The Cost of Optimally Acquired Information*

[Click here for most recent version.]

Alexander W. Bloedel† (Job Market Paper)

Weijie Zhong‡

First version: November 15, 2020
This version: June 24, 2021

Abstract

This paper develops a theory for the expected cost of optimally acquired information when information can be acquired sequentially. We study the "reduced-form" *Indirect Cost* functions for information generated by sequential minimization of a "primitive" *Direct Cost* function. The class of Indirect Costs is characterized by a recursive condition called *Sequential Learning-Proofness*. This condition is inconsistent with *Prior-Invariance*: Indirect Costs must depend on the decision-maker's prior beliefs.

We show that Sequential Learning-Proofness provides partial optimality foundations for the Uniformly Posterior Separable (UPS) cost functions used in the rational inattention literature: a cost function is UPS if and only if it is an Indirect Cost that (i) satisfies a mild regularity condition or, equivalently, (ii) is generated (only) by Direct Costs for which the optimal sequential strategy involves observing only Gaussian diffusion signals. We characterize the unique UPS cost function that is generated by a Prior-Invariant Direct Cost; it exists only when there are exactly two states.

We also propose two specific UPS cost functions based on additional optimality principles. We introduce and characterize *Total Information* as the unique Indirect Cost that is *Process-Invariant* when information can be decomposed both sequentially and "simultaneously": it is uniquely invariant to the "merging" and "splitting" of experiments. Under regularity conditions, *Mutual Information* is the unique Indirect Cost that is *Compression-Invariant* when aspects of the state space can be "freely ignored": it is uniquely invariant to the to the "merging" and "splitting" of states. We argue that Total Information and Mutual Information represent the normatively ideal costs of, respectively, "producing" and "processing" information.

JEL Codes: D80, D81, D83

Keywords: Information acquisition, cost of information, dynamic information acquisition, sequential sampling, rational inattention.

*This paper combines and supersedes the solo projects Bloedel (2020b) and Zhong (2020). We thank Doug Bernheim, Ben Hébert, Johannes Hörner, Ravi Jagadeesan, R. Vijay Krishna, Teddy Mekonnen, Paul Milgrom, Agathe Pernoud, Collin Raymond, and Ilya Segal for helpful feedback, and seminar participants at Florida State and Stanford for useful comments. Bloedel is especially grateful to Ilya Segal, collaboration with whom inspired some ideas developed in this paper. Bloedel gratefully acknowledges financial support from the Forman Fellowship through a grant to the Stanford Department of Economics, and from the B.F. Haley and E.S. Shaw Fellowship for Economics through a grant to the Stanford Institute for Economic Policy Research.

†Department of Economics, Stanford University. Email: abloedel@stanford.edu.

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

1 Introduction

1.1 Motivation and Framework

Economic models typically assume that individuals are exogenously endowed with information. This is plainly an abstraction. When faced with uncertainty, rational individuals *choose* what information to acquire by trading off its decision-relevant value against its cost. Consumers expend effort investigating the quality of products before making purchases. Firms spend billions of dollars each year in market research and R&D. Financial analysts are paid to conduct fundamental research on investment opportunities. The view that information is a costly choice variable that responds to incentives has proven important for understanding a variety of economic phenomena[1] and for designing new economic institutions.[2]

While the *value* of information is well-understood and often unambiguous,[3] there is little consensus on how to develop a general theory for its *cost*. This has proved challenging, in part, because "there is no general way of defining units for information" (Arrow (1996, p. 120)). Consequently, as Chade and Schlee (2002, p. 443) put it: "Unfortunately, we know precious little about how to choose functional forms for the production of information."[4]

In this paper, we develop a theory for the cost of information based on the premise that individuals flexibly choose not only *what information to learn*, but also the *optimal way in which to acquire it*. We are motivated by the following two desiderata for a general theory of information cost.

First, it should account for individuals' ability to acquire any given piece of information in a variety of different ways. For example, a pharmaceutical company aiming to test the efficacy of a new drug to a pre-specified confidence level has the option to run a single large clinical trial, multiple smaller trials simultaneously (in batch), or multiple smaller trials run sequentially (with the results of early trials potentially informing the design of later ones). It is typically optimal to acquire information in a piecemeal fashion for cost-smoothing purposes, and to do so sequentially for the option value that sequentiality provides (i.e., early observations are informative about both the payoff-relevant state and how to economize on continuation costs). This poses a challenge because, for reasons of tractability and portability, it is often necessary to model information acquisition in "reduced form" as a one-shot choice.

Second, it should provide a unified language for reasoning about the cost of information across different contexts. A key challenge is that the technology by which individuals acquire information

[1] An incomplete list of applications includes macroeconomic dynamics (Maćkowiak and Wiederholdt (2009); Flynn and Sastry (2020)), financial portfolio choice (Mondria (2010); Nieuwerburgh and Veldkamp (2010); Kacperczyk et al. (2016)), organizational structure (Dessein et al. (2016)), monopoly pricing (Matějka (2015); Ravid (2019, 2020)), oligopoly pricing (Matějka and McKay (2012)), coordination games (Yang (2015); Morris and Yang (2019); Denti (2019)), beauty contest games (Myatt and Wallace (2012); Hébert and La'O (2020)), and individuals' stochastic choice behavior (Matějka and McKay (2015); Caplin et al. (2019a); Köszegi and Matějka (2020)).

[2] An incomplete list of applications includes the design of optimal contracts for research and innovation (Yoder (2019); Rappaport and Somma (2017)), financial contracts (Yang (2020); Yang and Zeng (2019)), allocation mechanisms (Gleyze and Pernoud (2020); Mensch (2020)), and information disclosure policies (Gentzkow and Kamenica (2014); Bloedel and Segal (2020)).

[3] The well-known Blackwell (1951) theorem establishes that one Blackwell experiment is more informative than another if and only if the former yields higher expected utility in *all* decision problems. For a Bayesian decision-maker facing a *given* decision problem, the value of a Blackwell experiment is the expected utility gain it yields relative to having no information beyond the prior belief (e.g., Azrieli and Lehrer (2008); Frankel and Kamenica (2019)).

[4] This is a long-standing sentiment: Arrow (1985, p. 304) asserts that "it is an important and incompletely explored part of decision theory in general to formulate reasonable cost functions for information structures," while Pomatto et al. (2019, p. 1) observe more recently that "modeling the cost of producing information has remained an unsolved problem."

— or even what "acquiring information" means — depends on the context under study. For instance, to paraphrase Sims (2010, p. 161), the cost of *producing* new information (e.g., expending physical resources to conduct a clinical trial) need not bear any relation to the cost of *processing* already-available information (e.g., expending mental energy to read a report that summarizes the trial's findings).

Our framework considers a Bayesian decision-maker (DM) who aims to acquire information about an uncertain *state* distributed according to some *prior belief*. We view information acquisition as a two-stage process. In the first stage, DM decides what to learn, modeled as a choice of which *(Blackwell) experiment* (i.e., state-contingent distribution of *signals*) to obtain. In the second stage, DM decides how to acquire her chosen *target experiment*. Our central assumption is that information is available in many forms, so that DM need not acquire her target experiment in one shot; instead, she may *sequentially replicate* it through a strategy for sequentially acquiring "sub-experiments" that each may be less informative than her target, but collectively are sufficient for it. We impose no *a priori* parametric restrictions on DM's strategy space.[5] DM *optimizes* over such Sequential Replications with the goal of minimizing the total expected cost of replicating the target experiment, given her prior belief.

We pursue two parallel goals: (i) to characterize the *information cost functions* (i.e., cost functions over experiments and prior beliefs) that capture the second-stage sequential optimization in "reduced form," and (ii) for each such "reduced form" cost function, to characterize the "primitive" cost functions that could have generated it through sequential optimization.

To the first end, call a cost function *Sequential Learning-Proof (SLP)* if it cannot be reduced by any two-step Sequential Replication. We show that SLP cost functions are precisely the "fixed points" of the full sequential optimization process. Thus, they constitute the maximal class that satisfies our first desideratum: any non-SLP cost function has features that DM would "optimize away," and is therefore not "rationalizable." In this paper, we characterize the full class of SLP cost functions, as well as particular functions in this class that have additional normatively appealing properties.

To the second end, call the "primitive" cost function that DM uses at each step of the sequential optimization her *Direct Cost*. Given any Direct Cost, sequential optimization gives rise to a value function that we call DM's *Indirect Cost*. Every SLP cost function is its own Indirect Cost and, conversely, every Indirect Cost is SLP. However, each SLP cost function can be generated from many Direct Costs: only certain properties of the latter are preserved under optimization. For specific SLP cost functions that we consider in this paper, we also characterize the full set of Direct Costs that could have generated them. By imposing minimal assumptions on DM's Direct Cost and relying on the "meta-axiom" of optimality to impose discipline on her Indirect Cost, our framework achieves the robustness demanded by our second desideratum.

Implications for Rational Inattention. Our approach has implications for ongoing debates concerning the rational inattention (RI) model of Sims (1998, 2003), in which the cost of an experiment is given by the *Mutual Information* (i.e., expected reduction of Shannon entropy) between the uncertain state and signal. It is well known that Mutual Information approximates the expected length of

[5] However, it is important to note that many of our results do not rely on DM having complete flexibility, and extend to the case in which her strategy space is restricted. We revisit this point throughout the paper.

an optimally-encoded sequence of bits (i.e., answers to binary yes/no questions) needed to describe the observed signal; in our language, it is (approximately) the Indirect Cost generated by the specific Direct Cost that assigns equal cost to all bits.[6] As articulated by Sims (2010, p. 161), this foundation renders the RI model a sensible theory of information processing, but not production, costs. It therefore satisfies only our first desideratum.

Nonetheless, the RI model has become the benchmark theory of costly information in a range of applications, including some where its information-theoretic foundations have no clear relevance.[7] This has led many authors to critique fundamental aspects of the RI model and, towards our second desideratum, to propose a number of alternatives to the Mutual Information cost function (e.g., Caplin et al. (2019b); Pomatto et al. (2019); Hébert and Woodford (2020a)). A standing issue is that many of these proposed alternatives, unlike Mutual Information, do not have clear optimality foundations, and so may violate our first desideratum.

One contribution of this paper is to systematically provide such optimality foundations where they exist, and to point out where they do not. This facilitates a "meta-commentary" on the three leading critiques of the RI model — which we call the *Prior-Invariance*, *Returns-to-Scale*, and *Perceptual Distance* critiques — that clarifies the extent to which each is consistent with the idea of optimal information acquisition in various contexts. In doing so, we identify inherent modeling tradeoffs that applied researchers must confront.

1.2 Overview of Main Results

The paper consists of three main sets of characterization results, beginning with the most general and ending with the most specific. The first set provides a general characterization of all SLP cost functions and studies implications thereof. The second set provides optimality foundations for the (Uniformly) Posterior Separable cost functions that have become the default for modeling flexible information acquisition in the literature. The third set proposes two specific (Uniformly) Posterior Separable cost functions that we argue should be used in particular applied settings.

(1) Characterization of SLP and Indirect Costs. Our first set of results characterizes the full class of SLP and Indirect Cost functions. We show (Theorem 1) that a cost function is SLP if and only it satisfies two properties: (i) it is increasing in the Blackwell informativeness of an experiment and (ii) it exhibits *Preference for One-Shot Learning*. We also show that the Indirect Cost derived from *any* Direct Cost is SLP, and hence satisfies these properties.

Property (i) captures DM's ability to freely dispose of acquired information. Property (ii), Preference for One-Shot Learning, characterizes the additional restrictions imposed by sequential optimality: for any target experiment and prior belief, it is weakly cheaper to acquire information in one shot than via any two-step Sequential Replication. This latter condition implies that no SLP cost function is reducible by engaging in mixed strategies (i.e., is Randomizaton Averse), so that every SLP cost is

[6] See Cover and Thomas (2006, Ch. 10). The information-theoretic justification for Mutual Information is premised on the idea that costs can be averaged (or "amortized") across a large number of independent copies of the same information acquisition problem solved together in parallel (known as "block coding"). Due to integer problems arising from the indivisibility of bits, Mutual Information only approximates the minimum expected bit length in any individual problem.

[7] All of the papers cited in footnotes 1 and 2 above employ the RI model or generalizations thereof. Additional applications of the RI model are surveyed in Maćkowiak et al. (2018).

in the class of "canonical" cost functions that represent optimal *one-shot* information acquisition (cf. de Oliveira et al. (2017)). It also implies that every SLP cost function is linear in the probability of acquiring a given experiment (i.e., is Dilution Linear), which is a key axiom introduced by Pomatto et al. (2019). However, unlike these two necessary conditions, Preference for One-Shot Learning also constrains the cost of a given experiment *across different priors*.

An important implication (Proposition 1) is that, under suitable regularity conditions, no non-trivial SLP cost function is *Prior-Invariant* (i.e., independent of DM's prior belief). A leading critique of the RI model and many of its generalizations, which we call the *Prior-Invariance Critique*, is that the cost of producing (rather than processing) information should be Prior-Invariant.[8] The costs of information production, the argument goes, should not depend on prior beliefs because they correspond to the expenditure of "physical" or monetary resources. Mutual Information and many of its generalizations fail this criterion. We show that this reasoning relies on the implicit assumption that DM is constrained to one-shot information acquisition. Without this constraint, the optimal strategy is generically sequential and, under the criterion of *expected* cost minimization, naturally depends on DM's prior belief, leading to a prior-dependent Indirect Cost.

(2) Foundations for (Uniform) Posterior Separability. Our second set of results provides optimality foundations for the leading generalizations of the RI model. While the literature has moved beyond the Shannon entropy functional form, it almost universally maintains the assumption that the cost of information is measured by the expected reduction of some function of posterior beliefs.

Formally, an experiment is a mapping $\sigma : \Theta \to \Delta(S)$ from states $\theta \in \Theta$ to distributions over signals $s \in S$. Each prior belief $p \in \Delta(\Theta)$ and experiment σ induce, via Bayes' Rule, a distribution over DM's random posterior belief $q \in \Delta(\Theta)$ denoted by $\pi_{\langle \sigma | p \rangle}$. A cost function $C(\sigma \mid p)$ is called *Posterior Separable* if for each (full support) prior belief p there exists a convex *potential function* $F(\cdot \mid p)$ on posterior beliefs such that

$$C(\sigma \mid p) = \mathbb{E}_{\pi_{\langle \sigma | p \rangle}} [F(q \mid p) - F(p \mid p)], \tag{1}$$

and is *Uniformly Posterior Separable (UPS)* if the potential function is independent of the prior (Caplin et al. (2019b)). Mutual Information is the particular UPS cost function for which $F(q) = \sum_\theta q_\theta \log(q_\theta)$, the negative of Shannon entropy. For reasons of tractability, (Uniformly) Posterior Separable cost functions have become the default in applied models of flexible information acquisition.

We produce three characterization results: one for the class of Posterior Separable costs, and two that provide complementary perspectives on the UPS class. They are based on the familiar idea that, in continuous time, DM can engage in two main types of Sequential Replication: learning by *Poisson signals*, which arrive infrequently but are potentially quite informative, and learning via *Gaussian (diffusion) signals*, which arrive frequently but are only incrementally informative (cf. Zhong (2019); Hébert and Woodford (2020b)).[9]

Lemma 2 characterizes the class of Posterior Separable cost functions as the Indirect Costs arising from a *restricted* optimization problem in which DM is constrained to learn via direct Poisson signals and in which her Direct Cost function is *Locally Linear*, meaning that the cost of acquiring

[8] Versions of this view are expressed by Woodford (2012), Gentzkow and Kamenica (2014), Mensch (2018), Denti et al. (2020), and Rustichini (2020), among others. Nimark and Sundaresan (2019) and Ravid (2020) explore the importance of Prior-Invariance in economic applications.

[9] Our framework is cast in discrete time, but allows for taking the continuous-time limit.

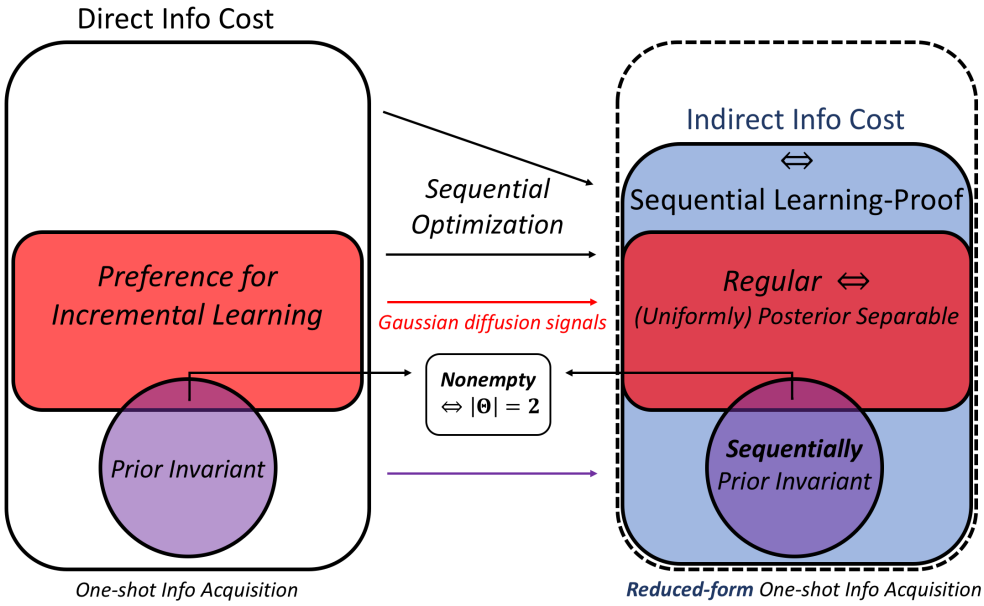

Figure 1: Characterization of (Uniformly) Posterior Separable cost functions.

infrequently-arriving Poisson signals is approximately Posterior Separable. We show that Local Linearity of corresponds to a standard "local continuous directional differentiability" condition on DM's Direct Cost. The constraints on DM's strategy space are necessary, as the Posterior Separable class is large enough to include many non-SLP functions. However, an implication of this result (Proposition 2) is that an SLP cost function is Posterior Separable if and only if it is Locally Linear.

By contrast, every UPS cost function is SLP. Indeed, the UPS class is characterized by the property of *Indifference to Sequential Learning* (Lemma 1): for each target experiment and prior belief, *all* Sequential Replications are equally costly. How stringent is this requirement? We provide two complementary perspectives.

We first establish (Theorem 2) that an SLP cost function is UPS if and only if it is *Regular*, i.e., Locally Linear with a suitably differentiable Posterior Separable approximation. While not completely innocuous, Regularity is a relatively mild smoothness assumption that is satisfied by nearly all information cost functions commonly used in applications. If one is willing to adopt such assumptions, then the SLP and UPS classes exactly coincide. The proof of this result generalizes known characterizations of the family of "Bregman divergences" (cf. Banerjee et al. (2005)).

However, we then show (Theorem 3) that requiring DM's Indirect Cost to be UPS amounts to making strong assumptions on her Direct Cost. Formally, a Direct Cost generates a UPS Indirect Cost if and only if the former exhibits *Preference for Incremental Learning* (see Figure 1). This imposes two requirements on DM's Direct Cost: for each target experiment and prior belief, (i) it is cheaper to sequentially replicate with Gaussian diffusion signals than to acquire information in one shot, and (ii) DM is indifferent among all such Gaussian replications. Our characterization shows that this limited form of "superadditivity" precisely counteracts the "subadditivity" of Preference for One-Shot Learning to achieve the global "additivity" that defines the UPS class. An important consequence of Theorem 3 is that, under Preference for Incremental Learning, DM's Indirect Cost

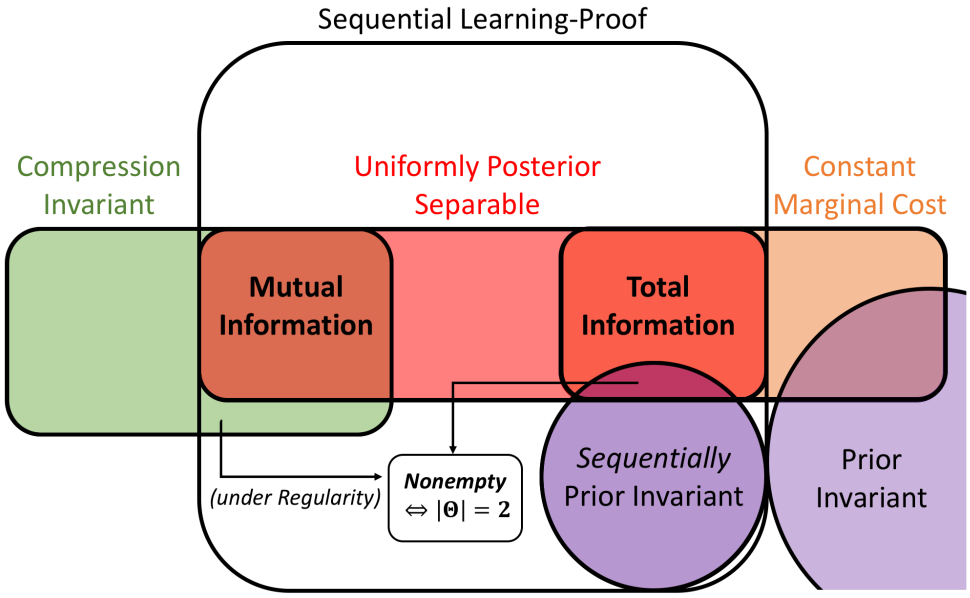

Figure 2: Hierarchy of SLP cost functions.

is completely characterized by a "local quadratic approximation" of her Direct Cost that represents the cost of an asymptotically uninformative Gaussian experiment. This quadratic approximation is simple to derive in many cases of interest, facilitating a tractable description of the family of Direct Costs consistent with a given UPS cost function.

While we view the main contribution of Theorem 3 as establishing the necessity of Preference for Incremental Learning, it is noteworthy that the sufficiency direction alone implies related results of Morris and Strack (2019) and Hébert and Woodford (2020b), which partially characterize UPS cost functions as Indirect Costs arising from more restricted optimization problems (see Subsection 4.6).

An important implication of Theorem 3 is that, generically, no UPS Indirect Cost is *Sequentially Prior-Invariant*, i.e., generated by a Prior-Invariant Direct Cost function. We refer to this strengthening of the Prior-Invariance Critique as the *Sequential Prior-Invariance Critique* of the UPS model. We provide (Proposition 3) necessary and sufficient conditions for a cost function to be both UPS and SPI: (i) the state space must be binary and (ii) the cost function must be Total Information with symmetric coefficients. We also characterize the full set of Prior-Invariant Direct Costs that generate this function.[10]

(3) Two Specific (and Mutually Exclusive) Proposals. Our final set of results characterizes two specific (classes of) SLP cost functions — each of which satisfies a different normatively ideal "invariance" property — as well as the Direct Costs that generate them (see Figure 2). An important implication of our characterizations is that these invariance properties are (generically) mutually exclusive — and also inconsistent with the criterion of Sequential Prior-Invariance — so that applied researchers must necessarily make substantive tradeoffs when modeling costly information acquisi-

[10] The two-state symmetric Total Information cost function is precisely the two-state Wald cost function introduced by Morris and Strack (2019). Proposition 3 can therefore be viewed as the maximal generalization of their Wald cost characterization.

tion.

(i) Information Production: Total Information and Process-Invariance. How should one model the cost of producing new information? We propose that the normative ideal is the new *Total Information* cost function

$$C_{TI}(\sigma \mid p) = \sum_{\theta,\theta'} p_\theta \gamma_{\theta,\theta'} D_{KL}(\sigma_\theta \mid \sigma_{\theta'}), \tag{2}$$

where $D_{KL}(\sigma_\theta \mid \sigma_{\theta'})$ is the Kullback-Leibler divergence between the θ- and θ'-contingent signal distributions, and $\gamma_{\theta,\theta'} \geq 0$ is a coefficient representing the marginal cost of distinguishing between states θ and θ'. Theorem 4 characterizes Total Information as the unique SLP cost function with *Constant Marginal Cost*: the cost of acquiring two conditionally independent experiments is the same as the sum of their costs, holding fixed the prior belief.

To understand this condition, note that there are two distinct ways to replicate a target experiment: *sequentially* or *simultaneously*. During Sequential Replication, on which we focus, DM conditions her continuation strategy on earlier signal realizations and uses her updated posterior belief to evaluate her continuation cost. During Simultaneous Replication, by contrast, DM must run all sub-experiments at once before observing any signals — hence the sub-experiments must generate signals that are independent conditional on the state — and evaluates the cost of each under the same prior belief. Simultaneous Replication is common in practice and, when DM's cost function is prior-dependent, may be preferable to Sequential Replication.[11]

Uniform Posterior Separability captures indifference with respect to Sequential Replication, while Constant Marginal Cost captures indifference with respect to Simultaneous Replication. Theorem 4 also shows that Total Information is UPS, meaning that it is also uniquely *Axiom 13*: it assigns the same cost to *all* replications of any given experiment. Put differently, "merging" or "splitting" experiments does not affect costs: only the totality of information produced, not the process by which it is acquired, matters.

Notably, special cases of Total Information have been introduced (in the binary-state case) as the *Wald* cost of Morris and Strack (2019) and (in the continuous-state limit) as the *Fisher Information* cost of Hébert and Woodford (2020a), albeit with substantively different foundations. Our analysis provides a unified perspective on, and new justification for, these specific proposals.

Our characterization of Total Information formally builds on Pomatto et al.'s (2019) characterization of the *Log-Likelihood Ratio (LLR)* cost function, which has a functional form similar to that of Total Information, satisfies a version of Constant Marginal Cost, and also aims to represent the cost of information production.[12] However, the LLR cost is Prior-Invariant and fails to be SLP. We argue that the "meaning" of Constant Marginal Cost is fundamentally different when information can be acquired sequentially than when it is restricted to be one-shot, as is implicit in Pomatto et al. (2019). In particular, we show that the Indirect Cost generated by the LLR cost function neither has Constant Marginal Cost nor is UPS; moreover, under suitable regularity conditions, *no* Direct Cost

[11] Most market research, political polling, and A/B testing is conducted via Simultaneous Replication (e.g., each polled voter can be viewed as drawing a conditionally independent sample). Simultaneous Replication also takes place in firms when information acquisition is decentralized among multiple employees.

[12] The LLR cost function is obtained by replacing each $p_\theta \gamma_{\theta,\theta'}$ term in (2) with a prior-independent coefficient $\beta_{\theta,\theta'}$.

with Constant Marginal Cost (except for Total Information itself) generates Total Information as its Indirect Cost.

Instead, we demonstrate that sufficiently flexible optimal information production generally leads to an Indirect Cost with *Decreasing Marginal Cost*: costs are not reducible by Simultaneous Replication. In an extension of our framework, we show that *Unrestricted Learning-Proofness (ULP)* — a condition that implies both SLP and Decreasing Marginal Cost — characterizes the class of cost functions that are robust to *both* Sequential Replication and Simultaneous Replication.[13] Even when Simultaneous Replication is not permitted, we show that an SLP cost function necessarily has Decreasing Marginal Cost if it is *Prior-Concave* (i.e., concave in DM's prior belief for each fixed experiment), which represents the ability to "freely dispose of already-available information." For a UPS cost function, Decreasing Marginal Cost, ULP, and Prior-Concavity are all equivalent properties. Total Information and Mutual Information satisfy these properties, but many commonly-used UPS cost functions do not; we fully characterize the set of UPS costs that do.

Our analysis speaks to what we call the *Returns-to-Scale Critique* of the Mutual Information cost function, which argues that its Decreasing Marginal Cost leads to unreasonable "corner solutions" in certain information acquisition problems. We contribute by showing that any sufficiently flexible optimization procedure necessarily results in an Indirect Cost with strictly Decreasing Marginal Cost, except in the case of Total Information.

(ii) Information Processing: Mutual Information and Compression-Invariance. How should one model the cost of processing already-available information? We propose that Mutual Information represents the normative ideal. Theorem 5 characterizes Mutual Information as the unique Bounded UPS cost function that is either *Weakly Compression-Invariant* or *Compression Monotone*; a fortiori, it is the unique Bounded and Regular SLP cost function satisfying either of these properties. These conditions capture the idea that DM is able to "freely ignore" aspects of the state space that she finds "irrelevant."

For example, suppose DM aims to learn about a political candidate's platform, which may be *l*eft-, *c*enter-, or *r*ight-leaning. Weak Compression-Invariance demands that if DM's target experiment is informative *only* about whether the candidate is right-leaning — e.g., it determines whether the true state is in $\{l, c\}$ or $\{r\}$ — then her cost remains the same when prior probability mass is shifted *within* the event $\{l, c\}$. Intuitively, "splitting" or "merging" the states l and c should not affect DM's cost when her (fixed) target experiment already ignores any distinction between them. Compression Monotonicity demands that learning about "coarser" events is cheaper. In particular, the cost of running experiment σ given prior p should be no smaller than the cost of running the experiment $\hat{\sigma}$ for which $\hat{\sigma}(s \mid l) = \hat{\sigma}(s \mid c) = p(l \mid \{l, c\}) \cdot \sigma(s \mid l) + p(c \mid \{l, c\}) \cdot \sigma(s \mid c)$ and $\hat{\sigma}(s \mid r) = \sigma(s \mid r)$ given the same prior. Intuitively, "merging" states l and c while generating the same information (i.e., conditional signal distributions) about the events $\{l, c\}$ and $\{r\}$ should not increase costs; in any decision problem where l and c are payoff-equivalent, DM should be able to freely ignore any distinction between

[13] More precisely, in Appendix A.1 we extend our baseline framework by allowing DM to engage in *Unrestricted Replication* in which information is acquired sequentially but she is able to "freely store and recall" previously-acquired information. What we have called Sequential Replication corresponds to the case of no storage (i.e., all acquired information is "immediately seen") while Simultaneous Replication corresponds to the case of full storage (i.e., no acquired information is "seen" until the end of the acquisition process).

them.[14]

Notice that Weak Compression-Invariance involves changing the prior while holding the target experiment fixed, while Compression Monotonicity does precisely the opposite. Caplin et al. (2019b) have shown that (the revealed-preference analogue to) the *combination of* these two conditions uniquely pins down Mutual Information within the UPS class. Perhaps surprisingly, Theorem 5 establishes that *either* condition separately suffices, while also admitting a comparatively elementary proof that elucidates the connection between these conditions and a "recursivity" property known to characterize Shannon entropy (e.g., Fadeev (1956)). We also provide optimality foundations for these conditions through an augmented optimization problem for DM in which she can freely "pre-garble" the state in certain ways before running an experiment, as envisioned by Shannon (1958) and is implicit in information-geometric notions of "invariance" (e.g., Amari (2016)). We show how allowing for richer forms of pre-garbling gives rise to stronger notions of "compression invariance," including that considered by Caplin et al. (2019b). Our analysis speaks to the prominent *Perceptual Distance Critique* of the Mutual Information cost function, which argues that its implication that it is equally costly to distinguish between "nearby" states (e.g., l and c in the above example) and "distant" states (e.g., l and r) runs counter to experimental evidence on human perception and choice.

Outline. The rest of the paper is organized as follows. We survey related literature in Subscetion 1.3. Section 2 presents the framework. Section 3 presents our general characterizations of SLP and Indirect Cost functions. Section 4 presents our characterization results for (Uniformly) Posterior Separable cost functions. Sections 5 and 6 present our characterizations of, respectively, Total Information and Mutual Information. Section 7 discusses applications and concludes. All omitted proofs, and a number of omitted results, are contained in the appendices.

1.3 Related Literature

Sequential Sampling Foundations for Cost of Information. As noted above, two related papers, Morris and Strack (2019) and Hébert and Woodford (2020b), share our focus on providing micro-foundations for information cost functions as reduced form representations of the expected cost of optimal sequential sampling. We discuss the relation between these papers and ours in depth in Subsections 4.6 and 5.2.1.

Axiomatic Foundations for Cost of Information. Several recent papers characterize information cost functions using normatively appealing axioms. Mensch (2018) and Denti et al. (2020) develop general theories of Prior-Invariant cost functions, which we show are not SLP. Hébert and Woodford (2020a) characterize a subclass of UPS cost functions called "neighborhood-based" costs with a set of axioms capturing the idea that "nearby" states are more costly to distinguish that "distant" states.[15] In the continuous-state limit, they show that these cost functions converge to the *Fisher Information* cost function, which they propose as an alternative to Sims' (2003) Mutual Information cost function. As discussed more fully in Subsection 5.2.1, our Total Information cost function is in

[14] Such "pre-garbling" of states does not generally lead to a Blackwell ranking between the experiments σ and $\hat{\sigma}$.

[15] Walker-Jones (2020) presents a different axiomatization for a subclass of neighborhood-based cost functions by suitably relaxing the Shannon (1948) axioms for entropy.

the neighborhood-based class and generalizes the Fisher Information cost function. Finally, as noted above, our Total Information cost function is closely related to the LLR cost function introduced by Pomatto et al. (2019), which is Prior-Invariant and therefore not SLP; our analysis of sequential information acquisition leads to a significant re-interpretation of their key Constant Marginal Cost axiom. We discuss the relation to their paper in depth in Subsection 5.

Dynamic Information Acquisition. Our focus on sequential cost-minimization connects our paper to the literature that studies the optimal way to dynamically acquire information before taking an action. However, the goals and methods of this paper differ fundamentally from those of that literature. While we characterize the implications of sequential optimization for "reduced form" Indirect Cost functions in settings without delay or discounting costs, that literature aims to explicitly characterize optimal learning dynamics in the presence of time costs and, often, subject to restrictions on DM's feasible set of sampling strategies.

In the seminal "sequential sampling" framework of Wald (1945, 1947) and Arrow et al. (1949), DM controls only the *duration* of learning by deciding when to stop observing conditionally independent draws from a fixed experiment, each with equal cost. In an extension of that framework, Moscarini and Smith (2001) allow DM to dynamically control the *speed* of learning by paying a cost to improve the precision of conditionally independent Gaussian signals. Recent work has studied settings in which DM dynamically controls the *direction* of learning by, for instance, choosing the skewness of Poisson experiments in a binary-state setting (Che and Mierendorff (2019)).[16] Woodford (2016), Zhong (2019, 2017), and Hébert and Woodford (2020b) are closest to the present paper in the sense that they endow DM with near complete flexibility in her choice of sequential strategy. However, in contrast to our work, these papers *assume* that DM's Direct Cost is derived from a (Uniformly) Posterior Separable cost function and focus on providing detailed characterizations of the optimal learning dynamics generated by time discounting, additive delay costs, constraints on the speed of learning, or inter-temporal cost smoothing (when the Direct Cost is convex in the underlying UPS measure).[17]

Revealed Preference. A burgeoning literature characterizes the testable implications of theories of costly information acquisition using a revealed preference approach. These studies take the perspective of an analyst who observes DM's choice behavior in various decision problems but not her information acquisition strategy or cost function, which must be identified from choice behavior. Caplin and Dean (2015) and de Oliveira et al. (2017) characterize behavior that is consistent with some "canonical" (i.e., Blackwell monotone and Randomizaton Averse) cost function, and show that a unique such cost function can be identified from sufficiently rich choice data.[18] Bloedel (2020a) characterizes behavior that is consistent with optimal dynamic information acquisition when the

[16] Other recent contributions in this vein allow DM to dynamically control which dimension of the state to learn about in settings with a 2×2 state space (Nikandrova and Pancs (2018); Mayskaya (2019); Ke and Villas-Boas (2019)), which of multiple sources to sample from in a multi-dimensional Gaussian setting (Liang et al. (2020, 2019); Liang and Mu (2020)), or the order in which to consult different sources in a sequential search setting (Doval (2018)).

[17] Without these additional features, DM would be indifferent among all sequential strategies given a UPS Direct Cost. Steiner et al. (2017) and Ravid (2019) study models in which DM has a Mutual Information Direct Cost function, but in which nontrivial dynamics emerge because the payoff-relevant state evolves over time.

[18] See also Dillenberger et al. (2014), Lu (2016), Ellis (2018), Lin (2018), and Chambers et al. (2019) for related analyses.

analyst observes both final choices and decision times and, in contrast to the present paper, DM's feasible set of experiments need not be separable across time periods and delay can be intrinsically costly. Most related to the present paper are Denti (2020), de Oliveira (2019), and Caplin et al. (2019b), which characterize choice behavior that is consistent with DM having a (Uniformly) Posterior Separable cost function. The latter two papers also characterize the Mutual Information cost function within this class.

The revealed preference approach, which places experimentally testable axioms on DM's observable choice behavior, is complementary to our approach, which places (not directly testable) axioms derived from optimality principles directly on DM's cost function. In particular, the methods of analysis in this paper are almost completely distinct from those of the revealed preference literature, and our characterization theorems elucidate complementary aspects of the cost functions under study.

Axiomatic Foundations for Mutual Information. In addition to the work of Caplin et al. (2019b), which motived our characterization of Mutual Information, several other papers provide foundations for this cost function based on related notions of "compression" of states. Tian (2019) introduces a "distraction free" axiom that is equivalent to our notion of Compression Monotonicity and proves that any Bounded UPS cost function must be Mutual Information, analogous to the equivalence of points (ii) and (iii) in our Theorem 5. Within information geometry, the work of Jiao et al. (2014a, 2015b) is closest to the equivalence of points (i) and (iii) in our Theorem 5. Bloedel and Segal (2020), Angeletos and Sastry (2019), and Hébert and La'O (2020) study the implications of these axioms in economic applications. See Subsection 6.3 for further discussion.

Value of Information. While this paper studies the *cost* of information, a complementary line of work studies the (instrumental) *value* of information for a Bayesian expected utility maximizer. The most closely related papers characterize value functions over experiment-prior pairs that represent the expected utility gain (relative to having no information beyond her prior belief) that DM achieves in a given decision problem by observing the experiment's signal. When the decision problem that DM faces is permitted to vary with her prior belief, the class of such value functions consists precisely of the Bounded Posterior Separable functions (Azrieli and Lehrer (2008)).[19] When the decision problem is fixed independently of DM's prior belief, the class of such value functions consists precisely of the Bounded UPS functions. Thus, perhaps surprisingly, (Uniformly) Posterior Separable functions can be derived as measures of either the value or the cost of information. However, there is a critical distinction: the structure of expected utility preferences implies that value functions for information are *necessarily* Posterior Separable while, as our analysis shows, the Indirect Cost of information is only Posterior Separable under additional conditions. We discuss this distinction, and an application of our results to the value of information, in Appendix K.

[19] See also Mensch (2018), which obtains an equivalent characterization, and Gilboa and Lehrer (1991), which obtains a similar characterization restricted to the class of partitional information structures. Jakobsen (2020) extends the Azrieli and Lehrer (2008) representation to settings where, as in Sender-Receiver games, the acquirer and user of information may have different preferences and prior beliefs. Cabrales et al. (2013) show that Mutual Information characterizes the value of information in a class of financial investment problems.

2 Framework

This section introduces our main framework. We first present model primitives in Subsection 2.1. In Subsection 2.2, we then introduce the model of sequential cost-minimization. Finally, in Subsection 2.3 we discuss the role of various model assumptions.

2.1 Primitives

States and Beliefs. A Bayesian decision-maker (DM) acquires information about an uncertain state θ drawn from a finite state space Θ. Generic states are denoted $\theta, \theta' \in \Theta$. Let $\Delta(\Theta)$ and $\Delta_\circ := \text{int}(\Delta(\Theta))$ denote, respectively, the sets of all and full-support probability measures on Θ (or *beliefs*). We will let p denote DM's *prior* belief and let q denote her *posterior* belief (conditional on having observed some information).

Experiments. Following Blackwell (1951), we model DM as acquiring information in the form of *experiments* $\langle S, \sigma \rangle$, where S is a Polish *signal space* and $\sigma : \Theta \to \Delta(S)$ is a measurable map.[20] Let $\sigma_\theta \in \Delta(S)$ denote the distribution of signals conditional on state θ. We will frequently suppress an experiment's signal space, denoting $\langle S, \sigma \rangle$ simply by σ.

An experiment σ is *bounded* if (i) the conditional signal distributions $\{\sigma_\theta\}_{\theta \in \Theta}$ are mutually absolutely continuous, and (ii) there exists a constant $B > 0$ such that the Radon-Nikodym derivatives $\frac{d\sigma_\theta}{d\sigma_{\theta'}} \in [1/B, B]$ for all $\theta, \theta' \in \Theta$. In other words, a bounded experiment does not definitively rule out any state and, moreover, has uniformly bounded likelihood ratios. Unless otherwise noted, we henceforth restrict attention to bounded experiments. Let $\mathcal{E}$ denote the collection of all experiments and let $\mathcal{E}_b \subset \mathcal{E}$ denote the collection of bounded experiments.

Posterior Distributions. When DM observes the signal drawn from an experiment, she updates her prior belief to its Bayesian posterior. Thus, from the *ex ante* perspective before a signal is drawn, DM's posterior belief is a random variable. Let $\tilde{q} \sim \pi_{\langle \sigma | p \rangle}$ denote the random posterior belief induced by experiment σ and prior p, and call its distribution $\pi_{\langle \sigma | p \rangle} \in \Delta(\Delta(\Theta))$ the induced *posterior distribution* (which is short for "distribution over posterior beliefs"). It is well known that, given prior belief $p \in \Delta(\Theta)$, the posterior distribution π is induced by some experiment (i.e., there exists some $\sigma \in \mathcal{E}$ such that $\pi = \pi_{\langle \sigma | p \rangle}$) if and only if $\pi \in \Pi(p) := \{\pi \in \Delta\Delta(\Theta) : \mathbb{E}_\pi[\tilde{q}] = p\}$, meaning that the random posterior $\tilde{q}$ averages to the prior p. Let $\Pi := \bigcup_{p \in \Delta} \Pi(p)$ denote the set of all posterior distributions.

Given a full-support prior belief $p \in \Delta_\circ$, it is easy to see that the experiment σ is bounded if and only if the induced posterior distribution satisfies $\text{supp}\left(\pi_{\langle \sigma | p \rangle}\right) \subseteq \Delta_\delta := \{q \in \Delta \mid q_\theta \geq \delta \ \forall \theta\}$ for some $\delta > 0$. Let $\Pi_\delta := \{\pi \in \Pi \mid \text{supp}(\pi) \subseteq \Delta_\delta\}$. Then $\Pi_b := \bigcup_{\delta > 0} \Pi_\delta$ denotes the set of posterior distributions that are inducible by some bounded experiment and full-support prior.

We endow Π with the weak* topology, rendering it a compact and separable topological space. The subsets of Π defined above are endowed with the appropriate relative topologies. Convergence of a sequence $\{\pi^{(n)}\}_{n \in \mathbb{N}} \subset \Pi$ to the limit point π^* in this topology is denoted by $\pi^{(n)} \to^{w^*} \pi^*$.

[20] A topological space is *Polish* if it is separable and completely metrizable. We restrict attention to Polish signal spaces without appreciable loss of generality in order to ensure that DM's posterior beliefs are well-defined in the sense of "regular conditional probabilities." A Polish signal space S will always be endowed with its Borel sigma-algebra.

Cost Functions. An *(information) cost function* is a map $C : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ satisfying the following conditions:

(i) If σ and τ are Blackwell equivalent, then $C(\sigma \mid p) = C(\tau \mid p)$ for all $p \in \Delta_\circ$.

(ii) If $\underline{\sigma}$ is uninformative,[21] then $C(\underline{\sigma} \mid p) = 0$ for all $p \in \Delta_\circ$.

(iii) Let $\{(\sigma^{(n)}, p^{(n)})\}_{n\in\mathbb{N}}$ be a sequence of experiment-prior pairs inducing posterior distributions $\pi^{(n)} := \pi_{\langle\sigma^{(n)}|p^{(n)}\rangle}$. If $\pi^{(n)} \to^{w^*} \pi^*$ and there exists some $\delta > 0$ such that $\{\pi^{(n)}\}_{n\in\mathbb{N}} \subset \Pi_\delta$, then $C(\sigma^{(n)} \mid p^{(n)}) \to C(\sigma^* \mid p^*)$, where $\pi^* = \pi_{\langle\sigma^*|p^*\rangle}$.

Let $\mathcal{C}$ denote the set of information cost functions. Note that cost functions are only defined (and finite-valued) for bounded experiments and full-support priors. The first two points of this definition are standard. Point (i) means that the cost of an experiment is fully determined by its information content (and DM's prior belief). This is without loss of generality when experiments are chosen optimally. Point (ii) means that DM has the option to "do nothing," which has zero cost. It also implies that DM has free access to mixed strategies, since the outcome of an uninformative experiment can always be used as a randomization device.

Point (iii) is a continuity condition. A standard way to define continuity for information cost functions is via weak* continuity of the induced cost function over posterior distributions $\hat{C} : \Pi_b \to \mathbb{R}_+$ defined by $\hat{C}\left(\pi_{\langle\sigma|p\rangle}\right) := C(\sigma \mid p)$ (which is well-defined by point (i)).[22] However, this notion of continuity is too strong for our purposes, as it is violated by important classes of unbounded cost functions (see Subsection 2.4 below). The weaker continuity condition in point (iii) of the above definition allows us to accommodate such cost functions.

2.2 Sequential Replication and Optimality

We model information acquisition as a two-stage process. In the first stage, DM decides *what* to learn, modeled as the choice of a *target experiment* to acquire. However, DM does not need to acquire this target experiment in one shot. Instead, in the second stage, she chooses the optimal *way in which to acquire* this target experiment through a strategy that *sequentially replicates* it: DM may sequentially acquire "sub-experiments" that are each less informative than her target experiment, but collectively are sufficient for it. In this subsection, we present two notions of what it means for an information cost function to be "rationalizable" with respect to sequential information acquisition of this form. (Per Theorem 1 below, these two notions are equivalent, but will be independently useful in the sequel.)

Sequential Learning-Proofness. The first notion of "rationalizability" states that the cost of acquiring a given experiment should not be reducible in expectation by acquiring information in two steps rather than one. We formalize this a fixed-point condition that a cost function must satisfy.

Given (Polish) signal spaces S', S'', define a *two-step sequential experiment* as an experiment $\sigma'' * \sigma' : \Theta \to \Delta(S' \times S'')$ with marginal distribution $\sigma'(\cdot \mid \theta) := \text{marg}_{S'}[\sigma'' * \sigma'](\cdot \mid \theta)$ and conditional marginal

[21] That is, $\underline{\sigma}(\cdot \mid \theta) = \underline{\sigma}(\cdot \mid \theta')$ for all $\theta, \theta' \in \Theta$ or, equivalently, the induced posterior distribution satisfies $\pi_{\langle\sigma|p\rangle} = \delta_p$ for all $p \in \Delta(\Theta)$.

[22] Perhaps the most natural way to define continuity is to use the weak* topology on $\mathcal{E}$ in the following sense: $C(\sigma^n \mid p) \to C(\sigma \mid p)$ whenever the experiments $\{\sigma^n\}$ and σ have the same (Polish) signal space S, and $\sigma^n(\cdot \mid \theta) \to^{w^*} \sigma(\cdot \mid \theta)$ for all $\theta \in \Theta$. However, as discussed by Torgersen (1991, p. 401) and Denti et al. (2020), this continuity requirement is much stronger than weak* continuity with respect to the induced posterior distributions.

distributions $\sigma''_{s'}(\cdot \mid \theta) := [\sigma'' * \sigma'](\cdot \mid \theta, s') \in \Delta(S'')$. Intuitively, $\sigma'' * \sigma'$ represents a two-step sequential information acquisition strategy where DM first acquires σ' and then, conditional on the realized signal s', acquires a second experiment $\sigma''_{s'}$. At the end of this process, she observes the tuple of realized signals (s', s'').

Definition 1 (SLP). *Cost function C is **Sequential Learning-Proof** (SLP) if $C = \Psi C$ where $\Psi C : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ is defined by*

$$\Psi(C)(\sigma \mid p) := \inf_{\sigma'' * \sigma' \gtrsim_B \sigma} C(\sigma' \mid p) + \mathbb{E}_{\langle \sigma' \mid p \rangle}\left[C\left(\sigma''_{\tilde{s}'} \mid q(\cdot \mid \tilde{s}')\right)\right] \tag{SLP}$$

Thus, a cost function is SLP if, and only if, it is weakly cheaper (in expectation) to acquire experiment σ in one shot than it is to acquire any two-step sequential experiment $\sigma'' * \sigma'$ that is at least as informative as σ. This formulation implicitly builds in the assumption that any "extra" information contained in $\sigma'' * \sigma'$ but not in σ can be freely discarded. Let $\mathcal{C}_{SLP}$ denote the set of all SLP cost functions.

