δ' of $[q_1', q_2']$ that contains p. Denote the grid by $G = (\widehat{q_i})_{i=1}^N$. Let ξ be $\min \|\widehat{q_i} - \widehat{q}_{i+1}\|$. We prove by induction that there exists π' s.t. $\text{Supp}(\pi') \subset G$ and $p_{\pi'} = p$ s.t. (i) $\mathbb{E}_{\pi'}[\|q-p\|^2]$ is arbitrarily close to $\mathbb{E}_{\widehat{\pi}}[\|q-p\|^2]$ and (ii) $\Phi(C)(\pi') \le \mathbb{E}_{\pi'}[H(q) - H(p)] + 2\epsilon \mathbb{E}_{\pi'}[\|q-p\|^2]$.

We iterate on $(P, \sigma) = (\pi'(\{q_1', q_2'\}), \mathbb{E}_{\pi'}[\|q-p\|^2])$. Obviously, such π' exists for $(P = 0, \sigma = 0)$. Then, we show that for every π' that satisfies the constraint (ii) with (P, σ), there exists π'' that satisfies the constraint (ii) and $(P' \ge P, \sigma' \ge \sigma + (1-P)\xi^2)$. Construct π'' as a contingent plan with $\pi_1 = \pi'$ and contingent on $\widehat{q_i}$ $(i \ne 1, N)$, π_2 being the binary experiment with support $\{\widehat{q}_{i-1}, \widehat{q}_{i+1}\}$. Therefore,

$$\begin{aligned}
\Phi(C)(\pi'') \le& \Phi(C)(\pi') + \sum_{i=2}^{N-1} \pi'(\widehat{q_i}) \Phi(C)(\pi_2|\widehat{q_i}) \\
\le& \mathbb{E}_{\pi'}[H(q) - H(p)] + 2\epsilon \mathbb{E}_{\pi'}[\|q-p\|^2] + \sum_{i=2}^{N-1} \pi'(\widehat{q_i}) \mathbb{E}_{\pi_2|\widehat{q_i}} \left[(q - \widehat{q_i})^T \frac{1}{2} \text{Hess}H(\widehat{q_i}) + \epsilon)(q - \widehat{q_i}) \right] \\
\le& \mathbb{E}_{\pi'}[H(q) - H(p)] + 2\epsilon \mathbb{E}_{\pi'}[\|q-p\|^2] + \sum_{i=2}^{N-1} \pi'(\widehat{q_i}) \mathbb{E}_{\pi_2|\widehat{q_i}} \left[H(q - \widehat{q_i}) + 2\epsilon \|q - \widehat{q_i}\|^2 \right] \\
=& \mathbb{E}_{\pi''} \left[(H(q) - H(p)) + 2\epsilon \|q-p\|^2 \right].
\end{aligned}$$

The first inequality is from $\Phi(C)$ being SLP. The second inequality is from the induction hypothesis and the definition of upper quadratic kernel. Note that since $\delta' \le \delta/2$, $\text{Supp}(\pi_2|\widehat{q_i}) \subset B_{\delta/2}(\widehat{q_i})$. Since $\widehat{q_i}$ lies in one of the open cover with radius δ, so does $B_{\delta/2}(\widehat{q_i})$.

The third inequality is from $\|\text{Hess}H(\widehat{q_i})-\text{Hess}H(q)\| \leq \epsilon$ for $q \in B_{\delta/2}(\widehat{q_i})$. On the other hand

$$\mathbb{E}_{\pi''}\left[\|q-p\|^2\right] = \mathbb{E}_{\pi'}[\|q-p\|^2] + \sum_{i=2}^{N-1} \pi'(\widehat{q_i})\mathbb{E}_{\pi_2|\widehat{q_i}}\left[\|q-\widehat{q_i}\|^2\right] \geq \sigma + (1-P)\xi^2.$$

