Modeling information acquisition via f-divergence and duality*

Alexander W. Bloedel† Tommaso Denti‡ Luciano Pomatto§

October 3, 2025

Abstract

We introduce a new cost function over experiments, *f-information*, based on the theory of multivariate statistical divergences, that generalizes Sims's classic model of rational inattention as well as the class of posterior-separable cost functions. We characterize its behavioral predictions by deriving optimality conditions that extend those of Matějka and McKay (2015) and Caplin, Dean, and Leahy (2019) beyond mutual information. Using these tools, we study the implications of f-information in a number of canonical decision problems. A strength of the framework is that it can be analyzed using familiar methods of microeconomics: convex duality and the Arrow-Pratt approach to expected utility.

*We thank Andrew Caplin, Mark Dean, Ben Hébert, Annie Liang, Elliot Lipnowski, Jay Lu, Massimo Marinacci, Filip Matějka, Stephen Morris, Jeffrey Mensch, Doron Ravid, Ilya Segal, Colin Stewart, Juuso Toikka, Weijie Zhong, and various conference and seminar audiences for helpful comments and discussions.

†UCLA. Email: abloedel@econ.ucla.edu.

‡NYU Stern. Email: td838@stern.nyu.edu.

§Caltech. Email: luciano@caltech.edu.

Contents

1 Introduction

Traditional models of information acquisition depict the decision maker as a statistician who observes a signal from a parametric family of experiments and can increase its precision at a cost. More recent models abandon this structure in favor of a non-parametric formulation, where the agent can select virtually any experiment (i.e. any mapping from states to signal distributions) as an information source. This captures the idea that the agent can fine-tune how they learn about the environment based on the decision problem at hand. Limitations on learning are then represented by an information cost function defined over experiments.

Following Sims (2003), much of the literature has assumed that the cost of information is given by Shannon's *mutual information*, due in large part to its tractability. In this case, as Matějka and McKay (2015) and Caplin, Dean, and Leahy (2019) have shown, optimal behavior resembles standard *multinomial logit* and the information acquisition problem can be solved via a basic variational condition.

Mutual information is a highly specific functional form, and a growing literature has begun to study alternative cost functions (Morris and Strack, 2019; Hébert and Woodford, 2021; Caplin, Dean, and Leahy, 2022; Pomatto, Strack, and Tamuz, 2023; Walker-Jones, 2023; Bloedel and Zhong, 2024, among others). Despite much progress in this direction, extending the analysis beyond mutual information has remained challenging. Unlike utility or production functions, which are defined over familiar economic objects, information costs are defined on the abstract, infinite-dimensional space of experiments, making them inherently harder to specify. Assumptions on learning technologies, which are rarely observed directly, are also more difficult to test. Finally, no other cost function in the literature leads to predictions that have a structure as simple as those of mutual information. For example, the link between mutual information and logit has found no immediate generalizations to these other costs.

In this paper, we introduce a new family of information costs, *f-information*. This family, which is parametrized by a convex function f, encompasses mutual information and many other cost functions in the literature as special cases. Our main result is a characterization of optimal behavior that extends those in Matějka and McKay (2015) and Caplin, Dean, and Leahy (2019) to f-information. Building on this characterization, we identify a number of tractable special cases of the framework, study their implications in a range of decision problems of interest, and relate the predicted behavior to well known models of random choice, such as *additive perturbed utility* (Fudenberg, Iijima, and Strzalecki, 2015) and *nested logit*.

Formally, given a finite set $\Theta = \{\theta_1, \dots, \theta_n\}$ of states, information is acquired by observing the outcome of an experiment $P = (\Omega, (P_\theta)_{\theta \in \Theta})$, where $P_\theta(\omega)$ is the probability of signal realization $\omega \in \Omega$ in state θ. The f-information cost of an experiment P is defined as

$$I_f(P) = \min_{\alpha \in \Delta(\Omega)} \sum_{\omega \in \Omega} \alpha(\omega) f\left(\frac{P_{\theta_1}(\omega)}{\alpha(\omega)}, \dots, \frac{P_{\theta_n}(\omega)}{\alpha(\omega)}\right), \tag{1}$$

where f is a non-negative convex function satisfying $f(1, \dots, 1) = 0$. For a fixed distribution

α over signal realizations, the map f assigns a penalty based on the likelihood ratios between the state-contingent distributions $P_{\theta_1}, \ldots, P_{\theta_n}$ and α. In statistics, this quantity is known as the *f-divergence* between P and α.[1] The cost of P is computed by selecting the measure α for which the average penalty is minimal. We call the solution to the minimization problem (1) the *f-mean* of P. Intuitively, it can be seen as a best approximation of the experiment P.

The notion of f-information formalizes the idea that an experiment P is informative when its state-contingent distributions $P_{\theta_1}, \ldots, P_{\theta_n}$ are far apart, and uninformative when they nearly coincide. When the state-contingent distributions cluster around their f-mean, the experiment conveys little information. When instead they vary across states, the experiment is more costly but also more informative. By varying the transformation f, we obtain a menu of cost functions that remain Blackwell monotone and convex.

We obtain mutual information when

$$f(x) = \sum_{\theta \in \Theta} \pi(\theta)(x(\theta) \log x(\theta) - x(\theta) + 1),$$

where π is the prior belief over states. Another special case of interest is the family of *posterior-separable* costs, introduced by Caplin, Dean, and Leahy (2022) as a generalization of mutual information, which includes most other cost functions that have been proposed in the literature. In all these cases, the f-mean coincides with the unconditional signal distribution $P_\pi = \sum_{\theta \in \Theta} \pi(\theta) P_\theta$.

In the first part of the paper, we characterize the behavioral implications of f-information. We study general decision problems where the agent must choose from a finite set A of actions, and describe the optimal stochastic choice rule $P = (A, (P_\theta)_{\theta \in \Theta})$, where $P_\theta(a)$ is the probability of taking action a in state θ.

To fix ideas, consider first the case of mutual information. As is well known, a stochastic choice rule P is optimal under mutual information if and only if it satisfies two conditions. First, each conditional probability P_θ is related to the unconditional distribution P_π by the modified logit formula

$$P_\theta(a) = \frac{P_\pi(a) \mathrm{e}^{a(\theta)}}{\sum_{b \in A} P_\pi(b) \mathrm{e}^{b(\theta)}}, \tag{2}$$

where $a(\theta)$ is the payoff that action a pays in state θ. Second, the unconditional probability P_π is the solution to an auxiliary concave optimization problem over the set $\Delta(A)$.[2]

For f-information, we obtain a parallel two-step characterization. Central to this result is the function $f^\star$, the *convex conjugate* of the transformation f. We show that a stochastic choice rule P is optimal if and only if each conditional probability P_θ satisfies:

$$P_\theta(a) = \alpha(a) \nabla_\theta f^\star(a\pi - \lambda), \tag{3}$$

[1] See Ali and Silvey (1966), Csiszár (1967), and Duchi, Khosravi, and Ruan (2018).

[2] The characterization has found wide application in models of information acquisition, including studies on labor economics (Acharya and Wee, 2020), optimal pricing (Boyacı and Akçay, 2018), insurance choice (Brown and Jeon, 2024), and industrial organization (Cusumano, Fabbri, and Pieroth, 2024), among many others.

where α is the f-mean of the stochastic choice rule, $\nabla_\theta f^\star$ is the partial derivative of $f^\star$ with respect to state θ, and $\lambda \in \mathbb{R}^\Theta$ is a vector of Lagrange multipliers ensuring that the conditional probabilities P_θ sum to 1. The prior π enters by multiplying the vector $a \in \mathbb{R}^\Theta$ of state-contingent payoffs statewise.

Under mutual information, α equals the unconditional distribution P_π and $\nabla_\theta f^\star(a\pi - \lambda)$ is proportional to $e^{a(\theta)}$, an exponential transformation of the payoff $a(\theta)$. In this case, condition (3) reduces to (2). Condition (3) establishes a more general relation between choice probabilities and the f-mean of P, with $\nabla_\theta f^\star$ replacing the exponential function. The map $\nabla_\theta f^\star$ is increasing, and it may depend on the entire payoff vector a rather than just $a(\theta)$.

In condition (2), solving for P_θ requires determining the endogenous term P_π via an auxiliary optimization problem. In condition (3), it requires solving for the quantities α and λ, which we show are the solutions to an auxiliary saddle-point problem. Once again, this auxiliary problem is of lower dimension than the original information acquisition problem.

A notable feature of our result is that the transformation f appears in (3) not directly, but through its convex conjugate $f^\star$. As in other instances of duality—Marshallian vs. Hicksian demand, cost vs. profit functions, or linear constraints vs. shadow prices—these two objects provide complementary perspectives on the problem. Assumptions stated in terms of f determine how the cost changes as a function of the experiment, whereas assumptions stated in terms of $f^\star$ determine how the primitives of the decision problem (i.e. the prior and action set) translate into choice probabilities. While these two perspectives are ultimately equivalent—there is a one-to-one relation between f and $f^\star$—the optimality condition (3) establishes that properties of the conjugate $f^\star$ are more directly related to behavior.

In the second part of the paper, we focus on a tractable special case of f-information and apply it to a number of canonical decision problems. We consider a specification that is additively separable and symmetric across states:

$$f(x) = \sum_{\theta \in \Theta} \pi(\theta)\phi(x(\theta)),$$

where ϕ is a univariate convex function. This functional form was first studied, in the context of information theory, by Csiszár (1972), and we accordingly refer to it as *Csiszár information*. Compared to the general case, it preserves much of the tractability of mutual information—for example, the Lagrange multiplier can be computed statewise. This is a new class of cost functions that, aside from the special case of mutual information, does not overlap with the family of posterior separable costs studied in most prior work.

In Section 5 we show that the optimal stochastic choice rule under Csiszár information closely resembles additive perturbed utility, a well-known model of discrete choice that generalizes logit (Fudenberg, Iijima, and Strzalecki, 2015). When the decision problem is symmetric, the predictions of the two models coincide. For more general problems, the optimal rule under Csiszár information differs by an endogenous term, α, the f-mean of the choice

rule. We interpret $\alpha(a)$ as the *salience* of action a, and characterize what it means for one action to be more salient than another.

Section 6 shows how the properties of the optimal choice rule can be analyzed through the degree of convexity of the conjugate function $\phi^\star$. To measure this convexity, we draw on tools from risk theory, in particular the Arrow-Pratt coefficient of $\phi^\star$. As we establish, the Arrow-Pratt coefficient measures the decision maker's response to a marginal increase in the stakes of the decision problem. We also connect the Arrow-Pratt coefficient to the ways behavior under Csiszár information can deviate from standard IIA properties.

In Section 7 we study *inconclusive evidence*, i.e. situations where informative and uninformative signals coexist, as in medical tests that yield not only positive or negative results but also inconclusive ones. Although common in practice, such signals cannot be rationalized by models of information acquisition based on mutual information or posterior separability (except for knife-edge cases), leading these models to generate counterfactual predictions (Denti, 2022). In contrast, we show that Csizár information can accommodate this phenomenon.

Section 8 concerns a classic question in psychology, namely, how increasing rewards for accuracy translate into a higher probability of making a correct choice. We analyze a standard task in which the decision maker's objective is to correctly identify the true state, and study how the predicted probability of a correct choice varies with the primitives of the problem. It has been observed that, in perceptual experiments, subjects tend to be less responsive to incentives than the benchmark model based on mutual information predicts (Dean and Neligh, 2023). We show that Csiszár information allows for a much wider range of predictions and demonstrate that properties of the resulting psychometric curve, such as it being S-shaped, can be directly linked to the prudence index of $\phi^\star$, another tool we borrow from risk theory.

In the last part of the paper, we apply the framework of f-information to address a well-known limitation of mutual information: the fact that states enter into the analysis only through the payoff consequences of different actions. This property rules out the possibility that distinguishing between more similar states, whether by physical characteristics or by their proximity, may be costlier. It also leads to unrealistic predictions, such as sharp discontinuities in behavior where smoother adjustments would be expected (e.g., Hébert and Woodford, 2021; Morris and Yang, 2022; Dean and Neligh, 2023; Pomatto, Strack, and Tamuz, 2023).

We propose two families of models, both instances of f-information, that take into account the structure of the state space. The common and central idea is that agents simplify the environment by representing states through a smaller set of attributes, and then acquire information as if attributes were the actual states. This attribute-based framework allows us to introduce interpretable parameters that capture how similarity between states shapes the cost of learning.

The first model, which we call *Perceptual Csiszár information*, extends Csiszár information by explicitly incorporating the decision maker's hardwired limitations in distinguishing between states. Although the resulting cost function admits a richer set of parameters and

loses the additively separable structure of standard Csiszár information, we develop a solution method tailored to this broader class and show that many of the analytical tools used in the separable case remain applicable. We illustrate the model in a canonical one-dimensional discrimination task and show that it yields intuitive sufficient conditions under which the predicted psychometric curve is S-shaped.

The second model, *Nested Shannon entropy*, is a posterior-separable cost function that generalizes mutual information by allowing the modeler to specify which subsets, or *nests*, of states share similar attributes. These costs describe the decision-maker as following an optimal two-step learning process in which they first learn about which nest contains the true state, and then learn about the states within that nest. We relate the resulting behavior to the well-known nested logit model, and show that the cost function connects closely to Hébert and Woodford's (2021) neighborhood-based costs and to Walker-Jones's (2023) multi-attribute Shannon entropy. We apply this model to a multi-dimensional discrimination task and show that it can capture the idea that learning about a multi-dimensional state can be harder than learning about a uni-dimensional one.

1.1 Related literature

Building on Sims' (2003) rational inattention framework and the optimality conditions for mutual information derived by Matějka and McKay (2015) and Caplin, Dean, and Leahy (2019), a burgeoning literature has examined the properties and behavioral implications of information costs (see Maćkowiak, Matějka, and Wiederholt, 2023; Strzalecki, 2025, for surveys). Our paper connects to several strands of this literature, as well as the adjacent literature on discrete choice.

The posterior-separable case. To date, most research on rational inattention has centered on the class of posterior-separable costs introduced by Caplin, Dean, and Leahy (2022). For an experiment $P = (\Omega, (P_\theta)_{\theta\in\Theta})$, these cost functions take the form

$$C(P) = \sum_{\omega\in\text{supp}(P_\pi)} P_\pi(\omega)H(p_\omega),$$

where $P_\pi \in \Delta(\Omega)$ is the unconditional signal distribution, $p_\omega \in \Delta(\Theta)$ is the posterior belief about the state following signal realization ω, $\pi \in \Delta(\Theta)$ is the prior, and H is a convex entropy function assigning a cost to each posterior. By allowing for general entropy functions, this formulation provides an extension of Sims' mutual information cost, which arises when H is proportional to Shannon entropy.[3]

We show that f-information, despite generalizing mutual information in a seemingly distinct way, includes the class of posterior separable costs as a special case. In the posterior-

[3]Applications include mechanism design (e.g., Mensch, 2022; Mensch and Ravid, 2022; Thereze, 2025; Bloedel and Segal, 2025), information design (e.g., Lipnowski, Mathevet, and Wei, 2020; Bloedel and Segal, 2021; Yoder, 2022), and macroeconomics (e.g., Hébert and La'O, 2023; Angeletos and Sastry, 2025).

separable case, the transformation f and its conjugate $f^\star$ can be expressed simply in terms of the entropy H and its conjugate $H^\star$, respectively. As a consequence, our analysis of f-information yields optimality conditions for information acquisition problems with posterior separable costs.

Given the body of work on posterior separable costs, we are obviously not the first to derive such conditions. Indeed, it is well known that optimal behavior under these costs can be characterized via concavification and related Lagrangian methods.[4] Nevertheless, our analysis offers a new perspective by shifting the focus from the entropy H to its conjugate $H^\star$.

To illustrate, fix a posterior-separable cost with entropy H. Concavification yields a primal first-order condition characterizing the optimal posterior p_a at which action a is chosen:

$$a - \lambda_\pi \in \partial H(p_a), \tag{4}$$

where action a is identified with the vector of state-contingent utilities it generates, the subdifferential $\partial H(p_a) \subseteq \mathbb{R}^\Theta$ represents the marginal cost of producing posterior p_a, and $\lambda_\pi \in \mathbb{R}^\Theta$ is a Lagrange multiplier ensuring Bayes plausibility with respect to prior π.[5] Versions of condition (4) appear in Caplin, Dean, and Leahy (2022, Lemma 1), Denti (2022, Lemma 10), Lipnowski and Ravid (2023, Proposition 3), and Bloedel and Segal (2025, Corollary 2), among others.

Condition (4) is particularly useful in revealed preference and mechanism design settings, where the goal is to construct a utility function or an entropy function to rationalize a given distribution of posteriors.[6] However, when the goal is to characterize the optimal behavior in a given decision problem—i.e. to *solve* an information acquisition problem—one must invert condition (4) to determine the posterior p_a as a function of the payoffs and multiplier. When $H^\star$ is differentiable, this inversion yields

$$p_a = \nabla H^\star(a - \lambda_\pi), \tag{5}$$

which is equivalent to the dual first-order condition (3) obtained via our approach.

These observations underscore that, for the purpose of studying the predictions of models of information acquisition, the central object is the conjugate $H^\star$. Indeed, equation (5) shows that what matters is not the tractability of H, but rather that of its conjugate. To appreciate this point, note that there is no guarantee that both H and $H^\star$ have simple closed forms. For instance, the posterior-separable costs in Hébert and Woodford (2021), Pomatto, Strack, and

[4] See, e.g., Gentzkow and Kamenica (2014), Caplin, Dean, and Leahy (2022), Denti (2022), Mensch (2022), Lipnowski and Ravid (2023), Muller-Itten, Armenter, and Stangebye (2024), and Bloedel and Segal (2025).

[5] As we show, this multiplier λ_π is equivalent to the multiplier λ in our optimality condition (3) divided statewise by the prior, i.e. $\lambda_\pi(\theta) = \lambda(\theta)/\pi(\theta)$.

[6] In revealed preference exercises, the distribution of posteriors can be inferred by the analyst from the decision maker's choice behavior (see, e.g., Caplin and Martin, 2015; Caplin and Dean, 2015). In design problems, the designer chooses the distribution to be implemented, subject to incentive compatibility (see, e.g., Mensch, 2022; Yoder, 2022; Bloedel and Segal, 2025).

Tamuz (2023), and Bloedel and Zhong (2024) have simple functional forms but, to the best of our knowledge, their conjugates do not. The family of nested entropies that we introduce in Section 10 offers an example of posterior-separable costs where both H admits a suggestive interpretation and $H^\star$ remains tractable.

Beyond posterior separability. While our analysis yields insights for the familiar posterior-separable case, the class of f-information costs is broader. Prior work has underscored the behavioral limitations of posterior separability (e.g., Denti, 2022), but our interest in nonposterior-separable costs extends beyond these critiques. Most of our applications focus on Csiszár information, a new and tractable family of f-information costs that intersects the posterior-separable class only in the special case of mutual information. Csiszár information is therefore of independent interest, distinct from the limitations of posterior separability.

Our paper thus contributes to a smaller but growing strand of the literature on non-posterior-separable cost functions. We emphasize connections to three lines of related work.

First, several papers develop revealed-preference analyses of costly information acquisition with general cost functions (e.g., Caplin and Dean, 2015; De Oliveira, Denti, Mihm, and Ozbek, 2017).[7] While our main objectives differ, the second part of our paper takes inspiration from this approach by studying the behavioral implications of f-information in canonical decision problems. Focusing on Csiszár information in particular, we provide a behavioral interpretation of the model's parameters, derive identification and comparative statics results, and study various IIA properties. A full revealed-preference characterization of f-information is left for future work.

Second, a number of papers propose non-posterior-separable costs by imposing structural restrictions directly on the cost function. Most closely related are Mu, Pomatto, Strack, and Tamuz (2021, Theorem 2) and Bordoli and Iijima (2025), which relax the linearity axioms of Pomatto, Strack, and Tamuz (2023) to derive costs based on Rényi divergences between state-contingent signal distributions. Although both these costs and f-information build on notions of statistical distance, we are not aware of a simple connection between them. Also related are the *sequential learning-proof* costs of Bloedel and Zhong (2024), which are defined via their robustness to dynamic optimization of the information acquisition process. Clarifying their relation to f-information remains an avenue for future research.

Finally, two recent papers share our interest in deriving optimality conditions for non-posterior-separable costs, albeit from complementary angles. Lipnowski and Ravid (2023) show that a version of the primal first-order condition (4) extends to the class of *iteratively differentiable* costs, which are locally—but not globally—posterior separable.[8] Their approach hinges on smoothness properties of the cost function itself, whereas our derivation of the dual condition (5) relies instead on the differentiability of the conjugate $f^\star$. This, in turn,

[7] See also Ellis (2018), Chambers, Liu, and Rehbeck (2020), Lin (2022), and Lipnowski and Ravid (2023).

[8] In contrast to our approach, Lipnowski and Ravid (2023) impose no functional form assumptions on the cost function aside from iterative differentiability, and also allow for infinite state spaces.

guarantees strict monotonicity of the underlying f-information cost, and hence captures the assumption that there is no free information.

Focusing on the class of sequential learning-proof costs, Muller-Itten, Armenter, and Stangebye (2024) introduce the concept of an *ignorance equivalent*: a vector of state-contingent payoffs that serves as a summary statistic in information acquisition problems and, in some contexts, obviates the need to fully solve for optimal strategies. In the special case of posterior separable costs, the ignorance equivalent collapses to (a normalized version of) the Lagrange multiplier in (4) and (5), which likewise plays an important role in our analysis.

Convex duality in choice theory. Our use of convex duality also connects to the literature on decision-making under uncertainty and discrete choice.

We employ convex conjugacy to analyze a choice model through two complementary representations, one focused on the properties of the information cost (via the transformation f), and another that emphasizes its behavioral implications (via $f^\star$). The use of dual representations has a long tradition in the robustness literature, both in decision theory (Hansen and Sargent, 2001; Maccheroni, Marinacci, and Rustichini, 2006; Strzalecki, 2011) and in robust optimization (Ben-Tal and Ben-Israel, 1991; Ben-Tal and Teboulle, 2007). A similar perspective was brought to rational inattention by De Oliveira, Denti, Mihm, and Ozbek (2017), who study the duality between values and costs in information acquisition problems.

With regard to discrete choice, dual optimality conditions analogous to (5) date back to the Williams-Daly-Zachary Lemma for additive random utility models.[9] Closest to our work is the family of *perturbed utility* models, in which stochastic choice arises from control costs of selecting the correct action. Hofbauer and Sandholm (2002) provide an analogue of (5) for such models. More recently, Fudenberg, Iijima, and Strzalecki (2015) introduce and characterize the *additive perturbed utility* model, where the control cost is separable across actions. We show that the special case of our framework based on Csiszár information is closely related to additive perturbed utility, and several aspects of our analysis are directly inspired by Fudenberg, Iijima, and Strzalecki (2015).[10]

A distinction between our paper and most of the discrete choice literature is that, in the latter, stochasticity in behavior arises for reasons unrelated to information acquisition (e.g., utility shocks or control costs). Fosgerau, Melo, De Palma, and Shum (2020) study the intermediate case in which the decision maker faces a *Bregman information* cost—a cost function over stochastic choice rules defined via a Bregman divergence. Using convex duality, they provide an elegant extension of Matějka and McKay (2015). However, their analysis connects only partially to information acquisition, as Bregman information costs are not generally Blackwell monotone (Cheng and Kim, 2025). Whenever Blackwell monotonicity

[9] See Strzalecki (2025) for a recent treatment.

[10] Flynn and Sastry (2023) extend the additive perturbed utility model to settings with an uncertain state.

fails, these cost functions cannot be interpreted as arising solely from an underlying process of costly information acquisition; instead, they capture other forms of costly stochastic choice.

2 Set up

2.1 Information acquisition problems

We consider the problem of an agent who is faced with a choice under uncertainty and who has the option to obtain costly information before committing to a specific course of action.

Let Θ be a finite set of *states*, and let A denote a finite set of *actions*. A state-dependent Bernoulli *utility function* represents the agent's preferences over actions. For brevity, we identify each action with the corresponding utility profile. We therefore view A as a finite subset of $\mathbb{R}^\Theta$, and normalize the utility function so that $a(\theta) \in \mathbb{R}$ is the utility from action a in state θ. The decision maker's *prior belief* is expressed through a probability distribution $\pi \in \Delta(\Theta)$ with full support.[11] We refer to each pair $\mathcal{D} = (\pi, A)$ as a *decision problem*.

Before taking an action, the agent can acquire additional information about the state. We model the acquisition of information as the choice of an experiment. An *experiment* $P = (\Omega, (P_\theta)_{\theta\in\Theta})$ consists of a finite set of outcomes Ω and a profile $(P_\theta)_{\theta\in\Theta}$ of distributions $P_\theta \in \Delta(\Omega)$ contingent on the state, with the interpretation that the experiment produces outcome $\omega \in \Omega$ with probability $P_\theta(\omega)$ depending on the true state θ. We denote by $P_\pi \in \Delta(\Omega)$ the resulting unconditional outcome distribution defined as $P_\pi(\omega) = \sum_{\theta\in\Theta} \pi(\theta)P_\theta(\omega)$.

We restrict attention to the class $\mathcal{E}$ of experiments with a finite outcome space.[12] Given our focus on decision problems with finite action sets, and the assumption of Blackwell monotonicity we will impose on information costs, the restriction to experiments with finite outcome spaces is without loss of generality and eases the exposition.

The cost of information is represented by a function $C\colon \mathcal{E} \to [0, +\infty]$ where an infinite cost corresponds to an infeasible experiment. Information costs are measured in the same units as the utility function and are additively separable from it. Therefore, conducting an experiment P and then taking an action a in state θ results in a net payoff of $a(\theta) - C(P)$. The value of information arises from the ability to tailor action choices to the realized outcome of the experiment. Given an experiment P with outcome space Ω, an *action strategy* $\sigma = (A, (\sigma_\omega)_{\omega\in\Omega})$ assigns to each possible outcome ω a probability distribution over actions, $\sigma_\omega \in \Delta(A)$.

The decision maker selects an experiment $P = (\Omega, (P_\theta)_{\theta\in\Theta})$ and an action strategy $\sigma = (A, (\sigma_\omega)_{\omega\in\Omega})$ to maximize their expected utility net of information costs:

$$\sum_{\theta\in\Theta} \pi(\theta) \sum_{\omega\in\Omega} P_\theta(\omega) \sum_{a\in A} \sigma_\omega(a)a(\theta) - C(P). \tag{6}$$

We refer to (6) as an *information acquisition problem*.

[11] We denote by $\mathbb{R}^\Theta$ the vector space of real-valued functions on Θ, and by $\Delta(\Theta)$ the set of probability distributions over Θ. Since Θ is finite, $\Delta(\Theta)$ can be identified with a convex subset of $\mathbb{R}^\Theta$.

[12] Note that we refer to $\mathcal{E}$ as a class, rather than a set, because $\mathcal{E}$ does not form a well-defined set (there is no such thing as the set of all finite sets). In doing so, we follow a common convention in set theory.

Prior dependence. We allow the cost function to depend on the prior π, but to ease the exposition we do not make this dependence explicit in the notation. Dependence on the prior enters in the analysis only when the same cost function is applied across decision problems that vary in the prior (as in Sections 5 and 6); this feature is otherwise irrelevant for our paper, where most results treat the prior as fixed. For a discussion of prior dependence, see Denti, Marinacci, and Rustichini (2022) and Bloedel and Zhong (2024).

2.2 Examples

As running examples, we focus on three classes of environments that feature prominently in the literature:

Example 1 (Binary choice)**.** The decision problem involves the choice between a risky action r, whose payoff varies with the state, and a safe action s, which yields a constant payoff of zero in all states.[13] Such decision problems are common in economic applications of rational inattention, including monopoly pricing and production (Ravid, 2020; Fabbri, 2024), coordination games (Yang, 2015; Morris and Yang, 2022; Denti, 2023), contract and information design (Yang, 2020; Bloedel and Segal, 2021; Ambuehl, Ockenfels, and Stewart, 2025).

Example 2 (Guess the state)**.** The action set $A = \{a_\theta : \theta \in \Theta\}$ consists of mutually exclusive bets on the state of nature: each action a_θ yields a winning payoff of $w > 0$ if the true state is θ, and zero otherwise. In experimental economics, *guess-the-state* problems have served as testbeds for models of rational inattention (Caplin, Csaba, Leahy, and Nov, 2020; Dewan and Neligh, 2020; Dean and Neligh, 2023).

Example 3 (Exchangeable actions)**.** Let $A = \{a_1, \dots, a_n\}$ be a set of n distinct actions. The state space has a product structure. The set of states Θ is a finite subset of $\mathbb{R}^n$, and the i-th dimension of the state corresponds to the utility of action i, so that $a_i(\theta) = \theta_i$ for all $\theta \in \Theta$.

The actions are said to be *exchangeable* if for every permutations $\gamma\colon \{1, \dots, n\} \to \{1, \dots, n\}$ and every state $\theta = (\theta_1, \dots, \theta_n)$,

$$\theta_\gamma = \left(\theta_{\gamma(1)}, \dots, \theta_{\gamma(n)}\right) \in \Theta \quad \text{and} \quad \pi(\theta) = \pi(\theta_\gamma).$$

Under this assumption the decision maker sees the actions as ex-ante homogeneous.

2.3 Background on kernels, Blackwell's order, and stochastic choice rules

We consider cost functions that are increasing with respect to Blackwell's informativeness order. To state this standard assumption, we first introduce some additional terminology.

[13] The restriction to a zero-payoff safe action is without loss of generality. Given any binary action set $A = \{a, b\}$, the decision maker's optimal information acquisition in (6) is unchanged if we redefine the action set as $B = \{r, s\}$, where $r(\theta) = a(\theta) - b(\theta)$ and $s(\theta) = 0$ for all $\theta \in \Theta$.

Given two finite sets Ω and Z, a *Markov kernel* $K = (Z, (K_\omega)_{\omega\in\Omega})$ specifies, for every $\omega \in \Omega$, a probability distribution $K_\omega \in \Delta(Z)$ (experiments and action strategies are examples of Markov kernels). We denote by $\Delta(Z)^\Omega$ the set of all Markov kernels which stochastically maps Ω into Z.

A Markov kernel $K \in \Delta(Z)^\Omega$ and a probability distribution $\alpha \in \Delta(\Omega)$ induce a distribution $K \circ \alpha \in \Delta(Z)$, defined for every $z \in Z$ as

$$(K \circ \alpha)(z) = \sum_{\omega\in\Omega} K_\omega(z)\alpha(\omega).$$

An experiment $Q \in \Delta(Z)^\Theta$ is a *garbling* of an experiment $P \in \Delta(\Omega)^\Theta$ if there exists a Markov kernel $K \in \Delta(Z)^\Omega$ such that $Q_\theta = K \circ P_\theta$ for every θ. In this case, we write $Q = K \circ P$. Intuitively, Q is a garbling of P if Q is obtained by compounding the experiment P with noise captured by K.

Definition 1. A cost function C is *Blackwell monotone* if $C(P) \geq C(Q)$ whenever Q is a garbling of P.

When the cost function is Blackwell monotone, it is without loss of generality, in the information acquisition problem (6), to restrict attention to experiments where the outcome space Ω coincides with the set of actions A, and where the action strategy is the identity function (see, e.g., Matějka and McKay, 2015, Corollary 1). Any such experiment $P = (A, (P_\theta)_{\theta\in\Theta})$ describes a *state-dependent stochastic choice rule* (Caplin and Martin, 2015; Caplin and Dean, 2015). In sum, when C is Blackwell monotone, the problem (6) simplifies to

$$\max_{P\in\Delta(A)^\Theta} \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} P_\theta(a)a(\theta) - C(P), \tag{7}$$

and a solution to this problem describes the decision maker's stochastic choice rule. Since $\Delta(A)^\Theta$ is compact, a solution exists provided that the restriction of C to $\Delta(A)^\Theta$ is lower semicontinuous and not identically equal to $+\infty$.

3 f-divergence and f-information

We study information acquisition problems under a new class of cost functions that extend mutual information as well as the more general posterior separable costs. These cost functions are based on a notion of statistical distance between probability distributions known as multivariate f-divergence (Györfi and Nemetz, 1978; García-García and Williamson, 2012; Duchi, Khosravi, and Ruan, 2018).

3.1 Multivariate f-divergences

Let $\mathbb{R}^n_+$ be the non-negative orthant of $\mathbb{R}^n$ and let $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$. We adopt the notation $\overline{\mathbb{R}} = (-\infty, +\infty]$, $\underline{\mathbb{R}} = [-\infty, +\infty)$, and $\overline{\mathbb{R}}_+ = [0, +\infty]$. An f-divergence is indexed by a function

$f\colon \mathbb{R}^n_+ \to \overline{\mathbb{R}}_+$ that is convex, lower semicontinuous, and satisfies $f(\mathbf{1}) = 0$. The effective domain of f, defined as $\operatorname{dom} f$, is the set of vectors $x \in \mathbb{R}^n_+$ such that $f(x) < +\infty$.

Definition 2. Let $P_1, \ldots, P_n$ and α be probability distributions over a finite set Ω. The *f-divergence* between $P_1, \ldots, P_n$ and α is

$$D_f(P_1, \ldots, P_n \| \alpha) = \sum_{\omega \in \Omega} \alpha(\omega) f\left(\frac{P_1(\omega)}{\alpha(\omega)}, \ldots, \frac{P_n(\omega)}{\alpha(\omega)}\right),$$

where we adopt the convention that $0f(\frac{x_1}{0}, \ldots, \frac{x_n}{0}) = \lim_{t \to +\infty} f(y + tx)/t$ for each $x = (x_1, \ldots, x_n) \in \mathbb{R}^n_+$ and any $y \in \operatorname{dom} f$.[14]

For $n = 1$, we obtain the classical notion of f-divergence for pairs of distributions (Ali and Silvey, 1966; Csiszár, 1967): for $\alpha, \beta \in \Delta(\Omega)$,

$$D_f(\beta \| \alpha) = \sum_{\omega \in \Omega} \alpha(\omega) f\left(\frac{\beta(\omega)}{\alpha(\omega)}\right).$$

The quantity $D_f(\beta \| \alpha)$ is a measure of how dissimilar the distributions β and α are. Under this index, two distributions are more dissimilar when their likelihood ratio, weighted by f, is higher in expectation. Binary f-divergences have found applications in many disciplines. In economics—and, specifically, in rational inattention—the most prominent example is *Kullback-Leibler divergence*, obtained by taking $f(t) = t \log t - t + 1$:[15]

$$D_f(\beta \| \alpha) = D_{\mathrm{KL}}(\beta \| \alpha) = \sum_{\omega \in \Omega} \beta(\omega) \log \frac{\beta(\omega)}{\alpha(\omega)}.$$

More generally, a multivariate f-divergences measures the dissimilarity between a collection of distributions $P_1, \ldots, P_n$ and a reference distribution α. As in the binary case, this dissimilarity is measured in terms of a weighted expectation of the likelihood ratios $(P_1/\alpha, \ldots, P_n/\alpha)$. These divergences enjoy several important properties, which generalize known features of binary f-divergences.[16] Next, we list the properties that will be relevant for this paper.

Lemma 1 (Duchi, Khosravi, and Ruan, 2018)**.** *f-divergences satisfy the following properties:*

(i). *For every Markov kernel $K \in \Delta(Z)^\Omega$,*

$$D_f(P_1, \ldots, P_n \| \alpha) \geq D_f(K \circ P_1, \ldots, K \circ P_n \| K \circ \alpha).$$

[14] This convention is standard and guarantees that D_f is lower semicontinuous over $\Delta(\Omega)^{n+1}$. The quantity $\lim_{t \to +\infty} f(y + tx)/t$ is well defined and independent of the choice of y; it is known as the *recession function* of f computed at x. See Rockafellar (1970, Theorem 8.5) and Combettes (2018).

[15] We adopt the conventions that $0 \log \frac{0}{0} = 0$ and $t \log \frac{t}{0} = 0 = +\infty$ for $t > 0$.

[16] The assumption that f takes positive values is without loss. Given a divergence D_f, with f convex but not necessarily non-negative, and a vector $y \in \mathbb{R}^n$, the map defined as $g(x) = f(x) + \sum_{i=1}^n y_i(x_i - 1)$ induces the same divergence, i.e. $D_g = D_f$. By choosing y appropriately, one can ensure that g is non-negative.

(ii). *The function*

$$(P_1, \ldots, P_n, \alpha) \mapsto D_f(P_1, \ldots, P_n \| \alpha)$$

is lower semicontinuous and convex on $\Delta(\Omega)^{n+1}$.

Property (i), also known as the *data processing inequality*, captures the idea that garbling the distributions $P_1, \ldots, P_n$, and α by a common kernel K makes the distributions $P_1, \ldots, P_n$ more similar to α. Property (ii) will allow us to employ tools from convex analysis in conjunction with f-divergences.

3.2 f-information

The next definition is central to the paper. Given an experiment $P \in \Delta(\Omega)^\Theta$ and a distribution $\alpha \in \Delta(\Omega)$, we denote by $D_f(P\|\alpha)$ the f-divergence between $(P_\theta)_{\theta\in\Theta}$ and α.

Definition 3. Let D_f be an f-divergence. The *f-information* of an experiment $P \in \Delta(\Omega)^\Theta$ is

$$I_f(P) = \inf_{\alpha\in\Delta(\Omega)} D_f(P\|\alpha).$$

A distribution $\alpha \in \Delta(\Omega)$ such that $I_f(P) = D_f(P\|\alpha)$ is an *f-mean* of P.

The principle behind f-information is that an experiment is more informative when its state-contingent outcome distributions P_θ are more distinct from one another. The f-mean of an experiment is a probability measure α that minimizes the f-divergence to $(P_\theta)_{\theta\in\Theta}$, and can be interpreted as a generalized average of these distributions. The informativeness of the experiment is then captured by the distance between the P_θ and their f-mean. Heuristically, the closer these distributions are to their f-mean, the closer they are to one another—hence, the less informative the experiment is about the underlying state.

A similar logic can be found in the more familiar definitions of mean and variance. Note that the arithmetic mean of n real numbers $x_1, \ldots, x_n$ is the unique minimizer of the quadratic distance $\sum_{i=1}^n (x_i - y)^2$ over all $y \in \mathbb{R}$. The variance of $x_1, \ldots, x_n$, a measure of how much these numbers differ from one another, is precisely the average quadratic distance from the arithmetic mean.

By varying the function f, we obtain a number of important special cases from statistics and rational inattention.

Example 4 (Mutual information)**.** Shannon's mutual information has been central to applications of rational inattention since Sims (2003) and is a special case of f-information. The quantity

$$I_{\mathrm{S}}(P) = \sum_{\theta\in\Theta} \pi(\theta) D_{\mathrm{KL}}(P_\theta \| P_\pi)$$

is the *mutual information* of the state and the experiment's outcome when their joint distribution is determined by prior π and experiment P. A well-known property of mutual information

is that the f-mean of any experiment P coincides with its unconditional distribution P_π (see, e.g., Steiner, Stewart, and Matějka, 2017):

$$I_{\mathrm{S}}(P) = \min_{\alpha \in \Delta(\Omega)} \sum_{\theta \in \Theta} \pi(\theta) D_{\mathrm{KL}}(P_\theta \| \alpha).$$

Thus, mutual information is a special case of f-information obtained by setting

$$f(x) = \sum_{\theta \in \Theta} \pi(\theta) \left(x(\theta) \log x(\theta) - x(\theta) + 1 \right).$$

Example 5 (Csiszár information)**.** More generally, suppose f is additively separable and takes the form

$$f(x) = \sum_{\theta \in \Theta} \pi(\theta) \phi(x(\theta))$$

where $\phi \colon \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ is a function that is convex, lower semicontinuous, and satisfies $\phi(1) = 0$. In this case, the f-information of an experiment P simplifies as

$$I_f(P) = \inf_{\alpha \in \Delta(\Omega)} \sum_{\theta \in \Theta} \pi(\theta) D_\phi(P_\theta \| \alpha), \tag{8}$$

where D_ϕ is the corresponding divergence defined over pairs of distributions. This special case was first introduced by Csiszár (1972), and for this reason we refer to (8) as *Csiszár information*. The definition of f-information extends Csiszár's notion beyond the additively separable case.[17] The importance of generalizing the additively separable case is illustrated in the next example.

Example 6 (Posterior separable)**.** The concept of f-information encompasses the class of posterior separable costs, which has been the focus of the rational inattention literature thus far. Let $H \colon \Delta(\Theta) \to \overline{\mathbb{R}}_+$ be a function that is convex and lower semicontinuous, with $H(\pi) = 0$. We will refer to H as an *entropy*.[18] For any such H, Caplin, Dean, and Leahy (2022) consider the cost function

$$C_H(P) = \sum_{\omega \in \mathrm{supp}(P_\pi)} P_\pi(\omega) H(p_\omega)$$

where $\mathrm{supp}(P_\pi) \subseteq \Omega$ is the support of the experiment's unconditional distribution P_π, and $p_\omega \in \Delta(\Theta)$ is the posterior following realization ω, given by Bayes' rule as

$$p_\omega(\theta) = P_\theta(\omega) \pi(\theta) / P_\pi(\omega)$$

[17] Csiszár's work was not motivated by information acquisition problems; rather, his primary aim was to develop a generalization of mutual information with desirable properties for statistical applications.

[18] The term *entropy* typically refers to concave functions of probability distributions, with the term *negentropy* reserved for their convex counterparts. For simplicity of exposition, we use *entropy* to refer to the convex case throughout.

for all $\theta \in \Theta$. The cost function C_H is termed *posterior separable*. Under this cost, an experiment is more costly if it induces more variability in the posterior belief, as measured by the expected variation of the entropy H.

For a suitable choice of f, a posterior separable cost function is a special case of f-information. Indeed, consider the transformation

$$f_H(x) = \begin{cases} H(x\pi) & \text{if } \sum_{\theta\in\Theta} x(\theta)\pi(\theta) = 1, \\ +\infty & \text{otherwise,} \end{cases}$$

where $x\pi = (x(\theta)\pi(\theta))_{\theta\in\Theta}$. Note that $D_{f_H}(P\|\alpha) < +\infty$ implies $\alpha = P_\pi$. Thus, $I_{f_H}(P) = D_{f_H}(P\|P_\pi) = C_H(P)$ and P_π is an f_H-mean of P. In general, f_H is not additively separable.

As is well known, mutual information (Example 4) can be represented as a posterior separable cost function by taking $H(p) = D_{\mathrm{KL}}(p\|\pi)$. Notably, the resulting f function differs from the one described in Example 4 above, illustrating that different functions f can generate the same cost function.

Next, we describe a few important properties of f-information that we will use in the analysis of information acquisition problems.

Lemma 2. *f-information has the following properties:*

(i). *I_f is Blackwell monotone.*

(ii). *For every experiment $P \in \Delta(\Omega)^\Theta$ there is $\alpha \in \Delta(\Omega)$ such that $I_f(P) = D_f(P\|\alpha)$.*

(iii). *Given an outcome space Ω, I_f is convex and lower semicontinuous on $\Delta(\Omega)^\Theta$.*

Property (i) is a fundamental requirement for $I_f(P)$ to be interpreted as a measure of the amount of information that P contains. Property (ii) states that each experiment admits an f-mean. Property (iii) will allow us to exploit tools from convex analysis.[19]

4 Optimality conditions

In this section, we characterize solutions and value functions of information acquisition problems in which cost is measured by f-information.

4.1 Information acquisition with mutual information

We first review the standard case in the literature, where the cost is given by mutual information:

$$\max_{P\in\Delta(A)^\Theta} \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} P_\theta(a)a(\theta) - \kappa I_{\mathrm{S}}(P). \tag{9}$$

[19]Given two experiment $P, Q \in \Delta(\Omega)$ and a weight $t \in [0,1]$, their convex combination is defined as $tP + (1-t)Q = (\Omega, (tP_\theta + (1-t)Q_\theta)_{\theta\in\Theta})$. A sequence of experiments (P^n) in $\Delta(\Omega)^\Theta$ converges to P if, for every $\theta \in \Theta$ and $\omega \in \Omega$, the sequence of real numbers $(P^n_\theta(\omega))$ converges to $P_\theta(\omega)$.

Here, $I_S(P)$ is the mutual information between the experiment's outcome and the state (Example 4), and $\kappa > 0$ is a constant that parametrizes the cost.

As discussed in Maćkowiak, Matějka, and Wiederholt (2023), the main appeal of mutual information lies in its tractability.[20] This tractability is well exemplified by the results of Matějka and McKay (2015) and Caplin, Dean, and Leahy (2019), who prove that the maximization problem (9) can be reduced to the simpler auxiliary problem

$$\max_{\alpha \in \Delta(A)} \kappa \sum_{\theta \in \Theta} \pi(\theta) \log \left(\sum_{a \in A} e^{\frac{a(\theta)}{\kappa}} \alpha(a) \right). \tag{10}$$

This is a lower-dimensional problem that involves maximization on unconditional distributions over actions, rather than maximization over experiments.

Theorem 1 (Matějka and McKay, 2015; Caplin, Dean, and Leahy, 2019)**.** *Information acquisition under mutual information has the following properties:*

(i). *A stochastic choice rule $P = (A, (P_\theta)_{\theta \in \Theta})$ is a solution to (9) if and only if there exists a solution $\alpha \in \Delta(A)$ of (10) such that for all $\theta \in \Theta$ and $a \in A$,*

$$P_\theta(a) = \frac{\alpha(a) e^{\frac{a(\theta)}{\kappa}}}{\sum_{b \in B} \alpha(b) e^{\frac{b(\theta)}{\kappa}}} \tag{11}$$

Moreover, for any such P and α, it holds that $\alpha = P_\pi$.

(ii). *The optimization problems (9) and (10) have the same value.*

This result describes a two-step recipe to solve information acquisition problems under mutual information. The first step is to find all distributions over actions $\alpha \in \Delta(A)$ that solve the auxiliary optimization problem (10). Then, from each such α, optimal choice rules can be derived mechanically from the formula (11). While the first step yields closed-form solutions only in specific settings—for instance, the unconditional distribution P_π is uniform in exchangeable decision problems (Example 3)—the auxiliary problem can be efficiently solved numerically using, e.g., the Blahut-Arimoto algorithm (Cover and Thomas, 2006).[21]

4.2 Duality

To study the behavioral implications of f-information, we associate to the transformation f a new object that is dual to it.