The Direct and Indirect Cost of Information. The second notion of "rationalizability" states that the "reduced form" cost of information should represent the expected cost-minimizing way to replicate a given experiment when DM is able to optimize over any sequential information acquisition strategy. We first describe DM's strategy space, which is completely flexible: she has access to any sequential information acquisition strategy, including one-shot learning, randomization, and free disposal of acquired information. We formalize sequential information acquisition as follows:

Definition 2 (Sequential Replication). *For $T \in \mathbb{N}$, **length**-$2T$ **Sequential Replication** of the **target experiment** σ consists of:*

(i) A collection of (Polish) signal spaces $\{S_t\}_{t=1}^{2T}$ satisfying $S_{2t-2} \times S_{2t-1} \subseteq S_{2t}$ (and where S_0 is singleton),
(ii) A collection of (even period) measurable maps $\sigma^{(2t)} : S_{2t} \times \Theta \to \Delta(S_{2t+1})$, and
(iii) A collection of (odd period) measurable maps $\gamma^{(2t+1)} : S_{2t} \times S_{2t+1} \to \Delta(S_{2t+2})$,

such that σ is Blackwell equivalent to the experiment $\sigma^R : \Theta \to \Delta(S_{2T})$ for which $\sigma^R(\cdot \mid \theta)$ is defined as the marginal distribution on S_{2T} of

$$\prod_{t=0}^{T-1} \sigma^{(2t)}(s_{2t+1} \mid s_{2t}, \theta)\gamma^{(2t+1)}(s_{2t+2} \mid s_{2t+1}, s_{2t}).$$

Definition 2 describes a sequential information acquisition process in which information is acquired in even periods (i.e., s_{2t+1} "adds" information about θ to s_{2t}) and can be disposed of odd periods (i.e., s_{2t+2} "discards" information from (s_{2t+1}, s_{2t})). Point (ii) states that, conditional on observing the non-discarded information contained in signal s_{2t}, DM acquires the experiment $\sigma^{(2t)}_{s_{2t}} : \Theta \to \Delta(S_{2t+1})$ defined by $\sigma^{(2t)}_{s_{2t}}(\cdot \mid \theta) := \sigma^{(2t)}(\cdot \mid s_{2t}, \theta)$. Point (iii) states that, conditional on the non-discarded information contained in s_{2t} and the newly acquired information contained in s_{2t+1}, DM "garbles" this information into s_{2t+2} according to the map $\gamma^{(2t+1)}$. Point (i) requires that the signal spaces be nested, which is a richness condition ensuring that it is feasible to choose full memory, i.e., let all of the garblings $\gamma^{(2t+1)}$ be fully informative. Under full memory, we may re-index time periods to get rid of the disposal rounds and recover the standard definition of a *sequential (Blackwell)*

experiment (Greenshtein (1996)). We discuss the role of free disposal of information in Subsection Section 2.3 below.

The final condition in Definition 2 requires that σ^R, the conditional distributions over terminal signals s_{2t} induced by this sequential process, must be Blackwell equivalent to the target experiment σ. That is, the sequential process must precisely replicate the information contained in the target experiment. Let $\langle \sigma, \gamma \rangle \rightharpoonup \sigma$ be shorthand notation for a Sequential Replication (of any length $2T$) of target experiment σ.

We now describe DM's cost of information. DM has a *Direct Cost* function C that represents her "primitive" technology for acquiring information; we do not impose any assumptions on C other than requiring that it is a well-defined cost function. If she acquires information in one shot, the cost is determined by C. If she acquires information via a Sequential Replication, the C serves as the "flow cost" for each (even period) acquisition round, while the (odd period) disposal rounds are free. The expected cost of optimally acquired information is represented as an *Indirect Cost* function defined as follows:

Definition 3 (Indirect Cost)**.** *Cost function C^* is the **Indirect Cost** generated by the **Direct Cost** function C if $C^* = \Phi C$, where ΦC is defined by*[23]

$$\Phi C(\sigma \mid p) := \inf_{\langle \sigma, \gamma \rangle} \mathbb{E}_{\langle \sigma, \gamma | p \rangle} \left[\sum_{t=0}^{T-1} C\left(\sigma^{(2t)}_{\tilde{s}_{2t}} \;\middle|\; q(\cdot \mid \tilde{s}_{2t}) \right) \right] \tag{IC}$$
$$s.t.\ \langle \sigma, \gamma \rangle \rightharpoonup \sigma.$$

Definition 3 states that $\Phi C(\sigma \mid p)$ is the minimum expected cost of Sequentially Replicating the target experiment $\sigma \in \mathcal{E}_b$ when the prior belief is $p \in \Delta_\circ$ (implicit in the program (IC) is that DM optimizes over the length $2T$ of Sequential Replications). Note that, while $\Phi C(\sigma \mid p)$ is a well-defined non-negative number, it is not *a priori* clear that the mapping $\Phi C : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ constitutes a well-defined cost function in the sense of Subsection Section 2.1 (which requires continuity, among other things). Thus, Definition 3 requires that ΦC be a well-defined cost function as part of the definition of Indirect Cost. Nonetheless, Theorem 1 below will establish that this additional qualifier is not needed: under our assumptions, ΦC is, in fact, a well-defined cost function.

Remark 1. *Note that the two-step operator Ψ used to define SLP cost functions corresponds to minimization over Sequential Replications with $T = 2$, and with free disposal of information* only *in the terminal period. In particular, given target experiment σ, sequential experiment $\sigma'' * \sigma' \gtrsim_B \sigma$ corresponds to the Sequential Replication in which (i) $S_1 = S_2 = S'$, $S_3 = S''$, and $S_4 = S' \times S''$, (ii) $\sigma^{(0)} = \sigma'$ and $\sigma^{(2)}(\cdot \mid \theta, s') = \sigma''_{s'}(\cdot \mid \theta)$, and (iii) $\gamma^{(1)}$ is fully informative while $\gamma^{(3)}$ discards the information in $\sigma'' * \sigma'$ that is superfluous for σ. The operator Φ can be equivalently defined by iterating on the Ψ operator (i.e., namely, $\Phi = \lim_{T \to \infty} \Psi^T$), yielding an equivalent "recursive" characterization of Indirect Cost functions.*

2.3 Discussion of Assumptions

Before proceeding with the analysis, we briefly discuss some important aspects of the model formulation.

[23] We implicitly assume in (IC) that $C(\sigma \mid p) = +\infty$ whenever $\sigma \notin \mathcal{E}_b$, so that only Sequential Replications with bounded experiments in each acquisition round are feasible at finite cost. Also, in (IC), the expectation operator $\mathbb{E}_{\langle \sigma, \gamma | p \rangle}$ is that induced by the joint probability measure over states and paths of signals induced by prior $p \in \Delta_\circ$ and Sequential Replication $\langle \sigma, \gamma \rangle \rightharpoonup \sigma$.

Cost of Sequential Replication. Our most substantive assumptions concern the cost of Sequential Replication in (IC), which specifies that (i) the total cost is additively separable across periods, (ii) there is no intrinsic cost to delay (either in form of additive time costs or time discounting), and (iii) the cost of information is additively separable from its value (in any un-modeled decision problem in the background). Assumption (iii) is standard — and also necessary to develop a theory of the cost of information separately from a theory of its value — so we do not comment on it further.

Assumption (i) is standard in economic and statistical models of sequential information sampling. It also allows for a straightforward comparison between Direct and Indirect Cost cost functions, since they are defined over the same domain of (static) experiments. While one could envision an alternative formulation in which the cost of Sequential Replication is determined by a (potentially non-additively separable) cost function over *sequential* Blackwell experiments, without substantive assumptions on this cost function the model would have very little predictive power for the Indirect Cost of information (see Appendix K).

One very literal interpretation of assumption (ii) is that our framework is best suited to settings in which (a) each round of Sequential Replication corresponds to negligibly short period of "calendar time" and (b) all information is gathered before DM makes a one-time decision. Environments satisfying these criteria include the experiments on human attention and perception considered in the mathematical psychology and neuroeconomics literatures, in which information is gathered over the course of seconds, and the consumer search settings considered in economics and marketing, in which information is gathered over the course of minutes. However, our favored interpretation of the framework is much broader. We view Sequential Replication simply as a way to formalize the idea that information can be acquired in a piecemeal fashion, which is arguably realistic in a wide range of economic environments — including those that do not fit the above criteria. In this view, assumption (ii) represents the idealized "frictionless limit" in which DM is not constrained to acquiring information at any particular rate relative to the passage of calendar time, while intrinsic delay costs would amount to assuming that DM's strategy space is more restricted in this respect.[24] See Appendix K for a discussion of how our results would change if we were to include additive delay costs or discounting.

Domain of (Direct) Costs. We assume that cost functions are finite-valued at all bounded experiments and full-support priors. This captures the idea that DM is not *a priori* restricted to a specific parametric class of experiments, and so has (essentially) full flexibility in *what* to learn. However, we do *not* require that cost functions be finite-valued — or even well-defined — at unbounded experiments or partial-support priors. There are two reasons for this restriction. First, it allows for a unified treatment of a very general class of cost functions, including those — such as Total Information and the LLR cost — that are infinite-valued at the fully informative experiment.[25] Second, it allows us to sidestep the question of how to assign costs starting from partial-support priors, in which case it is not clear whether costs should be determined by the experiment itself or by the posterior distribu-

[24] For instance, Hébert and Woodford (2020b) study a model that is related to ours but, motivated by experiments on human perception, in which DM is explicitly restricted to acquiring information gradually and pays a delay cost that is linear in the amount of time spent acquiring information. The idea of studying "frictionless limits" in which actions can be taken frequently with respect to calendar time and discounting vanishes is familiar from the theory of bargaining, renegotiation, and repeated games.

[25] As discussed in Appendix K, all of our results naturally extend, with minor technical qualifications, to larger domains of experiments.

tion that it induces, with the latter convention tacitly assuming that learning about zero-probability states has zero cost.[26] While some authors have argued that these two modeling conventions are inherently at odds (e.g., Denti et al. (2020)), we believe that there are compelling reasons to adopt either one depending on the context. By restricting to bounded experiments and full-support priors, we are ensured that DM's posterior beliefs will always have full support, even when information is acquired in multiple rounds.

Flexibility of Sequential Replication. Our definition of Sequential Replication endows DM with essentially complete flexibility in designing her sequential strategy; in particular, it does not restrict to her to any parametric class of sampling strategies, such as those comprised only of Gaussian or Poisson experiments (cf. Subsections 3.3.2 and 4.4). While our assumptions on cost functions ensure that DM has full flexibility in *what* to learn, this assumption on her strategy space captures the idea that she also has full flexibility in *how to acquire* the information she chooses to learn. This allows us to derive predictions that are robust to such restrictions: any SLP cost function function will remain as such even if such restrictions are imposed. That said, restrictions on the feasible set of experimentation strategies can be (partially) encoded into the Direct Cost function itself by making certain types of experiments prohibitively expensive.[27] Moreover, some of our results remain valid even when the strategy space itself is restricted; we note where this is the case throughout the paper. On the other hand, in Sections 5 and 6 we enlarge DM's strategy space even further and show how additional margins of optimization impose further structure on her Indirect Cost.

Free Disposal. Notably, the definition of Sequential Replication also allows for free disposal of acquired information, both during (in odd periods $t < 2T-1$) and at the end of (in period $2T-1$) the acquisition process. Similarly, the definition of SLP cost functions allows for free disposal at the end of the two-step acquisition process. These assumptions play two roles. First, allowing for disposal during the acquisition process allows us to accommodate Direct Cost functions that are non-monotone (in the sense of Axiom 1 below). Second, allowing for disposal at the end guarantees that DM's feasible sets are nested with respect to the Blackwell order: each Sequential Replication of experiment σ is also a valid Sequential Replication of any less informative experiment $\sigma' \preccurlyeq \sigma$. This is a standard assumption in models of one-shot information acquisition but takes on added significance in our sequential model, in which DM may want to acquire "superfluous information" that will be discarded in some later period.

There are two reasons for this. First, when the Direct Cost is prior-dependent — in particular, concave in the prior belief (see Subsection A.1.3) — DM may have an *intrinsic* preference for superfluous information because it decreases her continuation costs holding her continuation strategy fixed. Second, even when the Direct Cost is independent of the prior (see Subsection 3.4), she may have an *instrumental* preference for superfluous information because, by providing more information to condition on, it allows her access to more (and sometimes cheaper) continuation strategies. The latter phenomenon is illustrated in Appendix K. One might conjecture that it cannot arise when the Direct

[26] For instance, when $p = \delta_\theta$ puts full mass on state θ, all experiments induce the degenerate posterior distribution $\pi = \delta_p$. Therefore, in the latter convention, $C(\cdot \mid \delta_\theta) \equiv 0$. Subsection 3.4 discusses implications of this property.

[27] Examples of this appear in Proposition 2 and Theorem 3, which provide conditions on the Direct Cost function under which the optimal Sequential Replication resembles the acquisition of, respectively, Poisson and Gaussian diffusion signals.

Cost is monotone with respect to the Blackwell order, but the issue is that even then DM's induced cost function over *sequential* experiments (i.e., Sequential Replications) is generally not monotone with respect to the *sequential* Blackwell order (Greenshtein (1996)).

2.4 Specific Cost Functions

Three cost functions play especially important roles in our analysis. Each is based on the notion of *Kullback-Leibler (KL) divergence* between probability distributions. Given any (Polish) space X and probability measures $\mu, \nu \in \Delta(X)$, the KL divergence from μ to ν is defined as the expected log-likelihood ratio

$$D_{KL}(\nu \mid \mu) := \int_X \log\left(\frac{d\nu}{d\mu}(x)\right) d\nu(x) \tag{KL}$$

wherever this expression is well-defined and finite.

The first is the familiar Mutual Information cost function on which Sims' (2003) rational inattention model of information processing is based, and which has since become widespread in economic applications of costly information acquisition.

Definition 4 (Mutual Information)**.** *The **Mutual Information** cost function $C_{MI} : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ is defined by*

$$\begin{aligned} C_{MI}(\sigma \mid p) &:= \sum_\theta p_\theta D_{KL}(\sigma_\theta \mid \sigma \circ p) \\ &= \mathbb{E}_{\langle \sigma \mid p \rangle}[H(p) - H(\tilde{q})] \end{aligned} \tag{MI}$$

where $\sigma \circ p \in \Delta(S)$ denotes the marginal distribution of signals $[\sigma \circ p](\cdot) := \sum_\theta p_\theta \sigma(\cdot \mid \theta)$, and $H(p) := -\sum_\theta p_\theta \log(p_\theta)$ is Shannon entropy.

Thus, the Mutual Information cost function is defined as the expected KL divergence from the unconditional distribution of signals to the state-contingent distributions. It is easy to see that Mutual Information depends on DM's prior belief, which partially determines the unconditional signal distribution. It is well known that this definition is equivalent to the expected reduction of Shannon entropy.

The second cost function is the *Log-Likelihood Ratio (LLR)* cost function recently introduced by Pomatto et al. (2019) as a cost function for information production.

Definition 5 (LLR)**.** *The **Log-Likelihood Ratio (LLR)** cost function is the Prior-Invariant function $C_{LLR} : \mathcal{E}_b \to \mathbb{R}_+$ is defined by*

$$\begin{aligned} C_{LLR}(\sigma) &:= \sum_{\theta,\theta'} \beta_{\theta,\theta'} D_{KL}\left(\sigma_\theta \mid \sigma'_\theta\right) \\ &= \mathbb{E}_{\langle \sigma \mid p \rangle}[\mathcal{G}(\tilde{q} \mid p) - \mathcal{G}(p \mid p)] \end{aligned} \tag{LLR}$$

*for some vector $\boldsymbol{\beta} \in \mathbb{R}_+^{|\Theta|(|\Theta|-1)}$ of **discrimination coefficients**, where $\mathcal{G}(q \mid p) := \sum_{\theta,\theta'} \frac{q_\theta}{p_\theta} \beta_{\theta,\theta'} \log\left(\frac{q_\theta}{q_{\theta'}}\right)$.*

The third cost function, which is new to this paper, is a particular prior-dependent variant of the LLR cost function.

Definition 6 (Total Information). *The **Total Information** cost function $C_{TI} : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ is defined by*

$$C_{TI}(\sigma \mid p) := \sum_\theta p_\theta \left[\sum_{\theta'} \gamma_{\theta,\theta'} D_{KL}\left(\sigma_\theta \mid \sigma'_\theta\right) \right] \tag{TI}$$
$$= \mathbb{E}_{\langle \sigma | p \rangle} \left[G(\tilde{q}) - G(p) \right]$$

*for some vector $\boldsymbol{\gamma} \in \mathbb{R}_+^{|\Theta|(|\Theta|-1)}$ of **discrimination coefficients**, where $G(p) := \sum_{\theta,\theta'} p_\theta \gamma_{\theta,\theta'} \log\left(\frac{p_\theta}{p_{\theta'}}\right)$.*

We discuss the relationship between Total Information and the LLR cost functions further in Section 5. Special cases of Total Information include the Wald cost function of Morris and Strack (2019) and the Fisher Information cost function of Hébert and Woodford (2020a).

3 Characterization of SLP and Indirect Cost Functions

In this section, we characterize the full classes of SLP and Indirect Cost functions. Subsection 3.1 presents the main characterization result. Subsection 3.2 then presents leading examples of SLP cost functions. Finally, Subsections 3.3 and 3.4 present implications of the main characterization theorem.

3.1 Characterization

Axioms. We present the two axioms that will be used below to characterize the SLP and Indirect Cost functions. The first axiom is standard:

Axiom 1 (Blackwell monotone). *C is **Blackwell monotone** if $C(\cdot \mid p)$ is non-decreasing in the Blackwell order for all priors $p \in \Delta_\circ$.*

Axiom 1 states that acquiring more informative experiments is weakly more costly. In the context of one-shot information acquisition, it is easy to see that DM never benefits from free disposal of information if and only if her cost function is Blackwell monotone.

The second axiom, which is new, captures the restriction imposed by sequential optimality:

Axiom 2 (Preference for One-Shot Learning). *C exhibits **Preference for One-Shot Learning** if*

$$C(\sigma \mid p) \leq C(\sigma' \mid p) + \mathbb{E}_{\langle \sigma' | p \rangle} \left[C\left(\sigma''_{\tilde{s}'} \mid q(\cdot \mid \tilde{s}')\right) \right] \tag{POSL}$$

for all $\sigma'' * \sigma' \sim_B \sigma$ and $p \in \Delta_\circ$.

Axiom 2 states that it is cheaper (in expectation) to acquire all information at once rather than in two steps, *without* free disposal.[28] Mutual Information is the most well-known cost function that exhibits Preference for One-Shot Learning. As discussed in Subsection 3.2 below, Axiom 2 is also satisfied by Total Information and, more broadly, by the "uniformly posterior separable" cost functions used in the rational inattention literature.

[28] If the target experiment σ in (POSL) were fully informative, Preference for One-Shot Learning would formally resemble the "preference for one-shot resolution of uncertainty" axiom from Dillenberger (2010) and Dillenberger and Raymond (2020) (see also the "preference for clumped information" condition from Köszegi and Rabin (2009)). However, those papers study the *intrinsic* (i.e., decision-irrelevant) *value* of information for a DM with non-expected utility preferences, so there is little conceptual connection to these properties.

Characterization Theorem. The following result characterizes the full classes of SLP and Indirect Cost functions:

Theorem 1. *For cost function C^*, the following are equivalent:*

(i) C^* *is SLP, i.e.,* $C^* = \Psi C^*$.

(ii) C^* *is its own Indirect Cost, i.e.,* $C^* = \Phi C^*$.

(iii) C^* *is Blackwell monotone and exhibits Preference for One-Shot Learning.*

Moreover, given any Direct Cost C, the Indirect Cost $C^* = \Phi(C)$ is a well-defined cost function and is SLP. Thus, $\Phi(\mathcal{C}) = \mathcal{C}_{SLP}$.

Proof. See Appendix C. □

Theorem 1 has three takeaways. First, the equivalence of points (i) and (ii) means that SLP cost functions are, in fact, fixed points of the full sequential minimization process, i.e., not reducible in expectation by *any* Sequential Replication. This justifies our focus on SLP cost functions. It is also natural: any Sequential Replication can be decomposed into a sequence of two-step replications. Second, point (iii) gives the analyst a way to check whether a given cost function is SLP or not. We explore further implications of these axioms in Subsections 3.2–3.4 below. Third, Theorem 1 also establishes that sequential minimization of *any* Direct Cost generates a well-defined Indirect Cost function that is SLP. Henceforth, we therefore treat SLP and Indirect Cost functions interchangeably.

3.2 Examples of SLP Cost Functions

Theorem 1(iii) gives us a way to verify whether a given cost function is SLP. Here, we use this characterization to present examples of SLP cost functions that will be referenced throughout the paper.

3.2.1 Uniform Posterior Separability

The most important subclass of SLP costs are the "uniformly posterior separable" cost functions used in the rational inattention literature, which have become the default modeling tool in many economic applications of costly information acquisition (Caplin et al. (2019b)) and which are studied in detail in Section 4 below.

Definition 7 (UPS)**.** *Cost function C is **Uniformly Posterior Separable (UPS)** if there exists a convex **potential function** $F : \Delta_\circ \to \mathbb{R}$ such that*[29]

$$C(\sigma \mid p) = \mathbb{E}_{\langle \sigma | p \rangle}\left[F(\tilde{q}) - F(p)\right]$$

for all $\sigma \in \mathcal{E}_b$ and $p \in \Delta_\circ$.

[29] Convexity of the potential function implies that every UPS cost function is Blackwell monotone. The potential function is unique up to translation by affine functions, i.e., $F(q)$ represents C if and only if $F(q) + g(q)$ does as well, where $g : \Delta_\circ \to \mathbb{R}$ is affine.

Notice that, as is well known, Mutual Information is UPS with the potential function $F(q) := -H(q)$ given by the negative of Shannon entropy. Total Information is also UPS, with the potential function $F(q) := G(q)$ given in (TI). As suggested by the prior-dependence of the function $\mathcal{G}(q \mid p)$ in (LLR), the LLR cost function is not UPS; Corollary 1.3 below shows that it is not even SLP.

To see that every UPS cost function is SLP, it suffices to observe that the UPS class is characterized by the following strengthening of Preference for One-Shot Learning, which requires that the expected cost of two-step replication (without free disposal) is *equal to* the cost of one-shot learning.

Axiom 3 (Indifference to Sequential Learning)**.** *C exhibits **Indifference to Sequential Learning** if*

$$C(\sigma \mid p) = C(\sigma' \mid p) + \mathbb{E}_{\langle \sigma' \mid p \rangle}\left[C\left(\sigma''_{\tilde{s}'} \mid q(\cdot \mid \tilde{s}')\right)\right] \tag{ISL}$$

for all $\sigma'' * \sigma' \sim_B \sigma$ and $p \in \Delta_\circ$.

Notice that Indifference to Sequential Learning implies that C is Blackwell monotone because the expected second-stage cost in (ISL) is non-negative. By induction, it also implies that every Sequential Replication of a given experiment (without free disposal) is equally costly and, moreover, that it is without loss of optimality to restrict attention to Sequential Replications without free disposal.

Lemma 1. *Cost function C is UPS if and only if it exhibits Indifference to Sequential Learning.*

Proof. See Appendix K. □

Special cases of Lemma 1 have appeared elsewhere in the literature; we state and prove it here for completeness.[30] Lemma 1 and Theorem 1 together imply that every UPS cost function is SLP. Conversely, an SLP cost function is UPS if and only if it satisfies the opposite of Preference for One-Shot Learning, which we call *Preference for Sequential Learning*:

$$C(\sigma \mid p) \geq C(\sigma' \mid p) + \mathbb{E}_{\langle \sigma' \mid p \rangle}\left[C\left(\sigma''_{\tilde{s}'} \mid q(\cdot \mid \tilde{s}')\right)\right] \tag{PSL}$$

for all $\sigma'' * \sigma' \sim_B \sigma$ and $p \in \Delta_\circ$. Section 4 below is devoted to studying precisely how restrictive (PSL) is within the class of SLP cost functions.

3.2.2 Non-UPS Examples

Distance of Belief Movement. A natural class of cost functions that are SLP but not UPS are those derived from *quasi-metrics* $d(q,p)$ on the probability simplex.[31] Given a quasi-metric d, we may define the *distance-based* cost function

$$C(\sigma \mid p) = \mathbb{E}_{\langle \sigma \mid p \rangle}\left[d(\tilde{q}, p)\right], \tag{3}$$

where $d(q,p)$ is interpreted as the cost of a signal that moves DM's belief from the prior p to the posterior q. All distance-based cost functions exhibit Preference for One-Shot Learning because quasi-metrics satisfy the triangle inequality, and these functions are Blackwell monotone whenever $d(\cdot, p)$

[30] For instance, Frankel and Kamenica (2019) establish that if C has Full Domain and Posterior Separable, then it is UPS if and only if it exhibits Indifference to Sequential Learning (which they refer to as "combination invariance"). Zhong (2019) establishes the special case of Lemma 1 in which C is assumed to have Full Domain.

[31] Recall that $d : \Delta_\circ \times \Delta_\circ \to \mathbb{R}_+$ is a *quasi-metric* if (i) $d(q,p) = 0$ implies that $q = p$ and (ii) it satisfies the triangle inequality $d(q,p) \leq d(q,r) + d(r,p)$ for all $q, p, r \in \Delta_\circ$.

is convex for all $p \in \Delta_\circ$. Any quasi-metric derived from a norm satisfies the latter convexity requirement (e.g., the *total variation distance* $d_{TV}(q,p) := \frac{1}{2}\sum_\theta |q_\theta - p_\theta|$). It is easy to see that no non-zero cost function can be both UPS and distance-based, implying that these two subclasses of SLP costs are (essentially) disjoint.[32]

SLP-Preserving Operations. There are SLP cost functions that are neither UPS nor distance-based. To see this, note that we can always use certain combinations of extant SLP costs to generate new SLP costs. For instance, if cost functions C_1 and C_2 are both SLP, then the pointwise maximum $C(\sigma \mid p) := \max\{C_1(\sigma \mid p), C_2(\sigma \mid p)\}$ is also SLP.[33] However, even if each of C_1 and C_2 is either UPS or distance-based, C will generally fall outside of these classes. To see this, note that UPS and distance-based cost functions both exhibit "indifference to randomization" — i.e., are Randomizaton Averse but never strictly so (see (IR)) — while the pointwise maximum C will generally be sometimes-strictly Randomizaton Averse. For another example, the positive linear combination $aC_1 + bC_2$ (where $a, b \geq 0$) of two SLP cost functions is always SLP; if C_1 is UPS and C_2 is distance-based, then their linear combination will generally fall outside of both classes because these classes are disjoint (as argued above).[34]

3.3 Properties Implied by Preference for One-Shot Learning

In this subsection, we illustrate that Preference for One-Shot Learning implies two less restrictive axioms that the literature has used to characterize information cost functions. This exercise both sheds light on the meaning of Preference for One-Shot Learning and introduces new concepts that will be used in the sequel.

3.3.1 Randomization Aversion

Our definition of Sequential Replication endows DM with free access to mixed strategies: she can always use uninformative experiments (which have zero cost) as randomization devices for her continuation strategy. Naturally, then, SLP cost functions should not be further reducible (in expectation) by engaging in mixed strategies. Corollary 1.1 establishes that this is indeed the case.

To formalize the idea of mixed strategies, given weight $\alpha \in (0,1)$, define the *mixture* of experiments $\langle S_1, \sigma_1 \rangle$ and $\langle S_2, \sigma_2 \rangle$ as the experiment $\alpha\sigma_1 \oplus (1-\alpha)\sigma_2$ with signal space $[S_1 \cup S_2] \times \{1,2\}$ and conditional probabilities

$$[\alpha\sigma_1 \oplus (1-\alpha)\sigma_2]((s,i) \mid \theta) = \begin{cases} \alpha\sigma_1(s \mid \theta) \cdot \mathbf{1}(s \in S_1), & \text{if } i = 1 \\ (1-\alpha)\sigma_2(s \mid \theta) \cdot \mathbf{1}(s \in S_2), & \text{if } i = 2. \end{cases}$$

[32] If a distance-based cost were UPS, then by Lemma 1 the quasi-metric d must satisfy $d(q,p) = d(q,r) + d(r,p)$ for all $q, p, r \in \Delta_\circ$. But if p is a convex combination of r and q, we must have $d(q,p) \leq d(q,r)$ and therefore $d(r,p) = 0$. By extension, this implies that $d \equiv 0$. See Frankel and Kamenica (2019, Corollary 1) for a related observation.

[33] The pointwise maximum C is clearly a well-defined cost function and Blackwell monotone. Because both C_1 and C_2 satisfy Preference for One-Shot Learning, it is easy to show that C does as well. Thus, C is SLP by Theorem 1.

[34] More generally, it may be useful to note that SLP cost functions are analogous to *sublinear* functions $f : \mathbb{R}^n \to \mathbb{R}$ on Euclidean space, with Preference for One-Shot Learning corresponding to *subadditivity* (i.e., $f(x+y) \leq f(x) + f(y)$) and the Dilution Linear property corresponding to *positive homogeneity* (i.e., $f(ax) = af(x)$ for $a \in \mathbb{R}_{++}$). Thus, operations that preserve sublinearity can generally be translated into operations that preserve SLP. Also, recall that every sublinear function admits a variational representation as the supremum over a convex set of linear functions. Pushing the analogy further, we conjecture that every SLP cost function admits a variational representation as the supremum over a convex set of UPS cost functions.

In words, $\alpha\sigma_1 \oplus (1-\alpha)\sigma_2$ corresponds to flipping a (biased) coin, *observing* the result of this coin flip, and conditioning the choice of experiment σ_1 or σ_2 on this result. By construction, this notion of mixtures of experiments is equivalent to taking mixtures of the induced posterior distributions: given any prior $p \in \Delta_\circ$, we have $\pi_{\langle \alpha\sigma_1 \oplus (1-\alpha)\sigma_2 | p\rangle} = \alpha\pi_{\langle \sigma_1 | p\rangle} + (1-\alpha)\pi_{\langle \sigma_2 | p\rangle}$.[35] The following axiom states that this kind of randomization does not (further) reduce expected costs:

Axiom 4 (Randomizaton Averse)**.** *Cost function C is **Randomizaton Averse** if $C(\sigma \mid p) \leq \alpha C(\sigma' \mid p) + (1-\alpha)C(\sigma'' \mid p)$ for all $\sigma, \sigma'', \sigma'' \in \mathcal{E}_b$ and $\alpha \in (0,1)$ such that $\alpha\sigma' \oplus (1-\alpha)\sigma'' \sim_B \sigma$.*

Axiom 4 is standard in the literature under different names and, as noted above, is equivalent to assuming that DM's cost function is convex in the space of posterior distributions.[36]

Corollary 1.1. *If C^* is SLP, then it is Randomizaton Averse.*

Proof. Let $p \in \Delta_\circ$, $\sigma \in \mathcal{E}_b$, and $\alpha \in (0,1)$ be given, and let $\sigma', \sigma'' \in \mathcal{E}_b$ be such that $\alpha\sigma' \oplus (1-\alpha)\sigma'' \sim_B \sigma$. Define the uninformative first-stage experiment $\langle \{s', s''\}, \sigma_1 \rangle$ by $\sigma(s' \mid \theta) = \alpha$ and $\sigma(s'' \mid \theta) = 1-\alpha$ for all $\theta \in \Theta$. Define the second-stage experiments $\sigma_{2,s'} := \sigma'$ and $\sigma_{2,s''} := \sigma''$. Because C^* exhibits Preference for One-Shot Learning by Theorem 1 and $C(\sigma_1 \mid p) = 0$ by definition, we have $C(\sigma \mid p) \leq 0 + \alpha C(\sigma' \mid p) + (1-\alpha)C(\sigma'' \mid p)$, implying that C^* is Randomizaton Averse. □

Axiom 1 and Axiom 4 together characterize the class of "canonical" cost functions (de Oliveira et al. (2017); Caplin and Dean (2015)) that represent the expected cost of optimally acquired information under the restriction that information acquisition is one-shot. In our language, these axioms characterize the class of Indirect Cost functions generated from optimization over Sequential Replications with $T = 1$.

3.3.2 Dilution Linearity

Given an SLP cost function, DM finds it optimal to acquire all information in one shot. However, this is never a *uniquely* optimal strategy. Roughly speaking, DM can engage in two distinct types of Sequential Replication: (i) those in which partial information arrives over multiple periods and (ii) those in which information arrives "all at once" but not necessarily in the first period $t = 0$. Preference for One-Shot Learning states that the former type of replication is never (strictly) optimal, but it turns out that DM is indifferent among all replications of the latter type. Intuitively, the latter type of replication is only useful for inter-temporal cost-smoothing purposes, but an SLP cost function has already "optimized away" all potential gains to cost-smoothing. Thus, SLP cost functions render DM indifferent to the "speed" of learning.

[35] A natural alternative would be to consider *component-wise* mixtures by defining the experiment $\alpha\sigma_1 + (1-\alpha)\sigma_2$ to be $[\alpha\sigma_1 + (1-\alpha)\sigma_2](s \mid \theta) = \alpha\sigma_1(s \mid \theta) + (1-\alpha)\sigma_2(s \mid \theta)$. This would correspond to DM using a randomization device that flips a coin and chooses an experiment for her, but does *not* reveal the outcome of the coin flip. It is well known that the ranking $\alpha\sigma_1 \oplus (1-\alpha)\sigma_2 \succcurlyeq_B \alpha\sigma_1 + (1-\alpha)\sigma_2$ always holds, and that this ranking can be strict when $S_1 \cap S_2 \neq \emptyset$ (see, e.g., Lemma 4 and Example 3 of Denti et al. (2020)). Intuitively, this alternative form of randomization loses information when DM is unsure about which experiment a given signal was drawn from.

[36] Caplin and Dean (2015) refer to Axiom 4 as "mixture feasibility" and de Oliveira et al. (2017) refer to it simply as "convexity." As can be seen from the definitions, cost function C is Randomizaton Averse if and only if the function $\hat{C} : \Pi_b \to \mathbb{R}_+$ defined by $\hat{C}(\pi_{\langle \sigma | p\rangle}) := C(\sigma \mid p)$ satisfies $C(\alpha\pi' + (1-\alpha)\pi'') \leq \alpha C(\pi') + (1-\alpha)C(\pi'')$ whenever $\mathbb{E}_{\pi'}[q] = \mathbb{E}_{\pi''}[q]$. Hébert and Woodford (2020b, Lemma 1) show that, if C is Blackwell monotone, then Axiom 4 is equivalent to convexity in the space of experiments with respect to the alternative notion of mixtures described in footnote 35.

To formalize these ideas, for each experiment $\sigma \in \mathcal{E}_b$ and weight $\alpha \in (0,1]$, define the *α-dilution of* σ as

$$\alpha \cdot \sigma := \alpha\sigma \oplus (1-\alpha)\underline{\sigma}, \tag{4}$$

where $\underline{\sigma}$ is a completely uninformative experiment. In other words, $\alpha \cdot \sigma$ corresponds to running σ with probability α and learning nothing with complementary probability $1-\alpha$, which induces the posterior distribution $\pi_{\langle \alpha\cdot\sigma|p\rangle} = \alpha\pi_{\langle\sigma|p\rangle} + (1-\alpha)\delta_p$ given prior belief $p \in \Delta_\circ$.

Axiom 5 (Dilution Linear)**.** *Cost function C is **Dilution Linear** if $C(\alpha\cdot\sigma \mid p) = \alpha C(\sigma \mid p)$ for all $\sigma \in \mathcal{E}_b$, $p \in \Delta_\circ$, and $\alpha \in (0,1)$.*

Axiom 5 is one of two main axioms that Pomatto et al. (2019) use to characterize the LLR cost function. While Axiom 4 states that mixed strategies are never strictly optimal, Dilution Linear states that a particular kind of mixed strategy — in which DM simply randomizes over *whether* to acquire a given experiment — is also without loss of optimality. To interpret this condition, notice that DM can replicate σ by following the length-$2T$ strategy whereby she runs $\alpha\cdot\sigma$ in each acquisition period until success (or until the final period, in which case she runs σ for sure). In the infinite-horizon ($T \to \infty$) limit, this strategy requires an expected number $1/\alpha$ of acquisition rounds, and so has total expected cost $C(\alpha\cdot\sigma \mid p)/\alpha$. We call this *direct Poisson learning* because in the "continuous-time limit" in which $\alpha \to 0$, the flow cost $C(\alpha\cdot\sigma \mid p) \to 0$ and the dynamics of DM's posterior belief resembles a Poisson process that stops after the first jump away from the prior.

Direct Poisson learning can be profitable when DM has an incentive to smooth costs over time. For instance, it is common in applications to model the cost of information as a convex transformation of some well-known cost function — e.g., by setting $C(\sigma \mid p) = f(C_{MI}(\sigma \mid p))$ where $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a smooth strictly convex function and C_{MI} is Mutual Information.[37] Given such a Direct Cost function, the cost of following a direct Poisson strategy is $f'(0)\cdot C_{MI}(\sigma \mid p)$ in the continuous-time limit, which is strictly cheaper than the cost of one-shot learning. It is easy to see that DM is indifferent among all direct Poisson strategies — i.e., has no such cost-smoothing motive — if and only if her cost function is Dilution Linear. The following lemma shows that SLP cost functions satisfy this condition; intuitively, the potential gains to such cost-smoothing have already been "optimized away."

Corollary 1.2. *If C^* is SLP, then it is Dilution Linear.*

Proof. Let $\langle S,\sigma\rangle \in \mathcal{E}_b$, $p \in \Delta_\circ$, and $\alpha \in (0,1]$ be given. Theorem 1 implies that C^* satisfies Axiom 2. Define the uninformative first-stage experiment $\langle S',\sigma'\rangle$ by $S' = \{Y,N\}$ and $\sigma'(Y \mid \theta) = \alpha$ for all $\theta \in \Theta$. Define the second-stage experiments $\langle S'',\sigma''_Y\rangle := \langle S,\sigma\rangle$ and let $\langle S'',\sigma''_N\rangle$ be uninformative. Then $\sigma''*\sigma' \sim_B \alpha\cdot\sigma$ and, by Axiom 2, we have $C(\alpha\cdot\sigma \mid p) \le \alpha C(\sigma \mid p)$. On the other hand, C^* satisfies Axiom 4 by Corollary 1.1, so that $\alpha C(\sigma \mid p) \le C(\alpha\cdot\sigma \mid p)$. Combining these inequalities yields Axiom 5. □

This fact will be important in Subsection 4.2 below.

[37] For instance, this kind of convex transformation of Mutual Information is used in Myatt and Wallace (2012) to model effort substitution across different information sources in a game-theoretic setting, in Dean and Neligh (2019, Section 5) to fit experimental data on individual's stochastic choice behavior generated by inattention, and in Zhong (2019) to study optimal cost-smoothing in a dynamic learning model with time discounting.

3.4 Incompatibility with Prior-Invariance

The Prior Invariance Critique. A leading critique of the UPS model, which we call the *Prior-Invariance Critique*, is that the "physical" or monetary costs of information acquisition should not depend on DM's prior beliefs. This property is formalized as follows:

Axiom 6 (Prior-Invariant)**.** *Cost function C is **Prior-Invariant** if the map $C(\sigma \mid \cdot) : \Delta_\circ \to \mathbb{R}_+$ is constant for each $\sigma \in \mathcal{E}_b$.*

It is easy to see that no Bounded UPS cost function (such as Mutual Information) is Prior-Invariant. For instance, under such functions the cost of any experiment vanishes as the prior becomes dogmatic:

$$\lim_{p \to \delta_\theta} \mathbb{E}_{\pi_{\langle \sigma | p \rangle}} [F(q) - F(p)] = 0 \tag{5}$$

for all $\sigma \in \mathcal{E}_b$, where $\delta_\theta \in \Delta(\Theta)$ denotes the Dirac measure on state θ.[38] The failure of Prior Invariance has led many authors to criticize the UPS model, as exemplified by the following passage from Gentzkow and Kamenica (2014, p. 459):[39]

> "[The UPS] assumption would be incompatible with many interpretations of signal costs. In particular, the ... [UPS] ... assumption implies that the cost of a particular signal depends on the prior, i.e., on what previous information was observed. Even the answer to the question of whether one signal or another is more costly could depend on the prior. Thus, if $C(\sigma \mid p)$ represents some fixed cost of resources required to conduct an experiment that generates σ (e.g., a drug trial), the ... [UPS] ... assumption is inappropriate."

Based on this logic, Mensch (2018), Denti et al. (2020), and Pomatto et al. (2019) have developed theories of Prior-Invariant information cost functions and advocated for their use in place of the UPS model generally, and Mutual Information in particular.[40] This is not merely a philosophical modeling debate: whether or not the limit condition (5) holds can have important implications in economic applications, especially those in which the prior belief is endogenous. For instance, in dynamic environments where DM's prior belief is determined by previous rounds of information acquisition, (5) is a key determinant of the emergence of belief polarization (Nimark and Sundaresan (2019)). And in strategic settings where DM acquires information about other agent's endogenous actions, (5) can lead to (arguably) counterintuitive pure-strategy equilibria in which DM perfectly monitors others' actions but pays zero cost (because equilibrium actions are deterministic).[41] Indeed, these considerations have led several authors to altogether disavow the use of prior-dependent information cost functions in strategic (Mensch (2018); Denti et al. (2020)) and dynamic (Rustichini (2020)) settings.

[38] The proof is simple. First, any Bounded UPS must have a bounded potential function F, which can be uniquely extended to a continuous bounded convex function $\overline{F} : \Delta \to \mathbb{R}$ by Gale et al. (1968). Then, because the posterior distribution $\pi_{\langle \sigma | p \rangle} \to^{w^*} \delta_{\delta_\theta}$ as $p \to \delta_\theta$, Portmanteau's Theorem applied to $\overline{F}$ directly implies (5). Note that this argument depends crucially on boundedness. A slightly more general version of this result appears as Denti et al. (2020, Corollary 2).

[39] Gentzkow and Kamenica (2014, p. 459) define cost functions over posterior distibutions instead of experiment-prior pairs, and so write $C(\pi)$ instead of $C(\sigma \mid p)$. We maintain our notational convention here at the cost of slightly misquoting that paper.

[40] See Cabrales et al. (2013) and Shorrer (2018) for similar discussions in the context of the value of information.

[41] For instance, Mensch (2018) considers a principal-agent example in which DM, the principal, acquires costly information to monitor an agent's effort. He observes that first-best effort, full monitoring, and zero monitoring can be sustained under Mutual Information monitoring costs, which satisfy the limit condition (5). Ravid (2020) and Denti et al. (2020) construct analogous examples in the context of, respectively, bargaining and zero-sum games.

Prior-Invariance and SLP are Incompatible. How reasonable is the Prior-Invariance Critique? We argue that it has significantly less force than suggested by the literature. In particular, under relatively mild conditions, no nontrivial SLP cost functions is Prior-Invariant. We therefore conclude that the Prior-Invariance Critique implicitly rests on the assumption that DM is either (i) constrained to acquiring information in one-shot or (ii) does not gather information optimally.

The intuition for this incompatibility is straightforward: when DM acquires information sequentially to minimize her *expected* costs, her optimal sequential strategy — and thus the Indirect Cost that it induces — generically depends on her prior beliefs. However, converting this intuition into a formal proof is somewhat delicate because it is a tall order to characterize the form of DM's optimal sequential strategy. In addition, not all SLP cost functions — or even unbounded UPS cost functions, such as Total Information — satisfy a suitable variant of the limit condition (5), so a different proof technique is required.

To state the result, we require a few definitions. First, a Full Domain cost function is a map $C : \mathcal{E} \times \Delta_\circ \to \mathbb{R}_+$ that satisfies points (i) and (ii) in our definition of cost functions from Subsection 2.1, plus a slight strengthening of the continuity condition (iii) therein (see Appendix A.2 for details). The substantive part of this definition is that Full Domain requires *all* experiments to have finite cost. Most cost functions considered in the literature, including Mutual Information satisfy this condition, but Total Information and the LLR cost function do not. Second, call an experiment *nontrivial partitional* if it induces a partition of Θ and is not completely uninformative.[42] Finally, given bounded experiments $\langle S_1, \sigma_1 \rangle$ and $\langle S_2, \sigma_2 \rangle$, define the (bounded) *product experiment* $\langle S_1 \times S_2, \sigma_1 \otimes \sigma_2 \rangle$ by

$$[\sigma_1 \otimes \sigma_2](s_1, s_2 \mid \theta) := \sigma_1(s_1 \mid \theta) \cdot \sigma_2(s_2 \mid \theta), \tag{6}$$

which represents running σ_1 and σ_2 independently and observing the outcomes of both.

Proposition 1. *If C^* is SLP and Prior-Invariant, then the following hold:*

(i) If C^* has Full Domain, then it assigns equal cost to all nontrivial partitional experiments, and therefore is not strictly Blackwell monotone.

(ii) If $C^*(\sigma_1 \otimes \sigma_2) = C^*(\sigma_1) + C^*(\sigma_2)$ for all $\sigma_1, \sigma_2 \in \mathcal{E}_b$, then C^* is identically zero on $\mathcal{E}_b$.[43]

Proof. See Appendix D. □

Most Prior-Invariant cost functions suggested in the literature have Full Domain, including the "channel capacity" cost function suggested by Woodford (2012) and Nimark and Sundaresan (2019) and the "normalized" Mutual Information cost function suggested by Gentzkow and Kamenica (2014) and Denti et al. (2020).[44] Proposition 1(i) states that, subject to mild non-triviality condition, no such cost function is SLP. However, the Full Domain assumption does rule out some interesting and potentially important cost functions, including the LLR costs. Proposition 1(ii) therefore replaces

[42] Formally, $\sigma \in \mathcal{E}$ is *nontrivial partitional* if, for any $p \in \Delta_\circ$, if there exists a partition $\{E_i\}_{i=1}^n$ of Θ with $n \geq 2$ such that $\text{supp}(\pi_{\langle \sigma | p \rangle}) = \{p(\cdot \mid E_i)\}_{i=1}^n$. Note that nontrivial partitional experiments are not bounded because the conditional signal distributions are not mutually absolutely continuous.

[43] For emphasis, we write $C(\sigma)$ instead of $C(\sigma \mid p)$ when C is Prior-Invariant.

[44] Gentzkow and Kamenica (2014) and Denti et al. (2020) advocate for "normalizing" Mutual Information and other Bounded UPS cost functions by evaluating these functions at a *fixed* prior belief p^* that need not coincide with DM's actual prior belief p, thereby making such cost functions Prior-Invariant.

the Full Domain assumption with an additivity condition, under which the Prior-Invariant and SLP classes do not have any nontrivial intersection.[45] This additivity condition is the second main axiom that Pomatto et al. (2019) use, in conjunction with Axiom 5, to characterize the LLR cost function. We explore (a generalization of) this condition in depth in Section 5. For now, simply note that this condition is implied by Indifference to Sequential Learning when the cost function is Prior-Invariant, and therefore yields the following corollary:

Corollary 1.3. *The following hold:*

(i) No non-zero UPS cost function is Prior-Invariant.

(ii) No non-zero LLR cost function is SLP.

Proof. See Appendix K. □

Corollary 1.3(i) generalizes the above finding for Bounded UPS cost functions to all UPS cost functions, including those that violate the limit condition (5), such as Total Information and certain parameterizations of the "Tsallis cost functions" discussed in Caplin et al. (2019b) and Bloedel and Segal (2020).[46] Point (ii) of the corollary illustrates that the LLR cost function, a leading alternative to the rational inattention model based on Mutual Information, cannot be rationalized by flexible cost-minimization.

4 Foundations for (Uniform) Posterior Separability

Since Sims' (2003) introduction of the rational inattention model based on the Mutual Information cost function, a central goal of the subsequent literature has been to extend analyses of information acquisition beyond the Mutual Information functional form. The modern rational inattention literature has adopted the UPS and more general "posterior separable" classes of cost functions as the default modeling assumption. However, as discussed below, the justifications for focusing on these cost functions have, to date, remained somewhat ad hoc.

In this section, we apply the framework developed in Sections 2 and 3 to provide micro-foundations for these classes of cost functions based on sequential cost-minimization. Subsection 4.1 first provides background and a roadmap for the main analysis. Subsections 4.2–4.4 present our main characterization theorems, implications of which are developed in Subsection 4.5. We conlcude in Subsection 4.6 by comparing our results to those of two related papers in the literature.

4.1 Background

The UPS cost functions constitute a sublcass of the more general class of "posterior separable" cost functions. To define these functions, recall that a *divergence* is a function $D : \Delta_\circ \times \Delta_\circ \to \mathbb{R}_+$ that satisfies $D(p \mid p) = 0$; in this paper, we will also always assume that the maps $D(\cdot \mid p) : \Delta_\circ \to \mathbb{R}_+$ are convex for each $p \in \Delta_\circ$. The following definition is adapted from Caplin et al. (2019b):

[45] We conjecture that Proposition 1 can be strengthened to show that any Prior-Invariant and SLP cost function must be identically zero, without any additional conditions. However, to our knowledge, Proposition 1 and Corollary 1.3 cover all cost functions of applied interest in the literature.

[46] While it is asserted with some frequency in the literature that no UPS cost function is Prior-Invariant, to our knowledge this fact has not been formally shown, and certainly does not follow from the same argument used for Bounded UPS costs.

Definition 8 (Posterior Separable)**.** *Cost function C is **Posterior Separable** if there exists a divergence D such that*[47]

$$C(\sigma \mid p) \equiv \mathbb{E}_{\pi_{\langle \sigma|p\rangle}}[D(\tilde{q} \mid p)].$$

All Posterior Separable cost functions are Blackwell monotone because divergences are convex in the posterior. Also note that any Posterior Separable cost function C can be equivalently represented as $C(\sigma \mid p) = \mathbb{E}_{\pi_{\langle \sigma|p\rangle}}[F(q \mid p) - F(p \mid p)]$, where $F(\cdot \mid p) : \Delta_\circ \to \mathbb{R}_+$ is a *prior-dependent* (convex) potential function.[48] Thus, every UPS cost function is Posterior Separable. For a UPS cost function with potential F, the divergence $D_F(q \mid p) := F(q) - F(p) - \nabla F(p) \cdot (q - p)$ is called the *Bregman divergence* generated by F. However, te Posterior Separable class is quite large, and is not contained in the SLP class. For instance, it is easy to see from (LLR) that the LLR cost functions are Posterior Separable, although they are not SLP by Corollary 1.3. Note also that the distance-based cost functions (3) are Posterior Separable (but not UPS).