Therefore, we find a sequence of π_n', with (P_n, σ_n). By construction, $P_n \to 1$ or otherwise $\sigma_n \to \infty$, which is impossible since σ_n is bounded by $\mathbb{E}_{\widehat{\pi}}[\|q-p\|^2]$. Then, for P_n sufficiently close to 1, $\pi_n' \geq_{mps} \pi$. Since $\Phi(C)$ is Monotone,

$$\Phi(C)(\pi) \leq \Phi(C)(\pi_n') \leq \mathbb{E}_{\pi_n'}[H(q)-H(p)] + 2\epsilon\mathbb{E}_{\pi'}\left[\|q-p\|^2\right].$$

When $\epsilon, \delta \to 0$, the RHS converges to $\mathbb{E}_{\pi}[H(q)-H(p)]$.

Step 2. Prove the statement for $\pi \in \Delta(W)$ with finite support. We induce on the support size and assume that the statement is proved for $N-1$, Suppose $\text{Supp}(\pi) = \{q_1, \ldots, q_N\}$. Define $\pi_1(q_i) = \pi(q_i)$ for $i < N-1$ and $\pi_1\left(\frac{\pi(q_{N-1})q_{N-1}+\pi(q_N)q_N}{\pi(\{q_{N-1},q_N\})}\right) = \pi(\{q_{N-1}, q_N\})$. Define π_2 being degenerate contingent on q_i, $i < N-1$ and $\pi_2(q_{N-1}) = \frac{\pi(q_{N-1})}{\pi(\{q_{N-1},q_N\})}$, $\pi_2(q_N) = \frac{\pi(q_N)}{\pi(\{q_{N-1},q_N\})}$. Then, since $\Phi(C)$ is SLP,

$$\begin{aligned}\Phi(C)(\pi) \leq& \Phi(C)(\pi_1) + \pi(\{q_{N-1}, q_N\})\Phi(C)(\pi_2)\\ \leq& \mathbb{E}_{\pi_1}[H(q)-H(p)] + \pi(\{q_{N-1}, q_N\})\mathbb{E}_{\pi_2}[H(q)-H(\mathbb{E}_{\pi_2}[q])]\\ =& \mathbb{E}_{\pi}[H(q)-H(p)].\end{aligned}$$

Step 3. Prove the statement for general $\pi \in \Delta(W)$. It is sufficient to show that there exists finite support $\pi' \geq_{mps} \pi$ with $\mathbb{E}_{\pi'}[H(q)-H(p)]$ arbitrarily close to $\mathbb{E}_{\pi}[H(q)-H(p)]$. Pick arbitrary $\eta > 0$. $\forall p \in \text{supp}(\pi)$, since $p \in W$, it is contained in an open polygon spanned by $\{q_i(p)\} \subset W$ with diameter smaller than η.

The collection of the open polygons constitutes a collection of open cover of the compact set $\text{Supp}(\pi)$; hence, a finite cover exists, denoted by $\left(\left\{q_i^n\right\}\right)_{n=1}^N$. $\forall p \in \text{Supp}(\pi)$, there exists a unique cover with smallest index n that contains p. Since $\left\{q_i^n\right\}$ are linearly independent, there exists unique $(\pi_i) \in \Delta(|\Theta|)$ s.t. $\sum \pi_i q_i^n = p$. Define $\Pi((\pi_i)) = \pi(p)$. Then, the compound experiment $\pi' = \mathbb{E}_{\Pi}[\pi_2] \geq_{mps} \pi$ by construction and π' has finite support. Since $\Phi(C)$ is Monotone,

$$\begin{aligned}\Phi(C)(\pi) \leq& \Phi(C)(\pi') \leq \mathbb{E}_{\pi'}[H(q)-H(p)]\\ =& \mathbb{E}_{\pi}[H(q)-H(p)] + \mathbb{E}_{\pi}\left[\mathbb{E}_{\pi_2(q'|q)}[H(q')-H(q)]\right]\\ \leq& \mathbb{E}_{\pi}[H(q)-H(p)] + \frac{1}{2}\sup_{q \in \text{Conv}(\text{Supp}(\pi'))}\|\text{Hess}H(q)\| \times \eta^2.\end{aligned}$$

When $\eta \to 0$, it follows that $\Phi(C)(\pi) \leq \mathbb{E}_{\pi}[H(q)-H(p)]$.