Definition 4. The *Fenchel conjugate* of f is the function $f^\star \colon \mathbb{R}^\Theta \to \overline{\mathbb{R}}$ defined by

$$f^\star(x) = \sup_{y \in \mathbb{R}^\Theta_+} \sum_{\theta \in \Theta} x(\theta) y(\theta) - f(y).$$

[20] A few papers propose axiomatic motivations: among others, de Oliveira (2019), Mensch (2021), Caplin, Dean, and Leahy (2022), Cerreia-Vioglio, Maccheroni, Marinacci, and Rustichini (2023).

[21] See Armenter, Muller-Itten, and Stangebye (2024) for an alternative computational approach based on the observation that the objective function in the auxiliary problem (10) is concave.

Conjugation is one of the fundamental operations in convex analysis, with applications across different disciplines. In economics, conjugation appears most directly in the model of a competitive firm, where a firm's profit function is the Fenchel conjugate of the cost function.

The next lemma describes the properties of the Fenchel conjugate of the transformation f.

Lemma 3. *For a function $g\colon \mathbb{R}^\Theta \to \overline{\mathbb{R}}$, the following are equivalent:*

(i) *$g = f^\star$ for some f-information I_f;*

(ii) *g is convex, lower semi-continuous, and monotone. Moreover, $g(0) = 0$ and $1 \in \partial g(0)$.*

Given a function g that satisfies the conditions in (ii), the corresponding transformation f can be recovered as

$$f(x) = g^\star(x) = \sup_{y \in \mathbb{R}^\Theta} \sum_{\theta \in \Theta} x(\theta)y(\theta) - g(y).$$

Among these conditions, the monotonicity of $f^\star$ follows from f being defined on the non-negative orthant. The last property of $f^\star$, i.e. $g(0) = 0$ and $1 \in \partial g(0)$, is dual to the condition that f is non-negative and satisfies $f(1, \dots, 1) = 0$. All these results on Fenchel conjugates are standard (Rockafellar, 1970).

The result suggests two equivalent perspectives from which to study information acquisition problems, depending on whether one treats f or $f^\star$ as the main object of analysis. While the transformation f has a direct interpretation in terms of the cost of information, it will turn out to be mathematically and conceptually simpler to describe the resulting optimal behavior in terms of the conjugate $f^\star$.

Next, we illustrate the operation of conjugation in the context of our running examples:

Example 4 (continued)**.** Mutual information corresponds to the transformation $f(x) = \sum_{\theta \in \Theta} \pi(\theta) \left(x(\theta) \log x(\theta) - x(\theta) + 1\right)$. Direct computation show that the conjugate is

$$f^\star(x) = \sum_{\theta \in \Theta} \pi(\theta) e^{\frac{x(\theta)}{\pi(\theta)}} - 1.$$

Example 5 (continued)**.** In the case of Csiszár information, where $f(x) = \sum_{\theta \in \Theta} \pi(\theta)\phi(x(\theta))$, the conjugate of f can be expressed in terms of the conjugate of ϕ, the function $\phi^\star\colon \mathbb{R} \to \overline{\mathbb{R}}$ defined as $\phi^\star(t) = \sup_{s \in \mathbb{R}_+} ts - \phi(s)$. The conjugate of f is then given by

$$f^\star(x) = \sum_{\theta \in \Theta} \pi(\theta)\phi^\star\left(\frac{x(\theta)}{\pi(\theta)}\right).$$

Example 6 (continued)**.** Given a posterior separable cost, the conjugate of the transformation f_H can be expressed in terms of the conjugate of the entropy H. The conjugate of H is the map $H^\star\colon \mathbb{R}^\Theta \to \overline{\mathbb{R}}$ given by

$$H^\star(x) = \max_{p \in \Delta(\Theta)} \sum_{\theta \in \Theta} x(\theta)p(\theta) - H(p).$$

The conjugate of the transformation f_H is then

$$f_H^\star(x) = H^\star\left(\frac{x}{\pi}\right),$$

where the ratio x/π is intended statewise, i.e., $x/\pi = (x(\theta)/\pi(\theta))_{\theta\in\Theta}$. Since H is defined on the simplex, $f_H^\star$ is *translation invariant* with respect to the prior:

$$f_H^\star(x + c\pi) = f_H^\star(x) + c \tag{12}$$

for every constant $c \in \mathbb{R}$. Conversely, given any f-information cost, the conjugate $f^\star$ satisfies this translation invariance property only if $f = f_H$ for some entropy H.

4.3 Assumptions on f

Throughout the paper, we focus on functions f that satisfy the following assumption, which ensure that the associated conjugate is particularly tractable:

Assumption 1. *The function f satisfies:*

- *f is* co-finite*: $\lim_{t\to+\infty} f(y+tx)/t = +\infty$ for every $y \in \operatorname{dom} f$ and all non-zero $x \in \mathbb{R}_+^\Theta$.*
- *f is* essentially strictly convex*: f is strictly convex on every convex subset of $\{x \in \mathbb{R}_+^\Theta : \partial f(x) \neq \varnothing\}$.*
- $\mathbf{1}$ *belongs to the relative interior of* $\operatorname{dom} f$*.*

Under the first two assumptions, the conjugate function $f^\star$ is everywhere finite (i.e., $\operatorname{dom} f^\star = \mathbb{R}^\Theta$), and differentiable. Moreover, being $f^\star$ convex and differentiable, its gradient $\nabla f^\star$ is automatically continuous. Conversely, if a function $g : \mathbb{R}^\Theta \to \mathbb{R}$ is convex, monotone, and differentiable, then its conjugate $g^\star \colon \mathbb{R}_+^\Theta \to \overline{\mathbb{R}}$ is co-finite and essentially strictly convex. See Rockafellar (1970, Corollary 13.3.1 and Theorem 26.3).

In the case of mutual information, f is co-finite and essentially strictly convex. For Csiszár information, if ϕ is co-finite and strictly convex on its effective domain, then the corresponding transformation f is co-finite and essentially strictly convex. In the posterior-separable case, f_H is automatically co-finite; if the entropy function H is essentially strictly convex, then f_H is essentially strictly convex.

The final assumption that $\mathbf{1}$ lies in the relative interior of $\operatorname{dom} f$ will serve as a constraint qualification in our main theorem. In the more familiar posterior-separable case, this condition holds whenever H is finite in a neighborhood of the prior.

4.4 Characterization theorem

We now characterize the solutions and values of information acquisition problems under f-information. Mirroring the work of Matějka and McKay (2015) and Caplin, Dean, and

Leahy (2019) on mutual information, the key step in our analysis is to show that every optimization

$$\max_{P\in\Delta(A)^\Theta} \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} P_\theta(a)a(\theta) - I_f(P) \tag{13}$$

can be reduced to an auxiliary, lower-dimensional problem. This is now a maxmin optimization problem that takes the form

$$\max_{\alpha\in\Delta(A)} \min_{\lambda\in\mathbb{R}^\Theta} \sum_{a\in A} \alpha(a) f^\star(a\pi - \lambda) + \sum_{\theta\in\Theta} \lambda(\theta) \tag{14}$$

where the product $a\pi$ is intended statewise, i.e., $a\pi = (a(\theta)\pi(\theta))_{\theta\in\Theta}$.

Theorem 2. *Information acquisition under f-information has the following properties:*

(i). *A stochastic choice rule $P = (A, (P_\theta)_{\theta\in\Theta})$ is a solution to (13) if and only if there exists a saddle point (α, λ) of (14) such that*

$$P_\theta(a) = \alpha(a)\nabla_\theta f^\star(a\pi - \lambda)$$

for all $\theta \in \Theta$ and $a \in A$. Moreover, for any such P and (α, λ), α is an f-mean of P.

(ii). *The optimization problem (13) and the maxmin problem (14) have the same value.*

Condition (i) of Theorem 2 shows that the ratio of the choice probabilities of actions a and b in state θ takes the form

$$\frac{P_\theta(a)}{P_\theta(b)} = \frac{\alpha(a)}{\alpha(b)} \frac{\nabla_\theta f^\star(a\pi - \lambda)}{\nabla_\theta f^\star(b\pi - \lambda)}.$$

Similar to the case of mutual information, this expression is the product of two ratios. The first term $\alpha(a)/\alpha(b)$ pertains to the f-mean probabilities of a and b. The second ratio involves an increasing function $\nabla_\theta f^\star$ of the utility profiles a and b, scaled by the prior π and shifted by a vector λ that depends on the decision problem at hand. As we show in the proof, λ is in fact the Lagrange multiplier associated to the constraints $\sum_{a\in A} P_\theta(a) = 1$, for $\theta \in \Theta$.

Beyond characterizing optimal choice probabilities, Theorem 2 clarifies the significance of the conjugate $f^\star$. The map $\nabla f^\star$ maps each utility vector a—adjusted for the prior and the Lagrange multiplier—into vectors of likelihood ratios P_θ/α, succinctly capturing the behavioral implications of the f-information cost function.

The f-mean α and the Lagrange multiplier λ are characterized as a saddle point of the maxmin problem (14). Since the objective function in this problem is affine in α and convex in λ, it follows that a pair (α, λ) is a saddle point if and only if it satisfies the first-order conditions

$$f^\star(a\pi - \lambda) = \max_{b\in A} f^\star(b\pi - \lambda) \qquad \forall a \in \operatorname{supp}(\alpha), \tag{15}$$

$$\sum_{a\in A} \alpha(a)\nabla_\theta f^\star(a\pi - \lambda) = 1 \qquad \forall \theta \in \Theta. \tag{16}$$

Condition (15) disciplines the *consideration set* $\{a \in A : P_\pi(a) > 0\}$ since $\alpha(a) = 0$ implies $P_\pi(a) = 0$. Condition (16) ensures that

$$\sum_{a \in A} P_\theta(a) = \sum_{a \in A} \alpha(a) \nabla_\theta f^\star(a\pi - \lambda) = 1.$$

The multiplier $\lambda(\theta)$ can be viewed as the shadow price of acting in state θ, since relaxing the associated constraint would allow the total mass $\sum_{a \in A} P_\theta(a)$ to deviate from one. Therefore, $\lambda(\theta)$ can be seen as a measure of how profitable is for the decision maker to act in state θ.

Theorem 2 suggests a two-step approach to solve information acquisition problems. The first step, which involves identifying the saddle points of the maxmin problem (14), results in a closed-form solution only in specific cases. However, this problem can be solved efficiently using numerical methods, such as the Saddle-Point Mirror Prox algorithm (Bubeck, 2015, pp. 315–316).[22] The second step is to compute the conditional choice probabilities from the formula $P_\theta(a) = \alpha(a) \nabla_\theta f^\star(a\pi - \lambda)$.

4.5 Uniqueness

An inspection of the proof of Theorem 2 shows that any Lagrange multiplier λ can generate any optimal choice rule P:

Corollary 1. *For each saddle point $(\hat{\alpha}, \hat{\lambda})$ of (14) and each solution P to (13), there exists an action distribution α such that $(\alpha, \hat{\lambda})$ forms a saddle point of (14) and P can be expressed as*

$$P_\theta(a) = \alpha(a) \nabla_\theta f^\star(a\pi - \hat{\lambda}).$$

Consequently, whenever an action a is included in the consideration set, the corresponding *revealed posterior* (Caplin and Martin, 2015; Caplin and Dean, 2015) is uniquely determined:

$$p_a(\theta) = \frac{\pi(\theta) P_\theta(a)}{P_\pi(a)} = \frac{\pi(\theta) \nabla_\theta f^\star\left(a\pi - \hat{\lambda}\right)}{\sum_{\tau \in \Theta} \pi(\tau) \nabla_\tau f^\star\left(a\pi - \hat{\lambda}\right)}.$$

By standard arguments (see, e.g., Rockafellar 1970, Corollary 37.5.3), the saddle points of (14) constitutes a closed convex product set in $\Delta(A) \times \mathbb{R}^\Theta$. Therefore, the set of optimal choice rules can be identified with a closed convex subset of $\Delta(A)$, as in the case of mutual information.

4.6 Mutual Information

With mutual information, the f-mean of an experiment is the unconditional signal distribution, making $\alpha = P_\pi$. Furthermore, $\nabla_\theta f^\star(x) = e^{\frac{x(\theta)}{\pi(\theta)}}$. Thus, we have:

$$\frac{P_\theta(a)}{P_\theta(b)} = \frac{\alpha(a)}{\alpha(b)} \frac{\nabla_\theta f^\star(a\pi - \lambda)}{\nabla_\theta f^\star(a\pi - \lambda)} = \frac{P_\pi(a)}{P_\pi(b)} \frac{e^{a(\theta) - \frac{\lambda(\theta)}{\pi(\theta)}}}{e^{b(\theta) - \frac{\lambda(\theta)}{\pi(\theta)}}} = \frac{P_\pi(a)}{P_\pi(b)} \frac{e^{a(\theta)}}{e^{b(\theta)}}.$$

[22] For numerical computations, it is often convenient to bound the search domain of the Lagrange multiplier; we explain how to do this properly in Appendix A.

This is the same expression derived from condition (i) of Theorem 1, taking $\kappa = 1$. Thus, under mutual information, after accounting for the unconditional choice probabilities, a form of independence of irrelevant alternatives holds: the ratio of the choice probabilities of a and b in state θ depends solely on the utility difference between a and b in that state. Cost functions based on f-information, however, allow us to describe a broader range of behavior. The ratio $\nabla_\theta f^\star(a\pi - \lambda)/\nabla_\theta f^\star(b\pi - \lambda)$ potentially depends on the utilities of actions a and b in all states and, through λ, on what other actions are available.

In the case of mutual information, the Lagrange multiplier λ can be computed in closed form for each fixed α. Given that $\nabla_\theta f^\star(x) = e^{\frac{x(\theta)}{\pi(\theta)}}$, it follows from (16) that

$$\sum_{a\in A} \alpha(a) e^{a(\theta) - \frac{\lambda(\theta)}{\pi(\theta)}} = 1.$$

Simple algebra demonstrates that

$$\lambda(\theta) = \pi(\theta) \log \sum_{a\in A} \alpha(a) e^{a(\theta)}.$$

Thus, with mutual information, $\lambda(\theta)$ is a weighted average of the utility the available actions deliver in state θ. This serves as a measure of the desirability of acting in state θ.[23] By plugging the expression for the Lagrange multiplier into the maxmin problem (14) and using the fact that $f^\star(x) = \sum_{\theta\in\Theta} \pi(\theta) e^{\frac{x(\theta)}{\pi(\theta)}} - 1$, we obtain

$$\max_{\alpha\in\Delta(A)} \min_{\lambda\in\mathbb{R}^\Theta} \sum_{a\in A} \alpha(a) f^\star(a\pi - \lambda) + \sum_{\theta\in\Theta} \lambda(\theta) = \max_{\alpha\in\Delta(A)} \sum_{\theta\in\Theta} \pi(\theta) \log \sum_{a\in A} \alpha(a) e^{a(\theta)}.$$

This is the auxiliary maximization problem in Theorem 1, taking $\kappa = 1$.

In summary, a distinctive feature of mutual information is that the Lagrange multiplier can be found analytically, allowing the focus to be exclusively on finding α. As we will see in the analysis of Csiszár information, this is a distinctive but not a unique feature of mutual information.

4.7 Posterior-separable costs

For a general posterior separable cost, the stochastic choice rule takes the form:

$$P_\theta(a) = \frac{P_\pi(a) \nabla_\theta H^\star(a - \lambda/\pi)}{\pi(\theta)}$$

where, as before, $\lambda/\pi = (\lambda(\theta)/\pi(\theta))_{\theta\in\Theta}$. We obtain that the posterior following action a is given by

$$p_a(\theta) = \nabla_\theta H^\star(a - \lambda/\pi).$$

[23] In particular, the Lagrange multiplier λ coincides with what Muller-Itten, Armenter, and Stangebye (2024) call *ignorance equivalent*.

This expression gives special meaning to the gradient $\nabla H^\star$. This is a function mapping the utility vector of each action a, modified by the multiplier λ, into the posterior belief conditional on a being chosen. Therefore, assumptions on the conjugate of H translate directly into assumptions on posterior beliefs, and thus the decision maker's behavior. As discussed in 1.1, this optimality condition is dual to the more standard primal FOC arising from concavification.

4.8 Symmetric decision problems

Symmetry assumptions are often used to construct illustrative examples and simplify the analysis of applications. Under such assumptions, the solutions to information acquisition problems based on f-information inherit the symmetries of the underlying primitives, as we now explain.

We formalize symmetry through invariance with respect to a group of permutations Γ of the state space. Specifically, Γ is a set of bijective functions $\gamma\colon \Theta \to \Theta$ with the following properties: the composition of any two elements of Γ belongs to Γ, and the inverse of any element of Γ belongs to Γ as well. For each $x \in \mathbb{R}^\Theta$ and $\gamma \in \Gamma$, $x_\gamma \in \mathbb{R}^\Theta$ stands for the permuted vector $x_\gamma(\theta) = x(\gamma(\theta))$.

A decision problem $\mathcal{D} = (\pi, A)$ is said to be *invariant* with respect to a group Γ if $\pi_\gamma = \pi$ and $\{a_\gamma : a \in A\} = A$ for all $\gamma \in \Gamma$. A simple example arises when π is uniform and A is the set of bets that pay 1 in one state and 0 otherwise (Example 2). This decision problem is invariant under all bijections γ. The applications in the following sections will introduce further examples of decision problems with various forms of symmetry, including environments with exchangeable actions (Example 3).

A function $f : \mathbb{R}^\Theta_+ \to \overline{\mathbb{R}}$ is said to be *invariant* with respect to a group Γ if $f(x_\gamma) = f(x)$ for all $x \in \mathbb{R}^\Theta_+$ and $\gamma \in \Gamma$. An example is provided by Csiszár information (Example 5), whose associated transformation f is invariant under any permutation γ for which the prior is invariant.

Proposition 1. *Consider a decision problem $\mathcal{D}$ and a transformation f that are invariant with respect to a group Γ of permutations of the state space. Then, the maxmin problem (14) has an invariant saddle point (α, λ), meaning that:*

(i). $\alpha(a_\gamma) = \alpha(a)$ *for all* $a \in A$ *and* $\gamma \in \Gamma$.

(ii). $\lambda_\gamma = \lambda$ *for all* $\gamma \in \Gamma$.

The resulting choice rule has the following symmetry property: $P_{\gamma(\theta)}(a) = P_\theta(a_\gamma)$ *for all* $\theta \in \Theta$, $a \in A$, *and* $\gamma \in \Gamma$.

For instance, in the case in which Γ is the full group of permutations of the state space, α is uniform and λ a is constant vector. We will refer to choice rules that satisfy the property $P_{\gamma(\theta)}(a) = P_\theta(a_\gamma)$ simply as *symmetric*.

Pivotal to Proposition 1 is the following lemma, which shows that the conjugate $f^\star$ inherits the symmetry properties of f:

Lemma 4. *A transformation f is invariant under a permutation γ if and only if its conjugate $f^\star$ also is invariant, meaning that $f^\star(x_\gamma) = f^\star(x)$ for all $x \in \mathbb{R}^\Theta$. Moreover, $\nabla_{\gamma(\theta)} f^\star(x) = \nabla_\theta f^\star(x_\gamma)$ for all $\theta \in \Theta$.*

4.9 Essential smoothness

For some applications, we will study transformations f that are essentially smooth. This is an additional regularity condition which ensures that the Lagrange multiplier λ is unique and that the f-mean α and the predictive distribution P_π are mutually absolutely continuous.

Formally, a transformation f is *essentially smooth* if it satisfies the following properties: (i) $\operatorname{int}(\operatorname{dom} f)$ is not empty, (ii) f is differentiable on $\operatorname{int}(\operatorname{dom} f)$, and (iii) $\lim_{n\to+\infty} \|\nabla f(x_n)\| = +\infty$ whenever (x_n) is a sequence in $\operatorname{int}(\operatorname{dom} f)$ converging to a boundary point of $\operatorname{dom} f$.[24] These properties amount to a condition on the marginal cost of information:

Example 5 (continued)**.** In the case of a Csiszár cost based on a univariate transformation ϕ, the associated function f is essentially smooth if and only if ϕ is essentially smooth. A sufficient condition for ϕ to be essentially smooth is that ϕ is finite and differentiable on $(0, +\infty)$, and the derivative ϕ' is unbounded below—as in the case with mutual information (Example 4). This ensures that as the likelihood ratio P_θ/α in some state θ converges to 0, the marginal cost of further lowering the likelihood ratio diverges to infinity.

It is a classic result in convex analysis that f is essentially smooth if and only if its conjugate $f^\star$ is strictly convex (Rockafellar, 1970, Theorem 26.3). We collect other properties that will be helpful in the next sections.

Lemma 5. *Let f be essentially smooth. Then:*

(i) *$f^\star$ is strictly increasing.*

(ii) *If (α_1, λ_1) and (α_2, λ_2) are two saddle points of the maxmin problem (14), then $\lambda_1 = \lambda_2$.*

Strict monotonicity of $f^\star$ implies that the optimal choice rule P and its f-mean are mutually absolutely continuous: for all $\theta \in \Theta$ and $a \in A$, $P_\theta(a) > 0$ if and only $\alpha(a) > 0$. The second notable implication of essential smoothness is uniqueness of the multiplier.

Next we adapt these notions to the case in which costs are posterior separable. Since $H^\star$ is translation invariant, it cannot be strictly convex everywhere. We therefore introduce a minimal relaxation of strict convexity. We say that $H^\star$ is *strictly convex modulo translations* if for all $t \in (0, 1)$ and $x, y \in \mathbb{R}^\Theta$ such that $x \notin y + \mathbb{R}$,

$$H^\star(tx + (1-t)y) > tH^\star(x) + (1-t)H^\star(y).$$

[24] The notation $\operatorname{int}(\operatorname{dom} f)$ stands for the topological interior of the effective domain of f.

To characterize the dual property, we fix an enumeration of the state space, $\Theta = \{\theta_1, \ldots, \theta_n\}$, and denote by H_{n-1} the function

$$(p_1, \ldots, p_{n-1}) \mapsto H(p_1, \ldots, p_{n-1}, 1 - p_1 - \ldots - p_{n-1}).$$

Lemma 6. *The following statements are equivalent:*

(i). $H^\star$ *is strictly convex modulo translations.*

(ii). H_{n-1} *is essentially smooth.*

Motivated by Lemma 6, we say that H is *relatively smooth* if the function H_{n-1} is essentially smooth. Note that, in this definition, the specific enumeration of the state space is inconsequential. The next result shows that relatively smooth entropies share the same properties of essentially smooth transformations.

Lemma 7. *Let H be a relatively smooth entropy. Then:*

(i) $H^\star$ *is strictly increasing.*

(ii) *If (α_1, λ_1) and (α_2, λ_2) are two saddle points of the maximin problem (14), with prior $\pi \in \operatorname{ri}(\operatorname{dom} H)$), then $\lambda_1 \in \lambda_2 + \mathbb{R}$.*

5 Csiszár information and discrete choice

We now focus on Csiszár information, which is additively separable and symmetric across states. Its advantage, compared to the general case of f-information, is that its properties depend on a univariate rather than multivariate transformation. It encompasses mutual information cost as a special case and serves as a benchmark specification for the applications that follow.

In the next two sections, we establish structural properties of Csiszár information. We show it provides a new foundation for the perturbed utility model of discrete choice (Fudenberg, Iijima, and Strzalecki, 2015), and that mutual information is essentially the unique Csiszár cost that is also posterior separable.

5.1 Preliminaries

We assume that in the transformation

$$f(x) = \sum_{\theta \in \Theta} \pi(\theta) \phi(x(\theta)) \tag{17}$$

the map ϕ satisfies the following properties:

Assumption 2. *The map $\phi \colon \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ is strictly convex on its effective domain, is lower semicontinuous, and satisfies the conditions $\phi(1) = 0$, $1 \in \operatorname{ri}(\operatorname{dom} \phi)$, and $\lim_{t \to \infty} \frac{1}{t} \phi(t) = +\infty$.*

These assumptions guarantee that f satisfies the normalization $f(\mathbf{1}) = 0$ as well as the conditions in Assumption 1.

5.2 Optimality conditions

For the case of Csiszár information, the optimality conditions in Theorem 2 take a simple form. In particular, the optimal stochastic choice rule and Langrange multipliers can be determined state-by-state.

For brevity, from now on we denote by $\psi = \phi^\star$ the conjugate of ϕ. It is easy to see that $\psi\colon \mathbb{R} \to \mathbb{R}$ is increasing, convex, and differentiable, with $\psi(0) = 0$ and $\psi'(0) = 1$. In some instances, it will be convenient to assume that ψ is strictly convex, which corresponds to ϕ being essentially smooth (Section 4.9).

The conjugate of the state-separable transformation f defined in (17) is then given by

$$f^\star(x) = \sum_{\theta\in\Theta} \pi(\theta)\psi\left(\frac{x(\theta)}{\pi(\theta)}\right).$$

For convenience, we will work with the *prior-adjusted* Lagrange multiplier $\lambda_\pi \in \mathbb{R}^\Theta$ defined statewise as $\lambda_\pi(\theta) = \lambda(\theta)/\pi(\theta)$.[25]

Applying Theorem 2, the optimal stochastic choice rule is then given by

$$P_\theta(a) = \alpha(a)\psi'\left(a(\theta) - \lambda_\pi(\theta)\right), \tag{18}$$

while the optimality condition for the Lagrange multiplier given by (16) simplifies to

$$\sum_{a\in A} \alpha(a)\psi'\left(a(\theta) - \lambda_\pi(\theta)\right) = 1. \tag{19}$$

In words, (18) states that the probability of taking action a in state θ is the product of two terms: a baseline probability $\alpha(a)$ that is independent of the state, and an increasing function of the payoff $a(\theta)$ that a yields in state θ, minus the multiplier $\lambda_\pi(\theta)$. Moreover, taking α as given, we can determine $\lambda_\pi(\theta)$ as the solution of (19) without accounting for the multiplier $\lambda_\pi(\tau)$ or payoffs $\{a(\tau) : a \in A\}$ in any other states $\tau \neq \theta$.[26] In particular, since ψ' is increasing, we can interpret (19) as stating that $\lambda_\pi(\theta)$ represents a weighted average, under the probability distribution $\alpha \in \Delta(A)$, of the feasible payoffs $\{a(\theta) : a \in A\}$ in state θ.

As discussed in Section 4.6, for the special case of mutual information, we can solve (19) for the Lagrange multiplier in closed-form as a function of α, thereby reducing the saddle-point problem from Theorem 2 to the auxiliary maximization problem from Theorem 1. While such closed-form solutions are not always available, they can indeed be obtained in some other special cases of interest. The next example illustrates for the case in which ϕ is quadratic:

Example 6 (Chi-squared divergence)**.** Let $\phi(t) = \kappa(t-1)^2/2$ for all $t \in \mathbb{R}_+$, where $\kappa > 0$ is a constant. The corresponding ϕ-divergence is known as the *chi-squared divergence*.

[25] The multiplier λ_π represents the shadow cost of the constraint $\sum_{a\in A} \pi(\theta)P_\theta(a) = \pi(\theta)$ for every $\theta \in \Theta$, i.e., the joint state-action distribution must induce a marginal distribution over states equal to the prior, π.

[26] If ϕ is essentially smooth (i.e., ψ is strictly convex), then for each α there is a unique $\lambda_\pi(\theta)$ solving (19) since ψ' is strictly increasing.

In this case, the conjugate function $\psi = \phi^\star$ is given by $\psi(t) = \max\left\{t^2/(2\kappa) + t, -\kappa/2\right\}$ for all $t \in \mathbb{R}$, and its derivative is $\psi'(t) = \max\{t/\kappa + 1, 0\}$ for all $t \in \mathbb{R}$. Therefore, (19) reduces to

$$\sum_{a \in A} \alpha(a) \max\left\{a(\theta) - \lambda_\pi(\theta) + \kappa, 0\right\} = \kappa. \tag{20}$$

To solve this equation for $\lambda_\pi(\theta)$ as a function of α, it is convenient to rank the actions in the support of α in descending order of their payoffs in state θ. That is, we enumerate the consideration set as $\text{supp}(\alpha) = \{a_1, \ldots, a_n\}$ such that $a_1(\theta) \geq \cdots \geq a_n(\theta)$.[27] As we show in Appendix C.5.1, the unique solution to (20) can then be expressed as

$$\lambda_\pi(\theta) = \sum_{j=1}^{i^*(\theta)} \left(\frac{\alpha(a_j)}{\sum_{k=1}^{i^*(\theta)} \alpha(a_k)} \right) a_j(\theta) - \frac{\kappa}{\sum_{j=1}^{i^*(\theta)} \alpha(a_j)} + \kappa,$$

where the cutoff index $i^*(\theta) \in [n] := \{1, \ldots, n\}$ is given by

$$i^*(\theta) = \max\left\{ i \in [n] : \sum_{j=1}^{i} \alpha(a_j)\left(a_j(\theta) - a_i(\theta)\right) < \kappa \right\} = \max\left\{i \in [n] : a_i(\theta) > \lambda_\pi(\theta) - \kappa\right\}.$$

To interpret these expressions, notice that $\text{supp}(P_\theta) = \{a_i \in A : i \leq i^*(\theta)\}$, i.e., action a_i is *considered in state* θ if and only if $i \leq i^*(\theta)$. Therefore, $\lambda_\pi(\theta)$ represents an average of the payoffs to actions that are considered in state θ. For instance, if $\text{supp}(P_\theta) = \text{supp}(\alpha)$, then $\lambda_\pi(\theta) = \sum_{i=1}^n \alpha(a_i) a_i(\theta)$ is precisely the expected payoff in state θ under the distribution α.

5.3 Behavioral characterization of α and λ

Under Csiszár information, the saddle point (α, λ) can be given a transparent characterization in terms of the induced behavior. To this end, we begin by introducing two orderings—the first over states, the second over actions—defined by a stochastic choice rule.

Definition 5. Let $\mathcal{D} = (\pi, A)$ be a decision problem and $P = (A, (P_\theta)_{\theta \in \Theta})$ a choice rule. We say that choice is *bolder* in state θ than in state τ if, for every action $a \in A$,

$$a(\theta) = a(\tau) \quad \Longrightarrow \quad P_\theta(a) \leq P_\tau(a).$$

To build intuition, consider first the case in which a is a safe action that pays the same payoff in every state. Then, choice is bolder in state θ than in state τ if the decision maker is less likely to choose the safe action in θ. Definition 5 extends this logic to actions that are merely safe with respect to the event $\{\theta, \tau\}$.[28]

[27] If there are distinct actions $a, b \in \text{supp}(\alpha)$ with $a(\theta) = b(\theta)$, then we can rank a and b arbitrarily.

[28] Alternatively, we can interpret Definition 5 as stating that the menu $\{b(\theta) : b \in A\}$ of payoffs in state θ is stronger than the menu $\{b(\tau) : b \in A\}$ of payoffs in state τ, in the sense that any action a yielding the same payoff $a(\theta) = a(\tau)$ in both states faces stiffer competition, and thus is less likely to be chosen, in θ than in τ. Under this alternative interpretation, Definition 5 can be viewed as the analogue of the ranking of menus in the perturbed utility model of Fudenberg, Iijima, and Strzalecki (2015), suitably adapted to state-dependent stochastic choice.

Definition 6. Let $\mathcal{D} = (\pi, A)$ be a decision problem and $P = (A, (P_\theta)_{\theta\in\Theta})$ a choice rule. We say that action a is *more salient* than action b for if, for every state $\theta \in \Theta$,

$$a(\theta) = b(\theta) \quad \Longrightarrow \quad P_\theta(a) \geq P_\theta(b).$$

In words, a is more salient than b if the former is always chosen with higher probability in every state where the two actions are payoff equivalent.

For these two orderings to have bite, the decision problem must exhibit sufficient richness. Given a decision problem $\mathcal{D} = (\pi, A)$ and a choice rule $P = (A, (P_\theta)_{\theta\in\Theta})$, we say that states $\theta, \tau \in \Theta$ are *comparable* if there exists an action $a \in A$ such that $a(\theta) = a(\tau)$ and $P_\pi(a) > 0$. Analogously, we say that actions $a, b \in A$ are *comparable* if there exists a state $\theta \in \Theta$ such that $a(\theta) = b(\theta)$.

In such decision problems, the above orderings characterize the ordinal properties of the saddle point (α, λ).

Proposition 2. *Let $\mathcal{D} = (\pi, A)$ be a decision problem, and let $P = (A, (P_\theta)_{\theta\in\Theta})$ be a choice rule that is optimal under a Csiszár information with ϕ essentially smooth. Let (α, λ) be a corresponding saddle point. Then:*

(i) *If two states $\theta, \tau \in \Theta$ are comparable, then choice is bolder in state θ than in state τ if and only if $\lambda_\pi(\theta) \geq \lambda_\pi(\tau)$.*

(ii) *If two actions $a, b \in A$ are comparable, then action a is more salient than action b if and only if $\alpha(a) \geq \alpha(b)$.*

Proposition 2 provides a way to interpret behaviorally the endogenous variables (α, λ). We illustrate these definitions and the result in our running examples:

Example 1 (continued)**.** Suppose the decision maker must choose between a safe and a risky action, and assume P_π has full support. Due to the safe action, every pair of states is comparable. Using (19), it is easy to verify that $\lambda_\pi(\theta) \geq \lambda_\pi(\tau)$ if and only if $r(\theta) \geq r(\tau)$ (provided that ψ is strictly convex). Hence, Proposition 2(i) implies that $P_\theta(r) \geq P_\tau(r)$ if and only if $r(\theta) \geq r(\tau)$, i.e., the probability of choosing the risky action is a strictly increasing function of its reward.

Next, suppose there exists a state $\theta^* \in \Theta$ in which the risky and safe actions yield the same payoff: $r(\theta^*) = 0$. Then, the two actions are comparable, and Proposition 2(ii) implies that $\alpha(r) \geq \alpha(s)$ if and only if $P_{\theta^*}(r) \geq P_{\theta^*}(s)$. In fact, the distribution $\alpha \in \Delta(A)$ can be fully identified from observable choice behavior: the optimality condition (18) implies that $P_{\theta^*}(r)/P_{\theta^*}(s) = \alpha(r)/\alpha(s)$ and, since r and s are the only two available actions, it follows that $\alpha = P_{\theta^*}$.

Example 2 (continued)**.** Consider a guess-the-state problem with at least three distinct states and a uniform prior. The problem is invariant under the full group of permutations

of the state space. Therefore, it admits an optimal symmetric choice rule P for which α is uniform and λ is a constant vector (Proposition 1). As a result, every pair of states and every pair of actions are comparable, choice is equally bold in all states, and all actions are equally salient.

Example 3 (continued)**.** A problem with exchangeable actions is invariant under the subgroup of permutations $(\theta_1, \ldots, \theta_n) \mapsto (\theta_{\gamma(1)}, \ldots, \theta_{\gamma(n)})$, where γ is a permutation of the set $\{1, \ldots, n\}$. Thus, the information acquisition problem admits an optimal symmetric choice rule P such that α is uniform and $\lambda_\pi(\theta) = \lambda_\pi(\tau)$ for all pairs of states $\theta = (\theta_1, \ldots, \theta_n)$ and $\tau = (\tau_1, \ldots, \tau_n)$ that differ only by a permutation of their components. If, in addition, $\Theta = T^n$ for some finite $T \subset \mathbb{R}$, then any two states and any two actions are comparable, and all actions are equally salient. According to the optimality condition (19), choice is bolder in state θ than in state τ if and only if the ψ'-weighted average payoff is higher in the former state.

For instance, under mutual information (Section 4.6), we have $\lambda_\pi(\theta) \geq \lambda_\pi(\tau)$ if and only if $\sum_{i=1}^n e^{\theta_i} \geq \sum_{i=1}^n e^{\tau_i}$. Meanwhile, under the chi-squared divergence (Example 6), if all actions are taken with positive probability in all states, then we have $\lambda_\pi(\theta) \geq \lambda_\pi(\tau)$ if and only if $\sum_{i=1}^n \theta_i \geq \sum_{i=1}^n \tau_i$. Given any strictly convex ψ function, choice is bolder in state θ than in state τ if all actions yield weakly higher payoffs in the former, i.e., $\theta_i \geq \tau_i$ for all $i = 1, \ldots, n$.

Building on the binary-choice example, we now give a different, cardinal characterization of the optimal f-mean α. We now consider decision problems that include a state θ^* in which all actions yield the same payoff. This assumption can be easily made to hold in controlled experimental settings, where the existence of such states can be built in the design of the task at hand. We show below that in any state θ^* of this kind, the distribution $\alpha \in \Delta(A)$ coincides with the choice probability $P_{\theta^*} \in \Delta(A)$. This gives α a clear behavioral interpretation and makes it identifiable from observed choices.

Corollary 2. *Let $\mathcal{D} = (\pi, A)$ be a decision problem, and let $P = (A, (P_\theta)_{\theta \in \Theta})$ be a choice rule that is optimal under a Csiszár information with ψ strictly convex. Let (α, λ) be a corresponding saddle point. If there is a state $\theta^* \in \Theta$ such that $a(\theta^*) = b(\theta^*)$ for all $a, b \in A$, then it holds that $P_{\theta^*} = \alpha$.*

The result is an immediate implication of the optimality conditions (18). In state θ^*, (18) simplifies to $P_{\theta^*}(a)/P_{\theta^*}(b) = \alpha(a)/\alpha(b)$ for all $a, b \in A$. This implies that $P_{\theta^*} = \alpha$, as desired.

5.4 A foundation for additive perturbed utility

A central insight of Matějka and McKay (2015) is that optimal information acquisition can provide a new foundation for, and interpretation of, classic models of stochastic choice. For the special case of mutual information, their paper relates the stochastic choice rule from Theorem 1 to Luce's *multinomial logit* model. In our context, the multinomial logit model

posits that, in each state θ, the decision maker chooses each action $a \in A$ with probability

$$P_\theta(a) = \frac{e^{\frac{a(\theta)}{\kappa}}}{\sum_{b \in A} e^{\frac{b(\theta)}{\kappa}}}, \tag{21}$$

where $\kappa > 0$ is a parameter of the model.

Matějka and McKay (2015) observe that, in decision problems with exchangeable actions (Example 3), the stochastic choice rule in Theorem 1 reduces exactly to the classic logit formula (21). Beyond the exchangeable case, optimal behavior under mutual information costs follows what, in light of Proposition 2, can be seen as a *salience-adjusted* variant of the classic logit rule, whereby actions that are more salient are chosen with relatively higher probability conditional on the state. This adjustment implies, among other features, that strictly dominated actions are never chosen.

We now show that, more generally, optimal information acquisition under Csiszár information provides an analogous foundation for the *additive perturbed utility (APU)* model of discrete choice (Fudenberg, Iijima, and Strzalecki, 2015). In our notation, the APU model posits that, in each state θ, the decision maker's stochastic choice is given by the distribution $P_\theta \in \Delta(A)$ defined as

$$P_\theta = \operatorname*{arg\,max}_{p \in \Delta(A)} \sum_{a \in A} \left[p(a)a(\theta) - c(p(a)) \right], \tag{22}$$

where $c\colon [0,1] \to \overline{\mathbb{R}}_+$ is a perturbation function that incentivizes randomization. Fudenberg, Iijima, and Strzalecki (2015) assume that c is strictly convex and continuously differentiable on $(0,1)$. For our purposes, we make the weaker assumptions that c is strictly convex on its effective domain, is lower semicontinuous, and safisfies $1/n \in \operatorname{ri}(\operatorname{dom} c)$, where n is the cardinality of the action set.

The model, which has found applications in the discrete choice literature as well as in game theory, can be interpreted as representing ex-post optimization errors due to control costs (Mattsson and Weibull, 2002; Flynn and Sastry, 2023), deliberate randomization as a hedge against payoff uncertainty (Fudenberg, Iijima, and Strzalecki, 2015), or certain forms of additive random utility (Hofbauer and Sandholm, 2002). As is well known, APU generalizes multinomial logit: the model reduces to logit when the perturbation takes the form $c(t) = \kappa\,(t \log t - t + 1)$.

By analogy to Matějka and McKay (2015, Proposition 1), we show an equivalence between behavior under Csiszár information and the APU model in exchangeable-action settings:

Corollary 3. *In any exchangeable decision problem with n actions (as defined in Example 3), if P is a symmetric choice rule that is optimal under Csiszár information with transformation ϕ, then P coincides with that of an APU model in which the perturbation function is given by*

$$c(t) = \frac{1}{n}\phi\,(nt)\,. \tag{23}$$

Moreover, given any perturbation function c satisfying the normalizations $c(1/n) = c'(1/n) = 0$, there exists a transformation ϕ such that (23) holds for the corresponding Csiszár information.

Corollary 3 follows directly from comparing the optimality conditions (18) and (19) to those of the APU model (22), and recalling that the optimal f-mean α can be taken to be uniform in exchangeable decision problems.

For general, not necessarily exchangeable, decision problems, the optimal behavior under Csiszár information corresponds to a *salience-adjusted APU* model of discrete choice. Formally, consider a stochastic choice rule P that is optimal under Csiszár information, and let α be the corresponding f-mean. For simplicity, suppose that α has full support. Then, in each state θ, the optimal P_θ can be expressed as

$$P_\theta = \arg\max_{p \in \Delta(A)} \sum_{a \in A} \left[p(a)a(\theta) - \alpha(a)\phi\left(\frac{p(a)}{\alpha(a)}\right)\right], \tag{24}$$

The coefficient $\alpha(a) \in (0,1)$ affects the *salience* of action $a \in A$. By Proposition 2(ii), actions with higher salience are, all else equal, chosen with higher probability in (24). Therefore, Csiszár information permits two forms of context-dependence that the APU model does not: (i) *action-dependence* that arises when two actions $a, b \in A$ yield the same payoff $a(\theta) = b(\theta)$ in state θ but have different salience $\alpha(a) \neq \alpha(b)$, and (ii) *menu-dependence* arising from the fact that the vector α of saliences may depend on the full set A of available actions.

The discrete choice literature has considered versions of the salience-adjusted APU model in which α is treated as an exogenous parameter. For instance, Mattsson and Weibull (2002) and Cerreia-Vioglio, Maccheroni, Marinacci, and Rustichini (2023) study the special case of (24) corresponding to Shannon entropy (Example 4) and interpret α as the decision maker's default choice rule or initial bias, respectively. Meanwhile, Chambers, Masatlioglu, Natenzon, and Raymond (2025) study the special case of (24) corresponding to the chi-squared divergence (Example 6) and interpret α as representing the inherent salience of each action. These approaches are suited to modeling a decision maker's involuntary and automatic ("bottom up") allocation of attention.

By contrast, in our framework, α is determined endogenously. Thus, our approach is suited to modeling a decision maker's optimal and deliberate ("top down") allocation of attention. These optimality conditions also impose extra discipline on the salience weights. For instance, under Csizár information, strictly dominated actions are never chosen.

5.5 IIA properties

Luce's axiom of *independence of irrelevant alternatives (IIA)* is central to the theory of random choice because it provides a behavioral foundation for the logit model. In studying alternative models of discrete choice, a natural question is how they relate to IIA. In this section, we examine the connection between IIA and the predictions of Csiszár information, extending the analysis of Matějka and McKay (2015) beyond mutual information.

In the standard setting of random choice, IIA relates the behavior of a decision maker across different menus of options. In our framework, it translates into an assumption on the decision maker's behavior across decision problems, and hence on the underlying cost function.

We say that a cost function satisfies the IIA axiom if for any two decision problems (π, A) and (π, B), and any pair of corresponding optimal choice rules $P = (A, (P_\theta)_{\theta\in\Theta})$ and $Q = (B, (Q_\theta)_{\theta\in\Theta})$,

$$a(\theta) = c(\tau) \text{ and } b(\theta) = d(\tau) \quad \Longrightarrow \quad \frac{P_\theta(a)}{P_\theta(b)} = \frac{Q_\tau(c)}{Q_\tau(d)} \tag{25}$$

for all actions $a, b \in \text{supp}(P_\theta)$ and $c, d \in \text{supp}(Q_\tau)$, and all states $\theta, \tau \in \Theta$.

When Θ is a singleton, this reduces to Luce's IIA condition. In the more general, state-dependent case, the axiom requires that for any two payoffs $u, v \in \mathbb{R}$ that are both feasible in states θ and τ—that is, $u, v \in \{a(\theta) : a \in A\} \cap \{b(\tau) : b \in A\}$—the relative likelihood of choosing the action that yields u over the one that yields v must be invariant with respect to: (i) which actions implement the payoffs u and v, (ii) whether the realized state is θ or τ, and (iii) what other payoffs are available in those states.

As observed by Matějka and McKay (2015), IIA is generally too restrictive and is violated under mutual information. Specifically, given payoffs $u, v \in \mathbb{R}$ and actions $a, b \in \text{supp}(P_\theta)$ such that $a(\theta) = u$ and $b(\theta) = v$, the likelihood ratio

$$\frac{P_\theta(a)}{P_\theta(b)} = \frac{\alpha(a)}{\alpha(b)} e^{\frac{u-v}{\kappa}}$$

depends not only on the payoff difference $u - v$, but also on the relative salience of the actions a and b, as encoded by the f-mean $\alpha = P_\pi$. We therefore consider three relaxed variants of IIA that are more appropriate in environments with costly information acquisition.

The first axiom, which restates Axiom 1 from Matějka and McKay (2015), relaxes Luce's IIA by controlling for the specific actions that generate any given pair of payoff consequences, thereby addressing the aforementioned complication that arises with unequal salience.

Definition 7. A cost function C satisfies *IIA with respect to actions* if, for every decision problem $\mathcal{D} = (\pi, A)$ and optimal choice rule $P = (A, (P_\theta)_{\theta\in\Theta})$, it holds that

$$a(\theta) = a(\tau) \text{ and } b(\theta) = b(\tau) \quad \Longrightarrow \quad \frac{P_\theta(a)}{P_\theta(b)} = \frac{P_\tau(a)}{P_\tau(b)} \tag{26}$$

for all actions $a, b \in \text{supp}(P_\theta) \cap \text{supp}(P_\tau)$ and every pair of states $\theta, \tau \in \Theta$.

To interpret this condition, observe that

$$\frac{P_\theta(a)}{P_\theta(b)} = \frac{P_\tau(a)}{P_\tau(b)} \quad \Longleftrightarrow \quad \frac{p_a(\theta)}{p_a(\tau)} = \frac{p_b(\theta)}{p_b(\tau)},$$

where $p_a, p_b \in \Delta(\Theta)$ denote the decision maker's posterior beliefs upon taking actions a and b, respectively. Therefore, (26) states that, if actions a and b are both constant on $\{\theta, \tau\}$, then—conditional on the event $\{\theta, \tau\}$—they are *informationally equivalent* signals about the state.