The literature has focused on Posterior Separable, and in particular UPS, cost functions for several reasons. First, they retain much of the structure of the popular Mutual Information cost function but, by relaxing its specific functional form, are able to generate much richer patterns of behavior (cf. Caplin et al. (2019b); Dean and Neligh (2019)). Second, they are uniquely tractable: the Posterior Separable class is the largest class of cost functions for which optimal (one shot) strategies can be found via the concavification method popularized by the Bayesian persuasion literature (Gentzkow and Kamenica (2014); Caplin et al. (2019b)). Third, they are normatively appealing: it is well known that the Posterior Separable class is characterized by the *Indifference to Randomization* condition

$$C(\alpha\sigma_1 \oplus (1-\alpha)\sigma_2 \mid p) = \alpha C(\sigma_1 \mid p) + (1-\alpha)C(\sigma_2 \mid p) \tag{IR}$$

for all $\sigma_1, \sigma_2 \in \mathcal{E}_b$, $p \in \Delta_\circ$, and $\alpha \in [0,1]$, which is analogous to the linearity axiom that underlies expected utility theory (Torgersen (1991, pp. 353-54); Mensch (2018)). Fourth, they uniquely characterize the *value* of information for a Bayesian expected-utility maximizer (Azrieli and Lehrer (2008); Frankel and Kamenica (2019)).

Roadmap. This section develops four main results (illustrated schematically in Figure 1). In Subsection 4.2, we first characterize the Posterior Separable class as the Indirect Cost functions arising from a *restricted* optimization problem in which DM only has access to direct Poisson strategies. A corollary of this characterization is that every SLP cost function that is "locally once-differentiable" in the experiment is, in fact, Posterior Separable.

Subsections 4.3 and 4.4 then zoom in on, and provide two complementary characterization theorems for, the UPS class. The first characterization, Theorem 2, shows that any SLP cost function that is "locally once-differentiable" in the experiment *and* differentiable in the prior is, in fact, UPS. This suggests that, under standard regularity conditions assumed in applications, the SLP and UPS classes exactly coincide, i.e., the "preference for sequential learning" condition (PSL) condition automatically holds. Moreover, this implies sequential optimization wipes out any distinction between

[47] The divergence D is unique up to translation by mean-zero linear functions, i.e., $D(q \mid p)$ represents C if and only if $D(q \mid p) + \zeta \cdot (q - p)$ does as well, where $\zeta \in \mathbb{R}^\Theta$.

[48] Given a divergence representation as in Definition 8, we may set $F(q \mid p) := D(q \mid p)$. Conversely, given a potential function representation, $D(q \mid p) := F(q \mid p) - F(p \mid p) - \nabla_q F(q \mid p)|_{q=p} \cdot (q - p)$ can be used to generate a divergence representation. (If $F(\cdot \mid p)$ is not differentiable, $\nabla_q F(q \mid p)$ denotes a subgradient at q.)

Posterior Separable and UPS cost functions. However, the second characterization, Theorem 3, shows that these regularity conditions are economically meaningful: any UPS Indirect Cost is generated *only* by Direct Cost functions exhibiting a weaker version of (PSL) called "preference for incremental learning," meaning that it is cheaper to acquire information in the form of Gaussian diffusion signals than it is to acquire information in one shot. In Subsection 4.5, we use this latter characterization to show that, generically, UPS Indirect Cost functions cannot be generated by Prior-Invariant Direct Costs.

The relation between UPS cost functions and Gaussian diffusion learning was first discovered in two related papers, Morris and Strack (2019) and Hébert and Woodford (2020b). The results of this section build on, but significantly extend and maximally generalize, the results of those papers. We discuss the relation between those papers and ours in more detail in 4.6.

4.2 Posterior Separable Characterization

First-Order Approximation. Our characterization of Posterior Separable cost functions — as well as our first characterization of the UPS class in the next subsection — relies on a first-order approximation of DM's cost function that characterizes the cost of a "small amount of information" that arrives in the form of Poisson signals. To that end, define the *direct Poisson Indirect Cost* generated by Direct Cost function C as

$$\Phi_{DP}C(\sigma \mid p) := \lim_{\alpha \to 0} \frac{C_{RA}(\alpha \cdot \sigma \mid p)}{\alpha} \tag{DPIC}$$

where $\alpha \cdot \sigma$ is the α-dilution of σ defined in (4) and $C_{RA} \in \mathcal{C}$ denotes the *lower Randomizaton Averse envelope* of C, i.e., the pointwise largest Randomizaton Averse cost function that is majorized by C.[49] Thus, $\Phi_{DP}C$ represents the Indirect Cost generated by Direct Cost C in the *restricted* optimization problem in which DM only has access to mixed and direct Poisson strategies. The appropriate notion of first-order approximation is then defined as follows:

Definition 9 (Locally Linear)**.** *Cost function C is Locally Linear if there exists a divergence $D : \Delta^2_\circ \to \mathbb{R}_+$ such that*

$$\Phi_{DP}C(\sigma \mid p) = \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[D(\tilde{q} \mid p)] \tag{LL}$$

for all $(\sigma, p) \in \mathcal{E}_b \times \Delta_\circ$.

A Direct Cost function is Locally Linear if the expected cost of direct Poisson learning is Posterior Separable. An equivalent way of stating (LL) is that $C(\alpha \cdot \sigma \mid p) = \mathbb{E}_{\pi_{\langle \alpha \cdot \sigma | p \rangle}}[D(q \mid p)] + o(\alpha)$, meaning that the cost of acquiring σ with probability α is approximately Posterior Separable as $\alpha \to 0$. In this sense, we may interpret the divergence $D(q \mid p)$ as the cost of an infrequently-arriving Poisson signal that causes the prior p to jump to the posterior q. Clearly, any smooth transformation of a Posterior Separable cost function is Locally Linear.

Mathematically, (LL) states that $C_{RA}(\cdot \mid p)$ is directionally differentiable at the uninformative experiment $\underline{\sigma}$ and, moreover, that its directional derivative at this point is continuous and linear.[50]

[49] We show in the appendix that $\Phi_{DP}C$ and C_{RA} are well-defined cost functions.

[50] Because the spaces of experiments and posterior distributions are infinite-dimensional, continuity of the directional derivative does not imply that it is linear. For instance, the the pointwise maximum of two Posterior Separable cost functions is always Randomizaton Averse and Dilution Linear, but is typically not Locally Linear.

Indeed, we may equivalently view the direct Poisson operator Φ_{DP} as taking the directional derivative

$$\Phi_{DP}C(\sigma \mid p) = \frac{\partial}{\partial \epsilon} C_{RA}(\epsilon(\sigma - \underline{\sigma}) \oplus \underline{\sigma} \mid p)\Big|_{\epsilon=0},$$

and (LL) states that this directional derivative satisfies the linearity condition (IR), meaning that it is linear with respect the posterior distribution $\pi_{\langle \sigma|p \rangle}$ induced by the experiment σ at prior p. We emphasize that Definition 9 only a "local" differentiability condition — it is only required to hold at the uninformative experiment — and generally has no implications for global properties of a cost function.

It might seem like this notion of directional differentiability is specifically tailored to generate a Posterior Separable approximation. However, if we were to restrict attention to experiments that generate fewer than $n \in \mathbb{N}$ signals, Definition 9 would be implied by the standard notion of continuous directional differentiability in Euclidean space — *without* presupposing additive separability with respect to signals.[51] Such a finite-dimensional differentiability condition is quite weak: $C_{RA}(\cdot \mid p)$ is convex by construction, and finite-dimensional convex functions are known to be continuously differentiable almost everywhere. However, since we do not impose such an upper bound on the allowable number of signals, Definition 9 further requires that the approximation error generated by the standard differentiability notion be uniform in the number n of signals.

Characterization. The following lemma uses the above notion of first-order approximation to characterize the full Posterior Separable class as Indirect Cost functions from the restricted optimization problem in which DM only has access to mixed and direct Poisson strategies:

Lemma 2. *Given any cost function C^*, the following are equivalent:*

(i) C^* *is Posterior Separable.*

(ii) $C^* = \Phi_{DP}C^*$ *and is Locally Linear.*

(iii) Every Randomizaton Averse Direct Cost C for which $C^* = \Phi_{DP}C$ is Locally Linear.

Proof. See Appendix K. □

How does the Posterior Separable class relate to the SLP class, i.e., when DM is *not* restricted to direct Poisson strategies? The intersection is clearly nontrivial, as we have already seen that (a) every UPS cost function is SLP but not conversely, and (b) there exists a large class of Posterior Separable cost functions that are not SLP, such as those that are Prior-Invariant (with the LLR costs being a leading example).

Proposition 2. *Given an SLP cost function C^*, the following hold:*

(i) C^* *is Posterior Separable if and only if it is Locally Linear.*

(ii) Let C be a Direct Cost function. Then C^* is Posterior Separable and satisfies $C^* = \Phi_{DP}C$ if and only if (a) C is Locally Linear and (b) $\Phi_{DP}C$ exhibits Preference for One-Shot Learning.

Proof. See Appendix K. □

[51] Recall that an experiment with signal space S satisfying $|S| = n$ can be viewed as a $|\Theta| \times n$-dimensional Markov transition matrix. The space of such experiments is therefore a convex subset of $\mathbb{R}_+^{|\Theta| \times n}$.

Proposition 2(i) states that the intersection of these two classes of cost functions is precisely characterized by Local Linearity. Practically, this means that any mildly smooth SLP cost function must, in fact, be Posterior Separable. This provides an arguably compelling foundation for the use of Posterior Separable cost functions in applications — at least those that exhibit Preference for One-Shot Learning.

Proposition 2(ii) partially characterizes the class of Direct Cost functions that could generate a Posterior Separable Indirect Cost. It states that any Indirect Cost that is Posterior Separable and that can be attained by direct Poisson strategies must have been generated by a Locally Linear Direct Cost whose Posterior Separable approximation is itself SLP. In other words, the "first derivative" of a Direct Cost function is preserved under the full sequential optimization process if and only if direct Poisson learning is without loss of optimality.

4.3 UPS Characterization 1: Indirect Costs

In this subsection and the next, we characterize the UPS cost functions within the SLP class. Given Proposition 2(i), one might conjecture that every Locally Linear SLP cost function is not only Posterior Separable, but actually UPS. However, it is easy to construct counterexamples to this conjecture: witness the distance-based cost functions in (3).

In this subsection, we show that such counterexamples are special: any SLP cost function that is both Locally Linear *and* differentiable with respect to the prior (in a suitable sense) must, in fact, be UPS. This additional smoothness condition is formalized as follows:

Definition 10 (Regular)**.** *Cost function C is Regular if it is Locally Linear and the divergence D is continuously differentiable with respect to the prior, i.e., there exists a continuous vector-valued function $J : \Delta_\circ \times \Delta_\circ \to \mathbb{R}^\Theta$, denoted $J(q \mid p)$, such that*

$$\lim_{\epsilon \downarrow 0} \frac{D(q \mid p + \epsilon(r-p)) - D(q \mid p)}{\epsilon} = J(q \mid p) \cdot (r - p) \tag{7}$$

for all $p, q, r \in \Delta_\circ$.

Definition 10 is a fairly weak regularity condition that is satisfied by Total Information, Mutual Information and, to our knowledge, virtually all other information cost functions commonly used in economic applications. Under the hypotheses of the following theorem, it implies that the cost function itself satisfies $C^*(\sigma \mid \cdot) \in \mathbf{C}^1(\Delta_\circ)$ for all $\sigma \in \mathcal{E}_b$.

Theorem 2. *Given any cost function C^*, the following are equivalent:*

(i) C^* is SLP and Regular, with divergence D.

(ii) C^* is UPS with potential $F \in \mathbf{C}^2(\Delta_\circ)$, for which the Bregman divergence $D_F = D$.

Proof. See Appendix E. □

Theorem 2 states that any "smooth" SLP cost function must, in fact, be UPS. Thus, under regularity conditions typically assumed in applications, checking that a cost function is SLP is equivalent to testing that it is UPS. This provides an arguably compelling reason to focus on the UPS cost functions, especially given that Preference for One-Shot Learning may be difficult to verify while UPS is

easy to verify. In light of Proposition 2, the theorem also implies that, subject to the differentiability condition (7), the sequential optimization process eliminates all Posterior Separable cost functions outside of the UPS subclass.

Two aspects of the statement of Theorem 2 warrant emphasis. First, the restriction in point (ii) to twice continuously-differentiable potential functions, while not completely innocuous, is without loss of significant generality because such potential functions are dense in the space of all convex potential functions.[52] Second, as witnessed by examples in Subsection 3.2.2, the hypothesis in point (i) that C^* is Regular is the near-minimal smoothness condition needed for a result of this sort, even if we were to weaken the (generic) desideratum that $F \in \mathbf{C}^2(\Delta_\circ)$.

That point (ii) implies point (i) of the theorem is straightforward. The proof of the nontrivial direction, that point (i) implies point (ii), consists of several steps. First, we use the facts that C^* is Posterior Separable (by Proposition 2) and exhibits Preference for One-Shot Learning (by Theorem 1) to show that the derivative of the divergence in (7) must have the representation $J(q \mid p) = -k(p)(q-p)$ for some matrix-valued function $k(p)$, i.e., be linear in the posterior q. This allows us to express the divergence itself as an integral of this linear form. By successively differentiating this integral representation of the divergence, we show that $D(\cdot \mid p)$ is indeed twice continuously differentiable and, moreover, that its Hessian at q is $k(q)$, and therefore independent of p. This latter fact is used to establish that D is, in fact, a Bregman divergence.

Relation to Banerjee et al. (2005). Theorem 2 is related to a characterization of Bregman divergences established by Banerjee et al. (2005). They show that, subject to regularity conditions, a divergence D satisfies $\mathbb{E}_\pi[q] \in \arg\min_{\nu \in \Delta_\circ} \mathbb{E}_\pi[D(q \mid \nu)]$ for all $\pi \in \Pi_b$ if and only if it is a Bregman divergence. By way of comparison, the first step in our proof of Theorem 2 establishes that Preference for One-Shot Learning implies that $\nu^* = \mathbb{E}_\pi[q]$ must be a *critical point* — although not necessarily a minimizer of — the function $\mathbb{E}_\pi[D(q \mid \cdot)]$. The proof of Banerjee et al. (2005, Theorem 4) can be used to establish the following weaker version of our Theorem 2: If an SLP cost function is Locally Linear with a divergence D that is *jointly* twice continuously differentiable in (q, p), then it is UPS with potential function $F \in \mathbf{C}^3(\Delta_\circ)$.[53] However, our Theorem 2 establishes that many of these smoothness conditions are superfluous for the result.

4.4 UPS Characterization 2: Direct Costs

Theorem 2 seems to provide compelling foundations for use of UPS cost functions, as is standard in the rational inattention literature. However, it is important to note that the regularity conditions used to establish this theorem are placed on DM's Indirect Cost function, which is the output of the sequential optimization problem (IC). It is not *a priori* clear whether we should expect sequential

[52] In particular, given any convex potential function $F : \Delta_\circ \to \mathbb{R}$, there exists a sequence $\{F_n\}_{n \in \mathbb{N}}$ of convex potential functions with $F_n \in \mathbf{C}^2(\Delta_\circ)$ such that $F_n \to F$, and this convergence can be taken to be uniform on each Δ_δ. Also recall that every $\sigma \in \mathcal{E}_b$ and $p \in \Delta_\circ$ induces posterior distribution $\pi_{\langle \sigma | p \rangle} \in \Pi_\delta$ for some $\delta > 0$. Thus, the uniform convergence of the potential functions on compact subsets implies that $C_n(\sigma \mid p) \to C(\sigma \mid p)$ for each $\sigma \in \mathcal{E}_b$ and $p \in \Delta_\circ$, where C_n is the UPS cost function induced by F_n and C is the UPS cost function induced by F.

[53] In independent work, Hébert and Woodford (2020b, Lemma 4) establish a result that is equivalent to this weaker version of our Theorem 2 by appealing to Banerjee et al. (2005, Theorem 4). Aside from being established under more stringent technical conditions, their result has different conceptual content. Namely, they *assume* that DM's *Direct* Cost function is Posterior Separable and exhibits Preference for One-Shot Learning, rather than using optimality to derive these as properties of her Indirect Cost.

optimization to preserve such smoothness conditions and, therefore, what the economic meaning of Definition 10 actually is. To shed light on this question, in this subsection we characterize the class of Direct Cost functions that generate UPS Indirect Cost functions. We find that, in fact, Regularity of the Indirect Cost corresponds to a restrictive condition on the underlying Direct Cost function.

Preview: Gaussian Diffusion Learning. Our characterization is based on the following class of sequential strategies that arise in the continuous-time limit of our framework. We say that DM engages in *Gaussian diffusion learning* if her posterior belief follows a (continuous-time) diffusion process $(q_t)_{t\geq 0}$ in the simplex of the form

$$dq_t = v_t dW_t$$

where $(W_t)_{t\geq 0}$ is a $|\Theta|$-dimensional standard Brownian motion and $(v_t)_{t\geq 0}$ is a W-adapted matrix-valued volatility process chosen by DM, where $v_t \in \mathbb{R}^{\Theta\times\Theta}$ satisfies $\mathbf{1}^\top v_t = \mathbf{0}^\top$ because beliefs must stay in the simplex. In continuous-time models of sequential sampling, such belief dynamics arise when DM observes a real-valued signal process that itself follows a diffusion, and can dynamically control that processes' state-dependent drift and volatility matrix. In our framework, per the standard random walk approximation of diffusions, such belief dynamics arise in the limit of Sequential Replications in which (i) DM does not dispose of any information, (ii) DM acquires vanishingly-informative Bernoulli (i.e., binary-signal) experiments in each acquisition period, and (iii) the time horizon $T \to \infty$.[54]

If the Direct Cost C is "locally twice differentiable" in the sense that $C(\sigma \mid p) \approx \mathbb{E}_{\langle\sigma|p\rangle}[(\tilde{q}-p)k(p)(\tilde{q}-p)]$ for some matrix-valued function $k(p) \in \mathbb{R}^{\Theta\times\Theta}$ for experiments that induce posteriors q within an arbitrarily small distance of the prior p, we can use Ito's Lemma and the Optional Sampling Theorem to compute the *Gaussian Indirect Cost* $\Phi_G C$ generated by C as

$$\Phi_G C(\sigma \mid p) = \inf_{(v_t)_{t\geq 0},\tau} \mathbb{E}\left[\int_0^\tau \operatorname{tr}\left[v_t^\top k(q_t) v_t\right] dt\right], \quad \text{s.t. } q_\tau \sim \pi_{\langle\sigma|p\rangle} \tag{GIC}$$

where minimization is with respect to W-adapted volatility process $(v_t)_{t\geq 0}$ and W-adapted stopping times τ, and the expectation is with respect to the induced paths of posterior beliefs. If, moreover, $k(q) = \frac{1}{2}\mathcal{H}F(q)$ for some potential function $F \in \mathbf{C}^2(\Delta_\circ)$, the above expression simplifies to

$$\begin{aligned}\Phi_G C(\sigma \mid p) &= \mathbb{E}\left[\int_0^\tau \frac{1}{2}\operatorname{tr}\left[v_t^\top \mathcal{H}F(q_t) v_t\right] dt\right] \\ &= \mathbb{E}_{\pi_{\langle\sigma|p\rangle}}[F(q) - F(p)]\end{aligned} \tag{8}$$

for *any* volatility process and stopping time for which $q_\tau \sim \pi_{\langle\sigma|p\rangle}$, meaning that the Gaussian Indirect Cost (GIC) is UPS and, moreover, DM is indifferent among all Gaussian replications of a given experiment.

This calculation (formalized in Lemma 4 below) suggests a close connection between UPS cost functions and optimization over Gaussian diffusion strategies, versions of which have been discovered by Morris and Strack (2019) and Hébert and Woodford (2020b). In this subsection, we show

[54] This approximation is formalized in the proof of Theorem 3(i) (see Appendix F).

that, in a particular sense, this is the *only* way to generate a UPS Indirect Cost function. Formally, subject to mild technical conditions, the Indirect Cost of information is UPS if and only if the Direct Cost of information satisfies two properties: (i) Gaussian diffusion learning dominates one-shot learning and (ii) DM is indifferent among all Gaussian diffusion replications.

4.4.1 Axioms and Technical Conditions

The statement of our characterization theorem requires two technical conditions and one economically substantive axiom. We describe these conditions in turn.

Conditions on Direct Cost. To state our key axiom on the Direct Cost, we must first introduce a technical condition that formalizes the idea that C is "locally twice continuously differentiable" with respect to posterior distributions, where "locally" means that this differentiability is imposed only on posterior distributions supported on some sufficiently small δ-ball $B_\delta(p) := \{q \in \Delta_\circ : |q-p| \leq \delta\}$ around the prior p. Formally, we require that the Direct Cost admits the following kind of second-order Taylor expansion:

Definition 11 (Locally Quadratic)**.** *Cost function C is Locally Quadratic if there exists continuous matrix valued function $k : \Delta_\circ \to \mathbb{R}^{\Theta\times\Theta}$ such that for each $p_0 \in \Delta_\circ$ and $\epsilon > 0$, there exists some $\delta > 0$ such that*

$$\left|C(\sigma \mid p) - \mathbb{E}_{\pi_{\langle\sigma|p\rangle}}\left[(q-p)k(p)(q-p)\right]\right| \leq \epsilon \cdot \mathbb{E}_{\pi_{\langle\sigma|p\rangle}}[\|q-p\|^2] \tag{LQ}$$

*for all $\sigma \in \mathcal{E}_b$ and $p \in B_\delta(p_0)$ for which $\operatorname{supp}(\pi_{\langle\sigma|p\rangle}) \subseteq B_\delta(p_0)$. We refer to $k(p)$ as the **kernel** of $C(\cdot \mid p)$.*[55]

In words, C is Locally Quadratic if the cost of an experiment that only shifts beliefs locally is approximated by a quadratic form representing the weighted variance of belief movement. Definition 11 is essentially the minimal smoothness condition that allows us to compute the expected cost of Gaussian replication using Ito's lemma. We emphasize that, because we are only concerned with Gaussian replications and diffusion processes have continuous sample paths, Definition 11 is only a local condition that generally does not have any implications for global properties of the cost function.

When C is Posterior Separable, it is easy to check if it is Locally Quadratic:

Lemma 3. *The following hold:*[56]

(i) If C is Posterior Separable with divergence D, then it is Locally Quadratic if and only if the map $p \mapsto \mathcal{H}_q D(q \mid p)|_{q=p}$ is well-defined and continuous on $\Delta_\circ$, in which case its kernel is $k(q) = \frac{1}{2}\mathcal{H}_q D(q \mid p)|_{q=p}$.

(ii) If C is UPS with potential function F, then it is Locally Quadratic if and only if $F \in \mathbf{C}^2(\Delta_\circ)$, in which case its kernel is $k(q) = \frac{1}{2}\mathcal{H}F(q)$.

[55] It is without loss of generality to assume that, for each $p \in \Delta_\circ$, the matrix $k(p)$ is symmetric, satisfies $k(p)p = \mathbf{0}$, and is positive semi-definite on the tangent space of the probability simplex (i.e., satisfies $y^T k(p) y \geq 0$ for all $y \in \mathbb{R}^\Theta$ such that $y \cdot \mathbf{1} = 0$). Moreover, continuity of $k(\cdot)$ is in fact equivalent to the seemingly weaker condition that the quadratic form $y^T k(\cdot) y$ is continuous for all $y \cdot \mathbf{1} = 0$. Because $(q-p)$ in (LQ) always adds up to 0, we are essentially considering a $|\Theta|-1$-dimensional subspace, in which the bilinear form $\widetilde{k}(p) = [I, -\mathbf{1}] \cdot k(p) \cdot [I, -\mathbf{1}]^T \in \mathbb{R}^{(|\Theta|-1)\times(|\Theta|-1)}$ is uniquely pinned down. The full matrix $k(p)$ is only unique up to the addition of terms of the form $\mathbf{f}(p)\mathbf{1}^\top$ and $\mathbf{1}\mathbf{f}(p)^\top$, where $\mathbf{f} : \Delta_\circ \to \mathbb{R}^\Theta$.

[56] As in Footnote 55, the Hessian $\mathcal{H}F(q)$ of a function $F \in \mathbf{C}^2(\Delta_\circ)$ at point q is only unique up the addition of terms that preserve its value as a quadratic form on the tangent space to the probability simplex.

Proof. See Appendix K. □

For more general cost functions, the form of the Taylor expansion (LQ) might seem restrictive because (i) there is no first-order term, (ii) the second-order term is additively separable in posterior beliefs, and (iii) it corresponds to differentiability in the infinite-dimensional space of posterior distributions. However, we show in the appendix that Definition 11 is implied by the seemingly much weaker condition that C is locally twice continuously-differentiable with respect to experiments with (i) finitely-many signals and (ii) uniformly vanishing likelihood ratios, and that the approximation error induced by this second derivative is uniform in the number of signals. Except for the latter uniformity requirement, this alternative condition is equivalent to the standard definition of local twice differentiability in Euclidean space. At least when C is Randomizaton Averse, this is a weak requirement because finite-dimensional convex functions are known to be twice differentiable almost everywhere (Alexandrov's theorem). For future reference, we note that the local second derivative of C in experiment space is given by the "normalized kernel"

$$\overline{k}(p) := \text{Diag}(p)k(p)\text{Diag}(p), \tag{9}$$

where k is the kernel defined above and $\text{Diag}(p)$ is the matrix with diagonal entries determined by p and all other entries zero. As shown in Appendix K, the "normalization" of $\overline{k}(p)$ by these diagonal matrices corresponds to a change of variables from posterior distributions to experiments.

We may now state our main condition on the Direct Cost:

Axiom 7 (Preference for Incremental Learning)**.** *Suppose that C is Locally Quadratic with kernel k. We say that C exhibits Preference for Incremental Learning if:*

(i) The kernel satisfies $k(p) \equiv \frac{1}{2}\mathcal{H}F(p)$ for some potential function $F \in \mathbf{C}^2(\Delta_\circ)$.

(ii) The cost function C satisfies

$$C(\sigma \mid p) \geq \mathbb{E}_{\pi_{\langle \sigma|p\rangle}}[F(q) - F(p)]. \tag{PIL}$$

for all $\sigma \in \mathcal{E}_b$ and $p \in \Delta_\circ$.

We can understand the two conditions in Axiom 7 as follows. First, point (i) is an "integrability" condition on the kernel of C. This may seem like a technical condition. However, it has an intuitive economic characterization, which maximally generalizes Morris and Strack's (2019) characterization of UPS cost functions in the $|\Theta| = 2$ case:

Lemma 4. *Given a Locally Quadratic Direct Cost C with kernel k, the following are equivalent:*

(i) The kernel satisfies $k(q) \equiv \frac{1}{2}\mathcal{H}F(q)$ for potential function $F \in \mathbf{C}^2(\Delta_\circ)$.

(ii) The Gaussian Indirect Cost $\Phi_G C$ is UPS with potential $F \in \mathbf{C}^2(\Delta_\circ)$.

(iii) For each $\sigma \in \mathcal{E}_b$ and $p \in \Delta_\circ$, all Gaussian replications of $\pi_{\langle \sigma|p\rangle}$ have the same expected cost.

Proof. See Appendix K. □

Second, point (ii) of Axiom 7 states that the expected cost of Gaussian replication is always weakly lower than the cost of one-shot learning. This naturally suggests that Gaussian learning is the optimal form of Sequential Replication, as any individual step in a non-Gaussian replication could be replaced with its Gaussian replication (the proof of Theorem 3 formalizes this intuition). Thus, in sum,

Preference for Incremental Learning represents a restricted form of the "preference for sequential learning" condition (PSL) according to which (i) DM is indifferent among all Gaussian replications and (ii) weakly prefers Gaussian replication to one-shot learning. However, unlike (PSL), it does not directly restrict her preferences over other kinds of Sequential Replication.

Conditions on Indirect Cost. Before stating the theorem, we require one additional technical condition, which will be imposed on DM's Indirect Cost function:

Definition 12 (Locally Strongly Convex)**.** *C is Locally Strongly Convex if there exists $\delta > 0$ and $m > 0$ such that $C(\sigma \mid p) \geq m \cdot \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}\left[\|q - p\|^2\right]$ for all $(\sigma, p) \in \mathcal{E}_b \times \Delta_\circ$ such that $supp(\pi_{\langle \sigma | p \rangle}) \subseteq B_\delta(p)$.*

Intuitively, C is Locally Strongly Convex if the marginal cost of sampling from a Gaussian diffusion for an additional instant is strictly positive. The above definition formalizes this intuition by requiring that the "unit cost" of (sufficiently small) belief variance is bounded below by $m > 0$. As discussed in Appendix K, Definition 12 can be equivalently formulated in terms of the kernel k when C is Locally Quadratic, which makes it easy to check for Posterior Separable cost functions by virtue of Lemma 3. We note here that Mutual Information, Total Information, and LLR cost functions are all Locally Strongly Convex.[57]

4.4.2 Characterization

Characterization Theorem. We may now state our characterization theorem. It states that, subject to the aforementioned technical conditions, an Indirect Cost function is UPS if and only if every Direct Cost that generates it exhibits Preference for Incremental Learning:

Theorem 3. *Let the Direct Cost function C be Locally Quadratic with kernel k. The following hold:*

(i) If C exhibits Preference for Incremental Learning, then $\Phi(C)$ is UPS with potential $F \in \mathbf{C}^2(\Delta_\circ)$ where $\mathcal{H}F(q) \equiv 2k(q)$.

(ii) Let the Direct Cost function C also be Blackwell monotone. If $\Phi(C)$ is UPS and Locally Strongly Convex with potential F, then $F \in \mathbf{C}^2(\Delta_\circ)$ and C exhibits Preference for Incremental Learning with $k(q) \equiv \frac{1}{2}\mathcal{H}F(q)$ and is Locally Strongly Convex.

Proof. See Appendix F. □

Economically, Theorem 3 says that UPS Indirect Cost functions are generated *only* by Direct Costs for which it is always optimal to acquire information using *only* Gaussian diffusion signals. This is a stringent condition on the Direct Cost, for it rules out any strict cost savings from acquiring chunks of information.

The sufficiency direction, point (i), is a small conceptual step away from Lemma 4. That lemma tells us that Gaussian replication generates a UPS Indirect Cost when the Direct Cost satisfies the integrability condition Axiom 7(i). To establish Theorem 3(i), it therefore suffices to invoke the Preference for Incremental Learning inequality (PIL) to show that the restriction to Gaussian learning is without loss of optimality when DM's strategy space is not exogenously restricted (which includes

[57] An example of a cost function violating this condition is the Posterior Separable cost with divergence $D(q \mid p) = \|q - p\|^4$.

showing that free disposal is never optimal). The formal proof requires some technical work to, in effect, take a continuous-time limit of our discrete-time notion of Sequential Replication so as to approximate Gaussian diffusion signals.

As discussed further in Subsection 4.6 below, this sufficiency result generalizes (a) the aforementioned result of Morris and Strack (2019) by allowing for $|\Theta| \geq 3$ and by dropping the assumption that DM is exogenously restricted to Gaussian strategies, and (b) a related result of Hébert and Woodford (2020b) by, among other things, relaxing the technical conditions imposed on the Direct Cost function.

However, we view the necessity direction, point (ii), as the main contribution of Theorem 3. This necessity result, which has no parallels in the aforementioned papers, does two things. Most directly, it tells us that the optimality of Gaussian learning is the *only* way to generate an Indirect Cost in the UPS class. Indirectly, in conjunction with the sufficiency direction Theorem 3(i), it also provides a way to characterize the full set of (Locally Quadratic) Direct Costs that could have generated a given UPS Indirect Cost. We provide examples below.

In essence, the proof of the necessity direction, point (ii), shows that the kernel k of the Direct Cost C is preserved under sequential optimization whenever ΦC is UPS. To show this, we show that both the Direct and Indirect Cost of an incrementally informative experiment that only moves posterior beliefs locally away from the prior can be quadratically approximated (in the sense of Definition 11) by the same kernel k. A subtle technical difficulty arises from the fact that, in principle, the possibility of free information disposal means that DM's optimal Sequential Replication of such an incrementally informative experiment may involve acquiring superfluous information that moves posterior beliefs far away form the prior, but is ultimately discarded. The hypothesis that the Indirect Cost is Locally Strongly Convex allows us to place an upper bound on the amount of information that is discarded in an (approximately) optimal Sequential Replication.

Stronger Version of Necessity. The above discussion suggests that Preference for Incremental Learning is actually much more powerful than stated in Theorem 3(ii), the proof of which only utilizes the invariance of second derivatives under Φ but not this operator's fine details. In fact, the proof of Theorem 3(ii) yields the following stronger necessity result:

Corollary 3.1. *Let C be Locally Quadratic and Blackwell monotone. For any operator Φ' such that $\Phi \leq \Phi' \leq \mathrm{Id}$: If $\Phi'(C)$ is UPS and Locally Strongly Convex, then C exhibits Preference for Incremental Learning.*

Proof. See Appendix K. □

The function $\Phi'(C)$ in Corollary 3.1 represents the Indirect Cost from a more restricted optimization problem than (IC), for which we only assume that DM is able to do at least weakly better than acquiring information in one-shot. Therefore, Corollary 3.1 states that Preference for Incremental Learning is a necessary condition for a Direct Cost function to generate a UPS Indirect Cost almost independently of the underlying optimization problem. Thus, our standing assumption that DM's strategy space is fully flexible is not needed for the result.

Characterization of Rationalizing Direct Costs. We now show how Theorem 3 and Lemma 3 can be used to characterize the full set of (Locally Quadratic) Direct Cost functions that rationalize par-

ticular UPS Indirect Cost functions of interest.

Lemma 5. *Let C be a Locally Quadratic and Blackwell monotone Direct Cost function with kernel k. Then:*

(i) The Indirect Cost ΦC is Mutual Information if and only if C majorizes Mutual Information and k is equivalent to[58] the "Fisher information matrix"

$$g(p) := Diag(p)^{-1} - \mathbf{1}\mathbf{1}^\top. \tag{10}$$

(ii) The Indirect Cost ΦC is Total Information if and only if C majorizes Total Information and k is equivalent to $Diag(p)^{-1}\bar{k}_{TI}(p)Diag(p)^{-1}$, where the normalized kernel $\bar{k}_{TI}$ is defined componentwise by

$$\left[\bar{k}_{TI}(p)\right]_{\theta,\theta'} := \begin{cases} p_\theta \sum_{\theta'' \neq \theta} \gamma_{\theta,\theta''} + \sum_{\theta'' \neq \theta} p_{\theta''}\gamma_{\theta'',\theta}, & \text{if } \theta = \theta' \\ -p_\theta \gamma_{\theta,\theta'} - p_{\theta'}\gamma_{\theta',\theta}, & \text{if } \theta \neq \theta' \end{cases}$$

Proof. See Appendix K. □

4.4.3 Special Case: Locally Linear Direct Cost

Preference for Incremental Learning is a global property of the Direct Cost function that may be difficult to verify in practice. But when the Direct Cost is Locally Linear, it can be equivalently stated in a "local" form that is easier to check:

Lemma 6. *Let the Direct Cost C be Locally Linear with divergence D and Locally Quadratic with kernel k. If $k = \frac{1}{2}\mathcal{H}F$ for some convex function F, then the following are equivalent:*

(i) C exhibits Preference for Incremental Learning.

(ii) The Bregman divergence D_F generated by F satisfies

$$D(q \mid p) \geq D_F(q \mid p) \tag{PGL}$$

for all $q, p \in \Delta_\circ$.

Proof. See Appendix G. □

Recall from Subsection 4.2 that the divergence $D(q \mid p)$ in a Locally Linear approximation represents the cost of an infrequently-arriving Poisson signal that moves the prior p to the posterior q. Thus, we can interpret (PGL) as stating that it is always weakly cheaper to diffuse from p to q than it is to directly jump there. In a related setting (discussed further below), Hébert and Woodford (2020b) refer to the condition (PGL) as a *Preference for Gradual Learning* and provide an analogous interpretation.[59]

One might conjecture that, when the Direct Cost C is Locally Linear, the Indirect Cost is UPS if and only if the divergence approximation D of C is a Bregman divergence. Indeed, Proposition 2(ii) implies that D being a Bregman divergence is sufficient for the Indirect Cost to be UPS. Lemma 6

[58] In the sense that $y^\top k(p)y = y^\top g(p)y$ for all $p \in \Delta_\circ$ and all $y \in \mathbb{R}^\Theta$ such that $y \cdot \mathbf{1} = 0$.

[59] The definition of Preference for Gradual Learning in Hébert and Woodford (2020b) allows for the kernel k to violate the integrability condition in Axiom 7(i), but reduces to PGL when integrability holds.

also helps to elucidate that this conjecture is false: it is the *kernel* k of the Direct Cost — *not* its divergence approximation D — that determines whether the Indirect Cost is UPS. Namely, the following examples[60] show that the inequality (PGL) is strict for natural classes of Direct Cost functions:

Example 1 (Convex Transformation of Bregman Divergence)**.** Let $D_F(q \mid p)$ denote the Bregman divergence generated by $F \in \mathbf{C}^2(\Delta_\circ)$. Let C be the Posterior Separable cost function with divergence $D(q \mid p) := f(D_F(q \mid p))$, where $f : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing, convex, is twice differentiable at 0, and satisfies $f(0) = 0$ and $f''(0) = 1$. It is then immediate that $\mathcal{H}F(p) = \mathcal{H}_q D(q \mid p)\,|_{q=p}$, so that by Lemma 3 C is Locally Quadratic with kernel $k(p) = \frac{1}{2}\mathcal{H}F(q)$. Moreover, the divergences D and D_F satisfy PGL because f is convex, and satisfy PGL with strict inequality when f is strictly convex.

Example 2 (f-divergence)**.** The divergence D_f is called an ***f*-divergence** if there exists a convex function $f : \mathbb{R}_{++} \to \mathbb{R}_+$ with $f(1) = 0$ such that

$$D_f(q \mid p) = \sum_\theta f\left(\frac{p_\theta}{q_\theta}\right) q_\theta.$$

A cost function is called *f-based* if it is Posterior Separable with an f-divergence. By Lemma 3, it is easy to see that an f-based cost function is Locally Quadratic if and only if $f''(0)$ exists. It is well known that all Locally Quadratic f-based costs have the same kernel, namely, the Fisher information matrix (10) (Amari (2016, Lemma 3.5)). As shown above in Lemma 5, the Fisher information matrix is the Hessian of Shannon entropy, so that for any f-based cost PGL reduces to $D_f(q \mid p) \geq D_{KL}(q \mid p)$.

4.5 Sequential Prior-Invariance

In this subsection, we apply the preceding characterization results to revisit the Prior-Invariance Critique of the UPS model discussed in Subsection Section 3.4. Recall that Proposition 1 and Corollary 1.3 establish that essentially no Indirect Cost can be Prior-Invariant. While we conclude from this that Prior-Invariant is *prima facie* too strong of a property to demand of an Indirect Cost function, arguments in favor of Prior-Invariance are still compelling in some contexts when information costs are based on some "physical" device or process. A natural way to resolve this tension is to require that the *Direct* Cost of information be Prior-Invariant, even though the sequential optimization process will generate a prior-dependent Indirect Cost. This desideratum leads to the following class of Indirect Cost functions:

Definition 13 (Sequentially Prior-Invariant)**.** *Cost function C^* is Sequentially Prior-Invariant if there exists a Prior-Invariant direct cost C such that $C^* = \Phi(C)$.*

Note that the general theory developed in Sections 2 and 3 made no restrictions on DM's Direct Cost. Theorem 1 therefore implies that every Sequentially Prior-Invariant cost function is SLP. This raises two natural questions. First, to what extent is the weaker desideratum of *Sequential* Prior-Invariance consistent with the UPS model? Second, what properties define the Sequentially Prior-Invariant cost functions within the SLP class? We address the first question below. We return to the second question, the answer to which requires concepts introduced later, in Subsection A.1.4.

[60] Versions of these examples also appear in Hébert and Woodford (2020b).

A Sequential Prior-Invariance Critique. The main result of this subsection is the complete characterization of the Indirect Cost functions that are both Sequentially Prior-Invariant and UPS. We find that, generically, these two criteria are mutually exclusive. This finding can be interpreted as a *Sequential Prior-Invariance Critique* of the UPS model, which is stronger than the Prior-Invariance Critique common in the literature.

Proposition 3. *Let the Direct Cost function C be Prior-Invariant, Locally Quadratic, and nonzero. The following are equivalent:*

(i) $C^* = \Phi C$ *is UPS and Locally Strongly Convex.*

(ii) The state space $\Theta = \{\theta_1, \theta_2\}$ *is binary and there exists some* $\alpha > 0$ *such that*

$$C(\sigma) \geq \underline{C}(\sigma) := \alpha \max\left\{D_{KL}(\sigma_{\theta_1} \mid \sigma_{\theta_2}), D_{KL}(\sigma_{\theta_2} \mid \sigma_{\theta_1})\right\} \tag{11}$$

for all $\sigma \in \mathcal{E}_b$,

$$\Phi C(\sigma \mid p) \equiv \alpha\left[p_{\theta_1} D_{KL}(\sigma_{\theta_1} \mid \sigma_{\theta_2}) + p_{\theta_2} D_{KL}(\sigma_{\theta_2} \mid \sigma_{\theta_1})\right], \tag{Wald}$$

and the kernels of C and ΦC coincide.

Moreover, when $|\Theta| = 2$, *the Direct Cost* $\underline{C}$ *satisfies the above conditions.*

Proof. See Appendix H. □

The main message of Proposition 3 is that — subject to the technical condition that the Direct Cost be Locally Quadratic — the UPS and Sequentially Prior-Invariant classes of Indirect Costs intersect only when $|\Theta| = 2$. Thus, these classes are disjoint in all but the most stylized economic settings. A secondary message is that the unique such cost function, given in display (Wald), is the Total Information cost function with symmetric coefficients ($\gamma_{\theta_1,\theta_2} = \gamma_{\theta_2,\theta_1}$). The theorem also establishes, as a corollary to Theorem 3, that the only Prior-Invariant Direct Costs that could generate this Indirect Cost are those that majorize it for all prior beliefs (as stated in display (11)). This has immediate implications for our cost functions of primary interest:

Corollary 3.2. *The following hold:*

(i) If C is Prior-Invariant, Locally Quadratic, and Blackwell monotone, and if $|\Theta| \geq 3$, then ΦC is not Regular and exhibits sometimes-strict Preference for One-Shot Learning.

(ii) If C is Prior-Invariant, Locally Quadratic, and Blackwell monotone, and if $|\Theta| \geq 3$, then ΦC is not Total Information.

(ii) If C is Prior-Invariant, Locally Quadratic, and Blackwell monotone, then ΦC is not Mutual Information.

(iv) The Indirect LLR cost function ΦC_{LLR} is not UPS.

Proof. For point (i), ΦC cannot be UPS, because this would contradict Proposition 3. Thus, ΦC cannot be Regular by Theorem 2 and must exhibit sometimes-strict Preference for One-Shot Learning by Theorem 1 and Lemma 1. Points (ii) and (iii) are immediate from point (i) and the facts that Total Information and Mutual Information are UPS. Point (iv) is immediate from point (i) and the fact that the LLR cost function is Locally Quadratic (see Lemma 10). □

The proof of Proposition 3 builds directly on the necessity direction (point (ii)) of Theorem 3, which allows us to characterize the set of (Locally Quadratic) Prior-Invariant Direct Costs generating any Sequentially Prior-Invariant UPS Indirect Cost. For a Prior-Invariant Direct Cost, it can be shown that the second derivative with respect to experiments — the normalized kernel $\overline{k}(p)$ in (9) — must also be independent of the prior and, therefore, a constant matrix. When $|\Theta| > 3$, we show that this is inconsistent with $\overline{k}(p)$ being the Hessian of some potential function, and so violates the integrability condition Axiom 7(i). Notably, by Lemma 4, this implies that when $|\Theta| > 3$ the intersection between the Sequentially Prior-Invariant and UPS classes is empty *even if* DM is restricted to acquire only Gaussian diffusion signals.

By contrast, when $|\Theta| = 2$, the integrability condition Axiom 7(i) has no bite: *any* kernel satisfying the Taylor expansion (LQ) can be written as the Hessian of a convex potential function. The symmetric Total Information cost function in (Wald) arises from solving for the unique potential function that has a prior-independent Hessian, and the majorization condition (11) follows from (PIL).

4.6 Relation to Morris and Strack (2019) and Hébert and Woodford (2020b)

As noted above, our Theorem 3 and Proposition 3 build on — and maximally generalize — the results of two related papers, Morris and Strack (2019) and Hébert and Woodford (2020b). Here, we describe the relation between these results in more detail.

Relation to Morris and Strack (2019). Morris and Strack (2019) study a continuous-time variant of the Wald (1945) sequential sampling model, in which DM acquires information by choosing how long to observe an *exogenous* Gaussian diffusion signal process and pays a flow cost $c(q_t)dt$ at each instant before stopping. The key difference from our model is that their DM chooses only *when to stop* acquiring information, but *not* what information to acquire before stopping.[61]

The main result of that paper establishes that, when $|\Theta| = 2$: (i) every (bounded) experiment can be replicated by some stopping strategy and (ii) the expected cost of such replication is UPS with potential function $F \in \mathbf{C}^2(\Delta_\circ)$ whose second derivative is proportional to the flow cost $c(q)$. Aside from inessential notational differences, this finding is precisely the $|\Theta| = 2$ special case of our Lemma 4. When $|\Theta| \geq 3$, only a non-generic class of (bounded) experiments can be replicated by the sampling process considered in Morris and Strack (2019). They show that the expected cost of replication is UPS on that restricted domain of implementable experiments. By allowing DM to flexibly control the volatility of her posterior beliefs, our Lemma 4 extends that characterization to the full class of bounded experiments. More substantively, our Theorem 3 establishes that the connection between Gaussian learning and UPS is general, even when DM can undertake arbitrary Sequential Replication strategies.

Morris and Strack (2019) place particular emphasis on the special case of their model in which DM's flow cost of sampling is constant (i.e., is Prior-Invariant in our language). When $|\Theta| = 2$, they characterize the expected cost of sampling as the symmetric Total Information cost function in (Wald), which they refer to as the *Wald* cost function due to its connection to the Wald (1945)

[61] The flow cost $c(q)$ in Morris and Strack's (2019) model can be written in our notation as $c(q) = \frac{1}{2}\mathrm{tr}\left[v^\top(q)\mathcal{H}F(q)v(q)\right]$, as in (GIC), where $v(q)$ is the volatility of their DM's belief at posterior q, which is determined by parameters of her signal process and standard filtering formulae.

model. When $|\Theta| \geq 3$, that papers shows that: (i) only a non-generic subset of (bounded) experiments can be replicated by some stopping strategy in their model and (ii) the expected cost of any feasible experiment is a Total Information cost function with coefficients

$$\gamma_{\theta,\theta'} = \frac{1}{(|\Theta|-1)(m_\theta - m_{\theta'})^2}, \tag{12}$$

where $\boldsymbol{m} \in \mathbb{R}^\Theta$ is the vector of (fixed) state-dependent drifts of the diffusion signal process from which their DM samples.[62]

Our Proposition 3 maximally generalizes Morris and Strack's (2019) findings for Prior-Invariant flow costs in two respects. First, it shows that, when $|\Theta| = 2$, the Wald cost function is still Sequentially Prior-Invariant and, moreover, is the unique Sequentially Prior-Invariant and UPS cost function — even when DM's strategy space is fully flexible. Second, it shows that no Sequentially Prior-Invariant and UPS cost function exists when $|\Theta| \geq 3$, implying that Morris and Strack (2019) representation for this case cannot be extended to the full class of bounded experiments. In this sense, our Proposition 3 shows that their restriction to $|\Theta| = 2$ and Gaussian learning is actually without loss of generality. Importantly, none of these findings can be deduced from results in Morris and Strack (2019).

Relation to Hébert and Woodford (2020b). Hébert and Woodford (2020b) study a model of sequential sampling in which DM's strategy space is more flexible than in Morris and Strack (2019) but less flexible than in our model. They allow DM to acquire any type of experiment in each discrete time period, but impose an intertemporal constraint on her strategy space that bounds the average per-period cost that she can expend;[63] in the continuous-time limit on which they focus, DM is effectively constrained to optimizing over jump-diffusion belief processes under which she expends the same infinitesimal Direct Cost at each instant (i.e., perfectly smooths costs over time). To characterize their DM's optimal choice among Poisson and diffusion signals in continuous time, Hébert and Woodford (2020b) assume that their DM's Direct Cost satisfies slightly stronger versions of *both* our Locally Linear and Locally Quadratic conditions (the case discussed in Subsection 4.4.3).

Hébert and Woodford (2020b) introduce the "preference for gradual learning" condition PGL as a sufficient condition under which their DM finds Gaussian learning cheaper than Poisson learning. Our Lemma 6 slightly strengthens this observation by showing that PGL is (i) sufficient to render Gaussian learning optimal among *all* strategies, and (ii) also *necessary* for the optimality of Gaussian learning.

Hébert and Woodford (2020b, Theorem 8) establish a special case of our sufficiency result, Theorem 3(i). They show that when their DM's Direct Cost satisfies PGL, her optimal state-contingent choice probabilities in any decision problem with fewer actions than states[64] will be the same as would arise under one-shot information acquisition given the UPS Indirect Cost described in our

[62] Following Pomatto et al. (2019, Proposition 3), we could alternatively derive the coefficients (12) by requiring that the Total Information cost of standard Gaussian experiments is (i) independent of the prior belief and (ii) independent of the number of states.

[63] While we use sequential cost-minimization as a way to formalize DM's strategic flexibility, Hébert and Woodford (2020b) are primarily motivated by findings from neuroscience and mathematical psychology that human perception and attention allocation are gradual, dynamic processes. Consequently, much of Hébert and Woodford (2020b) is concerned with providing detailed characterizations of optimal learning dynamics in decision problems with delay costs and possibly time discounting, results that have no parallel in our work. For ease of comparison, we focus here on the special case of their model without discounting.