The statement about lower quadratic kernels.

We begin by showing that $\forall \xi > 0$ s.t. $k - \xi I \geq_{psd} 0$, there exists $H \in C^2\Delta(\Theta)$ s.t. (i)

$C(\pi) \geq \mathbb{E}_\pi[H(q) - H(p_\pi)]$ and (ii) $\text{Hess}H(p_0) \geq_{psd} k - \xi I$.

Since k is the lower quadratic kernel of C at p_0, choose $\varepsilon < \xi$ and choose δ as in Definition 8 corresponding to ε at p_0. ε, δ can be chosen sufficiently small that $\forall p \in B_\delta(p_0)$

$$(1-\varepsilon)(k - 2\varepsilon I) \geq_{psd} k - \xi I. \tag{8}$$

$\forall \chi \in (0, \delta)$, by Lemma 11, there exists $H_\chi \in C^2\Delta(\theta)$ s.t. (i) $0 \leq_{psd} \text{Hess}H_\chi(p) \leq_{psd} k - \xi I$,(ii) $\text{Hess}H_\chi(p_0) = k - \xi I$, and (iii) $\|\text{Hess}H_\chi(p)\| \leq \chi$ when $p \notin B_\chi(p_0)$.

In words, H_χ is locally quadratic with Hessian matrix $k - \xi I$ at p_0 and quickly becomes linear towards the direction from p_0 out of a small ball with radius χ around p_0. It suffices to verify that when χ is sufficiently small, the UPS function with potential H_χ is lower than C. Pick $\delta' \in (\chi, \delta)$. $\forall \pi \in \mathcal{R}$, let $p = p_\pi$. Suppose $p \in B_{\delta'}(p_0)$,

$$\begin{aligned}
&\mathbb{E}_\pi[H_\chi(q) - H_\chi(p)] = \mathbb{E}_\pi[H_\chi(q) - H_\chi(p) - \nabla H_\chi(p)(q-p)] \\
=& \int_{q \in B_\delta(p_0)} H_\chi(q) - H_\chi(p) - \nabla H_\chi(p)(q-p) \mathrm{d}\pi(q) + \int_{q \notin B_\delta(p_0)} H_\chi(q) - H_\chi(p) - \nabla H_\chi(p)(q-p) \mathrm{d}\pi(q) \\
\leq& \int_{q \in B_\delta(p_0)} \frac{1}{2}(q-p)^T (k - \xi)(q-p) \mathrm{d}\pi(q) + \int_{q \notin B_\delta(p_0)} (\chi \cdot \|k - \xi\| + \frac{1}{2}\chi) \cdot \|q - p\| \mathrm{d}\pi(q) \\
\leq& \int_{q \in B_\delta(p_0)} (q-p)^T \frac{1}{2} (k - \xi)(q-p) \mathrm{d}\pi(q) + \chi \cdot (\|k - \xi\| + 1) \pi(\Delta(\theta) \setminus B_\delta(p_0)) \\
\leq& (1-\varepsilon) \int_{q \in B_\delta(p_0)} (q-p)^T (\frac{1}{2}k - \varepsilon)(q-p) \mathrm{d}\pi(q) + \chi \cdot (\|k - \xi\| + 1) \frac{\int_{q \notin B_\delta(p_0)} \|q-p\|^2 \mathrm{d}\pi(q)}{(\delta - \delta')^2}. \\
\leq& (1-\varepsilon) C(\pi) + \chi \cdot (\|k - \xi\| + 1) \frac{C(\pi)}{m(\delta - \delta')^2}.
\end{aligned}$$