The second axiom postulates that the decision maker does not distinguish between states θ and τ that are payoff-equivalent, i.e., such that all actions in the decision problem are constant on the event $\{\theta, \tau\}$. This property is equivalent to Caplin, Dean, and Leahy's (2022) *invariance under compression* axiom for settings where the state space and prior are held fixed.[29]

Definition 8. The cost function C satisfies *IIA with respect to labels* if, for every decision problem $\mathcal{D} = (\pi, A)$ and optimal choice rule $P = (A, (P_\theta)_{\theta \in \Theta})$, it holds that

$$a(\theta) = a(\tau) \text{ for all } a \in A \quad \Longrightarrow \quad P_\theta = P_\tau$$

for every pair of states $\theta, \tau \in \Theta$.

Intuitively, this axiom captures two assumptions: that states are merely labels that index the payoff consequences of actions, and that the decision maker does not need to spend effort distinguishing these labels when doing so is payoff-irrelevant.

The third and final axiom, which restates Axiom 2 from Matějka and McKay (2015), is a separability property reminiscent of Savage's sure-thing principle, suitably adapted to stochastic choice. It posits that, if two actions a and b coincide on the event $\{\theta, \tau\}$, then the likelihood ratio of choosing a over b is the same in both states, regardless of how a and b differ on the complementary event $\Theta \backslash \{\theta, \tau\}$.

Definition 9. The cost function C satisfies *IIA with respect to states* if, for every decision problem $\mathcal{D} = (\pi, A)$ and optimal stochastic choice rule $P = (A, (P_\theta)_{\theta \in \Theta})$, it holds that

$$a(\theta) = b(\theta) \text{ and } a(\tau) = b(\tau) \quad \Longrightarrow \quad \frac{P_\theta(a)}{P_\theta(b)} = \frac{P_\tau(a)}{P_\tau(b)}$$

for all actions $a, b \in \operatorname{supp}(P_\theta) \cap \operatorname{supp}(P_\tau)$ and every pair of states $\theta, \tau \in \Theta$.

We note that IIA with respect to states is satisfied by the multinomial logit model (21), which further implies that the likelihood ratios are equal to one.

With these definitions in hand, we have the following result:

Proposition 3. *Given a Csiszár information cost with ϕ essentially smooth:*

(i). *The cost function satisfies IIA with respect to labels and states.*

(ii). *In a decision problem (π, A), condition (26) holds for states $\theta, \tau \in \Theta$ if $\lambda_\pi(\theta) = \lambda_\pi(\tau)$.*

(iii). *If $|\Theta| \geq 5$ and $\psi = \phi^\star$ is thrice continuously differentiable, the agent satisfies IIA with respect to actions if and only if the cost function is proportional to mutual information.*

[29] The invariance-under-compression axiom also applies to shifts in the prior, which we do not analyze here.

The result singles out mutual information as the only type of Csiszár information that satisfies IIA with respect to actions. This is a much stronger assumptions than IIA with respect to states or labels. That every Csiszár information satisfies IIA with respect to states follows from the additive separability of the transformation f. IIA with respect to labels is implied by the fact that the transformation ϕ is not a direct function of the state.

The characterization of mutual information in the last part of Proposition 3 is related to a result in the same spirit in Matějka and McKay (2015, Proposition 2), but differs in two respects. First, our primitives are different: unlike in their paper, we take the utility function as given, and moreover we start from the assumption that the cost function belongs to the Csiszár information class. Second, their result only shows that there exists a distribution α over action such that the stochastic choice rule takes the adjusted-logit formula 11, but does not ensure that this α is optimal or equal to the unconditional distribution P_π.

The proof of Proposition 3(iii) applies tools from risk theory to the study of information acquisition under Csiszár information. We introduce these tools in the following section.

6 Tools from risk theory and their applications

In this section we analyze the properties of Csiszár information by drawing on concepts from expected utility theory. We show that the degree of convexity of the conjugate $\psi = \phi^\star$ has a central place in characterizing the solutions to information acquisition problems, much like the concavity of a Bernoulli utility shapes behavior in expected utility theory.

To simplify the analysis, for the rest of this section, in addition to Assumption 2, we posit that ψ is twice continuously differentiable and strictly convex. Under these assumptions, we define

$$R_\psi(t) = \frac{\psi''(t)}{\psi'(t)}.$$

As in the study of utility functions, $R_\psi(t)$ is an index measuring the degree of convexity of the function ψ at the value t. With slight abuse of terminology, we refer to R_ψ as the *Arrow-Pratt coefficient* of ψ.

6.1 Behavioral characterization of the Arrow-Pratt coefficient

Our starting point is the following observation, which relates the solution to an information acquisition problem under Csiszár information and the coefficient R_ψ.

Corollary 4. *Given a decision problem (π, A) and a Csiszár information with transformation ϕ, if a stochastic choice rule P is optimal and (α, λ) is its corresponding saddle point, then*

$$\log \frac{P_\theta(a)}{P_\theta(b)} = \log \frac{\alpha(a)}{\alpha(b)} + \int_{b(\theta)}^{a(\theta)} R_\psi(t - \lambda_\pi(\theta))\, \mathrm{d}t, \tag{27}$$

for every state θ and pair of actions a and b in the support of P_π.

The result follows from the optimality conditions in Theorem 2—see in particular Equation (18)—together with the fact that R_ψ is the derivative of $\log \psi'$. It establishes that, at the optimum, the log-likelihood ratio between action a and b in a state θ is the sum of two terms: the log-likelihood ratio between the two actions under the f-mean α, and the integral of the Arrow-Pratt coefficient R_ψ between $a(\theta) - \lambda_\pi(\theta)$ and $b(\theta) - \lambda_\pi(\theta)$.

Building on this result, we give a behavioral interpretation of the Arrow-Pratt coefficient R_ψ and show that it measures how strongly the decision maker responds to a increase in incentives for information acquisition. To formalize this idea, we focus on a subclass of exchangeable decision problems (Example 3) that we call irreducible:

Definition 10. An n-action exchangeable decision problem is *irreducible* if there exists a payoff vector $d = (d_1, \dots, d_n) \in \mathbb{R}^n$ such that

$$\Theta = \left\{ \left(d_{\gamma(1)}, \dots, d_{\gamma(n)} \right) : \gamma \text{ is a permutation of } \{1, \dots, n\} \right\}.$$

We denote such a decision problem by $\mathcal{D}(d)$.

In this decision problem, every state (viewed as a payoff vector) is a permutation of the same state d. In an irreducible decision problem, the prior π is uniform; hence, an n-action irreducible problem is fully determined by its payoff vector d. A simple example is a guess-the-state problem (Example 2), corresponding to $d = (w, 0, \dots, 0)$ with $w > 0$ as the winning payoff.

Under Csiszár information, irreducible problems admit an optimal symmetric choice rule P corresponding to a saddle point (α, λ) where α is uniform and λ is a constant vector, i.e., $\lambda(\theta) = \lambda(d)$ for all $\theta \in \Theta$ (Proposition 1). The prior-adjusted Lagrange multiplier λ_π is also constant, with $\lambda_\pi(d)$ uniquely determined by

$$\frac{1}{n} \sum_{i=1}^{n} \psi'(d_i - \lambda_\pi(d)) = 1.$$

Uniqueness follows from the strict monotonicity of ψ'.

We consider two irreducible decision problems $D(d)$ and $\mathcal{D}(d')$ close if the Euclidean distance between the payoff vectors d and d' is small. This allows us to define perturbations of a given problem d that introduce a small additional incentive to acquire information.

Definition 11. Let $\mathcal{D}(d)$ be an irreducible decision problem, and let $i, j \in \{1, \dots, n\}$ be indices such that $d_i = d_j$. Given $\epsilon > 0$, we say that $d^\epsilon \in \mathbb{R}^n$ is an *ϵ-split* of $\mathcal{D}(d)$ along the dimensions i and j if

$$d^\epsilon_i = d_i + \epsilon, \quad d^\epsilon_j = d_j - \epsilon, \quad d^\epsilon_k = d_k \text{ for all } k \neq i, j.$$

In the original problem defined by d, the choice between actions a_i and a_j is inconsequential in state $\theta = d$, since the two actions yield the same payoff. The decision problem $\mathcal{D}(d^\epsilon)$ is a

perturbation where the choice between a_i and a_j is now made consequential in state d^ϵ while keeping fixed the payoffs of the other actions.

For example, consider the trivial problem $d = (0, \dots, 0)$, in which all actions yield zero payoff. A ϵ-split along the dimensions $i = 1$ and $j = 2$ produces the guess-the-state problem $d^\epsilon = (\epsilon, -\epsilon, 0, \dots, 0)$ with $\epsilon > 0$ as the winning payoff. Here, the perturbation injects a small incentive to acquire information. Table 1 presents a less trivial example.

$\mathcal{D}(d)$	a_1	a_2	a_3
$d = \theta_1$	1	0	0
θ_2	0	1	0
θ_3	0	0	1

(a) Original decision problem.

$\mathcal{D}(d^\epsilon)$	a_1	a_2	a_3
$d^\epsilon = \theta_1^\epsilon$	1	ϵ	$-\epsilon$
θ_2^ϵ	1	$-\epsilon$	ϵ
θ_3^ϵ	ϵ	1	$-\epsilon$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
θ_6^ϵ	$-\epsilon$	ϵ	1

(b) Perturbed problem corresponding to the ϵ-split d^ϵ along dimensions 2 and 3.

Table 1: Table (a) describes a guess-the-state problem. The set of states is $\Theta = \{\theta_1, \theta_2, \theta_3\}$, the action set is $A = \{a_1, a_2, a_3\}$, and each entry is the corresponding payoff. In each state, payoffs are permutations of the vector $d = (1, 0, 0)$. Table (b) describes a modified decision problem where the original state θ_1 is now split into two states θ_1^ϵ and θ_2^ϵ. In both states the agent's main goal is to play a_1, but they now face an additional incentive to choose a_2 in state θ_1^ϵ and a_3 in θ_2^ϵ. The same applies to states θ_2 and θ_3. Proposition 4 quantifies the decision maker's response to this incentive.

We are ready to present our behavioral characterization of the Arrow-Pratt coefficient R_ψ.

Proposition 4. *Consider an irreducible decision problem $\mathcal{D}(d)$, and let $i, j \in \{1, \dots, n\}$ be indices such that $d_i = d_j$. Consider a collection $(d^\epsilon)_{\epsilon \in (0,1)}$, where each d^ϵ is an ϵ-split of d along the dimensions i and j. Then:*

$$\log \frac{P^\epsilon_{d^\epsilon}(a_i)}{P^\epsilon_{d^\epsilon}(a_j)} = 2\epsilon R_\psi(d_i - \lambda_\pi(d)) + o(\epsilon),$$

where each P^ϵ is an optimal symmetric choice rule for $\mathcal{D}(d^\epsilon)$ and $\lambda_\pi(d)$ is the prior-adjusted Lagrange multiplier associated to d.

The perturbation d^ϵ modifies d by introducing a new, low-powered incentive for the decision maker to acquire information—specifically, to learn which of the two actions a_i or a_j is preferable in state d^ϵ. The parameter ϵ captures the scale of this incentive, and the log-likelihood ratio $\log P^\epsilon_{d^\epsilon}(a_i)/P^\epsilon_{d^\epsilon}(a_j)$ represents the predicted response of the decision maker. The proposition shows that this response, as a function of ϵ, is proportional to the Arrow-Pratt coefficient R_ψ evaluated at $d_i - \lambda_\pi(\theta)$, up to a first-order approximation.

6.2 Violations of IIA and the Arrow-Pratt coefficient

The IIA with respect to actions axiom requires that the likelihood ratio $P_\theta(a)/P_\theta(b)$ between two actions a and b depends only on the payoffs of those two actions in that state. Under Csiszár information, this property is generally violated, as the likelihood ratio can also depend on the payoffs of other available actions in that state. We now connect such violations of IIA to monotonicity properties of the Arrow-Pratt coefficient R_ψ.

The next definition, adapted to state-dependent stochastic choice, is inspired by the work of Fudenberg, Iijima, and Strzalecki (2015) on additive perturbed utility.

Definition 12. A cost function C exhibits *increasing selectivity* if, in every decision problem $\mathcal{D} = (\pi, A)$ and for every optimal choice rule $P = (A, (P_\theta)_{\theta\in\Theta})$, the following holds: for any two states $\theta, \tau \in \Theta$ such that choice is bolder in θ than in τ, and for every two actions $a, b \in A$ in the support of P_π,

$$a(\theta) = a(\tau) \geq b(\theta) = b(\tau) \quad \Longrightarrow \quad \frac{P_\theta(a)}{P_\theta(b)} \geq \frac{P_\tau(a)}{P_\tau(b)}.$$

Conversely, the agent exhibits *decreasing selectivity* if, under the same conditions,

$$a(\theta) = a(\tau) \geq b(\theta) = b(\tau) \quad \Longrightarrow \quad \frac{P_\theta(a)}{P_\theta(b)} \leq \frac{P_\tau(a)}{P_\tau(b)}.$$

Increasing and decreasing selectivity capture two patterns of violations of IIA with respect to actions. Recall that states in which the decision maker is bolder are associated with higher values of the multiplier: even though actions a and b yield the same payoff in states θ and τ, both are chosen with lower probability in the bolder state θ, i.e. $P_\theta(a) \leq P_\tau(a)$ and $P_\theta(b) \leq P_\tau(b)$. Increasing selectivity means that, in bolder states, the decision maker is relatively more likely to favor the better action: the likelihood ratio $P_\theta(a)/P_\theta(b)$ between the better action a and the worse action b is higher in θ than in τ. Decreasing selectivity describes the opposite pattern.

Increasing and decreasing selectivity are characterized by the monotonicity of the Arrow-Pratt coefficient:

Proposition 5. *Assume $|\Theta| \geq 5$. Let ψ be thrice continuously differentiable. Then:*

(i). *The agent exhibit increasing selectivity if and only if R_ψ is decreasing.*

(ii). *The agent exhibit decreasing selectivity if and only if R_ψ is increasing.*

A corollary of this result is that IIA with respect to actions characterizes the case where R_ψ is constant, i.e. ψ is exponential, in which case Csiszár information reduces to mutual information, as noted in Proposition 3(iii).

6.3 Relation to posterior separable costs

As noted in Examples 4–6, Csiszár information nests mutual information, which is also posterior separable. In fact, mutual information is essentially the unique cost function contained in both the Csiszár information and posterior separable classes. This implies that, generically, the two class of models lead to distinct predictions. Formally, we have the following characterization:

Proposition 6. *Assume $|\Theta| \geq 3$. For any Csiszár information cost function C with ψ thrice continuously differentiable, C is posterior separable if and only if it is proportional to mutual information.*

The proof of Proposition 6 builds on the idea of studying the transformation ψ as a Bernoulli utility function. Posterior-separable costs are characterized, in the dual space, by a property of translation invariance—see (12). In turn, this property is equivalent to the Arrow-Pratt coefficient R_ψ being constant, which implies ψ is exponential.

As we demonstrate in the next section, the behavioral predictions of Csiszárs information and those of posterior-separable costs can diverge even in very simple decision problems.

7 Inconclusive evidence and consideration sets

Inconclusive evidence refers to situations in which informative and uninformative signals co-exist, a common occurrence in many real-world scenarios. For example, medical test results often include not only positive and negative outcomes but also inconclusive ones. Except for knife-edge cases, inconclusive evidence is inconsistent with models of costly information acquisition based on mutual information or, more broadly, posterior separability (Denti, 2022). In this section, we demonstrate how Csizár cost can be used to analyze the possibility of inconclusive evidence in information choice. We maintain Assumption 2, as well as the hypothesis of Section 6: $\psi = \phi^\star$ is strictly convex and twice continuously differentiable.

7.1 Guess-the-state with outside option

To focus the discussion, we consider a guess-the-state problem, as in Example 2, with the addition of an outside option. Let $n \geq 2$ be the number of possible states, and assume the prior π is uniform. The decision maker has $n+1$ feasible actions. For each state θ, there is a risky action, a_θ, that corresponds to a bet on that state: $a_\theta(\theta) = w$, while $a_\theta(\tau) = 0$ for all $\tau \neq \theta$. The coefficient $w > 0$ is the reward for correctly guessing the state. In addition, there is a safe action, b, that yields a constant payoff of $c > 0$, independent of the state. This setup mirrors the structure of many economic applications, such as selecting between risky assets and bonds in a portfolio problem, or choosing whether to participate in projects with uncertain returns or take a known outside option.

In this decision problem, inconclusive evidence emerges when risky and safe actions are all chosen with positive probability: informative signals, prompting the selection of risky

actions, co-exists with uninformative signals, leading to the choice of the safe action. Except for knife-edge cases, such choice pattern is incompatible with mutual information:

Under mutual information, three distinct cases arise depending on the appeal of the safe action. To describe these cases, let $\hat{c}$ be the threshold defined by

$$\hat{c} = \log\left(\frac{1}{n}e^{\frac{w}{\kappa}} + \frac{n-1}{n}\right)^{\kappa}. \tag{28}$$

(i). For $c > \hat{c}$, no learning occurs and the decision maker never tries to guess the state: $P_\pi(b) = 1$ at the optimum.

(ii). For $c < \hat{c}$, the decision maker always tries to guess the state and never uses the safe action: $P_\pi(b) = 0$ at the optimum.

(iii). In the knife-edge case where $c = \hat{c}$, multiple solutions exist. The decision maker may exclusively choose the safe action, completely avoid it, or mix across all actions with positive probabilities.

Thus, under mutual information, inconclusive evidence emerges only in a knife-edge case and is never the unique prediction of the model. To give an intuition for this negative result and, more importantly, to address it, we next consider the case of Csizár information.

7.2 Predictions under Csiszár information

It will be useful once again to study ψ as if it was the Bernoulli utility function of a risk-loving agent. By the optimality condition for α in the maximin problem (14), both the risky and safe actions are part of consideration set only if

$$\frac{1}{n}\psi\left(w - \lambda_\pi(\theta)\right) + \frac{n-1}{n}\psi(0 - \lambda_\pi(\theta)) = \psi\left(c - \lambda_\pi(\,\theta)\right), \tag{29}$$

for all $\theta \in \Theta$. Mirroring the discussion in the previous section, the left-hand side of (29) can be seen as the expected utility of a lottery that pays w with probability $1/n$ and 0 with probability $(n-1)/n$, for an agent with wealth level equal to (the negative of) the prior-adjusted Lagrange multiplier $\lambda_\pi(\theta)$.[30] For (29) to hold, the quantity c must correspond to the certainty equivalent of the lottery.

The analogy with risk theory explains why inconclusive evidence is inconsistent with mutual information. Under mutual information, ψ is exponential, meaning that the certainty equivalent of a lottery is independent of the wealth level. As a result, (29) is independent of $\lambda_\pi(\theta)$, and the equation can hold only for a knife-edge configuration of the primitives of the problem.

Next we show that inconclusive evidence emerges as a robust prediction of the model as soon a we move away from the case of constant absolute risk seeking.

[30] Due to the symmetry of the environment and strict convexity of ψ, the Lagrange multiplier is unique and independent of the state—see Corollary 1 and Proposition 1.

Proposition 7. *Suppose $R_\psi = \psi''/\psi'$ is strictly monotone on the interval $(-w, w)$. Then, there are thresholds $\underline{c}$ and $\bar{c}$, with $\underline{c} < \bar{c}$, such that:*

(i). *If $c > \bar{c}$, then $P_\pi(b) = 1$ at the optimum.*

(ii). *If $c < \underline{c}$, then $P_\pi(b) = 0$ at the optimum.*

(iii). *If $c \in (\underline{c}, \bar{c})$, then $\operatorname{supp} P_\pi = A$ at the optimum.*

To generalize the result beyond the case in which the Arrow-Pratt coefficient is strictly monotone on a neighborhood of zero, we introduce a parametrization of the transformation ϕ: for all $k \in \operatorname{int}(\operatorname{dom}\phi)$, we define $\phi_k : \mathbb{R}_+ \to \overline{\mathbb{R}}$ by

$$\phi_k(t) = \frac{\phi(kt) - \phi(k)}{k} - (t-1)\phi'_+(k).$$

The original function ϕ corresponds to the case in which $k = 1$, meaning that $\phi_1 = \phi$. Note that the parameter k has no effect in the case of mutual information: if $\phi(t) = \kappa(t \log t - t + 1)$, then $\phi_k = \phi$ for all $k \in (0, +\infty)$.

The role of this parametrization is better understood through the conjugate of ϕ_k, which we denote by ψ_k. To elaborate, take $t_k \in \mathbb{R}$ such that $\psi'(t_k) = k$.[31] Then, simple calculations show that for all $t \in \mathbb{R}$,

$$\psi_k(t) = \frac{\psi(t + t_k) - \psi(t_k)}{k}.$$

In particular, $R_{\psi_k}(t) = R_\psi(t + t_k)$. Thus, the effect of the k parameter is to cause a shift of the Arrow-Pratt coefficient. Note that any shift can be generated in this way, as $\operatorname{int}(\operatorname{dom} k)$ coincides with the image of ψ'.

Proposition 8. *Suppose R_ψ is strictly monotone on a non-empty open interval. Then, there is an open set of parameters (k, w, c) such that under ϕ_k, $\operatorname{supp} P_\pi = A$ at the optimum.*

7.3 Posterior Separability

Finally, we emphasize that the inability to represent inconclusive evidence is inherent to all symmetric posterior-separable costs.[32] As in Example 6, let $H\colon \Delta(\Theta) \to \overline{\mathbb{R}}_+$ be an entropy function: convex, essentially strictly convex, lower semicontinuous function, with $\pi \in \operatorname{ri}(\operatorname{dom} H)$. We say that H is *symmetric* if $H(p) = H(q)$ for all posteriors $p, q \in \Delta(\Theta)$ such that the vectors $(p(\theta))_{\theta\in\Theta}$ and $(q(\theta))_{\theta\in\Theta}$ are permutations of each other.

Proposition 9. *Let information costs be posterior separable, with H symmetric. Then, for every w there exists a threshold $\hat{c}$ such that:*

(i). *If $c > \hat{c}$, then $P_\pi(b) = 1$ at the optimum.*

[31] The existence of t_k is ensured by the fact that $k \in \operatorname{int}(\operatorname{dom}\phi)$.

[32] If the cost is not symmetric, it may be possible for the safe action to be chosen alongside *some*, but not all, risky actions (see Appendix B).

(ii). *If $c < \hat{c}$, then $P_\pi(b) = 0$ at the optimum.*

(iii). *If $c = \hat{c}$, then for every $t \in [0,1]$ there is an optimal choice rule such that $P_\pi(b) = t$.*

As with mutual information, inconclusive evidence is a non-generic prediction. In Appendix B, we relate these observations to the more general, though more abstract, issue of studying the size of the consideration set under f-information and posterior-separable costs. As is well known, under posterior separability the size of the consideration set is at most the cardinality of the state space in generic decision problems (see, e.g., Denti, 2022, Proposition 4). We show that f-information can enlarge the consideration set, but by no more than one action. Hence, while f-information expands the consideration set to accommodate phenomena such as the use of inconclusive evidence, it does so in a parsimonious way, in line with the observation that decision-makers face limited consideration sets.

8 Choice accuracy and learning incentives

The rational inattention literature highlights two main shortcomings of mutual information as a model of information acquisition. As Dean and Neligh (2023) observe: first, "subjects are less responsive to incentives than the Shannon model would predict"; and second, "subjects do not behave identically in payoff-identical states when the environment admits a natural notion of perceptual distance." In the next three sections, we show that the f-information framework can address both limitations.

First, we examine responsiveness to incentives. In a canonical task in which the agent's objective is to correctly identify the true state (Example 2), we study how the predicted probability of a correct choice varies with the primitives of the problem. In our analysis, we show how to identify information costs non-parametrically and investigate the properties of the marginal cost of information.

8.1 Response functions

Let n and m be positive integers such that $1 \leq m < n$. The decision problem involves n equally likely states and n actions, where each action represents a bet on an event comprising m states. A successful bet—one where the realized state belongs to the chosen event—yields a reward of $w > 0$; otherwise, the payoff is zero. To ensure symmetry, we assume that in each state, exactly m actions yield the reward w, while the remaining $n - m$ actions result in zero payoff. Although this symmetric structure is somewhat special, it is well-suited for implementation in laboratory experiments.

We now present three concrete examples. In each case, we index the set of actions by the set of states, i.e., $A = \{a_\theta : \theta \in \Theta\}$:

- Suppose $m = 1$. Each action a_θ is a bet *on* state θ: it pays w if the realized state is θ and zero otherwise.

- Suppose $m = n - 1$. Each action a_θ is a bet *against* state θ: it pays zero if the realized state is θ and w otherwise.
- Suppose the states are points uniformly spaced on a circle. Each action a_θ pays w if the realized state is θ or one of its $m - 1$ immediate clockwise successors.

As in the previous sections, we work with Csiszár information and assume ϕ satisfies Assumption 2. We also assume that $(0, +\infty) \subseteq \operatorname{dom} \phi$. Equivalently, the conjugate function, $\psi = \phi^\star$, is strictly convex and the image of ψ' is $(0, +\infty)$.

A key quantity of interest is the probability of correctly guessing the state as a function of the learning incentive. The next proposition uses Theorem 2 to provide a characterization:

Proposition 10. *For every state θ,*

$$P_\theta(\{a : a(\theta) = w\}) = \frac{m}{n}\psi'(w - l),$$

where l is the unique solution of the equation

$$\frac{m}{n}\psi'(w - l) + \frac{n - m}{n}\psi'(-l) = \psi'(0).$$

The coefficient l is simply the multiplier $\lambda_\pi(\theta)$, which by the symmetry of the problem is independent of the state. Motivated by this result, for every $\gamma \in (0, 1)$ we define the decision maker's *response function* $\rho_\gamma : (0, +\infty) \to (0, 1)$ as

$$\rho_\gamma(w) = \gamma\psi'(w - l_\gamma(w)),$$

where $l_\gamma(w)$ is determined by the equation

$$\gamma\psi'(r - l_\gamma(w)) + (1 - \gamma)\psi'(-l_\gamma(w)) = \psi'(0).$$

Allowing all $\gamma \in (0, 1)$ is only a matter of notational convenience, since rational values of γ already provide a dense approximation.

The response function succinctly captures how the agent adjusts behavior in response to learning incentives. Next, we analyze the first- and second-order properties of response functions, comparing them to the benchmark case of mutual information.

8.2 First-order properties

In the case of mutual information the response function takes the form:

$$\rho_\gamma(w) = \frac{\gamma e^{\frac{w}{\kappa}}}{\gamma e^{\frac{w}{\kappa}} + 1 - \gamma}.$$

Dean and Neligh (2023) provide evidence that in the case of two states and two actions (i.e. $\gamma = 1/2$), the response function implied by mutual information fails to adequately fit experimental data. Intuition suggests that this issue may extend to other values of γ and alternative experimental designs, as the single parameter κ does not offer enough flexibility. The following result shows that Csizár information allows for a wider range of predictions.

Proposition 11. *For each γ, the response function satisfies the following properties:*

(i). $\rho_\gamma(w)$ *is strictly increasing in* w.

(ii). $\rho_\gamma(w)$ *is continuous in* w.

(iii). $\rho_\gamma(w) \to \gamma$ *as* $w \to 0$.

(iv). $\rho_\gamma(w) \to 1$ *as* $w \to +\infty$.

Conversely, any function that satisfies (i)–(iv) is a response function for γ for some ϕ.

Caplin, Csaba, Leahy, and Nov (2020), Dewan and Neligh (2020), and Dean and Neligh (2023) all provide experimental evidence that response functions are increasing.[33] Naturally, continuity cannot be directly tested with finite data. Focusing on a specific class of continuous functions, Dewan and Neligh (2020) offer mixed results on continuity. As the prize w approaches 0, and all actions yield almost identical payoffs, property (iii) shows that the agent's choice converges to a uniform randomization. Property (iv) implies that the state is learnable with arbitrary precision; it can be relaxed by dropping the hypothesis that ϕ is finite on $(0, +\infty)$.

Caplin, Csaba, Leahy, and Nov (2020) and Dewan and Neligh (2020) use response functions to estimate the cost of information. As the proof of Proposition 11 makes clear, non-parametric identification of ϕ cannot be achieved solely from observing the agent's behavior for a fixed γ, as multiple function ϕ can generate the same ρ_γ. However, we establish that ϕ can be identified by jointly varying both w and γ.

Proposition 12. *If ϕ_1 and ϕ_2 induce the same response function for every γ, then $\phi_1 = \phi_2$.*

The proof shows that identification is ensured even in the simpler case where ϕ_1 and ϕ_2 induce the same response function for every γ of the form $\gamma = 1/n$ or $\gamma = (n-1)/n$. It is therefore sufficient to focus on simple decision problems where the decision maker is asked to bet on or against a particular state. While exact identification requires observing the decision makers behavior for every n, informative bounds can still be obtained using the following expressions: for all $w > 0$,

$$\psi'(w) = \sup_{n>1} n\rho_{\frac{1}{n}}(w) \quad \text{and} \quad \psi'(-w) = \inf_{n>1} n\left(1 - \rho_{\frac{n-1}{n}}(w)\right). \tag{30}$$

We conclude the study of first-order properties by extending Proposition 11 to the case where both γ and w are allowed to vary. To this end, we introduce the concept of *inverse response function*. Given a response function ρ_γ, and given any $x \in (1, +\infty)$ and $y \in (0, 1)$, define $\gamma(x, y)$ and $w(x, y)$ as the unique γ and w that solve the system of equations:

$$\frac{\rho_\gamma(w)}{\gamma} = x \quad \text{and} \quad \frac{1 - \rho_\gamma(w)}{1 - \gamma} = y.$$

[33] It can be shown that, for any cost function, the marginal probability $P_\pi(\{a : a(\theta) = w\})$ of guessing correctly is non-decreasing in the reward w. See, e.g., Dewan and Neligh (2020).

That $\gamma(x, y)$ is well defined follows from the fact that ρ_γ is strictly increasing, continuous, and satisfies $\rho_\gamma(w) \to \gamma$ as $w \to 0$ and $\rho_\gamma(w) \to 1$ as $w \to \infty$. We refer to the mapping $(x, y) \mapsto w(x, y)$ as the *inverse response function*. While ρ_γ maps payoffs to choice probabilities, the inverse response function maps observed choice behavior—expressed in likelihood ratios—to the underlying payoff. Given $\gamma(x, y)$, the quantity $w(x, y)$ is the reward level that generates the likelihood ratios (x, y).

The inverse response function allows us to test and identify the transformation ϕ:

Proposition 13. *The inverse response function satisfies the following properties:*

(i). $w(x, y)$ *is strictly increasing in* x *and strictly decreasing in* y.

(ii). $w(x, y)$ *is continuous in* x *and* y.

(iii). $w(x, y) \to 0$ *as* $x \to 1$ *and* $y \to 1$.

(iv). $w(x, y) \to +\infty$ *as* $x \to +\infty$ *or* $y \to 0$.

(v). *For all* x *and* x', $w(x, y) - w(x', y)$ *is independent of* y.

(vi). *For all* y *and* y', $w(x, y) - w(x, y')$ *is independent of* x.

Conversely, any function that satisfies (i)–(vi) is an inverse response function for some ϕ. *Moreover, if* ϕ_1 *and* ϕ_2 *induce the same inverse response function, then* $\phi_1 = \phi_2$.

Properties (i)–(iv) of the inverse response function mirror those of the response function stated in Proposition 11, and they admit a similar interpretation. Properties (v) and (vi), in turn, reflect the separability inherent in Csiszár cost. Empirically testing these properties would shed light on the extent to which this separability assumption constrains the model. Finally, as the proof of the proposition illustrates, ϕ and the inverse response function are connected by the following equation: for all $x \in (1, +\infty)$ and $y \in (0, 1)$,

$$\phi'(x) = \inf_{z \in (0,1)} w(x, z) \quad \text{and} \quad \phi'(y) = -\inf_{z \in (1,+\infty)} w(z, y).$$

These formulas are the dual version of the expressions in (30).

8.3 Second-order properties

Second-order properties of the response function, such as concavity or convexity, reflect the decision maker's marginal sensitivity to learning incentives. In this section, we show that these properties reveal important characteristics of the marginal cost of information, such as whether it increases or decreases with information acquisition.

For the remainder of this section, we assume that ψ is thrice continuously differentiable. An important tool in our analysis is the Arrow-Pratt coefficient of ψ', defined as:

$$R_{\psi'} = \frac{\psi'''}{\psi''}.$$

In choice theory, $R_{\psi'}$ is known as the *prudence index* of ψ and figures prominently in the study of precautionary savings (Kimball, 1990). In a very different context, our findings connect $R_{\psi'}$ to the second-order properties of the response function.

We first investigate under what conditions the response function is concave:

Proposition 14. *The following statements are equivalent:*

(i). *For all γ, ρ_γ is concave.*

(ii). *$R_{\psi'}(t) \geq 0$ for $t < 0$, and $R_{\psi'}(t) \leq 0$ for $t > 0$.*

(iii). *$R_{\phi'}(t) \leq 0$ for $t \in (0,1)$, and $R_{\phi'}(t) \geq 0$ for $t > 1$.*

Thus, the response function is concave for every γ if and only if the prudence index $R_{\psi'}(t)$ is positive for $t < 0$ and negative for $t > 0$. Equivalently, this holds if ϕ' is concave on $(0,1)$ and convex on $(1,+\infty)$. This condition can be interpreted as stating that the marginal cost of information is increasing, for acquiring information means generating variability in the likelihood ratio below and above one.

The response function cannot be globally convex, as it is bounded above by 1. We therefore focus on the case where it is initially convex and later concave—an *S-shaped* profile. This is precisely the shape exhibited by the response function under mutual information:

Example 3 (Continued)**.** Under mutual information,

$$R_{\rho_\gamma}(w) = \frac{\rho_\gamma''(w)}{\rho_\gamma'(w)} = \frac{(1-\gamma) - \gamma e^w}{\gamma e^w + (1-\gamma)}.$$

Thus, the Arrow-Pratt coefficient of the response function is decreasing in learning incentives. In particular,

$$R_{\rho_\gamma}(w) \geq 0 \quad \iff \quad w \leq \log \frac{1-\gamma}{\gamma},$$

and hence ρ_γ is first convex and then concave.

The next result investigates conditions under which a response function is S-shaped. Formally, we say that ρ_γ is *S-shaped* if

$$w_1 \geq w_2 \quad \text{and} \quad \rho_\gamma''(w_1) \geq 0 \quad \Longrightarrow \quad \rho_\gamma''(w_2) \geq 0.$$

It is *inverse S-shaped* if $-\rho_\gamma$ is S-shaped. A sufficient condition for ρ_γ to be S-shaped is that its Arrow-Pratt coefficient of risk loving, R_{ρ_γ}, is decreasing.

Proposition 15. *The following properties hold:*

(i). *If $R_{\psi'}$ is decreasing, then ρ_γ is S-shaped. Moreover, ϕ' is inverse S-shaped.*

(ii). *If $R_{\psi'}$ is decreasing and ψ'' is monotone, then R_{ρ_γ} is decreasing. Moreover, $R_{\phi'}$ is increasing.*

(iii). *If R_{ρ_γ} is decreasing for all γ, then $R_{\psi'}$ is decreasing.*

Condition (i) shows that a decreasing prudence index $R_{\psi'}$ is a sufficient condition for the response function to be S-shaped. Conditions (ii) and (iii) provide partial converses.

9 Perceptual Csiszár information

Under the common assumption of mutual information cost, states enter the analysis only through their payoff consequences; other features of states, such as their physical characteristics and distance from each other, play no role. As a consequence, under mutual information, if two states θ_1 and θ_2 have the same prior probability, then exchanging the conditional distributions P_{θ_1} and P_{θ_2} of an experiment P leaves its cost unchanged. In decision problems, this is reflected in the property of IIA with respect to labels.

As several authors have noted (e.g., Hébert and Woodford, 2021; Morris and Yang, 2022; Dean and Neligh, 2023; Pomatto, Strack, and Tamuz, 2023), this invariance property leads to unrealistic predictions in decision problems where it is inherently more difficult to distinguish between states that are more similar. For example, in problems where an agent must bet on whether a one-dimensional state, such as the return of an asset, is positive or negative, under mutual information the optimal choice probability will display a jump exactly at the state equal to 0, rather than varying smoothly across nearby states as common sense suggests.

These observations apply not only to mutual information, but also to Csiszár information. For this reason, in the next two sections we study generalizations of Csiszár information that take into account the structure of the state space. Our goal is to identify a generalization with three features: (i) it remains a special case of f-information, with a conjugate that is analytically manageable; (ii) it has enough parameters to capture relevant features of the state space; and (iii) its parameters have transparent interpretations.

9.1 Encoding states as attributes

Our approach builds on the hypothesis that the decision maker learns by categorizing states through a simplified mental model that emphasizes a selected set of *attributes* of the state space. We interpret the attribute space as a subjective representation of the state space. Formally, learning proceeds in two stages: each state is first mapped into an attribute, and information is then acquired as if attributes were the primitive states.[34]

Definition 13. A *personal state space* consists of a finite set N and a kernel $K = (N, (K_\theta)_{\theta\in\Theta})$ such that for all $i \in N$ there is $\theta \in \Theta$ such that $K_\theta(i) > 0$. We refer to N as the set of *attributes* and to K as the *encoder*.

[34]The idea that decision makers may simplify their choice environments through a smaller set of attributes has several analogues in prior work. For example, see Gul, Natenzon, and Pesendorfer (2014) and Walker-Jones (2023).

Each attribute is a property of the state that the agent considers focal for reducing uncertainty about the environment. For instance, if $\theta \in \mathbb{R}$ is a one-dimensional variable, N could be a partition of Θ into "low," "medium," or "high" values. If $\theta \in \mathbb{R}^d$ is a high-dimensional vector describing the details of a health plan, N could consists of a set of coarse labels such as "cheap but minimal" or "expensive but comprehensive." The kernel K describes the probability $K_\theta(i)$ with which a state θ is perceived as belonging to attribute i. To avoid redundancy, we require that each attribute is associated with some state with positive probability.

Definition 14. Let $\phi\colon \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ be a function that satisfies Assumption 2, and let (N, K) be a personal state space. The *perceptual Csiszár information* is defined for every experiment $P = (\Omega, (P_\theta)_{\theta\in\Theta})$ as

$$I(P) = \inf_{Q\in\Delta(\Omega)^N} \left(\inf_{\alpha\in\Delta(\Omega)} \sum_{i\in N} \nu(i) D_\phi(Q_i \| \alpha) \right) \quad \text{s.t.} \quad Q \circ K = P,$$

where $\nu = \sum_{\theta\in\Theta} \pi(\theta) K_\theta$, and $Q \circ K\colon \Theta \to \Delta(\Omega)$ is the kernel defined as $\sum_{i\in N} Q_i(\omega) K_\theta(i) = P_\theta(\omega)$ for all $\omega \in \Omega, \theta \in \Theta$.

First, states are mapped to attributes via the kernel K. Second, the agent acquires information about the attribute via an experiment Q, subject to the standard Csiszár information

$$J(Q) = \inf_{\alpha\in\Delta(\Omega)} \sum_{i\in N} \nu(i) D_\phi(Q_i \| \alpha). \tag{31}$$

Here, ν represents the unconditional probability of attributes, obtained by combining the prior with the encoder. Given any target experiment P about the state, the agent chooses the cheapest experiment Q about the attribute that *replicates* P, in the sense that $Q \circ K = P$. If the target experiment cannot be replicated in this manner, then it is deemed infeasible and assigned infinite cost.

This interpretation closely parallels classic notions from information theory (see, e.g., Cover and Thomas, 2006, Chapter 7). In this analogy, states correspond to external messages to be processed; the experiment Q functions as a communication channel; attributes serve as codewords transmitted through the channel; and the kernel K maps messages to codewords. In light of this analogy, we refer to K as an encoder and we refer to Q as a *channel*.

The defining feature of the model is that the encoder K is exogenously given, while the channel Q is chosen optimally.[35] We therefore view the encoder K as modeling the agent's hardwired perceptual limitations. Formally, it delineates an upper bound on what the agent can learn about the state: an experiment P is replicable if and only if it is a garbling of K.

[35] By contrast, in Shannon's theory of channel coding, the channel Q is exogenously given and the encoder K is optimally designed, and in Sims's (2003) interpretation of the benchmark rational inattention model based on mutual information, both the encoder and channel are optimally chosen.

Meanwhile, the channel Q models the agent's deliberate allocation of attention, given these limitations.[36]

We illustrate these concepts through several examples:

- *Perfect perception:* When $N = \Theta$ and K is the identity map, every experiment P is replicable, and the perceptual Csiszár cost reduces to a standard Csiszár information with transformation ϕ.
- *Deterministic categorization:* Each $i \in N$ indexes an event $B_i \subseteq \Theta$ in a partition $\{B_i\}_{i \in N}$ of the state space. The encoder is defined by $K_\theta(i) = 1$ if $\theta \in B_i$, and $K_\theta(i) = 0$ otherwise. In this case, the agent can acquire information about which partition cell the state belongs to, but not the state itself. This setup captures an agent who bins states into coarse categories.
- *Perceptual distance:* Let $N = \Theta$, and let $d\colon \Theta \times \Theta \to \mathbb{R}_+$ be a metric on the state space. Given a decreasing function $\gamma\colon \mathbb{R}_+ \to \mathbb{R}_+$ with $\gamma(0) > 0$, define the encoder K as

$$K_\theta(\tau) = \frac{\gamma(d(\theta, \tau))}{\sum_{\sigma \in \Theta} \gamma(d(\theta, \sigma))}. \tag{32}$$

  This specification models an agent who struggles to distinguish between nearby states. It is flexible enough to nest (or approximate) the preceding examples as special cases, and will put it to work in Section 9.4.

Holding the transformation ϕ and the attribute set N fixed, the Blackwell ranking over encoders fully characterizes the ordinal ranking over perceptual Csiszár costs:

Proposition 16. *Consider two perceptual Csiszár costs, I_1 and I_2, with parameters (ϕ, N, K_1) and (ϕ, N, K_2), respectively. The following statements hold:*

(i). *If K_1 is a garbling of K_2, then $I_1(P) \geq I_2(P)$ for all $P \in \mathcal{E}$.*

(ii). *If $\mathrm{dom}(\phi) = \mathbb{R}_+$, then $I_1(P) \geq I_2(P)$ for all $P \in \mathcal{E}$ only if K_1 is a garbling of K_2.*

9.2 Optimality conditions

In solving for the optimal choice probabilities, the next assumption streamlines the analysis.

Assumption 3. *The set of vectors $\{(K_\theta(i))_{\theta \in \Theta} : i \in N\}$ is linearly independent.*

Interpreting the encoder as a matrix with states as rows and attributes as columns, Assumption 3 requires this matrix to have full column rank. This, in turn, implies that the

[36] In the language of cognitive psychology, the model parallels a hybrid earlylate selection theory of attention: the kernel K operates as an early-stage selective filter, involuntarily determining which stimuli are available for voluntary late-stage processing (Broadbent, 1958; Pashler, 1998; Bordalo, Gennaioli, and Shleifer, 2022).

number of attributes $|N|$ is weakly smaller than the number of states $|\Theta|$, consistent with the idea that attributes provide a coarse description of the state space.

Under Assumption 3, perceptual Csiszár information belongs to the class of f-information costs and admits a remarkably simple conjugate.

Proposition 17. *Consider a perceptual Csiszár information I with parameters (ϕ, N, K). Under Assumption 3, I coincides with an f-information with conjugate*

$$f^\star(x) = \sum_{i \in N} \nu(i) \psi \left(\sum_{\theta \in \Theta} \frac{\mu_i(\theta)}{\pi(\theta)} x(\theta) \right),$$

$\psi = \phi^\star$ and $\mu_i(\theta) = \frac{\pi(\theta) K_\theta(i)}{\nu(i)}$.

The distribution μ_i denotes the conditional distribution over states given attribute i.

Using Theorem 2, we obtain from Proposition 17 a characterization of the optimal choice rule in the perceptual Csiszár model. For a vector $x \in \mathbb{R}^\Theta$, we denote by $E[x] = (E_i[x])_{i \in N} \in \mathbb{R}^N$ the vector of conditional expectations $E_i[x] = \sum_{\theta \in \Theta} \mu_i(\theta) x(\theta)$. Given any saddle point (α, λ) of the maxmin problem (14), the optimal choice probabilities are

$$P_\theta(a) = \alpha(a) \sum_{i \in N} K_\theta(i) \psi' \left(E_i[a] - E_i[\lambda_\pi] \right). \tag{33}$$

This expression admits the decomposition

$$P_\theta(a) = \sum_{i \in N} K_\theta(i) Q_i(a),$$

where

$$Q_i(a) = \alpha(a) \psi' \left(E_i[a] - E_i[\lambda_\pi] \right)$$

is the probability that action a is chosen when attribute i is focal. Assumption 3 guarantees each Q_i is a valid probability distribution over actions.[37]

9.3 Working in the attribute space

The optimality condition described by (33) suggests that, in order to find the optimal choice rule, it is necessary to solve for the full saddle point (α, λ). Since I, unlike standard Csiszár information, is not additively separable across states, computing the multiplier λ may seem difficult. In particular, the value of λ in state θ may depend on the full profile of payoffs in the other states.

[37] Indeed,

$$1 = \sum_{i \in N} K_\theta(i) \left(\sum_{a \in A} Q_i(a) \right) \quad \text{for all } \theta \in \Theta \quad \Rightarrow \quad \sum_{a \in A} Q_i(a) = 1 \quad \text{for all } i \in N.$$

We show, however, that the problem can be simplified: it suffices to study a lower-dimensional saddle-point problem, where the original state space Θ is replaced by the space of attributes N, and the perceptual Csiszár information I is replaced by the standard Csiszár information J. The analysis can therefore be reduced to an auxiliary information acquisition problem in which information costs are separable and amenable to the tools developed in Sections 5 and 6.