[64] Recall that a decision problem (A, u) consists of an action set A and state-dependent Bernoulli utility function $u : A \times \Theta \to \mathbb{R}$. Hébert

Theorem 3(i). Given Lemma 6, their result can be (almost) equivalently restated in our language as follows: for Direct Cost functions that are both Locally Linear and Locally Quadratic, PGL is sufficient to guarantee that the Indirect Cost *of experiments with* $m \leq |\Theta|$ *signal realizations* is UPS.[65] However, this is *not* sufficient to pin down the entire Indirect Cost function.[66] Our Theorem 3(i) therefore strengthens their result by showing that PGL remains sufficient even when DM's strategy space is not restricted, dispensing with superfluous technical conditions on the Direct Cost function, and characterizing the Indirect Cost of *all* experiments.

Importantly, Hébert and Woodford (2020b) do not present any analogue to our necessity result, Theorem 3(ii). Without this necessity result, it is not possible to (i) characterize the full class of Direct Costs that generate UPS Indirect Cost functions, (ii) conclude that the restriction to Gaussian learning is without loss of generality for characterizing such Indirect Cost functions, or (iii) study the intersection of the Sequentially Prior-Invariant and UPS classes.

5 Foundations for Total Information

This section has two main purposes. First, we characterize Total Information as the unique SLP cost function exhibiting (a slight generalization of) the Constant Marginal Cost axiom recently introduced by Pomatto et al. (2019) to characterize the LLR cost function. Second, we argue that the meaning of Constant Marginal Cost is fundamentally different for Direct and Indirect Cost functions, and that sequential optimization tends to generate Indirect Cost functions with Decreasing Marginal Cost.

5.1 Simultaneous Replication and Marginal Cost

Many strategies for acquiring information used in practice involve two features. First, information is acquired in the form of multiple conditionally independent copies of a fixed experiment, with the precision of information determined by the number of copies (i.e., "sample size"). Second, these experiments are run "simultaneously," with the signals from each experiment observed only after all experiments have been run. These features are common, for instance, in political polling, market research, and A/B testing (e.g., Azevedo et al. (2019)), as well as within firms when information-gathering activities are decentralized among multiple employees. The following definition formalizes these kinds of information acquisition strategies, generalized to allow for conditionally independent but not necessarily identical experiments (recall the definition of the product experiment $\sigma_1 \otimes \sigma_2$ from (6)):

---

and Woodford (2020b) assume that $|A| \leq |\Theta|$ and rely on this assumption when constructing a Markovian (in the posterior belief) sequential replication for DM's optimal target experiment.

[65] Formally, there exists a convex function F such that $C(\sigma \mid p) = \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[F(q) - F(p)]$ for all $\sigma \in \mathcal{E}_b$ and $p \in \Delta_\circ$ such that $|\text{supp}(\pi_{\langle \sigma | p \rangle})| \leq |\Theta|$. Given any UPS cost function and decision problem (A, u), there always exists an optimal information acquisition strategy with no more than $\min\{|A|, |\Theta|\}$ signal realizations; if the cost function is strictly Blackwell monotone, then every optimal strategy has no more than $|A|$ signal realizations. Thus, the assumption that $|A| \leq |\Theta|$ nearly implies that DM will only choose experiments with $m \leq |\Theta|$ signals, and this implication is exact under strict Blackwell monotonicity. That *every* experiment with $m \leq |\Theta|$ signal realizations is optimal for some decision problem (A, u) with $|A| \leq |\Theta|$ actions follows from a simple extension of the duality arguments used to prove de Oliveira et al. (2017, Theorem 2).

[66] This can be shown to follow from de Oliveira et al. (2017, Theorem 2) or Denti (2020, Proposition 1).

Definition 14 (Simultaneous Replication). *The set of experiments* $\{\sigma_i\}_{i=1}^n$ *constitute a* ***Simultaneous Replication*** *of the target experiment* σ *if* $\sigma_1 \otimes \sigma_2 \otimes \cdots \otimes \sigma_n \succsim_B \sigma$.

The defining feature of Simultaneous Replication is that all "sub-experiments" σ_i must be conditionally independent. This captures the idea that, since each of these experiments is run "at the same time," DM cannot condition her strategy on the signals that they generate. A leading example of Simultaneous Replication is the acquisition of conditionally independent Gaussian experiments, for which the posterior precision varies linearly with the sample size n.

Formally, Simultaneous Replication is a special case of Sequential Replication, because in the latter DM can always choose to run a deterministic sequence of experiments. The distinction arises from the way that costs are assigned: because DM observes all signals at once, in a Simultaneous Replication must evaluate the cost of each σ_i under the same prior belief. In particular, we define DM's problem of choosing the optimal Simultaneous Replication given Direct Cost C as

$$\Phi_{sim}C(\sigma \mid p) := \inf_{\sigma_1 \otimes \cdots \otimes \sigma_n \succsim_B \sigma} \sum_{i=1}^{n} C(\sigma_i \mid p), \tag{SIC}$$

and refer to $\Phi_{sim}C$ as the *Simultaneous Indirect Cost* function. It is worth noting that there exist cost functions C for which $\Phi_{sim}C$ is not pointwise larger than ΦC (cf. Example 5 below).

The idea of Simultaneous Replication leads naturally to the following notions of the "marginal cost" of information.

Axiom 8 (Decreasing Marginal Cost). *Cost function* C *exhibits* ***Decreasing Marginal Cost*** *if*

$$C(\sigma \otimes \tau \mid p) \leq C(\sigma \mid p) + C(\tau \mid p) \tag{DMC}$$

for all $\sigma, \tau \in \mathcal{E}_b$ *and* $p \in \Delta_\circ$.

A cost function exhibits Decreasing Marginal Cost if it is cheaper to acquire information together than via any two-experiment Simultaneous Replication without disposal. It is easy to see that Decreasing Marginal Cost is the "simultaneous learning" analogue to Preference for One-Shot Learning: $\Phi_{sim}C(\sigma \mid p) = C(\sigma \mid p)$ for all $\sigma \in \mathcal{E}_b$ and $p \in \Delta_\circ$ if and only if C is Blackwell monotone and exhibits Decreasing Marginal Cost.

Axiom 9 (Constant Marginal Cost). *Cost function* C *exhibits* ***Constant Marginal Cost*** *if*

$$C(\sigma \otimes \tau \mid p) = C(\sigma \mid p) + C(\tau \mid p) \tag{CMC}$$

for all $\sigma, \tau \in \mathcal{E}_b$ *and* $p \in \Delta_\circ$.

By induction, a cost function exhibits Constant Marginal Cost if and only if the cost of all Simultaneous Replications (without disposal) of a given experiment are equally costly. In this sense, it is the "simultaneous learning" analogue to Indifference to Sequential Learning. From a different perspective, Constant Marginal Cost is the natural generalization of the assumption, common in statistical decision theory, that the cost of information is linear in sample size.[67] Clearly, any non-zero cost function exhibiting Constant Marginal Cost must be unbounded.

[67] This is the standard assumption, for instance, in Wald (1945), Arrow et al. (1949), and various models of Gaussian sampling in which costs are linear in the precision of Gaussian experiments (see Veldkamp (2011) for discussion).

Pomatto et al. (2019) introduce the special case of Axiom 9 for Prior-Invariant cost functions, based on an analogy to "constant returns to scale" in producer theory, and characterize the LLR cost functions as the unique Prior-Invariant and Dilution Linear cost functions exhibiting Constant Marginal Cost.

The Returns-to-Scale Critique. A leading critique of the Mutual Information cost function, which we call the *Returns-to-Scale Critique*, is based on the following observation:

Lemma 7. *Mutual Information exhibits sometimes-strictly Decreasing Marginal Cost.*

Proof. See Appendix K. □

It is easy to see that Mutual Information must exhibit Decreasing Marginal Cost "asymptotically" in the following sense. Consider the n-fold product of σ, denoted $\sigma^{(n)}$, which generates n conditionally independent draws from σ. Since Mutual Information is Bounded, the marginal cost of an additional draw $C_{MI}(\sigma^{(n)} \mid p) - C_{MI}(\sigma^{(n-1)} \mid p) \to 0$ as $n \to \infty$. Various authors have suggested that this property is inappropriate in many contexts and leads to counterintuitive predictions. For instance, Pomatto et al. (2019, p. 18) write:

> "Under LLR cost, the additivity axiom implies that the cost of observing k coin flips is linear in k. Hence the cost of observing a sequence of k flips goes to infinity with k. Under Mutual Information cost ... the cost of observing infinitely many coin flips is only approximately 3.6 times the cost of observing a single coin flip. The low — and arguably in many applications unrealistic — cost of acquiring perfect information is caused by the sub-additivity of Mutual Information as a cost-function ... These simple calculations suggest that using Sims' Mutual Information cost as a model of information production rather than information processing (as originally intended by Sims) may lead to counterintuitive predictions."

This difference between Decreasing Marginal Cost and Constant Marginal Cost can have important implications in economic applications. For instance, in a portfolio choice setting with Gaussian uncertainty, Nieuwerburgh and Veldkamp (2010) show that Mutual Information leads to "corner solutions" in which the investor learns about only one asset and consequently chooses an underdiversified portfolio, while with Constant Marginal Cost cost functions she learns about all assets and chooses a diversified portfolio (see also Morris and Strack (2019) for a related example). In strategic settings, it is known that Mutual Information cost functions can lead to multiple equilibria, even in settings where Constant Marginal Cost learning technologies induce unique equilibria (Myatt and Wallace (2012)).

Constant Marginal Cost and Sequential Replication. The Returns-to-Scale Critique is based on a particular vision of information acquisition that, evidently, does not account for the possibility of sequential optimization. For instance, the classical information-theoretic foundation for Mutual Information views this cost function as the expected cost of an optimal iterative search procedure in which DM sequentially asks binary yes/no questions about the state, and each such question has

equal cost. The optimal search procedure takes the form of a generalized "bisection" algorithm, which inevitably induces Decreasing Marginal Cost (cf. Veldkamp (2011)).

How general is the above logic? A main lesson of this section is that it is quite general: Constant Marginal Cost is generally not preserved under sequential optimization and, moreover, sufficiently rich strategy spaces for information acquisition inevitably lead to Indirect Cost functions with Decreasing Marginal Cost (see Subsection A.1 below). The following lemma illustrates an important special case of this lesson:

Lemma 8. *Every non-zero Indirect LLR cost function ΦC_{LLR} exhibits sometimes-strictly Decreasing Marginal Cost.*

Proof. See Appendix K. □

Lemma 8 illustrates that Constant Marginal Cost is generally not preserved under sequential optimization. Consequently, arguments in favor of or against Constant Marginal Cost that are relevant in contexts of one-shot information acquisition (i.e., for DM's Direct Cost function) may not be relevant when information can be acquired sequentially (i.e., for DM's Indirect Cost function).

The Case for Decreasing Marginal Cost. In Appendix A.1, we develop an extension of our framework that allows DM to engage in a general form of *Unrestricted Replication* that encompasses both our main notion of Sequential Replication and the above notion of Simultaneous Replication. We show that the resulting *Unrestricted Indirect Cost* necessarily exhibits Decreasing Marginal Cost, suggesting that this property is a necessary condition of optimality.

5.2 Total Information: Characterization

The following characterization of Total Information is the main result of this section. For technical reasons, we restrict attention to *Extensible* cost functions that can, in a suitable sense, be continuously extended to a class of experiments $\mathcal{E}_\circ$ that is strictly larger than $\mathcal{E}_b$, but strictly smaller than $\mathcal{E}$ (see Appendix I for formal details). We note that Mutual Information, Total Information, Mutual Information, and all other commonly-used cost functions that we know of are Extensible.

Theorem 4. *For an Extensible cost function C^*, the following are equivalent:*

(i) C^* is SLP and exhibits Constant Marginal Cost.

(ii) C^* is a Total Information cost function.

Moreover, if C is an Extensible and Dilution Linear Direct Cost function exhibiting Constant Marginal Cost, then ΦC is a Total Information cost function if and only if $C = \Phi C$.[68]

Proof. See Appendix I. □

Theorem 4 has two main takeaways. First, it characterizes Total Information as the unique SLP cost function exhibiting Constant Marginal Cost. Because Total Information is UPS, a notable — and perhaps surprising — implication of this result is that, within the SLP class, Constant Marginal

[68] Equation (55) in Appendix I notes that this result can, in fact, be generalized to a somewhat larger class of Direct Cost functions.

Cost automatically implies Indifference to Sequential Learning. This suggests that Total Information is uniquely "process invariant," in the sense that the expected cost of acquiring a given target experiment is the same no matter whether it is acquired in one shot, via Sequential Replication, or via Simultaneous Replication. Indeed, Corollary 6.2 in Appendix A.1shows that Total Information uniquely satisfies an even stronger *Process Invariance* condition. This explains why Total Information is named as such: "merging" or "splitting" experiments either sequentially or simultaneously does not affect costs. Put differently, costs depend only the totality of information produced, not the process by which it is acquired. No other cost function satisfies this property.

Second, the theorem also states that, within the Dilution Linear class, no Direct Cost function exhibiting Constant Marginal Cost — aside from Total Information itself — generates an Indirect Cost that also exhibits Constant Marginal Cost. Because DM's Direct Cost function is necessarily Dilution Linear if she has access to mixed and direct Poisson strategies before engaging in any more complicated form of Sequential Replication, this covers a broad class of Direct Cost functions. Thus, Theorem 4 significantly strengthens the messages of Lemma 8, namely, that Constant Marginal Cost is generally *not* preserved under sequential optimization. We therefore conclude that the meaning of Constant Marginal Cost is fundamentally different when it is imposed on the Indirect Cost, rather than the Direct Cost, of information.

Relation to Pomatto et al. (2019). The proof of Theorem 4 builds on Pomatto et al.'s (2019) characterization of the LLR cost function. In particular, by applying that paper's characterization prior-by-prior, under the hypotheses of the theorem we are able to restrict attention to cost functions resembling (LLR), but in which the discrimination coefficients $\beta_{\theta,\theta'}$ are replaced by prior-dependent functions $\hat{\beta}_{\theta,\theta'} : \Delta_\circ \to \mathbb{R}_+$. The combination of Preference for One-Shot Learning and Constant Marginal Cost implies that these prior-dependent coefficients must satisfy a series of linear inequalities, given which a linear separation argument implies that the coefficients must take the linear form $\hat{\beta}_{\theta,\theta'}(p) = p_\theta \gamma_{\theta,\theta'}$ characteristic of Total Information. However, in light of Lemma 8 and the final clause of Theorem 4, we conclude that the relation between Total Information and the LLR cost function is purely formal.

5.2.1 Special Cases of Total Information

Two recently proposed alternatives to the Mutual Information cost function — the ex ante Wald cost of Morris and Strack (2019) and the Fisher information cost of Hébert and Woodford (2020a) — are, in fact, special cases of the Total Information cost function. Our characterization of Total Information unifies and provides new foundations for these special cases. Conversely, economic applications considered in those and subsequent papers (e.g., Hébert and La'O (2020)) illustrate that Total Information is amenable to use in applied economic settings.

Wald Cost Function. In Subsection 4.5, we observed that the (essentially) unique Sequentially Prior-Invariant and UPS cost function, which exists only when $|\Theta| = 2$, is Total Information with symmetric coefficients. As noted in Subsection 4.6, this special case of what Morris and Strack (2019) call the Wald cost function. When $|\Theta| \geq 3$, Morris and Strack (2019) derive a variant of the Total Information with the coefficients in (12) from a Prior-Invariant Direct Cost when DM is constrained

to acquiring information in the form of exogenously-given Gaussian signals. Importantly, these restrictions on DM's strategy space imply that the cost function derived in that paper is only defined over a non-generic subset of bounded experiments. Thus, our Total Information cost function with the coefficients in (12) is the natural extension of Morris and Strack's (2019) cost function to the full domain of bounded experiments.

Fisher Information Cost Function. In recent work, Hébert and Woodford (2020a) axiomatize a particular class of UPS cost functions called the *Neighborhood-Based Costs* (NBCs).[69] Relative to our setup, Hébert and Woodford (2020a) also take as primitive a finite collection $\{N_\ell\}$ of non-emtpy *neighborhoods* $N_i \subseteq \Theta$, with the idea that it is costly to distinguish between states within each neighborhood. A UPS cost function C with potential function F is an NBC if there exists a corresponding collection $\{F_\ell\}$ of neighborhood-specific (bounded and twice continuously-differentiable) potential functions $F_\ell : \Delta(N_\ell) \to \mathbb{R}$ such that

$$F(q) = \sum_\ell \left(\sum_{\theta \in N_\ell} q_\theta \right) F_\ell \left(q(\cdot \mid N_\ell) \right) \tag{NBC}$$

where $q(\cdot \mid N_\ell) \in \Delta(N_\ell)$ denotes the conditional posterior on the neighborhood N_ℓ induced by $q \in \Delta(\Theta)$. Hébert and Woodford (2020a, Proposition 1) show that a UPS cost function (with potential $F \in \mathbf{C}^2(\Delta_\circ)$) is an NBC if and only if it satisfies two axioms (their Assumptions 2-3) capturing the idea that it is costly to distinguish between states $\theta, \theta' \in N_\ell$ within a common neighborhood, but not costly to distinguish between states that do not share any such common neighborhood.

This axiomatization of NBCs has no obvious connection to our characterization of Total Information. Yet, technical qualifiers aside, it is easy to see that every Total Information cost function is an NBC and, conversely, that an NBC is a Total Information cost function if and only if it exhibits Constant Marginal Cost within each neighborhood.[70] In this sense, our definition and derivation of Total Information bridges the gap between the distinct approaches of Hébert and Woodford (2020a) — who restrict to UPS costs and emphasize neighborhood structures — and Pomatto et al. (2019) — who emphasize Constant Marginal Cost and whose LLR cost functions are not UPS.

Hébert and Woodford (2020a) place particular emphasis on a specific limit of NBCs called the ***Fisher Information*** cost function. To derive this function, they assume that (i) the state space $\Theta \subset \mathbb{R}$ is linearly ordered with $\theta_1 < \theta_2 < \cdots < \theta_{|\Theta|}$, (ii) the collection of neighborhoods consists of pairs $\{\theta_i, \theta_{i+1}\}$ of "adjacent" states, (iii) each neighborhood-specific potential function is identical up to a neighborhood-specific scaling factor, and (iv) the number of states becomes unbounded (i.e., $|\Theta| \to \infty$). Under suitable technical conditions, they show that this limiting procedure yields the Fisher Information cost function over the continuous state space $\left(\underline{\theta}, \overline{\theta}\right) \subseteq \mathbb{R}$, defined as

$$C_{Fisher}(\hat{\sigma} \mid p) := \int_{\underline{\theta}}^{\overline{\theta}} \sum_{s \in S} \frac{\left(\frac{\partial}{\partial \theta} \hat{\sigma}(s \mid \theta) \right)^2}{\hat{\sigma}(s \mid \theta)} \hat{p}(\theta) d\theta, \tag{Fisher Information}$$

[69] The Neighborhood-Based Costs were first introduced, without axiomatic foundation, in the earlier working paper Hébert and Woodford (2017). The axiomatization discussed here, which first appeared in Hébert and Woodford (2020a), was obtained contemporaneously to and independently of the results in this paper.

[70] To see that Total Information is an NBC, let each pair of distinct states $\{\theta, \theta'\}$ define a neighborhood with neighborhood-specific potential function $F_{\{\theta,\theta'\}}(q) := q(\theta \mid \{\theta, \theta'\}) \log\left(\frac{q(\theta \mid \{\theta,\theta'\})}{q(\theta' \mid \{\theta,\theta'\})} \right) + q(\theta' \mid \{\theta, \theta'\}) \log\left(\frac{q(\theta' \mid \{\theta,\theta'\})}{q(\theta \mid \{\theta,\theta'\})} \right)$.

where $\hat{p} : \left(\underline{\theta}, \overline{\theta}\right) \to \mathbb{R}_+$ is the density (with respect to Lebesgue measure) of DM's absolutely continuous prior belief p, and the domain of C_{Fisher} is restricted to $\mathcal{E}_{Lip}$, the set of finite-support experiments $\hat{\sigma} : \left(\underline{\theta}, \overline{\theta}\right) \to \Delta(S)$ such that for each signal s the map $\theta \mapsto \hat{\sigma}(s \mid \theta)$ has Lipschitz continuous derivative. This cost function gets its name from the fact that it is the expected value of

$$\mathcal{I}(\hat{\sigma} \mid \theta) := \sum_s \left(\frac{\partial}{\partial \theta} \log(\hat{\sigma}(s \mid \theta)) \right)^2 \hat{\sigma}(s \mid \theta), \tag{13}$$

the Fisher information of the signal at state θ. Notably, Hébert and Woodford (2020a) and Hébert and La'O (2020) show that the Fisher Information cost function (suitably extended to experiments without finite support) is highly tractable in canonical Gaussian-Quadratic environments, where it leads to significantly different predictions than Mutual Information.

Importantly, Total Information is the natural discrete-state analogue to, and generalization of, the Fisher Information cost function. There are two ways to see this. First, observe that a near-defining feature of the Fisher Information cost function is that it is UPS and exhibits Constant Marginal Cost, the latter of which follows from the well-known additivity of the Fisher information $\mathcal{I}(\cdot \mid \theta)$ with respect to conditionally independent distributions.[71] While this property is not noted by Hébert and Woodford (2020a), it plays an important role in their characterization of optimal strategies in Gaussian-Quadratic environments. It is also curious that *none* of the finite-state NBCs that Hébert and Woodford (2020a) use to approximate the Fisher Information cost in the continuous-state limit exhibit Constant Marginal Cost.[72] In other words, that paper obtains Constant Marginal Cost of the Fisher Information cost function *only* in the continuous-state limit.

Second, we may heuristically derive the Fisher Information cost function as a particular limit of Total Information cost functions as follows. Suppose that the state space has the linear structure assumed by Hébert and Woodford (2020a), whereby $\Theta = \left\{\theta_1, \ldots, \theta_{|\Theta|}\right\} \subset (\underline{\theta}, \overline{\theta})$. Also assume that the discrimination coefficients in (TI) satisfy $\gamma_{\theta_i, \theta_j} = \mathbf{1}(j = i+1)/(\theta_i - \theta_{i+1})^2$. For any $\hat{\sigma} \in \mathcal{E}_{Lip}$ mapping $\hat{\sigma} : (\underline{\theta}, \overline{\theta}) \to \Delta(S)$, define its restriction to Θ by $\sigma : \Theta \to \Delta(S)$. Finally, recall the well-known characterization of Fisher information as the Hessian of the KL divergence, whereby $D_{KL}(\sigma_\theta \mid \sigma_{\theta'}) = \frac{(\theta - \theta')^2}{2} \mathcal{I}(\sigma \mid \theta) + o\left((\theta - \theta')^2\right)$ for any $\sigma \in \mathcal{E}_{Lip}$ and $\theta, \theta' \in (\underline{\theta}, \overline{\theta})$. Then in the continuous-state limit where $|\Theta| \to \infty$ and $|\theta_i - \theta_{i+1}| \to 0$, we have the approximation

$$\begin{aligned} C_{TI}(\sigma \mid p) &= \sum_{i=1}^{|\Theta|-1} \left[\frac{p_{\theta_i}}{(\theta_i - \theta_{i+1})^2} D_{KL}(\sigma_{\theta_i} \mid \sigma_{\theta_{i+1}}) + \frac{p_{\theta_{i+1}}}{(\theta_i - \theta_{i+1})^2} D_{KL}(\sigma_{\theta_{i+1}} \mid \sigma_{\theta_i}) \right] \\ &\approx \sum_{i=1}^{|\Theta|-1} \left(\frac{p_{\theta_i} + p_{\theta_{i+1}}}{2} \right) \cdot \mathcal{I}(\sigma \mid \theta_i) \\ &\to C_{Fisher}(\hat{\sigma} \mid p), \end{aligned}$$

assuming that the sequence of priors $p \in \Delta(\Theta)$ converges a probability measure with density $\hat{p}$ on the limiting state space $(\underline{\theta}, \overline{\theta})$. In our view, his approximation (suitably formalized) provides an appealing alternative to Hébert and Woodford's (2020a) derivation of the Fisher Information cost

[71] Formally, $\mathcal{I}(\sigma \otimes \tau \mid \theta) = \mathcal{I}(\sigma \mid \theta) + \mathcal{I}(\tau \mid \theta)$ for all $\sigma, \tau \in \mathcal{E}_{Lip}$, which implies that $C_{Fisher}(\sigma \otimes \tau \mid p) = C_{Fisher}(\sigma \mid p) + C_{Fisher}(\tau \mid p)$ as well.

[72] Hébert and Woodford (2020a) primarily focus on the case in which the approximating NBCs are proportional to Mutual Information within each neighborhood.

function because it (a) highlights the well-known fact that Fisher information arises as the Hessian of KL divergence, (b) makes more transparent the fact that Fisher Information costs exhibit Constant Marginal Cost, and (c) relates Fisher Information costs to our characterization of Total Information in Theorem 4. We believe that this connection to Total Information provides a compelling reason — independent of those discussed in Hébert and Woodford (2020a) — for focusing on the Fisher Information cost in continuous-state models. Of course, Total Information is well-defined also for finite-state settings and allows for a more general structure of the discrimination coefficients than used in this approximation, and so is not wedded to the interpretation that only adjacent states are costly to distinguish.

6 Foundations for Mutual Information

This section characterizes Mutual Information as the unique (subject to regularity conditions) SLP cost function consistent with the idea that DM is able to "freely ignore" aspects of the state space that she finds "irrelevant."[73]

The Perceptual Distance Critique. A common criticism of the Mutual Information cost function, which we call the *Perceptual Distance Critique*, is that this cost function treats all states symmetrically. For instance, Maćkowiak et al. (2018, p. 10) write that:

> "[E]ntropy does not depend on a metric, i.e., the distance between states does not matter. With entropy, it is as difficult to distinguish the temperature of 10°C from 20°C, as 1°C from 2°C. In each case the agent needs to ask one binary question, resolve the uncertainty of one bit. If, however, the agent needed to use a thermometer with inherent additive noise of a given size, then it is clear that distinguishing the more distant states 10°C and 20°C would be easier — reduction of entropy is not a good measure of information in that case."

Many authors have argued that this property is inappropriate for cost functions representing the cost of information production, for which the "distance" between states naturally affects costs (cf. examples in Pomatto et al. (2019)). This symmetry property is also known to have implications for equilibrium selection in coordination games with endogenous information choice (Morris and Yang (2019)), and is sometimes rejected in experimental tests of human attention allocation (Dean and Neligh (2019); Dewan and Neligh (2018)).

However, when it is costly for DM to "process" already-available information, it is natural to expect that she would learn how to "optimally encode" states before processing information about them. This idea underlies the classical information-theoretic foundations for Mutual Information (Cover and Thomas (2006, Ch. 10)). Consider, for instance, the problem of a DM who does not learn directly about the state, but rather from news outlets that generate information on her behalf. DM has limited attention and finds it costly to process the information presented by a news outlet. Which outlet will she choose to learn from? In general, if her cost function does not satisfy the same symmetry properties of Mutual Information, DM will find it more or less costly to learn from

[73] The material in this section is preliminary and subject to updates.

different outlets that present news in different "languages," even if the information that they convey about the payoff-relevant state is identical. Consequently, DM may choose to learn from an outlet that provides Blackwell *inferior* information about the state of interest, simply because she finds its language easier to process. As formalized and discussed further in Bloedel and Segal (2020), this kind of behavior appears to be inconsistent with standard notions of Bayesian rationality. The Mutual Information cost function does not suffer from this problem; below, we show that it the essentially unique SLP cost function with this property.

6.1 Axioms

Our characterization of Mutual Information is based on two key axioms, which formalize the idea that information costs are reduced under certain forms of compression of the state space. (In Appendix K, we interpret these axioms in terms of an extended optimization problem for DM.)

Definition 15. *A **compression** of Θ is a mapping $\kappa : \Theta \to 2^{\Theta} \setminus \emptyset$ such that:*

(i) The collection $\{\kappa(\theta)\}_{\theta \in \Theta}$ defines a partition of Θ.
(ii) $\theta \in \kappa(\theta)$ for all $\theta \in \Theta$.

Let $\mathcal{K}$ denote the set of all coarsenings of Θ.

The first axiom captures the idea that compressing states that DM's target experiment treats identically should not affect the cost of that experiment.

Axiom 10 (Weakly Compression Invariant)**.** *Cost function C is Weakly Compression Invariant if*

$$C(\sigma \mid p) = C(\sigma \mid p') \tag{14}$$

for all experiments $\sigma \in \mathcal{E}_b$ measurable with respect to compression $\kappa \in \mathcal{K}$,[74] *and priors $p' \in \Delta_\circ$ for which $p'(\kappa(\theta)) = p(\kappa(\theta))$ for all $\theta \in \Theta$.*

Another way to view this axiom is that Weakly Compression Invariant cost functions exhibit a restricted form of Prior-Invariance, whereby the cost of an experiment σ is not affected by shifting probability mass *within σ-measurable events*, but may be affected by shifting probability mass across these events. In other words, Axiom 10 states that DM's prior affects her cost of information only to the extent that it determines the probabilities of the events that she learns about.

For a simple example, suppose DM aims to learn about a political candidate's platform, which may be *l*eft-, *c*enter-, or *r*ight-leaning. Axiom 10 demands that if DM's target experiment is informative *only* about whether the candidate is right-leaning — e.g., it determines whether the true state is in $\{l, c\}$ or $\{r\}$ — then her cost remains the same when prior probability mass is shifted *within* the event $\{l, c\}$. For instance, the red and blue posterior distributions in the left-hand panel of Figure 3 have equal cost under a Weakly Compression Invariant cost function. Intuitively, "splitting" or "merging" the states l and c should not affect DM's cost when her (fixed) target experiment already ignores any distinction between them.

The second axiom captures the idea that DM finds it cheaper to learn about coarser events.

[74] Formally, $\sigma(\cdot \mid \theta) = \sigma(\cdot \mid \theta')$ whenever $\kappa(\theta) = \kappa(\theta')$.

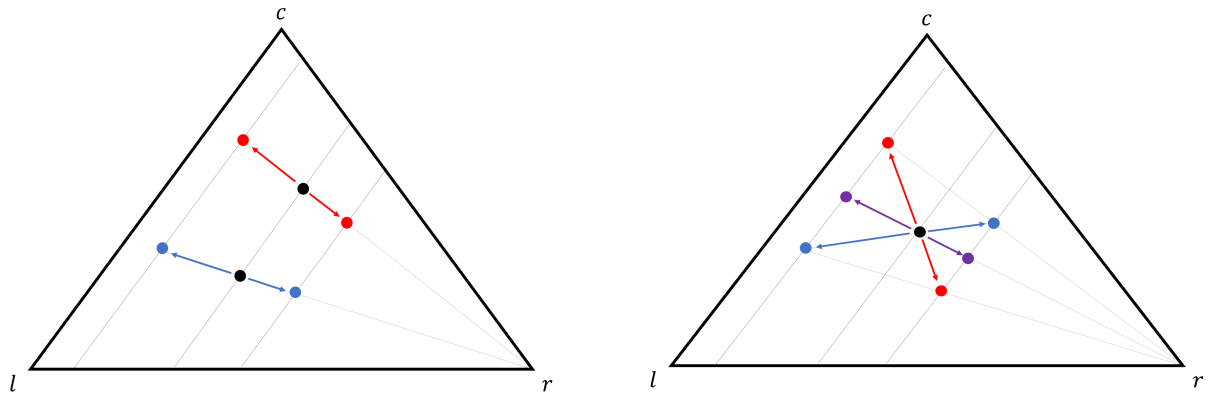

Figure 3: Illustrations of Axiom 10 (left) and Axiom 11 (right).

Definition 16. *Given coarsening κ and prior $p \in \Delta_\circ$, the $\langle \kappa \mid p \rangle$-**compression** of experiment $\sigma \in \mathcal{E}_b$, denoted by $\sigma_{\langle \kappa|p \rangle}$, is the bounded experiment defined by*

$$\sigma_{\langle \kappa|p \rangle}(s \mid \theta) := \frac{\sum_{\theta' \in \kappa(\theta)} \sigma(s \mid \theta') p_{\theta'}}{\sum_{\theta' \in \kappa(\theta)} p_{\theta'}} \tag{15}$$

For each state θ in a given cell of the compression κ, the compressed experiment $\sigma_{\langle \kappa|p \rangle}$ replaces the conditional distribution of signals $\sigma(\cdot\theta)$ with the average distribution of signals given the prior, conditional on the state being in that cell of κ. Intuitively, this operation captures the idea that DM "forgets" the distinction between all states within a given cell.

Axiom 11 (Compression Monotone)**.** *Cost function C is Compression Monotone if*

$$C(\sigma \mid p) \geq C(\sigma_{\langle \kappa|p \rangle} \mid p) \tag{16}$$

for all experiments $\sigma \in \mathcal{E}_b$, coarsenings $\kappa \in \mathcal{K}$, and priors $p \in \Delta_\circ$.

An equivalent "distraction free" axiom is studied in Tian (2019). Returning to the above example, Compression Monotonicity demands that the purple posterior distribution in the right-hand panel of Figure 3 costs less than either the red or blue posterior distributions. Intuitively, "merging" states l and c while generating the same information (i.e., conditional signal distributions) about the events $\{l, c\}$ and $\{r\}$ should not increase costs; in any decision problem where l and c are payoff-equivalent, DM should be able to freely ignore any distinction between them.

Notice that the experiments σ and $\sigma_{\langle \kappa|p \rangle}$ are generally not Blackwell comparable. This is illustrated in the right-hand panel of Figure 3, in which the red and blue posterior distributions share the same compression with respect to the coarsening $\{l, c\}$ and $\{r\}$, namely, the purple posterior distribution. However, none of these posterior distributions are mean-preserving spreads of the others because the convex hulls of their supports are not nested. Conversely, no garbling of the purple posterior distribution is a compression of the red or blue posterior distributions. Thus, a cost function that is Blackwell monotone may not be Compression Monotone, and vice versa. Intuitively, Blackwell monotone corresponds to monotonicity with respect to garblings of the signal space, while Compression Monotone corresponds to monotonicity with respect to garblings of the state space.

6.2 Mutual Information: Characterization

With the preceding definitions in hand, we may state our main characterization theorem for Mutual Information, which characterizes this cost function within the UPS class.

Theorem 5. *Let the cost function C^* be UPS and Bounded. If $|\Theta| \geq 3$, then the following are equivalent:*

(i) C^* *is Weakly Compression Invariant.*

(ii) C^* *is Compression Monotone.*

(iii) C^* *is a Mutual Information cost function.*

Proof. See Appendix J. □

The equivalence between points (ii) and (iii) was first noted in Tian (2019) and proved using methods very similar to ours. An immediate corollary of Theorem 5 is that, subject to regularity conditions, Mutual Information is also the unique such cost function within the larger SLP class.

Corollary 5.1. *Let the cost function C^* be SLP and Bounded. If $|\Theta| \geq 3$, then he following are equivalent:*

(i) C^* *is Regular and either (a) Weakly Compression Invariant or (b) Compression Monotone.*

(ii) C^* *is a Mutual Information cost function.*

Proof. Immediate from Theorems 2 and 5. □

The requirement that there are at least three states is needed to rule out trivial cases: when $|\Theta| = 2$, every cost function is Weakly Compression Invariant and every Blackwell monotone cost function is Compression Monotone. It is also worth noting that the boundedness qualifier is necessary for parts of both Theorem 5 and Corollary 5.1, and that the Regularity qualifier is necessary for Corollary 5.1. The follow examples illustrate this:

Example 3 (Symmetric Total Information)**.** Consider the Total Information cost function with symmetric coefficients: $\gamma_{\theta,\theta'} = \overline{c} > 0$ for all pairs of states. It is then easy to see from (TI) that this cost function is Weakly Compression Invariant, in particular, because it is linear in prior beliefs. However, it is not Compression Monotone.

Example 4 (Total Variation Cost)**.** Recall from Subsection 3.2.2 the class of distance-based cost functions (3), which are SLP but not UPS and, hence, not Regular by Theorem 2. The distance-based cost function generated by the total variation distance $d_{TV}(q \mid p) := \frac{1}{2}\sum_\theta |q_\theta - p_\theta|$ is both Weakly Compression Invariant and Compression Monotone. This follows from the well-known fact that the total variation distance is, in fact, an f-divergence (see, e.g., Amari (2016, Section 3.2)).

The proof of Theorem 5 is based on Shannon's (1948) and Fadeev's (1956) classic characterizations of the Shannon entropy function $H(p) := -\sum_\theta p_\theta \log(p_\theta)$.[75] Building on those classic characterizations, we show that any function $F : \Delta(\Theta) \to \mathbb{R}$ that is continuous and satisfies the *recursivity* condition

$$F(p) = F\left(p + p_{\theta'}(\delta_\theta - \delta_{\theta'})\right) + (p_\theta + p_{\theta'})F\left(\frac{p_\theta}{p_\theta + p_{\theta'}}\delta_\theta + \frac{p_{\theta'}}{p_\theta + p_{\theta'}}\delta_{\theta'}\right) \tag{R}$$

[75] See Csiszár (2008) for a survey of functional equation characterizations of Shannon entropy.

for all $\theta, \theta' \in \Theta$ must be proportional to Shannon entropy. To understand the meaning of (R), it is convenient to interpret $F(p)$ as a measure of the "uncertainty" contained in the belief p. Roughly speaking, the recursivity condition (R) states that the total uncertainty in belief p is equal to the "total expected uncertainty" from first merging states θ, θ' into the event $\{\theta, \theta'\}$, and then differentiating between θ and θ' within that event. Note that, while (R) is in some sense analogous to Indifference to Sequential Learning, the former is a recursivity property of the potential function F while the latter is a recursivity property of the cost function C, so that these conditions are not directly comparable. In essence, the proof of Theorem 5 establishes that, given the existence of a UPS representation with potential F, either Weakly Compression Invariant or Compression Monotone imply that (the unique continuous extension of F to the entire simplex) satisfies (R).

6.3 Related Characterizations of Mutual Information

Relation to Caplin et al. (2019b). Theorem 5 is inspired, in large part, by a related characterization of Mutual Information due to Caplin et al. (2019b). In essence, that paper establishes that Mutual Information is the unique Bounded UPS cost function satisfying *both* Axioms 10 and 11. However, a major difference is that Caplin et al. (2019b) take a revealed preference approach, wherein the analyst observes (only) DM's optimal choice of experiment in every decision problem and must draw inferences about her cost function (which is unobserved). CDL therefore impose their key "invariance under compression" (IUC) axiom on DM's choice behavior, whereas we impose our axioms on the cost function itself. Roughly speaking, Caplin et al.'s (2019b) IUC axiom states that DM's optimal state-contingent action probabilities (i) are identical in all payoff-relevant states and (ii) do not change when probability mass is shifted within payoff-equivalent events. Clearly, Axiom 11 implies the former property, while the conjunction of Axioms 10 and 11 implies the latter property.

The marginal contributions of Theorem 5 are twofold. First, we show that Axioms 10 and 11 are *each* sufficient to characterize Mutual Information within the UPS class, while Caplin et al.'s (2019b) IUC axiom combines (the revealed preference analogous of) these two properties, which are logically distinct. In fact, it can be shown that our Theorem 5 implies CDL's characterization. Second, by working directly with the cost function rather than the optimal choice behavior that it induces, our Theorem 5 is amenable to a comparatively elementary proof that elucidates the connection between invariance conditions of DM's information cost function and Fadeev's (1956) classic characterization of Shannon entropy. We believe that these innovations relative to Caplin et al.'s (2019b) have both conceptual and pedagogical value.

Information Theory and Geometry. Since the pioneering work of Shannon (1948), a large literature has provided axiomatic foundations for Shannon entropy, mutual information, and their generalizations using tools from the theory of functional equations. Csiszár (2008) and Ebanks et al. (1998) survey these contributions. To our knowledge, the closest result in this literature to our Theorem 5 is the recent characterization of mutual information by Jiao et al. (2015a).[76] In our language, that paper establishes that any Full Domain UPS cost function that is "invariant to sufficient statistics" — i.e., both Weakly Compression Invariant and invariant to permutations of the state space — must

[76] See also the closely related characterization of KL divergence in Jiao et al. (2014b).

be a Mutual Information cost function. Their result, which is proved via more sophisticated functional equation techniques than we use, is implied by the equivalence of points (i) and (iii) in our Theorem 5.

The related information geometry literature initiated by Čencov (1982) (see also Amari (2016)) studies the differential geometric structure of the probability simplex. Čencov (1982, Chapter 11) (see also Amari (2016, Section 3.5)) characterizes the metric induced by the Fisher information matrix (10) as the essentially unique "invariant" metric on (the tangent space to) the probability simplex, meaning roughly that the distance between two probability distributions p and q with fixed support size $|\text{supp}(p)| = |\text{supp}(q)| = n$ does not depend on (a) the size $|\Theta| \geq n$ of the state space over which they are defined, or (b) the precise states over which they are supported. As discussed in Amari (2016, Section 3), this notion of invariance is closely related to versions of Axioms 10 and 11 studied in the information theory literature. In independent work, Hébert and Woodford (2020a, Proposition 2) use this characterization of the Fisher information matrix to characterize Mutual Information within the class of Full Domain UPS cost functions that (a) have smooth potentials $F \in \mathbf{C}^2(\Delta_\circ)$ and (b) satisfy a strong monotonicity condition that implies both Axiom 10 and Axiom 11, as well as invariance with respect to permutations of the state space. Their result is therefore implied by our Theorem 5.[77]

7 Concluding Remarks

Summary. This paper develops a theory for the cost of optimally acquired information when information can be acquired sequentially. It makes three main contributions. First, we introduce and characterize the class of SLP information cost functions, which are precisely the cost functions that are "rationalizable" by an underlying sequential optimization process. Second, we show that every SLP cost function satisfying mild regularity conditions is a UPS cost function, of the sort used in the rational inattention literature. However, these regularity conditions are economically meaningful: any UPS cost function must have been generated by a Direct Cost for which it is optimal to learn only by Gaussian diffusion signals. Third, within the SLP (and UPS) class, we characterize the new Total Information and the familiar Mutual Information cost functions as the unique cost functions satisfying additional normatively appealing properties. These cost functions are tractable for use in economic applications.

Open Questions. Aside from the obvious possibilities of (i) exploring economic applications of the Total Information cost function and (ii) relaxing various regularity conditions used to obtain our characterization results, our analysis leaves open three main questions for future research.

First, our analysis leaves open the question of how, in general, to compute the Indirect Cost ΦC generated by an arbitrary Direct Cost C. Proposition 2(ii) shows that ΦC is characterized by the "first derivative" (i.e., Locally Linear approximation) of C if and only if direct Poisson learning is an optimal strategy. More subtly, Theorem 3 shows that ΦC is characterized by the "second derivative" (i.e., Locally Quadratic approximation) of C if (and under additional conditions only if) Gaussian

[77] A technical difference is that Čencov's (1982) theorem requires considering probability simplices with arbitrarily high dimension, which means that this result cannot be directly applied to our setting with a fixed finite state space. Consequently, Hébert and Woodford (2020a) assume that the state space is uncountably infinite and consider prior beliefs with (partial) finite support, assuming that it is costless to learn about zero-probability states (cf. Subsections 2.3 and 3.4). Caplin et al. (2019b) make similar assumptions.

diffusion learning is always optimal. It would clearly be desirable to have an algorithmic approach for characterizing ΦC when neither of these (stringent) conditions are met.

Second, our analysis leaves open the full characterization of the Sequentially Prior-Invariant class of cost functions. Given the importance attributed to Prior-Invariance in the literature, it would be highly desirable to have such a characterization, as well as specific and tractable functional forms for use in applications. Proposition 3, Corollary 3.2, and Corollary 6.4 take important initial steps in this direction, but perhaps the main takeaway from these results is that a complete characterization is likely to be quite difficult because, given a Prior-Invariant Direct Cost, one cannot restrict attention to either direct Poisson or Gaussian diffusion strategies. Moreover, while Corollary 3.2(i) establishes that (generically) no Sequentially Prior-Invariant cost function is Regular — which, in our view, suggests that there is little hope for analytically tractable Sequentially Prior-Invariant cost functions — it remains to be seen whether the Sequentially Prior-Invariant and Posterior Separable classes intersect.

Finally, it would be useful to characterize the testable revealed preference implications of SLP information costs, and Total Information in particular, so that our theory can be taken to economic data. This would complement recent advances studying the revealed preference implications of the Posterior Separable and UPS models (Caplin et al. (2019b); Denti (2020)) and of more general models of sequential information sampling (Bloedel (2020a)).

References

Amari, S.-i. (2016). *Information Geometry and its Applications*. Springer.

Angeletos, G.-M. and Sastry, K. (2019). Inattentive economies. Working paper, MIT.

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

Arrow, K. J. (1996). The economics of information: An exposition. *Empirica*, 23(2):119–128.

Arrow, K. J., Blackwell, D., and Girshick, M. A. (1949). Bayes and minimax solutions to sequential decision problems. *Econometrica*, pages 213–244.

Azevedo, E. M., Deng, A., Olea, J. L. M., Rao, J., and Weyl, E. G. (2019). A/b testing with fat tails. Working paper.

Azrieli, Y. and Lehrer, E. (2008). The value of a stochastic information structure. *Games and Economic Behavior*, 63(2):679–693.

Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

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

Bloedel, A. W. (2020a). Augmented stochastic choice and optimal sequential learning. Technical report, Stanford University.

Bloedel, A. W. (2020b). The cost of optimally-acquired information. Mimeo, Stanford University.

Bloedel, A. W. and Segal, I. (2020). Persuading a rationally inattentive agent. Working paper, Stanford University.

Cabrales, A., Gossner, O., and Serrano, R. (2013). Entropy and the value of information for investors. *American Economic Review*, 103(1):360–377.

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

Caplin, A., Dean, M., and Leahy, J. (2019a). Rational inattention, optimal consideration sets and stochastic choice. *Review of Economic Studies*, forthcoming.

Caplin, A., Dean, M., and Leahy, J. (2019b). Rationally inattentive behavior: Characterizing and generalizing shannon entropy. Working paper.

Chade, H. and Schlee, E. (2002). Another look at the radner-stiglitz nonconcavity in the value of information. *Journal of Economic Theory*, 107(2):421–452.

Chambers, C. P., Liu, C., and Rehbeck, J. (2019). Costly information acquisition. Technical report.

Che, Y.-K. and Mierendorff, K. (2019). Optimal dynamic allocation of attention. *American Economic Review*, 109(8):2993–3029.

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

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

de Oliveira, H. (2019). Axiomatic foundations for entropic costs of attention. Working Paper, Penn State.

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

Dean, M. and Neligh, N. (2019). Experimental tests of rational inattention. Technical report.

Denti, T. (2019). Unrestricted information acquisition. Working paper, Cornell University.

Denti, T. (2020). Posterior-separable cost of information. Technical report.

Denti, T., Marinacci, M., and Rustichini, A. (2020). Experimental cost of information. Working paper.

Dessein, W., Galeotti, A., and Santos, T. (2016). Rational inattention and organizational focus. *American Economic Review*, 106(6):1522–1536.

Dewan, A. and Neligh, N. (2018). Estimating information cost functions in models of rational inattention. Technical report.

Dillenberger, D. (2010). Preferences for one-shot resolution of uncertainty and allais-type behavior. *Econometrica*, 78(6):1973–2004.

Dillenberger, D., Lleras, J. S., Sadowski, P., and Takeoka, N. (2014). A theory of subjective learning. *Journal of Economic Theory*, 153:287–312.

Dillenberger, D. and Raymond, C. (2020). Additive-belief-based preferences. Technical report, Working Paper.

Doval, L. (2018). Whether or not to open pandora's box. *Journal of Economic Theory*, 175:127–158.

Ebanks, B., Sahoo, P. K., and Sander, W. (1998). *Characterization of Information Measures*. World Scientific.

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

Ely, J. C., Frankel, A., and Kamenica, E. (2015). Suspense and surprise. *Journal of Political Economy*, 123(1):215–260.

Fadeev, D. K. (1956). On the concept of entropy of a finite probabilistic scheme. *Uspekhi Matematicheskikh Nauk*, 11(1):227–231.

Flynn, J. P. and Sastry, K. (2020). Attention cycles. Working paper, MIT.

Frankel, A. and Kamenica, E. (2019). Quantifying information and uncertainty. *American Economic Review*, 109(10):3650–3680.

Gale, D., Klee, V., and Rockafellar, R. (1968). Convex functions on convex polytopes. *Proceedings of the American Mathematical Society*, 19(4):867–873.

Gentzkow, M. and Kamenica, E. (2014). Costly persuasion. *American Economic Review, Papers and Proceedings*, 104(5):457–462.

Gilboa, I. and Lehrer, E. (1991). The value of information: An axiomatic approach. *Journal of Mathematical Economics*, 20(5):443–459.

Gleyze, S. and Pernoud, A. (2020). Informationally simple incentives. Working paper.

Greenshtein, E. (1996). Comparison of sequential experiments. *The Annals of Statistics*, 24(1):436–448.

Hébert, B. and La'O, J. (2020). Information acquisition, efficiency, and non-fundamental volatility. Working paper, Columbia and Stanford GSB.

Hébert, B. and Woodford, M. (2017). Rational inattention with sequential information sampling. Working paper, Stanford GSB and Columbia University.

Hébert, B. and Woodford, M. (2020a). Neighborhood-based information costs. Working paper, Stanford GSB and Columbia University.

Hébert, B. and Woodford, M. (2020b). Rational inattention when decisions take time. Working paper, Stanford GSB and Columbia University.

Jakobsen, A. (2020). An axiomatic model of persuasion. Working paper, University of Calgary.

Jiao, J., Courtade, T., No, A., Venkat, K., and Weissman, T. (2014a). Information measures: The curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60(12):7616–7626.