The first inequality is from the bound on the Hessian matrix of H_χ. The second inequality is from $\|q - p\| \leq 1$. The third inequality is from Equation (8). The last inequality is from k being a lower quadratic kernel of C at p_0 (first term) and C being Strongly Positive (second term). Fixing ε, δ and δ' and picking χ s.t. $\frac{\chi \cdot (\|k-\xi\|+1)}{m(\delta-\delta')^2} \leq \varepsilon$, the last line is lower than $C(\pi)$. Suppose $p \notin B_{\delta'}(p_0)$,

$$\begin{aligned}
& H_\chi(q) - H_\chi(p) - \nabla H_\chi(p)(q-p) \\
\leq& \mathbf{1}_{q \in B_{\delta'-\chi}(p)} \chi \cdot \|q-p\|^2 + \mathbf{1}_{q \notin B_{\delta'-\chi}(p)} \chi (\|k - \xi\| + 1) \cdot \|q - p\| \\
\leq& \mathbf{1}_{q \in B_{\delta'-\chi}(p)} \chi \cdot \|q-p\|^2 + \mathbf{1}_{q \notin B_{\delta'-\chi}(p)} \chi (\|k - \xi\| + 1) \cdot \|q - p\| \\
\leq& \mathbf{1}_{q \in B_{\delta'-\chi}(p)} \chi \cdot \|q-p\|^2 + \mathbf{1}_{q \notin B_{\delta'-\chi}(p)} \chi (\|k - \xi\| + 1) \cdot \frac{\|q - p\|^2}{\delta' - \chi} \\
\Longrightarrow& \mathbb{E}_\pi[H_\chi(q) - H_\chi(p)] \leq \max\left\{\chi, \frac{\chi(\|k - \xi\| + 1)}{\delta' - \chi}\right\} \cdot \frac{C(\pi)}{m}
\end{aligned}$$

Fixing ε, δ and δ' and picking χ s.t. $\max\left\{\frac{\chi}{m}, \frac{\chi(\|k-\xi\|+1)}{m(\delta'-\chi)}\right\} \leq 1$, the last line is lower than $C(\pi)$.

Finally, because $\mathbb{E}_\pi[H(q) - H(p_\pi)]$ is SLP and is below C, and since Φ is isotone, we have that $\Phi(C)(\pi) \geq \mathbb{E}_\pi[H(q) - H(p_\pi)]$. Therefore, $k - \xi I$ is a lower quadratic kernel of $\Phi(C)$ at p_0. Now, $\forall \epsilon > 0$, pick $\xi = \frac{1}{2}\epsilon$ and pick parameter δ defining the lower quadratic kernel $k - \xi I$ corresponding to $\frac{1}{2}\epsilon$,

$$\Phi(C)(\pi) \geq \int_{B_\delta(p_0)} \frac{1}{2}(q-p)^T(k - \xi I - \epsilon I)(q-p)\pi(\mathrm{d}q).$$

□

B.2 Proof of Theorem 4

Proof. ($\Rightarrow$ direction) C and H have open domain $W \subset \Delta(\Theta)^\circ$. Theorem 3 implies that $\forall \pi \in \Delta(W)$, $\Phi(C)(\pi) \leq C^H_{ups}(\pi)$. $\forall \pi \notin \Delta(W)$, $\Phi(C)(\pi) \leq \infty = C^H_{ups}(\pi)$. Therefore, $\Phi(C) \leq C^H_{ups}$.

FLIEs and Proposition 2 implies $\forall \pi \in \Delta(W)$, $C(\pi) \geq \Phi_{IE}(C) \geq C^H_{ups}(\pi)$. $\forall \pi \notin \Delta(W)$, $C(\pi) = \infty \geq C^H_{ups}(\pi)$. Therefore, $C \geq C^H_{ups}$. Applying Φ to both side, $\Phi(C) \geq C^H_{ups}$.

Combining both inequalities, $\Phi(C) = C^H_{ups}$.