Specifically, given any decision problem $\mathcal{D} = (\Theta, \pi, A)$, we define the *reduced problem* $\bar{\mathcal{D}} = (N, \nu, \bar{A})$, where

$$\bar{A} = \left\{ E[a] \in \mathbb{R}^N : a \in A \right\}.$$

That is, $\bar{\mathcal{D}}$ is the projection of $\mathcal{D}$ onto the attribute space. To simplify the exposition, in the next proposition we assume that $a \neq b$ implies $E[a] \neq E[b]$, so that the sets A and $\bar{A}$ are in a one-to-one correspondence.[38]

Proposition 18. *Let $\mathcal{D} = (\Theta, \pi, A)$ be a decision problem and consider a perceptual Csiszár information cost I with parameters (ϕ, N, K). Assume $a \neq b$ implies $E[a] \neq E[b]$.*

Let J be the associated Csiszár information as defined in (31). Then, the following statements are equivalent:

(i). *$P = (A, (P_\theta)_{\theta \in \Theta})$ is optimal in $\mathcal{D} = (\Theta, \pi, A)$ under the perceptual Csiszár cost I.*

(ii). *There exists a choice rule $\bar{Q} = (\bar{A}, (\bar{Q}_i)_{i \in N})$, which is optimal in $\bar{\mathcal{D}} = (N, \nu, \bar{A})$ under the standard Csiszár cost J, such that*

$$P_\theta(a) = \sum_{i \in N} K_\theta(i) \bar{Q}_i(E[a]) \quad \text{for all } a \in A, \theta \in \Theta. \tag{34}$$

Proposition 18, which does not require Assumption 3, shows that the optimal choice rule P can be computed in two steps. First, solve for the optimal rule $\bar{Q}$ in the reduced decision problem $\bar{\mathcal{D}} = (N, \nu, \bar{A})$ with Csiszár information cost J. Second, recover P from $\bar{Q}$ using (34). Since the second step is purely mechanical, the perceptual Csiszár model retains the tractability of the standard Csiszár framework.

In particular, Theorem 2 implies that the first step reduces to finding a saddle point $(\bar{\alpha}, \bar{\lambda}) \in \Delta(\bar{A}) \times \mathbb{R}^N$. Moreover, as shown in Section 5, the multiplier $\bar{\lambda}$ can be computed attribute-by-attribute as the solution to

$$\sum_{\bar{a} \in \bar{A}} \bar{\alpha}(\bar{a}) \psi' \left(\bar{a}(i) - \bar{\lambda}_\nu(i) \right) = 1 \quad \text{for all } \; i \in N, \tag{35}$$

where $\bar{\lambda}_\nu(i) = \bar{\lambda}(i)/\nu(i)$.

The following example illustrates:

[38] The result extends to cases with $a \neq b$ and $E[a] = E[b]$, which can be resolved using any tie-breaking rule between actions with the same projection.

Example 8 (Perceptual mutual information)**.** Suppose $\psi(t) = e^t - 1$, so that J is mutual information. Then, every saddle point $(\bar{\alpha}, \bar{\lambda}) \in \Delta(\bar{A}) \times \mathbb{R}^N$ in $\bar{\mathcal{D}}$ satisfies

$$\bar{\lambda}_\nu(i) = \log \sum_{\bar{a} \in \bar{A}} \bar{\alpha}(\bar{a}) e^{\bar{a}(i)} \quad \text{for all } i \in N.$$

As a result, all optimal choice rules in $\mathcal{D}$ take the form

$$P_\theta(a) = \sum_{i \in N} K_\theta(i) \frac{\bar{\alpha}(E[a]) e^{E_i[a]}}{\sum_{b \in A} \bar{\alpha}(E[b]) e^{E_i[b]}}.$$

Thus, we obtain a perceptual version of Matějka and McKay (2015). The choice rule resembles a state-dependent mixed logit model.[39]

Finally, as a corollary, we also obtain a continuity result on $P_\theta(a)$ as a function of θ:

Corollary 5. *For all $a \in A$ and $\theta, \tau \in \Theta$,*

$$|P_\theta(a) - P_\tau(a)| \leq \bar{\alpha}(E[a]) \cdot \|K_\theta - K_\tau\|_1 \cdot \max_{i \in N} \psi' \left(E_i[a] - \bar{\lambda}_\nu(i) \right),$$

where $\| \cdot \|$ is the L^1-norm.

A notable implication of Corollary 5 is the coarser bound:

$$|P_\theta(a) - P_\tau(a)| \leq \|K_\theta - K_\tau\|_1.$$

That is, the encoder K bounds the slope of the map $\theta \mapsto P_\theta$, uniformly across all decision problems. This implies that the perceptual Csiszár model can generate the discrete-state analogue of the continuous-choice property from Morris and Yang (2022). It achieves this by placing hard constraints on what the agent is able to learn; as discussed in Lipnowski and Ravid (2023), this would be the only way to achieve continuous choice, uniformly across all decision problems, in a continuous-state version of the model.

9.4 Application: perceptual distance in one-dimensional problems

We conclude our presentation of perceptual Csiszar information with an application to a canonical one-dimensional discrimination task. The state space is a finite, equally spaced subset of the real line, $\Theta \subset \mathbb{R}$. For clarity, we index the states in increasing order and write $\Theta = \{\theta_1, \ldots, \theta_n\}$, with $\theta_{i+1} - \theta_i = \Delta > 0$ for all $i = 1, \ldots, n-1$. Since discrimination tasks are typically formulated in continuous settings, this construction can be viewed as a uniform discretization.

The agent chooses between a risky actions r and a safe action s. The payoff of the risky action varies monotonically with the state: $\theta \geq \tau$ implies $r(\theta) \geq r(\tau)$. A simple example is a

[39] A version of this choice rule appears in a sender-receiver context in Bloedel and Segal (2021).

binary bet where action r pays 1 if the state is positive, and -1 if the state is negative. As in Example 1, the the safe action's payoff is normalized to zero.

We consider a decision maker whose perceptual acuity diminishes with proximity between states. To encode this structure, we set $N = \{1, \dots, n\}$ and interpret $K_{\theta_i}(j)$ as the probability of encoding state θ_i as θ_j.

A central object of interest in discrimination tasks is the relationship between stimulus intensity and choice frequency. In our framework, this is captured by the function $\theta \mapsto P_\theta(r)$, commonly referred to as *psychometric function*. Pyshchometric functions observed in experiments are typically S-shaped (Khaw, Li, and Woodford, 2021). This means that $P_\theta(r)$ increases with θ, consistent with action r being more appealing in high states, and that this function is convex at low stimulus levels and concave at high ones. In our discrete setting, we say that the psychometric function is *convex at* θ_i if

$$P_{\theta_i}(r) - P_{\theta_{i-1}}(r) \leq P_{\theta_{i+1}}(r) - P_{\theta_i}(r),$$

and *concave at* θ_i if the inequality is reversed.

The next proposition relates these features of the psychometric function to properties of the encoder.

Proposition 19. *(i). The psychometric function is monotone increasing if the encoder satisfies the monotone likelihood ratio property (MLRP):*

$$\theta \geq \tau \text{ and } i \geq j \quad \text{implies} \quad K_\theta(i)K_\tau(j) \geq K_\tau(i)K_\theta(j).$$

(ii). Assume the encoder satisfies the MLRP. The psychometric function is convex at θ_i if $\frac{1}{2}K_{\theta_{i-1}} + \frac{1}{2}K_{\theta_{i+1}}$ first-order stochastically dominates K_{θ_i}.

(ii). Assume the encoder satisfies the MLRP. The psychometric function is concave at θ_i if K_{θ_i} first-order stochastically dominates $\frac{1}{2}K_{\theta_{i-1}} + \frac{1}{2}K_{\theta_{i+1}}$.

The MLRP captures the idea that higher states are more likely to be encoded as higher attributes, reflecting perceptual consistency with the ordering of states. Next we provide an example of a class of encoders that satisfies the MLRP.

Example 9. For all $\theta \in \Theta$ and $i \in N$, define the encoder

$$K_\theta(i) = \frac{\gamma(|\theta - \theta_i|)}{\sum_{j \in N} \gamma(|\theta - \theta_j|)},$$

where $\gamma \colon \mathbb{R}_+ \to (0, +\infty)$ is a decreasing function. This specification assigns higher encoding probability to nearby states, with the decay governed by γ (cf. Equation (32)). The encoder satisfies the MRLP if γ is log-concave.

Convexity and concavity of the psychometric function can be derived from primitive properties of convexity and concavity of the encoder. A simple example follows:

Example 10. Let ξ and χ be two probability distributions over attributes satisfying the MLRP: for all $i, j \in N$ with $i \geq j$, $\xi(i)\chi(j) \geq \xi(j)\chi(i)$. For each state θ and attribute i, define the encoder as

$$K_\theta(i) = \gamma(\theta)\xi(i) + (1 - \gamma(\theta))\chi(i)$$

where $\gamma\colon \mathbb{R} \to (0, 1)$ is an increasing function. In this specification, the encoder forms a convex combination of two baseline perceptual modes, ξ and χ. The distribution ξ represents perception biased toward high states, while χ represents perception biased toward low states. The mixing function γ governs the relative weight of these modes: as the true state θ increases, more weight is placed on the high-state mode ξ.

One can verify that the encoder inherits the MLRP. In addition, $\frac{1}{2}K_{\theta_{i-1}} + \frac{1}{2}K_{\theta_{i+1}}$ first-order stochastically dominates K_{θ_i} whenever γ is convex at θ_i. In the case where γ is concave at θ_i, the reverse dominance relation holds. Consequently, an increasing psychometric function with an S-shape arises when γ is convex for low values of θ and concave for high values.

10 Nested entropies

We build on the idea of encoding states into attributes to introduce a new class of posterior-separable costs based on *nested entropies*. These entropy functions combine analytical tractability—via a well-behaved conjugate—with a suggestive interpretation in terms of "nests" of states sharing similar attributes. As we show, they connect closely to Hébert and Woodford's (2021) neighborhood-based costs and to Walker-Jones's (2023) multi-attribute Shannon entropy, as well as to the nested logit model from discrete choice.

10.1 Nested Shannon entropy

Let (N, K) be a personal state space, consisting of a finite set of attributes N and a Markov kernel $K = (N, (K_\theta)_{\theta\in\Theta})$ that encodes states into attributes (Definition 13). We assume that for every attribute i there exists a state θ such that $K_\theta(i) > 0$. Given a prior π over the state space, the induced distribution over attributes is $\nu = \sum_{\theta\in\Theta} \pi(\theta)K_\theta$. For each attribute i, the conditional distribution of states given i is denoted by μ_i, with $\mu_i(\theta) = K_\theta(i)\pi(\theta)/\nu(i)$.

Definition 15. Let (N, K) be a personal state space, and fix weights $\zeta > 0$ and $\eta_i > 0$ for each $i \in N$. The *nested Shannon entropy* $H_{\mathrm{NS}}\colon \Delta(\Theta) \to \mathbb{R}_+$ is defined as

$$H_{\mathrm{NS}}(p) = \inf \left\{ \zeta\, D_{\mathrm{KL}}(r\|\nu) + \sum_{i\in N} r(i)\eta_i D_{\mathrm{KL}}(q_i\|\mu_i) \right\} \tag{36}$$

where the infimum is taken over all attribute distributions r and Markov kernels $q = (\Theta, (q_i)_{i\in N})$ such that $\sum_{i\in N} r(i)q_i = p$.

As with perceptual Csiszár information, the decision maker is envisioned as learning by categorizing states into attributes: the cost of a posterior $p \in \Delta(\Theta)$ is computed indirectly,

as the cost of the cheapest extension of p to state-attribute pairs. Such an extension is represented by a pair (r, q), consisting of an attribute distribution $r \in \Delta(N)$ and a Markov kernel $q \in \Delta(\Theta)^N$, such that $\sum_{i \in N} r(i) q_i = p$. The pair (r, q) induces a joint distribution over states and attributes whose marginal over states is p.

The cost of a candidate extension (r, q) decomposes into *across-attribute* and *within-attribute* components:

$$\zeta D_{\mathrm{KL}}(r \| \nu) + \sum_{i \in N} r(i) \eta_i D_{\mathrm{KL}}(q_i \| \mu_i).$$

This expression can itself be viewed as an entropy function over joint distributions of states and attributes, measuring the divergence from the "prior" determined by π and the encoder K. The parameters ζ and $(\eta_i)_{i \in N}$ govern the relative importance of the across-attribute and within-attribute components.

10.2 Special cases

Several special cases illustrate the logic of nested Shannon entropy and clarify the interpretation of its parameters. To simplify the exposition, we assume throughout that

$$\eta_i = \eta \quad \text{for all } i \in N.$$

When learning across attributes is less costly than learning within attributes ($\zeta \leq \eta$), the nested Shannon entropy is bounded above by the standard Shannon entropy, scaled by η:

$$H_{\mathrm{NS}}(p) \leq \eta D_{\mathrm{KL}}(p \| \pi).$$

Grouping states into attributes allows the decision maker to simplify the learning problem and thereby incur lower information costs. In the special case where the costs of learning across and within attributes are identical ($\eta = \zeta$), the nested Shannon entropy coincides with the standard Shannon entropy:

$$H_{\mathrm{NS}}(p) = \eta D_{\mathrm{KL}}(p \| \pi).$$

These results follow directly from the chain rule for KL divergence (Cover and Thomas, 2006, Chapter 2).

In the extreme case where learning within attributes is prohibitively costly ($\eta \to +\infty$), the decision maker can acquire information about states only indirectly, through attributes. In this limit, an extension (r, q) of a posterior p to state-attribute pairs has finite cost only if $q_i = \mu_i$ for all $i \in N$. Consequently, the limiting entropy is

$$H_{\mathrm{NS}}(p) = \inf \left\{ \zeta D_{\mathrm{KL}}(r \| \nu) : r \in \Delta(N),\ \sum_{i \in N} r(i) \mu_i = p \right\}. \tag{37}$$

This special case aligns closely with perceptual Csiszár information:

Proposition 20. *Let C denote the posterior-separable cost function induced by the entropy in (37). If the set of vectors $\{(K_\theta(i))_{\theta\in\Theta} : i \in N\}$ is linearly independent, then C coincides with the perceptual Csiszár information cost parametrized by (N, K, ϕ), where $\phi(t) = \zeta(t\log t - t + 1)$ for all $t \in \mathbb{R}_+$.*

Nested Shannon entropy thus relaxes some of the rigidities inherent in perceptual Csiszár information by introducing a trade-off between learning indirectly through attributes and directly about states, governed by the parameters ζ and η. Under perceptual Csiszár information, learning is restricted to the attribute space, which forces many information structures to have infinite cost: an experiment is feasible only if it is a garbling of the encoder. By contrast, nested Shannon entropy assigns finite cost to every posterior (except in the limiting cases $\zeta \to +\infty$ or $\eta \to +\infty$).

For fixed values of ζ and η, the encoder determines the structural relationship between states and attributes. As with perceptual Csiszár information, this relationship is subjective, reflecting the agent's perceptual limitations.

In the extreme case of *perfect perception*—when $K_\theta(i) > 0$ implies $K_\tau(i) = 0$ for all $\tau \neq \theta$—nested Shannon entropy reduces standard Shannon entropy scaled by ζ:

$$H_{\mathrm{NS}}(p) = \zeta D_{\mathrm{KL}}(p\|\pi).$$

Here, because attributes fully reveal the underlying states, only the attributes themselves are costly to learn.

At the opposite extreme of *null perception*—$K_\theta = K_\tau$ for all $\theta, \tau \in \Theta$—nested Shannon entropy reduces to standard Shannon entropy scaled by η:

$$H_{\mathrm{NS}}(p) = \eta D_{\mathrm{KL}}(p\|\pi).$$

In this case, attributes convey no information about the states, so the decision maker optimally learns directly about the states instead.

Finally, we highlight an intermediate case of imperfect perception: *deterministic categorization*. Here, each $i \in N$ corresponds to a cell $B_i \subseteq \Theta$ in a partition $\{B_i\}_{i\in N}$ of the state space. The attribute reveals exactly which cell contains the state, and nothing more: $K_\theta(i) = 1$ if $\theta \in B_i$, and $K_\theta(i) = 0$ otherwise. In this case, the minimization problem defining nested Shannon entropy admits a closed-form solution:

Proposition 21. *Under deterministic categorization, for each $p \in \Delta(\Theta)$, the infimum in (36) is achieved by*

$$r(i) = p(B_i) \quad \textit{and} \quad q_i(\theta) = p(\theta|B_i).$$

Under deterministic categorization, each partition cell B_i can be interpreted as a nest of states with shared attributes. This interpretation is reinforced by the close relationship between $H^\star_{\mathrm{NS}}$, the conjugate of H_{NS}, and the nested logit model in discrete choice, which we detail next.

10.3 Conjugate function and optimality conditions

The conjugate of the nested Shannon entropy admits a tractable closed-form expression:

Proposition 22. *For every* $x \in \mathbb{R}^{\Theta}$,

$$H^{\star}_{NS}(x) = \zeta \log \left(\sum_{i \in N} \nu(i) \left(\sum_{\theta \in \Theta} \mu_i(\theta) e^{x(\theta)/\eta_i} \right)^{\eta_i/\zeta} \right).$$

In discrete choice theory, this functional form is known as the surplus function of the *generalized nested logit* model (Wen and Koppelman, 2001). Whereas in discrete choice nests group alternatives that consumers regard as substitutes, here nests capture states that share similar attributes in the learning process, such as perceptual proximity.

Under deterministic categorization, this expression collapses to the surplus function of the canonical *nested logit* model:

$$H^{\star}_{\mathrm{NS}}(x) = \zeta \log \left(\sum_{i \in N} \pi(B_i) \left(\sum_{\theta \in B_i} \pi(\theta | B_i) e^{x(\theta)/\eta_i} \right)^{\eta_i/\zeta} \right).$$

In this formulation, each state either belongs to a nest or not. By contrast, the more general specification above allows for graded participation across nests, with the degree of overlap determined by the encoder's noise.

In discrete choice applications, it is standard to restrict attention to the parameter region $\eta_i \leq \zeta$, ensuring a random-utility interpretation. In our setting, however, no such restriction is warranted: the case $\eta_i > \zeta$ corresponds to situations where learning about attributes/nests is less costly than learning about states within attributes/nests.

Leveraging the closed-form expression for the conjugate, we can apply Theorem 2 to derive explicit optimality conditions (up to the f-mean $\alpha = P_\pi$ and the Lagrange multiplier λ). In particular, for every action a in the consideration set, the posterior p_a at which a is chosen is given by

$$p_a(\theta) = \nabla_\theta H^{\star}_{\mathrm{NS}}(a - \lambda_\pi) = \frac{\sum_{i \in N} \nu(i) \mu_i(\theta) e^{\frac{a(\theta) - \lambda_\pi(\theta)}{\eta_i}} \left(\sum_{\tau \in \Theta} \mu_i(\tau) e^{\frac{a(\tau) - \lambda_\pi(\tau)}{\eta_i}} \right)^{\frac{\eta_i - \zeta}{\zeta}}}{\sum_{i \in N} \nu(i) \left(\sum_{\tau \in \Theta} \mu_i(\tau) e^{\frac{a(\tau) - \lambda_\pi(\tau)}{\eta_i}} \right)^{\frac{\eta_i}{\zeta}}},$$

where λ_π denotes the prior-adjusted Lagrange multiplier. This expression admits a suggestive decomposition:

$$p_a(\theta) = \sum_{i \in N} r_a(i) q_{(a,i)}(\theta),$$

where $r_a(i)$ represents the probability of nest i,

$$r_a(i) = \frac{\nu(i) \left(\sum_{\tau \in \Theta} \mu_i(\tau) e^{\frac{a(\tau) - \lambda_\pi(\tau)}{\eta_i}} \right)^{\frac{\eta_i}{\zeta}}}{\sum_{j \in N} \nu(j) \left(\sum_{\tau \in \Theta} \mu_j(\tau) e^{\frac{a(\tau) - \lambda_\pi(\tau)}{\eta_j}} \right)^{\frac{\eta_j}{\zeta}}},$$

and $q_{(a,i)}(\theta)$ is the probability of state θ conditional on nest i,

$$q_{(a,i)}(\theta) = \frac{\mu_i(\theta) e^{\frac{a(\theta)-\lambda_\pi(\theta)}{\eta_i}}}{\sum_{\tau\in\Theta} \mu_i(\tau) e^{\frac{a(\tau)-\lambda_\pi(\tau)}{\eta_i}}}.$$

The pair (r_a, q_a) thus extends p_a to state-attribute pairs, mirroring the two-stage structure of nested logit models.

10.4 Relation to neighborhood-based costs

Nested Shannon entropy bears a close resemblance to two other families of cost functions in the literature: the *neighborhood-based costs* of Hébert and Woodford (2021) and the *multi-attribute Shannon entropy* (MASE) of Walker-Jones (2023). Like our approach, these families embed structural features of the state space into the cost function. To facilitate comparison, we focus on the leading parametric specification of neighborhood-based cost, which is also built on KL divergence and encompasses MASE as a special case.[40]

Given a finite index set I, a covering $\mathcal{B} = \{B_i\}_{i\in I}$ of the state space,[41] and constants $\kappa_i > 0$, Hébert and Woodford (2021) define the entropy function

$$H_{\mathrm{HW}}(p) = \sum_{i\in I} \kappa_i \, \bar{p}(i) D_{\mathrm{KL}}(p_i \| \pi_i), \tag{38}$$

where $\bar{p}(i) = p(B_i)$ is the posterior probability of event B_i, $p_i \in \Delta(B_i)$ is the corresponding conditional posterior given by $p_i(\theta) = p(\theta|B_i)$ for all $\theta \in B_i$, and $\pi_i \in \Delta(B_i)$ is the analogous conditional prior. Hébert and Woodford interpret each event B_i as a *neighborhood of states* that are costly to distinguish. These neighborhoods are analogous to nests or attributes in the nested Shannon model, and (38) resembles the $\zeta \to 0$ limit of (36), but without the minimization step.

The connection between nested Shannon entropy, neighborhood-based costs, and MASE is most transparent when the neighborhood structure takes the form $\mathcal{B} = \{B_0\} \cup \{B_i\}_{i\in N}$, where N is a set of attributes, $B_0 = \Theta$, and $\{B_i\}_{i\in N}$ is a partition of Θ. In this setting, the chain rule for KL divergence yields

$$H_{\mathrm{HW}}(p) = \kappa_0 \, D_{\mathrm{KL}}(\bar{p} \| \bar{\pi}) + \sum_{i\in N} (\kappa_0 + \kappa_i) \, \bar{p}(i) D_{\mathrm{KL}}(p_i \| \pi_i).$$

This expression is exactly the MASE entropy function of Walker-Jones (2023). It also coincides with the nested Shannon entropy under deterministic categorization, with nests $\{B_i\}_{i\in N}$ and scaling parameters $\zeta = \kappa_0$ and $\eta_i = \kappa_0 + \kappa_i$ (Proposition 21). Thus, this special case of the

[40] One can extend Definition 15 by nesting more general entropy functions—that is, general convex transformations of probability distributions. We restrict attention to the Shannon case (KL divergence) for clarity of exposition, but our main results, such as the closed-form expression for the conjugate, remain valid.

[41] Each B_i is a subset of states, and $\Theta = \cup_{i\in I} B_i$.

nested Shannon model aligns with the subclass of neighborhood-based models that exhibit tree-like neighborhood structures, and is equivalent to MASE.

Beyond this special case, however, nested Shannon entropy and neighborhood-based costs diverge in subtle but important ways. When neighborhoods overlap, there are multiple ways to extend a posterior belief p to state-neighborhood pairs. For example, if a state θ lies in two distinct neighborhoods $B_i \neq B_j$, nested Shannon entropy splits the probability mass $p(\theta)$ across the two events and—when several such splits are possible—selects the allocation that minimizes cost. By contrast, the neighborhood-based entropy accounts for the probability $p(\theta)$ twice, since θ is included in both events.[42] In the next section, we show a simple class of decision problems where the two cost functions lead to qualitatively different predictions.

10.5 Application: the challenge of multi-dimensional learning

To conclude our presentation of the nested Shannon model, we apply it to a simple multi-dimensional discrimination task. This serves two purposes: to highlight a novel connection between optimal information acquisition and concepts in psychology, and to illustrate behavioral differences between the nested Shannon and neighborhood-based models.

We consider a setting where the state is two-dimensional and the decision maker finds it hard to engage in *multi-dimensional learning*: it is easy to learn about each dimension of the state separately, but difficult to learn about both simultaneously. For instance, in the perceptual experiments of Tversky and Russo (1969), it is easy for lab subjects to correctly determine which of two rectangles has the larger area when they differ only by width or height, but harder to do so when they differ along both dimensions. Similarly, in a market setting, it may be easy for a consumer to choose correctly between products that differ only in terms of quality or price, but harder for them to do so when the products differ in both respects.

The premise that multi-dimensional comparisons are more difficult than uni-dimensional ones—while largely absent from the rational inattention literature—is familiar from several lines of research in psychology and economics. For instance, this theme is central to recent work on similarity and comparison complexity in the stochastic choice literature (e.g., He and Natenzon, 2024; Shubatt and Yang, 2024).[43]

Setting. Formally, we consider the following simplified setting. The state space is a four-element product set, $\Theta = \{u, d\} \times \{l, r\}$, and the prior is uniform, $\pi(\theta) = 1/4$ for all $\theta \in \Theta$. For mnemonic convenience, we interpret the state as the location of a visual stimulus, where the first dimension indexes its vertical position ("up" or "down") and the second dimension

[42] In particular, the induced measure $\bar{p}$ on I in (38) typically has total mass strictly greater than one.

[43] Under the standard mutual information cost, the decision maker may endogenously simplify a multi-dimensional state by optimally learning only about a particular linear combinations of its dimensions (e.g., Kőszegi and Matějka, 2020), but there is no sense in which multi-dimensional learning is intrinsically harder than uni-dimensional learning.

indexes its horizontal position ("left" or "right"). It is convenient to define the events

$$U = \{(u,l),(u,r)\}, \quad D = \Theta \backslash U, \quad L = \{(u,l),(d,l)\}, \quad R = \Theta \backslash L.$$

That is, $\{U, D\}$ defines a partition of states based on their vertical positions ("Up" or "Down"), while $\{L, R\}$ defines a partition based on their horizontal positions ("Left" or "Right").

For each event $i \in \{U, D, L, R\}$, we define $a_i \in \mathbb{R}^\Theta$ as the action that pays a reward of 1 if $\theta \in i$ and pays 0 otherwise. We also define the actions $a_{\text{diag}}, a_{\text{off}} \in \mathbb{R}^\Theta$ as

$$a_{\text{diag}}(\theta) = \begin{cases} 1, & \text{if } \theta \in \{(u,l),(d,r)\} \\ 0, & \text{otherwise,} \end{cases} \qquad \text{and} \qquad a_{\text{off}}(\theta) = \begin{cases} 1, & \text{if } \theta \in \{(u,r),(d,l)\} \\ 0, & \text{otherwise.} \end{cases}$$

That is, a_{diag} pays a reward of 1 when θ lies on the diagonal, and pays 0 otherwise; symmetrically, a_{diag} pays a reward of 1 when θ lies on the off-diagonal, and pays 0 otherwise.[44] We consider the three binary-choice decision problems defined via the action sets

$$A_1 = \{a_U, a_D\}, \quad A_2 = \{a_L, a_R\}, \quad A_3 = \{a_{\text{diag}}, a_{\text{off}}\}.$$

In decision problem 1 (resp. 2), the decision maker faces a symmetric bet on the first (resp., second) dimension of the state. Meanwhile, in problem 3, the decision maker faces a symmetric bet on whether the state lies in the diagonal or off-diagonal of the state space.

Note that these decision problems are permutations of each other. Therefore, a decision maker whose cost function is symmetric with respect to all permutations of the state space (e.g., mutual information) will have the same choice accuracy in all three problems. However, for a decision maker who finds multi-dimensional learning challenging, intuition suggests that choice accuracy should be higher in decision problems 1 and 2, which only require learning one dimension of the state, than in decision problem 3, which requires learning both dimensions.

Focusing on the limiting case where learning about a single dimension is nearly costless, we show that this behavioral pattern arises from a natural specification of the nested Shannon model, but cannot be generated by any specification of the neighborhood-based model.

Nested Shannon costs. We consider a nested Shannon cost with the following parameters. The set of attributes is $N = \{U, D, L, R\}$, where each attribute $i \in N$ indexes the corresponding event defined above; the prior $\nu \in \Delta(N)$ over attributes is uniform, so that $\nu(i) = 1/4$ for all $i \in N$; the conditional distributions $\mu_i \in \Delta(\Theta)$ are uniform on the associated events, so that $\mu_i(\theta) = \frac{1}{2}\mathbf{1}(\theta \in i)$ for all $i \in N$ and $\theta \in \Theta$; and there is $\eta > 0$ such that $\eta_i = \eta$ for all $i \in N$. We treat $\zeta > 0$ as a parameter that can be varied, and focus our analysis on the limit $\zeta \to 0$.[45]

This specification captures the idea that the decision maker's subjective representation of the environment treats each dimension of the state as a separate source of uncertainty. Note that ν and $(\mu_i)_{i \in N}$ correspond to the marginal and conditional distributions of the prior π.

[44] For the purposes of this application, setting the reward to equal 1 is just a normalization.

[45] The proof of Proposition 23 provides a full characterization of behavior for all values of $\zeta > 0$.

We assume that $\eta > 0$ is constant across attributes and take $\zeta \to 0$ in order to isolate the effect of multi-dimensionality. In particular, these parametric restrictions ensure that the cost function is symmetric with respect to permutations of each dimension of the state and imply that it is nearly costless for the decision maker to learn about each dimension separately.

Proposition 23. *For each decision problem $j \in \{1,2,3\}$, let $P^j \in \Delta(A_j)^\Theta$ be an optimal stochastic choice rule under the above nested Shannon cost. As $\zeta \to 0$, it holds that:*

$$P^1_\theta(a) \to \mathbf{1}(a(\theta)=1), \qquad P^2_\theta(a) \to \mathbf{1}(a(\theta)=1), \qquad P^3_\theta(a) \to \begin{cases} \frac{e^{1/\eta}}{e^{1/\eta}+1}, & \text{if } a(\theta)=1 \\ \frac{1}{e^{1/\eta}+1}, & \text{otherwise.} \end{cases}$$

The behavioral pattern in Proposition 23 is intuitive. In the limit $\zeta \to 0$, where it becomes nearly free to perfectly distinguish between the events in $\{U, D\}$ and $\{L, R\}$, the choice accuracy in both problems 1 and 2 becomes nearly perfect. Meanwhile, in problem 3, the choice accuracy is governed by the parameter $\eta \in (0, +\infty]$, which determines the cost of learning jointly about both dimensions. Note that this choice accuracy is decreasing in η, with

$$\lim_{\eta \to 0} P^3_\theta(a) = \mathbf{1}(a(\theta)=1) \quad \text{and} \quad \lim_{\eta \to +\infty} P^3_\theta(a) = \frac{1}{2}.$$

Therefore, the parameter $\eta \in (0, +\infty]$ fully controls the difficulty of multi-dimensional learning.

Neighborhood-based costs. We now present an impossibility result demonstrating that the neighborhood-based model (38) cannot produce this behavioral pattern, regardless of the neighborhood structure. In particular, we show that, under any such cost function, if the choice accuracy in both problems 1 and 2 is nearly perfect, then choice accuracy in problem 3 must also be nearly perfect and the cost function itself must be nearly identically zero.

Formally, we call a neighborhood structure $\mathcal{B}$ *nonredundant* if it contains no singleton neighborhoods, i.e., $B \in \mathcal{B}$ implies $|B| \geq 2$. Since singleton neighborhoods do not contribute to the entropy (38), nonredundancy is an innocuous assumption that merely simplifies notation.

Proposition 24. *Fix any index set I, nonredundant neighborhood structure $\mathcal{B}$, and convergent sequence of coefficients $(\kappa^n_i)_{i \in I} \to (\kappa^*_i)_{i \in I} \in \overline{\mathbb{R}}^I_+$. For each decision problem $j \in \{1,2,3\}$ and $n \in \mathbb{N}$, let $P^{j,n} \in \Delta(A_j)^\Theta$ be an optimal stochastic choice rule under the neighborhood-based cost defined via (38) with this neighborhood structure and coefficients $(\kappa^n_i)_{i \in I}$. If it holds that*

$$\lim_{n \to \infty} P^{1,n}_\theta(a) = \mathbf{1}(a(\theta)=1) \quad \text{and} \quad \lim_{n \to \infty} P^{2,n}_\theta(a) = \mathbf{1}(a(\theta)=1),$$

then it also holds that

$$\kappa^*_i = 0 \text{ for all } i \in I \quad \text{and} \quad \lim_{n \to \infty} P^{3,n}_\theta(a) = \mathbf{1}(a(\theta)=1).$$

The contrast between Propositions 23 and 24 reflects the difference between the ways nested Shannon and neighborhood-based cost functions aggregate costs across nests/neighborhoods.

Namely, the nested Shannon model allows us to decouple the operations of learning *about* nests and learning *within* nests, while the neighborhood-based model generally does not.[46]

To illustrate, consider a neighborhood-based cost with neighborhood structure $\mathcal{B}' = \{U, D\}$ and strictly positive coefficients. Under this cost function, it is free to learn about the first dimension of the state, as doing so does not require distinguishing the states within U and D. However, it is costly to learn about the second dimension, which does require distinguishing the states within U and D. This implies that choice accuracy is perfect in problem 1, and imperfect in problems 2 and 3. By symmetric reasoning, the neighborhood-based cost with neighborhood structure $\mathcal{B}'' = \{L, R\}$ makes it costly to learn about the first dimension and free to learn about the second dimension; this implies perfect choice accuracy in problem 2, and imperfect choice accuracy in problems 1 and 3. In either case, learning *about* one dimension of the state requires distinguishing between states *within* the neighborhoods that hold the other dimension fixed, rendering both uni- and multi-dimensional learning costly.

By contrast, under the nested Shannon cost with nests $\{U, D, L, R\}$, the premise of optimal encoding implies that the decision maker can learn exclusively about one dimension of the state while learning nothing about the other. In problems 1 and 2, this flexibility effectively allows the decision maker to *choose* between facing the neighborhood structure $\mathcal{B}'$ or $\mathcal{B}''$, resulting in perfect choice accuracy in both problems. Choice accuracy is only imperfect in problem 3, where learning about both dimensions is necessary. We conclude that this feature of the nested Shannon cost is crucial for modeling the challenges of multi-dimensional learning.

[46] The "deterministic categorization" special case discussed in Section 10.4 is an exception.

Appendix

A Bounds on Lagrange multipliers

In this section we derive bounds on Lagrange multipliers that are useful for both analysis and computations. We denote by $\|\cdot\|_\infty$ the uniform norm on $\mathbb{R}^\Theta$:

$$\|x\|_\infty = \max_{\theta\in\Theta} |x(\theta)|.$$

Lemma 8. *Under f-information, if λ is a Lagrange multiplier for a decision problem $\mathcal{D} = (\pi, A)$, then for all $y \in \mathbb{R}^\Theta$,*

$$\sum_{\theta\in\Theta} \lambda(\theta)y(\theta) \leq (2 + \|y\|_\infty)\left(\max_{a\in A} \|a\|_\infty + |f(\mathbf{1} - y)|\right).$$

Proof. Let $x \in \mathbb{R}^\Theta_+$. As in the proof of Theorem 2, we denote by $\mathcal{P}_x$ the set of vectors $P \in \mathbb{R}^{\Theta\times A}_+$ such that $\sum_{a\in A} P_\theta(a) = x(\theta)$ for all $\theta \in \Theta$. Furthermore, let $V(x)$ be the value of the following optimization problem:

$$\max_{P\in\mathcal{P}_x} \quad \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} P_\theta(a)a(\theta) - I_f(P),$$

where, given $\Theta = \{\theta_1, \ldots, \theta_n\}$,

$$I_f(P) = \inf_{\alpha\in\Delta(A)} \sum_{a\in A} \alpha(a) f\left(\frac{P_{\theta_1}(a)}{\alpha(a)}, \ldots, \frac{P_{\theta_n}(a)}{\alpha(a)}\right).$$

As shown in the proof of Theorem 2, the value function $V\colon \mathbb{R}^\Theta_+ \to \underline{\mathbb{R}}$ is concave and λ is a supergradient of V at $x = \mathbf{1}$. Moreover, for all $x \in \mathbb{R}^\Theta_+$,

$$\max_{a\in A} \sum_{\theta\in\Theta} a(\theta)x(\theta)\pi(\theta) - f(x) \leq V(x) \leq \sum_{\theta\in\Theta} \pi(\theta)x(\theta) \max_{a\in A} a(\theta) - f(x).$$

The lower bound is achieved by restricting attention to choice rules P for which there is $\alpha \in \Delta(A)$ such that $P_\theta(a) = \alpha(a)x(\theta)$ for all $a \in A$ and $\theta \in \Theta$. The upper bound follows from $I_f(P) \geq f(x)$ for all $P \in \mathcal{P}_x$.

Let $y \in \mathbb{R}^\Theta$. By the definition of supergradient, we have:

$$\sum_{\theta\in\Theta} \lambda(\theta)y(\theta) \geq V(\mathbf{1} + y) - V(\mathbf{1}).$$

Using the bounds on V described above, we obtain

$$\begin{aligned} V(\mathbf{1} + y) &\geq -\left(1 + \max_{\theta\in\Theta} |y(\theta)|\right) \max_{a\in A,\theta\in\Theta} |a(\theta)| - |f(\mathbf{1} + y)|, \\ V(\mathbf{1}) &\leq \max_{a\in A,\theta\in\Theta} |a(\theta)|. \end{aligned}$$

The desired result follows. □

Let $B_\epsilon(x)$ be the closed ball of radius $\epsilon > 0$ centered around x, under the uniform norm:

$$B_\epsilon(x) = \left\{ y \in \mathbb{R}^\Theta : \|x - y\|_\infty \leq \epsilon \right\}.$$

Proposition 25. (i) *If f is essentially smooth, then for all decision problems*

$$\mathcal{D}$$

and all $\epsilon > 0$ such that $B_\epsilon(\mathbf{1}) \subseteq \operatorname{dom} f$,

$$\|\lambda\|_\infty \leq \left(\frac{2}{\epsilon} + 1\right) \left(\max_{a \in A} \|a\|_\infty + \max_{x \in B_\epsilon(\mathbf{1})} |f(x)| \right)$$

where λ is the unique Lagrange multiplier associated with $\mathcal{D}$.

(ii) *If H is relatively smooth, then for all decision problems $\mathcal{D}$, with $\pi \in \operatorname{ri}(\operatorname{dom} H)$, and all $\epsilon > 0$ such that $B_\epsilon(\pi) \cap \Delta(\Theta) \subseteq \operatorname{dom} H$,*

$$\|\lambda\|_\infty \leq \left(\frac{2}{\epsilon} + \frac{1}{\min_{\theta \in \Theta} \pi(\theta)}\right) \left(\max_{a \in A} \|a\|_\infty + \max_{p \in B_\epsilon(\pi) \cap \Delta(\Theta)} |H(p) - H(\pi)| \right),$$

where λ is the unique Lagrange multiplier associated with $\mathcal{D}$ such that $\sum_{\theta \in \Theta} \lambda(\theta) = 0$.

Proof. For each $\theta \in \Theta$, let δ_θ be the Dirac measure concentrated on Θ.

(i). The desired result follows from applying Lemma 8 with $y = \pm\epsilon\delta_\theta$.

(ii). Recall that, under posterior separable costs, $f(x) = H(x\pi) - H(\pi)$ for all $x \in \mathbb{R}^\Theta_+$ such that $\sum_\theta x(\theta)\pi(\theta) = 1$. If we apply Lemma 8 with $\pi y = \epsilon(\delta_\theta - \delta_\tau)$ for a pair of states θ and τ, we obtain:

$$\epsilon\frac{\lambda(\theta)}{\pi(\theta)} - \epsilon\frac{\lambda(\tau)}{\pi(\tau)} \leq \left(2 + \frac{\epsilon}{\min_{\rho \in \Theta} \pi(\rho)}\right) \left(\max_{a \in A} \|a\|_\infty + \max_{p \in B_\epsilon(\pi) \cap \Delta(\Theta)} |H(p) - H(\pi)| \right).$$

Using the normalization $\sum_{\rho \in \Theta} \lambda(\rho) = 0$, we obtain:

$$\begin{aligned} -\frac{\lambda(\tau)}{\pi(\tau)} &\leq \left(\frac{2}{\epsilon} + \frac{1}{\min_{\rho \in \Theta} \pi(\rho)}\right) \left(\max_{a \in A} \|a\|_\infty + \max_{p \in B_\epsilon(\pi) \cap \Delta(\Theta)} |H(p) - H(\pi)| \right) \\ \frac{\lambda(\theta)}{\pi(\theta)} &\leq \left(\frac{2}{\epsilon} + \frac{1}{\min_{\rho \in \Theta} \pi(\rho)}\right) \left(\max_{a \in A} \|a\|_\infty + \max_{p \in B_\epsilon(\pi) \cap \Delta(\Theta)} |H(p) - H(\pi)| \right). \end{aligned}$$

Since θ and τ are arbitrary, we obtain

$$|\lambda(\theta)| \leq \left|\frac{\lambda(\theta)}{\pi(\theta)}\right| \leq \left(\frac{2}{\epsilon} + \frac{1}{\min_{\rho \in \Theta} \pi(\rho)}\right) \left(\max_{a \in A} \|a\|_\infty + \max_{p \in B_\epsilon(\pi) \cap \Delta(\Theta)} |H(p) - H(\pi)| \right).$$

The desired result follows. □

The proposition allows us to search for the Lagrange multiplier within a compact set of vectors, instead of the entire $\mathbb{R}^\Theta$. This permits the direct application of computational techniques to find the saddle points of (14)—see, e.g., Bubeck (2015, Chapter 4).

B The size of the consideration set

The findings in Section 7 on inconclusive evidence suggest a broader distinction between posterior separable costs and f-information in terms of the size of the consideration set. This section explores these differences by moving beyond the guess-the-state setting and analyzing abstract decision problems. A corollary of this analysis will be a proof that mutual information is the essentially unique Csiszár cost that is posterior separable.

For the remainder of this section, we fix a state space Θ with cardinality n. We denote by $\mathfrak{D}(\Theta, \pi)$ the set of decision problems with state space Θ and prior π. Each $\mathcal{D} \in \mathfrak{D}(\Theta, \pi)$ can be represented by a $n \times m$ payoff matrix, where m is the number of feasible actions in $\mathcal{D}$. This representation allows us to defined a topology on $\mathfrak{D}(\Theta, \pi)$ as follows: a sequence of decision problems $(\mathcal{D}^l)$ converges to a decision problem $\mathcal{D}$ in $\mathfrak{D}(\Theta, \pi)$ if (i) each $\mathcal{D}^l$ has the same number of feasible actions as $\mathcal{D}$, and (ii) the payoff matrix associated with $\mathcal{D}^l$ converges to the payoff matrix associated with $\mathcal{D}$ as l approaches infinity.

The next result provides a bound to the size of the consideration set under posterior separable costs.

Proposition 26. *Under posterior separable costs, with $\pi \in \operatorname{ri}(\operatorname{dom} H)$, the consideration set has the following properties:*

(i). *Every decision problem $\mathcal{D} \in \mathfrak{D}(\Theta, \pi)$ admits an optimal choice rule P such that $|\operatorname{supp} P_\pi| \leq n$.*

(ii). *If H is relatively smooth, then the set of decision problems $\mathcal{D} \in \mathfrak{D}(\Theta, \pi)$ that admit an optimal choice rule P such that $|\operatorname{supp} P_\pi| > n$ is nowhere dense in $\mathfrak{D}(\Theta, \pi)$.*

Thus, under posterior separable costs, the size of the consideration set is at most the cardinality of the state space, modulo knife-edge cases. Part (i) of the proposition is known in the literature (see, e.g., Denti 2022, Proposition 4). We provide a proof for part (i) and (ii) based on our characterization theorem of optimal information acquisition.

To connect Proposition 26 with the findings on inconclusive evidence from Section 7, observe that in the guess-the-state problem with outside option, there are n possible states and $n+1$ feasible of actions—comprising n risky actions and one safe action. Under symmetric costs (assumed in Proposition 9), all risky actions are taken with the same probability at the optimum. Thus, for inconclusive evidence to emerge, all $n+1$ actions must be taken with positive probability. This requirement conflicts with the fact that there are only n states, as Proposition 26 demonstrates in a broader context.

Under f-information, the consideration set expands in a precise sense:

Proposition 27. *Under f-information, the consideration set has the following properties:*

(i). *Every decision problem $\mathcal{D} \in \mathfrak{D}(\Theta, \pi)$ admits an optimal choice rule P such that $|\operatorname{supp} P_\pi| \leq n+1$.*

(ii). *If f is essentially smooth, then the set of decision problems $\mathcal{D} \in \mathfrak{D}(\Theta, \pi)$ that admit an optimal choice rule P such that $|\operatorname{supp} P_\pi| > n+1$ is nowhere dense in $\mathfrak{D}(\Theta, \pi)$.*

The intuition behind Propositions 26 and 27 is as follows. Under f-information, the optimality conditions for α in the max-min problem (14) require that, for all $a, b \in \operatorname{supp} P_\pi$,

$$f^\star(a\pi - \lambda) = f^\star(b\pi - \lambda).$$

This defines a system with $m = \operatorname{supp} |P_\pi| - 1$ equations and n unknown variables, corresponding to the values of the Lagrange multiplier in each state. If $\operatorname{supp} |P_\pi| > n+1$, the system becomes overdetermined and, generically, has no solution. Under posterior separability, since $H^\star$ is translation invariant, any Lagrange multiplier λ can be shifted by an arbitrary constant $c \in \mathbb{R}$, meaning that $\lambda + \pi c$ is also a valid multiplier. This eliminates one degree of freedom, making the system overdetermined whenever $\operatorname{supp} |P_\pi| > n$.

As the analysis on inconclusive evidence demonstrates, there are settings where, under Csiszár information, the size of the consideration is exactly $n+1$, highlighting a distinction from posterior separable costs. The next result generalizes these findings.

To state the result, let $\mathfrak{D}(\Theta)$ denote the set of decision problems with state space Θ. Each $\mathcal{D} \in \mathfrak{D}(\Theta)$ can be represented by a prior $\pi \in \Delta(\Theta)$ and a $n \times m$ payoff matrix, where m is the number of feasible actions in $\mathcal{D}$. This representation allows us to defined a topology on $\mathfrak{D}(\Theta)$ as follows: a sequence of decision problems $(\mathcal{D}^l)$ converges to a decision problem $\mathcal{D}$ in $\mathfrak{D}(\Theta)$ if (i) each $\mathcal{D}^l$ has the same number of feasible actions as $\mathcal{D}$, and (ii) both the prior and the payoff matrix associated with $\mathcal{D}^l$ converge to those of $\mathcal{D}$ as l approaches infinity.

Proposition 28. *Let $\psi = \phi^\star$ be strictly convex and twice differentiable, with $R_\psi = \psi''/\psi'$ be strictly monotone on a non-empty open interval. If $n \geq 3$, then there exists an open set of decision problems $\mathcal{D} \in \mathfrak{D}(\Theta)$ such that $|\operatorname{supp} P_\pi| = n+1$ at the optimum under ϕ-informativity.*

The proof is constructive: the critical decision problem retains the structure of the guess-the-state problem with an outside option from the Section 7, but with an additional state and an extra action to regulate the value of the Lagrange multiplier. If R_ϕ is strictly monotone on a neighborhood of zero (as in Proposition 7), these additional state and action are unnecessary, and the result holds for $n \geq 2$.