Jiao, J., Courtade, T., Venkat, K., and Weissman, T. (2015a). Justification of logarithmic loss via the benefit of side information. *IEEE Transactions on Information Theory*, 61(10):5357–5365.

Jiao, J., Courtade, T. A., No, A., Venkat, K., and Weissman, T. (2014b). Information measures: the curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60(12):7616–7626.

Jiao, J., Courtade, T. A., Venkat, K., and Weissman, T. (2015b). Justification of logarithmic loss via the benefit of side information. *IEEE Transactions on Information Theory*, 61(10):5357–5365.

Kacperczyk, M., Nieuwerburgh, S. V., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocations. *Econometrica*, 84(2):571–626.

Ke, T. T. and Villas-Boas, J. M. (2019). Optimal learning before choice. *Journal of Economic Theory*, 180:383–437.

Köszegi, B. and Matějka, F. (2020). Choice simplification: A theory of mental budgeting and naive diversification. *Quarterly Journal of Economics*, 135(2):1153–1207.

Köszegi, B. and Rabin, M. (2009). Reference-dependent consumption plans. *American Economic Review*, 99(3):909–936.

Lee, P. (1964). On the axioms of information theory. *The Annals of Mathematical Statistics*, 35(1):415–418.

Liang, A. and Mu, X. (2020). Complementary information and learning traps. *Quarterly Journal of Economics*, 135(1):389–448.

Liang, A., Mu, X., and Syrgkanis, V. (2019). Optimal and myopic information acquisition. Technical report.

Liang, A., Mu, X., and Syrgkanis, V. (2020). Dynamically aggregating diverse information. Technical report.

Lin, Y.-H. (2018). Stochastic choice and rational inattention. Technical report.

Lu, J. (2016). Random choice and private information. *Econometrica*, 84(6):1983–2027.

Maćkowiak, B., Matějka, F., and Wiederholdt, M. (2018). Survey: Rational inattention, a disciplined behavioral model. Working paper.

Maćkowiak, B. and Wiederholdt, M. (2009). Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803.

Matějka, F. (2015). Rigid pricing and rationally inattentive consumer. *Journal of Economic Theory*, 158:656–678.

Matějka, F. and McKay, A. (2012). Simple market equilibria with rationally inattentive consumers. *American Economic Review, Papers and Proceedings*, 102(3):24–29.

Matějka, F. and McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298.

Mayskaya, T. (2019). Dynamic choice of information sources. Technical report.

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

Mensch, J. (2020). Screening inattentive agents. Working paper.

Mondria, J. (2010). Porfolio choice, attention allocation, and price comovement. *Journal of Economic Theory*, 145:1837–1864.

Morris, S. and Strack, P. (2019). The wald problem and the equivalence of sequential sampling and ex-ante information costs. Working paper, MIT and Yale University.

Morris, S. and Yang, M. (2019). Coordination and continuous stochastic choice. Technical report, Princeton University.

Moscarini, G. and Smith, L. (2001). The optimal level of experimentation. *Econometrica*, 69(6):1629–1644.

Myatt, D. P. and Wallace, C. (2012). Endogenous information acquisition in coordination games. *Review of Economic Studies*, 79:340–374.

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

Nikandrova, A. and Pancs, R. (2018). Dynamic project selection. *Theoretical Economics*, 13(1):115–143.

Nimark, K. P. and Sundaresan, S. (2019). Inattention and belief polarization. *Journal of Economic Theory*, 180:203–228.

Pomatto, L., Strack, P., and Tamuz, O. (2019). The cost of information. Working paper, Caltech and Yale University.

Rappaport, D. and Somma, V. (2017). Incentivizing information design. Working paper.

Ravid, D. (2019). Bargaining with rational inattention. Technical report, University of Chicago.

Ravid, D. (2020). Ultimatum bargaining with rational inattention. *American Economic Review*, forthcoming.

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

Rustichini, A. (2020). Neural and normative theories of stochastic choice. Working paper, University of Minnesota.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Shannon, C. E. (1958). A note on a partial ordering for communication channels. *Information and Control*, 1:390–397.

Shorrer, R. I. (2018). Entropy and the value of information for investors: The prior-free implications. *Economics Letters*, 164:62–64.

Sims, C. A. (1998). Stickiness. In *Carnegie-Rochester Conference Series on Public Policy*, volume 49, pages 317–356.

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

Sims, C. A. (2010). Rational inattention and monetary economics. In *Handbook of Monetary Economics*, volume 3, pages 155–181. Elsevier.

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

Tian, J. (2019). Attention costs without distraction and entropy. Working paper.

Torgersen, E. (1991). *Comparison of Statistical Experiments*. Cambridge University Press.

Tverberg, H. (1958). A new derivation of the information function. *Mathematica Scandinavica*, 6:297–298.

Čencov, N. (1982). *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society.

Veldkamp, L. (2011). *Information Choice in Macroeconomics and Finance*. Princeton University Press.

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

Wald, A. (1947). Foundations of a general theory of sequential decision functions. *Econometrica*, pages 279–313.

Walker-Jones, D. (2020). Rational inattention and perceptual distance. Working paper, University of Toronto.

Woodford, M. (2012). Inattentive valuation and reference-dependent choice. Technical report, Columbia University.

Woodford, M. (2016). Optimal evidence accumulation and stochastic choice. Technical report.

Yang, M. (2015). Coordination with flexible information acquisition. *Journal of Economic Theory*, 158:721–738.

Yang, M. (2020). Optimality of debt under flexible information acquisition. *Review of Economic Studies*, 87(1):487–536.

Yang, M. and Zeng, Y. (2019). Financing entreprenerial production: Security design with flexible information acquisition. *Review of Financial Studies*, 32(3):819–863.

Yoder, N. (2019). Designing incentives for heterogeneous researchers. Working paper.

Zhong, W. (2017). Time preference and information acquisition. Technical report.

Zhong, W. (2019). Optimal dynamic information acquisition. Working paper, Stanford GSB.

Zhong, W. (2020). The indirect cost of information. Working paper, Stanford GSB.

Appendix

A Material Omitted from the Main Text

A.1 A Theory of Unrestricted Replication

The analysis of Section 5 leaves open two questions. First, is Mutual Information somehow special in exhibiting Decreasing Marginal Cost, or is that property the norm under sequential optimization ? Second, is there a general optimization framework that encompasses both sequential and simultaneous replication?

Example 5 (Residual Variance)**.** Consider the UPS cost function based on the potential function $F(q) = \sum_\theta q_\theta(1-q_\theta)$, which Ely et al. (2015) and Frankel and Kamenica (2019) refer to as the "residual variance" of belief q. The associated Bregman divergence is $D_F(q \mid p) = \|q-p\|^2$, meaning that the cost function is simply the variance of belief movement. In Appendix K, we show that this cost function strictly satisfies (DMC) for certain pairs of experiments and strictly satisfies the opposite "increasing marginal cost" inequality for others.

Unrestricted Replication. As noted above in Subsection 5.1, the notion of Simultaneous Replication (paired with the cost specification (SIC)) that underlies Axioms 8 and 9 lies outside the scope of our main framework based on Sequential Replication. The reason for this is that Sequential Replication requires that DM observe the outcome from her first experiment before running a second experiment, whereas Simultaneous Replication allows DM to run two experiments simultaneously without observing the outcomes of either until the end. However, it is natural to argue that Simultaneous Replication *should* be permitted in a sufficiently flexible model of sequential information acquisition, because DM should have the ability to "ignore" signals from early experiments until they are used.

Here, we introduce a notion of "unrestricted replication" that allows DM to do precisely this, thereby nesting the natural notions of simultaneous and sequential information gathering. The formal definition largely mirrors the formal definition of Sequential Replication. To state it, given a sequence of signal spaces $\{S_t\}_{t=1}^{2T}$, denote the set of length-$2t$ histories by $S^{2t-1} := \prod_{\tau=1}^{2t-1} S_\tau$ with generic element $s^{2t-1} \in S^{2t-1}$.

Definition 17 (Unrestricted Replication)**.** *For $T \in \mathbb{N}$, **length**-$2T$ **Unrestricted Replication** of the **target experiment** σ consists of:*

(i) A collection of (Polish) signal spaces $\{S_t\}_{t=1}^{2T}$ satisfying $S^{2t-1} \subseteq S_{2t}$ (and where S_0 is singleton),
(ii) A collection of (even period) measurable maps $\sigma^{(2t)} : S_{2t} \times \Theta \to \Delta(S_{2t+1})$, and
(iii) A collection of (odd period) measurable maps $\gamma^{(2t+1)} : S^{2t+1} \to \Delta(S_{2t+2})$,

such that σ is Blackwell equivalent to the experiment $\sigma^R : \Theta \to \Delta(S_{2T})$ for which $\sigma^R(\cdot \mid \theta)$ is defined as the marginal distribution on S_{2T} of

$$\prod_{t=0}^{T-1} \sigma^{(2t)}(s_{2t+1} \mid s_{2t}, \theta)\gamma^{(2t+1)}(s_{2t+2} \mid s^{2t+1}, s_{2t}).$$

The only difference between Definition 17 and the definition of Sequential Replication is that point (iii) of the former allows the "garblings" $\gamma^{(2t+1)}$ to provide information about the *entire history* s^{2t+1} of signals, while the latter requires that $\gamma^{(2t+1)}$ provides information only about the past two signals (s_{2t}, s_{2t+1}).[78] Intuitively, in a Sequential Replication, DM can choose to *discard* some previously-acquired information in period $2t+1$ and, once she does so, that information is gone forever. By contrast, in an Unrestricted Replication, DM can choose to *store* some previously-acquired information in period $2t+1$, which can be *recalled* at some later period $2t'+1$ (where $t' > t$). DM does not "observe" stored information before it is recalled, meaning that in the interim periods (a) she cannot condition her strategy on it and (b) also does not incorporate it into her updated beliefs.

It is therefore clear that any valid Sequential Replication is also a valid Unrestricted Replication. Similarly, Simultaneous Replication is the special case of Unrestricted Replication in which all information is stored and only recalled at the end of the acquisition process (i.e., only the final garbling $\gamma^{(2T-1)}$ is informative), which forces all experiments to be conditionally independent and leads to total cost (SIC).

In analogy to the notion of an Indirect Cost function, the following definition formalizes the expected cost of optimally acquired information when DM optimizes over all Unrestricted Replications. Let $\langle \sigma, \gamma \rangle \rightharpoonup_U \sigma$ be shorthand notation for a Unrestricted Replication (of any length $2T$) of target experiment σ.

Definition 18 (Unrestricted Indirect Cost)**.** *Cost function C^* is the **Unrestricted Indirect Cost** generated by the Direct Cost function C if $C^* = \Phi C$, where $\Phi_U C$ is defined by*

$$\Phi_U C(\sigma \mid p) := \inf_{\langle \sigma, \gamma \rangle} \mathbb{E}_{\langle \sigma, \gamma \mid p \rangle} \left[\sum_{t=0}^{T-1} C\left(\sigma^{(2t)}_{\tilde{s}_{2t}} \;\middle|\; q(\cdot \mid \tilde{s}_{2t}) \right) \right] \qquad \text{(UIC)}$$
$$s.t.\ \ \langle \sigma, \gamma \rangle \rightharpoonup_U \sigma.$$

As illustrated by Example 5, the additional flexibility to store and recall acquired information afforded by Unrestricted Replication can lead to strict cost reductions even for natural SLP cost functions. Thus, even when C is SLP, we may have $\Phi_U C < C$.

Unrestricted Learning-Proofness. We can similarly formulate "rationalizabilty" with respect to Unrestricted Replication via a fixed-point condition that resembles Sequential Learning-Proofness. Given (Polish) signal spaces S', R', S'', define a *two-step augmented sequential experiment* as an experiment $\sigma'' * \gamma * \sigma' : \Theta \to \Delta(S' \times R' \times S'')$ with marginal distribution $\sigma'(\cdot \mid \theta) := \text{marg}_{S'}[\sigma'' * \gamma * \sigma'](\cdot \mid \theta)$, garbling function $\gamma(\cdot \mid s') := \text{marg}_{R'}[\sigma'' * \gamma * \sigma'](\cdot \mid \theta, s')$ for all $\theta \in \Theta$, and conditional marginal distributions $\sigma''_{r'}(\cdot \mid \theta) := \text{marg}_{S''}[\sigma'' * \gamma * \sigma'](\cdot \mid \theta, r') \in \Delta(S'')$. Intuitively, $\sigma'' * \gamma * \sigma'$ represents a two-step sequential information acquisition strategy where DM first acquires σ', garbles the realized signal s' into r', and conditional on r' updates her beliefs and acquires a second experiment $\sigma''_{r'}$. At the end of this process, she observes the entire tuple of realized signals (s', r', s''), meaning that she "recalls" s' at the end of the acquisition process.

[78] Point (i) of Definition 17 requires that $S^{2t-1} \subseteq S_{2t}$, but this is actually equivalent (by induction) to Point (i) of Definition 2 which requires that $S_{2t-2} \times S_{2t-1} \subseteq S_{2t}$. We make this notational change in Definition 17 to emphasize the history-dependence in point (iii).

Definition 19 (ULP). *Cost function C is **Unrestricted Learning-Proof (ULP)** if $C = \Psi_U C$ where $\Psi_U C$ is the cost function defined by*

$$\Psi_U(C)(\sigma \mid p) := \inf_{\sigma'' * \gamma * \sigma' \succsim_B \sigma} C(\sigma' \mid p) + \mathbb{E}_{\langle \gamma \circ \sigma' \mid p \rangle}\left[C\left(\sigma''_{\tilde{r}'} \mid q(\cdot \mid \tilde{r}')\right)\right] \tag{ULP}$$

Thus, a cost function is ULP if, and only if, it is weakly cheaper (in expectation) to acquire experiment σ in one shot than it is to acquire any two-step augmented sequential experiment $\sigma'' * \gamma * \sigma'$ that is at least as informative as σ. This formulation implicitly builds in the assumption that any "extra" information contained in $\sigma'' * \gamma * \sigma'$ but not in σ can be freely discarded. The difference between (ULP) and (SLP) is that the former allows DM to manipulate her second-step beliefs by garbling some information acquired in the first step, which is then recalled after the second acquisition step. Let $\mathcal{C}_{ULP}$ denote the set of all ULP cost functions.

A.1.1 Characterization of ULP and Unrestricted Indirect Cost Functions

In analogy to the characterization of SLP and Indirect Cost functions in Subsection 3.1, we may characterize the ULP and Unrestricted Indirect Cost functions via Axiom 1 and the following strengthening of Preference for One-Shot Learning.

Axiom 12 (Robust POSL). *C exhibits **Robust POSL** if*

$$C(\sigma \mid p) \leq C(\sigma' \mid p) + \mathbb{E}_{\langle \gamma \circ \sigma' \mid p \rangle}\left[C\left(\sigma''_{\tilde{r}'} \mid q(\cdot \mid \tilde{r}')\right)\right]$$

for all $\sigma'' * \gamma * \sigma' \sim_B \sigma$ and $p \in \Delta_\circ$.

Axiom 12 states that it is cheaper (in expectation) to acquire all information at once rather than via any Unrestricted Replication with $T = 2$ and without free disposal.

Theorem 6. *For cost function C^*, the following are equivalent:*

(i) C^* *is ULP, i.e.,* $C^* = \Psi_U C^*$.

(ii) C^* *is its own Unrestricted Indirect Cost, i.e.,* $C^* = \Phi_U C^*$.

(iii) C^* *is Blackwell monotone and exhibits Robust POSL.*

Moreover, given any Direct Cost C, the Unrestricted Indirect Cost $\Phi_U(C)$ is a well-defined cost function that is Blackwell monotone and exhibits Robust POSL. Thus, $\Phi_U(\mathcal{C}) = \mathcal{C}_{ULP}$.

Proof. See Appendix K. □

Theorem 6 is the "unrestricted learning" analogue to Theorem 1. The intuitions for and interpretations of these results are similar, so we refrain from commenting further here.

The main importance of Theorem 6 vis a vis the Returns-to-Scale Critique is that every ULP cost function exhibits Decreasing Marginal Cost:

Corollary 6.1. *Every ULP cost function is SLP and exhibits Decreasing Marginal Cost.*

Proof. Immediate from the definitions: (SLP) is implied by (ULP) when the garbling γ is fully informative, while (DMC) is implied by (ULP) when the garbling γ is completely uninformative. □

Thus, the expected cost of optimally acquired information necessarily exhibits Decreasing Marginal Cost when DM's space of information acquisition strategies is sufficiently rich. Moreover, we show below that, except in the case of Total Information, this Decreasing Marginal Cost is sometimes-strict.

A.1.2 Process-Invariance and Total Information

We may now formalize the idea, suggested by Theorem 4, that Total Information is uniquely "process invariant." We formalize "process invariance" via the following strengthening of Robust POSL, which requires that the expected cost of two-step unrestricted replication (without free disposal at the end) is *equal* to the cost of one-shot learning.

Axiom 13 (Process-Invariant)**.** *C is **Process-Invariant** if*

$$C(\sigma \mid p) = C(\sigma' \mid p) + \mathbb{E}_{\langle \gamma \circ \sigma' \mid p \rangle}\left[C\left(\sigma''_{\tilde{r}'} \mid q(\cdot \mid \tilde{r}')\right)\right] \tag{PI}$$

for all $\sigma'' * \gamma * \sigma' \sim_B \sigma$ and $p \in \Delta_\circ$.

The relationship between Process-Invariant and ULP cost functions is analogous to the relationship between UPS and SLP cost functions, but with Unrestricted Replication replacing Sequential Replication. Indeed, by analogy to Corollary 6.1, it is easy to see that any Process-Invariant cost function is both UPS (by letting γ in (PI) be fully informative and invoking Lemma 1) and exhibits Constant Marginal Cost (by letting γ in (PI) be completely uninformative).

Corollary 6.2. *Total Information is the unique Process-Invariant cost function.*

Proof. See Appendix K. □

As discussed in Subsection Section 5.2, we believe that Corollary 6.2 provides a normatively compelling foundation for the Total Information cost function.

A.1.3 Relation to Sequential Replication

We have seen above (e.g., in Example 5) that the Unrestricted Indirect Cost generated by a given Direct Cost function can, in general, be strictly lower than its Indirect Cost. However, in this subsection, we show that the restriction to Sequential Replication is without loss of optimality under the following condition on the Direct Cost:

Axiom 14 (Prior-Concave)**.** *Cost function C is **Prior-Concave** if for each $\sigma \in \mathcal{E}_b$ the map $C(\sigma \mid \cdot) : \Delta_\circ \to \mathbb{R}_+$ is concave.*

We interpret Axiom 14 as capturing DM's ability to "freely dispose of freely-available information." In particular, suppose DM is freely endowed with an experiment τ, either from some external source or from prior rounds of acquisition (the costs of which are already sunk), and can acquire further information *after* observing the realized signal generated by τ. This extra information has *instrumental* value because it allows DM to tailor her continuation strategy on the observed signal. When DM's cost function is prior-dependent, this extra information may also have an *intrinsic* value or cost because it alter's DM's continuation costs even when her continuation strategy does not condition on the realized signal from τ, simply by changing her beliefs. DM's cost function is Prior-Concave if and only if such extra information is never intrinsically costly, namely, $\mathbb{E}_{\langle \tau \mid p \rangle}[C(\sigma \mid \tilde{q})] \leq C(\sigma \mid p)$ for all $\sigma, \tau \in \mathcal{E}_b$ and $p \in \Delta_\circ$. The following examples illustrate why this condition may be normatively desirable:

Example 6 (Value of Information Before Acquisition)**.** Suppose that DM faces a standard (one-shot) information acquisition problem

$$V(p) := \sup_{\sigma \in \mathcal{E}_b} \left[\mathbb{E}_{\pi_{\langle \sigma | p \rangle}} [U(q)] - C(\sigma \mid p) \right] \tag{17}$$

where $U(q) := \max_{a \in A} \mathbb{E}_q [u(a, \theta)]$ for some decision problem (A, u) and C is an information cost function. What is the value of DM to freely observing the outcome of an experiment τ before solving (17)? DM assigns positive value to all such τ if and only if the value function $V(p)$ is convex in her belief p. We show in Appendix K that $V(p)$ is guaranteed to be convex if C is Prior-Concave and that, conversely, violations of Prior-Concavity can lead to $V(p)$ being non-convex. Thus, Axiom 14 is a tight sufficient condition for a Bayesian DM to have a globally positive value of information before solving a standard information acquisition problem.

Example 7 (Prior-Invariance)**.** Every Prior-Invariant cost function is Prior-Concave by definition. As we shall see below (Corollary 6.4), all Sequentially Prior-Invariant cost functions are also Prior-Concave. Thus, any cost function that is not Prior-Concave cannot be reconciled with a Prior-Invariant Direct Cost, for which freely-available information has no intrinsic value or cost. This suggests that violations of Axiom 14 correspond to non-instrumental motives for information avoidance.

The following lemma uses Axiom 14 to bridge the gap between SLP and ULP costs:

Lemma 9. *For cost function C, the following hold:*

(i) If C is Prior-Concave, then so is the Indirect Cost $\Phi(C)$.

(ii) If C is SLP and Prior-Concave, then it is ULP and, hence, exhibits Decreasing Marginal Cost.

(iii) If C is UPS, then the following are equivalent: (a) C is ULP, (b) C is Prior-Concave, and (c) C exhibits Decreasing Marginal Cost.

Proof. See Appendix K. □

Taken together, points (i) and (ii) of Lemma 9 establish that the Indirect Cost ΦC is, in fact, ULP when the underlying Direct Cost function C is Prior-Concave. Meanwhile, point (iii) states that Prior-Concavity is a necessary and sufficient condition for a UPS cost function to be ULP or to exhibit Decreasing Marginal Cost.

Thus, if one accepts the desideratum that "reduced form" information cost functions should be ULP, then one should use (only) those UPS cost functions that are Prior-Concave. The following result, which echoes Theorem 2, shows that every Regular ULP cost function is, in fact, of this form:

Corollary 6.3. *Given any cost function C^*, the following are equivalent:*

(i) C^* is ULP and Regular, with divergence D.

(ii) C^* is UPS and Prior-Concave with potential $F \in \mathbf{C}^2(\Delta_\circ)$, for which the Bregman divergence $D_F = D$.

Proof. Immediate from Corollary 6.1, Theorem 2, and Lemma 9(iii). □

It is easy to see that Total Information is Prior-Concave because it is, in fact, linear in prior beliefs. Note also that Lemmas 7 and 9 imply that Mutual Information is sometimes-strictly Prior-Concave. However, other seemingly natural UPS cost functions (such as the residual variance cost function from Example 5) violate this condition.

In general, it may be difficult to determine whether a given cost function is Prior-Concave because, even for a fixed experiment, the induced posterior distribution varies with the prior. However, for UPS cost functions with potentials $F \in \mathbf{C}^2(\Delta_\circ)$, Prior-Concavity can be equivalently formulated as a checkable concavity condition on the Hessian of F (see Appendix K). The following example illustrates this Hessian condition in the simplest case of binary states:

Example 8 (Binary States)**.** When $\Theta = \{\theta_1, \theta_2\}$, we may parameterize the belief $q = (q_{\theta_1}, q_{\theta_2})$ by the probability $q_{\theta_1} \in [0,1]$ and re-write the potential function $F(q)$ as $\hat{F}(q_{\theta_1}) = F\big((q_{\theta_1}, q_{\theta_2})\big)$. In this case, provided that $F \in \mathbf{C}^2(\Delta_\circ)$, the cost function is Prior-Concave if and only if the map $q_{\theta_1} \mapsto q_{\theta_1}^2\left(1 - q_{\theta_1}\right)^2 \hat{F}''(q_{\theta_1})$ is concave. Notice that this quantity is the marginal cost of sampling from a Gaussian diffusion for an additional instant (cf. Section 4.4 and Morris and Strack (2019)).

In Appendix K, we apply Lemma 9(iii) to characterize the marginal *value* of information for a Bayesian DM.

A.1.4 Sequential Prior-Invariance Revisited

Proposition 3 from Subsection 4.5 establishes that, aside from the $|\Theta| = 2$ case, Sequentially Prior-Invariant cost functions are the UPS and Sequentially Prior-Invariant classes do not intersect. However, that result leaves open the question of how to characterize the Sequentially Prior-Invariant cost functions within the larger SLP class. The following result uses concepts from this subsection to provide a partial characterization of this class:

Corollary 6.4. *If C^* is an Sequentially Prior-Invariant cost function, then:*

(i) C^* is Prior-Concave and ULP.

(ii) If $C^* = \Phi C$ for a Locally Quadratic and Prior-Invariant Direct Cost C, then unless $|\Theta| = 2$ and C^* is the Wald cost function, it exhibits sometimes-strictly Preference for One-Shot Learning and Decreasing Marginal Cost

Proof. Since C^* is Sequentially Prior-Invariant and every Prior-Invariant cost function is Prior-Concave, Lemma 9(i) implies that C^* is Prior-Concave, given which Lemma 9(ii) implies that it is also ULP. This proves point (i). Under the hypotheses of point (ii), Proposition 3 and Corollary 3.2(iii) imply that C^* is not a Total Information cost function. It then follows from Theorem 4 that C^* does not exhibit Constant Marginal Cost and, hence, exhibits sometimes-strictly Decreasing Marginal Cost by Corollary 6.1. This proves point (ii). □

Note, however, that the necessary conditions in Corollary 6.4(i) are *not* sufficient to characterize the Sequentially Prior-Invariant class of cost functions; in particular, both Mutual Information and Total Information satisfy these conditions, but are not Sequentially Prior-Invariant. We leave a full characterization of the Sequentially Prior-Invariant class to future work.

A.2 Omitted Definitions

Definition 20 (Bounded). *Cost function C is **Bounded** if $\sup_{(\sigma,p)\in\mathcal{E}_b\times\Delta_\circ} C(\sigma \mid p) < \infty$.*

Definition 21 (Full Domain). *A **Full Domain cost function** is a map $C : \mathcal{E}\times\Delta_\circ \to \mathbb{R}_+$ such that:*

(i) If σ and τ are Blackwell equivalent, then $C(\sigma \mid p) = C(\tau \mid p)$ for all $p \in \Delta(\Theta)$.

(ii) If $\underline{\sigma}$ is uninformative, then $C(\underline{\sigma} \mid p) = 0$ for all $p \in \Delta(\Theta)$.

(iii) Let $\{(\sigma^{(n)}, p^{(n)})\}_{n\in\mathbb{N}}$ be a sequence of experiment-prior pairs inducing posterior distributions $\pi^{(n)} := \pi_{\langle\sigma^{(n)}|p^{(n)}\rangle} \in \Pi_\circ$. If $\pi^{(n)} \to^{w^*} \pi^* \in \Pi(p^*)$ and $p^* \in \Delta_\circ$, then $C(\sigma^{(n)} \mid p^{(n)}) \to C(\sigma^* \mid p^*)$, where $\pi^* = \pi_{\langle\sigma^*|p^*\rangle}$.

A.3 Auxilliary Lemmas

Lemma 10. *The LLR cost function (LLR) is Locally Quadratic with normalized kernel $\overline{k}_{LLR}(p)$ from (9) defined componentwise by*

$$\left[\overline{k}_{LLR}(p)\right]_{\theta,\theta'} := \begin{cases} \sum_{\theta''\neq\theta}(\beta_{\theta,\theta''} + \beta_{\theta'',\theta}), & \text{if } \theta = \theta' \\ -\beta_{\theta,\theta'} - \beta_{\theta',\theta}, & \text{if } \theta \neq \theta' \end{cases} \tag{18}$$

Proof. Follows from straightforward calculation given the definition of the normalized kernel in (9), Lemma 3(i), and the potential function $\mathcal{G}(q \mid p)$ in Definition 5. □

Lemma 11. *For Locally Quadratic Direct Cost function C with kernel k:*

(i) The kernel k is positive semidefinite on the tangent space to the simplex, i.e., $y^\top k(p)y \geq 0$ for all $y \cdot \mathbf{1} = 0$ and $p \in \Delta_\circ$.

(ii) There exists a kernel $\hat{k}$ of C that is symmetric and satisfies $k(p)p = \mathbf{0}$ for all $p \in \Delta_\circ$.

(iii) A continuous matrix-valued function $\hat{k} : \Delta_\circ \to \mathbb{R}^{\Theta\times\Theta}$ satisfies (LQ) for C if and only if there exist continous functions $\boldsymbol{f}, \boldsymbol{g} : \Delta_\circ \to \mathbb{R}^\Theta$ such that $\hat{k}(p) \equiv k(p) + \boldsymbol{f}(p)\mathbf{1}^\top + \mathbf{1}\boldsymbol{g}(p)^\top$.

Proof. See Appendix K. □

Lemma 12. *Given any Direct Cost function C, the map $\Phi_{DP}C : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ defined in (DPIC) is a well-defined cost function.*

Proof. See Appendix K. □

B An Equivalent Belief-Based Formulation

Now we consider the sequential minimization of information cost. $\forall \pi \in \Delta^2(X)$, we first define the belief processes that replicate the information structure π.

Definition 22. *$\forall \pi \in \Delta^2(X)$, A $2T$-period Markov chain $\langle q_t \rangle$ (define on $(\Omega, \mathcal{F}, \mathcal{P})$) **replicates** π if 1) $q_{2T} \sim \pi$, 2) $\mathbb{E}[q_{2t+1}|q_{2t}] = q_{2t}$ and 3) $\mathbb{E}[q_{2t-1}|q_{2t}] = q_{2t}$.*

The first condition means $\langle q_t \rangle$ eventually replicates π. The second condition means from any period $2t$ to $2t+1$, information is acquired and belief is updated according to Bayes rule. The third condition means from any period $2t-1$ to $2t$, information is "discarded" and belief contracts. By defining $\langle q_t \rangle$ as in Definition 22, it is implicitly assumed that 1) acquiring information only measurable to belief and time is sufficient (for optimality, which will be defined and proved later) and 2) information is freely disposable. The conditions in Definition 22 are denoted by $\langle q_t \rangle \rightharpoonup \pi$.

The we define the indirect cost of information:

Definition 23. *$C^* : \Delta^2(X) \to \mathbb{R}^+$ is an **indirect information cost** function if $\exists$ direct information cost $C : \Delta^2(X) \to \mathbb{R}^+$ s.t $\forall \pi \in \Delta^2(X)$:*

$$C^*(\pi) = \inf_{\langle q_t \rangle} \mathbb{E}\left[\sum_{t=0}^{T-1} C(\pi_{2t}(q_{2t+1}|q_{2t}))\right] \tag{19}$$
$$s.t.\ \langle q_t \rangle \rightharpoonup \pi$$

where $\pi_{2t}(q_{2t+1}|q_{2t})$ denotes the conditional distribution of q_{2t+1} on q_{2t}.

Equation (19) defines a program that searches for a cost minimizing belief process that replicates the target information structure π. In the objective function, only in even periods the cost of belief change is counted because by definition in even periods information is acquired and in odd periods information is freely discarded. The optimization over $\langle q_t \rangle$ implicitly allows T to be chosen as well. The integrability of Equation (19) is in general not guaranteed, but for $C \in \mathcal{C}$ the expression is well defined. $\forall C \in \mathcal{C}$, Equation (19) is a well defined non-negative real number, and hence a map ϕ can be defined as $\Phi(C) = C^*$ indicating that $\Phi(C)$ is the indirect information cost derived from solving Equation (19) with direct information cost C.

In the appendix, I prove Lemma 16 which shows that we can consider a more complicated sequential signal structure which is not necessarily Markovian. However, the terminal belief distribution can always be replicated by a process satisfying Definition 22 and with weakly lower total cost. Therefore, Equation (19) could be thought as the maximal flexibility benchmark.

B.0.1 Key Lemmas

Lemma 13. *Given cost function C, the map $C_{RA} : \Pi_b \to \mathbb{R}_+$ defined by*

$$C_{RA}(\pi) := \inf_{P \in \Delta(\Pi_b)} \mathbb{E}_P[C(\tilde{\pi})] \tag{RA}$$
$$s.t.\ \mathbb{E}_P[\tilde{\pi}] = \pi$$

is (i) a well-defined Randomizaton Averse cost function, and (ii) the pointwise largest Randomizaton Averse cost function that is majorized by C.

Proof. See Appendix K. □

Lemma 14. *Given any cost function C, we have $\Phi C = \Phi C_{RA}$.*

Proof. See Appendix K. □

Lemma 15. *If C is a Locally Quadratic cost function, then its lower random-averse envelope C_{RA} is also Locally Quadratic and has the same kernel as C.*

Proof. See Appendix K. □

Lemma 16. *For all Randomizaton Averse Direct Cost functions C, $\forall$ $2T$-period signal process $\langle s_t \rangle$ defined as before, there exists $2T$-period $\langle q_t \rangle$ satisfying Definition 22 s.t. the posterior induced by s_{2t} is distributed as q_{2T} and:*

$$\sum_{t=0}^{T-1} \mathbb{E}_{q_{2t}}[C(\pi_t(q_{2t+1}|q_{2t}))] \leq \sum_{t=0}^{T-1} \mathbb{E}_{s_{2t}}[C(\nu(s_{2t+1}|s_{2t}))]$$

Proof. Given any process $\langle s_t \rangle$, it induces a joint probability measure $m(q_0, q_1, \cdots, q_{2T})$, where each q_{2t} is the conditional measure of x on s_{2t} and each q_{2t+1} is the conditional measure of x on (s_{2t}, s_{2t+1}). Now we convert this measure to get a process $\langle \widehat{q_t} \rangle$ satisfying Definition 22. Define the conditional marginals of joint probability measure $\widehat{m}$ by:

$$\begin{cases} m(q_{2t}, q_{2t+1}) = m(q_{2t})\widehat{m}(q_{2t+1}|q_{2t}) \\ m(q_{2t+1}, q_{2t+2}) = m(q_{2t+1})\widehat{m}(q_{2t+2}|q_{2t+1}) \end{cases}$$

Then let $\widehat{m}(q_0, q_1, \cdots, q_{2t}) = \prod \widehat{m}(q_{2t+1}|q_{2t})\widehat{m}(q_{2t+2}|q_{2t+1})$. It is easy to verify by induction that $\widehat{m}(q_{2t}) = m(q_{2t})$. Now we verify that the process $\langle \widehat{q_t} \rangle$ defined according to $\widehat{m}$ satisfies the conditions in Definition 22.

First, by definition $\langle \widehat{m_t} \rangle$ is Markov — the distribution of q_{t+1} solely depends on q_t. Second, we verify the martingale property. By definition:

$$\begin{aligned} \mathbb{E}_{\widehat{m}}[q_{2t+1}|q_{2t}] \cdot \widehat{m}(q_{2t}) &= \int \widehat{m}(q_{2t+1}, q_{2t}) q_{2t+1} \mathrm{d}q_{2t+1} \\ &= \int m(\nu, q_{2t}) \nu \mathrm{d}\nu \\ &= \int \left(\int_{(s_{2t}, s_{2t+1}) \to (q_{2t}, \nu)} f(s_{2t}, s_{2t+1}) \frac{q_{2t}(\cdot) f(s_{2t+1}|s_{2t}, \cdot)}{f(s_{2t}, s_{2t+1})} \mathrm{d}s_{2t}, s_{2t+1} \right) \mathrm{d}\nu \\ &= \int_{s_{2t} \to q_{2t}} \int q_{2t}(\cdot) f(s_{2t+1}|s_{2t}, \cdot) \mathrm{d}s_{2t+1} \mathrm{d}s_{2t} \\ &= q_{2t} \cdot m(q_{2t}) = q_{2t} \cdot \widehat{m}(q_{2t}) \end{aligned}$$

Notation $s \to q$ means q is the posterior belief induced by signal s. The second equality is by definition of $\widehat{m}$. The third equality is by the Bayes rule that determines q_{2t+1}. The forth equality is by 1) s_{2t} determines q_{2t} 2) Fubini theorem. The last equality is straight forward.

$$\begin{aligned} &\mathbb{E}_{\widehat{m}}[q_{2t+1}|q_{2t+2}] \cdot \widehat{m}(q_{2t+2}) \\ =& \int \widehat{m}(q_{2t+1}, q_{2t+2}) q_{2t+1} \mathrm{d}q_{2t+1} \\ =& \int m(q_{2t}, q_{2t+1}) \cdot \frac{m(q_{2t+1}, q_{2t+2})}{m(q_{2t+1})} q_{2t+1} \mathrm{d}q_{2t}, q_{2t+1} \\ =& \int m(\nu, q_{2t+2}) \nu \mathrm{d}\nu \end{aligned}$$

$$
\begin{aligned}
&= \int \left(\int \left(\int_{(s_{2t}, s_{2t+1}, s_{2t+2}) \to (q, \nu, q_{2t+2})} f(s_{2t}, s_{2t+1}, s_{2t+2}) \cdot \frac{q(\cdot) f(s_{2t+1} | s_{2t}, \cdot)}{f(s_{2t}, s_{2t+1})} \mathrm{d}s_{2t}, s_{2t+1}, s_{2t+2} \right) \mathrm{d}q \right) \mathrm{d}\nu \\
&= \int \left(\int \left(\int_{(s_{2t}, s_{2t+1}, s_{2t+2}) \to (q, \nu, q_{2t+2})} f(s_{2t}, s_{2t+1}, s_{2t+2}) \cdot q_{2t+2} \mathrm{d}s_{2t}, s_{2t+1}, s_{2t+2} \right) \mathrm{d}q \right) \mathrm{d}\nu \\
&= q_{2t+2} \cdot \widehat{m}(q_{2t+2})
\end{aligned}
$$

The second equality is by definition of $\widehat{m}$. The forth equality is by the Bayes rule that determines q_{2t+1}. The fifth equality is by the Bayes rule that determines q_{2t+2}. The last equality is straight forward. Moreover, by definition $\widehat{m}$ always has the same marginal distributions as m. So the distributions of induced belief at period $2T$ are the same. Therefore, $\widehat{m}$ defines a process $\langle \widehat{q_t} \rangle$ satisfying Definition 22.

Now we show that the cost of $\langle \widehat{q_t} \rangle$ is weakly lower than that of $\langle s_t \rangle$:

$$
\begin{aligned}
\mathbb{E}_{s_{2t}} [C(\nu(s_{2t+1} | s_{2t}))] &= \mathbb{E}_{q_{2t}} \left[\mathbb{E}_{s_{2t}} [C(\nu(s_{2t+1} | s_{2t})) | q_{2t}] \right] \\
&\geq \mathbb{E}_{q_{2t}} \left[C \left(\mathbb{E}_{s_{2t}} [\nu(s_{2t+1} | s_{2t}) | q_{2t}] \right) \right] \\
&= \mathbb{E}_{q_{2t}} [C(\widehat{m}(\cdot | q_{2t}))]
\end{aligned}
$$

The first equality is law of iterated expectation. The inequality is because C is Randomizaton Averse. □

C Proof of Theorem 1

C.1 Proof that Points (i)–(iii) are Equivalent

The implications (i) $\implies$ (iii) and (ii) $\implies$ (i) are immediate from the definitions. Thus, it suffices to show that (iii) $\implies$ (ii). Thus, suppose that C^* is Blackwell monotone and exhibits Preference for One-Shot Learning. Fix any $p \in \Delta_\circ$, $\sigma \in \mathcal{E}_b$, and length-$2T$ Sequential Replication $\langle \boldsymbol{\sigma}, \boldsymbol{\gamma} \rangle \rightharpoonup \sigma$. Suppose that the $t = 1$ garbling $\gamma^{(1)}$ is not fully informative. Then because C^* is Blackwell monotone, it is weakly cheaper to replace $\sigma^{(0)}$ and $\gamma^{(1)}$ with $\tilde{\sigma}^{(0)} := \gamma^{(1)} \circ \sigma^{(0)} : \Theta \to \Delta(S_2)$ and fully informative $\tilde{\gamma}^{(1)} : S_2 \to \Delta(S_2)$, while leaving the rest of the Sequential Replication unchanged. Then, because C^* satisfies Preference for One-Shot Learning, it is weakly cheaper to acquire $\sigma^{(2)} * \tilde{\sigma}^{(1)}$ in one shot rather than in two steps. Proceeding inductively in this manner, it is easy to see that

$$
C^*(\sigma \mid p) \leq \mathbb{E}_{\langle \boldsymbol{\sigma}, \boldsymbol{\gamma} | p \rangle} \left[\sum_{t=0}^{T-1} C^* \left(\sigma^{(2t)}_{\tilde{s}_{2t}} \;\middle|\; q(\cdot \mid \tilde{s}_{2t}) \right) \right].
$$

Since this holds for all $\langle \boldsymbol{\sigma}, \boldsymbol{\gamma} \rangle \rightharpoonup \sigma$, it follows that $C^*(\sigma \mid p) \leq \Phi C^*(\sigma \mid p) \leq C^*(\sigma \mid p)$, where the second inequality is by definition of the Φ operator. Thus, $C^* = \Phi C^*$, as desired.

C.2 Proof that $\Phi(\mathcal{C}) = \mathcal{C}_{SLP}$

By the equivalence of points (i) and (iii) in the present Theorem 1, it suffices to show that $C^* := \Phi C$ is a well-defined cost function that satisfies Axioms 1 and 2.

We do so in a series of lemmas. In fact, we prove a slightly stronger set of results that pertain to Full Domain Direct and Indirect Cost functions. As the reader can verify, the corresponding statements for cost functions defined on $\mathcal{E}_b \times \Delta_\circ$ follow from the same arguments by restricting all

posterior distributions to be supported on some Δ_δ with $\delta > 0$. By Lemmas 13 and 14, it is without loss of generality to assume that the Direct Cost function C is Randomizaton Averse. We do so throughout the proof, which uses the belief-based notation laid out in Appendix B.

Lemma 17. *For any Full Domain Direct Cost C, the map $C^* : \Pi \to \mathbb{R}_+$ is a well-defined Indirect Cost function.*

Proof. It suffices to establish continuity; the other two properties are automatic. By Prokhorov's theorem, $\Delta^2(\Theta)$ is a compact and separable metric space equipped with the Lévy-Prokhorov metric (henceforth, *l-p* metric). Wlog, we consider the open set of generic information structures in $\Delta^2(\Theta)$. The analysis applies to any generic subspace. $\forall \pi$, since the set of generic information structures is open, there exists an interior closed ball $B_{\delta_0}(\pi)$. Then, Heine-Cantor theorem implies that $C(\pi)$ is uniformly continuous on $B_{\delta_0}(\pi)$. $\forall \epsilon > 0$, $\forall \pi \in \Delta^2(x)$, let $\delta < \delta_0$ be the uniform continuity parameter of C w.r.t. $\epsilon' < \epsilon$.

Upper semi-continuity: $\forall \pi$ and $\pi' \in B_\delta(\pi)$. Let $q_0 = \mathbb{E}_\pi[\nu]$ and $q_0' = \mathbb{E}_{\pi'}[\nu]$. Then $\|q_0 - q_0'\| < 2\delta$.[79] Since $B_{\delta_0}(\pi)$ is interior, δ can be picked small enough s.t. $\forall \pi'$, there exists ν s.t. $q_0' = \alpha\nu + (1-\alpha)q_0$ and $\alpha < \epsilon'$. Now pick $\langle q_t \rangle \rightharpoonup \pi$ with total cost lower than $C^*(\pi) + \epsilon$. We construct a sequential learning strategy replicating π': 1) acquire some information and get posterior q_0 and ν, 2) conditional on q_0, follow $\langle q_t \rangle$; conditional on ν, stays. The terminal belief q_{2T} is exactly π with $(1-\alpha)$ probability and ν with α probability. Let π'' be its distribution, then π'' has the same mean as π' and $d_{l-p}(\pi', \pi'') \le d_{l-p}(\pi, \pi') + d_{l-p}(\pi, \pi'') < 2\delta + \alpha$. 3) Denote $\pi' - \pi'' = m^+ - m^-$ where both are positive measures, bounded by $2\delta + \alpha$ and satisfy $\mathbb{E}_{m^+}[\nu] = \mathbb{E}_{m^-}[\nu]$. Contract m^- and then acquire posterior according to m^+. By construction, we replicated π' through 1)-3). Count the total cost: step 1) acquires information structure within $B_{2\delta+\epsilon'}(\delta_{q_0})$; step 2) incurs cost weakly less than $C^*(\pi) + \epsilon$; step 3) acquires some information with less than $2\delta + \alpha$ probability. By Lemma 16, this process can always be modified to satisfy Definition 22, replicate π and has weakly lower cost. Therefore, if we choose ϵ' sufficiently small, the total cost is bounded above by $C^*(\pi) + 3\epsilon$, hence $\overline{\lim}_{\pi' \to \pi} C^*(\pi') \le C^*(\pi)$.

lower semi-continuity: δ can be picked sufficiently small that $\forall q_0$, there also exists ν' and α' s.t. $q_0 = \alpha\nu' + (1-\alpha)q_0'$. Then previous argument also shows that $C^*(\pi) \le \underline{\lim}_{\pi' \to \pi} C^*(\pi')$.

Therefore, since $C^*(\pi)$ is both upper semi-continuous and lower semi-continuous, $C^*(\pi)$ is continuous at any generic π. Since $0 \le C^*(\pi) \le C(\pi)$, $C^*(\pi)$ is bounded. □

Lemma 18. *For any Full Domain Direct Cost C, the Full Domain Indirect Cost function C^* is Blackwell monotone.*

Proof. $\forall \pi, \pi' \in \Delta^2(\Theta)$ and $\pi \le_{BW} \pi'$, by definition, there exists $\pi''(\nu|q)$ s.t. $\pi'(\nu) = \mathbb{E}[\pi(q)\pi''(\nu|q)]$ and $\mathbb{E}[\pi''(\nu|q)] = q$. From joint distribution $\pi(q)\pi''(\nu|q)$, we can obtain marginal distribution $\widehat{\pi}(q|\nu)$. Now $\forall$ 2T-period $\langle q_t \rangle$ replicating π', define $2T$-period $\langle \widehat{q}_t \rangle$ replicating π: $\widehat{q}_t = q_t$ when $t < 2T$ and $\widehat{q}_{2T}|\widehat{q}_{2T-1} \sim \mathbb{E}[\widehat{\pi}(\widehat{q}_{2T}|q_{2T})|\widehat{q}_{2T-1}]$. It is easy to verify that $\langle \widehat{q}_t \rangle$ satisfies the conditions in Definition 22 and hence $\langle \widehat{q}_t \rangle$ replicates π. Noticing that $\sum C(\pi_t(q_{2t+1}|q_{2t})) = \sum C(\pi_t(\widehat{q}_{2t+1}|\widehat{q}_{2t}))$ and therefore $C^*(\pi) \le C^*(\pi')$. Axiom 1 is verified. □

[79] $\pi' - \pi$ can be written as $m^+ - m^-$ where both are positive measure and bounded by δ by the definition of $l-p$ metric. Then $\|q_0 - q_0'\| = \|\int \nu(\mathrm{d}m^+ - \mathrm{d}m^-)\| \le 2\delta$.

Lemma 19. *For any Full Domain Direct Cost C, the Full Domain Indirect Cost function C^* is Randomizaton Averse.*

Proof. For any $p \in \Delta_\circ$, $\Pi(p)$ is a compact and separable subset of Π. $\forall \epsilon$, there exists a finite ϵ-net of $\Pi(p)$. Now $\forall P \in \Delta(\Pi(p))$, discretizing P on the ϵ-net gives finite distribution $\widehat{P}$ ϵ-close to P (under $l-p$ metric). Therefore, there exists finite distributions $\widehat{P}$ converging to P. Now given $\widehat{P}$, $\forall \epsilon$, there exists a uniform upper-bound T for all π in $\text{supp}(\widehat{P})$ such that $\langle q_t \rangle_{t=0}^{2T}$ replicates π and the total cost is lower than $C^*(\pi)+\epsilon$. This implies $C^*(\mathbb{E}_{\widehat{P}}) \leq \mathbb{E}_{\widehat{P}}[C^*(\pi)]+\epsilon$. By continuity of C^* (shown in Lemma 17), $\mathbb{E}_{\widehat{P}}[C^*(\pi)] \to \mathbb{E}_P[C^*(\pi)]$ and $C^*(\mathbb{E}_{\widehat{P}}[\pi]) \to C^*(\mathbb{E}_P[\pi])$. To sum up, $C^*(\mathbb{E}_P[\pi]) \leq \mathbb{E}_P[C^*(\pi)]$. □

Lemma 20. *For any Full Domain Direct Cost C, the Full Domain Indirect Cost function C^* satisfies Preference for One-Shot Learning.*

Proof. $\forall \pi(q), \pi'(\nu|q)$ and $\pi''(\nu) = \mathbb{E}_\pi[\pi'(\nu|q)]$. $\forall \epsilon > 0$. Pick any $\delta > 0$ and take the closures of δ-interior points of all $\Delta(\Theta)$'s, denote it by D^δ. Then open set $\Delta(\Theta) \setminus D^\delta$ is shrinking to an empty set and hence there exists δ s.t. $\pi(D^\delta) > 1-\epsilon$. Now we construct a sequence of information structures that replicates π''.