($\Leftarrow$ direction) $\Phi(C) = C^H_{ups}$ implies that Hess$H$ is the kernel of $\Phi(C)$ on W. Theorem 3 implies that k_C is a lower quadratic kernel of $\Phi(C)$ on W. Therefore, $k_C \leq_{psd}$ Hess$H$. On the other hand, k_C being the upper quadratic kernel of C implies Hess$H \leq_{psd} k_C$; hence, Hess$H = k_C$. Then, Proposition 2 implies $\forall \pi \in \Delta(W)$, $\Phi_{IE}(C)(\pi) \leq C^H_{ups}(\pi)$. $\forall \pi \notin \Delta(W)$, $\Phi_{IE}(C)(\pi) \leq \infty = C^H_{ups}(\pi)$. Therefore, $\Phi_{IE}(C) \leq C^H_{ups} \leq \Phi(C) \leq C$. □

C Main Proofs for Section 5

C.1 Proof of Theorem 6

Step 1. We begin with proving that the Wald cost is the indirect cost of a Prior Invariant, Strongly Positive and Locally Quadratic direct cost. Let

$$C(h_B(\sigma, p)) := \max\{D_{KL}(\sigma_0|\sigma_1), D_{KL}(\sigma_1|\sigma_0)\}.$$

Evidently, C is Prior Invariant. To show that C is Locally Quadratic, it is sufficient to show that $D_{KL}(\sigma_0|\sigma_1)$ and $D_{KL}(\sigma_0|\sigma_1)$ are Locally Quadratic and have the same kernel. Parameterize $\Delta(\Theta)^\circ$ using the probability of $\theta = 1$; $D_{KL}(\sigma_0|\sigma_1)$ can be rewritten as $\mathbb{E}_\pi[D(q|p_\pi)]$, where $D(q|p) = \frac{q}{p}\log\left(\frac{q}{1-q}\right)$. Direct calculation implies that its kernel is

$$D''_{qq}(q|p)\Big|_{q=p} = \frac{1}{p^2(1-p)^2}.$$

Note that the term does not change when the indices of the two states are flipped. Therefore, C is Locally Quadratic with kernel $\frac{1}{p^2(1-p)^2}$, same as C_{Wald}. It remains to prove that C FLIEs, i.e. $C \geq C_{\text{Wald}}$. Note that $H^*(q) = pD(q|p) + (1-p)D(1-q|1-p)$; thus, $\forall \sigma, p$, $C_{\text{Wald}}(h_B(\sigma, p) = pD_{KL}(\sigma_0|\sigma_1) + (1-p)D_{KL}(\sigma_1|\sigma_0) \leq \max\{D_{KL}(\sigma_0|\sigma_1), D_{KL}(\sigma_1|\sigma_0)\} = C(h_B(\sigma, p))$. Then, Theorem 4 implies that $\Phi(C) = C_{\text{Wald}}$.

Step 2. By Theorem 5, SPI and CMC$^{©}$ implies UPS. Therefore, to prove the theorem, it suffices to show when C is SPI, UPS and Locally Quadratic, C must be proportional to the Wald cost and $|\Theta| = 2$. The proof invokes the following lemma:

Lemma 6. *Suppose that $W \subseteq \Delta^\circ(\Theta)$ is open and that $C \in \mathcal{C}$ is Strongly Positive and Prior Invariant on W. For any $p^*, p_0 \in W$, if $k(p^*) \gg_{psd} 0$ is a lower kernel of C at p^*, then*

$$k(p_0) := \operatorname{diag}(p_0)^{-1}\operatorname{diag}(p^*)k(p^*)\operatorname{diag}(p^*)\operatorname{diag}(p_0)^{-1} \tag{9}$$

is a lower kernel of C at p_0.

Proof. See Appendix F.2. □

Let C_0 be the Prior Invariant direct cost such that $C = \Phi(C_0)$. Consider $\kappa_C(p) := \operatorname{diag}(p)k_C(p)\operatorname{diag}(p)$. Pick arbitrary $p_1, p_2 \in \Delta(\Theta)^\circ$. Since $C \le C_0$, k_C must be the lower kernel of C_0 as well. Then, Lemma 6 implies that $\operatorname{diag}(p_i)^{-1}\kappa_C(p_j)\operatorname{diag}(p_i)^{-1}$ is also a lower kernel of C_0 at p_i, for $i, j \in \{1, 2\}$. Theorem 3 implies that $\operatorname{diag}(p_i)^{-1}\kappa_C(p_j)\operatorname{diag}(p_i)^{-1}$ is a lower kernel of C at p_i, i.e. $\forall i, j \in \{1, 2\}$