A corollary of Propositions 26 and 28 is that mutual information essentially is the unique intersection of class of posterior separable costs with Csiszár information.

Corollary 6. *A Csiszár cost with $\psi = \phi^\star$ strictly convex and thrice continuously differentiable is posterior separable if and only if it is mutual information, i.e., there is some $\kappa > 0$ such that $\psi(t) = \kappa(\mathrm{e}^{t/\kappa} - 1)$ for all $t \in \mathbb{R}$.*

B.1 Proofs

B.1.1 Proof of Proposition 26-(i)

Let P be an optimal choice rule; denote by m and l the cardinalities of A and $\operatorname{supp} P_\pi$, respectively. If $l \leq n$, then the desired result holds. Suppose therefore that $l > n$. Next we construct another optimal choice rule Q such that $|\operatorname{supp} Q_\pi| < l$. By induction on l, this implies that there exists an optimal choice rule whose consideration set has no more than n actions.

Let (α, λ) be a saddle point of (14) that generates P. Notice that α, which is equal to P_π, is a solution of the following system of linear equations (label the system's independent variable by $\beta \in \mathbb{R}^A$):

$$\sum_{a \in A} \beta(a) \nabla_\theta H^\star(a - \lambda_\pi) = \pi(\theta), \qquad \theta \in \Theta, \tag{39}$$

$$\beta(a) = 0, \qquad a \notin \operatorname{supp} P_\pi. \tag{40}$$

This linear system has $n + m - l$ equations and m unknowns. Since $l > n$, there must be a non-zero vector β such that

$$\sum_{a \in A} \beta(a) \nabla_\theta H^\star(a - \lambda_\pi) = 0, \qquad \theta \in \Theta, \tag{41}$$

$$\beta(a) = 0, \qquad a \notin \operatorname{supp} P_\pi. \tag{42}$$

Note that both β and $-\beta$ are non-zero solutions of (41) and (42). Hence, we can assume without loss of generality that $\beta(a) > 0$ for some $a \in A$.

We define $\gamma \in \mathbb{R}^A$ as follows: for all $a \in A$,

$$\gamma(a) = \alpha(a) - \beta(a) \min_{b : \beta(b) > 0} \frac{\alpha(b)}{\beta(b)}.$$

Claim 1. *The vector γ has the following properties:*

(i). $\gamma(a) \geq 0$ *for all* $a \in A$.

(ii). $\gamma(a) = 0$ *for some* $a \in \operatorname{supp} P_\pi$.

(iii). γ *is a solution of (39) and (40).*

(iv). $\gamma \in \Delta(A)$.

(v). (γ, λ) *is a saddle point of* (14).

Proof. (i). If $\beta(a) \leq 0$, then $\gamma(a) \geq \alpha(a) \geq 0$. If $\beta(a) > 0$, then

$$\gamma(a) \geq 0 \quad \iff \quad \frac{\alpha(a)}{\beta(a)} \geq \min_{b : \beta(b) > 0} \frac{\alpha(b)}{\beta(b)}.$$

Thus, $\gamma(a) \geq 0$ also when $\beta(a) > 0$. This proves (i).

(ii). Take any a, with $\beta(a) > 0$, such that

$$\frac{\alpha(a)}{\beta(a)} = \min_{b:\beta(b)>0} \frac{\alpha(b)}{\beta(b)}.$$

Then, $\gamma(a) = 0$. Moreover, (42) ensures that $a \in \operatorname{supp} P_\pi$. This proves (ii).

(iii). This follows from α being a solution of (39)–(40) and β being a solution of (41)–(42).

(iv). By (i), $\gamma(a) \geq 0$ for all $a \in A$. Since $H^\star$ is translation invariant,

$$\sum_{\theta\in\Theta} \nabla_\theta H^\star(a - \lambda_\pi) = 1.$$

It follows from (39) that

$$\begin{aligned}\sum_{a\in A} \gamma(a) &= \sum_{a\in A} \gamma(a) \left(\sum_{\theta\in\Theta} \nabla_\theta H^\star(a - \lambda_\pi) \right) \\ &= \sum_{\theta\in\Theta} \left(\sum_{a\in A} \gamma(a) \nabla_\theta H^\star(a - \lambda_\pi) \right) = \sum_{\theta\in\Theta} \pi(\theta) = 1.\end{aligned}$$

We conclude that $\gamma \in \Delta(A)$.

(v). Since $\operatorname{supp}\gamma \subseteq \operatorname{supp} P_\pi$ and (P_π, λ) is a saddle point, we have:

$$\min_{a\in\operatorname{supp}\gamma} H^\star(a - \lambda_\pi) \geq \min_{a\in\operatorname{supp} P_\pi} H^\star(a - \lambda_\pi) = \max_{a\in A} H^\star(a - \lambda_\pi).$$

Hence, it follows from (39) that (γ, λ) is a saddle point. □

Let Q be the optimal choice rule generated by (γ, λ). Under posterior separability, $Q_\pi = \gamma$. Thus, $\operatorname{supp} Q_\pi \subseteq \operatorname{supp} P_\pi$ by (40). Moreover, $\operatorname{supp} Q_\pi \neq \operatorname{supp} P_\pi$. Indeed, by (ii) of Claim 1, there exists $a \in \operatorname{supp} P_\pi$ such that $\gamma(a) = 0$. Given that $\gamma(a) = 0$, we must have $Q_\pi(a) = 0$. It follows that $\operatorname{supp} Q_\pi \neq \operatorname{supp} P_\pi$. Overall, we conclude that $\operatorname{supp} Q_\pi$ is a proper subset of $\operatorname{supp} P_\pi$. This shows that $|\operatorname{supp} Q_\pi| < |\operatorname{supp} P_\pi| = l$, as desired. This concludes the proof of part (i) of Proposition 26.

B.1.2 Proof of Proposition 26-(ii)

For every decision problem $\mathcal{D} \in \mathcal{D}(\Theta, \pi)$, we fix an enumeration of the action set, $A = \{a_1, \ldots, a_m\}$, where m is the number of feasible actions. With a slight abuse of notation, we identify α with an element of $\Delta(\{1, \ldots, m\})$.

We first prove a continuity property of the Lagrange multiplier. We denote by $\lambda_\mathcal{D}$ the unique Lagrange multiplier associated with a decision problem $\mathcal{D}$, under the normalization that $\sum_{\theta\in\Theta} \lambda_\mathcal{D}(\theta) = 0$. Uniqueness comes from H being relatively smooth.

Claim 2. *If $\mathcal{D}^l \to \mathcal{D}$, then $\lambda_{\mathcal{D}^l} \to \lambda_\mathcal{D}$.*

Proof. By the definition of convergence between decision problems each $\mathcal{D}^l$ has the same number of action as $\mathcal{D}$, which we denote by m. By Proposition 25, the sequence $(\lambda_{\mathcal{D}^l})$ is bounded. Thus, we can assume that it converges to some λ without loss of generality. For each l, let $(\alpha^l, \lambda_{\mathcal{D}^l})$ be a saddle point for the decision problem $\mathcal{D}^l$. Since the sequence (α^l) is bounded, we can assume that it converges to some α in $\Delta(\{1,\dots,m\})$ without loss of generality. By the continuity of $H^\star$ and $\nabla H^\star$, the pair (α, λ) is a saddle point for the decision problem $\mathcal{D}$. Thus, $\lambda = \lambda_{\mathcal{D}}$. This proves that $\lambda_{\mathcal{D}^l} \to \lambda_{\mathcal{D}}$. □

Let $\mathbb{D}$ be the set of decision problems that admits an optimal choice rule P such that $|\operatorname{supp} P_\pi| > n$. Let $\operatorname{cl}\mathbb{D}$ be the closure of $\mathbb{D}$.

Claim 3. *For each $\mathcal{D} \in \operatorname{cl}\mathbb{D}$, there is a set of actions B, with $|B| > n$, such that*

$$\max_{a\in A} H^\star(a - (\lambda_{\mathcal{D}})_\pi) = \min_{a\in B} H^\star(a - (\lambda_{\mathcal{D}})_\pi).$$

Proof. Let $(\mathcal{D}^l)$ be a sequence in $\mathbb{D}$ such that $\mathcal{D}^l \to \mathcal{D}$. By the definition of convergence between decision problems, each $\mathcal{D}^l$ has the same number of action as $\mathcal{D}$, which we denote by m. Each decision problem $\mathcal{D}^l = (\Theta, \pi, A^l)$ has a saddle point $(\alpha^l, \lambda_{\mathcal{D}^l})$ such that $|\operatorname{supp}\alpha^l| > n$. Possibly passing to a subsequence, we can assume that $\operatorname{supp}\alpha^l = \operatorname{supp}\alpha^{l+1} \subseteq \{1,\dots,m\}$ for all l; accordingly, We define $I = \operatorname{supp}\alpha^1$. For all l, since $(\alpha^l, \lambda_{\mathcal{D}^l})$ is a saddle point, we obtain:

$$\max_{i=1,\dots,m} H^\star(a_i^l - (\lambda_{\mathcal{D}^l})_\pi) = \min_{i\in I} H^\star(a_i^l - (\lambda_{\mathcal{D}^l})_\pi).$$

By Claim 5, $\lambda_{\mathcal{D}^l} \to \lambda_{\mathcal{D}}$. Since $H^\star$ is continuous,

$$\max_{i=1,\dots,m} H^\star(a_i - (\lambda_{\mathcal{D}})_\pi) = \min_{i\in I} H^\star(a_i - (\lambda_{\mathcal{D}})_\pi).$$

Hence, we can choose $B = \{a_i : i \in I\}$. □

Take an arbitrary decision problem $\mathcal{D} \in \operatorname{cl}\mathbb{D}$ and an arbitrary $\epsilon > 0$. Let $A = \{a_1,\dots,a_m\}$ be an enumeration of the action set. By Proposition 26-(i), the decision problem $\mathcal{D}$ has a saddle point $(\alpha, \lambda_{\mathcal{D}})$ such that $|\operatorname{supp}\alpha| \le n$. We define $\mathcal{D}^\epsilon = (\Theta, \pi, A^\epsilon)$ as follows: for all $\theta \in \Theta$ and $i = 1,\dots,m$,

$$a_i^\epsilon(\theta) = \begin{cases} a_i(\theta) & \text{if } i \in \operatorname{supp}\alpha, \\ a_i(\theta) - \epsilon & \text{if } i \notin \operatorname{supp}\alpha. \end{cases}$$

Note that $(\alpha, \lambda_{\mathcal{D}})$ is a saddle point of $\mathcal{D}^\epsilon$. Thus, in particular, $\lambda_{\mathcal{D}} = \lambda_{\mathcal{D}^\epsilon}$ and $(\alpha, \lambda_{\mathcal{D}^\epsilon})$ is a saddle point of $\mathcal{D}^\epsilon$. It follows that, for every $i \notin \operatorname{supp}\alpha$,

$$\begin{aligned} H^\star(a_i^\epsilon - (\lambda_{\mathcal{D}^\epsilon})_\pi) = H^\star(a_i^\epsilon - (\lambda_{\mathcal{D}})_\pi) &< H^\star(a_i - (\lambda_{\mathcal{D}})_\pi) \\ &\le \min_{j\in\operatorname{supp}\alpha} H^\star(a_j - (\lambda_{\mathcal{D}})_\pi) \\ &= \min_{j\in\operatorname{supp}\alpha} H^\star(a_j^\epsilon - (\lambda_{\mathcal{D}}^\epsilon)_\pi) = \max_{i=1,\dots,n} H^\star(a_i^\epsilon - (\lambda_{\mathcal{D}^\epsilon})_\pi), \end{aligned}$$

where we use the fact that $H^\star$ is strictly increasing (Lemma 7). We deduce from Claim 6 that $\mathcal{D}^\epsilon \notin \operatorname{cl}\mathbb{D}$. Since $\mathcal{D} \in \operatorname{cl}\mathbb{D}$ and $\epsilon > 0$ are arbitrary, we conclude that $\operatorname{cl}\mathbb{D}$ has empty interior.

B.1.3 Proof of Proposition 27-(i)

The structure of this proof parallels that of Proposition 26-(i). We repeat several steps to emphasize both the analogies and the differences.

Let P be an optimal choice rule, and denote by m and l the cardinalities of A and $\operatorname{supp} P_\pi$, respectively. If $l \leq n+1$, then the desired result holds. Suppose therefore that $l > n+1$. Next we construct another optimal choice rule Q such that $|\operatorname{supp} Q_\pi| < l$. By induction on l, this implies that there exists an optimal choice rule whose consideration set has no more than $n+1$ actions.

Let (α, λ) be a saddle point of (14) that generates P. Notice that α is a solution of the following system of linear equations (label the system's independent variable by $\beta \in \mathbb{R}^A$):

$$\sum_{a \in A} \beta(a) \nabla_\theta f^\star(a\pi - \lambda) = 1, \qquad \theta \in \Theta, \tag{43}$$

$$\sum_{a \in \operatorname{supp} P_\pi} \beta(a) = \sum_{a \in \operatorname{supp} P_\pi} \alpha(a), \tag{44}$$

$$\beta(a) = \alpha(a), \qquad a \notin \operatorname{supp} P_\pi. \tag{45}$$

This linear system has $n+1+m-l$ equations and m unknowns. Since $l > n+1$, there must be a non-zero vector β such that

$$\sum_{a \in A} \beta(a) \nabla_\theta f^\star(a\pi - \lambda) = 0, \qquad \theta \in \Theta, \tag{46}$$

$$\sum_{a \in \operatorname{supp} P_\pi} \beta(a) = 0, \tag{47}$$

$$\beta(a) = 0, \qquad a \notin \operatorname{supp} P_\pi. \tag{48}$$

Note that both β and $-\beta$ are non-zero solutions of (46)–(48). Hence, we can assume without loss of generality that $\beta(a) > 0$ for some $a \in A$.

We define $\gamma \in \mathbb{R}^A$ as follows: for all $a \in A$,

$$\gamma(a) = \alpha(a) - \beta(a) \min_{b : \beta(b) > 0} \frac{\alpha(b)}{\beta(b)}.$$

Claim 4. *The vector γ has the following properties:*

(i). $\gamma(a) \geq 0$ *for all* $a \in A$.

(ii). $\gamma(a) = 0$ *for some* $a \in \operatorname{supp} P_\pi$.

(iii). γ *is a solution of (43)–(45).*

(iv). $\gamma \in \Delta(A)$.

(v). *If* $\gamma(a) > 0$, *then* $\alpha(a) > 0$.

(vi). (γ, λ) *is a saddle point of* (14).

Proof. (i). If $\beta(a) \leq 0$, then $\gamma(a) \geq \alpha(a) \geq 0$. If $\beta(a) > 0$, then

$$\gamma(a) \geq 0 \quad \Longleftrightarrow \quad \frac{\alpha(a)}{\beta(a)} \geq \min_{b:\beta(b)>0} \frac{\alpha(b)}{\beta(b)}.$$

Thus, $\gamma(a) \geq 0$ also when $\beta(a) > 0$. This proves (i).

(ii). Take any a, with $\beta(a) > 0$, such that

$$\frac{\alpha(a)}{\beta(a)} = \min_{b:\beta(b)>0} \frac{\alpha(b)}{\beta(b)}.$$

Then, $\gamma(a) = 0$. Moreover, (48) ensures that $a \in \operatorname{supp} P_\pi$. This proves (ii).

(iii). This follows from α being a solution of (43)–(45) and β being a solution of (46)–(48).

(iv). By (i), $\gamma(a) \geq 0$ for all $a \in A$. By (44) and (45),

$$\begin{aligned}\sum_{a\in A} \gamma(a) &= \sum_{a\in \operatorname{supp} P_\pi} \gamma(a) + \sum_{a\notin \operatorname{supp} P_\pi} \gamma(a) \\ &= \sum_{a\in \operatorname{supp} P_\pi} \alpha(a) + \sum_{a\notin \operatorname{supp} P_\pi} \alpha(a) = \sum_{a\in A} \alpha(a) = 1.\end{aligned}$$

It follows that $\gamma \in \Delta(A)$.

(v). For $a \notin \operatorname{supp} P_\pi$, we have $\gamma(a) = \alpha(a)$. For $a \in \operatorname{supp} P_\pi$, we have:

$$P_\pi(a) = \alpha(a) \sum_{\theta\in\Theta} \pi(\theta) \nabla_\theta f^\star(a\pi - \lambda).$$

Thus, $\operatorname{supp} P_\pi(a) \subseteq \operatorname{supp} \alpha$. Thus, in any case, $\gamma(a) > 0$ implies $\alpha(a) > 0$, as desired.

(vi). Since $\operatorname{supp} \gamma \subseteq \operatorname{supp} \alpha$ and (α, λ) is a saddle point, we have:

$$\min_{a\in\operatorname{supp}\gamma} f^\star(a\pi - \lambda) \geq \min_{a\in\operatorname{supp}\alpha} f^\star(a\pi - \lambda) = \max_{a\in A} f^\star(a\pi - \lambda).$$

Hence, it follows from (43) that (γ, λ) is a saddle point. □

Let Q be the optimal choice rule generated by (γ, λ). We claim that $|\operatorname{supp} Q_\pi| < l$. To verify this claim, first we observe that $\operatorname{supp} Q_\pi \subseteq \operatorname{supp} P_\pi$. Indeed, $Q_\pi(a) > 0$ implies

$$\gamma(a) > 0 \quad \text{and} \quad \sum_{\theta\in\Theta} \pi(\theta) \nabla_\theta f^\star(a\pi - \lambda) > 0.$$

Since $\gamma(a) > 0$ implies $\alpha(a) > 0$ (see (v) of Claim 4), we obtain that $Q_\pi(a) > 0$ implies

$$P_\pi(a) = \alpha(a) \sum_{\theta\in\Theta} \pi(\theta) \nabla_\theta f^\star(a\pi - \lambda) > 0.$$

This proves that $\operatorname{supp} Q_\pi \subseteq \operatorname{supp} P_\pi$. We also note that $\operatorname{supp} Q_\pi \neq \operatorname{supp} P_\pi$. Indeed, by (ii) of Claim 4, there exists $a \in \operatorname{supp} P_\pi$ such that $\gamma(a) = 0$. Given that $\gamma(a) = 0$, we must have $Q_\pi(a) = 0$. It follows that $\operatorname{supp} Q_\pi \neq \operatorname{supp} P_\pi$. Overall, we conclude that $\operatorname{supp} Q_\pi$ is a proper subset of $\operatorname{supp} P_\pi$. This shows that $|\operatorname{supp} Q_\pi| < |\operatorname{supp} P_\pi| = l$, as desired. This concludes the proof of part (i) of Proposition 27.

B.1.4 Proof of Proposition 27-(ii)

The structure of this proof parallels that of Proposition 26-(ii). We repeat several steps to emphasize both the analogies and the differences.

For every decision problem $\mathcal{D} \in \mathcal{D}(\Theta, \pi)$, we fix an enumeration of the action set, $A = \{a_1, \dots, a_m\}$, where m is the number of feasible actions. With a slight abuse of notation, we identify α with an element of $\Delta(\{1, \dots, m\})$.

We first prove a continuity property of the Lagrange multiplier. We denote by $\lambda_{\mathcal{D}}$ the unique Lagrange multiplier associated with a decision problem $\mathcal{D}$. Uniqueness comes from the fact that f is essentially smooth.

Claim 5. *If $\mathcal{D}^l \to \mathcal{D}$, then $\lambda_{\mathcal{D}^l} \to \lambda_{\mathcal{D}}$.*

Proof. By the definition of convergence between decision problem, each $\mathcal{D}^l$ has the same number of action as $\mathcal{D}$, which we denote by m. By Proposition 25, the sequence $(\lambda_{\mathcal{D}^l})$ is bounded. Thus, we can assume that it converges to some λ without loss of generality. For each l, let $(\alpha^l, \lambda_{\mathcal{D}^l})$ be a saddle point for the decision problem $\mathcal{D}^l$. Since the sequence (α^l) is bounded, we can assume that it converges to some α in $\Delta(\{1, \dots, m\})$ without loss of generality. By the continuity of $f^\star$ and $\nabla f^\star$, the pair (α, λ) is a saddle point for the decision problem $\mathcal{D}$. Thus, $\lambda = \lambda_{\mathcal{D}}$. This proves that $\lambda_{\mathcal{D}^l} \to \lambda_{\mathcal{D}}$. □

Let $\mathbb{D}$ be the set of decision problems that admits an optimal choice rule P such that $|\operatorname{supp} P_\pi| > n + 1$. Let $\operatorname{cl} \mathbb{D}$ be the closure of $\mathbb{D}$.

Claim 6. *For each $\mathcal{D} \in \operatorname{cl} \mathbb{D}$, there is a set of actions $B \subseteq A$, with $|B| > n$, such that*

$$\max_{a \in A} f^\star(a\pi - \lambda_{\mathcal{D}}) = \min_{a \in B} f^\star(a\pi - \lambda_{\mathcal{D}}).$$

Proof. Let $(\mathcal{D}^l)$ be a sequence in $\mathbb{D}$ such that $\mathcal{D}^l \to \mathcal{D}$. By the definition of convergence between decision problems, each $\mathcal{D}^l$ has the same number of action as $\mathcal{D}$, which we denote by m. Each decision problem $\mathcal{D}^l$ has a saddle point $(\alpha^l, \lambda_{\mathcal{D}^l})$ such that $|\operatorname{supp} \alpha^l| > n$. Possibly passing to a subsequence, we can assume that $\operatorname{supp} \alpha^l = \operatorname{supp} \alpha^{l+1}$ for all l; accordingly, we define $I = \operatorname{supp} \alpha^1$. For all l, since $(\alpha^l, \lambda_{\mathcal{D}^l})$ is a saddle point, we must have:

$$\max_{i=1,\dots,m} f^\star(a_i^l \pi - \lambda_{\mathcal{D}^l}) = \min_{i \in I} f^\star(a_i^l \pi - \lambda_{\mathcal{D}^l}).$$

By Claim 5, $\lambda_{\mathcal{D}^l} \to \lambda_{\mathcal{D}}$. Since $f^\star$ is continuous,

$$\max_{i=1,\dots,m} f^\star(a_i \pi - \lambda_{\mathcal{D}}) = \min_{i \in I} f^\star(a_i \pi - \lambda_{\mathcal{D}}).$$

Hence, we can choose $B = \{a_i : i \in I\}$. □

Take an arbitrary decision problem $\mathcal{D} \in \operatorname{cl} \mathbb{D}$ and an arbitrary $\epsilon > 0$. Let $A = \{a_1, \dots, a_m\}$ be an enumeration of the action set. By Proposition 27-(i), the decision problem $\mathcal{D}$ has a

saddle point $(\alpha, \lambda_{\mathcal{D}})$ such that $|\operatorname{supp}\alpha| \leq n+1$. (Recall that, since f is essentially smooth, $\operatorname{supp}\alpha$ coincides with the consideration set). We define $\mathcal{D}^\epsilon = (\Theta, \pi, A^\epsilon)$ as follows: for all $\theta \in \Theta$ and $i = 1, \ldots, m$,

$$a_i^\epsilon(\theta) = \begin{cases} a_i(\theta) & \text{if } i \in \operatorname{supp}\alpha, \\ a_i(\theta) - \epsilon & \text{if } i \notin \operatorname{supp}\alpha. \end{cases}$$

Note that $(\alpha, \lambda_{\mathcal{D}})$ is a saddle point of $\mathcal{D}^\epsilon$. Thus, in particular, $\lambda_{\mathcal{D}} = \lambda_{\mathcal{D}^\epsilon}$. It follows that, for every $i \notin \operatorname{supp}\alpha$,

$$\begin{aligned} f^\star(a_i^\epsilon\pi - \lambda_{\mathcal{D}^\epsilon}) = f^\star(a_i^\epsilon\pi - \lambda_{\mathcal{D}}) &< f^\star(a_i\pi - \lambda_{\mathcal{D}}) \\ &\leq \min_{j \in \operatorname{supp}\alpha} f^\star(a_j\pi - \lambda_{\mathcal{D}}) \\ &= \min_{j \in \operatorname{supp}\alpha} f^\star(a_j^\epsilon\pi - \lambda_{\mathcal{D}^\epsilon}) = \max_{i=1,\ldots,n} f^\star(a_i^\epsilon\pi - \lambda_{\mathcal{D}^\epsilon}), \end{aligned}$$

where we use the fact $f^\star$ is strictly increasing. We deduce from Claim 6 that $\mathcal{D}^\epsilon \notin \operatorname{cl}\mathbb{D}$. Since $\mathcal{D} \in \operatorname{cl}\mathbb{D}$ and $\epsilon > 0$ are arbitrary, we conclude that $\operatorname{cl}\mathbb{D}$ has empty interior.

B.1.5 Proof of Proposition 28

Let $\Theta = \{\theta_1, \ldots, \theta_n\}$ and $A = \{a_1, \ldots, a_m\}$, with $m > n \geq 3$. The core of the proof is constructing a decision problem $\mathcal{D} = (\Theta, \pi, A)$ and a pair $(\alpha, \lambda) \in \Delta(A) \times \mathbb{R}^\Theta$ such that:

(i). (α, λ) is the unique saddle point of $\mathcal{D}$.

(ii). $\operatorname{supp}\alpha = \{a_1, \ldots, a_{n+1}\}$.

(iii). For all $j > n+1$,

$$\sum_{i=1}^n \psi(a_j(\theta_i) - \lambda_\pi(\theta_i)) < \sum_{i=1}^n \psi(a_{n+1}(\theta_i) - \lambda_\pi(\theta_i)),$$

where $\lambda_\pi(\theta_i) = \lambda(\theta_i)/\pi(\theta_i)$.

Toward this goal, we introduce parametrizations for $\mathcal{D}$ and (α, λ). The decision problem $\mathcal{D}$ is parametrized as follows:

- Given $\bar{\pi} \in (0,1)$, each state $\theta_1, \ldots, \theta_{n-1}$ has prior probability $\bar{\pi}/(n-1)$, and state θ_n has prior probability $1 - \bar{\pi}$.
- For $j = 1, \ldots, n-1$, action a_j pays $\rho > 0$ in state θ_j, pays $z \in \mathbb{R}$ in state θ_n, and pays 0 in every other state.
- Action a_n pays z in state θ_n and $\sigma \in (0, \rho)$ in every other state.
- Action a_{n+1} pays $y \in \mathbb{R}$ in state θ_n and $x \in \mathbb{R}$ in every other state.
- For $j = n+2, \ldots, m$, action a_j pays $y - 1$ in state θ_n and $x - 1$ in every other state.

The pair (α, λ) is parametrized as follows:

- Given $\bar{\alpha} \in (0,1)$, $\alpha(a_1) = \ldots = \alpha(a_n) = \bar{\alpha}/n$ and $\alpha(a_{n+1}) = 1 - \bar{\alpha}$.
- Given $\bar{\lambda} \in \mathbb{R}$, $\lambda(\theta_1) = \ldots = \lambda(\theta_{n-1}) = \bar{\lambda}(n-1)/\bar{\pi}$ and $\lambda(\theta_n) = 0$

To sum up, $\mathcal{D}$ is parametrized by $\bar{\pi} \in (0,1)$, $\rho > 0$, $\sigma \in (0, \rho)$, and $x, y, x \in \mathbb{R}$. The pair (α, λ) is parametrized by $\bar{\alpha} \in (0,1)$ and $\bar{\lambda} \in \mathbb{R}$. By construction, $\operatorname{supp} \alpha = \{a_1, \ldots, a_{n+1}\}$. In addition, action a_j, with $j = n+2, \ldots, m$, is dominated by action a_{n+1}. Thus,

$$\sum_{i=1}^{n} \psi(a_j(\theta_i) - \lambda_\pi(\theta_i)) < \sum_{i=1}^{n} \psi(a_{n+1}(\theta_i) - \lambda_\pi(\theta_i)), \tag{49}$$

It remains to choose parameter values so that (α, λ) is the unique saddle point of $\mathcal{D}$.

For (α, λ) to be a saddle point, the necessary and sufficient conditions are as follows. For λ to be optimal given α, we need:

$$\frac{\bar{\alpha}}{n}\psi'\left(\rho - \bar{\lambda}\right) + \frac{\bar{\alpha}(n-2)}{n}\psi'\left(0 - \bar{\lambda}\right) + \frac{\bar{\alpha}}{n}\psi'\left(\sigma - \bar{\lambda}\right) + (1-\bar{\alpha})\psi'\left(x - \bar{\lambda}\right) = \psi'(0), \tag{50}$$

$$\bar{\alpha}\psi'(z) + (1-\bar{\alpha})\psi'(y) = \psi'(0). \tag{51}$$

For α to be optimal given λ, we need:

$$\frac{\bar{\pi}}{n-1}\psi\left(\rho - \bar{\lambda}\right) + \frac{\bar{\pi}(n-2)}{n-1}\psi\left(0 - \bar{\lambda}\right) + (1-\bar{\pi})\psi(z) = \bar{\pi}\psi\left(x - \bar{\lambda}\right) + (1-\bar{\pi})\psi(y), \tag{52}$$

$$\bar{\pi}\psi\left(\sigma - \bar{\lambda}\right) + (1-\bar{\pi})\psi(z) = \bar{\pi}\psi\left(x - \bar{\lambda}\right) + (1-\bar{\pi})\psi(y). \tag{53}$$

We denote by X a non-empty open interval of the real line such that R_ψ is strictly monotone on X. To simplify the exposition, we assume that $X \cap (-\infty, 0) \neq \varnothing$. Similar arguments apply to the case in which $X \cap (0, +\infty) \neq \varnothing$.

We choose $\bar{\lambda} > 0$ such that R_ψ is strictly monotone on

$$\left(-\bar{\lambda} - \epsilon, -\bar{\lambda} + \epsilon\right)$$

for all ϵ sufficiently small. Take $\rho \in (0, \epsilon)$ and $\sigma \in (0, \rho)$ such that

$$\frac{1}{n-1}\psi\left(\rho - \bar{\lambda}\right) + \frac{n-2}{n-1}\psi\left(0 - \bar{\lambda}\right) = \psi\left(\sigma - \bar{\lambda}\right). \tag{54}$$

Note that σ admits an explicit expression:

$$\sigma = \bar{\lambda} + \psi^{-1}\left(\frac{1}{n-1}\psi\left(\rho - \bar{\lambda}\right) + \frac{n-2}{n-1}\psi\left(0 - \bar{\lambda}\right)\right).$$

The fact that $\sigma \in (0, \rho)$ comes from ψ being strictly convex and increasing. By choosing ϵ sufficiently small, we can be sure that

$$\rho - \bar{\lambda} < 0 \quad \text{and} \quad \sigma - \bar{\lambda} < 0.$$

To satisfy (50), we impose the restriction that $x > \bar{\lambda}$, and we define $\bar{\alpha}$ by:

$$\bar{\alpha} = \frac{n\left(\psi'\left(x - \bar{\lambda}\right) - \psi'(0)\right)}{n\psi'\left(x - \bar{\lambda}\right) - \psi'\left(\rho - \bar{\lambda}\right) - (n-2)\psi'\left(0 - \bar{\lambda}\right) - \psi'\left(\sigma - \bar{\lambda}\right)}.$$

Note that, as $x \downarrow \bar{\lambda}$, we have $\bar{\alpha} \downarrow 0$. Next, to satisfy (53), we impose the restrictions that $z > 0$ and $y < 0$, and we define $\bar{\pi}$ by:

$$\bar{\pi} = \frac{\psi(z) - \psi(y)}{\psi\left(x - \bar{\lambda}\right) + \psi(z) - \psi\left(\sigma - \bar{\lambda}\right) - \psi(y)}.$$

Then, thanks to (54), equation (52) is automatically satisfied. Finally, to satisfy (51), we define y by:

$$y = \frac{(\psi')^{-1}\left(\psi'(0) - \bar{\alpha}\psi'(z)\right)}{1 - \bar{\alpha}}.$$

Note that y is well defined as long as we choose z sufficiently close to zero. Overall, this parameter choice ensures that (α, λ) is a saddle point of $\mathcal{D}$.

Now we prove that (α, λ) is the unique saddle point of $\mathcal{D}$. We will use the following result:

Claim 7. *We have:*

$$\frac{1}{n-1}\psi'\left(\rho - \bar{\lambda}\right) + \frac{n-2}{n-1}\psi'\left(0 - \bar{\lambda}\right) \neq \psi'\left(\sigma - \bar{\lambda}\right). \tag{55}$$

Proof. Notice that

$$\frac{\partial}{\partial x}\psi'\left(\psi^{-1}(x)\right) = R_\psi\left(\psi^{-1}(x)\right).$$

Since R_ψ is strictly monotone on the interval $\left(-\bar{\lambda} - \epsilon, -\bar{\lambda} + \epsilon\right)$, the composite function $\psi' \circ \psi^{-1}$ is strictly convex or strictly concave on the interval $\left(\psi\left(-\bar{\lambda} - \epsilon\right), \psi\left(-\bar{\lambda} + \epsilon\right)\right)$. Then, the desired result follows from applying $\psi' \circ \psi^{-1}$ to both sides of (54). □

Take any other saddle point (β, λ)—since f is essentially smooth, the Lagrange multiplier is unique. By (49), we must have $\beta(a_i) = 0$ for all $i > n+2$. Next, we verify that $\beta(a_i) = \beta(a_j)$ for all $i, j = 1, \ldots, n-1$. To show this, we can use the optimality conditions for the Lagrange multiplier in states θ_i and θ_j, which imply:

$$\beta(a_i)\left(\psi'\left(\rho - \bar{\lambda}\right) - \psi'\left(0 - \bar{\lambda}\right)\right) = \beta(a_j)\left(\psi'\left(\rho - \bar{\lambda}\right) - \psi'\left(0 - \bar{\lambda}\right)\right).$$

We conclude that $\beta(a_i) = \beta(a_j)$.

Next we argue that $\beta(a_{n+1}) = 1 - \bar{\alpha}$. This follows immediately from the optimality condition for the Lagrange multiplier in state θ_n:

$$(1 - \beta(a_{n+1}))\psi'(z) + \beta(a_{n+1})\psi'(y) = \psi'(0) \quad \Longrightarrow \quad \beta(a_{n+1}) = 1 - \bar{\alpha}.$$

Hence, we are done with proving uniqueness as soon as we show that $\beta(a_n) = \beta(a_1)$. To do so, we use the optimality condition for the Lagrange multiplier in the first state, which, together with (50), imply:

$$\begin{aligned}&\frac{(n-1)\beta(a_1)}{\bar{\alpha}}\left(\frac{1}{n-1}\psi'\left(\rho-\bar{\lambda}\right)+\frac{n-2}{n-1}\psi'\left(0-\bar{\lambda}\right)\right)+\frac{\beta(a_n)}{\bar{\alpha}}\psi'\left(\sigma-\bar{\lambda}\right)\\=&\frac{n-1}{n}\left(\frac{1}{n-1}\psi'\left(\rho-\bar{\lambda}\right)+\frac{n-2}{n-1}\psi'\left(0-\bar{\lambda}\right)\right)+\frac{1}{n}\psi'\left(\sigma-\bar{\lambda}\right).\end{aligned}$$

By (55), we must have $\beta(a_1) = \beta(a_n)$. This proves that (α, λ) is the unique saddle point of $\mathcal{D}$.

The next result concludes the proof of the proposition.

Claim 8. *If $(\mathcal{D}^l)$ is a sequence of decision problems that converges to $\mathcal{D}$ in $\mathfrak{D}(\Theta)$, then, for all l sufficiently large, the consideration set has $n+1$ elements at the optimum.*

Proof. By the definition of convergence between decision problems, each $\mathcal{D}^l$ has the same number of action as $\mathcal{D}$, which we denote by m. For each l, let (α^l, λ^l) be a saddle point of $\mathcal{D}^l$. We identify each α^l with an element $\Delta(\{1, \ldots, m\})$. We apply the same convention to α.

We claim that $\alpha^l \to \alpha$ and $\lambda^l \to \lambda$. The sequence (α^l) is bounded because $\Delta(\{1, \ldots, m\})$ is compact. The sequence (λ^l) is bounded by Proposition 25. Thus, we can assume that $\alpha^l \to \alpha^*$ and $\lambda^l \to \lambda^*$ for some $\alpha^* \in \Delta(\{1, \ldots, m\})$ and $\lambda^* \in \mathbb{R}^\Theta$. By continuity of ψ and ψ', (α^*, λ^*) is a saddle point of $\mathcal{D}$. Since (α, λ) is the unique saddle point of $\mathcal{D}$, we deduce that $(\alpha^*, \lambda^*) = (\alpha, \lambda)$.

Since $\alpha^l \to \alpha$, $\operatorname{supp} \alpha^l \supseteq \operatorname{supp} \alpha$ for all l sufficiently large. In addition, by (49), for all l sufficiently large,

$$\max_{j=n+2,\ldots,m} \sum_{i=1}^{n} \psi\left(a_j^l(\theta_i) - \lambda_{\pi^l}^l(\theta_i)\right) < \sum_{i=1}^{n} \psi\left(a_{n+1}^l(\theta_i) - \lambda_{\pi^l}^l(\theta_i)\right).$$

This proves that $\operatorname{supp} \alpha^l \subseteq \alpha$ for all l sufficiently large. It follows that $\operatorname{supp} \alpha^l = \alpha$ for all l sufficiently large. □

B.1.6 Proof of Corollary 6

The "only if" direction is well known. For the "if" direction, suppose that ϕ-informativity is posterior separable. By Propositions 26 and 28, the Arrow-Pratt coefficient is not strictly monotone on any open interval. Since ψ is thrice continuously differentiable, this implies that R_ψ is continuously differentiable, and therefore constant: there exists $\kappa > 0$ such that $R_\psi(x) = 1/\kappa$ for all $x \in \mathbb{R}$. We obtain that $\psi(x) = \kappa(e^{x/\kappa} - 1)$ for all $x \in \mathbb{R}$, which in turn implies that $\phi(x) = \psi^\star(x) = \kappa(x \log x - x + 1)$ for all $x \in \mathbb{R}_+$.

C Proofs of the results in the main text

C.1 Proof of Lemma 2

(i). The result is a consequence of the data-processing inequality for f-divergences (Lemma 1):

$$\begin{aligned} I_f(K \circ P) &= \inf_{\beta \in \Delta(Z)} D_f(K \circ P \| \beta) \\ &\leq \inf_{\alpha \in \Delta(\Omega)} D_f(K \circ P \| K \circ \alpha) \leq \inf_{\alpha \in \Delta(\Omega)} D_f(P \| \alpha) = I_f(P). \end{aligned}$$

(ii). If $I_f(P) = +\infty$, then $I_f(P) = D_f(P\|\alpha)$ for all $\alpha \in \Delta(\Omega)$. Suppose instead that $I_f(P) < +\infty$. By Lemma 1, $D_f(P\|\alpha)$ is lower semicontinuous in α. Thus, since $\Delta(\Omega)$ is compact, there exists $\alpha \in \Delta(\Omega)$ such that $D_f(P\|\alpha) = \min_{\beta \in \Delta(\Omega)} D_f(P\|\beta)$.

(iii). To verify convexity, take $P, Q \in \Delta(\Omega)^\Theta$ and $t \in [0,1]$. By (ii) above, there are $\alpha, \beta \in \Delta(\Omega)$ such that $I_f(P) = D_f(P\|\alpha)$ and $I_f(Q) = D_f(Q\|\beta)$. Then, since D_f is convex on $\Delta(\Omega)^\Theta \times \Delta(\Omega)$ (Lemma 1),

$$\begin{aligned} tI_f(P) + (1-t)I_f(Q) &= tD_f(P\|\alpha) + (1-t)D_f(Q\|\beta) \\ &\geq D_f(tP + (1-t)Q \| t\alpha + (1-t)\beta) \\ &\geq I_f(tP + (1-t)Q). \end{aligned}$$

We conclude that I_f is convex.

To verify lower semicontinuity, let (P^n) be a sequence in $\Delta(\Omega)^\Theta$ with limit P. By (i), for every n there is $\alpha^n \in \Delta(\Omega)$ such that $I_f(P^n) = D_f(P^n\|\alpha^n)$. Since $\Delta(\Omega)$ is compact, we can assume that the sequence (α^n) is convergent without loss of generality. Setting $\alpha = \lim_{n\to+\infty} \alpha^n$, we obtain

$$\liminf_{n\to+\infty} I_f(P^n) = \liminf_{n\to+\infty} D_f(P^n\|\alpha^n) \geq D_f(P\|\alpha) \geq I_f(P)$$

where we use the lower semicontinuity of D_f (Lemma 1). This demonstrates that I_f is lower semicontinuous.

C.2 Proof of Theorem 2

We begin by recasting (13) as a constrained optimization problem. With a slight abuse of notation, we write $P = (A, (P_\theta)_{\theta\in\Theta})$ to denote the *improper choice rule* that specifies, for every $\theta \in \Theta$, a non-negative measure over actions $P_\theta \in \mathbb{R}_+^A$.

For every $x \in \mathbb{R}_+^\Theta$, let $\mathcal{P}_x$ the set of improper choice rules $P \in \mathbb{R}_+^{A\times\Theta}$ such that $\sum_{a\in A} P_\theta(a) = x(\theta)$ for all $\theta \in \Theta$.

The f-divergence $D_f(P\|\alpha)$ between an improper choice rule P and a probability distribution $\alpha \in \Delta(A)$ is defined in the obvious way, extending Definition 2 to non-negative measures. Similar to the case of proper choice rules (cf. Lemma 1), the function $(P, \alpha) \mapsto D_f(P\|\alpha)$

is lower semicontinuous and convex on $\mathbb{R}_+^{\Theta\times A}\times\Delta(A)$. In addition, $D_f(P\|\alpha)\geq f(x)$ for all $x\in\mathbb{R}_+^{\Theta}$ and $P\in\mathcal{P}_x$.

Let $I_f(P)$ be the f-information of an improper choice rule P:

$$I_f(P)=\inf_{\alpha\in\Delta(A)}D_f(P\|\alpha).$$

Similar to the case of proper choice rules, the function $P\mapsto I_f(P)$ is lower semicontinuous and convex on $\mathbb{R}_+^{\Theta\times A}$. In addition, for every $P\in\mathbb{R}_+^{\Theta\times A}$, there exists $\alpha\in\Delta(A)$ such that $I_f(P)=D_f(P\|\alpha)$.

For every $x\in\mathbb{R}_+^{\Theta}$, we consider the constrained optimization problem

$$\max_{P\in\mathcal{P}_x}\sum_{\theta\in\Theta}\pi(\theta)\sum_{a\in A}P_\theta(a)a(\theta)-I_f(P). \tag{56}$$

We go back to (13) when $x=\mathbf{1}$. We denote by $V(x)$ the value of (56). We say that $\lambda\in\mathbb{R}^{\Theta}$ is a *Lagrange multiplier* for (13) if

$$V(\mathbf{1})=\sup_{P\in\mathbb{R}_+^{A\times\Theta}}\sum_{\theta\in\Theta}\pi(\theta)\sum_{a\in A}P_\theta(a)a(\theta)-I_f(P)-\sum_{\theta\in\Theta}\lambda(\theta)\left(\sum_{a\in A}P_\theta(a)-1\right).$$

Lemma 9. *The value function $V:\mathbb{R}_+^{\Theta}\to\underline{\mathbb{R}}$ satisfies the following properties:*

(i). $\operatorname{dom}V=\operatorname{dom}f$.

(ii). *For every $x\in\mathbb{R}_+^{\Theta}$, there exists $P\in\mathcal{P}_x$ such that*

$$V(x)=\sum_{\theta\in\Theta}\pi(\theta)\sum_{a\in A}P_\theta(a)a(\theta)-I_f(P).$$

(iii). *V is concave.*

Proof. (i). Fix $x\in\mathbb{R}_+^{\Theta}$. If $V(x)>-\infty$, then there exist $P\in\mathcal{P}_x$ and $\alpha\in\Delta(A)$ such that $D_f(P\|\alpha)<+\infty$. Since $f(x)\leq D_f(P\|\alpha)$, we obtain $x\in\operatorname{dom}f$.

Conversely, suppose that $f(x)<+\infty$. Given a distribution $\alpha\in\Delta(A)$, we define $P\in\mathbb{R}_+^{A\times\Theta}$ by $P_\theta(a)=\alpha(a)x(\theta)$. Note that $\sum_{a\in A}P_\theta(a)=x(\theta)$ for all $\theta\in\Theta$. Moreover, $D_f(P\|\alpha)=f(x)$. Thus, $P\in\mathcal{P}_x$ and $I_f(P)<+\infty$. We deduce that $x\in\operatorname{dom}V$.

(ii). If $V(x)=-\infty$, then $f(x)=+\infty$ by (i). It follows that $I_f(P)\geq f(x)=+\infty$ for all $P\in\mathcal{P}_x$. We obtain that

$$V(x)=-\infty=\sum_{\theta\in\Theta}\pi(\theta)\sum_{a\in A}P_\theta(a)a(\theta)-I_f(P).$$

for all for all $P\in\mathcal{P}_x$.

Suppose instead that $V(x)\in\mathbb{R}$. Since,the function

$$P\mapsto\sum_{\theta\in\Theta}\pi(\theta)\sum_{a\in A}P_\theta(a)a(\theta)-I_f(P)$$

is upper semicontinuous, the desired result follows from the compactness of $\mathcal{P}_x$.

(iii). Take $x, y \in \mathbb{R}_+^\Theta$ and $t \in (0,1)$. By (ii), there are $P \in \mathcal{P}_x$ and $Q \in \mathcal{P}_y$ such that

$$\begin{aligned} V(x) &= \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} P_\theta(a) a(\theta) - I_f(P), \\ V(y) &= \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} Q_\theta(a) a(\theta) - I_f(Q). \end{aligned}$$

Using the fact that I_f is convex, we obtain

$$\begin{aligned} &(1-t)V(x) + tV(y) \\ =&\sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} [(1-t)P_\theta(a) + tQ_\theta(a)]a(\theta) - (1-t)I_f(P) - tI_f(Q) \\ \leq&\sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} [(1-t)P_\theta(a) + tQ_\theta(a)]a(\theta) - I_f((1-t)P + tQ) \\ \leq&V((1-t)x + ty). \end{aligned}$$

This demonstrates that V is concave. □

Since $\operatorname{dom} f = \operatorname{dom} V$ and $\mathbf{1} \in \operatorname{ri}(\operatorname{dom} f)$ by Assumption 1, we have $\mathbf{1} \in \operatorname{ri}(\operatorname{dom} f)$. Given that V is concave, the superdifferential of V at $x = \mathbf{1}$, defined as $\partial V(\mathbf{1})$, is nonempty (Rockafellar, 1970, Theorem 23.4): there exists $\lambda \in \mathbb{R}^\Theta$ such that for all $x \in \mathbb{R}_+^\Theta$,

$$V(\mathbf{1}) - \sum_{\theta\in\Theta} \lambda(\theta) \geq V(x) - \sum_{\theta\in\Theta} \lambda(\theta) x(\theta)$$

Note that $\lambda \in \partial V(\mathbf{1})$ if and only λ is a Lagrange multiplier for (13).