First, let $q_0 = \mathbb{E}_\pi[\nu]$. Let $q_0' = \mathbb{E}_\pi[\nu|\nu \in D^\delta]$ and $q_0'' = \mathbb{E}_\pi[\nu|\nu \notin D^\delta]$. Then $q_0 = \pi(D^\delta)q_0' + \pi(\Delta(\Theta) \setminus D^\delta)q_0''$. Define information structure π_0^δ with support $\left\{q_0', q_0''\right\}$ and the corresponding probabilities. Define information structures $\pi_0'^\delta(\nu) = \pi(\nu|\nu \in \Delta(\Theta) \setminus D^\delta)$. Now partition D^δ to finite Borel subsets each of diameter $\eta < \delta$, denote the partition by $\left\{D_i^{\delta,\eta}\right\}$. Define $\widetilde{\pi}^{\delta,\eta}$ with support $\left\{\nu_i = \mathbb{E}_\pi[\nu|\nu \in D_i^{\delta,\eta}]\right\}$ and distribution $\widetilde{\pi}^{\delta,\eta}(\nu_i) = \pi(D_i^{\delta,\eta})$. Define $\widetilde{\pi}_i'^{\delta,\eta} = \mathbb{E}_\pi\left[\pi'(\nu|q)|q \in D_i^{\delta,\eta}\right]$. Now consider the following sequential information structure: 1) acquire π_0^δ. 2) acquire $\pi_0'^\delta$ conditional on q_0''; acquire $\widetilde{\pi}^{\delta,\eta}$ conditional on q_0'. 3) acquire π' following $\pi_0'^\delta$; following $\widetilde{\pi}^{\delta,\eta}$, acquire $\widetilde{\pi}_i'^{\delta,\eta}$ conditional on ν_i. Now we verify that the sequential information structure replicates π'': $\forall$ Borel set $U \subset \Delta(\Theta)$,

$$\begin{aligned}
\text{Prob}(U) =& \text{Prob}(U|q_0')\pi_0^\delta(q_0') + \text{Prob}(U|q_0'')\pi_0^\delta(q_0'') \\
=& \sum_i \text{Prob}(U|\nu_i, q_0')\widetilde{\pi}^{\delta,\eta}(\nu_i)\pi_0^\delta(q_0') + \mathbb{E}_{\pi_0'^\delta}[\pi'(U|q)]\pi_0^\delta(q_0'') \\
=& \sum_i \widetilde{\pi}_i'^{\delta,\eta}(U)\widetilde{\pi}^{\delta,\eta}(\nu_i)\pi_0^\delta(q_0') + \mathbb{E}_\pi\left[\pi'(U|q)|q \in \Delta(\Theta) \setminus D^\delta\right]\pi(\Delta(\Theta) \setminus D^\delta) \\
=& \sum_i \mathbb{E}_\pi\left[\pi'(U|q)|q \in D_i^{\delta,\eta}\right]\pi(D_i^{\delta,\eta})\pi(D^\delta) + \mathbb{E}_\pi\left[\pi'(U|q)|q \in \Delta(\Theta) \setminus D^\delta\right]\pi(\Delta(\Theta) \setminus D^\delta) \\
=& \mathbb{E}_\pi[\pi'(\nu|q)] = \pi''(U)
\end{aligned}$$

By definition, when $\delta, \eta \to 0$, $\pi_0^\delta \xrightarrow{w-*} \delta_{q_0}$, $\widetilde{\pi}^{\delta,\eta} \xrightarrow{w-*} \pi$. By continuity of C and C^*, $C(\pi_0^\delta) \to 0$ and $C^*(\widetilde{\pi}^{\delta,\eta}) \to C^*(\pi)$. Now we calculate the cost of $\widetilde{\pi}_i'^{\delta,\eta}$. $\forall i$, $\forall q \in D_i^{\delta,\eta}$, by definition of δ and η, $\|q-\nu_i\| \leq \eta$ and there exists $q' \in \Delta(\Theta)$ s.t. $\|q-q'\| \geq \delta$ and ν_i is a linear combination of q, q'. Define information structure $\pi_{i,q}'(\cdot) = \frac{\|q'-\nu_i\|}{\|q'-q\|}\pi'(\cdot|q) + \frac{\|\nu_i-q\|}{\|q'-q\|}\delta_{q'}$. Then $d_{l-p}(\pi'(\cdot|q), \pi_{i,q}'(\cdot)) \leq \frac{\eta}{\eta+\delta}$. Now consider information structure $\widetilde{\pi}_i'^{\delta,\eta} = \mathbb{E}_\pi[\pi_{i,q}'(\nu)|q \in D_i^{\delta,\eta}]$, then $d_{l-p}(\widetilde{\pi}_i'^{\delta,\eta}, \widetilde{\pi}_i'^{\delta,\eta}) \leq \frac{\eta}{\eta+\delta}$ (because conditional on each $q \in D_i^{\delta,\eta}$, the measure on any Borel set differs by at most $\frac{\eta}{\eta+\delta}$). Since C^* is continuous on D^δ, and hence uniformly continuous by Heine-Cantor, there exists η s.t. $\frac{\eta}{\eta+\delta}$ is the uniform continuity parameter

w.r.t. ϵ for C^*. Therefore:

$$\begin{aligned}\mathbb{E}_\pi\left[C^*(\pi'(\nu|q))|q\in D_i^{\delta,\eta}\right]\geq&\mathbb{E}_\pi\left[C^*(\pi'_{i,q}(\nu))|q\in D_i^{\delta,\eta}\right]-\epsilon\\ \geq&C^*\left(\mathbb{E}_\pi\left[\pi'_{i,q}(\nu)|q\in D_i^{\delta,\eta}\right]\right)-\epsilon\\ \geq&C^*\left(\widetilde{\pi}_i'^{\delta,\eta}\right)-2\epsilon\end{aligned}$$

The two ϵ each comes from the distance between $\pi',\pi'_{i,q}$ and $\widetilde{\pi}_i'^{\delta,\eta},\widetilde{\pi}_i'^{\delta,\eta}$. The second inequality is implied by the fact that C^* is Randomizaton Averse, which is verified before.

Now we construct a belief process replicating π'' and satisfy Definition 22. $\forall\widetilde{\pi}_i'^{\delta,\eta}$, there exists a $2T_i$ process $\langle q_t^i\rangle\rightharpoonup\widetilde{\pi}_i'^{\delta,\eta}$ such that $\sum C(\pi_t(q_{2t+1}^i|q_{2t}^i))\leq C^*(\widetilde{\pi}_i'^{\delta,\eta})+\epsilon$. There also exists a $2T_0$ process $\langle q_t^0\rangle\rightharpoonup\widetilde{\pi}^{\delta,\eta}$ s.t. $\sum C(\pi_t(q_{2t+1}^0|q_{2t}^0))\leq C^*(\widetilde{\pi}^{\delta,\eta})+\epsilon$. Let $T=\max\{T_i\}+T_0+1$.

First, let $q_1\sim\pi_0^\delta$, $q_2=q_1$. If $q_2=q_0''$, $q_3\sim\mathbb{E}_{\pi_0'^\delta}[\pi'(\nu|q)]$ and all $q_t\equiv q_3$ for $t>3$. If $q_2=q_0'$, $q_{t+2}\equiv q_t^0$, and $q_{t+T_0+2}\equiv q_t^i$ conditional on $q_{T_0+2}=\nu_i$. The total direct cost of this process is:

$$\begin{aligned}&C(\pi_0^\delta)+\pi_0^\delta(q_0'')C\left(\mathbb{E}_{\pi_0'^\delta}[\pi'(\nu|q)]\right)\\&+\pi_0^\delta(q_0')\left(\sum_{t=0}^{2T_0}C(\pi_t(q_{2t+1}^0|q_{2t}^0))+\sum_i\widetilde{\pi}^{\delta,\eta}(\nu_i)\left(\sum_{t=0}^{2T_i}C(\pi_t(q_{2t+1}^i|q_{2t}^i))\right)\right)\\\leq&C(\pi_0^\delta)+(1-\pi(D^\delta))C\left(\mathbb{E}_{\pi_0'^\delta}[\pi'(\nu|q)]\right)\\&+\pi(D^\delta)\left(C^*(\widetilde{\pi}^{\delta,\eta})+\epsilon+\sum_i\pi(D_i^{\delta,\eta}|D^\delta)C^*(\widetilde{\pi}_i'^{\delta,\eta})+\epsilon\right)\\\leq&C(\pi_0^\delta)+(1-\pi(D^\delta))C\left(\mathbb{E}_{\pi_0'^\delta}[\pi'(\nu|q)]\right)\\&+\pi(D^\delta)\left(C^*(\widetilde{\pi}^{\delta,\eta})+\sum_i\pi(D_i^{\delta,\eta}|D^\delta)\mathbb{E}_\pi[C^*(\pi'(\nu|q))|q\in D_i^{\delta,\eta}]+4\epsilon\right)\\\to&C^*(\pi)+\mathbb{E}_\pi[C^*(\pi'(\nu|q))]\text{ when }\delta\to0\&\frac{\eta}{\delta}\to0\end{aligned}$$

By Lemma 16, the process $\langle q_t\rangle$ can always be transformed to one satisfying Definition 22 with lower total direct cost. This suggests that $C^*(\pi'')\leq C^*(\pi)+\mathbb{E}_\pi[C^*(\pi'(\nu|q))]$. □

D Proof of Proposition 1

D.1 Proof of Point (i)

Let $|\Theta|=n$. If $n=2$, the result is trivial because the fully informative experiment $\overline{\sigma}$ is the unique nontrivial partitional experiment. We therefore focus on the case $n\geq3$.

Let $E\subsetneq\Theta$ be some event with $|E|\geq2$, which we may without loss (by relabelling states appropriately) take to be $E=\{\theta_1,\ldots,\theta_m\}$ for some m satisfying $1<m<n$. Let $\langle S_E,\sigma_E\rangle$ denote the partitional experiment that identifies E, i.e., $S_E:=\{0,1\}$ and $\sigma_E(1\mid\theta):=\mathbf{1}(\theta\in E)$. Let $S_2\equiv\{\phi,\theta_1,\ldots,\theta_m\}$, and for each $s\in\{0,1\}$ define the partitional experiments $\langle S_2,\sigma_2|_i\rangle$ by

$$\sigma_2|_i(\cdot\mid\theta)=\begin{cases}\delta_\phi(\cdot), & \text{if } i=0\\ \delta_\theta(\cdot), & \text{if } i=1.\end{cases}$$

That is, $\sigma_2(\cdot \mid 1) = \overline{\sigma}$ and $\sigma_2(\cdot \mid 0) = \underline{\sigma}$. Let $\sigma := \sigma_2 * \sigma_E$. Because C is SLP and has Full Domain, it follows from Theorem 1 that C exhibits the Preference for One-Shot Learning inequality

$$\begin{aligned} C(\sigma) &\leq C(\sigma_E) + p(E)C(\overline{\sigma}) + (1 - p(E))C(\underline{\sigma}) \\ &= C(\sigma_E) + p(E)C(\overline{\sigma}) \end{aligned}$$

for all $p \in \Delta_\circ$, where where the equality follows from the fact that $C(\underline{\sigma}) = 0$. Thus, we have

$$C(\sigma) \leq \inf_{p \in \Delta_\circ} [C(\sigma_E) + p(E)C(\overline{\sigma})] = C(\sigma_E).$$

Because C is Blackwell monotone, we also have $C(\sigma_E) \leq C(\sigma)$. It follows that $C(\sigma) = C(\sigma_E)$. The remainder of the proof proceeds by induction on the size of E.

D.2 Proof of Point (ii)

Let the state space be $\Theta = \{1, \ldots, n\}$. Let C^* be SLP, Prior-Invariant, and satisfy Constant Marginal Cost. By Corollary 1.2, C^* is also Dilution Linear. Then by Pomatto et al. (2019, Theorem 1), C^* is a LLR cost function. If C^* is non-zero, then there exists at least one coefficient $\beta_{ij} > 0$. Without loss of generality, suppose that $\beta_{12} > 0$. Let $\mathcal{E}_1 \subseteq \mathcal{E}_b$ denote the collection of bounded experiments for which $\sigma_i = \sigma_j$ for all $i, j \neq 1$. That is, $\sigma \in \mathcal{E}_1$ is informative only about whether the state is θ_1 or not. Then for any $\sigma \in \mathcal{E}_1$, we have

$$C^*(\sigma) = B_{12} D_{KL}(\sigma_1 \mid \sigma_2) + B_{21} D_{KL}(\sigma_2 \mid \sigma_1)$$

where $B_{12} := \sum_{j=2}^n \beta_{1j} > 0$ and $B_{21} := \sum_{j=2}^n \beta_{j1} \geq 0$. By Lemma 10 and Proposition 3, it can be shown that the Gaussian Indirect Cost (GIC) induced by C^*

$$\Phi_G C^*(\sigma \mid p) = (B_{12} + B_{21}) \cdot [p_1 D_{KL}(\sigma_1 \mid \sigma_2) + (1 - p_1) D_{KL}(\sigma_2 \mid \sigma_1)].$$

By Lemma 29, it can be shown that there exit $p \in \Delta_\circ$ and $\sigma \in \mathcal{E}_1$ for which $\Phi_G C^*(\sigma \mid p) < C^*(\sigma)$. Since $\Phi C^* \leq \Phi_G C^*$, it follows that $C^* \neq \Phi C^*$. Thus, by Theorem 1, C^* is not SLP. Contradiction. Therefore, any such C^* must be identically zero.

E Proof of Theorem 2

E.1 Proof that (ii) $\implies$ (i)

(i) $\implies$ (ii). Let $C^*(\sigma \mid p) \equiv \mathbb{E}_{\pi_{\langle \sigma \mid p \rangle}}[D_F(q \mid p)]$, where $D_F(q \mid p) := F(q) - F(p) - \nabla F(p) \cdot (q - p)$ is the Bregman divergence generated by $F \in \mathbf{C}^2(\Delta_\circ)$. C^* is clearly Locally Linear. To see that it is Regular, we compute

$$\begin{aligned} \nabla_p D_F(q \mid p) &= -\nabla F(p) + \nabla F(p) - \mathcal{H}F(p) \cdot (q - p) \\ &= -\mathcal{H}F(p) \cdot (q - p), \end{aligned}$$

which is continuous.

E.2 Proof that (i) $\implies$ (ii)

Let C be SLP and Locally Linear. The first property implies C is Dilution Linear. This and the second property imply that C is Posterior Separable.

Step 1: Integral Representation. Because C is SLP, it satisfies POSL. Take any $p \in \Delta_\circ$ and $\pi \in \Pi_\circ(p)$ with binary support $\text{supp}(\pi) = \{q', q''\} \subset \Delta_\circ$. Take any finite-support $\pi' \in \Pi_\circ(q')$ and let $\pi'' \in \Pi_\circ(q'')$ be defined by $\pi'' := \delta_{q''}$. Define $\hat{\pi} \in \Pi_\circ(p)$ by $\hat{\pi}(\cdot) := \pi(q')\pi'(\cdot) + \pi(q'')\pi''(\cdot)$. Then

$$\hat{C}(\hat{\pi}) = \pi(q'')D(q'' \mid p) + \pi(q')\mathbb{E}_{\pi'}[D(\nu \mid p)] \tag{20}$$

$$\leq \hat{C}(\pi) + \pi(q'')D(q'' \mid q'') + \pi(q')\mathbb{E}_{\pi'}[D(\nu \mid q')] \tag{21}$$

$$= \pi(q')D(q' \mid p) + \pi(q'')D(q'' \mid p) + \pi(q')\mathbb{E}_{\pi'}[D(\nu \mid q')], \tag{22}$$

where (20) is by definition of $\hat{\pi}$, (21) is by POSL, and (22) is by definition of π and the fact that $D(q'' \mid q'') = 0$. Combining (20) and (22) yields

$$0 \leq D(q' \mid p) + \mathbb{E}_{\pi'}[D(\nu \mid q') - D(\nu \mid p)]. \tag{23}$$

Note that the RHS of (23) and $D(q' \mid p)$, viewed as functions of p, are both minimized (and equal to 0) at $p = q'$. Differentiating (23) with respect to p at $p = q'$ in direction $y := p - q'$ therefore delivers

$$\frac{\partial}{\partial \epsilon}\mathbb{E}_{\pi'}[D(\nu \mid q' + \epsilon y)] = \mathbb{E}_{\pi'}[J(\nu \mid q') \cdot y] = 0, \tag{24}$$

where passing the derivative through the expectation operator is permitted because π' has finite and interior support. Note that the preceding argument works for all $p \in \Delta_\circ$ holding $q' \in \Delta_\circ$ fixed, which implies that (24) holds for all $y \in \mathcal{T}(\Delta)$. It follows that

$$\mathbb{E}_{\pi'}[J(\nu \mid q')] = \alpha \mathbf{1} \tag{25}$$

for some $\alpha \in \mathbb{R}$. Note that for any $\theta \in \Theta$ the function $\hat{J}(\nu \mid q') := J(\nu \mid q') - J_\theta(\nu \mid q')\mathbf{1}$ is also a valid derivative of $D(\nu \mid q')$ (i.e., satisfies Definition 10). This fact combined with (25) implies that

$$\mathbb{E}_{\pi'}\left[\hat{J}(\nu \mid q')\right] = \mathbf{0}. \tag{26}$$

Note that (25) holds for all finite-support $\pi' \in \Pi(q')$ and that $\hat{J}(\cdot \mid q')$ is continuous on $\Delta_\circ$ by Definition 10. Thus, Lemma 21 (stated and proved below) applied to $J(\cdot \mid q')$ implies that there exists a matrix-valued function $q' \mapsto k(q') \in \mathbb{R}^{|\Theta| \times |\Theta|}$ such that $\hat{J}(\nu \mid q') = -k(q')(\nu - q')$.

An application of the Gradient Theorem then yields

$$D(\nu \mid q') = -\int_a^b A(r(x))(\nu - r(x)) \cdot r'(x)dx \tag{27}$$

for any smooth curve $r : [a, b] \to \Delta_\circ$ for which $r(0) = \nu$ and $r(1) = q'$, where $a, b \in \mathbb{R}$. Note that $r'(x) \in \mathcal{T}(\Delta)$.

Step 2: Continuous Differentiability. Fix $y \in \mathcal{T}(\Delta)$. Consider any smooth curve $r : [0,1] \to \Delta_\circ$ for which (i) $r(0) = \nu - \delta y$ and $r(1) = q'$, (ii) $r(t^*) = \nu$ and $r(t^* + \epsilon) = \nu + \eta y$ for some $t^* \in (0,1)$ and all $\eta \in [-\delta, \delta]$, and (iii) $r(\cdot)$ is uniformly bounded away from $\mathrm{bd}(\Delta)$, the boundary of the simplex. It is clear that such curves exist whenever $\delta > 0$ is sufficiently small because $\nu, q' \in \Delta_\circ$, i.e., are interior. Note that the truncated curve $r|_{[t^*,1]}$ serves as a smooth path from ν to q', for which the representation (27) applies. We have

$$\frac{\partial}{\partial \epsilon} D(\nu + \epsilon y \mid q')\Big|_{\epsilon=0} = \frac{d}{dt} D(r(t) \mid q')\Big|_{t=t^*} \tag{28}$$

$$= -\frac{d}{dt}\left[\int_t^1 A(r(x))(r(a) - r(x)) \cdot r'(x) dx\right]\Big|_{t=t^*} \tag{29}$$

$$= k(r(t^*))(\nu - r(t^*)) \cdot r'(t^*) - \int_{t^*}^1 k(r(x)) r'(t^*) \cdot r'(x) dx \tag{30}$$

$$= -\int_{t^*}^1 k(r(x)) y \cdot r'(x) dx, \tag{31}$$

where (28) is by definition of the curve r, (29) follows from (27), (30) follows from the standard Leibniz rule,[80] and (31) follows from the facts that $r(t^*) = \nu$ and $r'(t^*) = y$.

The preceding argument establishes that the (two-sided) directional derivative of $D(\cdot \mid q)$ at ν in direction y exists and is finite, for any $\nu \in \Delta_\circ$ and $y \in \mathcal{T}(\Delta)$. By a simple adaptation of Theorem 25.2 and Corollary 2.5.5.1 of Rockafellar (1970), this implies that $D(\cdot \mid q')$ is continuously differentiable on $\Delta_\circ$. That is, there exists a continuous function $D(\cdot \mid q') : \Delta_\circ \to \mathbb{R}^\Theta$ such that

$$\frac{\partial}{\partial \epsilon} D(\nu + \epsilon y \mid q')\Big|_{\epsilon=0} = \nabla_\nu D(\nu \mid q') \cdot y$$

for all $y \in \mathcal{T}(\Delta)$.

Step 3: Twice Continuous Differentiability. Differentiating $D(\cdot \mid q)$ at ν in direction y a second time, we obtain

$$\frac{\partial^2}{\partial \epsilon' \partial \epsilon} D(\nu + \epsilon y + \epsilon' y \mid q')\Big|_{\epsilon=\epsilon'=0} = -\frac{d}{dt}\left[\int_t^1 k(r(x)) y \cdot r'(x) dx\right]\Big|_{t=t^*} \tag{32}$$

$$= k(r(t^*)) y \cdot r'(t^*) \tag{33}$$

$$= y^\top k(\nu) y \tag{34}$$

where (32) follows from (31), (33) follows from the fact that $r'(t + \eta) = y$ for all $\eta \in (-\delta, \delta)$ and the Leibniz rule, and (34) is by definition of the curve r (namely that $r(t^*) = \nu$ and $r'(t^*) = y$).

Step 4: UPS Representation. We now show that C has a UPS representation. Let $p^* \in \Delta_\circ$ be given and define the convex function $F : \Delta_\circ \to \mathbb{R}_+$ by $F(q) := D(q \mid p^*)$. Letting $L(q,p) := D(q \mid p) - D(q \mid p^*)$, we have $D(q \mid p) = F(q) + L(q,p)$.

[80] Note that the Leibniz rule applies because the derivative of the integrand in (30), namely the function $(x,a) \mapsto A(r(x)) r'(a) \cdot r'(x) \in \mathbb{R}$, is continuous by construction.

We claim that $F(\cdot, p)$ is affine. To see this, first note that displays (32)–(34) imply that, for each $y \in \mathcal{T}(\Delta)$, the map

$$p \mapsto \frac{\partial^2}{\partial \epsilon' \partial \epsilon} D(q + \epsilon y + \epsilon' y \mid p)\Big|_{\epsilon=\epsilon'=0}$$

is constant, which in turn implies that

$$\frac{\partial^2}{\partial \epsilon' \partial \epsilon} L(q + \epsilon y + \epsilon' y, p) = 0. \tag{35}$$

Given any $q_1, q_2 \in \Delta_\circ$, it is then easy to see that (35) applied at all $q \in [q_1, q_2]$ and with $y := q - q'$ implies that $F(\alpha q_1 + (1-\alpha) q_2, p) = \alpha F(q_1, p) + (1-\alpha) F(q_2, p)$. That is, $F(\cdot, p)$ is affine, as claimed.

Because $L(\cdot, p)$ is affine, we have $\mathbb{E}_\pi[L(q,p)] = \mathbb{E}_{\delta_p}[L(q,p)] = L(p,p)$ for all $\pi \in \Pi_b(p)$. Therefore, $\hat{C}(\pi) = \mathbb{E}_\pi[F(q) + L(q,p)] = \mathbb{E}_\pi[F(q)] + L(p,p)$ for all $\pi \in \Pi_b(p)$. Because $\hat{C}(\delta_p) = 0$, it must be that $L(p,p) = -F(p)$. This establishes the desired UPS representation $\hat{C}(\pi) = \mathbb{E}_\pi[F(q) - F(p)]$. Finally, it is easy to see that $F \in \mathbf{C}^2(\Delta_\circ)$.

E.3 Supporting Lemmas

Lemma 21. *Let $f : \Delta_\circ \to \mathbb{R}^n$ be continuous. Let $p \in \Delta_\circ$ be given. If $\mathbb{E}_\pi[f(q)] = \mathbf{0}$ for all $\pi \in \Pi_\circ(p)$ with $|supp(\pi)| \leq 3$, then there exists a matrix $k \in \mathbb{R}^{n \times \Theta}$ such that $f(q) = k \cdot (q - p)$.*

Proof. Towards a contradiction, suppose that there exist $\nu, \mu \in \Delta_\circ$, $\alpha \in (0,1)$, and $i \in [n]$ such that $f_i(\alpha\nu + (1-\alpha)\mu) > \alpha f_i(\nu) + (1-\alpha) f_i(\mu)$. If $\alpha\nu + (1-\alpha)\mu = p$, then we are done: letting $\pi := \alpha\delta_\nu + (1-\alpha)\delta_\mu$, we have $\pi \in \Pi_\circ(p)$ and $\mathbb{E}_\pi[f_i(q)] < f_i(p)$, which contradicts the hypothesis of the lemma.

Suppose instead that $\alpha\nu + (1-\alpha)\mu \neq p$. Define $\pi := \alpha\delta_\nu + (1-\alpha)\delta_\mu$ as above. Define $\hat{p} := p + \epsilon[p - (\alpha\nu + (1-\alpha)\mu)]$, where $\epsilon > 0$ is taken small enough that $\hat{p} \in \Delta_\circ$. Define $\pi' := \frac{1}{1+\epsilon}\delta_{\hat{p}} + \frac{\epsilon}{1+\epsilon}\delta_{\alpha\nu+(1-\alpha)\mu}$, which satisfies $\pi' \in \Pi_\circ(p)$ by construction. Define $\pi'' := \frac{1}{1+\epsilon}\delta_{\hat{p}} + \frac{\epsilon}{1+\epsilon}\pi$, which is an MPS of π' and satisfies $\pi'' \in \Pi_\circ(p)$. It follows that

$$\mathbb{E}_{\pi''}[f_i(q)] - \mathbb{E}_{\pi'}[f_i(q)] = \frac{\epsilon}{1+\epsilon}[\alpha f_i(\nu) + (1-\alpha) f_i(\mu) - f_i(\alpha\nu + (1-\alpha)\mu)] < 0$$

by our supposition, which again violates the hypothesis of the lemma.

A symmetric argument establishes that $f_i(\alpha\nu + (1-\alpha)\mu) < \alpha f_i(\nu) + (1-\alpha) f_i(\mu)$ is also impossible. It follows that $f(\alpha\nu + (1-\alpha)\mu) = \alpha f(\nu) + (1-\alpha) f(\mu)$ for all $\nu, \mu \in \Delta_\circ$ and $\alpha \in (0,1)$. By standard arguments, this implies the existence of $k \in \mathbb{R}^{n \times \Theta}$ and $z \in \mathbb{R}^n$ such that $f(q) = kq + z$; because $\mathbb{E}_{\delta_p}[f(q)] = f(p) = \mathbf{0}$, it must be that $z = -kp$. □

Lemma 22. *If cost function C^* satisfies the conditions of Theorem 2 (i.e., is SLP and Regular), then for every $\sigma \in \mathcal{E}_b$ it satisfies $C^*(\sigma \mid \cdot) \in \mathbf{C}^1(\Delta_\circ)$.*

Proof. See Appendix K. □

F Proof of Theorem 3

By Lemmas 13, 14 and 15, it is without loss of generality to assume that the Direct Cost function C is Randomizaton Averse. We do so throughout the proof, which uses the belief-based notation laid out in Appendix B.

F.1 Proof of Point (i)

Suppose that the Direct Cost function C is Locally Quadratic and exhibits Preference for Incremental Learning with kernel k. Define the UPS cost function $\underline{C}$ by

$$\underline{C}(\pi) := \mathbb{E}_\pi[F(\nu)] - F(\mathbb{E}_\pi[\nu]) \leq C(\pi)$$

where the convex function $F(q)$ has Hessian matrix $2k(q)$. By Lemma 1 and Theorem 1, we have $\underline{C} = \Phi\underline{C}$. Therefore, because $\Phi : \mathcal{C} \to \mathcal{C}$ is an increasing map, it is sufficient to show that $\Phi(C) \leq \underline{C}$.

First, we show that $\Phi(C)(\pi) \leq \underline{C}(\pi)$ for $\pi \in \Pi_b$ have binary support. We prove this by finding π' arbitrarily close to π under L–P metric and $\langle q_t \rangle \rightharpoonup \pi'$ with cost arbitrarily close to $\underline{C}$. Since $|\text{supp}(\pi)| = 2$, denote the two posterior beliefs by ν_1, ν_2. Pick $M \in \mathbb{N}$, $\forall i \in \mathbb{N}$, $i \leq M$ define $\lambda_i = \frac{i}{M}$. Consider the subspace $\{q_i = \lambda_i \nu_1 + (1-\lambda_i)\nu_2\}$. Let q_{m_0} be the closest point to $\mathbb{E}_\pi[\nu]$. Define information structure $\widehat{\pi}$ to be with prior q_{m_0} and posteriors q_0, q_M. Then $\lim_{M\to\infty} d(\pi, \widehat{\pi})_{lp} = 0$ By continuity of $\Phi(C)$ and $\underline{C}$, $\forall \epsilon > 0$, $\exists M$ large enough that $|\Phi(C)(\pi) - \Phi(C)(\widehat{\pi})| \leq \epsilon$ and $|\underline{C}(\pi) - \underline{C}(\widehat{\pi})| \leq \epsilon$.

Now consider the following process $\langle q_t \rangle$, defined as follows: $q_0 = q_{m_0}$, $\pi_t(q_{t+1}|q_i) = \frac{1}{2}\delta_{q_{i+1}} + \frac{1}{2}\delta_{q_{i-1}}$ when $i \in [1, M-1]$. $\pi_t(q_{t+1}|q_0) = \delta_{q_0}$ and $\pi_t(q_{t+1}|q_M) = \delta_{q_M}$. In other words, $\langle q_t \rangle$ is a standard random walk in $\{q_i\}$ stopped at absorbing boundary $\{q_0, q_M\}$. Let $T \in \mathbb{N}$ be the length of the process $\langle q_t \rangle$. Then it is easy to verify that $\text{prob}(q_T \in \{q_0, q_M\}) \to 1$ when $T \to \infty$.[81] Therefore, $\exists T$ large enough, s.t. if we let $q_T \sim \pi'$, then $|\Phi(C)(\pi') - \Phi(C)(\widehat{\pi})| \leq \epsilon$ and $|\underline{C}(\pi') - \underline{C}(\widehat{\pi})| \leq \epsilon$. By definition, finite process $\langle q_t \rangle \rightharpoonup \pi'$. Notice that $\langle q_t \rangle$ does not have the information disposal periods. The cost of $\langle q_t \rangle$ is:

$$\begin{aligned}
\sum C(\pi_t(q_{t+1}|q_t)) \leq & \mathbb{E}\Big[\sum (F(q_{t+1}) - F(q_t))\Big] && (36)\\
& + \mathbb{E}\Big[\sum \Big|(F(q_{t+1}) - F(q_t)) - \frac{1}{M^2}(\nu_2 - \nu_1)^T k(q_t)(\nu_2 - \nu_1)\Big|\Big] && (37)\\
& + \mathbb{E}\Big[\sum \Big|\frac{1}{M^2}(\nu_2 - \nu_1)^T k(q_t)(\nu_2 - \nu_1) - C(\pi_t(q_{t+1}|q_t))\Big|\Big] && (38)
\end{aligned}$$

The first term (36) is exactly $\mathbb{E}_{\pi'}[F(\nu)] - F(\mathbb{E}_{\pi'}[\nu]) = \underline{C}(\pi')$. Now consider the second term (37). $\forall q_i$, let $f(\alpha) = F(q_i + \frac{\alpha}{M}(\nu_2 - \nu_1))$. Then:

$$\begin{aligned}
& \frac{1}{2}\inf_{\alpha\in[0,1]} f''(\alpha) \leq f(1) - f(0) - f'(0) \leq \frac{1}{2}\sup_{\alpha\in[0,1]} f''(\alpha)\\
\iff & \inf_{\alpha\in[0,1]} \frac{1}{2}\frac{1}{M^2}(\nu_2 - \nu_1)^T \mathcal{H}F(q_i + \frac{\alpha}{M}(\nu_1 - \nu_1))(\nu_2 - \nu_1)\\
& \leq F(q_{i+1}) - F(q_i) - \frac{1}{M}\nabla F(q_i)(\nu_2 - \nu_1)\\
& \leq \sup_{\alpha\in[0,1]} \frac{1}{2}\frac{1}{M^2}(\nu_2 - \nu_1)^T \mathcal{H}F(q_i + \frac{\alpha}{M}(\nu_2 - \nu_1))(\nu_2 - \nu_1)\\
\implies & \Big|F(q_{i+1}) - F(q_i) - \frac{1}{M}\nabla F(q_i)(\nu_2 - \nu_1) - \frac{1}{M^2}(\nu_2 - \nu_1)^T k(q_i)(\nu_2 - \nu_1)\Big|\\
& \leq \frac{1}{M^2}\|\nu_2 - \nu_1\|^2 \cdot \sup_{q'\in[q_{i-1}, q_{i+1}]} \|k(q') - k(q)\|
\end{aligned}$$

[81] Let P_T be such probability, then i) P_T is increasing, ii) $P_{T+M} \geq P_T + (1 - P_T)\frac{1}{2^M}$.

Since the kernel $k(q)$ is continuous on $\Delta_\circ$, it is uniformly continuous on $[q_0, q_m]$. Therefore, M can be picked large enough that $\sup_{q' \in [q_{i-1}, q_{i+1}]} \|k(q') - k(q)\| \le \epsilon$. This implies

$$(37) \le \epsilon \cdot \mathbb{E}\Big[\sum \|q_{t+1} - q_t\|^2\Big] = \epsilon \cdot \mathbb{E}_{\pi'}[\|\nu - q_{m_0}\|^2]$$

Now consider (38). Since $[q_0, q_M]$ is compact, there exists uniform δ satisfying (LQ). Therefore, when M is picked larger than $\frac{1}{\delta}$, q_{t+1} is always within $B_\delta(q_t)$ and:

$$(38) \le \epsilon \cdot \mathbb{E}\Big[\sum \|q_{t+1} - q_t\|^2\Big] = \epsilon \cdot \mathbb{E}_{\pi'}[\|\nu - q_{m_0}\|^2]$$

To sum up:

$$\begin{aligned}
\sum C(\pi_t(q_{t+1}|q_t)) \le& (36) + (37) + (38) \\
\le& \underline{C}(\pi') + 2\epsilon \cdot \mathbb{E}_{\pi'}[\|\nu - q_{m_0}\|^2] \\
\implies \Phi(C)(\pi') \le& \underline{C}(\pi') + 2\epsilon \cdot \mathbb{E}_{\pi'}[\|\nu - q_{m_0}\|^2] \\
\implies \Phi(C)(\pi) \le& \underline{C}(\pi') + 2\epsilon \cdot \mathbb{E}_{\pi'}[\|\nu - q_{m_0}\|^2] + 2\epsilon \\
\le& \underline{C}(\pi) + 2\epsilon \cdot \mathbb{E}_{\pi'}[\|\nu - q_{m_0}\|^2] + 4\epsilon
\end{aligned}$$

Since ϵ can be arbitrarily small, $\Phi(C)(\pi) \le \underline{C}(\pi)$. Therefore, $\Phi(C)(\pi) = \underline{C}(\pi)$.

Next, we prove the statement for general $\pi \in \Pi_b$. $\forall \epsilon > 0$, let δ be the continuity parameter of $\Phi(C)$ and $\underline{C}$ at π. First consider any finite (Borel) partition $\cup_{i=1}^M D_i$ of $\Delta(\Theta)$ where the diameter of any D_i is bounded by δ. Without loss of generality, we consider the case $\pi(D_i) > 0$ and $M \ge 3$. Let $q_i = \mathbb{E}_\pi[\nu|\nu \in D_i]$. Define $\pi'(q) = \sum_i \pi(D_i) \cdot \delta_{q_i}(q)$. Then $d(\pi, \pi')_{lp} \le \delta$ and hence $|\Phi(C)(\pi) - \Phi(C)(\pi')| \le \epsilon$. Let $q_0 = \mathbb{E}_\pi[\nu]$. Now we consider a decomposition of π':

$$\begin{cases}
\pi_1 = \pi(D_1)\delta_{q_1} + (1 - \pi(D_1))\delta_{\mathbb{E}_\pi[\nu|\nu \notin D_1]} & \\
\pi_i(\cdot|q) = \dfrac{\pi(D_i)}{\sum_{j \ge i} \pi(D_j)}\delta_{q_i}(\cdot) + \dfrac{\sum_{j>i} \pi(D_j)}{\sum_{j \ge i} \pi(D_j)}\delta_{\mathbb{E}_\pi[\nu|\nu \in \cup_{j>i} D_i]}(\cdot) & \text{when } q = \mathbb{E}_\pi[\nu|\nu \in \cup_{j>i} D_j] \\
\pi_i(\cdot|q) = q & \text{otherwise}
\end{cases}$$

By definition $\pi'(\nu) = \mathbb{E}[\prod_{i=1}^{M-1} \pi_i]$. Therefore by recursively applying Axiom 2:

$$\begin{aligned}
\Phi(C)(\pi') \le& \mathbb{E}\Big[\sum \Phi(C)(\pi_i)\Big] \\
=& \pi(D_1)F(q_1) + \sum_{j>1} \pi(D_j) F(\mathbb{E}_\pi[\nu|\nu \in \cup_{j>1} D_j]) - F(q_0) \\
&+ \left(\sum_{j>1} \pi(D_j)\right)\left(\frac{\pi(D_2)}{\sum_{j>1}\pi(D_j)}F(q_2) + \frac{\sum_{j>2}\pi(D_j)}{\sum_{j>1}\pi(D_j)}F(\mathbb{E}_\pi[\nu|\nu \in \cup_{j>2} D_j]) - F(\mathbb{E}_\pi[\nu|\nu \in \cup_{j>1} D_j])\right) \\
&+ \cdots \\
&+ \left(\sum_{j>1} \pi(D_j)\right)\prod_{i=1}^{M-2}\frac{\sum_{j>i}\pi(D_j)}{\sum_{j \ge i}\pi(D_j)}\left(\frac{\pi(D_{M-1})}{\pi(D_{M-1}) + \pi(D_M)}F(q_{M-1}) + \frac{\pi(D_M)}{\pi(D_{M-1}) + \pi(D_M)}F(q_M)\right. \\
&\left. - F(\mathbb{E}_\pi[\nu|\nu \in D_{M-1} \cup D_M])\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum \pi(D_i) F(q_i) - F(q_0) \\
&= \underline{C}(\pi')
\end{aligned}
$$

The first equality utilizes the result with binary support information structures. The second equality is from cancelling out terms. Since ϵ can be chosen arbitrarily, $\text{supp}(\pi') \subseteq \text{co}[\text{supp}(\pi)]$ for all $\delta > 0$, and $\pi' \to^{w^*} \pi$ as $\delta \to 0$, we have $\Phi(C)(\pi) \leq \underline{C}(\pi)$.

F.2 Proof of Point (ii)

Let C be Locally Quadratic and $C^* = \Phi C$ be UPS and Locally Strongly Convex. We begin with an important auxiliary lemma.

Lemma 23. *For any $\pi \in \Pi_b$ and $\epsilon > 0$, if $\langle q_t \rangle \rightharpoonup \pi$ is ϵ-optimal in the belief-based problem (19), then*

$$
\mathbb{E}\Big[\sum \|q_{2t+2} - q_{2t+1}\|^2\Big] \leq \frac{\epsilon}{m} \tag{39}
$$

Proof. Let $\langle q_t \rangle \rightharpoonup \pi$ satisfy the hypothesis of the lemma. Using the fact that C^* is UPS, we have

$$
\begin{aligned}
\mathbb{E}\Big[\sum C^*(\pi_{2t}(q_{2t+1}|q_{2t}))\Big] =& \mathbb{E}\Big[\sum \mathbb{E}_{\pi_{2t}}[F(q_{2t+1}) - F(q_{2t})|q_{2t}]\Big] \\
=& \mathbb{E}\Big[\sum \mathbb{E}_{\pi_t}[F(q_{t+1}) - F(q_t)|q_t]\Big] - \mathbb{E}\Big[\sum \mathbb{E}_{\pi_{2t+1}}[F(q_{2t+2}) - F(q_{2t+1})|q_{2t+1}]\Big] \\
=& \mathbb{E}\Big[\sum (F(q_{t+1}) - F(q_t))\Big] - \mathbb{E}\Big[\sum \mathbb{E}_{\pi_{2t+1}}[F(q_{2t+2}) - F(q_{2t+1})|q_{2t+1}]\Big] \\
=& \mathbb{E}[F(q_{2T}) - F(q_0)]] - \mathbb{E}\Big[\sum \mathbb{E}_{\pi_{2t+1}}[F(q_{2t+2}) - F(q_{2t+1})|q_{2t+1}]\Big] \\
\geq& \mathbb{E}\Big[\sum C(\pi_{2t}(q_{2t+1}|q_{2t}))\Big] - \epsilon - \mathbb{E}\Big[\sum \mathbb{E}_{\pi_{2t+1}}[F(q_{2t+2}) - F(q_{2t+1})|q_{2t+1}]\Big]
\end{aligned}
$$

By definition $C^* \leq C$, therefore this implies:

$$
\mathbb{E}\Big[\sum (F(q_{2t+1}) - F(q_{2t+2}))\Big] \leq \epsilon
$$

Moreover, because C^* is Locally Strongly Convex, it can be shown that

$$
C^*(\text{Dist}(q_{2t+1}|q_{2t+2})) \geq m\mathbb{E}\Big[\|q_{2t+1} - q_{2t+2}\|^2\Big]
$$

where $\text{Dist}(q_{2t+1}|q_{2t+1})$ denotes the conditional distribution of q_{2t+1} on q_{2t+2}. Combining the two inequalities above yields (39). □

Lemma 23 provides an upper bound for the amount of "discarded" information. Its total variance must be bounded above by $\frac{\epsilon}{m}$, which can be arbitrarily small when we choose ϵ small. The key intuition here is that because C^* is Locally Strongly Convex, all information is costly, including that which is discarded in the end. Moreover, because C^* is UPS, those costs on discarded information are avoidable—if we replace each information structure in the sequence with a replicating process. Therefore, the total amount of discarded cost is bounded above by the approximation error of the replicating processes.

We now show that for any replicating process that approximates C^*, the probability of a path leaving a small neighbourhood around q_0 is bounded. $\forall \pi \in \Pi_b$, let $q_0 = \mathbb{E}_\pi[\nu]$. Suppose the diameter of $\text{supp}(\pi)$ is less than δ_0. Pick an arbitrary $\epsilon > 0$ and consider $\langle q_t \rangle \rightharpoonup \pi$ and $C^*(\pi) \geq$

$\mathbb{E}[\sum C(\pi_{2t}(q_{2t+1}|q_2t))] - \epsilon$. The previous analysis implies (39): $\mathbb{E}\left[\sum \|q_{2t+2} - q_{2t+1}\|^2\right] \leq \frac{\epsilon}{m}$. Now take any path of $\langle q_t \rangle$, denoted by $q_t[\omega]$, such that $q_{t_0}[\omega]$ first leaves $B_{\delta_1}(q_0)$ at period t (here we choose $\delta_1 > \delta_0$). In other words, $\forall t < t_0$, $q_t[\omega] \in B_{\delta_1}(q_0)$ and $q_{t_0}[\omega] \notin B_{\delta_1}(q_0)$. Collect all paths $q_t[\omega']$ s.t. $q_t[\omega'] = q_t[\omega]$ when $t \leq t_0$. Let Ω_0 denote the set of events corresponding to theses paths.

Now we construct a process $\langle \widehat{q_t} \rangle$ in $\mathbb{R}^{|X|}$, satisfying $\sum_x \widehat{q_t}(x) \equiv 1$. The process is defined on event space Ω_0 (with corresponding sigma algebra $\mathcal{F}_0$ and probability measure P_0 restricted to Ω_0). For notational simplicity, I label $\langle \widehat{q_t} \rangle$ using t from t_0 to $2T$. If t_0 is even, let $T_0 = T + \frac{t_0}{2}$:

$$\begin{cases} \widehat{q}_{t_0+s+1}[\omega] - \widehat{q}_{t_0+s}[\omega] = q_{t_0+2s+1}[\omega] - q_{t_0+2s}[\omega] \\ \widehat{q}_{T_0+s+1}[\omega] - \widehat{q}_{T_0+s}[\omega] = q_{t_0+2s+2}[\omega] - q_{t_0+2s+1}[\omega] \end{cases}$$

where s is from 0 to $T - \frac{t_0}{2} - 1$. If t_0 is odd, let $T_0 = T + \frac{t_0-1}{2}$:

$$\begin{cases} \widehat{q}_{t_0+s+1}[\omega] - \widehat{q}_{t_0+s}[\omega] = q_{t_0+2s}[\omega] - q_{t_0+2s-1}[\omega] \\ \widehat{q}_{T_0+s+1}[\omega] - \widehat{q}_{T_0+s}[\omega] = q_{t_0+2s+1}[\omega] - q_{t_0+2s}[\omega] \end{cases}$$

where s is from 0 to $T - \frac{t_0-1}{2}$. $\langle \widehat{q_t} \rangle$ essentially reorders the belief changes of $\langle q_t \rangle$ by grouping all even periods (acquisition periods) together and then all odd periods (disposal periods) together. Now we verify that $\langle \widehat{q_t} \rangle$ is a martingale up to period T_0. We only show the case with even t_0 and the case with odd t_0 follows:

$$\begin{aligned}
&\mathbb{E}[\widehat{q}_{t_0+s+1} - \widehat{q}_{t_0+s} | \widehat{q}_{t_0}, \cdots, \widehat{q}_{t_0+s}] \\
=&\int \widehat{q}_{t_0+s+1}[\omega] - \widehat{q}_{t_0+s}[\omega] \mathrm{d}P_0(\omega | (\widehat{q}_{t_0}, \ldots, \widehat{q}_{t_0+s})[\omega] = \widehat{q}_{t_0}, \cdots, \widehat{q}_{t_0+s}) \\
=&\int (q_{t_0+2s+1}[\omega] - q_{t_0+2s}[\omega]) \mathrm{d}P_0(\omega | (\widehat{q}_{t_0}, \ldots, \widehat{q}_{t_0+s})[\omega] = \widehat{q}_{t_0}, \cdots, \widehat{q}_{t_0+s}) \\
=&\int (q_{t_0+2s+1}[\omega] - q_{t_0+2s}[\omega]) \mathrm{d}P_0(\omega | q_{t_0+2s'+1}[\omega] - q_{t_0+2s'}[\omega] = \widehat{q}_{t_0+s'+1} - \widehat{q}_{t_0+s'}) \\
=&\int \left(\int q_{t_0+2s+1}[\omega] - q_{t_0+2s}[\omega] \mathrm{d}P_0(\omega) \Big| (q_{t_0}, \cdots, q_{t_0+2s})[\omega] = q_{t_0}, \cdots, q_{t_0+2s} \right) \\
&\mathrm{d}P_0((q_{t_0}, \cdots, q_{t_0+2s})[\omega] = q_{t_0}, \cdots, q_{t_0+2s} \Big| q_{t_0+2s'+1}[\omega] - q_{t_0+2s'}[\omega] = \widehat{q}_{t_0+s'+1} - \widehat{q}_{t_0+s'}) \\
=&0
\end{aligned}$$

The last equality is by the Markov property of $\langle q_t \rangle$ and martingale property of $\langle q_t \rangle$ at even t's. Therefore, $\langle \widehat{q_t} \rangle_{t=t_0,\ldots,T_0}$ is a martingale process.

Since the previous analysis is done in the event space Ω_0 where q_t first crosses $B_{\delta_1}(q_0)$ at q_{t_0}. Ω can actually been partitioned into Ω_0^α's plus Ω_1, where each α indexes a path first crossing $B_{\delta_1}(q_0)$,[82] and Ω_1 contains events when the path never crosses $B_{\delta_1}(q_0)$. Let $t_0(\alpha)$ denote the first crossing time of each α.