$$\begin{aligned}
&\operatorname{diag}(p_i)^{-1}\kappa_C(p_j)\operatorname{diag}(p_i)^{-1} \le_{psd} \operatorname{diag}(p_i)^{-1}\kappa_C(p_i)\operatorname{diag}(p_i)^{-1}\\
\iff &\kappa_C(p_j) \le_{psd} \kappa_C(p_i)\\
\iff &\kappa_C(p_j) = \kappa_C(p_i).
\end{aligned}$$

The first equivalent is due to $\operatorname{diag}(p_i)$ being full rank. The second equivalent is from the fact that i, j can be swapped. Therefore, wlog, $\kappa_C(p) \equiv \kappa$. There are two possible cases:

- *Case 1:* $|\Theta| > 2$. Let $\Theta = \{1, \dots n\}$ and parameterized $p \in \Delta(\Theta)^\circ$ by its first $n-1$ entries,

  $$k_C(p) = \operatorname{diag}(p)^{-1}\kappa\operatorname{diag}(p)^{-1} = \left[\frac{\kappa_{ij}}{p_ip_j} - \frac{\kappa_{in}}{p_i(1-\sum_{\ell=1}^{n-1}p_\ell)} - \frac{\kappa_{jn}}{p_j(1-\sum_{\ell=1}^{n-1}p_\ell)} + \frac{\kappa_{nn}}{(1-\sum_{\ell=1}^{n-1}p_\ell)^2}\right]_{ij}$$

  Since C is UPS and Locally Quadratic, $k_C(p) = \operatorname{Hess}H$ for some $H \in \boldsymbol{C}^2(\Delta(\Theta)^\circ)$. Note that $k_C(p)$ is $\boldsymbol{C}^\infty$ smooth. Therefore, $\forall i \neq j$, by the symmetry of cross-partial derivatives, we have

  $$\begin{aligned}
  &\frac{\partial^3}{\partial p_i^2 \partial p_j}H(p) \equiv \frac{\partial}{\partial p_i}k_C(p)_{ij} \equiv \frac{\partial}{\partial p_j}k_C(p)_{ii}\\
  \iff &-\frac{\kappa_{ij}}{p_i^2p_j} + \frac{\kappa_{in}}{p_i^2(1-\sum_{\ell=1}^{n-1}p_\ell)} - \frac{\kappa_{jn}}{p_j^2(1-\sum_{\ell=1}^{n-1}p_\ell)} + \frac{\kappa_{nn}}{p_i(1-\sum_{\ell=1}^{n-1}p_\ell)^2} \equiv 0\\
  \iff &\kappa_{ij}p_n^2 + \kappa_{jn}p_i^2 - \kappa_{in}(p_i + p_n)p_j =\equiv 0.
  \end{aligned}$$

  By varying p, the final equality holds only when $\kappa_{ij} = \kappa_{in} = \kappa_{jn} = 0$. This implies that κ is a full rank diagonal matrix $\operatorname{diag}(z)$. Then, $k_C p \cdot p = p^{-1} \cdot z$, which contradicts $k_c(p) \cdot p \equiv \mathbf{0}$. Therefore, this case is impossible.

- *Case 2:* $\Theta = \{0,1\}$. $k_c(p) \cdot p \equiv \mathbf{0}$ implies that $\kappa_{11} = \kappa_{22} = -\kappa_{12}$, denoted by some $\gamma > 0$. Therefore, $k_C(p) = \frac{\gamma}{p_1^2(1-p_1^2)} = \gamma \cdot \mathrm{Hess} H^*(p)$. Since C is UPS, $C = \gamma \cdot C_{\mathrm{Wald}}$.

□

References

Angeletos, G.-M. and K. A. Sastry (2024). "Inattentive Economies." In.