We define the Lagrangian function $\mathcal{L} : \mathbb{R}_+^{A\times\Theta} \times \mathbb{R}^\Theta \to \underline{\mathbb{R}}$ by

$$\begin{aligned} \mathcal{L}(P,\lambda) &= \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A} P_\theta(a) a(\theta) - I_f(P) - \sum_{\theta\in\Theta} \lambda(\theta) \sum_{a\in A} (P_\theta(a) - 1) \\ &= \sum_{a\in A} \sum_{\theta\in\Theta} (a(\theta)\pi(\theta) - \lambda(\theta)) P_\theta(a) - I_f(P) + \sum_{\theta\in\Theta} \lambda(\theta). \end{aligned}$$

The Lagrangian function is concave in P and affine in α. It defines the maxmin problem

$$\max_{P\in\mathbb{R}_+^{\Theta\times A}} \min_{\lambda\in\mathbb{R}^\Theta} \mathcal{L}(P,\lambda). \tag{57}$$

By standard arguments (see, e.g., Rockafellar, 1970, Theorem 28.3), a pair (P,λ) is a saddle point of (57) if and only P is a solution of (13) and λ is a Lagrange multiplier for (13). We have shown that (13) admits a solution and a Lagrange multiplier. Thus, the maxmin problem (57) admits a saddle point. Moreover, the saddle value of (57) is $V(\mathbf{1})$.

Next is a key step in the proof: it allows us to connect the Lagrangian function $\mathcal{L}$ to the conjugate function $f^\star$.

Lemma 10. *For all $\lambda \in \mathbb{R}^{\Theta}$ and $\alpha \in \Delta(A)$,*

$$\max_{P \in \mathbb{R}_+^{A\times\Theta}} \sum_{a\in A}\sum_{\theta\in\Theta}(a(\theta)\pi(\theta) - \lambda(\theta))P_\theta(a) - D_f(P\|\alpha) = \sum_{a\in A}\alpha(a)f^\star(a\pi - \lambda).$$

The maximum is achieved by $P \in \mathbb{R}_+^{A\times\Theta}$ such that, for all $\theta \in \Theta$ and $a \in A$,

$$P_\theta(a) = \alpha(a)\nabla f^\star(a\pi - \lambda).$$

Proof. Since f is co-finite (Assumption 1), if $D_f(P\|\alpha) < +\infty$ and $\alpha(a) = 0$, then $P_\theta(a) = 0$ for all $\theta \in \Theta$. Thus, direct computations show that

$$\begin{aligned}
&\sup_{P \in \mathbb{R}_+^{A\times\Theta}} \sum_{a\in A}\sum_{\theta\in\Theta}(a(\theta)\pi(\theta) - \lambda(\theta))P_\theta(a) - \lambda(\theta)) - D_f(P\|\alpha)\\
&= \sum_{a\in\mathrm{supp}(\alpha)} \alpha(a) \sup_{x\in\mathbb{R}_+^\Theta} \sum_{\theta\in\Theta}(a(\theta)\pi(\theta) - \lambda(\theta))\frac{x(\theta)}{\alpha(a)} - f\left(\frac{x}{\alpha(a)}\right)\\
&= \sum_{a\in\mathrm{supp}(\alpha)} \alpha(a) \sup_{y\in\mathbb{R}_+^\Theta} \sum_{\theta\in\Theta}(a(\theta)\pi(\theta) - \lambda(\theta))y(\theta) - f(y)\\
&= \sum_{a\in\mathrm{supp}(\alpha)} \alpha(a)f^\star(a\pi - \lambda) = \sum_{a\in A}\alpha(a)f^\star(a\pi - \lambda).
\end{aligned}$$

The second part of the statement follows from the fact that

$$f^\star(a\pi - \lambda) = \sum_{\theta\in\Theta}(a(\theta)\pi(\theta) - \lambda(\theta))y(\theta) - f(y) \quad\iff\quad y \in \partial f^\star(a\pi - \lambda).$$

See Rockafellar (1970, Theorem 23.5). Since $f^\star$ is differentiable (being f co-finite and essentially strictly convex, see Assumption 1), $\partial f^\star(a\pi - \lambda) = \{\nabla f^\star(a\pi - \lambda)\}$. □

The next lemmas establish a relationship between the maxmin problems (14) and (57). To ease the exposition, we denote by L the function

$$(\alpha, \lambda) \mapsto \sum_{a\in A}\alpha(a)f^\star(a\pi - \lambda) + \sum_{\theta\in\Theta}\lambda(\theta).$$

Lemma 11. *The maxmin problems (14) and (57) have the same value, $V(\mathbf{1})$.*

Proof. By a minimax theorem (Rockafellar, 1970, Corollary 37.3.1), the maxmin problem (14) has a saddle value. We have argued above that the saddle value of (57) is $V(\mathbf{1})$. By Lemma 10,

$$\inf_{\lambda\in\mathbb{R}^\Theta} \sup_{P\in\mathbb{R}_+^{A\times\Theta}} \mathcal{L}(P, \lambda) = \inf_{\lambda\in\mathbb{R}^\Theta} \sup_{\alpha\in\Delta(A)} L(\alpha, \lambda).$$

Hence, (14) and (57) have the same value, $V(\mathbf{1})$. □

Lemma 12. *Let (P, λ) be saddle point of (57). Take any $\alpha \in \Delta(A)$ such that $D_f(P\|\alpha) = I_f(P)$. Then, (α, λ) is a saddle point of (14). Moreover, $P_\theta(a) = \alpha(a)\nabla_\theta f^\star(a\pi - \lambda)$ for all $\theta \in \Theta$ and $a \in A$.*

Proof. Since (P, λ) is saddle point of (57), $\mathcal{L}(P, \lambda) \geq \mathcal{L}(Q, \lambda)$ for all $Q \in \mathbb{R}_+^{A \times \Theta}$. In other terms,

$$\begin{aligned}&\sum_{a \in A} \sum_{\theta \in \Theta} (a(\theta)\pi(\theta) - \lambda(\theta)) P_\theta(a) - D_f(P \| \alpha) \\ &= \sup_{\beta \in \Delta(A)} \sup_{Q \in \mathbb{R}_+^{A \times \Theta}} \sum_{a \in A} \sum_{\theta \in \Theta} (a(\theta)\pi(\theta) - \lambda(\theta)) Q_\theta(a) - D_f(Q \| \beta).\end{aligned}$$

It follows from Lemma 10 that $L(\alpha, \lambda) \geq L(\beta, \lambda)$ for all $\beta \in \Delta(A)$. Moreover, $P_\theta(a) = \alpha(a) \nabla_\theta f^\star(a\pi - \lambda)$ for all $\theta \in \Theta$ and $a \in A$.

It remains to verify that $L(\alpha, \lambda) \leq L(\alpha, l)$ for all $l \in \mathbb{R}^\Theta$. Since (P, λ) is saddle point of (57), P is a solution of (13). Thus, in particular, $\sum_{a \in A} P_\theta(a) = 1$ for all $\theta \in \Theta$. We have just argued that $P_\theta(a) = \alpha(a) \nabla_\theta f^\star(a\pi - \lambda)$ for all $\theta \in \Theta$ and $a \in A$. Thus, for all $\theta \in \Theta$,

$$\sum_{a \in A} \alpha(a) \nabla_\theta f^\star(a\pi - \lambda) = 1.$$

This is the first-order condition for the problem of minimizing $L(\alpha, l)$ over $l \in \mathbb{R}^\Theta$. We conclude that $L(\alpha, \lambda) \leq L(\alpha, l)$ for all $l \in \mathbb{R}^\Theta$. □

Lemma 13. *Let (α, λ) be a saddle point of (14). Define $P \in \mathbb{R}_+^{A \times \Theta}$ by $P_\theta(a) = \alpha(a) \nabla_\theta f^*(a\pi - \lambda)$. Then, (P, λ) is a saddle point of (57). Moreover, $I_f(P) = D_f(P \| \alpha)$.*

Proof. By Lemma 10,

$$\sum_{a \in A} \sum_{\theta \in \Theta} (a(\theta)\pi(\theta) - \lambda(\theta)) P_\theta(a) - D_f(P \| \alpha) + \sum_{\theta \in \Theta} \pi(\theta)\lambda(\theta) = L(\alpha, \lambda).$$

Since (α, λ) is a saddle point of L, $L(\alpha, \lambda) \geq L(\beta, \lambda)$ for all $\beta \in \Delta(A)$. It follows from Lemma 10 that $L(\alpha, \lambda)$ is equal to

$$\sup_{\beta \in \Delta(A)} \sup_{Q \in \mathbb{R}_+^{A \times \Theta}} \sum_{a \in A} \sum_{\theta \in \Theta} (a(\theta)\pi(\theta) - \lambda(\theta)) Q_\theta(a) - D_f(Q \| \beta) + \sum_{\theta \in \Theta} \lambda(\theta).$$

Overall, we deduce that $\mathcal{L}(P, \lambda) \geq \mathcal{L}(Q, \lambda)$ for all $Q \in \mathbb{R}_+^{A \times \Theta}$. Moreover, $I_f(P) = D_f(P \| \alpha)$.

It remains to show that $\mathcal{L}(P, \lambda) \leq \mathcal{L}(P, l)$ for all $l \in \mathbb{R}^\Theta$. The first-order condition for the problem of minimizing $L(\alpha, l)$ over $l \in \mathbb{R}^\Theta$ is

$$\sum_{a \in A} \alpha(a) \nabla_\theta f^\star(a\pi - \lambda) = 1.$$

Thus, $P \in \mathcal{P}_\mathbf{1}$. As a result, $\mathcal{L}(P, \lambda) = \mathcal{L}(P, l)$ for all $l \in \mathbb{R}^\Theta$. □

Theorem 2 follows from Lemmas 11–13.

C.3 Proofs of the results in Section 4.8

Proof of Lemma 4. If f is invariant, then for all $x \in \mathbb{R}^{\Theta}$

$$f^\star(x_\gamma) = \sup_{y \in \mathbb{R}^{\Theta}_+} \sum_{\theta \in \Theta} x_\gamma(\theta) y(\theta) - f(x_\gamma) = \sup_{y \in \mathbb{R}^{\Theta}_+} \sum_{\theta \in \Theta} x(\theta) y_{\gamma^{-1}}(\theta) - f(x) = f^\star(x).$$

Thus, f invariant implies $f^\star$ invariant. An analogous argument shows that $f^\star$ invariant implies f invariant. To prove the last part of the claim, set $x^\star = \nabla f^\star(x)$. By the subdifferential inequality, for all $y \in \mathbb{R}^{\Theta}$,

$$f^\star(y) - f^\star(x) \geq \sum_{\theta \in \Theta} x^\star(\theta)(y(\theta) - x(\theta)).$$

Since $f^\star$ is invariant, $f^\star(y_\gamma) = f^\star(y)$ and $f^\star(x_\gamma) = f^\star(x)$. Moreover, simple algebra shows that

$$\sum_{\theta \in \Theta} x^\star(\theta)(y(\theta) - x(\theta)) = \sum_{\theta \in \Theta} x^\star_\gamma(\theta)(y_\gamma(\theta) - x_\gamma(\theta)).$$

We obtain that for all $y \in \mathbb{R}^{\Theta}$,

$$f^\star(y_\gamma) - f^\star(x_\gamma) \geq \sum_{\theta \in \Theta} x^\star_\gamma(\theta)(y_\gamma(\theta) - x_\gamma(\theta)).$$

Since $\mathbb{R}^{\Theta} = \{y_\gamma : y \in \mathbb{R}^{\Theta}\}$, we deduce that $x^\star_\gamma = \nabla f^\star(x_\gamma)$, as desired. □

Proof of Proposition 1. Let (α, λ) be a saddle point of the maxmin problem (14). For every $\gamma \in \Gamma$, we define α_γ as follows: $\alpha_\gamma(a) = \alpha(a_{\gamma^{-1}})$ for all $a \in A$. We claim that $(\alpha_\gamma, \lambda_\gamma)$ is also a saddle point of (14).

First we show that α_γ is a best response to λ_γ, that is,

$$\sum_{a \in A} \alpha_\gamma(a) f^\star(a\pi - \lambda_\gamma) = \max_{a \in A} f^\star(a\pi - \lambda_\gamma).$$

We begin by observing that

$$\begin{aligned}
\sum_{a \in A} \alpha_\gamma(a) f^\star(a\pi - \lambda_\gamma) &= \sum_{a \in A} \alpha_\gamma(a) f^\star(a\pi_\gamma - \lambda_\gamma) \\
&= \sum_{a \in A} \alpha_\gamma(a) f^\star((a_{\gamma^{-1}}\pi - \lambda)_\gamma) \\
&= \sum_{a \in A} \alpha(a_{\gamma^{-1}}) f^\star(a_{\gamma^{-1}}\pi - \lambda) = \sum_{a \in A} \alpha(a) f^\star(a\pi - \lambda)
\end{aligned}$$

where the first equality uses the invariance of π, the third equality the invariance of f (which implies the invariance of $f^\star$), and the last equality the invariance of A. An analogous argument demonstrates that

$$\max_{a \in A} f^\star(a\pi - \lambda_\gamma) = \max_{a \in A} f^\star(a\pi - \lambda).$$

Since α is a best response to λ,

$$\sum_{a\in A}\alpha(a)f^\star(a\pi-\lambda)=\max_{a\in A}f^\star(a\pi-\lambda).$$

It follows that α_γ is a best response to λ_γ.

Next we show that λ_γ is a best response to α_γ, that is,

$$\lambda_\gamma\in\arg\min_{l\in\mathbb{R}^\Theta}\sum_{a\in A}\alpha_\gamma(a)f^\star(a\pi-l)+\sum_{\theta\in\Theta}l(\theta).$$

Reasoning as above,

$$\sum_{a\in A}\alpha_\gamma(a)f^\star(a\pi-\lambda_\gamma)+\sum_{\theta\in\Theta}\lambda_\gamma(\theta)=\sum_{a\in A}\alpha(a)f^\star(a\pi-\lambda)+\sum_{\theta\in\Theta}\lambda(\theta).$$

In addition, for all $l\in\mathbb{R}^\Theta$,

$$\sum_{a\in A}\alpha_\gamma(a)f^\star(a\pi-l)+\sum_{\theta\in\Theta}l(\theta)=\sum_{a\in A}\alpha(a)f^\star(a\pi-l_{\gamma^{-1}})+\sum_{\theta\in\Theta}l_{\gamma^{-1}}(\theta).$$

Hence, since λ is a best response to α,

$$\lambda\in\arg\min_{l\in\mathbb{R}^\Theta}\sum_{a\in A}\alpha_\gamma(a)f^\star(a\pi-l_{\gamma^{-1}})+\sum_{\theta\in\Theta}l_{\gamma^{-1}}(\theta).$$

Since $\mathbb{R}^\Theta=\{l_{\gamma^{-1}}:l\in\mathbb{R}^\Theta\}$, it follows that λ_γ is a best response to α_γ.

Since the choice $\gamma\in\Gamma$ was arbitrary, any pair $(\alpha_\gamma,\lambda_\gamma)$ is a saddle point of (14). We define $\bar\alpha\in\Delta(A)$ and $\bar\lambda\in\mathbb{R}^\Theta$ as follows:

$$\bar\alpha=\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\alpha_\gamma\quad\text{and}\quad\bar\lambda=\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\lambda_\gamma$$

where $|\Gamma|$ is the cardinality of Γ. Since the saddle points of (14) form a convex product set in $\Delta(A)\times\mathbb{R}^\Theta$ (see, e.g., Rockafellar 1970, Corollary 37.5.3), we deduce that $(\bar\alpha,\bar\lambda)$ is a saddle-point as well. Since Γ is a group, $\bar\alpha(a)=\bar\alpha(a_\gamma)$ for all $a\in A$ and $\gamma\in\gamma$, and $\bar\lambda_\gamma=\bar\lambda$ for all $\gamma\in\gamma$. We conclude that $(\bar\alpha,\bar\lambda)$ is an invariant saddle point of (14).

The resulting optimal choice rule is given by

$$P_\theta(a)=\bar\alpha(a)\nabla_\theta f^\star(a\pi-\bar\lambda).$$

For every $\gamma\in\Gamma$, we have that

$$\begin{aligned}P_{\gamma(\theta)}(a)=\bar\alpha(a)\nabla_{\gamma(\theta)}f^\star(a\pi-\bar\lambda)&=\bar\alpha(a)\nabla_\theta f^\star(a_\gamma\pi_\gamma-\bar\lambda_\gamma)\\&=\bar\alpha(a_\gamma)\nabla_\theta f^\star(a_\gamma\pi-\bar\lambda)=P_\theta(a_\gamma)\end{aligned}$$

where the first line uses the relation $\nabla_{\gamma(\theta)}f^\star(x)=\nabla_\theta f^\star(x_\gamma)$ (Lemma 4), and the second line the invariance of $\bar\alpha$, π, and $\bar\lambda$. □

C.4 Proofs of the results in Section 4.9

Proof of Lemma 5. (i). It suffices to show that $\nabla_\theta f^\star(x) > 0$ for all $\theta \in \Theta$ and $x \in X$. Define $y = \nabla f^\star(x)$. Then, $x \in \partial f(y)$ (Rockafellar, 1970, Theorem 23.5). Since f is essentially smooth, $\partial f(z) = \varnothing$ for all $z \notin \operatorname{int}(\operatorname{dom} f)$ (Rockafellar, 1970, Theorem 26.1). Thus, $y \in \operatorname{int}(\operatorname{dom} f)$. Since $\operatorname{dom} f \subseteq \mathbb{R}_+^\Theta$, we conclude that $y(\theta) > 0$ for all $\theta \in \Theta$.

(ii). Let (α_1, λ_1) and (α_2, λ_2) be two saddle points $\mathcal{D}$. By the product structure of the set of saddle points, (α_1, λ_2) is a saddle point as well. This means that λ_1 and λ_2 are two solutions of the following optimization problem:

$$\min_{\lambda \in \mathbb{R}^\Theta} \quad \sum_{a \in A} \alpha_1(a) f^\star(a\pi - \lambda) - \sum_{\theta \in \Theta} \lambda(\theta).$$

Since $f^\star$ is strictly convex, the objective function of this optimization problem is strictly convex. Thus, the solution must be unique: $\lambda_1 = \lambda_2$. □

Proof of Lemma 6. Let $\Theta = \{\theta_1, \ldots, \theta_n\}$ be an enumeration of the state space. For each $x \in \mathbb{R}^{n-1}$, we define

$$H_{n-1}^\star(x_1, \ldots, x_{n-1}) = H^\star(x_1, \ldots, x_{n-1}, 0).$$

The function $H_{n-1}^\star : \mathbb{R}^{n-1} \to \mathbb{R}$ inherits the properties of $H^\star$. It is monotone increasing, convex, and differentiable. In addition, $H_{n-1}^\star$ is strictly convex if and only if $H^\star$ is strictly convex modulo translations. Direct computations show that the conjugate of $H_{n-1}^\star$ is the function H_{n-1}. The desired result follows. □

Proof of Lemma 7. (i) It suffices to show that $\nabla_\theta H^\star(x) > 0$ for all $\theta \in \Theta$ and $x \in X$. Define $p = \nabla H^\star(x)$. Then, $x \in \partial H(p)$ (Rockafellar, 1970, Theorem 23.5). Note that

$$\nabla H_{n-1}(p_1, \ldots, p_{n-1}) = (x_1 - x_n, \ldots, x_{n-1} - x_n).$$

Since H_{n-1} is essentially smooth, $(p_1, \ldots, p_{n-1}) \in \operatorname{int}(\operatorname{dom} H_{n-1})$ (Rockafellar, 1970, Theorem 26.1). Thus, $p_1, \ldots, p_{n-1} > 0$ and $p_1 + \ldots + p_{n-1} < 1$. We conclude that $p_i > 0$ for all $i = 1, \ldots, n$, as desired.

(ii) Let (α_1, λ_1) and (α_2, λ_2) be two saddle points. By the product structure of the set of saddle points, (α_1, λ_2) is a saddle point as well. This means that λ_1 and λ_2 are two solutions of the following optimization problem:

$$\min_{\lambda \in \mathbb{R}^\Theta} \quad \sum_{a \in A} \alpha_1(a) H^\star(a - \lambda/\pi) - \sum_{\theta \in \Theta} \lambda(\theta).$$

Since $H^\star$ is strictly convex modulo translations, the objective function of this optimization problem is also strictly convex modulo translations. Thus, the solution must be unique up to translations: $\lambda_1 - \lambda_2 \in \mathbb{R}$. □

C.5 Proofs of the results in Section 5

C.5.1 Details for Example 6

Let $\alpha \in \Delta(A)$ be given. Enumerate $\text{supp}(\alpha) = \{a_1, \ldots, a_n\}$ so that $a_1(\theta) \geq \cdots \geq a_n(\theta)$.

First, note that the map $t \in \mathbb{R} \mapsto \ell(t) = \sum_{j=1}^{n} \alpha(a_j) \max\{a_j(\theta) - t + \kappa, 0\} \in \mathbb{R}_+$ is unbounded above and strictly decreasing on $(-\infty, \bar{t})$, where $\bar{t} = \sup\{t \in \mathbb{R} : \ell(t) > 0\}$. It follows that there exists a unique $\lambda_\pi(\theta) \in \mathbb{R}$ such that $\ell(\lambda_\pi(\theta)) = \kappa$, i.e., such that (20) holds. Moreover, for this value of $\lambda_\pi(\theta)$, there exists at least one $i \in [n]$ such that $a_i(\theta) > \lambda_\pi(\theta) - \kappa$ (for otherwise we would obtain the contradiction that $\ell(\lambda_\pi(\theta)) = 0$). Therefore, the index

$$i^*(\theta) = \max\{i \in [n] : a_i(\theta) > \lambda_\pi(\theta) - \kappa\} \tag{58}$$

is well-defined. Since $a_1(\theta) \geq \cdots \geq a_n(\theta)$ by convention, $\max\{a_i(\theta) - \lambda_\pi(\theta) + \kappa, 0\} \neq 0$ if and only if $i \in \{1, \ldots, i^*(\theta)\}$. Thus, (20) can be equivalently written as the linear equation

$$\sum_{j=1}^{i^*(\theta)} \alpha(a_j)\left(a_j(\theta) - \lambda_\pi(\theta) + \kappa\right) = \kappa,$$

which delivers the expression for $\lambda_\pi(\theta)$ stated in Example 6:

$$\lambda_\pi(\theta) = \sum_{j=1}^{i^*(\theta)} \left(\frac{\alpha(a_j)}{\sum_{k=1}^{i^*(\theta)} \alpha(a_k)}\right) a_j(\theta) - \frac{\kappa}{\sum_{j=1}^{i^*(\theta)} \alpha(a_j)} + \kappa.$$

Now, plugging this value of $\lambda_\pi(\theta)$ into (58) implies that

$$i \leq i^*(\theta) \iff a_i(\theta) > \lambda_\pi(\theta) - \kappa \iff \kappa > \sum_{j=1}^{i^*(\theta)} \alpha(a_j)\left(a_j(\theta) - a_i(\theta)\right).$$

Consequently, we have

$$i > i^*(\theta) \implies \kappa \leq \sum_{j=1}^{i^*(\theta)} \alpha(a_j)\left(a_j(\theta) - a_i(\theta)\right) \leq \sum_{j=1}^{i} \alpha(a_j)\left(a_j(\theta) - a_i(\theta)\right),$$

where the final inequality holds because $a_1(\theta) \geq \cdots \geq a_n(\theta)$. We conclude that

$$i^*(\theta) = \max\left\{i \in [n] : \kappa > \sum_{j=1}^{i} \alpha(a_j)\left(a_j(\theta) - a_i(\theta)\right)\right\}.$$

This completes our analysis of Example 6.

C.5.2 Proof of Proposition 2

We use the optimality condition (18) to prove both parts of the proposition. Note that, since ψ is increasing and strictly convex, both ψ and ψ' are strictly increasing.

First, for part (i), suppose that states $\theta, \tau \in \Theta$ are comparable, where action $a \in A$ satisfies $a(\theta) = a(\tau) = k \in \mathbb{R}$ and $P_\pi(a) > 0$. Then (18) implies that $\alpha(a) > 0$ and

$$P_\theta(a) - P_\tau(a) = \alpha(a) \cdot \left(\psi'\left(k - \lambda_\pi(\theta)\right) - \psi'\left(k - \lambda_\pi(\tau)\right)\right).$$

If $\lambda_\pi(\theta) \geq \lambda_\pi(\tau)$, then since ψ' is increasing, it follows that $P_\theta(a) \leq P_\tau(a)$. Conversely, if $P_\theta(a) \leq P_\tau(a)$, then since ψ' is strictly increasing, it follows that $\lambda(\theta) \geq \lambda(\tau)$, as desired.

Next, for part (ii), suppose that actions $a, b \in A$ are comparable, where state $\theta \in \Theta$ satisfies $a(\theta) = b(\theta) = k \in \mathbb{R}$. Then (18) implies that

$$P_\theta(a) - P_\theta(b) = \psi'\left(k - \lambda(\theta)\right) \cdot \left(\alpha(a) - \alpha(b)\right),$$

where $\psi'\left(k - \lambda(\theta)\right) > 0$ because ψ is strictly increasing. It follows that $P_\theta(a) \geq P_\theta(b)$ if and only if $\alpha(a) \geq \alpha(b)$, as desired.

C.5.3 Proof of Proposition 3

(i). IIA with respect to states follows directly from the optimality condition (18). To verify IIA with respect to labels, let $\theta, \tau \in \Theta$ satisfy $a(\theta) = a(\tau)$ for all $a \in A$. Take any saddle point (α, λ). Since ψ' is strictly increasing (as ϕ is essentially smooth), the prior-adjusted Lagrange multipliers $\lambda_\pi(\theta)$ and $\lambda_\pi(\tau)$ are the unique solutions to condition (19) in states θ and τ, respectively. Therefore, since $a(\theta) = a(\tau)$ for all $a \in A$, (19) implies that $\lambda_\pi(\theta) = \lambda_\pi(\tau)$. By Corollary 1, any optimal choice rule is generated by a saddle point of the form $(\hat{\alpha}, \hat{\lambda})$ with $\hat{\lambda} = \lambda$. IIA with respect to labels then follows directly from the optimality condition (18).

(ii). The result follows directly from the optimality condition (18).

(iii). By inspection, it is easy to see that mutual information satisfies IIA with respect to actions. For the converse, suppose that $|\Theta| \geq 5$ and take any Csiszàr information cost for which ψ is thrice continuously differentiable and strictly convex (recall that ψ is strictly convex if and only if ϕ is essentially smoooth). If this cost satisfies IIA with respect to actions, then Proposition 5 (proved separately in Appendix C.6.2 below) implies that the Arrow-Pratt coefficient is constant: $R_\psi = 1/\kappa$ for some $\kappa > 0$. The desired result then follows from the next lemma:

Lemma 14. *For all $t \in \mathbb{R}$,*

$$\psi(t) = \begin{cases} \int_0^t \mathrm{e}^{\int_0^s R_\psi(u)\,\mathrm{d}u}\,\mathrm{d}s & \text{if } t \geq 0, \\ -\int_t^0 \mathrm{e}^{-\int_s^0 R_\psi(u)\,\mathrm{d}u}\,\mathrm{d}s & \text{if } t < 0. \end{cases}$$

Proof. Since R_ψ is the derivative of $\log \psi'$, and $\psi'(0) = 1$, it follows from the fundamental theorem of calculus that

$$\psi'(s) = \begin{cases} \mathrm{e}^{\int_0^s R_\psi(u)\,\mathrm{d}u} & \text{if } t \geq 0, \\ \mathrm{e}^{-\int_s^0 R_\psi(u)\,\mathrm{d}u} & \text{if } s < 0. \end{cases}$$

Using the normalization $\psi(0) = 0$, we obtain the desired result from another application of the fundamental theorem of calculus. □

C.6 Proofs of the results in Section 6

C.6.1 Proof of Proposition 4

By the optimality conditions (18) and (19),

$$\log \frac{P^{\epsilon}_{\theta^{\epsilon}}(a^{\epsilon}_i)}{P^{\epsilon}_{\theta^{\epsilon}}(a^{\epsilon}_j)} = \log \frac{\psi'(d_i + \epsilon u - \lambda_\pi(d^{\epsilon}))}{\psi'(d_i + \epsilon v - \lambda_\pi(d^{\epsilon}))},$$

where $\lambda_\pi(d^{\epsilon})$ is determined by the equation

$$\frac{1}{n}\psi'(d_i + \epsilon u - \lambda_\pi(d^{\epsilon})) + \frac{1}{n}\psi'(d_i + \epsilon v - \lambda_\pi(d^{\epsilon})) + \frac{1}{n}\sum_{k \neq i,j} \psi'(d_k - \lambda_\pi(d_\epsilon)) = 1.$$

By the implicit function theorem, $\lambda_\pi(d^{\epsilon})$ is a differentiable function of $\epsilon \in (0,1)$. Moreover, $\lambda_\pi(d^{\epsilon}) \to \lambda_\pi(d)$ as $\epsilon \to 0$. Then, the desired result follows from a first-order Taylor expansion of the map

$$\epsilon \mapsto \log \frac{\psi'(d_i + \epsilon u - \lambda_\pi(d^{\epsilon}))}{\psi'(d_i + \epsilon v - \lambda_\pi(d^{\epsilon}))}$$

at $\epsilon = 0$, using the fact that R_ψ is the derivative of $\log \psi'$.

C.6.2 Proof of Proposition 5

We prove (i). The proof of (ii) is specular and left to the reader.

"If." Let (Θ, π, A) be a decision problem. Let P an optimal choice rule, with corresponding saddle point (α, λ). Suppose that choice is bolder in state θ than in state τ. Take actions $a, b \in A$ in the support of P_π such that

$$a(\theta) = a(\tau) \geq b(\theta) = b(\tau).$$

Note that states θ and τ are comparable (because $a(\theta) = a(\tau)$). Thus, $\lambda_\pi(\theta) \geq \lambda_\pi(\tau)$ (Proposition 2). Using the optimality condition (27), we obtain:

$$\begin{aligned}
\log \frac{P_\theta(a)\alpha(b)}{P_\theta(b)\alpha(a)} &= \int_{b(\theta)}^{a(\theta)} R_\psi(x - \lambda_\pi(\theta))\,\mathrm{d}x \\
&\geq \int_{b(\theta)}^{a(\theta)} R_\psi(x - \lambda_\pi(\tau))\,\mathrm{d}x \\
&= \int_{b(\tau)}^{a(\tau)} R_\psi(x - \lambda_\pi(\tau))\,\mathrm{d}x = \log \frac{P_\tau(a)\alpha(b)}{P_\tau(b)\alpha(a)},
\end{aligned}$$

where the inequality follows from R_ψ be decreasing. We deduce that $\frac{P_\theta(a)}{P_\theta(b)} \geq \frac{P_\tau(a)}{P_\tau(b)}$, as desired.

"Only if." By contraposition, suppose R_ψ is *not* decreasing. Then, there exist $x_1, x_2 \in \mathbb{R}$ such that $x_1 > x_2$ and $R_\psi(x_1) > R_\psi(x_2)$. Since ψ'' is differentiable, $R_\psi = \psi''/\psi'$ is differentiable as well. By the mean value theorem, there exists $x_3 \in (x_1, x_2)$ with $R'_\psi(x_3) > 0$. As ψ'' is continuously differentiable, R'_ψ is continuous, so $R'_\psi(x) > 0$ for all x sufficiently close to x_3.

Thus, there is a nonempty open interval X on which R_ψ is strictly increasing. Choose $\bar{x}, \underline{x} \in X$ such that $\bar{x} > \underline{x}$. By slightly perturbing these points if necessary, we can ensure

$$\frac{1}{2}\psi'(\bar{x}) + \frac{1}{2}\psi'(\underline{x}) \neq 1.$$

For concreteness, we focus on the case

$$\frac{1}{2}\psi'(\bar{x}) + \frac{1}{2}\psi'(\underline{x}) > 1, \tag{59}$$

the other case being analogous (see comment at the end of the proof).

We now construct a decision problem (Θ, π, A) and an optimal choice rule $P = (A, (P_\theta)_{\theta\in\Theta})$ with saddle point (α, λ) such that the agent fails to satisfy increasing selectivity.

Let the state space and the action set be:

$$\Theta = \{1, 2, 3, 4, 5\} \quad \text{and} \quad A = \{a, b, c\}.$$

We specify prior, payoffs, f-mean, and Lagrange multiplier state by state.

State $\theta = 1$. In the first state, action a pays $\bar{x}$, action b pays $\underline{x}$, and action c pays $y < 0$. The prior-adjusted Lagrange multiplier takes value 0. For each $y < 0$, we select $\xi(y) \in (0, 1)$ such that

$$\xi(y)\left(\frac{1}{2}\psi'(\bar{x}) + \frac{1}{2}\psi'(\underline{x})\right) + (1 - \xi(y))\psi'(y) = 1 = \psi'(0).$$

Equation (59) guarantees the existence of such $\xi(y)$, with $\xi(y) \to 0$ as $y \to 0$. Set $\alpha(a) = \alpha(b) = \xi(y)/2$ and $\alpha(c) = 1 - \xi(y)$. Then (19) holds.

State $\theta = 2$. Pick $\epsilon > 0$ sufficiently small so that $\bar{x} - \epsilon \in X$, $\underline{x} - \epsilon \in X$, and

$$\frac{1}{2}\psi'(\bar{x} - \epsilon) + \frac{1}{2}\psi'(\underline{x} - \epsilon) > 1. \tag{60}$$

By choosing y close to zero (so $\xi(y)$ is close to 0), we can ensure that there is z such that

$$\psi'(z) = \frac{1}{1 - \xi(y)}\left(1 - \xi(y)\left(\frac{1}{2}\psi'(\bar{x}) + \frac{1}{2}\psi'(\underline{x})\right)\right)$$

Then, in state $\theta = 2$, action a pays $\bar{x}$, action b pays $\underline{x}$, and action c pays $z + \epsilon$. The prior-adjusted Lagrange multiplier takes value ϵ. Then (19) holds.

State $\theta = 3$. Same as $\theta = 1$, but swap the payoffs of a and b.

State $\theta = 4$. Same as $\theta = 2$, but swap the payoffs of a and b.

State $\theta = 5$. Here, a and b pays -1, and action c pays 1. By the intermediate value theorem, there exists $w \in [-1, 1]$ such that

$$\xi(y)\psi'(-1 + w) + (1 - \xi(y))\psi'(1 + w) = \psi'(0). \tag{61}$$

Set w as the prior-adjusted Lagrange multiplier in state $\theta = 5$. Then (19) holds.

We now complete our construction by selecting the prior. From (59) and (60) we have:

$$\frac{1}{4}\left(\psi'(\bar{x}) + \psi'(\underline{x}) + \psi'(\bar{x} - \epsilon) + \psi'(\underline{x} - \epsilon)\right) > 1 > \frac{1}{2}\left(\psi'(y) + \psi'(z)\right).$$

Since $\psi'(-1+w) < \psi'(1+w)$, there exists $\zeta \in (0,1)$ such that

$$\begin{aligned}&\frac{\zeta}{4}(\psi'(\bar{x}) + \psi'(\underline{x}) + \psi'(\bar{x}-\epsilon) + \psi'(\underline{x}-\epsilon)) + (1-\zeta)\psi'(-1+w)\\ =&\frac{\zeta}{2}(\psi'(y) + \psi'(z)) + (1-\zeta)\psi'(1+w). \end{aligned} \tag{62}$$

We set the prior as follows:

$$\pi(1) = \pi(2) = \pi(3) = \pi(4) = \frac{\zeta}{4} \quad \text{and} \quad \pi(5) = 1-\zeta.$$

It follows from (62) that α is a best response to λ in the maxmin problem (14). This concludes our construction.

Note that states $\theta = 1$ and $\theta = 2$ are comparable (indeed, $a(1) = a(2)$). Moreover, the prior-adjusted Lagrange multiplier is larger in the second state: $\lambda_\pi(1) = 0 < \epsilon = \lambda_\pi(2)$. Thus, choice is bolder in the state $\theta = 2$ than in state $\theta = 1$ (Proposition 2). Moreover,

$$\log \frac{P_1(a)}{P_1(b)}\frac{\alpha(b)}{\alpha(a)} = \int_{\underline{x}}^{\bar{x}} R_\psi(x)\,\mathrm{d}x > \int_{\underline{x}}^{\bar{x}} R_\psi(x-\epsilon)\,\mathrm{d}x = \log \frac{P_2(a)}{P_2(b)}\frac{\alpha(b)}{\alpha(a)}$$

where we use the fact that R_ψ is strictly increasing on X. We deduce that $\frac{P_1(a)}{P_1(b)} > \frac{P_2(a)}{P_2(b)}$. Hence, the agent *does not* exhibit increasing selectivity, as claimed.

In the case where $\frac{1}{2}\psi'(\bar{x}) + \frac{1}{2}\psi'(\underline{x}) < 1$, we require $y > 0$. Furthermore, the payoffs in state $\theta = 5$ are reversed: actions a and b yield a payoff of 1, while action c yields -1. The remainder of the proof follows with these modifications almost verbatim.

C.6.3 Proof of Proposition 6

See Corollary 6 in Appendix B.

C.7 Proof of the results in Section 7

C.7.1 Proof of Propositions 7 and 8

We begin by characterizing optimal information acquisition for a fixed function ϕ using Theorem 2. By symmetry of the environment, we can assume the Lagrange multiplier is independent of the state without loss of generality (see Proposition 1 and Corollary 1). Consequently, we identify λ with an element of the real line. Since ψ is strictly convex, λ is unique (see Section 4.9).

Claim 9. *For all $\theta, \tau \in \Theta$, $\alpha(a_\theta) = \alpha(a_\tau)$ in any saddle point of (14).*

Proof. By the optimality condition for λ,

$$\begin{aligned}\alpha(a_\theta)\psi'(w-\lambda_\pi) + (1-\alpha(a_\theta)-\alpha(b))\psi'(0-\lambda_\pi) + \alpha(b)\psi'(c-\lambda_\pi) &= \psi'(0),\\ \alpha(a_\tau)\psi'(w-\lambda_\pi) + (1-\alpha(a_\tau)-\alpha(b))\psi'(0-\lambda_\pi) + \alpha(b)\psi'(c-\lambda_\pi) &= \psi'(0).\end{aligned}$$

Combining these two equations, we obtain:

$$\alpha(a_\theta)\psi'(w-\lambda_\pi)+(1-\alpha(a_\theta))\psi'(0-\lambda_\pi)=\alpha(a_\tau)\psi'(w-\lambda_\pi)+(1-\alpha(a_\tau))\psi'(0-\lambda_\pi)$$

Since $\psi'(w-\lambda_\pi)>\psi'(0-\lambda_\pi)$, we deduce that $\alpha(a_\theta)=\alpha(a_\tau)$. □

Thus, we can identify α with a single number, an element of the unit interval $[0,1]$, with the convention that α is the f-mean probability of the safe action. Inconclusive evidence corresponds to the case in $\alpha\in(0,1)$.

Claim 10. *For every w, there exists a unique $\bar{c}$ such that*

$$\frac{1}{n}\psi(w-\bar{c})+\frac{n-1}{n}\psi(-\bar{c})=\psi(0).$$

The threshold value $\bar{c}$ has the following properties:

(i). $\frac{w}{n}<\bar{c}<w$.

(ii). *If $c\geq\bar{c}$, then the max-min problem (14) has a saddle point $(\alpha,\lambda)=(1,c/n)$.*

(iii). *If $c>\bar{c}$, then $P_\pi(b)=1$ at the optimum.*

(iv). *If $c<\bar{c}$, then $P_\pi(b)<1$ at the optimum.*

Proof. To verify the existence of $\bar{c}$, note that $c\geq w$ implies

$$\frac{1}{n}\psi(w-c)+\frac{n-1}{n}\psi(-c)>\psi(0)$$

by strict monotonicity of ψ. If instead $c\leq\frac{w}{n}$, then

$$\frac{1}{n}\psi(w-c)+\frac{n-1}{n}\psi(-c)>\psi(w-nc)\geq\psi(0),$$

where we use the fact that ψ is strictly convex and monotone. Thus, by the intermediate value, there exists $\bar{c}\in(\frac{w}{n},w)$ such that

$$\frac{1}{n}\psi(w-\bar{c})+\frac{n-1}{n}\psi(-\bar{c})=\psi(0).$$

The uniqueness of $\bar{c}$ follows from $\frac{1}{n}\psi(w-c)+\frac{n-1}{n}\psi(-c)$ being strictly decreasing in c. This demonstrates the first part of the statement, as well as property (i).

To prove properties (ii)–(iv), note that $(1,\lambda)$ is a saddle point of (14) if and only if

$$\psi'(c-\lambda_\pi)=\psi'(0),$$
$$\frac{1}{n}\psi(w-\lambda_\pi)+\frac{n-1}{n}\psi(-\lambda_\pi)\leq\psi(c-\lambda_\pi).$$

Equivalently, $\lambda_\pi=c$ and

$$\frac{1}{n}\psi(w-c)+\frac{n-1}{n}\psi(-c)\leq\psi(0).$$

Thus, $(1, \lambda)$ is a saddle point of (14) if and only if $\lambda_\pi = c$ and $c \geq \bar{c}$. This shows (ii) and (iv).

To prove also (iii), suppose $c > \bar{c}$ and let (α, λ) be a saddle point of (14). As shown above, $\lambda_\pi = c$. Since $c > \bar{c}$,

$$\frac{1}{n}\psi(w - c) + \frac{n-1}{n}\psi(-c) < \psi(0).$$

This implies that $\alpha = 1$. We deduce that (iii) hold. □

Claim 11. *For every w, there exist unique $\underline{\lambda}$ and $\underline{c}$ such that*

$$\frac{1}{n}\psi'(w - \underline{\lambda}_\pi) + \frac{n-1}{n}\psi'(-\underline{\lambda}_\pi) = \psi'(0),$$
$$\frac{1}{n}\psi(w - \underline{\lambda}_\pi) + \frac{n-1}{n}\psi(-\underline{\lambda}_\pi) = \psi(\underline{c} - \underline{\lambda}_\pi).$$

The threshold values $\underline{\lambda}$ and $\underline{c}$ have the following properties:

(i). $0 < \underline{\lambda}_\pi < w$ *and* $\frac{w}{n} < \underline{c} < w$.

(ii). *If $c \leq \underline{c}$, then the max-min problem (14) has a saddle point $(\alpha, \lambda) = (0, \underline{\lambda})$.*

(iii). *If $c < \underline{c}$, then $P_\pi(b) = 0$ at the optimum.*

(iv). *If $c > \underline{c}$, then $P_\pi(b) > 0$ at the optimum.*

Proof. Existence and uniqueness of $\underline{\lambda}$, as well as the fact that $\underline{\lambda}_\pi \in (0, w)$, follow from ψ' being strictly increasing and continuous. The value of $\underline{c}$ is obtained by inverting the second equation:

$$\underline{c} = \psi^{-1}\left(\frac{1}{n}\psi(w - \underline{\lambda}_\pi) + \frac{n-1}{n}\psi(-\underline{\lambda}_\pi)\right) + \underline{\lambda}_\pi.$$

To prove the bounds for $\underline{c}$, it suffices to observe that

$$\psi^{-1}\left(\frac{1}{n}\psi(w - \underline{\lambda}_\pi) + \frac{n-1}{n}\psi(-\underline{\lambda}_\pi)\right) + \underline{\lambda}_\pi < \psi^{-1}\left(\psi(w - \underline{\lambda}_\pi)\right) + \underline{\lambda}_\pi = w,$$
$$\psi^{-1}\left(\frac{1}{n}\psi(w - \underline{\lambda}_\pi) + \frac{n-1}{n}\psi(-\underline{\lambda}_\pi)\right) + \underline{\lambda}_\pi > \psi^{-1}\left(\psi\left(\frac{w}{n} - \underline{\lambda}_\pi\right)\right) + \underline{\lambda}_\pi = \frac{w}{n},$$

where we use the fact that ψ is strictly increasing and convex. This demonstrates the first part of the statement, as well as property (i).

To prove properties (ii)–(iv), note that $(0, \lambda)$ is a saddle point of (14) if and only if

$$\frac{1}{n}\psi'(w - \lambda_\pi) + \frac{n-1}{n}\psi'(-\lambda_\pi) = \psi'(0),$$
$$\frac{1}{n}\psi(w - \lambda_\pi) + \frac{n-1}{n}\psi(-\lambda_\pi) \geq \psi(c - \lambda_\pi).$$

Equivalently, $\lambda = \underline{\lambda}$ and

$$\underline{c} = \psi^{-1}\left(\frac{1}{n}\psi(w - \underline{\lambda}_\pi) + \frac{n-1}{n}\psi(-\underline{\lambda}_\pi)\right) + \underline{\lambda}_\pi \geq c.$$

This shows (ii) and (iv). To prove also (iii), suppose $c < \underline{c}$ and let (α, λ) be a saddle point of (14). As shown above, $\lambda = \underline{\lambda}$. Since $c < \underline{c}$, we obtain:

$$\frac{1}{n}\psi(w - \underline{\lambda}_\pi) + \frac{n-1}{n}\psi(-\underline{\lambda}_\pi) > \psi(c - \underline{\lambda}_\pi).$$

This implies that $\alpha = 0$. We deduce that (iii) hold. □

Claim 12. *If R_ψ is strictly monotone on $(-w, w)$, then $\underline{c} < \bar{c}$.*

Proof. It is easy to see that $\underline{c} \leq \bar{c}$. This is because $c > \bar{c}$ implies $P_\pi(b) = 1$ at the optimum, while $c < \underline{c}$ implies $P_\pi(b) = 0$ at the optimum. Thus, to prove that $\underline{c} < \bar{c}$, we only need to rule out the case in which $\underline{c} = \bar{c}$.