Now we calculate the total amount of information discarded in event space Ω_0. Again we only show the case for t_0 even.

$$\mathbb{E}\left[\sum_{t=0}^{T} \|q_{2t+2} - q_{2t+1}\|^2\right] = \mathbb{E}\left[\left\|\sum_{t=0}^{T} (q_{2t+2} - q_{2t+1})\right\|^2\right]$$

[82] Ω_0^α's are clearly disjoint since a path can only first cross $B_{\delta_1}(q_0)$ once.

$$
\begin{aligned}
\geq & \mathbb{E}\left[\left\|\sum_{t=0}^{T}(q_{2t+2}-q_{2t+1})\right\|^2 \middle| \cup\Omega_0^\alpha\right]\cdot P(\omega\in\cup\Omega_0^\alpha)\\
= & \mathbb{E}\left[\left\|\widehat{q}_{2T}-\widehat{q}_{T_0}\right\|^2 \middle| \cup\Omega_0^\alpha\right]\cdot P(\omega\in\cup\Omega_0^\alpha)\\
\geq & \mathbb{E}\left[\mathbb{E}\left[\frac{1}{3}\left(\left\|\widehat{q}_{T_0}-\widehat{q}_{t_0}\right\|^2+\left\|\widehat{q}_{t_0}-q_0\right\|^2+\|\widehat{q}_{2T}-q_0\|^2\right)\middle|\Omega_0^\alpha\right]\middle|\cup\Omega_0^\alpha\right]\cdot P(\omega\in\cup\Omega_0^\alpha)\\
\geq & \left(\frac{1}{3}\mathbb{E}\left[\mathbb{E}\left[\left\|\widehat{q}_{T_0}-\widehat{q}_{t_0}\right\|^2\middle|\Omega_0^\alpha\right]\middle|\cup\Omega_0^\alpha\right]+\frac{1}{3}\delta_1^2+\frac{1}{3}\delta_0^2\right)\cdot P\left(\omega\in\cup\Omega_0^\alpha\right)\\
= & \left(\frac{1}{3}\mathbb{E}\left[\mathbb{E}\left[\sum_{2t\geq t_0(\alpha)}\|q_{2t+1}-q_{2t}\|^2\middle|\Omega_0^\alpha\right]\middle|\cup\Omega_0^\alpha\right]+\frac{1}{3}\delta_1^2+\frac{1}{3}\delta_0^2\right)\cdot P\left(\omega\in\cup\Omega_0^\alpha\right)
\end{aligned}
$$

The first equality is by $\mathbb{E}[q_{2t+1}|q_{2t+2}]=q_{2t+2}$ from Definition 22. The first inequality is from the non-negativity of norm. The second equality is by definition of $\langle\widehat{q}_t\rangle$. The second inequality is from Cauchy-Schwarz inequality. The last inequality is from $q_{t_0}\notin B_{\delta_1}(q_0)$ and definition of δ_0. The last equality is by definition of $\langle\widehat{q}_t\rangle$. Combining the result with (39), we obtain:

$$
\left(\mathbb{E}\left[\left\|\sum_{2t\geq t_0(\alpha)}(q_{2t+1}-q_{2t})\right\|^2\middle|\cup\Omega_0^\alpha\right]+\delta_1^2+\delta_0^2\right)\cdot P(\omega\in\cup\Omega_0^\alpha)\leq\frac{3\epsilon}{m} \tag{40}
$$

Now we state the main proof for Theorem 3. $\forall q_0\in\Delta_\circ$, $\forall\eta$, let δ be the parameter pinned down in Definition 11. Pick $\delta_0<\delta_1<\delta$. Consider arbitrary $\pi\in\Pi_b$ s.t. $\mathrm{supp}(\pi)\subset B_{\delta_0}(q_0)$ and $\mathbb{E}_\pi[\nu]=q_0$. Since $k(q)$ is continuous, let ξ be $\sup\|k(q)-k(q_0)\|$ when $q\in B_{\delta_1}(q_0)$. Now we show that $C^*(\pi)$ is differentiable at q_0 and:

$$
\left|C^*(\pi)-\int(\nu-q_0)^T k(q_0)(\nu-q_0)\mathrm{d}\pi(\nu)\right|\leq\eta\int\|\nu-q_0\|^2\mathrm{d}\pi(\nu) \tag{41}
$$

where $k(q)$ locally characterize $C(\pi)$. Consider any $\langle q_t\rangle\rightharpoonup\pi$ and $\mathbb{E}[\sum C(q_{2t+1}|q_{2t})]\leq C^*(\pi)+\epsilon$ (ϵ is for now a free parameter that we will pin down later in the proof). The total cost of $\langle q_t\rangle$ can be written as:

$$
\begin{aligned}
\mathbb{E}[\sum_t C(q_{2t+1}|q_{2t})] = & \mathbb{E}\left[\sum_t C(q_{2t+1}|q_{2t})\mathbf{1}_{\forall t'\leq 2t, q_{t'}[\omega]\in B_{\delta_1}(q_0)}\right] && \text{(I)}\\
& +\mathbb{E}\left[\sum_t C(q_{2t+1}|q_{2t})\mathbf{1}_{\forall t'<2t, q_{t'}[\omega]\in B_{\delta_1}(q_0)\&q_{2t}\notin B_{\delta_1}(q_0)}\right] && \text{(II)}\\
& +\mathbb{E}\left[\sum_t C(q_{2t+1}|q_{2t})\mathbf{1}_{\exists t'<2t, q_{t'}[\omega]\notin B_{\delta_1}(q_0)}\right] && \text{(III)}
\end{aligned}
$$

In words, we partition paths of $\langle q_t\rangle$ into three groups. (I) includes paths that never leaves $B_{\delta_1}(q_0)$ until $2t$ for each t. (II) includes paths that first leaves $B_{\delta_2}(q_0)$ at $2t$ for each t. (III) includes paths that have left $B_{\delta_1}(q_0)$ before $2t$ for each t. We bound the total cost of the three groups separately.

We first study (I):

$$
(I)=\mathbb{E}\left[\sum_t C(q_{2t+1}|q_{2t})\mathbf{1}_{\forall t'\leq 2t, q_{t'}[\omega]\in B_{\delta_1}(q_0)\&\mathrm{supp}(q_{2t+1}|q_{2t})\in B_{\delta_1}(q_0)}\right] \tag{I-a}
$$

$$+\mathbb{E}\left[\sum_t C(q_{2t+1}|q_{2t})\mathbf{1}_{\forall t'\le 2t, q_{t'}[\omega]\in B_{\delta_1}(q_0)\&\mathrm{supp}(q_{2t+1}|q_{2t})\notin B_{\delta_1}(q_0)}\right] \tag{I-b}$$

We further partition group (I) into two sub-groups. In group (I-a), the paths never leaves $B_{\delta_1}(q_0)$ until $2t+1$ and in group (I-b) the paths first leaves $B_{\delta_1}(q_0)$ in period $2t+1$.

$$\begin{aligned}
(I-a) =&\mathbb{E}\Bigg[\sum_t\Bigg(\int (q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t}) + (q_{2t+1}-q_{2t})^T(k(q_{2t})-k(q_0))(q_{2t+1}-q_{2t})\\
&+ C(q_{2t+1}|q_{2t}) - (q_{2t+1}-q_{2t})^T k(q_{2t})(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\Bigg)\\
&\times \mathbf{1}_{\forall t'\le 2t, q_{t'}[\omega]\in B_{\delta_1}(q_0)\&\mathrm{supp}(q_{2t+1}|q_{2t})\in B_{\delta_1}(q_0)}\Bigg]\\
\ge&\mathbb{E}\Bigg[\sum_t\Bigg(\int (q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t}) - (\eta+\xi)\int \|q_{2t+1}-q_{2t}\|^2\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\Bigg)\\
&\times \mathbf{1}_{\forall t'\le 2t, q_{t'}[\omega]\in B_{\delta_1}(q_0)\&\mathrm{supp}(q_{2t+1}|q_{2t})\in B_{\delta_1}(q_0)}\Bigg]
\end{aligned}$$

The inequality is implied by the differentiability condition of C and definition of continuity parameter ξ.

To calculate (I-b), we modify the distribution of q_{2t+1} for any event that satisfies the restriction, namely $q_{2t}\in B_{\delta_1}(q_0)$ but there exists $q_{2t+1}\notin B_{\delta_1}(q_0)$. For any such q_{2t}, let π_{2t} still be the distribution of q_{2t+1}. Now let $q'=\mathbb{E}_{\pi_{2t}}[\nu|\nu\notin B_{\delta_1}(q_0)]$. Let $\pi''(\nu|q')=\pi_{2t}(\nu|\nu\notin B_{\delta_1}(q_0))$ and $\pi''(\nu|q)=\delta_\nu$ otherwise. Let $\pi'(\nu)=\mathbf{1}_{\nu\in B_{\delta_1}(q_0)}\pi_{2t}(\nu)+\pi_{2t}(\Delta(X)\setminus B_{\delta_1}(q_0))\delta_{q'}$. It is easy to verify that $\pi_{2t}(\nu)=\mathbb{E}_{\pi'}[\pi''(\nu|q)]$. Suppose $q'\notin B_\delta(q_0)$, let q'' be a linear combination of q_{2t} and q' that is on the boundary of $B_\delta(q_0)$. Then we construct $\widetilde{\pi}'$ by shifting q' to q''. Let $q_1'=\mathbb{E}_{\pi_{2t}}[\nu|\nu\in B_{\delta_1}(q_0)]$. Define:

$$\widetilde{\pi}'(\nu)=\mathbf{1}_{\nu\in B_{\delta_1}(q_0)}\pi_{2t}(\nu)\cdot\frac{\|q'-q_1'\|}{\|q'-q_{2t}\|}\cdot\frac{\|q''-q_{2t}\|}{\|q''-q_1'\|}+\frac{\|q_1'-q_{2t}\|}{\|q''-q_1'\|}\delta_{q''}(\nu)$$

By definition, $\mathbb{E}_{\widetilde{\pi}'}[\nu]=q_{2t}$ and $\widetilde{\pi}'\le_{BW}\pi'$. When $q'\in B_\delta(q_0)$, let $\widetilde{\pi}'=\pi'$. Now we calculate the cost of π_{2t}:

$$\begin{aligned}
C(\pi_{2t})\ge&C(\pi')\ge C(\widetilde{\pi}')\\
&\ge\int(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}\widetilde{\pi}'(\nu)-(\eta+\xi)\int\|\nu-q_{2t}\|^2\mathrm{d}\widetilde{\pi}'(\nu)\\
&\ge\int(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}\widetilde{\pi}'(\nu)-(\eta+\xi)\int\|\nu-q_{2t}\|^2\mathrm{d}\pi_{2t}(\nu)\\
&=\int(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}\pi_{2t}(\nu)+\int(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}(\widetilde{\pi}'-\pi_{2t})(\nu)\\
&\quad-(\eta+\xi)\int\|\nu-q_{2t}\|^2\mathrm{d}\pi_{2t}(\nu)\\
&=\int(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}\pi_{2t}(\nu)-(\eta+\xi)\int\|\nu-q_{2t}\|^2\mathrm{d}\pi_{2t}(\nu)\\
&\quad+\int_{\nu\in B_\delta(q_0)}(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}(\widetilde{\pi}'-\pi_{2t})(\nu)
\end{aligned}$$

$$
\begin{aligned}
&+\int_{\nu\notin B_\delta(q_0)}(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}(\widetilde{\pi}'-\pi_{2t})(\nu)\\
\geq&\int(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}\pi_{2t}(\nu)-(\eta+\xi)\int\|\nu-q_{2t}\|^2\mathrm{d}\pi_{2t}(\nu)\\
&-\left(1-\frac{\delta-\delta_1}{\delta+\delta_1}\right)(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}\pi_{2t}(\nu)\\
&-\pi_{2t}(\Delta(X)\setminus B_{\delta_1}(q_0))\left(\int(\nu-q_{2t})k(q_0)(\nu-q_{2t})\mathrm{d}\pi''(\nu|q')+(q'-q_{2t})^T k(q_0)(q'-q_{2t})\right)
\end{aligned}
$$

The first two inequalities are by $\pi_{2t}\geq_{BW}\pi'\geq_{BW}\widetilde{\pi}'$ and Axiom 1. The third inequality is by Definition 11. The forth inequality: $\frac{\|q'-q_1'\|}{\|q'-q_{2t}\|}<1$; $\frac{\|q''-q_{2t}\|}{\|q''-q_1'\|}\geq\frac{\delta-\delta_1}{\delta+\delta_1}$ bounds the second line, the third line is calculated by ignoring the weakly positive term provided by $\widetilde{\pi}'$. Finally, since $k(q_0)$ is fixed, there exists $M=\sup_{q,\nu\in\Delta(X)}(\nu-q)^T k(q_0)(\nu-q)<\infty$. To sum up:

$$
\begin{aligned}
C(\pi_{2t})\geq&\int(\nu-q_{2t})^T k(q_0)(\nu-q_{2t})\mathrm{d}\pi_{2t}(\nu)-\left(\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}\right)\int\|\nu-q_{2t}\|^2\mathrm{d}\pi_{2t}(\nu)\\
&-\pi_{2t}(\Delta(X)\setminus B_{\delta_1}(q_0))M
\end{aligned}
\tag{42}
$$

Plug (42) into (I-b), we get:

$$
\begin{aligned}
(I-b)\geq&\mathbb{E}\Bigg[\sum_t\Bigg(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\\
&-\left(\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}\right)\int\|q_{2t+1}-q_{2t}\|^2\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})+\pi_{2t}\left(\Delta(X)\setminus B_{\delta_1}(q_0)|q_{2t}\right)\cdot M\Bigg)\\
&\times\mathbf{1}_{\forall t'\leq 2t,q_{t'}[\omega]\in B_{\delta_1}(q_0)\&\mathrm{supp}(q_{2t+1}|q_{2t})\notin B_{\delta_1}(q_0)}\Bigg]\\
=&\mathbb{E}\Bigg[\sum_t\Bigg(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\\
&-\left(\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}\right)\int\|q_{2t+1}-q_{2t}\|^2\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\Bigg)\\
&\times\mathbf{1}_{\forall t'\leq 2t,q_{t'}[\omega]\in B_{\delta_1}(q_0)\&\mathrm{supp}(q_{2t+1}|q_{2t})\notin B_{\delta_1}(q_0)}\Bigg]\\
&-P(\cup\Omega_0^\alpha)\cdot M\\
\geq&\mathbb{E}\Bigg[\sum_t\Bigg(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\\
&-\left(\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}\right)\int\|q_{2t+1}-q_{2t}\|^2\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\Bigg)\\
&\times\mathbf{1}_{\forall t'\leq 2t,q_{t'}[\omega]\in B_{\delta_1}(q_0)\&\mathrm{supp}(q_{2t+1}|q_{2t})\notin B_{\delta_1}(q_0)}\Bigg]\Bigg]\\
&-\frac{3\epsilon}{(\delta_1^2+\delta_0^2)m}
\end{aligned}
$$

The equality is by rewriting the event space at which some path first crosses $B_{\delta_1}(q_0)$, The last inequality is implied by (42).

Now we study (II) and (III):

$$
\begin{aligned}
&(II)+(III)\\
\geq&\mathbb{E}\left[\sum_t\left(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\right)\mathbf{1}_{\exists t'\leq 2t, q_{t'}[\omega]\notin B_{\delta_1}(q_0)}\right]\\
&-\mathbb{E}\left[\sum_t\left(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\right)\mathbf{1}_{\exists t'\leq 2t, q_{t'}[\omega]\notin B_{\delta_1}(q_0)}\right]\\
\geq&\mathbb{E}\left[\sum_t\left(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\right)\mathbf{1}_{\exists t'\leq 2t, q_{t'}[\omega]\notin B_{\delta_1}(q_0)}\right]\\
&-\mathbb{E}\left[\sum_t\left(\sum_{s\geq t}\int(q_{2s+1}-q_{2t})^T k(q_0)(q_{2s+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2s+1}|q_{2s})\right)\mathbf{1}_{\forall t'<2t, q_{t'}[\omega]\in B_{\delta_1}(q_0)\&q_{2t}[\omega]\notin B_{\delta_1}(q_0)}\right]\\
=&\mathbb{E}\left[\sum_t\left(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\right)\mathbf{1}_{\exists t'\leq 2t, q_{t'}[\omega]\notin B_{\delta_1}(q_0)}\right]\\
&-\mathbb{E}\left[\sum_{2s\geq t_0(\alpha)}\int(q_{2s+1}-q_{2t})^T k(q_0)(q_{2s+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2s+1}|q_{2s})\middle|\cup\Omega_0^\alpha\right]P(\cup\Omega_0^\alpha)\\
\geq&\mathbb{E}\left[\sum_t\left(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\right)\mathbf{1}_{\exists t'\leq 2t, q_{t'}[\omega]\notin B_{\delta_1}(q_0)}\right]\\
&-\|k(q_0)\|\mathbb{E}\left[\sum_{2s\geq t_0(\alpha)}\int\|q_{2s+1}-q_{2s}\|^2\mathrm{d}\pi_{2t}(q_{2s+1}|q_{2s})\middle|\cup\Omega_0^\alpha\right]P(\cup\Omega_0^\alpha)\\
\geq&\mathbb{E}\left[\sum_t\left(\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\right)\mathbf{1}_{\exists t'\leq 2t, q_{t'}[\omega]\notin B_{\delta_1}(q_0)}\right]\\
&-\|k(q_0)\|\frac{3\epsilon}{m}
\end{aligned}
$$

The first inequality is by definition of M. The second equality is by definition any path which ever crosses $B_{\delta_1}(q_0)$ must have first crossed it at some history. The equality is rewriting the second term in the language of Ω_0^α and $t_0(\alpha)$. The last inequality is implied by (40).

Now combine (I), (II), and (III) together and we get:

$$
\begin{aligned}
\mathbb{E}[\sum_t C(q_{2t+1}|q_{2t})]\geq&\mathbb{E}\left[\sum_t\int(q_{2t+1}-q_{2t})^T k(q_0)(q_{2t+1}-q_{2t})\mathrm{d}q_{2t}(q_{2t+1}|q_{2t})\right]\\
&-\left(\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}\right)\mathbb{E}\left[\sum_t\int\|q_{2t+1}-q_{2t}\|^2\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\right]\\
&-\left(\frac{1}{\delta_1^2+\delta_0^2}+\|k(q_0)\|\right)\frac{3\epsilon}{m}\\
\geq&\int(\nu-q_0)^T k(q_0)(\nu-q_0)\mathrm{d}\pi(\nu)\\
&-\left(\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}\right)\mathbb{E}\left[\sum_t\int\|q_{2t+1}-q_{2t}\|^2\mathrm{d}\pi_{2t}(q_{2t+1}|q_{2t})\right]
\end{aligned}
$$

$$
\begin{aligned}
&-\left(\frac{1}{\delta_1^2+\delta_0^2}+\|k(q_0)\|\right)\frac{3\epsilon}{m}\\
\geq&\int(\nu-q_0)^T k(q_0)(\nu-q_0)\mathrm{d}\pi(\nu)\\
\geq&\int(\nu-q_0)^T k(q_0)(\nu-q_0)\mathrm{d}\pi(\nu)\\
&-\left(\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}\right)\left(\int\|\nu-q_0\|^2\mathrm{d}\pi(\nu)+\frac{\epsilon}{m}\right)\\
&-\left(\frac{1}{\delta_1^2+\delta_0^2}+\|k(q_0)\|\right)\frac{3\epsilon}{m}
\end{aligned}
$$

Fix all other parameters and let $\epsilon\to 0$ (ϵ is the approximation error defined in (19)), then this implies:

$$
C^*(\pi)\geq\int(\nu-q_0)^T k(q_0)(\nu-q_0)\mathrm{d}\pi(\nu)-\left(\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}\right)\left(\int\|\nu-q_0\|^2\mathrm{d}\pi(\nu)\right)
$$

Therefore, $\forall\epsilon>0$, let δ be its corresponding parameter define in Definition 11, we can pick δ_0,δ_1 small enough such that $\eta+\xi+\frac{2\delta_1}{\delta+\delta_1}<\epsilon$.[83] Hence we proved that we find δ_0 s.t. $\forall\pi$ s.t. $\mathbb{E}_\pi[\nu]=q_0$ and $\mathrm{supp}(\pi)\subset B_{\delta_0}(q_0)$:

$$
C^*(\pi)\geq\int(\nu-q_0)^T k(q_0)(\nu-q_0)\mathrm{d}\pi(\nu)-\epsilon\left(\int\|\nu-q_0\|^2\mathrm{d}\pi(\nu)\right)
$$

On the other hand:

$$
C^*(\pi)\leq C(\pi)\leq\int(\nu-q_0)^T k(q_0)(\nu-q_0)\mathrm{d}\pi(\nu)+\epsilon\left(\int\|\nu-q_0\|^2\mathrm{d}\pi(\nu)\right)
$$

Therefore,

$$
\left|C^*(\pi)-\int(\nu-q_0)^T k(q_0)(\nu-q_0)\mathrm{d}\pi(\nu)\right|\leq\epsilon\left(\int\|\nu-q_0\|^2\mathrm{d}\pi(\nu)\right)
$$

We verified the twice differentiability of $C^*(\pi)$ and (41). By definition, if $C^*(\pi)=\mathbb{E}_\pi[F(\nu)]-F(\mathbb{E}_\pi[\nu])$, then $2k(q)\equiv\mathcal{H}F(q)$ is the Hessian matrix of $F(q)$. In other words, $k(q)$ also locally characterizes $C^*(q)$. Therefore $C(\pi)\geq C^*(\pi)=\mathbb{E}_\pi[F(\nu)]-F(\mathbb{E}_\pi[\nu])$, which completes the proof.

G Proof of Lemma 6

We begin with a preliminary lemma.

Lemma 24. *Let C_1 and $\hat{C}$ be Posterior Separable cost functions with divergences D_1 and D_2, respectively. The following are equivalent:*

(i) $C_1(\sigma\mid p)\geq C_2(\sigma\mid p)$ *for all* $\sigma\in\mathcal{E}_b$ *and* $p\in\Delta_\circ$.

(ii) $D_1(q\mid p)\geq D_2(q\mid p)$ *for all* $q,p\in\Delta_\circ$.

[83] Here we recycled symbol ϵ, now it is used to denote the parameter defining the differentiability condition of C^*.

Proof of Lemma 24. That (ii) implies (i) is immediate. We prove that (i) implies (ii) by contraposition. Suppose that there exist $p, \nu \in \Delta_\circ$ such that $D_2(\nu \mid p) > D_1(\nu \mid p)$. For each $\epsilon > 0$, define $r(\epsilon) := p + \epsilon(p - \nu)$. Henceforth, we take $\epsilon > 0$ to be sufficiently small that $r(\epsilon) \in \Delta_\circ$. Define the posterior distribution $\pi(\epsilon) \in \Pi(p)$ by

$$\pi(\epsilon) := \frac{1}{1+\epsilon}\delta_{r(\epsilon)} + \frac{\epsilon}{1+\epsilon}\delta_\nu.$$

We then have

$$\mathbb{E}_{\pi(\epsilon)}[D_1(q \mid p) - D_2(q \mid p)] = \frac{1}{1+\epsilon}[D_1(r(\epsilon) \mid p) - D_2(r(\epsilon) \mid p)] + \frac{\epsilon}{1+\epsilon}[D_1(\nu \mid p) - D_F(\nu \mid p)].$$

Dividing through by $\epsilon > 0$ and sending $\epsilon \to 0$ yields

$$\begin{aligned}\lim_{\epsilon \to 0} \frac{\mathbb{E}_{\pi(\epsilon)}[D_1(q \mid p) - D_2(q \mid p)]}{\epsilon} &= \frac{\partial}{\partial \epsilon} D_1(r(\epsilon) \mid p)\,|_{\epsilon=0} - \frac{\partial}{\partial \epsilon} D_2(r(\epsilon) \mid p)\,|_{\epsilon=0} + D_1(\nu \mid p) - D_2(\nu \mid p) \\ &= D_1(\nu \mid p) - D_2(\nu \mid p) \\ &< 0,\end{aligned}$$

where the first equality follows from the definition of directional derivative, the second equality follows from the fact that $D_1(\cdot \mid p)$ and $D_2(\cdot \mid p)$ are minimized at $q = p$ and the inequality follows from the supposition that $D_2(\nu \mid p) > D_1(\nu \mid p)$. Consequently, for $\epsilon' > 0$ sufficiently small, we have that $\mathbb{E}_{\pi(\epsilon')}[D_1(q \mid p) - D_2(q \mid p)] < 0$. Let σ' be such that $\pi_{\langle \sigma' | p \rangle} = \pi(\epsilon')$. Then $C_1(\sigma' \mid p) < C_2(\sigma' \mid p)$, as desired. □

We may now prove Lemma 6 itself.

Proof of Lemma 6. We first show that (i) $\implies$ (ii) by contraposition. Suppose that PGL does not hold. Then by Lemma 24 below, there exist $p \in \Delta_\circ$ and $\sigma \in \mathcal{E}_b$ such that $G := \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[D_F(q \mid p) - D(q \mid p)] > 0$. Because C is Locally Linear, there exists an $\alpha \in (0,1)$ sufficiently small that $C_{RA}(\alpha \cdot \sigma \mid p)/\alpha - \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[D(q \mid p)] < G$. It follows that

$$\mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[D_F(q \mid p)] > \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[D(q \mid p)] + G > \frac{C_{RA}(\alpha \cdot \sigma \mid p)}{\alpha}. \tag{43}$$

Because $\mathbb{E}_{\pi_{\langle \alpha \cdot \sigma | p \rangle}}[D_F(q \mid p)] = \alpha \cdot \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[D_F(q \mid p)]$, rearranging (43) yields

$$C_{RA}(\alpha \cdot \sigma \mid p) < \mathbb{E}_{\pi_{\langle \alpha \cdot \sigma | p \rangle}}[D_F(q \mid p)],$$

which implies that there exists some $\tau \in \mathcal{E}_b$ for which $C(\tau \mid p) < \mathbb{E}_{\pi_{\langle \tau | p \rangle}}[D_F(q \mid p)]$. This witnesses that Preference for Incremental Learning does not hold.

Next, we show that (ii) $\implies$ (i). Let $p \in \Delta_\circ$ and $\sigma \in \mathcal{E}_b$ be given. Recall that the map $\alpha \mapsto C_{RA}(\alpha \cdot \sigma \mid p)/\alpha$ is non-increasing on $(0,1]$ because C_{RA} is Randomizaton Averse by definition. Therefore,

$$C(\sigma \mid p) \geq C_{RA}(\sigma \mid p) \geq \lim_{\alpha \to 0} \frac{C(\alpha \cdot \sigma \mid p)}{\alpha} = \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[D(q \mid p)] \geq \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[D_F(q \mid p)],$$

where the first equality is from the definition of C being Locally Linear and the second equality is immediate from PGL. □

H Proof of Proposition 3

We proceed via a series of lemmas.

Lemma 25. *If the Direct Cost function C be Prior-Invariant and Locally Quadratic with kernel k, then the normalized kernel $\overline{k} : \Delta_\circ \to \mathbb{R}^{\Theta\times\Theta}$ is a constant matrix (denoted simply by $\overline{k} \in \mathbb{R}^{\Theta\times\Theta}$).*

Proof. See Appendix K. □

Lemma 26. *Let the Direct Cost function C be Prior-Invariant, Locally Quadratic, and nonzero. If the Indirect Cost function $C^* = \Phi C$ is UPS, then $|\Theta| \leq 2$.*

Proof. Let $\Theta = \{1,\dots,n\}$ and parametrize $p \in \Delta_\circ$ by its first $n-1$ elements, so that $p = (p_1, p_2, \cdots, p_{n-1}, 1 - \sum_{i=1}^{n-1} p_i)$. Let $k(p)$ be the kernel of C and define its $(n-1)\times(n-1)$ dimensional projection $\tilde{k}(p_1,\dots,p_{n-1})$ as

$$\begin{aligned}\tilde{k}(p_1,\cdots,p_{n-1}) &:= \begin{bmatrix} I_{n-1} & -\mathbf{1}\end{bmatrix} \cdot k(p) \cdot \begin{bmatrix} I_{n-1} \\ -\mathbf{1}^T\end{bmatrix} \\ &= \left[\frac{\overline{k}_{ij}}{p_i p_j} - \frac{\overline{k}_{in}}{p_i(1-\Sigma_{\ell=1}^{n-1} p_\ell)} - \frac{\overline{k}_{jn}}{p_j(1-\Sigma_{\ell=1}^{n-1} p_\ell)} + \frac{\overline{k}_{nn}}{(1-\Sigma_{\ell=1}^{n-1} p_\ell)^2}\right]_{ij}\end{aligned}$$

where $I_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ is the identity matrix, $\mathbf{1} \in \mathbb{R}^{n-1}$ is a vector consisting of all 1's, and the second line in the display represents $\tilde{k}$ by its ij^{th} element. The normalized kernel $\overline{k}$ is constant by the hypotheses of the lemma and Lemma 25.

Let F be a potential function in the UPS representation of C^*, which satisfies $F \in \mathbf{C}^2(\Delta_\circ)$ by Theorem 3. Let $\tilde{F}(p_1,\dots,p_{n-1}) := F(p)$. Again by Theorem 3, we have $\tilde{k} \equiv \frac{1}{2}\mathcal{H}\tilde{F}$. Because each ij^{th} element of $\tilde{k}(p_1,\dots,p_{n-1})$ is $\mathbf{C}^\infty$ smooth, it follows that $\tilde{F}$ is also $\mathbf{C}^\infty$ smooth.

Suppose, towards contradiction, that $n \geq 3$. Then we may calculate the higher order cross-partial derivatives of $\tilde{F}$ as follows. For any $i \neq j$, by symmetry of cross-partials for smooth functions, we have:

$$\begin{aligned}&\frac{\partial^3}{\partial p_i^2 \partial p_j}\tilde{F}(p_1,\cdots,p_{n-1}) = \frac{\partial}{\partial p_i}\tilde{k}(p_1,\cdots,p_{n-1})_{ij} = \frac{\partial}{\partial p_j}\tilde{k}(p_1,\cdots,p_{n-1})_{ii} \\ \iff &\frac{\partial}{\partial p_i}\left(\frac{\overline{k}_{ij}}{p_i p_j} - \frac{\overline{k}_{in}}{p_i(1-\sum_{\ell=1}^{n-1} p_\ell)} - \frac{\overline{k}_{jn}}{p_j(1-\sum_{\ell=1}^{n-1} p_\ell)} + \frac{\overline{k}_{nn}}{(1-\Sigma p_k)^2}\right) = \frac{\partial}{\partial p_j}\left(\frac{\overline{k}_{ii}}{p_i^2} - 2\frac{\overline{k}_{in}}{p_i(1-\sum_{\ell=1}^{n-1} p_\ell)} + \frac{\overline{k}_{nn}}{(1-\sum_{\ell=1}^{n-1} p_\ell)^2}\right) \\ \iff &-\frac{\overline{k}_{ij}}{p_i^2 p_j} - \frac{\overline{k}_{jn}}{(1-\sum_{\ell=1}^{n-1} p_\ell)^2 p_j} + \frac{\overline{k}_{in}}{p_i^2(1-\sum_{\ell=1}^{n-1} p_\ell)} = -\frac{\overline{k}_{in}}{p_i(1-\sum_{\ell=1}^{n-1} p_\ell)^2} \\ \iff &\overline{k}_{ij}p_n^2 + \overline{k}_{jn}p_i^2 - \overline{k}_{in}(p_i+p_n)p_j = 0\end{aligned}$$

By varying p over $\Delta_\circ$, it is easy to see that this final equality holds only when $\overline{k}_{ij} = \overline{k}_{in} = \overline{k}_{jn} = 0$. This implies $\overline{k}$ is a diagonal matrix, which is not permitted by the condition $\overline{k} \cdot \mathbf{1} = \mathbf{0}$ (which must hold without loss of generality by Lemma 11). Therefore there does not exist a non-trivial normalized kernel $\overline{k}$ such that the corresponding kernel $k(p)$ is a Hessian matrix, which is the desired contradiction. □

Lemma 27. *Let the Direct Cost function C be Prior-Invariant, Locally Quadratic, and nonzero. If $|\Theta| = 2$, the Indirect Cost function $C^* = \Phi C$ is UPS, then it is a Total Information cost function with symmetric coefficients, as described in* (Wald).

Proof. Let the state space be $\Theta = \{1, 2\}$. Let $\overline{k}$ denote the normalized kernel of C, which is constant by Lemma 25. Because we may without loss of generality let $\overline{k}$ be PSD and satisfy $\overline{k}\mathbf{1} = \mathbf{0}$ by Lemma 11, it follows that there exists some $\overline{\alpha} > 0$ such that

$$\overline{k} = \begin{bmatrix} \overline{\alpha} & -\overline{\alpha} \\ -\overline{\alpha} & \overline{\alpha} \end{bmatrix}$$
$$\implies k(p_1, p_2) = \begin{bmatrix} \frac{\overline{\alpha}}{p_1^2} & -\frac{\overline{\alpha}}{p_1 p_2} \\ -\frac{\overline{\alpha}}{p_1 p_2} & \frac{\overline{\alpha}}{p_2^2} \end{bmatrix}$$

Parametrizing $p = (p_1, 1 - p_1)$ by $p_1 \in (0, 1)$, we may define the scalar projection $\tilde{k}(p_1)$ of $k(p)$ by

$$\begin{aligned} \widetilde{k}(p_1) &= [1 \ \ -1] \begin{bmatrix} \frac{\overline{\alpha}}{p_1^2} & \frac{-\overline{\alpha}}{p_1(1-p_1)} \\ \frac{-\overline{\alpha}}{p_1(1-p_1)} & \frac{\overline{\alpha}}{(1-p_1)^2} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= \frac{\overline{\alpha}}{p_1^2(1-p_1)^2} \end{aligned}$$

It is then easy to verify that $\widetilde{k}(p_1) = \frac{1}{2}F''(p_1)$ for the convex function $F : (0, 1) \to \mathbb{R}$ defined by

$$F(p_1) := 2\overline{\alpha} \cdot \left(p_1 \log\left(\frac{p_1}{1-p_1}\right) + (1-p_1)\log\left(\frac{1-p_1}{p_1}\right)\right)$$

This is precisely the potential function in the UPS representation of the cost function (Wald) with $\alpha := 2\overline{\alpha}$. □

Lemma 28. *The cost function $\underline{C}(\sigma) := \alpha \max\left\{D_{KL}(\sigma_{\theta_1} \mid \sigma_{\theta_2}), D_{KL}(\sigma_{\theta_2} \mid \sigma_{\theta_1})\right\}$ is Locally Quadratic and has the same kernel as the Wald cost function.*

Proof. We may define Prior-Invariant cost functions $C_1(\sigma) := D_{KL}(\sigma_{\theta_1} \mid \sigma_{\theta_2})$ and $C_2(\sigma) := D_{KL}(\sigma_{\theta_2} \mid \sigma_{\theta_1})$. Note that C_1 is the LLR cost function with $\beta_{12} = \alpha$ and $\beta_{21} = 0$. Symmetrically, C_2 is the LLR cost function with $\beta_{12} = 0$ and $\beta_{21} = \alpha$. Thus, C_1 and C_2 are both Posterior Separable with respective prior-dependent potential functions $F_1, F_2 : (0, 1) \to \mathbb{R}$ defined by

$$F_1(q_1 \mid p_1) := \alpha \frac{q_1}{p_1} \log\left(\frac{q_1}{1-q_1}\right)$$
$$F_2(q_1 \mid p_1) := \alpha \frac{1-q_1}{1-p_1} \log\left(\frac{1-q_1}{q_1}\right)$$

By direct calculation, we see that

$$\frac{\partial^2}{\partial q_1^2} F_1(q_1 \mid p_1)\,|_{q_1=p_1} = \frac{\partial^2}{\partial q_1^2} F_2(q_1 \mid p_1)\,|_{q_1=p_1} = \frac{\alpha}{p_1^2(1-p_1)^2}.$$

Thus, Lemma 3(i) implies that C_1 and C_2 are both Locally Quadratic and, without loss of generality, have the same kernel k, which also coincides with the kernel of the Wald cost function. It is then easy to verify that k is also a valid kernel of $\underline{C}$, which proves the lemma. □

I Proof of Theorem 4

The main proof invokes Pomatto et al.'s (2019) characterization of the LLR cost function. That paper's result pertains to cost functions that (i) are defined on a slightly larger domain of experiments than $\mathcal{E}_b$ and (ii) satisfy a continuity condition that, by itself, is neither weaker nor stronger than what we have assumed (in Subsection 2.1). Below, we first present definitions from Pomatto et al. (2019) and then introduce a suitable notion of extensibility for cost function that allows us to apply that paper's result. We emphasize that these definitions are purely technical and, consistent with a conjecture in Pomatto et al. (2019, Appendix A), we believe that these additional conditions could be dispensed with.

Let $\mathcal{E}_\circ \subset \mathcal{E}$ denote the collection of experiments σ satisfying the following two properties:

(i) The measures $(\sigma_\theta)_{\theta\in\Theta}$ are mutually absolutely continuous.

(ii) The log-likelihood ratios $\boldsymbol{L}(s) := (L_{\theta,\theta'}(s))_{\theta\neq\theta'\in\Theta}$, where $_{\theta,\theta'}(s) := \log\left(\frac{d\sigma_\theta}{d\sigma_\theta}(s)\right)$, have finite moments: For every $\theta \in \Theta$ and $\boldsymbol{\alpha} \in \mathbb{N}^\Theta$,

$$M^\sigma_\theta(\boldsymbol{\alpha}) := \int_S \prod_{\theta'\neq\theta} \left|L_{\theta,\theta'}(s)^{\alpha'_\theta}\right| d\sigma(s \mid \theta) < \infty.$$

It is easy to see that $\mathcal{E}_b \subsetneq \mathcal{E}_\circ \subsetneq \mathcal{E}$. Define the measures $\Lambda^{(\sigma)}_\theta \in \Delta(\mathbb{R}^{|\Theta|(|\Theta|-1)})$ by $\Lambda^{(\sigma)}_\theta(B) := \sigma_\theta(\{s \in S : \boldsymbol{L}(s) \in B\})$. That is, $\Lambda^{(\sigma)}_\theta$ denotes the distribution of the vector of log-likelihood ratios generated by experiment σ conditional on state θ. For each $N \in \mathbb{N}$, define the pseudo-metric $d^{(N)}_{PST} : \mathcal{E}_\circ \times \mathcal{E}_\circ \to \mathbb{R}_+$ by

$$d^{(N)}_{PST}(\sigma,\tau) := \max_{\theta\in\Theta} d_{TV}\left(\Lambda^{(\sigma)}_\theta, \Lambda^{(\tau)}_\theta)\right) + \max_{\theta\in\Theta} \max_{\boldsymbol{\alpha}\in\{0,\ldots,N\}^\Theta} |M^\sigma_\theta(\boldsymbol{\alpha}) - M^\tau_\theta(\boldsymbol{\alpha})|$$

where $d_{TV}\left(\Lambda^{(\sigma)}_\theta, \Lambda^{(\tau)}_\theta)\right) := \sup_{B\in\mathcal{B}(\mathbb{R}^{|\Theta|(|\Theta|-1)})} |\Lambda^{(\sigma)}_\theta(B) - \Lambda^{(\tau)}_\theta(B)|$ is the total variation distance on $\Delta(\mathbb{R}^{|\Theta|(|\Theta|-1)})$. We say that a function $f : \mathcal{E}_\circ \to \mathbb{R}$ is *PST-continuous* if it is uniformly continuous with respect to $d^{(N)}_{PST}$ for some N.

The preceding definitions all appear in Pomatto et al. (2019), albeit with slightly different notation. The following definition allows us to suitably bridge the gap between cost functions defined on the domain $\mathcal{E}_b$ of bounded experiments and cost functions defined on the larger domain $\mathcal{E}_\circ$.

Definition 24 (Extensible). *Cost function $C : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ is **Extensible** if there exists a mapping $\overline{C} : \mathcal{E}_\circ \times \Delta_\circ \to \mathbb{R}_+$ such that:*

(i) The restriction $\overline{C}\,|_{\mathcal{E}_b\times\Delta_\circ} = C$.

(ii) For each $\sigma \in \mathcal{E}_\circ$ and $p \in \Delta_\circ$, if a sequence $\left\{\sigma^{(n)}\right\} \subset \mathcal{E}_b$ satisfies (a) $\sigma^{(n)} \gtrsim_B \sigma^{(n-1)}$ for all n and (b) $\pi_{\langle\sigma^{(n)}|p\rangle} \to^{w^*} \pi_{\langle\sigma|p\rangle}$ as $n \to \infty$, then $\lim_{n\to\infty} C(\sigma^{(n)} \mid p) = \overline{C}(\sigma \mid p)$.

(iii) For each $p \in \Delta_\circ$, $\overline{C}(\cdot \mid p)$ is PST-continuous.

In Definition 24, point (i) states that $\overline{C}(\cdot \mid p)$ indeed extends $C(\cdot \mid p)$ from $\mathcal{E}_b$ to $\mathcal{E}_\circ$. Given our standing assumption that $C(\cdot \mid p)$ assigns equal cost to Blackwell equivalent experiments, point (ii) implies that $\overline{C}(\cdot \mid p)$ also satisfies this property. Notably, this equal-cost property is also implied by point (iii).

More generally, point (ii) says that the cost of any (monotonic) sequence of bounded experiments whose posterior distributions converge weakly to that of $\sigma \in \mathcal{E}_\circ$ converges to the cost of σ. As shown below, this allows us to extend properties of C, such as Constant Marginal Cost, to the larger domain on which $\overline{C}$ is defined. When C is Blackwell monotone, so that $C(\sigma^{(n)} \mid p) \geq C(\sigma^{(n-1)} \mid p)$, it can be viewed as a mild lower semi-continuity requirement on the cost function.

Point (iii) is a technical condition required by Pomatto et al. (2019). Note that for sequences of (uniformly) bounded experiments, PST-continuity is much less demanding than the weak* continuity on posterior distributions that we have assumed. However, PST-continuity has implications for the cost function along sequences of experiments that are *not* uniformly bounded, for which our standing continuity assumption has no such implications.

We note that Mutual Information, Total Information, and the LLR cost functions are all Extensible. Indeed, we do not know of any cost functions used in applications that are *not* Extensible.

I.1 Preliminary Lemmas

Define the map $\boldsymbol{D}_{KL} : \mathcal{E}_b \to \mathbb{R}_+^{|\Theta|(|\Theta|-1)}$ by $\boldsymbol{D}_{KL}(\sigma) := (D_{KL}(\sigma_\theta \mid \sigma_{\theta'}))_{\theta \neq \theta' \in \Theta}$. That is, $\boldsymbol{D}_{KL}(\sigma)$ denotes the vector of KL divergences between state-contingent signal distributions generated by experiment σ. Let $\mathcal{D} := \boldsymbol{D}_{KL}[\mathcal{E}_b] \subseteq \mathbb{R}_+^{|\Theta|(|\Theta|-1)}$ denote the image of this map. In the sequel, we will appeal to the following lemmas, which are due to Pomatto et al. (2019).

Lemma 29. $\mathbb{R}_{++}^{|\Theta|(|\Theta|-1)} \subseteq \mathcal{D}$.

Proof. This fact is established during the proof of Pomatto et al. (2019, Lemma 2). □

Lemma 30. *Let C be an Extensible and Dilution Linear cost function. Then C exhibits Constant Marginal Cost if and only if there exists a continuous coefficient function $\beta : \Delta_\circ \to \mathbb{R}_+^{|\Theta|(|\Theta|-1)}$ such that*

$$C(\sigma \mid p) = \sum_{\theta,\theta'} \beta_{\theta,\theta'}(p) D_{KL}(\sigma_\theta \mid \sigma_{\theta'}), \tag{44}$$

for all $\sigma \in \mathcal{E}_b$ and $p \in \Delta_\circ$, in which case C is Posterior Separable with prior-dependent potential function F defined by

$$F(q \mid p) := \sum_{\theta,\theta'} \frac{q_\theta}{p_\theta} \beta_{\theta,\theta'}(p) \log\left(\frac{q_\theta}{q_{\theta'}}\right). \tag{45}$$

Proof. Let $\sigma \in \mathcal{E}_\circ$ and $p \in \Delta_\circ$. Let $\pi := \pi_{\langle \sigma \mid p \rangle}$. For each $n \in \mathbb{N}$, let $\{D_i\}$ denote a (Borel) partition of $\Delta(\Theta)$ such that $\operatorname{diam}(D_i) \leq 1/n$ for all i. Define the posterior distribution π^n via $\operatorname{supp}(\pi^n) := \{q_i = \mathbb{E}_\pi[\tilde{q} \mid \tilde{q} \in D_i]\}$ and $\pi^n(D_i) := \pi(D_i)$. Because $\pi(\Delta \backslash \Delta_\circ) = 0$, it is easy to see that for each n there exists a $\delta_n > 0$ such that $\operatorname{supp}(\pi^n) \subset \Delta_{\delta_n}$. Thus, any experiment σ^n for which $\pi_{\langle \sigma^{(n)} \mid p \rangle} = \pi^n$ satisfied $\sigma^n \in \mathcal{E}_b$. We also have $\sigma^{(n)} \gtrsim_B \sigma^{(n-1)}$ by construction. Therefore, every $\sigma \in \mathcal{E}_\circ$ can be approximated in this manner by a sequence of bounded experiments. It is then easy to see that, because C is Dilution Linear and Constant Marginal Cost, $\overline{C}$ inherits these properties (on all of $\mathcal{E}_\circ$) by Definition 24(ii).

Therefore, in conjunction with Definition 24(iii), for each $p \in \cdot_\circ$ the function $\overline{C}(\sigma \mid p)$ satisfies the hypotheses of Pomatto et al. (2019, Theorem 1). Applying that result prior-by-prior delivers that $\overline{C}(\sigma \mid p) = \sum_{\theta,\theta'} \beta_{\theta,\theta'}(p) D_{KL}(\sigma_\theta \mid \sigma_{\theta'})$ for all $\sigma \in \mathcal{E}_\circ$ and $p \in \Delta_\circ$. The representation (44) then follows from Definition 24(i). Continuity of the function β can then be shown to follow from continuity of

$C(\sigma \mid \cdot) : \Delta_\circ \to \mathbb{R}_+$ for each $\sigma \in \mathcal{E}_b$ and Lemma 29. Finally, the Posterior Separable representation (45) can be shown by direct calculation. □

I.2 Proof that (i) $\iff$ (ii)

The fact that (ii) $\implies$ (i) is immediate from the facts that Total Information is UPS and is a special case of the representation in Lemma 30. It therefore suffices to prove that (ii) $\implies$ (i). To that end, let C^* be Extensible, SLP, and exhibit Constant Marginal Cost. By Corollary 1.2 and Lemma 30, C^* is Posterior Separable with the prior-dependent potential function $F(q \mid p)$ given in (45).

For notational ease in what follows, for any $p \in \Delta_\circ$ let $(\tilde{q}, \tilde{\nu}) \sim \pi \in \Delta(\Delta_\circ \times \Delta_\circ)$, with marginal posterior distributions $\tilde{q} \sim \pi' \in \Pi(p)$ and $\tilde{\nu} \sim \pi'' \in \Pi$ and conditional posterior distributions $\tilde{\nu} \mid q \sim \pi(\cdot \mid q) \in \Pi(q)$. Suppose also that each of these posterior distributions is induced by a bounded experiment. In words, π corresponds to the joint distribution of first- and second-stage posterior beliefs induced by a sequential experiment $\sigma'' * \sigma' \sim_B \sigma$, where $\pi_{\langle \sigma | p \rangle} = \pi'$ and there is no extraneous randomization (i.e., the second stage experiment is Markov in the first-stage belief, as in Appendix B).

Then by Theorem 1, C^* satisfies the Preference for One-Shot Learning inequality:

$$\mathbb{E}_{\pi''}[F(\tilde{\nu} \mid p) - F(p \mid p)] \leq \mathbb{E}_{\pi'}[F(\tilde{q} \mid p) - F(p \mid p)] + \mathbb{E}_{\pi}[F(\tilde{\nu} \mid \tilde{q}) - F(\tilde{q} \mid \tilde{q})] \tag{46}$$

The LHS of (46) may be expanded as

$$\mathbb{E}_{\pi''}[F(\tilde{\nu} \mid p) - F(p \mid p)] = \mathbb{E}_{\pi'}[F(\tilde{q} \mid p) - F(p \mid p)] + \mathbb{E}_{\pi}[F(\tilde{\nu} \mid p) - F(\tilde{q} \mid p)],$$

which upon substitution into (46) yields:

$$\mathbb{E}_{\pi}[F(\tilde{\nu} \mid p) - F(\tilde{q} \mid p)] \leq \mathbb{E}_{\pi}[F(\tilde{\nu} \mid \tilde{q}) - F(\tilde{q} \mid \tilde{q})] \tag{47}$$

Define the functions $\hat{\boldsymbol{\beta}}, \boldsymbol{\ell} : \Delta_\circ \to \mathbb{R}_+^{|\Theta|(|\Theta|-1)}$ by

$$\hat{\boldsymbol{\beta}}(p) := \left(\frac{\beta_{\theta,\theta'}(p)}{p_\theta}\right)_{\theta \neq \theta' \in \Theta} \qquad \text{and} \qquad \boldsymbol{\ell}(q) := \left(q_\theta \log\left(\frac{q_\theta}{q_{\theta'}}\right)\right)_{\theta \neq \theta' \in \Theta}$$

Note that we may write the potential function as $F(q \mid p) := \hat{\boldsymbol{\beta}}(p) \cdot \boldsymbol{\ell}(q)$. Substituting this into (47) yields

$$\begin{aligned} 0 &\geq \mathbb{E}_\pi[(F(\tilde{\nu} \mid p) - F(\tilde{\nu} \mid \tilde{q})) + (F(\tilde{q} \mid \tilde{q}) - F(\tilde{q} \mid p))] \\ &= \mathbb{E}_\pi\left[\boldsymbol{\ell}(\tilde{\nu}) \cdot \left(\hat{\boldsymbol{\beta}}(p) - \hat{\boldsymbol{\beta}}(\tilde{q})\right) + \boldsymbol{\ell}(\tilde{q}) \cdot \left(\hat{\boldsymbol{\beta}}(\tilde{q}) - \hat{\boldsymbol{\beta}}(p)\right)\right] \\ &= \mathbb{E}_\pi\left[(\boldsymbol{\ell}(\tilde{\nu}) - \boldsymbol{\ell}(\tilde{q})) \cdot \left(\hat{\boldsymbol{\beta}}(p) - \hat{\boldsymbol{\beta}}(\tilde{q})\right)\right] \end{aligned} \tag{48}$$

Let $\sigma^{(q)} \in \mathcal{E}_b$ satisfy $\pi_{\langle \sigma^{(q)} | q \rangle}(\cdot) = \pi(\cdot \mid q)$. Now, for any pair of states $\theta \neq \theta'$, we may calculate:

$$\begin{aligned} \mathbb{E}_{\pi(\cdot|q)}[\ell_{\theta,\theta'}(\tilde{\nu}) - \ell_{\theta,\theta'}(q)] &= \int_S \left[\nu_\theta(s) \log\left(\frac{q_\theta}{q_{\theta'}} \frac{d\sigma_\theta^{(q)}}{d\sigma_{\theta'}^{(q)}}(s)\right) - q_\theta \log\left(\frac{q_\theta}{q_{\theta'}}\right)\right] d\left(\sum_{\theta'' \in \Theta} q_{\theta''} \sigma^{(q)}(s \mid \theta'')\right) \\ &= \mathbb{E}_{\pi(\cdot|q)}\left[(\tilde{\nu}_\theta - q_\theta) \log\left(\frac{q_\theta}{q_{\theta'}}\right)\right] \end{aligned} \tag{49}$$

$$+\int_S \frac{q_\theta d\sigma^{(q)}(s\mid\theta)}{d\left(\sum_{\theta''\in\Theta} q_{\theta''}\sigma^{(q)}(s\mid\theta'')\right)}\log\left(\frac{d\sigma_\theta^{(q)}}{d\sigma_{\theta'}^{(q)}}(s)\right)d\left(\sum_{\theta''\in\Theta} q_{\theta''}\sigma^{(q)}(s\mid\theta'')\right) \tag{50}$$

$$= 0 + q_\theta\int_S \log\left(\frac{d\sigma_\theta^{(q)}}{d\sigma_{\theta'}^{(q)}}(s)\right)d\sigma^{(q)}(s\mid\theta) \tag{51}$$

$$= q_\theta D_{KL}\left(\sigma_\theta^{(q)}\mid\sigma_{\theta'}\right) \tag{52}$$

The equality in (49) is by a change of variables from posteriors to signals, and by Bayes' rule in the first logarithm term. Then (50) is from regrouping terms and another change of variable back to posteriors in the first term. Next, (51) follows from the martingale condition $\mathbb{E}_{\pi(\cdot|p)}[\tilde{\nu}_\theta] = q_\theta$ and from a change of measure in the second integral. Finally, (52) is by the definition of KL divergence.