Arrow, K. J. (1996). "The Economics of Information: An Exposition." In: *Empirica* 23.2, pp. 119–128.

Arrow, K. J., D. Blackwell, and M. A. Girshick (1949). "Bayes and Minimax Solutions to Sequential Decision Problems." In: *Econometrica*, pp. 213–244.

Banerjee, A., X. Guo, and H. Wang (2005). "On the Optimality of Conditional Expectation as a Bregman Predictor." In: *IEEE Transactions on Information Theory* 51.7, pp. 2664–2669.

Blackwell, D. A. (1951). "Comparison of Experiments." In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, pp. 93–102.

Bloedel, A. W. and W. Zhong (2020). "The Cost of Optimally Acquired Information." Working paper.

Caplin, A. and M. Dean (2013). *Behavioral Implications of Rational Inattention with Shannon Entropy*. Tech. rep. NYU.

— (2015). "Revealed Preference, Rational Inattention, and Costly Information Acquisition." In: *American Economic Review* 105.7, pp. 2183–2203.

Caplin, A., M. Dean, and J. Leahy (2019). "Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy." Working paper.

— (2022). "Rationally inattentive behavior: Characterizing and generalizing Shannon entropy." In: *Journal of Political Economy* 130.6, pp. 1676–1715.

Che, Y.-K. and K. Mierendorff (2019). "Optimal Dynamic Allocation of Attention." In: *American Economic Review* 109.8, pp. 2993–3029.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*. 2nd edition. Wiley-Interscience.

Dean, M. and N. Neligh (2023). "Experimental tests of rational inattention." In: *Journal of Political Economy* 131.12, pp. 3415–3461.

Denti, T. (2022). "Posterior separable cost of information." In: *American Economic Review* 112.10, pp. 3215–3259.

Denti, T., M. Marinacci, and A. Rustichini (2022). "Experimental cost of information." In: *American Economic Review* 112.9, pp. 3106–3123.

Denti, T. and D. Ravid (2023). "Robust Predictions in Games with Rational Inattention." In: *arXiv preprint arXiv:2306.09964*.

Dewan, A. and N. Neligh (2020). "Estimating information cost functions in models of rational inattention." In: *Journal of Economic Theory* 187, p. 105011.

Dillenberger, D., R. V. Krishna, and P. Sadowski (2023). "Subjective information choice processes." In: *Theoretical Economics* 18.2, pp. 529–559.

FDA, U. (2019). *Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry.

Fehr, E. and A. Rangel (2011). "Neuroeconomic foundations of economic choice—recent advances." In: *Journal of economic perspectives* 25.4, pp. 3–30.

Fudenberg, D., P. Strack, and T. Strzalecki (2018). “Speed, accuracy, and the optimal timing of choices.” In: *American Economic Review* 108.12, pp. 3651–3684.

Gentzkow, M. and E. Kamenica (2014). “Costly Persuasion.” In: *American Economic Review, Papers and Proceedings* 104.5, pp. 457–462.

Georgiadis, G. and B. Szentes (2020). “Optimal monitoring design.” In: *Econometrica* 88.5, pp. 2075–2107.

Hébert, B. and J. La’O (2023). “Information acquisition, efficiency, and nonfundamental volatility.” In: *Journal of Political Economy* 131.10, pp. 2666–2723.

Hébert, B. and M. Woodford (2021). “Neighborhood-based information costs.” In: *American Economic Review* 111.10, pp. 3225–3255.

— (2023). “Rational inattention when decisions take time.” In: *Journal of Economic Theory* 208, p. 105612.

Huffman, D. A. (1952). “A method for the construction of minimum-redundancy codes.” In: *Proceedings of the IRE* 40.9, pp. 1098–1101.

Johari, R. et al. (2022). “Always valid inference: Continuous monitoring of a/b tests.” In: *Operations Research* 70.3, pp. 1806–1821.

Kuczma, M. (2009). “An introduction to the theory of functional equations and inequalities.” In: *(No Title)*.