By contradiction, suppose that $\underline{c} = \bar{c}$. Then, the maxmin problem (14) has saddle points $(1, \bar{c}/n)$ and $(0, \underline{\lambda})$. By uniqueness of the Lagrange multiplier, $\bar{c} = n\underline{\lambda}$. We obtain:

$$\frac{1}{n}\psi(w - \bar{c}) + \frac{n-1}{n}\psi(-\bar{c}) = \psi(0), \tag{63}$$

$$\frac{1}{n}\psi'(w - \bar{c}) + \frac{n-1}{n}\psi'(-\bar{c}) = \psi'(0). \tag{64}$$

We now use the fact that R_ψ is strictly monotone on $(-w, w)$ to reach a contradiction. The key observation is that

$$\frac{\mathrm{d}}{\mathrm{d}x}\psi'(\psi^{-1}(x)) = \frac{\psi''(\psi^{-1}(x))}{\psi'(\psi^{-1}(x))} = R_\psi(\psi^{-1}(x)).$$

Hence, since R_ψ is strictly monotone on $(-w, w)$ and ψ^{-1} is strictly increasing on its entire domain, the composite function $\psi' \circ \psi^{-1}$ is either strictly convex or strictly concave on the interval $(\psi(-w), \psi(w))$. In any case, (63) implies that

$$\frac{1}{n}\psi'(w - \bar{c}) + \frac{n-1}{n}\psi'(-\bar{c}) \neq \psi'(0),$$

which contradicts (64). We conclude that $\underline{c} < \bar{c}$, as desired. □

This proves Proposition 7. To prove Proposition 8, suppose that R_ψ is strictly monotone on a non-empty open interval. Thus, there must exists $\underline{x}, \bar{x} \in \mathbb{R}$, with $\underline{x} < \bar{x}$, such that R_ψ is strictly monotone on $(\underline{x} - \epsilon, \bar{x} + \epsilon)$ for any $\epsilon > 0$ sufficiently small. Define $\underline{k} = \psi'(\underline{x})$ and $\bar{k} = \psi'(\bar{x})$. Since R_ψ is strictly monotone on $(\underline{x} - \epsilon, \bar{x} + \epsilon)$, R_{ψ_k} is strictly monotone on $(-\epsilon, +\epsilon)$. Thus, for all $k \in (\underline{k}, \bar{k})$ and $w \in (0, \epsilon)$, the Arrow-Pratt coefficient of ψ_k is strictly monotone on $(-w, w)$. Proposition 8 follows.

C.7.2 Proof of Proposition 9

By symmetry of the environment, we can assume that the Lagrange multiplier is independent of the state—see Proposition 1 and Corollary 1. Thus, we identify λ with an element of the

real line. Consequently, given that $f_H^\star$ is translation invariant with respect to the prior (see page 21), we obtain: for all $\theta \in \Theta$,

$$f_H^\star(b\pi - \lambda) \geq f_H^\star(a_\theta \pi - \lambda) \quad \iff \quad H^\star(c, \ldots, c) \geq H^\star(w, 0, \ldots, 0).$$

Claim 13. *There exists a unique $\hat{c}$ such that*

$$H^\star(\hat{c}, \ldots, \hat{c}) = H^\star(w, 0, \ldots, 0)$$

Moreover, $c \geq \hat{c}$ if and only if $H^\star(c, \ldots, c) \geq H^\star(w, 0, \ldots, 0)$.

Proof. To prove the existence of $\hat{c}$, note that $c \geq w$ implies

$$H^\star(c, \ldots, c) \geq H^\star(w, 0, \ldots, 0)$$

by monotonicity of $H^\star$. If instead $c \leq \frac{w}{n}$, then

$$H^\star(w, 0, \ldots, 0) = \frac{1}{n} \sum_{\theta \in \Theta} H^*(a_\theta) \geq H\left(\frac{w}{n}, \ldots, \frac{w}{n}\right) \geq H^\star(c, \ldots, c),$$

where the first equality uses the symmetry of $H^\star$ (which follows from the symmetry of H), the second inequality the convexity of $H^\star$, and the third inequality the monotonicity of $H^\star$. Hence, by the intermediate value theorem, there exist $\hat{c}$ such that

$$H^\star(\hat{c}, \ldots, \hat{c}) = H^\star(w, 0, \ldots, 0).$$

The fact that $\hat{c}$ is uniquely pinned follows from the fact that $H^\star(c, \ldots, c)$ is strictly increasing in c. This, in turn, comes from the fact that $H^\star$ is translation invariant:

$$H^\star(c, \ldots, c) = H^\star(0, \ldots, 0) + c.$$

This also proves the last part of the proposition. □

Applying Theorem 2, we obtain (i) and (ii) of Proposition 9. If $c > \hat{c}$, $\alpha(a_\theta) = 0$ for all $\theta \in \Theta$, which implies $\alpha(b) = P_\pi(b) = 1$. If $c < \hat{c}$, then $\alpha(b) = P_\pi(b) = 0$.

Regarding (iii), take any $t \in [0, 1]$. Define $\alpha(b) = t$ and, for every $\theta \in \Theta$, $\alpha(a_\theta) = (1 - t)/n$.

Claim 14. *For $c = \hat{c}$, the pair $(\alpha, 0)$ is a saddle point of (14).*

Proof. Since $c = \hat{c}$, (15) holds. Now we check that (16) also holds. Notice that for all $\theta \in \Theta$,

$$\frac{1}{n} \sum_{\tau \in \Theta} \nabla_\theta f_H^\star(a_\tau \pi) = \sum_{\tau \in \Theta} \pi(\tau) \nabla_\tau f_H^\star(a_\theta \pi) = \sum_{\tau \in \Theta} \nabla_\tau H^\star(a_\theta) = 1$$

where the first equality follows from Lemma 4. Similarly, $\nabla_\theta f_H^\star(s\pi) = 1$ for all $\theta \in \Theta$. Hence,

$$t \nabla_\theta f_H^\star(b\pi) + \frac{1 - t}{n} \sum_{\tau \in \Theta} \nabla_\theta f_H^\star(a_\tau \pi) = t + 1 - t = 1.$$

This shows that also (16) is satisfied. We conclude $(\alpha, 0)$ is a saddle point of (14) for $c = \hat{c}$. □

The choice rule corresponding to $(\alpha, 0)$ has $P_\pi(b) = \alpha(b) = t$. This proves (iii).

C.8 Proofs of the results in Section 8

C.8.1 Proof of Proposition 10

Let (α, λ) be a saddle point of (14). Since ϕ is essentially smooth, the Lagrange multiplier is unique (Lemma 5). Moreover, by the symmetry of the environment, it is independent of the state (Proposition 1).

For every state θ, let A_θ be the set of actions that pays w if the realized state is θ. The optimality condition for the Lagrange multiplier in states θ is:

$$\alpha(A_\theta)\psi'(w - \lambda_\pi(\theta)) + (1 - \alpha(A_\theta))\psi'(-\lambda_\pi(\theta)) = \psi'(0).$$

Thus, taking any two states θ and τ,

$$\alpha(A_\theta)\left[\psi'(w - \lambda_\pi(\theta)) - \psi'(-\lambda_\pi(\theta))\right] = \alpha(A_\tau)\left[\psi'(w - \lambda_\pi(\theta)) - \psi'(-\lambda_\pi(\theta))\right],$$

where we use the fact that $\lambda_\pi(\theta) = \lambda_\pi(\tau)$. Since ψ' is strictly increasing and $r > 0$, we conclude that $\alpha(A_\theta) = \alpha(A_\tau)$. Furthermore, because each action pays w in exactly m states, we have:

$$\sum_{\theta \in \Theta} \alpha(A_\theta) = m.$$

Since $\alpha(A_\theta) = \alpha(A_\tau)$ for all $\theta, \tau \in \Theta$, we conclude that $\alpha(A_\theta) = m/n$ for all $\theta \in \Theta$. The desired result follows.

C.8.2 Properties of $l_\gamma(w)$

Lemma 15. *The Lagrange multiplier has the following properties:*

(i). $l_\gamma(w)$ *is strictly increasing in* w.

(ii). $l_\gamma(w)$ *is strictly increasing in* γ.

(iii). $l_\gamma(w)$ *is continuous in* w.

(iv). $l_\gamma(w) \to 0$ *as* $w \to 0$.

(v). $l_\gamma(w) \to +\infty$ *as* $w \to +\infty$.

(vi). $l_\gamma(w) \to 0$ *as* $\gamma \to 0$.

(vii). $l_\gamma(w) \to w$ *as* $\gamma \to 1$.

Proof. (i). Suppose $w^1 > w^2$. Using the fact that ψ' is strictly increasing, we obtain:

$$\begin{aligned}\gamma\psi'\left(w^1 - l_\gamma(w^1)\right) + (1-\gamma)\psi'\left(-l_\gamma(w^2)\right) &> \gamma\psi'\left(w^2 - l_\gamma(w^2)\right) + (1-\gamma)\psi'\left(-l_\gamma(w^2)\right) \\ &= \gamma\psi'\left(w^1 - l_\gamma(w^1)\right) + (1-\gamma)\psi'\left(-l_\gamma(w^1)\right).\end{aligned}$$

We conclude that $l_\gamma(w^1) > l_\gamma(w^2)$.

(ii). Suppose $\gamma^1 > \gamma^2$. Using the fact that ψ' is strictly increasing, we obtain:

$$\begin{aligned}\gamma^1\psi'\left(w - l_{\gamma^2}(w)\right) + (1-\gamma^1)\psi'\left(-l_{\gamma^2}(w)\right) &> \gamma^2\psi'\left(w - l_{\gamma^2}(w)\right) + (1-\gamma^2)\psi'\left(-l_{\gamma^2}(w)\right) \\ &= \gamma^1\psi'\left(r - l_{\gamma^1}(w)\right) + (1-\gamma^1)\psi'\left(-l_{\gamma^1}(w)\right).\end{aligned}$$

We conclude that $l_{\gamma^1}(w) > l_{\gamma^2}(w)$.

(iii). Let (w^m) be a sequence of rewards with limit w. Each $l_\gamma(w^m)$ satisfies $0 \leq l_\gamma(w^m) \leq w^m$. Thus, the sequence $(l_\gamma(w^m))$ is bounded. Without loss of generality, we can assume it has a limit, l. For every m,

$$\gamma\psi'\left(w^m - l_\gamma(w^m)\right) + (1-\gamma)\psi'\left(-l_\gamma(w^m)\right) = \psi'(0).$$

Taking the limit as $m \to \infty$, we obtain from the continuity of ψ' that:

$$\gamma\psi'(w-l) + (1-\gamma)\psi'(-l) = \psi'(0).$$

Since $l_\gamma(w)$ is the unique solution of this equation, we conclude that $l = l_\gamma(w)$.

(iv). By (i), $l_\gamma(w)$ is increasing in w. Define $l_\gamma(0) = \inf_{w>0} l_\gamma(w)$. Since $l_\gamma(w) > 0$ for all w, we have $l_\gamma(0) \geq 0$. Furthermore, for every $w > 0$,

$$\gamma\psi'(w - l_\gamma(w)) + (1-\gamma)\psi'(-l_\gamma(w)) = \psi'(0).$$

Taking the limit as $w \to 0$, we obtain from the continuity of ψ' that:

$$\psi'(-l_\gamma(0)) = \psi'(0).$$

We conclude that $l_\gamma(0) = 0$.

(v). By (i), $l_\gamma(w)$ is increasing in w. Define $\bar{l_\gamma} = \sup_{w>0} l_\gamma(w)$. By contradiction, suppose $\bar{l_\gamma} < +\infty$. Recall that $\psi'(w) \to +\infty$ as $w \to +\infty$. Thus, by choosing w sufficiently large, we can ensure that

$$\gamma\psi'\left(w - \bar{l}\right) + (1-\gamma)\psi'\left(-\bar{l}\right) > \psi'(0).$$

This implies that $l_\gamma(w) > \bar{l_\gamma}$, a contradiction with the definition of $\bar{l_\gamma}$. We conclude that $\bar{l_\gamma} = +\infty$, as desired.

(vi). By (ii), $l_\gamma(w)$ is increasing in γ. Define $l_0(w) = \inf_{\gamma\in(0,1)} l_\gamma(w)$. Since $l_\gamma(w) > 0$ for all γ, we have $l_0(w) \geq 0$. Furthermore, for every $\gamma \in (0,1)$,

$$\gamma\psi'(w - l_\gamma(w)) + (1-\gamma)\psi'(-l_\gamma(w)) = \psi'(0).$$

Taking the limit as $\gamma \to 0$, we obtain from the continuity of ψ' that:

$$\psi'(-l_0(w)) = \psi'(0).$$

We conclude that $l_0(w) = 0$.

(vii). By (ii), $l_\gamma(w)$ is increasing in γ. Define $l_1(w) = \sup_{\gamma\in(0,1)} l_\gamma(w)$. Since $l_\gamma(w) < w$ for all γ, we have $l_1(w) \leq w$. Furthermore, for every $\gamma \in (0,1)$,

$$\gamma\psi'(r - l_\gamma(w)) + (1-\gamma)\psi'(-l_\gamma(w)) = \psi'(0).$$

Taking the limit as $\gamma \to 1$, we obtain from the continuity of ψ' that:

$$\psi'(w - l_1(w)) = \psi'(0).$$

We conclude that $l_1(w) = w$. □

C.8.3 Proof of Proposition 11

(i). We have:

$$\rho_\gamma(w) = \psi'(0) - (1-\gamma)\psi'(-l_\gamma(w)).$$

Since $l_\gamma(w)$ is strictly increasing in w (Lemma 15) and ψ' is a strictly increasing function, the right-hand side of the above equation is strictly increasing in w. We conclude that $\rho_\gamma(w)$ is strictly increasing in w.

(ii). It follows from the facts that l_γ is a continuous function (Lemma 15), and ψ' is a continuous function.

(iii). The property that $\rho_\gamma(w) \to \gamma$ as $w \to 0$ follows from the facts that $l_\gamma(w) \to 0$ as $w \to 0$ (Lemma 15), ψ' is a continuous function, and $\psi'(0) = 1$.

(iv). The property that $\rho_\gamma(w) \to 1$ as $w \to +\infty$ follows from the equation

$$\lim_{w\to+\infty} \rho_\gamma(w) = 1 - (1-\gamma) \lim_{w\to+\infty} \psi'(-l_\gamma(w)) = 1,$$

where we use the fact that $l_\gamma(w) \to +\infty$ as $w \to +\infty$ (Lemma 15), and the assumption that $\psi'(t) \to 0$ as $t \to -\infty$.

To prove the last part of the proposition, we use a guess-and-verify argument. Let $\rho_\gamma : (0, +\infty) \to (0,1)$ be a strictly increasing, continuous function, with $\rho_\gamma(w) \to \gamma$ as $w \to 0$ and $\rho_\gamma(w) \to 1$ as $w \to +\infty$. To simplify the exposition, set $\rho_\gamma(0) = \gamma$.

If ρ_γ is generated by some ϕ, the two functions are related by the following equations:

$$\rho_\gamma(w) = \gamma\psi'(w - l_\gamma(w)), \tag{65}$$

$$1 - \rho_\gamma(w) = (1-\gamma)\psi'(-l_\gamma(w)), \tag{66}$$

where $\psi = \phi^\star$. We guess a functional form for the Lagrange multiplier:

$$l_\gamma(w) = w - \frac{w}{1+w}.$$

This guess allows us to define ψ' using (65) and (66) for $t \in (-\infty, 1)$:

$$\psi'(t) = \begin{cases} \frac{1}{\gamma}\rho_\gamma\left(\frac{t}{1-t}\right) & \text{if } t \in [0,1) \\ \frac{1}{1-\gamma}\left(1 - \rho_\gamma\left(\frac{\sqrt{t^2-4t}-t}{2}\right)\right) & \text{if } t < 0. \end{cases}$$

To complete the construction, we define $\psi'(t) = t/\gamma$ for $t \in [1, +\infty)$.

Using the properties of ρ_γ, one can verify that ψ' is strictly increasing and continuous. Moreover, the image of ψ' is $(0, +\infty)$. We also have that $\psi'(0) = 1$. Consequently, we can define ψ as follows:

$$\psi(t) = \begin{cases} \int_0^t \psi'(s)\,\mathrm{d}s & \text{if } t \geq 0, \\ -\int_t^0 \psi'(s)\,\mathrm{d}s & \text{if } t < 0. \end{cases}$$

Setting $\phi = \psi^\star$, one can easily verify that ρ_γ is the response function generated by ϕ, with Lagrange multiplier $l_\gamma(w) = w - \frac{w}{1+w}$.

Note that the value of $\psi'(t)$ for $t > 1$ is essentially undetermined. Any other completion of ψ' that preserves continuity, monotonicity, and full range would work. This indicates that multiple ϕ can generate the same response function for a fixed γ.

C.8.4 Proof of Proposition 12

Suppose ϕ_1 and ϕ_2 induce the same response function for all γ. By Lemma 15, $l_\gamma(w) \to 0$ as $\gamma \to 0$ and $l_\gamma(w) \to w$ as $\gamma \to 1$. Thus, for all $w > 0$ and $i \in \{1, 2\}$,

$$\lim_{\gamma \to 0} \frac{\rho_\gamma(w)}{\gamma} = \lim_{\gamma \to 0} \psi_i'(w - l_\gamma(w)) = \psi_i'(w),$$
$$\lim_{\gamma \to 1} \frac{1 - \rho_\gamma(w)}{1 - \gamma} = \lim_{\gamma \to 1} \psi_i'(-l_\gamma(w)) = \psi_i'(-w),$$

where we use the fact that ψ_i' is continuous. Since $\psi_1'(0) = 1 = \psi_2'(0)$, we obtain that $\psi_1' = \psi_2'$. Given that $\psi_1(0) = 0 = \psi_2(0)$, $\psi_1' = \psi_2'$ implies $\psi_1 = \psi_2$, which in turn implies $\phi_1 = \phi_2$.

C.8.5 Proof of Proposition 13

Under Csiszár costs,

$$\frac{\rho_\gamma(w)}{\gamma} = \psi'(r - l_\gamma(w)) \quad \text{and} \quad \frac{1 - \rho_\gamma(w)}{1 - \gamma} = \psi'(-l_\gamma(w)).$$

Using the fact that $\phi' = (\psi')^{-1}$, we obtain:

$$w(x, y) = \phi'(x) - \phi'(y). \tag{67}$$

It follows that properties (i)–(vi) are satisfied. In addition, the inverse response function identifies ϕ:

$$\phi'(x) = \inf_{z \in (0,1)} w(x, z) \quad \text{and} \quad \phi'(y) = -\inf_{z \in (1, +\infty)} w(z, y).$$

To prove the second part of the proposition, let $(x, y) \mapsto w(x, y)$ be a function that satisfies (i)–(vi). We define $\phi' : (0, +\infty) \to \mathbb{R}$ by:

$$\phi'(t) = \begin{cases} \inf_{y \in (0,1)} w(t, y) & \text{if } t > 1, \\ 0 & \text{if } t = 1, \\ -\inf_{x \in (1, +\infty)} w(x, t) & \text{if } t < 1. \end{cases}$$

Claim 15. *The function ϕ' satisfies the following properties:*

(i). *ϕ' is strictly increasing.*

(ii). *ϕ' is continuous.*

(iii). *The image of ϕ' is $\mathbb{R}$.*

Proof. (i). First, take $t > s > 1$. We have:

$$\begin{aligned}\phi'(t) = \inf_{y\in(0,1)} w(t,y) &= \inf_{y\in(0,1)} w(t,y) - w(s,y) + w(s,y)\\ &= w(t,1/2) - w(s,1/2) + \inf_{y\in(0,1)} w(s,y)\\ &> \inf_{y\in(0,1)} w(s,y) = \phi'(s),\end{aligned}$$

where we use the facts that $w(t,y) - w(s,y)$ is independent of y and $w(t,y) > w(s,y)$. This also shows that $\phi'(t) > 0$ for all $t > 1$.

Now, take $1 > t > s$. We have:

$$\begin{aligned}-\phi'(t) = \inf_{x\in(1,+\infty)} w(x,t) &= \inf_{x\in(1,+\infty)} w(x,t) - w(x,s) + w(x,s)\\ &= w(2,t) - w(2,s) + \inf_{x\in(1,+\infty)} w(x,s)\\ &< \inf_{x\in(1,+\infty)} w(x,s) = -\phi'(s),\end{aligned}$$

where we use the facts that $w(x,t) - w(x,s)$ is independent of x and $w(x,t) < w(x,s)$. This also shows that $\phi'(t) < 0$ for all $t < 1$. We conclude that ϕ' is strictly increasing.

(ii). First, we verify continuity at $t > 1$:

$$\begin{aligned}\lim_{s\to t}\phi'(s) = \lim_{s\to t}\left(\inf_{y\in(0,1)} w(s,y)\right) &= \lim_{s\to t}\left(\inf_{y\in(0,1)} w(s,y) - w(t,y) + w(t,y)\right)\\ &= \lim_{s\to t}\left(w(s,1/2) - w(t,1/2) + \inf_{y\in(0,1)} w(t,y)\right)\\ &= \left(\lim_{s\to t} w(s,1/2) - w(t,1/2)\right) + \inf_{y\in(0,1)} w(t,y) = \phi'(t),\end{aligned}$$

where we use the facts that $w(t,y) - w(s,y)$ is independent of y and $w(s,y) \to w(t,y)$ as $s \to t$.

Next, we verify continuity at $t < 1$:

$$\begin{aligned}\lim_{s\to t}\phi'(s) = -\lim_{s\to t}\left(\inf_{x\in(1,+\infty)} w(x,s)\right) &= \lim_{s\to t}\left(\inf_{x\in(1,+\infty)} w(x,s) - w(x,t) + w(x,t)\right)\\ &= \lim_{s\to t}\left(w(2,s) - w(2,t) + \inf_{x\in(1,+\infty)} w(x,t)\right)\\ &= \left(\lim_{s\to t} w(2,s) - w(2,t)\right) + \inf_{x\in(1,+\infty)} w(x,t) = \phi'(t),\end{aligned}$$

where we use the facts that $w(x,t) - w(x,s)$ is independent of x and $w(x,s) \to w(x,t)$ as $t \to s$.

Now, we verify right-continuity at $t = 1$:

$$\lim_{s \downarrow 1} \phi'(s) = \inf_{s \in (1,+\infty)} \phi'(s) = \inf_{s \in (1,+\infty)} \inf_{y \in (0,1)} w(s,y) = 0,$$

where we use the facts that ϕ is decreasing and $w(s,y) \to 0$ as $s \to 1$ and $y \to 1$.

Finally, we verify left-continuity at $t = 1$:

$$\lim_{s \uparrow 1} \phi'(s) = \sup_{s \in (0,1)} \phi'(s) = \sup_{s \in (0,1)} \left(- \inf_{x \in (1,+\infty)} w(x,s) \right) = - \inf_{s \in (0,1)} \inf_{x \in (1,+\infty)} w(x,s) = 0,$$

where we use the facts that ϕ is decreasing and $w(s,y) \to 0$ as $s \to 1$ and $y \to 1$. Overall, we conclude that ϕ' is continuous.

(iii). Since ϕ' is continuous, it is enough to show that $\phi'(t) \to +\infty$ as $t \to +\infty$ and $\phi'(t) \to -\infty$ as $t \to 0$. We have:

$$\begin{aligned}
\lim_{t \to +\infty} \phi'(t) &= \lim_{t \to +\infty} \left(\inf_{y \in (0,1)} w(t,y) \right) \\
&= \lim_{t \to +\infty} \left(\inf_{y \in (0,1)} w(t,y) - w(2,y) + w(2,y) \right) \\
&= \left(\lim_{t \to +\infty} w(t,1/2) - w(2,1/2) \right) + \left(\inf_{y \in (0,1)} w(2,y) \right) = +\infty,
\end{aligned}$$

where we use the facts that $w(t,y) - w(s,y)$ is independent o y as $w(t,y) \to +\infty$ as $t \to +\infty$. Moreover:

$$\begin{aligned}
- \lim_{t \to 0} \phi'(t) &= \lim_{t \to 0} \left(\inf_{x \in (1,+\infty)} w(x,t) \right) \\
&= \lim_{t \to 0} \left(\inf_{x \in (1,+\infty)} w(x,t) - w(x,1/2) + w(x,1/2) \right) \\
&= \left(\lim_{t \to 0} w(2,t) - w(2,1/2) \right) + \left(\inf_{x \in (1,+\infty)} w(x,1/2) \right) = +\infty,
\end{aligned}$$

where we use the facts that $w(x,t) - w(x,s)$ is independent o x as $w(x,t) \to +\infty$ as $t \to 0$. We conclude that the range of ϕ' is $\mathbb{R}$. □

We define $\psi' = (\phi')^{-1}$. Using the properties of ϕ', one can verify that ψ' is strictly increasing and continuous. Moreover, the image of ψ' is $(0, +\infty)$. We also have that $\psi'(0) = 1$. Consequently, we can define ψ as follows:

$$\psi(t) = \begin{cases} \int_0^t \psi'(s) \, \mathrm{d}s & \text{if } t \geq 0, \\ - \int_t^0 \psi'(s) \, \mathrm{d}s & \text{if } t < 0. \end{cases}$$

Setting $\phi = \psi^\star$, we observe that ϕ' is the derivative of ϕ on $(0, +\infty)$.

Let $(x, y) \mapsto \tilde{w}(x, y)$ be the inverse response function generated by ϕ. As shown in (67),

$$\tilde{w}(x, y) = \phi'(x) - \phi'(y) = \inf_{t\in(0,1)} w(x, t) + \inf_{s\in(1,+\infty)} w(s, y).$$

We obtain: for all s',

$$\begin{aligned}\tilde{w}(x, y) &= \inf_{t\in(0,1)} \left(w(x, t) - w(x, y) + w(x, y)\right) + \inf_{s\in(1,+\infty)} w(s, y)\\ &= \inf_{t\in(0,1)} \left(w(s', t) - w(s', y) + w(x, y)\right) + \inf_{s\in(1,+\infty)} w(s, y)\\ &= w(x, y) - w(s', y) + \inf_{t\in(0,1)} w(s', t) + \inf_{s\in(1,+\infty)} w(s, y),\end{aligned}$$

where we use the fact that $w(x, t) - w(x, y)$ is independent of x. It follows that:

$$\tilde{w}(x, y) + \inf_{s'\in(1,+\infty)} w(s', y) = w(x, y) + \inf_{s'\in(1,+\infty)} \inf_{t\in(0,1)} w(s', t) + \inf_{s\in(1,+\infty)} w(s, y).$$

Since $\inf_{s'\in(1,+\infty)} \inf_{t\in(0,1)} w(s', t) = 0$, we conclude that $\tilde{w}(x, y) = w(x, y)$. Thus, $(x, y) \mapsto w(x, y)$ is the inverse response function generated by ϕ.

C.8.6 Proofs of the results in Section 8.3

We begin with deriving properties of the Lagrange multiplier. With respect to Lemma 15, we use the additional hypothesis that ψ is thrice continuously differentiable.

Claim 16. *The Lagrange multiplier $l_\gamma(w)$ is twice continuously differentiable in w. Moreover:*

(i). *For all $w \in (0, +\infty)$, $l'_\gamma(w) \in (0, 1)$.*

(ii). *$l'_\gamma(w) \to 0$ as $\gamma \to 0$.*

(iii). *$l'_\gamma(w) \to 1$ as $\gamma \to 1$.*

(iv). *For all $w \in (0, +\infty)$,*

$$l''_\gamma(w) = R_{\psi'}(w - l_\gamma(w))l'_\gamma(w)(1 - l'_\gamma(w))^2 + R_{\psi'}(-l_\gamma(w))(l'_\gamma(w))^2(1 - l'_\gamma(w)). \quad (68)$$

Proof. By the implicit function theorem,

$$l'_\gamma(w) = \frac{\gamma\psi''(w - l_\gamma(w))}{\gamma\psi''(w - l_\gamma(w)) + (1 - \gamma)\psi''(-l_\gamma(w))}.$$

Since $\psi'' > 0$, we deduce that $l'_\gamma(w) \in (0, 1)$. In addition, because $l_\gamma(w) \to 0$ as $\gamma \to 0$ and $l_\gamma(w) \to w$ as $\gamma \to 1$, we obtain that $l'_\gamma(w) \to 0$ as $\gamma \to 0$ and $l'_\gamma(w) \to 1$ as $\gamma \to 1$. Finally, differentiating $l'_\gamma(w)$ in w, we obtain the desired formula for $l''_\gamma(w)$ after some elementary algebraic manipulation. □

Next, we compute the Arrow-Pratt coefficient of the response function. We have:

$$\begin{aligned}
\rho_\gamma(w) &= \gamma\psi'(w - l_\gamma(w)),\\
\rho'_\gamma(w) &= \gamma\psi''(w - l_\gamma(w))(1 - l'_\gamma(w)),\\
\rho''_\gamma(w) &= \gamma\psi'''(w - l_\gamma(w))(1 - l'_\gamma(w))^2 - \gamma\psi''(w - l_\gamma(w))l''_\gamma(w).
\end{aligned}$$

We obtain:

$$\begin{aligned}
R_{\rho_\gamma}(w) = \frac{\rho''_\gamma(w)}{\rho'_\gamma(w)} &= R_{\psi'}(w - l_\gamma(w))(1 - l'_\gamma(w)) - \frac{l''_\gamma(w)}{1 - l'_\gamma(w)}\\
&= R_{\psi'}(w - l_\gamma(w))(1 - l'_\gamma(w))^2 - R_{\psi'}(-l_\gamma(w))(l'_\gamma(w))^2, \qquad (69)
\end{aligned}$$

where the last line uses the formula for $l''_\gamma(w)$. In addition, we observe:

$$R_{\psi'}(w) = \lim_{\gamma\to 0} R_{\rho_\gamma}(w), \qquad (70)$$

$$R_{\psi'}(-w) = -\lim_{\gamma\to 1} R_{\rho_\gamma}(w), \qquad (71)$$

where we use the facts that $l'_\gamma(w) \to 0$ as $\gamma \to 0$, and $l'_\gamma(w) \to r$ as $\gamma \to 1$. Finally, using the formula $\phi' = (\psi')^{-1}$, we obtain: for all $t \in (0, +\infty)$,

$$R_{\phi'}(t) = -\frac{R_{\psi'}(\phi'(t))}{\psi''(\phi'(t))}. \qquad (72)$$

Proof of Proposition 14. *(i) implies (ii).* Suppose ρ_γ is concave for all γ, namely, $R_{\rho_\gamma} \le 0$ for all γ. Then, (ii) follows from (70) and (71).

(ii) implies (i). Suppose $R_{\psi'}(t) \le 0$ for $t > 0$, and $R_{\psi'}(t) \ge 0$ for $t < 0$. Using (69), we have:

$$R_{\rho_\gamma}(w) = R_{\psi'}(w - l_\gamma(w))(1 - l'_\gamma(w))^2 - R_{\psi'}(-l_\gamma(w))(l'_\gamma(w))^2 \le 0,$$

where we use the facts that $l_\gamma(w) \in (0, w)$ and $l'_\gamma(w) \in (0, 1)$. This proves that ρ_γ is concave.

(ii) if and only if (iii). We obtain from (72) that:

$$R_{\phi'}(t) \ge 0 \quad \Longleftrightarrow \quad R_{\psi'}(\phi'(t)) \le 0.$$

The equivalence of (ii) and (iii) follows from the fact that $t \ge 1$ if and only if $\phi'(t) \ge 0$. □

Next we prove Proposition 15, in successive claims.

Claim 17. *If $R_{\psi'}$ is decreasing, then ρ_γ is S-shaped for all γ.*

Proof. To ease the exposition, we drop the subscript γ. Suppose $w_1 \ge w_2$ and $\rho''(w_1) \ge 0$. Then, $R_\rho(w_1) \ge 0$. Using (69), we obtain:

$$R_{\psi'}(w_1 - l(w_1))(1 - l'(w_1))^2 \ge R_{\psi'}(-l(w_1))(l'(w_1))^2. \qquad (73)$$

Next we distinguish two cases.

Case (i): $R_{\psi'}(w_1 - l(w_1)) \geq 0$. Since $R_{\psi'}$ is decreasing, it follows that all $w_3 \in [w_2, w_1]$,

$$R_{\psi'}(-l(w_1)) \geq R_{\psi'}(-l(w_3)) \geq R_{\psi'}(w_3 - l(w_3)) \geq R_{\psi'}(w_1 - l(w_1)) \geq 0,$$

where we use the fact that both $w - l(w)$ and $l(w)$ are increasing in r (indeed, $1 > l'(w) > 0$). Using (68), we deduce that $l''(w_3) \geq 0$ for all $w_3 \in [w_2, w_1]$. It follows that $l'(w_2) \leq l'(w_1)$. We obtain:

$$\begin{aligned} R_{\psi'}(w_2 - l(w_2))(1 - l'(w_2))^2 &\geq R_{\psi'}(w_1 - l(w_1))(1 - l'(w_1))^2 \\ &\geq R_{\psi'}(-l(w_1))(l'(w_1))^2 \\ &\geq R_{\psi'}(-l(w_2))(l'(w_2))^2. \end{aligned}$$

We conclude that $R_\rho(w_2) \geq 0$, which implies $\rho''(w_2) \geq 0$.

Case (ii): $R_{\psi'}(w_1 - l(w_1)) \leq 0$. By (73), $R_{\psi'}(-l(w_1)) \leq 0$. Since $R_{\psi'}$ is decreasing, it follows that all $w_3 \in [w_2, w_1]$,

$$0 \geq R_{\psi'}(-l(w_1)) \geq R_{\psi'}(-l(w_3)) \geq R_{\psi'}(w_3 - l(w_3)) \geq R_{\psi'}(w_1 - l(w_1)),$$

where we use the fact that both $w - l(w)$ and $l(w)$ are increasing in w (indeed, $1 > l'(w) > 0$). Using (68), we deduce that $l''(w_3) \leq 0$ for all $w_3 \in [w_2, w_1]$. It follows that $l'(w_2) \geq l'(w_1)$. We obtain:

$$\begin{aligned} R_{\psi'}(w_2 - l(w_2))(1 - l'(w_2))^2 &\geq R_{\psi'}(w_1 - l(w_1))(1 - l'(w_1))^2 \\ &\geq R_{\psi'}(-l(w_1))(l'(w_1))^2 \\ &\geq R_{\psi'}(-l(w_2))(l'(w_2))^2. \end{aligned}$$

We conclude that $R_\rho(w_2) \geq 0$, which implies $\rho''(w_2) \geq 0$. □

Claim 18. *If $R_{\psi'}$ is decreasing, then ϕ' is inverse S-shaped.*

Proof. Suppose $t_1 \geq t_2$ and $\phi'''(t_1) \leq 0$. Then, $R_{\phi'}(t_1) \leq 0$. We deduce from (72) that $R_{\psi'}(\phi'(t_1)) \geq 0$. Since ϕ' is increasing, $\phi'(t_1) \geq \phi'(t_2)$. Since $R_{\psi'}$ is decreasing, $R_{\psi'}(\phi'(t_2)) \geq R_{\psi'}(\phi'(t_1)) \geq 0$. Using (72) again, we deduce that $R_{\phi'}(t_2) \leq 0$, which in turn implies $\phi'''(t_2) \leq 0$. We conclude that ϕ' is inverse S-shaped. □

Claim 19. *If $R_{\psi'}$ is decreasing and ψ' is convex or concave, then R_{ρ_γ} is decreasing for all γ.*

Proof. To ease the exposition, we drop the subscript γ. First, we consider the case in which ψ' is convex, namely, $R_{\psi'} \geq 0$. If $w_1 \geq w_2$, then,

$$\begin{aligned} R_\rho(w_1) &= R_{\psi'}(w_1 - l(w_1))(1 - l'(w_1))^2 - R_{\psi'}(-l(w_1))(l'(w_1))^2 \\ &\leq R_{\psi'}(w_2 - l(w_2))(1 - l'(w_2))^2 - R_{\psi'}(-l(w_2))(l'(w_2))^2 = R_\rho(w_2), \end{aligned}$$

where we use the facts that $w - l(w)$ and $l(w)$ are increasing in w, $R_{\psi'}$ is decreasing and non-negative, and $l'(w)$ is increasing in w—see (68). Thus, R_ρ is decreasing when ψ' is convex and $R_{\psi'}$ is decreasing.

Now we consider the case in which ψ' is concave, namely, $R_{\psi'} \leq 0$. If $w_1 \geq w_2$, then,

$$\begin{aligned} R_\rho(w_1) &= R_{\psi'}(w_1 - l(w_1))(1 - l'(w_1))^2 - R_{\psi'}(-l(w_1))(l'(w_1))^2 \\ &\leq R_{\psi'}(w_2 - l(w_2))(1 - l'(w_2))^2 - R_{\psi'}(-l(w_2))(l'(w_2))^2 = R_\rho(w_2), \end{aligned}$$

where we use the facts that $w - l(w)$ and $l(w)$ are increasing in w, $R_{\psi'}$ is decreasing and non-positive, and $l'(w)$ is decreasing in w—see (68). Thus, R_ρ is decreasing when ψ' is concave and $R_{\psi'}$ is decreasing. □

Claim 20. *If $R_{\psi'}$ is decreasing and ψ' is convex or concave, then $R_{\phi'}$ is increasing.*

Proof. Suppose $t_1 \geq t_2$. Using (72), we have

$$R_{\phi'}(t_1) \geq R_{\phi'}(t_2) \quad \Longleftrightarrow \quad R_{\psi'}(\phi'(t_1))\psi''(\phi'(t_2)) \leq R_{\psi'}(\phi'(t_2))\psi''(\phi'(t_1)).$$

Recall that ϕ' is increasing. Thus, since $R_{\psi'}$ is decreasing,

$$R_{\psi'}(\phi'(t_1)) \leq R_{\psi'}(\phi'(t_2)).$$

If ψ' is convex, then $R_{\psi'} \geq 0$ and ψ'' is increasing. If instead ψ' is concave, then $R_{\psi'} \leq 0$ and ψ'' is decreasing. In any case,

$$R_{\psi'}(\phi'(t_1))\psi''(\phi'(t_2)) \leq R_{\psi'}(\phi'(t_2))\psi''(\phi'(t_1)),$$

as desired. □

Claim 21. *If R_{ρ_γ} is decreasing for all γ, then $R_{\psi'}$ is decreasing.*

Proof. Suppose $w_1 \geq w_2 > 0$. Since R_{ρ_γ} is decreasing,

$$R_{\rho_\gamma}(w_1) \leq R_{\rho_\gamma}(w_2).$$

Taking the limit as $\gamma \to 0$, we obtain from (70) that:

$$R_{\psi'}(w_1) \leq R_{\psi'}(w_2).$$

Suppose now that $0 > w_1 \geq w_2$. Since R_{ρ_γ} is decreasing,

$$-R_{\rho_\gamma}(-w_1) \leq -R_{\rho_\gamma}(-w_2).$$

Taking the limit as $\gamma \to 1$, we obtain from (71) that:

$$R_{\psi'}(w_1) \leq R_{\psi'}(w_2).$$

Finally, suppose $w_1 \geq 0 \geq w_2$. For all $w_3 \in (0, w_1]$, $R_{\psi'}(w_1) \leq R_{\psi'}(w_3)$. By continuity, $R_{\psi'}(w_1) \leq R_{\psi'}(0)$. Analogously, for all $w_4 \in [w_2, 0)$, $R_{\psi'}(w_4) \leq R_{\psi'}(w_2)$. By continuity, $R_{\psi'}(0) \leq R_{\psi'}(w_2)$. We deduce that $R_{\psi'}(w_1) \leq R_{\psi'}(w_2)$. We conclude that R_ψ is decreasing. □

C.9 Proofs of the results in Section 9

C.9.1 Proof of Proposition 16

We prove each point in turn.

Point (i). Suppose that K_1 is a garbling of K_2, i.e., there exists $\Gamma : N \to \Delta(N)$ such that $K_1 = \Gamma \circ K_2$. We proceed in two steps.

First, we claim that every P that is replicable under K_1 is also replicable under K_2. Let $P = (\Omega, (P_\theta)_{\theta\in\Theta})$ be given. Suppose there exists a $Q \in \Delta(\Omega)^N$ that replicates P under K_1, i.e., such that $P_\theta = \sum_{i\in N} K_{1,\theta}(i)Q_i$ for all $\theta \in \Theta$. Then, for every $\theta \in \Theta$, we have

$$P_\theta \;=\; \sum_{i\in N}\left(\sum_{j\in N} K_{2,\theta}(j)\Gamma_j(i)\right)Q_i \;=\; \sum_{j\in N} K_{2,\theta}(j)\left(\sum_{i\in N}\Gamma_j(i)Q_i\right),$$

where the first equality is by $K_1 = \Gamma \circ K_2$ and the second equality interchanges the order of summation. Define $Q \circ \Gamma \in \Delta(\Omega)^N$ as $[Q \circ \Gamma]_j := \sum_{i\in N}\Gamma_j(i)Q_i$ for all $j \in N$. Then $Q \circ \Gamma$ replicates P under K_2, i.e., $P_\theta = \sum_{j\in N} K_{2,\theta}(j)[Q \circ \Gamma]_j$ for all $\theta \in \Theta$. This proves the claim.

Next, we claim that, for every P that is replicable under K_1, it holds that $I_{f_1}(P) \geq I_{f_2}(P)$. Let $P = (\Omega, (P_\theta)_{\theta\in\Theta})$ and $Q \in \Delta(\Omega)^N$ that replicates P under K_1 be given. By the above work, $Q \circ \Gamma \in \Delta(\Omega)^N$ replicates P under K_2. Define $\nu_1 := \sum_{\theta\in\Theta}\pi(\theta)K_{1,\theta}$ and $\nu_2 := \sum_{\theta\in\Theta}\pi(\theta)K_{2,\theta}$. Given any $\alpha \in \Delta(\Omega)$, it holds that

$$\begin{aligned}
\sum_{i\in N}\nu_1(i)D_\phi(Q_i\|\alpha) \;&=\; \sum_{i\in N}\left(\sum_{\theta\in\Theta}\pi(\theta)\sum_{j\in N}K_{2,\theta}(j)\Gamma_j(i)\right)D_\phi(Q_i\|\alpha)\\
&=\; \sum_{j\in N}\nu_2(j)\left(\sum_{i\in N}\Gamma_j(i)D_\phi(Q_i\|\alpha)\right)\\
&\geq\; \sum_{j\in N}\nu_2(j)D_\phi\left([Q\circ\Gamma]_j\|\alpha\right),
\end{aligned}$$

where the inequality holds because Lemma 1(ii) implies that the map $D_\phi(\cdot\|\alpha) : \Delta(\Omega) \to \overline{\mathbb{R}}_+$ is convex. Taking α to be the f_1-mean of P, we obtain the claim:

$$I_{f_1}(P) \;\geq\; \sum_{j\in N}\nu_2(j)D_\phi\left([Q\circ\Gamma]_j\|\alpha\right) \;\geq\; \inf_{\beta\in\Delta(\Omega)}\sum_{j\in N}\nu_2(j)D_\Phi\left([Q\circ\Gamma]_j\|\beta\right) \;=\; I_{f_2}(P).$$

Finally, to complete the proof, note that: (a) $I_{f_1}(P) < +\infty$ implies P is replicable under K_1, and (b) $I_{f_1}(P) = +\infty$ implies $I_{f_1}(P) \geq I_{f_2}(P)$. Thus, $I_{f_1}(P) \geq I_{f_2}(P)$ for all $P \in \mathcal{E}$.

Point (ii). Suppose that $\mathrm{dom}(\phi) = \mathbb{R}_+$. We prove the contrapositive. To this end, suppose that K_1 is not a garbling of K_2. Take $P = K_1$. Then $Q \in \Delta(N)^N$ defined as $Q_i(j) = \mathbf{1}(j = i)$ replicates P under K_1. Moreover, since $\mathrm{dom}(\phi) = \mathbb{R}_+$, we have $I_{f_1}(P) < +\infty$. Meanwhile, by the supposition, there does not exist any $R \in \Delta(N)^N$ that replicates $P = K_1$ under K_2 (as any such R would witness that K_1 is a garbling of K_2). Therefore, $I_{f_2}(P) = +\infty$. It follows that $I_{f_2}(P) > I_{f_1}(P)$, which proves the contrapositive.

C.9.2 Proof of Proposition 17

We begin by recalling a general fact about Fenchel conjugates of composite functions due to Hiriart-Urruty (2006). For each $i \in N$, let $h_i^\star : \mathbb{R}^\Theta \to \mathbb{R}$ be a convex function. Let $g^\star : \mathbb{R}^N \to \mathbb{R}$ be an increasing convex function. Define $f^\star : \mathbb{R}^\Theta \to \mathbb{R}$ as the composition $f^\star(x) = g^\star\left((h_i^\star(x))_{i\in N}\right)$ for all $x \in \mathbb{R}^\Theta$. By construction, $f^\star$ is convex. Letting $f = (f^\star)^\star$, $g = (g^\star)^\star$, and $h_i = (h_i^\star)^\star$ for each $i \in N$, we have the following result:

Lemma 16 (Hiriart-Urruty, 2006)**.** *For all $x \in \mathbb{R}_+^\Theta$,*

$$f(x) = \inf\left\{g(y) + \sum_{i\in N} y(i) h_i\left(\frac{z_i}{y(i)}\right)\right\},$$

where the infimum is over all $y \in \mathbb{R}_+^N$ and $z = (z_i)_{i\in N} \in (\mathbb{R}_+^\Theta)^N$ such that $\sum_{i\in N} z_i = x$.

We now use Lemma 16 to prove the proposition. Let $f^\star$ be the asserted conjugate function for the perceptual Csiszár model stated in Proposition 17. Note that it can be written as the composition $f^\star = g^\star \circ (h_i^*)_{i\in N}$ of the increasing convex functions $g^\star$ and $h_i^\star$ defined as

$$g^\star(y) = \sum_{i\in N} \nu(i)\psi\left(\frac{y(i)}{\nu(i)}\right) \quad \text{and} \quad h_i^\star(x) = \sum_{\theta\in\Theta} K_\theta(i)x(\theta),$$

the primal functions of which are given by

$$g(y) = \sum_{i\in N} \nu(i)\phi\left(y(i)\right) \quad \text{and} \quad h_i(x) = \begin{cases} 0 & \text{if } x(\theta) = K_\theta(i) \text{ for all } \theta \in \Theta, \\ +\infty & \text{otherwise.} \end{cases}$$

Applying Lemma 16, we obtain: for every $x \in \mathbb{R}^\Theta$,

$$f(x) = \inf\left\{\sum_{i\in N} \nu(i)\phi\left(y(i)\right)\right\},$$

where the infimum is over all $y \in \mathbb{R}_+^N$ such that $\sum_{i\in N} y(i)K_\theta(i) = x(\theta)$ for all $\theta \in \Theta$. Consequently, given $P \in \Delta(\Omega)^\Theta$ and $\alpha \in \Delta(\Omega)$, simple algebra shows that

$$D_f(P\|\alpha) = \inf\left\{\sum_{i\in N} \nu(i) \sum_{\omega\in\Omega} \alpha(\omega)\phi\left(\frac{Q_i(\omega)}{\alpha(\omega)}\right)\right\}$$

where the infimum is over all $Q \in (\mathbb{R}_+^\Omega)^N$ such that $\sum_{i\in N} Q_i(\omega)K_\theta(i) = P_\theta(\omega)$ for all $\theta \in \Theta$ and $\omega \in \Omega$. Summing over ω for each θ in the replication constraint, we obtain that:

$$\sum_{i\in N}\left(\sum_{\omega\in\Omega} Q_i(\omega)\right) K_\theta(i) = 1 \quad \text{for all } \theta \in \Theta.$$

Assumption 3 then implies that $\sum_{\omega\in\Omega} Q_i(\omega) = 1$ for all $i \in N$. The desired result follows.