Therefore, define the vector $\boldsymbol{D}_{KL}^{(q)}(\sigma) := (q_\theta D_{KL}(\sigma_\theta\mid\sigma_{\theta'}))_{\theta,\theta'\in\Theta} \in \mathbb{R}_+^{|\Theta|(|\Theta|-1)}$. By (49)–(52) and the Law of Iterated Expectation, we may equivalently rewrite (48) as

$$0 \geq \mathbb{E}_\pi\left[\boldsymbol{D}_{KL}^{(\tilde{q})}(\sigma)\cdot\left(\hat{\boldsymbol{\beta}}(p)-\hat{\boldsymbol{\beta}}(\tilde{q})\right)\right] \tag{53}$$

Suppose, towards a contradiction, that there exist $p,q\in\Delta_\circ$ such that $\hat{\boldsymbol{\beta}}(q)\neq\hat{\boldsymbol{\beta}}(p)$; in particular, suppose without loss of generality that $\hat{\beta}_{\theta,\theta'}(q) > \hat{\beta}_{\theta,\theta'}(p)$. (If $\hat{\boldsymbol{\beta}}(q)\lneq\hat{\boldsymbol{\beta}}(p)$, then we may simply interchange the role of p and q in the subsequent argument.) Pick any $\alpha\in(0,1)$ and $q'\in\Delta_\circ$ such that $p=\alpha q+(1-\alpha)q'$. Construct a joint distribution $\pi\in\Delta(\Delta_\circ\times\Delta_\circ)$ as follows. First, let $\pi' := \alpha\delta_q+(1-\alpha)\delta_{q'}$. Then let $\pi(\cdot\mid q') := \delta_q$ and let $\pi(\cdot\mid q) := \pi_{\langle\sigma^{(q)}|q\rangle}$ for some $\sigma^{(q)}\in\mathcal{E}_b$ for which $D_{KL}(\sigma_\theta^{(q)}\mid\sigma_{\theta'}^{(q)}) = M > 0$ and $D_{KL}(\sigma_{\theta'}^{(q)}\mid\sigma_\theta^{(q)}) = D_{KL}(\sigma_{\hat{\theta}}^{(q)}\mid\sigma_{\hat{\theta}'}^{(q)}) = m > 0$ for all $\hat{\theta},\hat{\theta}'\notin\{\theta,\theta'\}$. The existence of such an experiment for any choice of $M,m>0$ is guaranteed by Lemma 29. For this choice of joint distribution, (53) reduces to

$$0 \geq \alpha\boldsymbol{D}_{KL}^{(q)}(\sigma^{(q)})\cdot\left(\hat{\boldsymbol{\beta}}(p)-\hat{\boldsymbol{\beta}}(\tilde{q})\right) = Mq_\theta\left(\hat{\beta}_{\theta,\theta'}(q)-\hat{\beta}_{\theta,\theta'}(p)\right)+Km \tag{54}$$

where $K := q_\theta\left(\hat{\beta}_{\theta',\theta}(q)-\hat{\beta}_{\theta',\theta}(p)\right)+\sum_{\hat{\theta},\hat{\theta}'\notin\{\theta,\theta'\}} q_{\hat{\theta}}\left(\hat{\beta}_{\hat{\theta},\hat{\theta}'}(q)-\hat{\beta}_{\hat{\theta},\hat{\theta}'}(p)\right)$. Because $q_\theta>0$ by $q\in\Delta_\circ$ and $\hat{\beta}_{\theta,\theta'}(q)-\hat{\beta}_{\theta,\theta'}(p)>0$ by assumption, there exists $M>0$ large enough and $m>0$ small enough that the inequality in (54) is violated, which is the desired contradiction.

Consequently, there exists some $\boldsymbol{\gamma}\in\mathbb{R}_+^{|\Theta|(|\Theta|-1)}$ such that $\hat{\boldsymbol{\beta}}(\cdot)\equiv\boldsymbol{\gamma}$. But this is equivalent to $\beta_{\theta,\theta'}(p)\equiv p_\theta\gamma_{\theta,\theta'}$, so that C is a Total Information cost function, as desired.

I.3 Remainder of Proof

Let C be an Extensible and Dilution Linear Direct Cost function exhibiting Constant Marginal Cost. If ΦC is a Total Information cost function with coefficient vector $\boldsymbol{\gamma}$, then Theorem 3 and the representation (44) of C in Lemma 30 imply that we must have the (PIL) inequality

$$\sum_{\theta,\theta'}\left(\beta_{\theta,\theta'}(p)-p_\theta\gamma_{\theta,\theta'}\right)D_{KL}(\sigma_\theta\mid\sigma_{\theta'}) \geq 0 \tag{55}$$

for all $\sigma\in\mathcal{E}_b$ and $p\in\Delta_\circ$.

Suppose there exists some $p\in\Delta_\circ$ and some pair of states $\theta\neq\theta'$ for which $\beta_{\theta,\theta'}(p)-p_\theta\gamma_{\theta,\theta'}<0$. Then by Lemma 29, for every $M,m>0$ there exists some $\sigma\in\mathcal{E}_b$ for which $D_{KL}(\sigma_\theta\mid_{\theta'}) = M$ and all

other KL divergences are equal to m. By the supposition, there exist M sufficiently large and m sufficiently small that (55) is violated. Thus, it follows that we must have

$$\beta_{\theta,\theta'}(p) - p_\theta \gamma_{\theta,\theta'} \geq 0 \tag{56}$$

for all $p \in \Delta_\circ$ and all $\theta \neq \theta'$.

Now, Theorem 3 and Lemma 6 also demand that $\mathcal{H}_q F(q \mid p)|_{q=p} = \mathcal{H}G(p)$, where F is the potential function in the Posterior Separable representation (45) of C and G is the Total Information potential function from (TI). This implies that $\text{Diag}(p)\mathcal{H}_q F(q \mid p)|_{q=p}\text{Diag}(p) = \text{Diag}(p)\mathcal{H}G(p)\text{Diag}(p)$. Direct computation shows that, for $\theta \neq \theta'$, the $(\theta,\theta')^{th}$ entry of this matrix equation is

$$\beta_{\theta,\theta'}(p) + \beta_{\theta',\theta}(p) = p_\theta \gamma_{\theta,\theta'} + p_{\theta'} \gamma_{\theta',\theta}. \tag{57}$$

Combining (56) with (57) yields $\beta_{\theta,\theta'}(p) = p_\theta \gamma_{\theta,\theta'}$ for all θ, θ' and $p \in \Delta_\circ$, which is equivalent to $C = \Phi C$, completing the proof.

Remark 2. *Notice that Lemma 30 implies that any Extensible and Dilution Linear Direct Cost function exhibiting Constant Marginal Cost is, in fact, Posterior Separable. A weaker requirement on the Direct Cost function, therefore, would be that it is Locally Linear, and that its Locally Linear approximation is Extensible and exhibits Constant Marginal Cost. The same proof would then go through by applying Lemma 30 to the Direct Cost's Locally Linear approximation and using Lemma 6 to derive the inequality* (55) *from* (PGL).

J Proof of Theorem 5

We omit a formal proof of the necessity direction of the theorem — that Mutual Information is Weakly Compression Invariant and Compression Monotone — which follows from the well-known fact that Mutual Information satisfies the "data processing inequality" (e.g., Cover and Thomas (2006, Theorem 2.8.1)). Below, we show that these properties are sufficient for Mutual Information, i.e., points (i) and (ii) each imply point (iii) of the theorem.

J.1 Preliminary Lemmas

In Appendix J.1.1, we first show that any Bounded UPS cost function can be continuously extended to a "full domain" UPS cost function that is defined on the full domain $\mathcal{E} \times \Delta(\Theta)$ of experiment-prior pairs. In Appendix J.1.2, we then present a characterization of Shannon entropy in terms of a "recursivity" condition for "full domain" potential functions. The bulk of the proof, in Appendices J.2 and J.3 below, shows that C being Weakly Compression Invariant or Compression Monotone implies that the "full domain" potential function in the continuous extension of C satisfies the "recursivity" condition and is therefore proportional to Shannon entropy, which implies that C itself is a Mutual Information cost function. A proof method very similar to ours, based on the recursivity condition (R) and its relation to Fadeev (1956) (Lemma 33), as well as the geometric argument from Lemma 41 (illustrated in Figure 6), was first provided in Tian (2019, Lemmas 2',3,4, Figure 1).

J.1.1 Extension

Towards the first extension step, we begin with the following definition:

Definition 25 (Full Domain UPS)**.** *A Full Domain cost function C is **Full Domain UPS** if there exists a (convex) potential function $F : \Delta(\Theta) \to \mathbb{R}$ such that*

$$C(\sigma \mid p) = \mathbb{E}_{\pi_{\langle \sigma|p \rangle}}[F(\tilde{q}) - F(p)] \tag{FD-UPS}$$

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

We note that the literature typically refers to Full Domain UPS simply as "Uniformly Posterior Separable." However, it can be shown that the restriction of a Full Domain UPS cost function to $\mathcal{E}_b \times \Delta_\circ$ is necessarily Bounded (cf. Gale et al. (1968)), which is strictly less general than our UPS definition. The following lemma shows a converse of this fact:

Lemma 31. *Let $C : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ be a Bounded UPS cost function with potential $F : \Delta_\circ \to \mathbb{R}$. There exists a continuous convex function $\overline{F} : \Delta(\Theta) \to \mathbb{R}$ of F, which generates (via (FD-UPS)) a Full Domain UPS cost function $\overline{C} : \mathcal{E} \times \Delta(\Theta) \to \mathbb{R}_+$ such that:*

(i) $\overline{F}(\delta_\theta) = 0$ for all $\theta \in \Theta$.[84]

(ii) The restriction $\overline{C}\,|_{\mathcal{E}_b \times \Delta_\circ} = C$.

(iii) $\overline{C}$ is weak continuous: If (a) $\{\pi^n\} \subset \Pi$ satisfies $\pi^n \to^{w^*} \pi^*$ and (b) $\{(\sigma^n, p^n)\} \subseteq \mathcal{E} \times \Delta(\Theta)$ and $(\sigma^*, p^*) \in \mathcal{E} \times \Delta(\Theta)$ satisfy $\pi_{\langle \sigma^n|p^n \rangle} = \pi^n$ and $\pi_{\langle \sigma^*|p^* \rangle} = \pi^*$, the $\lim_{n\to\infty} \overline{C}(\sigma^n \mid p^n) = \overline{C}(\sigma^* \mid p^*)$.*

(iv) If $\sigma, \tau \in \mathcal{E}$ and $p \in \Delta(\Theta)$ satisfy $\pi_{\langle \sigma|p \rangle} = \pi_{\langle \tau|p \rangle}$, then $\overline{C}(\sigma \mid p) = \overline{C}(\sigma \mid p)$.

Proof. We first show that F admits a convex continuous extension $\overline{F} : \Delta(\Theta) \to \mathbb{R}$ (which need not satisfy point (i) of the lemma). If F is bounded, then the existence (and uniqueness) of such an extension follows from Gale et al. (1968), so it suffices to show that F is bounded, i.e., $\sup_{p \in \Delta_\circ} |F(p)| < \infty$. To that end, let $p \in \Delta_\circ$ be given and let $\nabla F(p)$ denote a subgradient of F at p. Define the convex function $\tilde{F} : \Delta_\circ \to \mathbb{R}_+$ by $\tilde{F}(q) := F(q) - F(p) - \nabla F(p) \cdot (q - p)$, which is non-negative because F is convex. Because $\tilde{F}(q)$ differs from $F(q)$ by only an affine term, $\tilde{F}$ is also a valid potential function for C. Also, it is clear that $\tilde{F}$ is bounded if and only if F is bounded, so it suffices to show the former. Suppose, towards contradiction, there exists a sequence $\{q^n\} \subset \Delta_\circ$ such that $\limsup_{n\to\infty} \tilde{F}(q^n) = +\infty$. We may assume without loss of generality that $q^n \to q^* \in \Delta(\Theta) \backslash \Delta_\circ$. Pick any $\alpha \in (0, 1)$, and define $r^n := \alpha q^n + (1 - \alpha)p \in \Delta_\circ$. Let $\sigma^n \in \mathcal{E}_b$ satisfy $\pi_{\langle \sigma^n|r^n \rangle} = \alpha \delta_{q^n} + (1 - \alpha)\delta_p$. Note that $\sup_{n \in \mathbb{N}} \tilde{F}(r^n) < \infty$ because, by construction, there exists some $\delta > 0$ such that $\{r^n\} \subset \Delta_\delta$ and $\tilde{F}$ is continuous on the compact set Δ_δ. Therefore, $\limsup_{n\to\infty} C(\sigma^n \mid r^n) = \limsup_{n\to\infty} \left[\alpha \tilde{F}(q^n) + (1 - \alpha)\tilde{F}(p) - \tilde{F}(r^n)\right] = \infty$, contradicting that C is Bounded.

Now let $\overline{C}$ denote the Full Domain UPS cost function corresponding to $\overline{F}$. Define $\mathbf{f} \in \mathbb{R}^\Theta$ by $f_\theta := F(\delta_\theta)$. Then $\hat{F}(q) := \overline{F}(q) - \mathbf{f} \cdot q$ satisfies property (i) and is continuous by construction. Moreover, because $\overline{F}$ and $\hat{F}$ differ by an affine term, they are both valid potentials for $\overline{C}$. So simply re-define $\overline{F}$ as $\hat{F}$, so that point (i) of the lemma holds. Point (ii) of the lemma is immediate. Point (iii) follows from the continuity of $\overline{F}$ and the Portmanteau Theorem. Point (iv) follows from the definition of Full Domain UPS. □

[84] Frankel and Kamenica (2019) refer to this property of potential functions as "null uncertainty."

For completeness, we note that the following lemma, which is a slight variant of Lemma 1 from the main text, provides an equivalent characterization of the class of Full Domain UPS cost functions:

Lemma 32. *The Full Domain cost function C is Full Domain UPS if and only if it satisfies*

$$C(\overline{\sigma} \mid p) = C(\sigma \mid p) + \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[C(\overline{\sigma} \mid \tilde{q})], \tag{58}$$

in which case $F(p) := -C(\overline{\sigma} \mid p)$ is convex and satisfies (FD-UPS).

Proof. The "only if" direction is trivial. Suppose that C satisfies (58) and define $F(p) := -C(\overline{\sigma} \mid p)$. Define the Full Domain cost function $\hat{C}$ by $\hat{C}(\sigma \mid p) = \mathbb{E}_{\pi_{\langle \sigma | p \rangle}}[F(\tilde{q}) - F(p)]$. It is easy to see from (FD-UPS) that $\hat{C} = C$. It remains to show that F is convex, which in turn implies that $\hat{C}$ and hence C is Full Domain UPS. But this follows immediately from the fact that $C \geq 0$ and an argument analogous to that in the proof of Lemma 21. □

J.1.2 Recursivity

The following definition is a variant of the recursivity axiom introduced in Shannon's (1948) characterization of Shannon entropy, and which has appeared repeatedly in the subsequent literature (see Csiszár (2008) and Ebanks et al. (1998, Ch. 3) for surveys).

Definition 26 (Recursive)**.** *The function $F : \Delta(\Theta) \to \mathbb{R}$ is Recursive if*

$$F(p) = F\left(p + p_{\theta'}(\delta_\theta - \delta_{\theta'})\right) + (p_\theta + p_{\theta'}) F\left(\frac{p_\theta}{p_\theta + p_{\theta'}}\delta_\theta + \frac{p_{\theta'}}{p_\theta + p_{\theta'}}\delta_{\theta'}\right) \tag{R}$$

for all $p \in \Delta(\Theta)$ and $\theta, \theta' \in \Theta$ with $p_\theta + p_{\theta'} > 0$.

The following lemma constitutes a key step in the proof of Theorem 5:

Lemma 33. *Suppose that $|\Theta| \geq 3$. Let $F : \Delta(\Theta) \to \mathbb{R}$ be continuous. If F is Recursive, then there exists a constant $\alpha \in \mathbb{R}$ such that $F(p) \equiv \alpha H(p)$, where $H(p) := -\sum_\theta p_\theta \log(p_\theta)$ is Shannon entropy.*

Lemma 33 is closely related to Shannon's (1948) characterization of entropy. It illustrates that Shannon's (1948) second axiom, which requires that F is monotone increasing in the support size of uniform distributions, is needed only to ensure that $\alpha > 0$. We establish Lemma 33 as a corollary to Fadeev's (1956) celebrated characterization of Shannon entropy, which replaces Shannon's (1948) monotonicity axiom with a symmetry assumption (and adopts a slightly weaker recursivity assumption):

Lemma 34. *Suppose that $|\Theta| \geq 3$. Consider a collection $\{F_k\}_{k=2}^{|\Theta|}$ of functions $F_k : \Delta(\{1, \ldots, k\}) \to \mathbb{R}$ satisfying the following properties:*

(i) The function F_2 is continuous.

(ii) The functions F_2 and F_3 are symmetric.

(iii) For all $k \in \{2, \ldots, |\Theta|\}$ and $(p_1, \ldots, p_k) \in \text{int}\Delta(\{1, \ldots, k\})$, we have:

$$F_k(p_1, \ldots, p_k) = F_{k-1}(p_1 + p_2, p_3, \ldots, p_k) + (p_1 + p_2) F_2\left(\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)\right) \tag{59}$$

Then there exists a constant $\alpha \in \mathbb{R}$ such that $F_k(p_1,\dots,p_k) = -\alpha \sum_{i=1}^{k} p_i \log(p_i)$.

Proof. See Fadeev (1956), or the more general results of Tverberg (1958) and Lee (1964). □

There are three main differences between the statements of Lemmas 33 and 34. First, Lemma 33 assumes the recursivity condition (R), which applies to all pairs of states, whereas Lemma 34 assumes the recursivity condition (59), which applies only to a single predetermined pair of indices. Second, Lemma 34 works with the collection of functions $\{F_k\}$, where F_k is defined over all probability distributions with a given support size without specifying which states those probabilities are assigned to. Third, Lemma 34 also assumes that F_2 and F_3 are symmetric, while Lemma 33 does not assume symmetry. Thus, relative to Fadeev's (1956) result, Lemma 33 assumes a stronger version of recursivity but does not assume any form of symmetry. The content of Lemma 33 is that, in fact, this stronger recursivity property implies symmetry. However, once symmetry is established, Lemma 33 follows directly from Lemma 34.

Proof of Lemma 33. Let F satisfy the hypotheses of Lemma 33. Suppose for now that it is also symmetric (this will proved separately below). Let $\left\{\theta_1,\dots,\theta_{|\Theta|}\right\}$ be an enumeration of Θ. For each $k \in \{2,\dots,|\Theta|\}$, let $\Delta^{(k)} := \{p \in \Delta(\Theta) : \operatorname{supp}(p) = \{\theta_1,\dots,\theta_k\}\}$ and define the function $F_k : \Delta^{(k)} \to \mathbb{R}$ by the restriction $F_k := F\,|_{\Delta^{(k)}}$, which inherits continuity and (assumed) symmetry from F. Thus, Lemma 34 implies that $F_k(p_1,\dots,p_k) = -\alpha \sum_{i=1}^{k} p_i \log(p_i)$ for all such k. Because F is (assumed) symmetric, it is then easy to see that $F(p) = -\alpha H(p)$ for all $p \in \Delta(\Theta)$.

To prove the lemma, it therefore suffices to show that F being Recursive implies that it is symmetric. We first observe that $F(\delta_\theta) = 0$ for all $\theta \in \Theta$. Take any $\theta \neq \theta'$ and $p \in \Delta(\Theta)$ with $\operatorname{supp}(p) = \{\theta,\theta'\}$, for which the equation (R) is equivalent to $F(p) = F(\delta_\theta) + F(p)$. This implies that $F(\delta_\theta) = 0$, as desired.[85]

The proof of symmetry is then by induction on $|\operatorname{supp}(p)| \geq 2$. For the base step, take any $p \in \Delta(\Theta)$ with $\operatorname{supp}(p) = \{\theta,\theta'\}$. Let $\theta'' \notin \{\theta,\theta'\}$. By the recursivity condition (R) applied to the pair of states (θ'',θ), we have

$$\begin{aligned} F(p) &= F\left(p + p_\theta(\delta_{\theta''} - \delta_\theta)\right) + p_\theta F(\delta_\theta) \\ &= F\left(p + p_\theta(\delta_{\theta''} - \delta_\theta)\right) \\ &= F(p \circ \lambda_{\theta \leftrightarrow \theta''}) \end{aligned}$$

where the second line is because $F(\delta_\theta) = 0$ for all $\theta \in \Theta$ and, in the third line, $\lambda_{\theta \leftrightarrow \theta''}$ denotes the transposition that switches (only) the states θ,θ''. The same argument applies to the pair of states (θ',θ''). Hence, F is symmetric with respect to transpositions λ for which $\lambda(\operatorname{supp}(p)) \neq \operatorname{supp}(p)$ (at beliefs p with binary support). But by the same logic, we also have

$$\begin{aligned} F\left(p \circ \lambda_{\theta \leftrightarrow \theta''}\right) &= F\left(\left[p \circ \lambda_{\theta \leftrightarrow \theta''}\right] \circ \lambda_{\theta \leftrightarrow \theta'}\right) \\ &= F\left(\left[\left[p \circ \lambda_{\theta \leftrightarrow \theta''}\right] \circ \lambda_{\theta \leftrightarrow \theta'}\right] \circ \lambda_{\theta' \leftrightarrow \theta''}\right) \\ &= F\left(p \circ \lambda_{\theta \leftrightarrow \theta'}\right), \end{aligned}$$

where the first equality follows from the fact that $\operatorname{supp}\left(p \circ \lambda_{\theta \leftrightarrow \theta''}\right) = \{\theta',\theta''\}$, the second equality follows from the fact that $\operatorname{supp}\left(\left[p \circ \lambda_{\theta \leftrightarrow \theta''}\right] \circ \lambda_{\theta \leftrightarrow \theta'}\right) = \{\theta,\theta''\}$, and the final equality follows from

[85] Note that when $|\Theta| = 2$, this is the only implication of Definition 26. Thus, the $|\Theta| \geq 3$ hypothesis is necessary for Lemmas 33 and 34.

composing the various transpositions. This establishes that F is also symmetric with respect to transpositions λ for which $\lambda(\text{supp}(p)) = \text{supp}(p)$ (at beliefs p with binary support). Thus, F is symmetric with respect to all transpositions (at beliefs p with binary support). Since every permutation can be achieved as the composition of transpositions, this proves that F is symmetric at beliefs p with binary support.

For the inductive step, take $k \in \{2, \dots, |\Theta| - 1\}$ and suppose that $F(p) = F(p \circ \lambda)$ for all permutations λ and all $p \in \Delta(\Theta)$ with $|\text{supp}(p)| \leq k$. Let $p \in \Delta(\Theta)$ with $|\text{supp}(p)| = k + 1$ be given, and take any $\theta, \theta' \in \text{supp}(p)$. For any permutation λ, we have:

$$
\begin{aligned}
F(p \circ \lambda) &= F\left(\left[p + [p \circ \lambda]_{\lambda(\theta')}(\delta_\theta - \delta_{\theta'})\right] \circ \lambda\right) && (60)\\
&\quad + \left([p \circ \lambda]_{\lambda(\theta)} + [p \circ \lambda]_{\lambda(\theta')}\right) F\left(\left[\frac{[p \circ \lambda]_{\lambda(\theta)}}{[p \circ \lambda]_{\lambda(\theta)} + [p \circ \lambda]_{\lambda(\theta')}}\delta_\theta + \frac{[p \circ \lambda]_{\lambda(\theta')}}{[p \circ \lambda]_{\lambda(\theta)} + [p \circ \lambda]_{\lambda(\theta')}}\delta_{\theta'}\right] \circ \lambda\right)\\
&= F\left([p + p_{\theta'}(\delta_\theta - \delta_{\theta'})] \circ \lambda\right) + (p_\theta + p_{\theta'}) F\left(\left[\frac{p_\theta}{p_\theta + p_{\theta'}}\delta_\theta + \frac{p_{\theta'}}{p_\theta + p_{\theta'}}\delta_{\theta'}\right] \circ \lambda\right) && (61)\\
&= F\left(p + p_{\theta'}(\delta_\theta - \delta_{\theta'})\right) + (p_\theta + p_{\theta'}) F\left(\frac{p_\theta}{p_\theta + p_{\theta'}}\delta_\theta + \frac{p_{\theta'}}{p_\theta + p_{\theta'}}\delta_{\theta'}\right) && (62)\\
&= F(p) && (63)
\end{aligned}
$$

where (60) (the first two lines) is (R) applied at belief $p \circ \lambda$ to the pair of states $\lambda(\theta)$ and $\lambda(\theta')$, (61) is by the identify $[p \circ \lambda]_{\lambda(\theta'')} \equiv p_{\theta''}$, (62) is by the assumed symmetry of F at beliefs with support size $\leq k$, and (63) is (R) applied at belief p to the pair of states θ and θ'. Therefore, F is symmetric for beliefs with support size $k + 1$. This completes the inductive step and thus the proof of the lemma. □

J.2 Proof that (i) $\implies$ (iii)

Lemma 35. *Let $\overline{C}$ be a Full Domain UPS cost function with continuous potential $\overline{F} : \Delta(\Theta) \to \mathbb{R}$. If the restriction $\overline{C}\,|_{\mathcal{E}_b \times \Delta_\circ}$ is Weakly Compression Invariant, then $\overline{C}$ satisfies the following strengthening of Axiom 10:*

$$\overline{C}(\sigma \mid p) = \overline{C}(\sigma \mid p') \qquad \text{(FD-WCI)}$$

all experiments $\sigma \in \mathcal{E}$ measurable with respect to coarsening $\kappa \in \mathcal{K}$, and priors $p' \in \Delta(\Theta)$ for which $p'(\kappa(\theta)) = p(\kappa(\theta))$ for all $\theta \in \Theta$.

Proof. Immediate from the weak* continuity of $\overline{C}$ established in Lemma 31(iii), the fact that for every $\sigma \in \mathcal{E}$ and prior $p \in \Delta(\Theta)$ there exits a sequence $\{\sigma^n\} \subset \mathcal{E}_b$ with $\lim_{n\to\infty} \pi_{\langle \sigma^n | p \rangle} = \pi_{\langle \sigma | p \rangle}$, and the fact that for every $\sigma \in \mathcal{E}$ the mapping $p \mapsto \pi_{\langle \sigma | p \rangle}$ is weak* continuous. □

The following lemma (which is illustrated in Figure 4) constitutes the bulk of the proof:

Lemma 36. *Suppose that $|\Theta| \geq 3$. Let C be a Full Domain UPS cost function with continuous potential $F : \Delta(\Theta) \to \mathbb{R}$ that satisfies $F(\delta_\theta) = 0$ for all $\theta \in \Theta$. If C satisfies (FD-WCI), then F is Recursive.*

Proof. Let C and F satisfy the hypotheses of the lemma. Let $p \in \Delta(\Theta)$ and $\theta, \theta' \in \Theta$ with $p_\theta + p_{\theta'} > 0$ be given. Define the experiment $\langle S, \sigma \rangle$ by $S = \{1, \dots, |\Theta| - 1\}$ and $\sigma(1 \mid \theta) = \sigma(1 \mid \theta') = 1$ and $\sigma(k \mid \theta_k) = 1$ for all $k \in \{2, \dots, |\Theta| - 1\}$, where $\{\theta_k\}_{k=2}^{|\Theta|-1}$ is any enumeration of $\Theta \setminus \{\theta, \theta'\}$. Thus, σ is measurable with

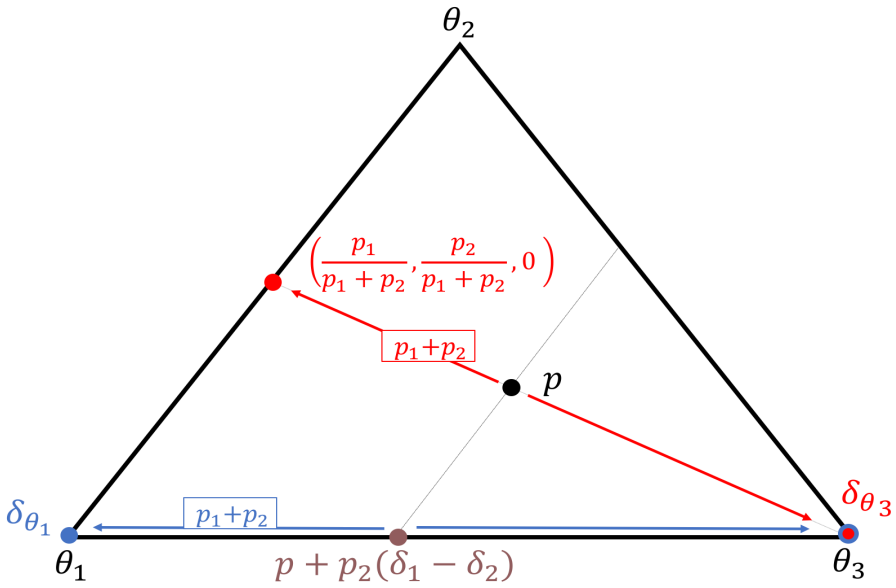

Figure 4: Illustration of Lemma 36 when $|\Theta| = 3$, $\theta = \theta_1$, and $\theta' = \theta_2$.

respect to the compression κ defined by $\kappa(\theta) = \kappa(\theta') = \{\theta, \theta'\}$ and $\kappa(\theta'') = \{\theta''\}$ otherwise. Given any prior $p \in \Delta(\Theta)$, σ induces the posterior distribution

$$\pi_{\langle \sigma | p \rangle} = \sum_{\theta'' \neq \theta, \theta'} p_{\theta''} \delta_{\theta''} + (p_\theta + p_{\theta'}) \left(\frac{p_\theta}{p_\theta + p_{\theta'}} \delta_\theta + \frac{p_{\theta'}}{p_\theta + p_{\theta'}} \delta_{\theta'} \right). \tag{64}$$

Because C satisfies Lemma 35, we have $C(\sigma \mid p) = C(\sigma \mid p + p_{\theta'}(\delta_\theta - \delta_{\theta'}))$ (see the right-hand panel of Figure 4). By the assumed Full Domain UPS form of C, (64), and the assumption that $F(\delta_{\theta)} = 0$ for all $\theta \in \Theta$, we have:

$$\begin{aligned} C(\sigma \mid p) &= \sum_{\theta'' \neq \theta, \theta'} p_{\theta''} F(\delta_{\theta''}) + (p_\theta + p_{\theta'}) F\left(\frac{p_\theta}{p_\theta + p_{\theta'}} \delta_\theta + \frac{p_{\theta'}}{p_\theta + p_{\theta'}} \delta_{\theta'} \right) - F(p) \\ &= (p_\theta + p_{\theta'}) F\left(\frac{p_\theta}{p_\theta + p_{\theta'}} \delta_\theta + \frac{p_{\theta'}}{p_\theta + p_{\theta'}} \delta_{\theta'} \right) - F(p) \end{aligned} \tag{65}$$

and

$$\begin{aligned} C(\sigma \mid p + p_{\theta'}(\delta_\theta - \delta_{\theta'})) &= \sum_{\theta'' \neq \theta, \theta'} p_{\theta''} F(\delta_{\theta''}) + (p_\theta + p_{\theta'}) F(\delta_\theta) - F(p + p_{\theta'}(\delta_\theta - \delta_{\theta'})) \\ &= -F(p + p_{\theta'}(\delta_\theta - \delta_{\theta'})) \end{aligned} \tag{66}$$

Equating (65) and (66) yields the recursivity condition (R), which proves that F is Recursive. □

Now, to prove this portion of Theorem 5, it suffices to apply Lemma 36 to the continuous extensions of the cost function:

Proof that (i) $\implies$ (iii) in Theorem 5. Suppose that $|\Theta| \geq 3$. Let $C : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ be a Bounded UPS cost function with potential F Suppose that C is Weakly Compression Invariant. Let $\overline{C}$ and $\overline{F}$ denote

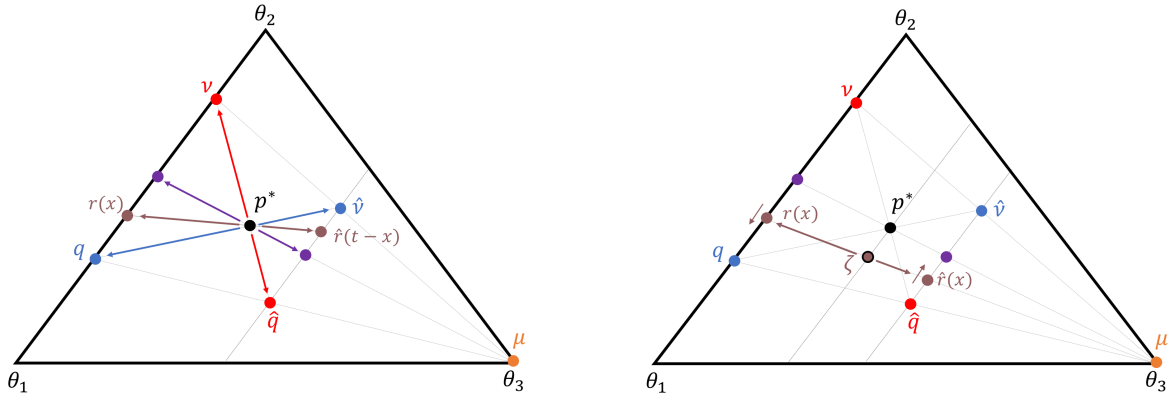

Figure 5: Illustration of Lemma 38 (left) and Lemma 39 (right) when $|\Theta| = 3$, $\theta = \theta_1$, and $\theta' = \theta_2$.

the (continuous) extensions of C and F from Lemma 31. Lemma 36 establishes that $\overline{F}$ is Recursive. Then Lemma 33 implies that there exists some $\alpha \in \mathbb{R}$ such that $\overline{F}(q) \equiv \alpha H(p)$ (it must be that $\alpha < 0$ because $\overline{F}$ is convex and H is concave). By definition of $\overline{C}$, we have $\overline{C}\,|_{\mathcal{E}_b \times \Delta_\circ} = C$, implying that C is a Mutual Information cost function, as desired. □

J.3 Proof that (ii) $\implies$ (iii)

Lemma 37. *Let $\overline{C}$ be a Full Domain UPS cost function with continuous potential $\overline{F} : \Delta(\Theta) \to \mathbb{R}$. If the restriction $\overline{C}\,|_{\mathcal{E}_b \times \Delta_\circ}$ is Compression Monotone, then $\overline{C}$ satisfies the following strengthening of Axiom 11:*

$$\overline{C}(\sigma \mid p) \geq \overline{C}(\sigma_{\langle \kappa | p \rangle} \mid p) \tag{FD-CM}$$

for all compressions κ, experiments $\sigma \in \mathcal{E}$, and priors $p \in \Delta(\Theta)$.

Proof. Immediate from the weak* continuity of $\overline{C}$ established in Lemma 31(iii) and the fact that for any experiment $\sigma \in \mathcal{E}$ and compression κ, the map $p \mapsto \pi_{\langle \sigma_{\langle \kappa | p \rangle} | p \rangle}$ is weak* continuous. □

Lemma 38. *Suppose that $|\Theta| \geq 3$. Let C be a Full Domain UPS cost function with continuous potential $F : \Delta(\Theta) \to \mathbb{R}$. Let $\theta \neq \theta'$, $\nu \in \Delta(\Theta)$ with $\nu_{\theta'} > 0$, $t \in (0, \nu_{\theta'}]$, $\mu \in \Delta(\Theta)$ with $\theta, \theta' \notin supp(\mu)$, and $\alpha \in (0,1)$ be given. Define the beliefs $q, \hat{q}, \hat{\nu}, p^* \in \Delta(\Theta)$ as follows:*

$$q := \nu + t(\delta_\theta - \delta_{\theta'}), \quad \hat{q} := \alpha q + (1-\alpha)\mu, \quad \hat{\nu} := \alpha \nu + (1-\alpha)\mu, \quad p^* := \left(\frac{\alpha}{1+\alpha}\right)\nu + \left(1 - \frac{\alpha}{1+\alpha}\right)\hat{q}$$

Also define the maps $r, \hat{r} : [0,t] \to \Delta(\Theta)$ by $r(x) := \nu + x(\delta_\theta - \delta_{\theta'})$ and $\hat{r}(x) := \alpha r(x) + (1-\alpha)\mu$. Then the family of posterior distributions $\left\{\pi^{(x)}\right\}_{x \in [0,t]}$ defined by

$$\pi^{(x)} := \left(\frac{\alpha}{1+\alpha}\right)\delta_{r(x)} + \left(1 - \frac{\alpha}{1+\alpha}\right)\delta_{\hat{r}(t-x)} \tag{67}$$

satisfies $\pi^{(x)} \in \Pi(p^*)$ for all $x \in [0,t]$. Thus, there exists a family of experiments $\left\{\sigma^{(x)}\right\}_{x \in [0,t]} \subset \mathcal{E}$ such that $\pi_{\langle \sigma^{(x)} | p^* \rangle} = \pi^{(x)}$ for all $x \in [0,t]$.

Proof. For all $x \in [0,t]$, we have

$$\begin{aligned}\left(\frac{\alpha}{1+\alpha}\right)r(x) + \left(1-\frac{\alpha}{1+\alpha}\right)\hat{r}(t-x) &= \left(\frac{\alpha}{1+\alpha}\right)[\nu + x(\delta_\theta - \delta_{\theta'})] + \left(1-\frac{\alpha}{1+\alpha}\right)[\alpha(\nu + (t-x)(\delta_\theta - \delta_{\theta'})) + (1-\alpha)\mu] \\ &= \left(\frac{\alpha}{1+\alpha}\right)\nu + \left(1-\frac{\alpha}{1+\alpha}\right)[\alpha(\nu + t(\delta_\theta - \delta_{\theta'})) + (1-\alpha)\mu] \\ &= \left(\frac{\alpha}{1+\alpha}\right)\nu + \left(1-\frac{\alpha}{1+\alpha}\right)[\alpha q + (1-\alpha)\mu]\end{aligned}$$

which is the definition of p^*, as desired. (See the left-hand panel of Figure 5 for illustration.) □

Lemma 39. *Suppose that $|\Theta| \geq 3$. Let C be a Full Domain UPS cost function with continuous potential $F : \Delta(\Theta) \to \mathbb{R}$, and suppose that C satisfies (FD-CM). Define the maps $f, \hat{f} : [0,t] \to \mathbb{R}$ by*

$$\begin{aligned}f(x) &:= \left(\frac{\alpha}{1+\alpha}\right)F(r(x)) \\ \hat{f}(x) &:= \left(1-\frac{\alpha}{1+\alpha}\right)F(\hat{r}(x)),\end{aligned}$$

where $\alpha, r(\cdot), \mu$ are as defined in Lemma 38 above. Then these functions are absolutely continuous and, at almost every $x \in [0,t]$, are differentiable and satisfy $f'(x) = \hat{f}'(x)$.

Proof. By construction, $\operatorname{supp}(r(x)) = \operatorname{supp}(\hat{r}(x))$ for all $x \in (0,t)$. Since $f, \hat{f}$ are convex and continuous by construction, it follows that each is absolutely continuous, and hence almost-everywhere differentiable, on $[0,t]$. Let $\mathcal{D}$ denote the (full measure) set of points in $(0,t)$ at which both functions are differentiable. Let $x \in \mathcal{D}$ be given. For all $\epsilon \in (-x, t-x)$, define the posterior distribution $\rho^{(\epsilon)} \in \Pi$ by

$$\rho^{(\epsilon)} := \left(\frac{\alpha}{1+\alpha}\right)\delta_{r(x+\epsilon)} + \left(1-\frac{\alpha}{1+\alpha}\right)\delta_{\hat{r}(x-\epsilon)}.$$

By direct calculation, we see that $\mathbb{E}_{\rho(\epsilon)}[\tilde{q}] = \left(\frac{\alpha}{1+\alpha}\right)r(x) + \left(1-\frac{\alpha}{1+\alpha}\right)\hat{r}(x) =: \zeta$ for all $\epsilon \in (-x, t-x)$. Thus, for every such ϵ there exists an experiment $\tau^{(\epsilon)} \in \mathcal{E}$ for which $\pi_{\langle \tau^{(\epsilon)}|\zeta\rangle} = \rho^{(\epsilon)}$.

Let κ denote the compression for which $\kappa(\theta) = \kappa(\theta') = \{\theta, \theta'\}$ and $\kappa(\theta'') = \{\theta''\}$ otherwise. A short calculation delivers that $\tau^{(0)} \sim_B \tau^{(\epsilon)}_{\langle \kappa|\zeta\rangle}$ for all $\epsilon \in (-x, t-x)$. (See the right-hand panel of Figure 5 for illustration.) Now consider the minimization problem

$$\inf_{\epsilon \in (-x,t-x)} C(\tau^{(\epsilon)} \mid \zeta) = \inf_{\epsilon \in (-x,t-x)} \left[f(x+\epsilon) + \hat{f}(x-\epsilon)\right], \tag{68}$$

where the equality follows from the Full Domain UPS hypothesis on C and the definitions of $f, \hat{f}$. The fact that C satisfies (FD-CM) implies that $\epsilon = 0$ is a solution to the program (68). Since $x \in \mathcal{D}$, the necessary first-order condition

$$\frac{d}{d\epsilon}\left[f(x+\epsilon) + \hat{f}(x-\epsilon)\right]\Big|_{\epsilon=0} = f'(x) - \hat{f}'(x) = 0$$

holds, delivering the lemma. □

Lemma 40. *Suppose that $|\Theta| \geq 3$. Let C be a Full Domain UPS cost function with continuous potential $F : \Delta(\Theta) \to \mathbb{R}$. Let the belief $p^* \in \Delta(\Theta)$ and experiments $\sigma^{(0)}, \sigma^{(t)} \in \mathcal{E}$ be defined as in Lemma 38. If C satisfies (FD-CM), then $C\left(\sigma^{(0)} \mid p^*\right) = C\left(\sigma^{(t)} \mid p^*\right)$.*

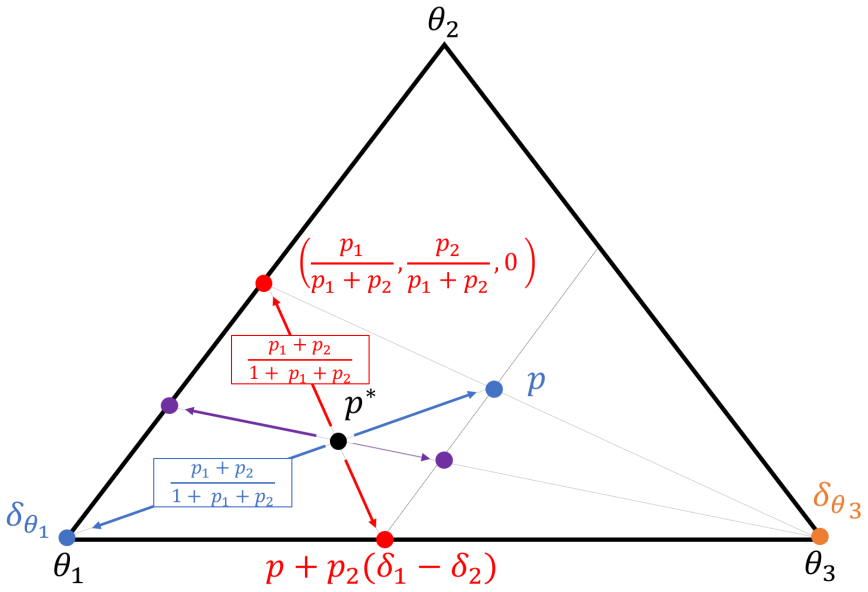

Figure 6: Illustration of Lemma 41 when $|\Theta| = 3$, $\theta = \theta_1$, and $\theta' = \theta_2$.

Proof. Define the function $c : [0,t] \to \mathbb{R}$ by $c(x) := C\left(\sigma^{(x)} \mid p^*\right)$, where $\sigma^{(x)} \in \mathcal{E}$ is as defined in Lemma 38. Because C is Full Domain UPS and by (67), we have $c(x) := f(x) + \hat{f}(t-x)$. By Lemma 39, $f(\cdot)$, $\hat{f}(\cdot)$, and hence $c(\cdot)$ are absolutely continuous on $(0,t)$ and, by hypothesis, are continuous on $[0,t]$. Therefore:

$$c(t) - c(0) = \int_0^t \left[f'(x) + \hat{f}'(t-x)\right] dx \tag{69}$$

$$= \int_0^t f'(x)dx - \int_0^t \hat{f}'(y)dy \qquad \text{where } y = t-x \tag{70}$$

$$= \int_0^t \left[f'(x) - \hat{f}'(x)\right] dx \tag{71}$$

$$= 0 \tag{72}$$

where (69) follows from continuity of $c(x)$ and the Fundamental Theorem of Calculus, (70) is a standard change of variable, and (72) follows from the fact that the integrand in (71) is almost-everyone zero by Lemma 39. Thus, by definition of $c(\cdot)$, we obtain the desired equality $C\left(\sigma^{(0)} \mid p^*\right) = C\left(\sigma^{(t)} \mid p^*\right)$. □

Lemma 41. *Suppose that* $|\Theta| \geq 3$. *Let* C *be a Full Domain UPS cost function with continuous potential* $F : \Delta(\Theta) \to \mathbb{R}$ *that satisfies* $F(\delta_\theta) = 0$ *for all* $\theta \in \Theta$. *If* C *satisfies* (FD-CM), *then* F *is Recursive.*

Proof. Let $\theta \neq \theta'$ be given. Notice that because $F(\delta_\theta) = 0$, the Recursive condition (R) is vacuously satisfied when $p_{\theta'} = 0$. Thus, it suffices to consider $p \in \Delta(\Theta)$ with $p_{\theta'} > 0$. Consider the setting of Lemmas 38, 39 and 40 in which $\hat{\nu} := p$, $\nu := \left(\frac{p_\theta}{p_\theta + p_{\theta'}}\right)\delta_\theta + \left(\frac{p_{\theta'}}{p_\theta + p_{\theta'}}\right)\delta_{\theta'}$, $q := \delta_\theta$, and $\hat{p} = p + p_{\theta'}(\delta_\theta - \delta_{\theta'})$. This corresponds to $t := p_{\theta'}$ and $\alpha := p_\theta + p_{\theta'}$; the choice of μ does not matter as long as $\theta, \theta' \notin \text{supp}(\mu)$. (See Figure 6 for illustration.) By the assumed Full Domain UPS form of C and the assumption that

$F(\delta_\theta) = 0$ for all $\theta \in \Theta$, we have

$$C(\sigma^{(0)} \mid p^*) = \left(\frac{p_\theta + p_{\theta'}}{1 + p_\theta + p_{\theta'}}\right) F\left(\left(\frac{p_\theta}{p_\theta + p_{\theta'}}\right)\delta_\theta + \left(\frac{p_{\theta'}}{p_\theta + p_{\theta'}}\right)\delta_{\theta'}\right) + \left(\frac{1}{1 + p_\theta + p_{\theta'}}\right) F\left(p + p_{\theta'}(\delta_\theta - \delta_{\theta'})\right) - F(p^*) \tag{73}$$

$$\begin{aligned} C(\sigma^{(p_{\theta'})} \mid p^*) &= \left(\frac{p_\theta + p_{\theta'}}{1 + p_\theta + p_{\theta'}}\right) F(\delta_\theta) + \left(\frac{1}{1 + p_\theta + p_{\theta'}}\right) F(p) - F(p^*) \\ &= \left(\frac{1}{1 + p_\theta + p_{\theta'}}\right) F(p) - F(p^*) \end{aligned} \tag{74}$$

and by Lemma 40 we have $C(\sigma^{(0)} \mid p^*) = C(\sigma^{(p_{\theta'})} \mid p^*)$. Equating (73) and (74) yields the recursivity condition (R), which proves that F is Recursive. □

Finally, to prove this portion of Theorem 5, it suffices to apply Lemma 41 to the continuous extensions of the cost function:

Proof that (ii) $\implies$ (iii) in Theorem 5. Suppose that $|\Theta| \geq 3$. Let $C : \mathcal{E}_b \times \Delta_\circ \to \mathbb{R}_+$ be a Bounded UPS cost function with potential F Suppose that C is Weakly Compression Invariant. Let $\overline{C}$ and $\overline{F}$ denote the (continuous) extensions of C and F from Lemma 31. Lemma 41 establishes that $\overline{F}$ is Recursive. Then Lemma 33 implies that there exists some $\alpha \in \mathbb{R}$ such that $\overline{F}(q) \equiv \alpha H(p)$ (it must be that $\alpha < 0$ because $\overline{F}$ is convex and H is concave). By definition of $\overline{C}$, we have $\overline{C}\,|_{\mathcal{E}_b \times \Delta_\circ} = C$, implying that C is a Mutual Information cost function, as desired. □

K Auxiliary Results and Proofs

[Additional material to be posted soon. Click here for most recent version.]