Li, Y. (2022). “Selling data to an agent with endogenous information.” In: *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 664–665.

Liang, A. and X. Mu (2020). “Complementary Information and Learning Traps.” In: *Quarterly Journal of Economics* 135.1, pp. 389–448.

Liang, A., X. Mu, and V. Syrgkanis (2022). “Dynamically Aggregating Diverse Information.” In: *Econometrica* 90.1.

Maćkowiak, B., F. Matějka, and M. Wiederholt (2023). “Rational inattention: A review.” In: *Journal of Economic Literature* 61.1, pp. 226–273.

Matějka, F. and A. McKay (2015). “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model.” In: *American Economic Review* 105.1, pp. 272–298.

Mensch, J. (2018). *Cardinal Representations of Information*. Tech. rep.

— (2021). “Rational inattention and the monotone likelihood ratio property.” In: *Journal of Economic Theory* 196, p. 105284.

Miao, J. and H. Xing (2024). “Dynamic discrete choice under rational inattention.” In: *Economic Theory* 77.3, pp. 597–652.

Morris, S. and P. Strack (2019). “The Wald Problem and the Equivalence of Sequential Sampling and Ex-Ante Information Costs.” Working paper, MIT and Yale University.

Morris, S. and M. Yang (2019). *Coordination and Continuous Stochastic Choice*. Tech. rep. Princeton University.

Moscarini, G. and L. Smith (2001). “The Optimal Level of Experimentation.” In: *Econometrica* 69.6, pp. 1629–1644.

Mu, X. et al. (2021). “From Blackwell dominance in large samples to Rényi divergences and back again.” In: *Econometrica* 89.1, pp. 475–506.

Müller-Itten, M., R. Armenter, and Z. R. Stangebye (2023). “Rational Inattention via Ignorance Equivalence.” In.

Myatt, D. P. and C. Wallace (2012). “Endogenous Information Acquisition in Coordination Games.” In: *Review of Economic Studies* 79, pp. 340–374.

Oliveira, H. de (2019). “Axiomatic Foundations for Entropic Costs of Attention.” Working Paper, Penn State.

Oliveira, H. de et al. (2017). “Rationally Inattentive Preferences and Hidden Information Costs.” In: *Theoretical Economics* 12.2, pp. 621–654.

Pomatto, L., P. Strack, and O. Tamuz (2023). “The cost of information: The case of constant marginal costs.” In: *American Economic Review* 113.5, pp. 1360–1393.

Ratcliff, R. (1978). “A theory of memory retrieval.” In: *Psychological review* 85.2, p. 59.

Ratcliff, R. and G. McKoon (2008). “The diffusion decision model: theory and data for two-choice decision tasks.” In: *Neural computation* 20.4, pp. 873–922.

Ravid, D. (2020). “Ultimatum Bargaining with Rational Inattention.” In: *American Economic Review* forthcoming.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Royden, H. and P. Fitzpatrick (2010). “Real Analysis.” In.

Shannon, C. E. (1948). “A Mathematical Theory of Communication.” In: *The Bell System Technical Journal* 27, pp. 379–423.

Sims, C. A. (2003). “Implications of rational inattention.” In: *Journal of Monetary Economics* 50.3, pp. 665–690.

— (2010). “Rational Inattention and Monetary Economics.” In: *Handbook of Monetary Economics*. Vol. 3. Elsevier, pp. 155–181.

Wald, A. (1945). “Sequential Tests of Statistical Hypotheses.” In: *The Annals of Mathematical Statistics* 16.2, pp. 117–186.

— (1947). “Foundations of a General Theory of Sequential Decision Functions.” In: *Econometrica*, pp. 279–313.

Wong, Y. F. (2023). “Dynamic monitoring design.” In: *Available at SSRN 4466562*.

Woodford, M. (2012). *Inattentive Valuation and Reference-Dependent Choice*. Tech. rep. Columbia University.

Zhong, W. (2022). “Optimal dynamic information acquisition.” In: *Econometrica* 90.4, pp. 1537–1582.