C.9.3 Proof of Proposition 18

Let $\mathcal{D} = (\Theta, \pi, A)$ be given. For any (ϕ, N, K), we have

$$\begin{aligned}
&\max_{P \in \Delta(A)^\Theta} \sum_{\theta \in \Theta} \pi(\theta) \sum_{a \in A} P_\theta(a) a(\theta) - I(P) \\
&= \max_{P \in \Delta(A)^\Theta} \sup_{Q \in \Delta(A)^N} \sum_{\theta \in \Theta} \pi(\theta) \sum_{a \in A} P_\theta(a) a(\theta) - J(Q) \quad \text{s.t.} \quad Q \circ K = P \\
&= \max_{Q \in \Delta(A)^N} \sum_{\theta \in \Theta} \pi(\theta) \sum_{a \in A} [Q \circ K]_\theta(a) a(\theta) - J(Q) \\
&= \max_{Q \in \Delta(A)^N} \sum_{i \in N} \nu(i) \sum_{a \in A} Q_i(a) E_i[a] - J(Q) \\
&= \max_{\bar{Q} \in \Delta(\bar{A})^N} \sum_{i \in N} \nu(i) \sum_{\bar{a} \in \bar{A}} \bar{Q}_i(\bar{a}) \bar{a}(i) - J(\bar{Q}),
\end{aligned}$$

where the first equality follows from Definition 14, the second equality is by substitution of the constraint, and the remaining equalities rearrange terms. The desired result follows.

C.9.4 Proof of Proposition 19

(i). Assume the encoder satisfies the MLRP. By standard arguments (Karlin and Rinott, 1980; Milgrom, 1981), K_θ and μ_i are increasing in θ and i, respectively, according to first-order stochastic dominance. Since $r(\theta)$ is increasing in θ, $E_i[r]$ is increasing in i.

By Proposition 18, the reduced Lagrange multiplier $E_i[\lambda_\pi]$ is a solution of the equation

$$\alpha(r)\psi'(E_i[r] - E_i[\lambda_\pi]) + \alpha(s)\psi'(0 - E_i[\lambda_\pi]) = \psi'(0).$$

Since ψ' is a strictly increasing function, $E_i[r]$ is increasing in i, the quantity $E_i[r - \lambda_\pi]$ must be increasing in i. It follows that

$$Q_i(r) = \alpha(r)\psi'(E_i[r - \lambda_\pi])$$

is increasing in i. We obtain that

$$P_\theta(r) = \sum_{i \in N} K_\theta(i) Q_i(r)$$

is increasing in θ, given that K_θ is increasing in θ according to first-order stochastic dominance.

(ii). As shown above,

$$P_\theta(r) = \sum_{i \in N} K_\theta(i) Q_i(r)$$

where the quantity $Q_i(r)$ is increasing in i. Simple algebra shows that the psychometric function is convex at θ_i if and only if

$$\sum_{i \in N} \left(\frac{1}{2} K_{\theta_{i-1}} + \frac{1}{2} K_{\theta_{i+1}} \right) Q_i(r) \geq \sum_{i \in N} K_{\theta_i} Q_i(r).$$

Consequently, if $\frac{1}{2}K_{\theta_{i-1}} + \frac{1}{2}K_{\theta_{i+1}}$ first-order stochastically dominates K_{θ_i}, then the psychometric function is convex at θ_i.

(iii). Same argument as in (ii), with the directions reversed.

C.10 Proofs of the results in Section 10

C.10.1 Proof of Proposition 20

For notational convenience, we let $K_{\mathrm{NS}} \in \Delta(N)^{\Theta}$ denote the encoder associated with (37). The linear independence assumption implies that $\{\mu_i\}_{i \in N}$ is affinely independent.

By direct calculation, C can be equivalently written, for every experiment $P \in \Delta(\Omega)^{\Theta}$, as

$$C(P) = \min \left\{ \sum_{\omega \in \mathrm{supp}(R_\rho)} R_\rho(\omega)\, \zeta\, D_{\mathrm{KL}}(\overline{r}_\omega \| \nu) \right\},$$

where the minimum is taken over all $K \in \Delta(N)^{\Theta}$ and $R \in \Delta(\Omega)^{\Theta \times N}$ subject to

$$\sum_{i \in N} K_\theta(i) R_{(\theta,i)} = P_\theta \text{ for all } \theta \in \Theta \tag{74}$$

and

$$r_{\omega,i} = \mu_i \quad \text{for all } \omega \in \mathrm{supp}(R_\rho) \text{ and } i \in \mathrm{supp}(\overline{r}_\omega). \tag{75}$$

In the above, $\rho \in \Delta(\Theta \times N)$ is an induced prior belief on $\Theta \times N$ given by $\rho(\theta, i) = \pi(\theta) K_\theta(i)$, $R_\rho \in \Delta(\Omega)$ is given by $R_\rho(\omega) = \sum_{(\theta,i) \in \Theta \times N} \rho(\theta, i) R_{(\theta,i)}(\omega)$, and $r_\omega \in \Delta(\Theta \times N)$ is the posterior belief on $\Theta \times N$ conditional on signal ω. Moreover, $\overline{r}_\omega \in \Delta(N)$ is the marginal distribution of r_ω on N, and $r_{\omega,i} \in \Delta(\Theta)$ is the conditional distribution on Θ conditional on attribute $i \in N$, which can be expressed as

$$r_{\omega,i}(\theta) = \frac{\rho(\theta, i) R_{(\theta,i)}(\omega)}{\sum_{\tau \in \Theta} \rho(\tau, i) R_{(\tau,i)}(\omega)}. \tag{76}$$

Take any P for which the constraint set is nonempty, and any (K, R) and associated ρ satisfying (74) and (75). Let $\overline{\rho} \in \Delta(N)$ be the marginal of ρ on N, defined as $\overline{\rho}(i) = \sum_{\theta \in \Theta} \rho(\theta, i)$. Plugging (76) into (75) yields, for all $\theta \in \Theta$, $\omega \in \mathrm{supp}(R_\rho)$, and $i \in \mathrm{supp}(\overline{r}_\omega)$,

$$\rho(\theta, i) R_{(\theta,i)}(\omega) = \mu_i(\theta) \cdot \sum_{\tau \in \Theta} \rho(\tau, i) R_{(\tau,i)}(\omega). \tag{77}$$

Summing over $\omega \in \mathrm{supp}(R_\rho)$, this delivers

$$\frac{\rho(\theta, i)}{\overline{\rho}(i)} = \mu_i(\theta) \quad \text{for all } \theta \in \Theta \text{ and } i \in \mathrm{supp}(\overline{\rho}). \tag{78}$$

Plugging this back into (77) and defining $\rho_i \in \Delta(\Theta)$ as $\rho_i(\theta) = \rho(\theta, i)/\overline{\rho}(i)$, we obtain

$$R_{(\theta,i)} = \sum_{\tau \in \Theta} \rho_i(\tau) R_{(\tau,i)} \quad \text{for all } (\theta, i) \in \mathrm{supp}(\rho).$$

For each $i \in \mathrm{supp}(\overline{\rho})$, this implies that there exists $\widehat{R}_i \in \Delta(\Omega)$ such that $\widehat{R}_i = R_{(\theta,i)}$ for all $(\theta, i) \in \mathrm{supp}(\rho)$; moreover, for any $(\theta, i) \notin \mathrm{supp}(\rho)$, we can replace $R_{(\theta,i)}$ with $\widehat{R}_i$ without affecting the constraints (74) and (76) or the cost. Finally, for each $i \notin \mathrm{supp}(\overline{\rho})$, let $\widehat{R}_i \in \Delta(\Omega)$

be arbitrary. Denote the resulting experiment on N as $\widehat{R} \in \Delta(\Omega)^N$. By construction, when $\widehat{R}$ is viewed as a θ-independent experiment on $\Theta \times N$, the pair $(K, \widehat{R})$ satisfies (74) and (75) and attains the same cost as (K, R).

Meanwhile, note that constraint (74) implies $\pi(\theta) = \sum_{i \in N} \rho(\theta, i)$ for all $\theta \in \Theta$. Together with (78), this yields $\pi = \sum_{i \in N} \overline{\rho}(i)\mu_i$. Since we also have $\pi = \sum_{i \in N} \nu(i)\mu_i$ and $\{\mu_i\}_{i \in N}$ is affinely independent, it follows that $\overline{\rho} = \nu$. We thus obtain $\rho(\theta, i) = \nu(i)\mu_i(\theta) = \pi(\theta)K_{\mathrm{NS},\theta}(i)$ for all $\theta \in \Theta$ and $i \in N$. We conclude that $K = K_{\mathrm{NS}}$ and $R_\rho = \widehat{R}_\nu := \sum_{i \in N} \nu(i)\widehat{R}_i$.

Overall, we obtain: for any $P \in \Delta(\Omega)^\Theta$,

$$\begin{aligned} C(P) &= \min_{R \in \Delta(\Omega)^N} \left\{ \sum_{\omega \in \mathrm{supp}(R_\nu)} R_\nu(\omega)\, \zeta\, D_{\mathrm{KL}}(\overline{r}_\omega \| \nu) \right\} \quad \text{subject to } R \circ K_{\mathrm{NS}} = P. \\ &= \min_{R \in \Delta(\Omega)^N} \left\{ \sum_{i \in N} \nu(i)\, \zeta\, D_{\mathrm{KL}}(R_i \| R_\nu) \right\} \quad \text{subject to } R \circ K_{\mathrm{NS}} = P, \end{aligned}$$

where we use a standard change of variables for KL divergence. The result follows.

C.10.2 Proof of Proposition 21

Under deterministic categorization, we have $\nu(i) = \pi(B_i)$ and $\mu_i = \pi(\cdot \mid B_i)$ for all $i \in N$.

Fix any $p \in \Delta(\Theta)$. Note that the extension (r, q) defined as $r(i) = p(B_i)$ and $q_i = p(\cdot \mid B_i)$ for all $i \in N$ satisfies $\sum_{i \in N} r(i)q_i = p$ and achieves a finite value in (36). Hence, problem (36) is feasible. Thus, since the feasible set is compact and the objective is lower semi-continuous, a minimizer in (36) exists. We show that (r, q) is the essentially unique minimizer; formally, any minimizer (r', q') must satisfy $r' = r$ and $q'_i = q_i$ for all $i \in N$ such that $\mathrm{supp}(p) \cap B_i \neq \emptyset$.[47]

To this end, take any extension (r', q') that achieves the infimum in (36). Since it achieves a finite value, we must have $q'_i \ll \mu_i$ for all $i \in \mathrm{supp}(r')$. Therefore, feasibility implies that: (i) $i \in \mathrm{supp}(r')$ if only if $\mathrm{supp}(p) \cap B_i \neq \emptyset$, and (ii) for every $i \in \mathrm{supp}(r')$ and $\theta \in B_i$, $q'_i(\theta) = p(\theta)/r'(i)$. It follows that $r' = r$ and $q'_i = q_i$ for all $i \in N$ such that $\mathrm{supp}(p) \cap B_i \neq \emptyset$.

C.10.3 Proof of Proposition 22

By inspection, the function $H^\star_{\mathrm{NS}}$ in Proposition 22 can be written as the composition $H^\star_{\mathrm{NS}} = g^\star \circ (h^\star_i)_{i \in N}$ of the functions $g^\star : \mathbb{R}^N \to \mathbb{R}$ and $h^\star_i : \mathbb{R}^\Theta \to \mathbb{R}$ defined as

$$g^\star(y) = \zeta \log \left(\sum_{i \in N} \nu(i) e^{y(i)/\zeta} \right) \quad \text{and} \quad h^\star_i(x) = \eta_i \log \left(\sum_{\theta \in \Theta} \mu_i(\theta) e^{x(\theta)/\eta_i} \right).$$

It is easy to see that $g^\star$ and all the h^*_i are increasing. It can also be verified, via Hölder's inequality, that all these functions are convex. Hence, Lemma 16 implies that $(H^\star_{\mathrm{NS}})^\star = H_{\mathrm{NS}}$.

[47] If there is a nest B_i such that $\mathrm{supp}(p) \cap B_i = \emptyset$ and hence $r(i) = 0$, we can define $q_i, q'_i \in \Delta(\Theta)$ arbitrarily.

C.10.4 Proof of Proposition 23

For the specified parameters, the conjugate function in Proposition 22 simplifies as

$$H^\star(x) = c(\zeta,\eta) + \zeta \log\left(\sum_{i\in N}\left(\sum_{\theta\in i} e^{x(\theta)/\eta}\right)^{\eta/\zeta}\right), \tag{79}$$

where $N = \{U, D, L, R\}$ and $c(\zeta,\eta) \in \mathbb{R}$ is a constant that depends only on ζ and η. As in Example 6, we also define $f^\star : \mathbb{R}^\Theta \to \mathbb{R}$ as $f^\star(x) = H^\star(\frac{x}{\pi})$ and $f : \mathbb{R}^\Theta_+ \to \overline{\mathbb{R}}$ as $f = (f^\star)^\star$.

Claim 22. *$H^\star$ is strictly convex modulo translations.*

Proof. By construction, $H^\star$ is convex and translation invariant. Thus, it suffices to show that $H^\star$ is non-affine modulo translations. To this end, take any $t \in (0,1)$ and $x, y \in \mathbb{R}^\Theta$ such that $x - y \notin \mathbb{R}$. There must exist some $i \in \{U, D, L, R\}$ such that, letting $i = \{\theta, \tau\}$, $x(\theta) - y(\theta) \neq x(\tau) - y(\tau)$. Hölder's (strict) inequality then implies that

$$\begin{aligned} e^{(tx(\theta)+(1-t)y(\theta))/\eta} + e^{(tx(\tau)+(1-t)y(\tau))/\eta} &= \left(e^{x(\theta)/\eta}\right)^t \left(e^{y(\theta)/\eta}\right)^{1-t} + \left(e^{x(\tau)/\eta}\right)^t \left(e^{y(\tau)/\eta}\right)^{1-t} \\ &< \left(e^{x(\theta)/\eta} + e^{x(\tau)/\eta}\right)^t \left(e^{y(\theta)/\eta} + e^{y(\tau)/\eta}\right)^{1-t}, \end{aligned}$$

where the inequality is strict because the hypothesis that $x(\theta) - y(\theta) \neq x(\tau) - y(\tau)$ implies that the vectors $(e^{x(\theta)/\eta}, e^{x(\tau)/\eta}), (e^{y(\theta)/\eta}, e^{y(\tau)/\eta}) \in \mathbb{R}^2_+$ are linearly independent. Therefore,

$$\begin{aligned} H^\star(tx + (1-t)y) - c(\zeta,\eta) &= \zeta \log\left(\sum_{i\in N}\left(\sum_{\theta'\in i} e^{(tx(\theta')+(1-t)y(\theta'))/\eta}\right)^{\eta/\zeta}\right) \\ &< \zeta \log\left(\sum_{i\in N}\left(\sum_{\theta'\in i} e^{x(\theta')/\eta}\right)^{t\eta/\zeta}\left(\sum_{\theta'\in i} e^{y(\theta')/\eta}\right)^{(1-t)\eta/\zeta}\right) \\ &\leq \zeta\left(\left[\sum_{i\in N}\left(\sum_{\theta'\in i} e^{x(\theta')/\eta}\right)^{\eta/\zeta}\right]^t \times \left[\sum_{i\in N}\left(\sum_{\theta'\in i} e^{y(\theta')/\eta}\right)^{\eta/\zeta}\right]^{1-t}\right), \end{aligned}$$

where the strict inequality follows from the preceding display and an analogous application of Hölder's (weak) inequality to each term of the outer sum, and the final line follows from applying Hölder's (weak) inequality to the entire outer sum. Upon simplification, we obtain the desired strict inequality $H^\star(tx + (1-t)y) < tH^\star(x) + (1-t)H^\star(y)$. □

Next, denote by Γ the group of permutations generated by $\gamma_1, \gamma_2 : \Theta \to \Theta$, where each γ_i permutes the ith component of the state.[48] That is, $\gamma_1(u,\cdot) = (d,\cdot)$, $\gamma_1(d,\cdot) = (u,\cdot)$, $\gamma_2(\cdot,l) = (\cdot,r)$, and $\gamma_2(\cdot,r) = (\cdot,l)$. By inspection, each decision problem $\mathcal{D}_j = (\pi, A_j)$ with $j \in \{1,2,3\}$ is invariant with respect to Γ. In particular:

[48] The permutation group Γ includes γ_1, γ_2, the composition $\gamma_1 \circ \gamma_2 = \gamma_2 \circ \gamma_1$, and the identity map.

- In problem 1, $a_U = a_{U,\gamma_2} = a_{D,\gamma_1}$ and $a_D = a_{D,\gamma_2} = a_{U,\gamma_1}$.
- In problem 2, $a_L = a_{L,\gamma_1} = a_{R,\gamma_2}$ and $a_R = a_{R,\gamma_1} = a_{L,\gamma_2}$.
- In problem 3, $a_{\text{diag}} = a_{\text{off},\gamma_1} = a_{\text{off},\gamma_2}$ and $a_{\text{off}} = a_{\text{diag},\gamma_1} = a_{\text{diag},\gamma_2}$.

Moreover, by inspection, the conjugate $H^\star$ in (79) is invariant with respect to Γ. Therefore, $f^\star$ is also invariant because π is uniform. By Lemma 4, it follows that f is also invariant with respect to Γ. Using these facts, Proposition 1 then implies that, for each decision problem $j \in \{1,2,3\}$, there exists a saddle point $(\alpha^j, \lambda^j) \in \Delta(A_j) \times \mathbb{R}^\Theta$ such that $\alpha^j(a) = 1/2$ for all $a \in A_j$ and $\lambda^j = \lambda^j_{\gamma_1} = \lambda^j_{\gamma_2}$, which implies that $\theta \mapsto \lambda^j(\theta)$ is constant; by translation invariance, we can set $\lambda^j = \mathbf{0}$ without loss of generality. Moreover, letting $P^j \in \Delta(A^j)^\Theta$ denote the associated optimal choice rule for each problem $j \in \{1,2,3\}$, Proposition 1 implies:

- In problem 1, $P^1_{(u,l)} = P^1_{(u,r)}$ and $P^1_{(d,l)} = P^1_{(d,r)}$; it suffices to find $P^1_{(u,l)}(a_U)$ and $P^1_{(d,l)}(a_U)$.
- In problem 2, $P^2_{(u,l)} = P^2_{(d,l)}$ and $P^2_{(u,r)} = P^2_{(d,r)}$; it suffices to find $P^2_{(u,l)}(a_L)$ and $P^2_{(u,r)}(a_L)$.
- In problem 3, $P^3_{(u,l)}(a_{\text{diag}}) = P^3_{(d,r)}(a_{\text{diag}}) = P^3_{(u,r)}(a_{\text{off}}) = P^3_{(d,l)}(a_{\text{off}})$ and $P^3_{(u,r)}(a_{\text{diag}}) = P^3_{(d,l)}(a_{\text{diag}}) = P^3_{(u,l)}(a_{\text{off}}) = P^3_{(d,r)}(a_{\text{off}})$; it suffices to find $P^3_{(u,l)}(a_{\text{diag}})$ and $P^3_{(d,l)}(a_{\text{diag}})$.

Claim 23. *For each problem $j \in \{1,2,3\}$, P^j is the unique optimal stochastic choice rule.*

Proof. We focus here on problem $j = 1$; the other cases are analogous and hence omitted. Take any saddle point $(\beta^1, \ell^1) \in \Delta(A_1) \times \mathbb{R}^\Theta$ for problem 1. Claim 22, Lemma 6, and Lemma 7 together imply that $\ell^1 - \lambda^1 \in \mathbb{R}$, i.e., $\ell^1 = \mathbf{0}$ modulo translations. Hence, $(\beta^1, \mathbf{0})$ is also a saddle point for problem 1. Since $f^\star$ is translation invariant (as π is uniform), this saddle point generates the same choice rule P^1 if $\beta^1 = \alpha^1$. Thus, it suffices to show that $\beta^1 = \alpha^1$.

For saddle point $(\beta^1, \mathbf{0})$, the optimality condition (16) reads

$$\beta^1(a_U)\nabla_\theta f^\star(a_{U,\pi}) + \Big(1 - \beta^1(a_U)\Big)\nabla_\theta f^\star(a_{D,\pi}) = 1 \quad \text{for all } \theta \in \Theta.$$

Combining these conditions for states $\theta \in \{(u,l),(d,l)\}$, we obtain

$$\beta^1(a_{U,\pi})\Big[\nabla_{(u,l)} f^\star(a_{U,\pi}) - \nabla_{(d,l)} f^\star(a_{U,\pi})\Big] = \Big(1 - \beta^1(a_{U,\pi})\Big)\Big[\nabla_{(d,l)} f^\star(a_{D,\pi}) - \nabla_{(u,l)} f^\star(a_{D,\pi})\Big].$$

Lemma 4 with the permutation $\gamma_1 \in \Gamma$ implies that $\nabla_{(u,l)} f^\star(a_{U,\pi}) = \nabla_{(d,l)} f^\star(a_{D,\pi})$ and $\nabla_{(d,l)} f^\star(a_{U,\pi}) = \nabla_{(u,l)} f^\star(a_{D,\pi})$. Moreover, direct calculation yields $\nabla_{(u,l)} f^\star(a_{U,\pi}) > \nabla_{(d,l)} f^\star(a_{U,\pi})$. It follows from the above display that $\beta^1(a_U) = 1 - \beta^1(a_U)$. Hence, we obtain $\beta^1 = \alpha^1$. □

In the posterior separable case, as noted in Section 4.7 the optimal choice rule is given by

$$P^j_\theta(a) = \frac{\alpha^j(a)}{\pi(\theta)}\nabla_\theta H^\star(a) = 2\,\nabla_\theta H^\star(a)$$

for all $\theta \in \Theta$, $j \in \{1,2,3\}$, and $a \in A_j$. Observe that the gradient $\nabla H^\star$ is given by

$$\nabla_\theta H^\star(x) = \sum_{i \in N\,:\,\theta \in i} \frac{e^{x(\theta)/\eta}}{\sum_{\tau \in i} e^{x(\tau)/\eta}} \times \frac{\left(\sum_{\tau \in i} e^{x(\tau)/\eta}\right)^{\eta/\zeta}}{\sum_{j \in N} \left(\sum_{\tau \in j} e^{x(\tau)/\eta}\right)^{\eta/\zeta}}.$$

We now specialize this formula to the three decision problems, considering each in turn.

Problem 1. Per the above, it suffices to find $\nabla_{(u,l)} H^\star(a_U)$ and $\nabla_{(d,l)} H^\star(a_U)$. First, we have

$$\begin{aligned}\nabla_{(u,l)} H^\star(a_U) = &\left[\frac{e^{1/\eta}}{2 \cdot e^{1/\eta}} \times \frac{\left(2 \cdot e^{1/\eta}\right)^{\eta/\zeta}}{\left(2 \cdot e^{1/\eta}\right)^{\eta/\zeta} + 2\left(e^{1/\eta}+1\right)^{\eta/\zeta} + 2^{\eta/\zeta}}\right] \\ &+ \left[\frac{e^{1/\eta}}{e^{1/\eta}+1} \times \frac{\left(e^{1/\eta}+1\right)^{\eta/\zeta}}{\left(2 \cdot e^{1/\eta}\right)^{\eta/\zeta} + 2\left(e^{1/\eta}+1\right)^{\eta/\zeta} + 2^{\eta/\zeta}}\right],\end{aligned}$$

where the first term in brackets corresponds to $i = U$ and the second term in brackets corresponds to $i = L$. After simplification, this becomes

$$\nabla_{(u,l)} H^\star(a_U) = \left[\frac{1}{2} \times \frac{e^{1/\zeta}}{e^{1/\zeta} + 2\left(\frac{e^{1/\eta}+1}{2}\right)^{\eta/\zeta} + 1}\right] + \left[\frac{e^{1/\eta}}{e^{1/\eta}+1} \times \frac{\left(\frac{e^{1/\eta}+1}{2}\right)^{\eta/\zeta}}{e^{1/\zeta} + 2\left(\frac{e^{1/\eta}+1}{2}\right)^{\eta/\zeta} + 1}\right]. \quad (80)$$

Next, we have

$$\begin{aligned}\nabla_{(d,l)} H^\star(a_U) = &\left[\frac{1}{2} \times \frac{2^{\eta/\zeta}}{\left(2 \cdot e^{1/\eta}\right)^{\eta/\zeta} + 2\left(e^{1/\eta}+1\right)^{\eta/\zeta} + 2^{\eta/\zeta}}\right] \\ &+ \left[\frac{e^{1/\eta}}{e^{1/\eta}+1} \times \frac{\left(e^{1/\eta}+1\right)^{\eta/\zeta}}{\left(2 \cdot e^{1/\eta}\right)^{\eta/\zeta} + 2\left(e^{1/\eta}+1\right)^{\eta/\zeta} + 2^{\eta/\zeta}}\right],\end{aligned}$$

where the first term in brackets corresponds to $i = D$ and the second term in brackets corresponds to $i = L$. After simplification, this becomes

$$\nabla_{(d,l)} H^\star(a_U) = \left[\frac{1}{2} \times \frac{1}{e^{1/\zeta} + 2\left(\frac{e^{1/\eta}+1}{2}\right)^{\eta/\zeta} + 1}\right] + \left[\frac{e^{1/\eta}}{e^{1/\eta}+1} \times \frac{\left(\frac{e^{1/\eta}+1}{2}\right)^{\eta/\zeta}}{e^{1/\zeta} + 2\left(\frac{e^{1/\eta}+1}{2}\right)^{\eta/\zeta} + 1}\right]. \quad (81)$$

To calculate the desired limits, we note that, for any $\eta > 0$,

$$\left(\frac{e^{1/\eta}+1}{2}\right)^{\eta} < e.$$

Therefore, (80) implies that

$$\lim_{\zeta \to 0} \nabla_{(u,l)} H^\star(a_U) = \frac{1}{2} \cdot 1 + \frac{e^{1/\eta}}{e^{1/\eta}+1} \cdot 0 = \frac{1}{2},$$

and (81) implies that

$$\lim_{\zeta\to 0} \nabla_{(d,l)} H^\star(a_U) = \frac{1}{2}\cdot 0 + \frac{e^{1/\eta}}{e^{1/\eta}+1}\cdot 0 = 0.$$

By the symmetry properties noted above, we conclude that $\lim_{\zeta\to 0} P^1_\theta(a) = \mathbf{1}(a(\theta)=1)$.

Problem 2. The calculations are symmetric to those for Problem 1, and hence omitted.

Problem 3. Per the above, it suffices to find $\nabla_{(u,l)} H^\star(a_{\text{diag}})$ and $\nabla_{(d,l)} H^\star(a_{\text{diag}})$. Noting that a_{diag} pays the reward in exactly one state within each nest $i \in \{U, D, L, R\}$, we obtain

$$\nabla_{(u,l)} H^\star(a_{\text{diag}}) = 2\cdot\left[\frac{e^{1/\eta}}{e^{1/\eta}+1} \times \frac{\left(e^{1/\eta}+1\right)^{\eta/\zeta}}{4\cdot\left(e^{1/\eta}+1\right)^{\eta/\zeta}}\right] = \frac{1}{2}\cdot\frac{e^{1/\eta}}{e^{1/\eta}+1},$$

$$\nabla_{(d,l)} H^\star(a_{\text{diag}}) = 2\cdot\left[\frac{1}{e^{1/\eta}+1} \times \frac{\left(e^{1/\eta}+1\right)^{\eta/\zeta}}{4\cdot\left(e^{1/\eta}+1\right)^{\eta/\zeta}}\right] = \frac{1}{2}\cdot\frac{1}{e^{1/\eta}+1}.$$

Note that both of these expressions are independent of $\zeta > 0$. Hence, by the symmetry properties noted above, they yield the desired form of P^3 for all $\zeta > 0$, viz., as $\zeta \to 0$.

C.10.5 Proof of Proposition 24

Define $P^1 \in \Delta(A_1)^\Theta$ and $P^2 \in \Delta(A_2)^\Theta$ as $P^1_\theta(a_U) = \mathbf{1}(\theta \in U)$ and $P^2_\theta(a_L) = \mathbf{1}(\theta \in L)$ for all $\theta \in \Theta$. The associated unconditional action probabilities and posteriors are given by

$$\begin{aligned} P^1_\pi(a) = 1/2 \quad &\text{and} \quad p^1_a(\theta) = \frac{1}{2}\mathbf{1}(a(\theta)=1) \quad \text{for all } a \in A_1,\ \theta \in \Theta, \\ P^2_\pi(a) = 1/2 \quad &\text{and} \quad p^2_a(\theta) = \frac{1}{2}\mathbf{1}(a(\theta)=1) \quad \text{for all } a \in A_2,\ \theta \in \Theta. \end{aligned}$$

Suppose that $\lim_{n\to\infty} P^{1,n}_\theta(a_U) = P^1_\theta(a_U)$ and $\lim_{n\to\infty} P^{2,n}_\theta(a_L) = P^2_\theta(a_L)$, which implies that

$$\begin{aligned} \lim_{n\to\infty} P^{1,n}_\pi(a) = P^1_\pi(a) \quad &\text{and} \quad \lim_{n\to\infty} p^{1,n}_a = p^1_a \quad \text{for all } a \in A_1, \\ \lim_{n\to\infty} P^{2,n}_\pi(a) = P^2_\pi(a) \quad &\text{and} \quad \lim_{n\to\infty} p^{2,n}_a = p^2_a \quad \text{for all } a \in A_2. \end{aligned}$$

Throughout the proof, we adopt the following notational conventions. First, for each vector of coefficients $\kappa \in \mathbb{R}^I_+$, we denote by $C(\cdot;\kappa) : \mathcal{E} \to \mathbb{R}_+$ the associated neighborhood-based cost defined via (38). Second, for each $i \in I$, we extend the KL divergence $D_{\text{KL}}(\cdot\|\pi_i)$ on $\Delta(B_i)$ to the orthant $\mathbb{R}^{B_i}_+$ by defining (with minor abuse of notation) the map $D_{\text{KL}}(\cdot\|\pi_i) : \mathbb{R}^{B_i}_+ \to \mathbb{R}_+$ as

$$D_{\text{KL}}(x\|\pi_i) = \sum_{\theta\in B_i} x(\theta)\log\left(\frac{x(\theta)}{\pi_i(\theta)}\right).$$

This extension, which is without loss of generality, allows us to take derivatives of $D_{\mathrm{KL}}(\cdot\|\pi_i)$ in the usual way on $\mathbb{R}_+^{B_i}$. Finally, we define the maps $H^n : \mathbb{R}_+^\Theta \to \mathbb{R}_+$ and $H : \mathbb{R}_+^\Theta \to \mathbb{R}_+$ as

$$H^n(x) = \sum_{i\in I} \kappa_i^n\, \overline{x}(i) D_{\mathrm{KL}}(x_i\|\pi_i) \quad \text{and} \quad H(x) = \sum_{i\in I} \kappa_i^*\, \overline{x}(i) D_{\mathrm{KL}}(x_i\|\pi_i),$$

where $\overline{x}(i) = \sum_{\theta\in B_i} x(\theta)$ and $x_i(\theta) = x(\theta)/\overline{x}(i)$ for all $i \in I$ and $\theta \in B_i$. This is shorthand notation for the entropy (38) with coefficients κ^n and κ^*, respectively, extended to the orthant.

We first show, via three claims, that the coefficients converge to $\kappa_i^* = 0$ for all $i \in I$.

Claim 24. *$\kappa_i^* < +\infty$ for every $i \in I$.*

Proof. Suppose, towards a contradiction, that there exists $i \in I$ with $\kappa_i^* = +\infty$. Since $|B_i| \geq 2$ by the nonredudancy assumption, there is some $E \in \{U, D, L, R\}$ such that $B_i \cap E \neq \emptyset$ and $B_i \backslash E \neq \emptyset$. We suppose here that $E = U$; the other cases are specular and hence omitted.

Consider the decision maker's cost in problem 1. For each $n \in \mathbb{N}$, it holds that

$$\begin{aligned} C(P^{1,n};\kappa^n) &= P_\pi^{1,n}(a_U) H^n(p_{a_U}^{1,n}) + (1 - P_\pi^{1,n}(a_U)) H^n(p_{a_D}^{1,n}) \\ &\geq \kappa_i^n \left[P_\pi^{1,n}(a_U)\, \overline{p}_{a_U}^{1,n}(i)\, D_{\mathrm{KL}}\left(p_{a_U,i}^{1,n} \| \pi_i \right) \right]. \end{aligned}$$

Note that $\lim_{n\to\infty} \overline{p}_{a_U}^{1,n}(i) = \frac{1}{2}|B_i \cap U| \geq 1/2$, where the inequality is by $B_i \cap U \neq \emptyset$. Moreover, $\lim_{n\to\infty} p_{a_U,i}^{1,n} = p_{a_U,i}^{1,}$ and $\mathrm{supp}(p_{a_U,i}^1) \subsetneq \mathrm{supp}(\pi_i) = B_i$, where the strict inclusion is by $B_i \backslash U \neq \emptyset$. Thus, by continuity of KL divergence, given any $\epsilon \in (0,1)$ and sufficiently large n,

$$C(P^{1,n};\kappa^n) \geq \kappa_i^n \left[(1-\epsilon)\, \frac{1}{4}\, D_{\mathrm{KL}}\left(p_{a_U,i}^1 \| \pi_i \right) \right].$$

Since the term in brackets is strictly positive and $\kappa_i^n \to \kappa_i^* = +\infty$, we obtain $C(P^{1,n};\kappa^n) \to +\infty$. This contradicts the optimality of $P^{1,n}$ in decision problem 1 for large n, as desired. □

Claim 25. *It holds that*

$$\lim_{n\to\infty} C(P^{1,n};\kappa^n) = C(P^1;\kappa^*) = \frac{1}{2} H(p_{a_U}^1) + \frac{1}{2} H(p_{a_D}^1), \tag{82}$$

$$\lim_{n\to\infty} C(P^{2,n};\kappa^n) = C(P^2;\kappa^*) = \frac{1}{2} H(p_{a_L}^1) + \frac{1}{2} H(p_{a_R}^1). \tag{83}$$

Moreover, in each decision problem $j \in \{1,2\}$, it holds that

$$P^j \in \operatorname*{arg\,max}_{Q\in\Delta(A_j)^\Theta} \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A_j} Q_\theta(a) a(\theta) - C(Q;\kappa^*). \tag{84}$$

Proof. For each $j \in \{1,2\}$, define the map $C^j : \Delta(A_j)^\Theta \times \mathbb{R}_+^I \to \mathbb{R}_+$ as

$$C^j(Q;\kappa) = \sum_{a\in A^j} Q_\pi(a) \left[\sum_{i\in I} \kappa_i\, \overline{q}_a(i)\, D_{\mathrm{KL}}\left(q_{a,i} \| \pi_i \right) \right],$$

where $\{q_a\}_{a\in A^j} \subseteq \Delta(\Theta)$ are the posteriors induced by Q. In words, $C^j(\cdot;\kappa)$ is the neighborhood-based cost with coefficients κ, restricted to the subdomain of stochastic choice rules on A_j.

Take any $j \in \{1,2\}$. We assert that C^j is jointly continuous. To this end, take any convergent sequence (Q^k,κ^k) in $\Delta(A_j)\times\mathbb{R}^I_+$ with limit point (Q,κ). By the triangle inequality,

$$\left|C^j(Q^k,\kappa^k) - C^j(Q,\kappa)\right| \leq \left|C^j(Q^k,\kappa^k) - C^j(Q^k,\kappa)\right| + \left|C^j(Q^k,\kappa) - C^j(Q,\kappa)\right|. \tag{85}$$

We consider each term on the RHS of (85) in turn. For the first term, we have

$$\begin{aligned}\left|C^j(Q^k,\kappa^k) - C^j(Q^k,\kappa)\right| &\leq \sum_{a\in A_j} Q^k_\pi(a) \sum_{i\in N} \left|\kappa^k_i - \kappa_i\right| \overline{q}^k_a(i)\, D_{\mathrm{KL}}\left(q^k_{a,i}\|\pi_i\right)\\ &\leq \sum_{i\in N}\left|\kappa^k_i - \kappa_i\right| \times \sup_{p_i\in\Delta(B_i)} D_{\mathrm{KL}}\left(p_i\|\pi_i\right)\\ &\to 0 \quad \text{as } k\to+\infty,\end{aligned}$$

where the first line is by the triangle inequality and the final line uses the fact that $D_{\mathrm{KL}}(\cdot\|\pi_i)$ is bounded on $\Delta(B_i)$. For the second term, note that $Q^k \to Q$ implies $Q^k_\pi \to Q_\pi$ and $(q^k_a)_{a\in A_j} \to (q_a)_{a\in A_j}$ (being that A_j is finite). It follows that

$$\begin{aligned}\lim_{k\to\infty} C^j(Q^k;\kappa) &= \sum_{a\in A^j} \lim_{k\to\infty} Q^k_\pi(a) \sum_{i\in I} \kappa_i \lim_{k\to\infty}\left[\overline{q}^k_a(i)\, D_{\mathrm{KL}}\left(q^k_{a,i}\|\pi_i\right)\right]\\ &= \sum_{a\in A^j} Q_\pi(a) \sum_{i\in I} \kappa_i\, \overline{q}_a(i)\, D_{\mathrm{KL}}\left(q_{a,i}\|\pi_i\right) = C^j(Q,\kappa).\end{aligned}$$

Since both terms on the RHS of (85) converge to 0 as $k\to\infty$, we obtain $\lim_{k\to\infty} C^j(Q^k;\kappa^k) = C^j(Q;\kappa)$. We conclude that C^j is jointly continuous, as asserted.

Since Claim 24 establishes that $\kappa^* \in \mathbb{R}^I_+$, continuity of the C^j directly implies (82) and (83). Moreover, note that continuity of the C^j also implies, via Berge's Theorem of the Maximum, that the correspondences $\mathcal{Q}^j : \mathbb{R}^I_+ \rightrightarrows \Delta(A_j)^\Theta$ defined as

$$\mathcal{Q}^j(\kappa) = \operatorname*{arg\,max}_{Q\in\Delta(A_j)^\Theta} \sum_{\theta\in\Theta}\pi(\theta)\sum_{a\in A^j} Q_\theta(a)a(\theta) - C^j(Q;\kappa)$$

are upper hemi-continuous. Since each $P^{j,n} \in \mathcal{Q}^j(\kappa^n)$ by hypothesis, this implies that $P^j \in \mathcal{Q}^j(\kappa^*)$ for each $j\in\{1,2\}$. This establishes (84), completing the proof of the claim. □

Claim 26. *$\kappa^*_i = 0$ for every $i \in I$.*

Proof. Suppose, towards a contradiction, that there exists $k\in I$ with $\kappa^*_k > 0$. Since $|B_k| \geq 2$ by the nonredundancy hypothesis, there is some $E \in \{U,D,L,R\}$ such that $B_k \cap E \neq \emptyset$ and $B_k\backslash E \neq \emptyset$. We suppose here that $E = U$; the other cases are specular and hence omitted.

Consider decision problem 1. Since $\mathrm{supp}(p^1_{a_U}) = U$ and $B_k\backslash U \neq \emptyset$, it follows that $\mathrm{supp}(p_{a_U,k}) \subsetneq B_k = \mathrm{supp}(\pi_k)$. We show that this yields a contradiction to (84) in Claim 25.

To this end, for each $\epsilon \in (0,1)$, define $Q^\epsilon \in \Delta(A_1)^\Theta$ as $Q^\epsilon_\theta(\cdot) = \epsilon/2 + (1-\epsilon)P^1_\theta(\cdot)$ for all $\theta \in \Theta$, so that $Q^\epsilon_\pi(\cdot) = 1/2$ and the associated posteriors are $q^\epsilon_a = \epsilon\pi + (1-\epsilon)p^1_a \in \Delta(\Theta)$ for each $a \in A_1$. For the limit coefficients κ^*, the value of decision problem 1 under Q^ϵ is

$$V(\epsilon) := \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A^1} Q^\epsilon_\theta(a)a(\theta) - C(Q^\epsilon) = 1 - \epsilon/2 - C(Q^\epsilon;\kappa^*).$$

Note that $V(0) = 1 - C(P^1)$ is the value under P^1. Moreover, $V(\epsilon) > V(0)$ if and only if

$$\frac{2}{\epsilon} \cdot [C(P;\kappa^*) - C(Q^\epsilon;\kappa^*)] > 1. \tag{86}$$

Thus, to obtain a contradiction to (84), it suffices to show (86) for some $\epsilon > 0$. Note that

$$\begin{aligned} C(P^1;\kappa^*) - C(Q^\epsilon;\kappa^*) &= \frac{1}{2}\left(H(p^1_{a_U}) - H(q^\epsilon_{a_U})\right) + \frac{1}{2}\left(H(p^1_{a_D}) - H(q^\epsilon_{a_D})\right) \\ &\geq \frac{1}{2}\nabla H(q^\epsilon_{a_U}) \cdot \left(p^1_{a_U} - q^\epsilon_{a_U}\right) + \frac{1}{2}\nabla H(q^\epsilon_{a_D}) \cdot \left(p^1_{a_D} - q^\epsilon_{a_D}\right) \\ &= \frac{\epsilon}{2}\left(\nabla H(q^\epsilon_{a_U}) \cdot \left(p^1_{a_U} - \pi\right) + \nabla H(q^\epsilon_{a_D}) \cdot \left(p^1_{a_D} - \pi\right)\right), \end{aligned}$$

where the inequality holds because H is convex and differentiable at full-support beliefs. Therefore, to show that (86) holds for some $\epsilon > 0$, it suffices to show that

$$\nabla H(q^\epsilon_{a_U}) \cdot \left(p^1_{a_U} - \pi\right) + \nabla H(q^\epsilon_{a_D}) \cdot \left(p^1_{a_D} - \pi\right) \to +\infty \quad \text{as } \epsilon \to 0. \tag{87}$$

We establish (87) in what follows.

First, note that for any $\theta \in \Theta$ and full-support $p \in \Delta(\Theta)$, it holds that

$$\begin{aligned} \nabla_\theta H(p) &= \frac{\partial}{\partial p(\theta)} \sum_{i\in I} \kappa_i \overline{p}(i) D_{\mathrm{KL}}(p_i\|\pi_i) \\ &= \sum_{i\in I\,:\,\theta\in B_i} \kappa_i \left(\frac{\partial \overline{p}(i)}{\partial p(\theta)} D_{\mathrm{KL}}(p_i\|\pi_i) + \overline{p}(i) \sum_{\tau\in B_i} \nabla_\tau D_{\mathrm{KL}}(p_i\|\pi_i) \frac{\partial p_i(\tau)}{\partial p(\theta)} \right) \\ &= \sum_{i\in I\,:\,\theta\in B_i} \kappa_i (D_{\mathrm{KL}}(p_i\|\pi_i) + \nabla_\theta D_{\mathrm{KL}}(p_i\|\pi_i) - \nabla D_{\mathrm{KL}}(p_i\|\pi_i) \cdot p_i) \\ &= \sum_{i\in I\,:\,\theta\in B_i} \kappa_i \log\left(\frac{p_i(\theta)}{\pi_i(\theta)}\right), \end{aligned}$$

where the second line is by the chain rule and the third and fourth lines are by direct calculation. Now, take any $a \in A_1$. The above display implies that

$$\begin{aligned} \nabla H(q^\epsilon_a) \cdot \left(p^1_a - \pi\right) &= \sum_{\theta\in\Theta} \nabla_\theta H(q^\epsilon_a) \left(p^1_a(\theta) - \pi(\theta)\right) \\ &= \sum_{i\in I} \kappa_i \sum_{\theta\in B_i} \log\left(\frac{q^\epsilon_{a,i}(\theta)}{\pi_i(\theta)}\right) \left(p^1_a(\theta) - \pi(\theta)\right), \end{aligned} \tag{88}$$

where the second line is by the preceding display and interchanging the order of summation. Moreover, note that

$$q^{\epsilon}_{a,i}(\theta) = \frac{\epsilon\pi(\theta) + (1-\epsilon)p^1_a(\theta)}{\epsilon\overline{\pi}(i) + (1-\epsilon)\overline{p}^1_a(i)} \quad \text{for all } i \in I,\, \theta \in \Theta. \tag{89}$$

We assert that each term in the sum in (88) is non-negative. Fix any $i \in I$ and $\theta \in B_i$. We prove the assertion for $a = a_U$; the case where $a = a_D$ is specular. Note that $\overline{p}^1_{a_U}(i) = \frac{1}{2}|B_i \cap U|$.

Case 1: $p^1_{a_U}(\theta) > \pi(\theta)$. This implies $p^1_{a_U}(\theta) = 1/2$. Plugging this and $\overline{\pi}(i) = 1/|B_i|$ into (89), a short calculation reveals that $q^{\epsilon}_{a,i}(\theta) \geq \pi_i(\theta)$ if and only if $|B_i| \geq |B_i \cap U|$. Since the latter inequality trivially holds, we conclude that $\log\left(q^{\epsilon}_{a_U,i}(\theta)/\pi_i(\theta)\right)\left(p^1_{a_U}(\theta) - \pi(\theta)\right) \geq 0$.

Case 2: $p^1_{a_U}(\theta) < \pi(\theta)$. This implies $p^1_{a_U}(\theta) = 0$. If $B_i \cap U = \emptyset$, then (89) implies $q^{\epsilon}_{a_U,i}(\theta) = \pi_i(\theta)$. If $B_i \cap U \neq \emptyset$, then (89) implies $q^{\epsilon}_{a_U,i}(\theta) < \pi_i(\theta)$. In either case, we obtain $\log\left(q^{\epsilon}_{a_U,i}(\theta)/\pi_i(\theta)\right) \leq 0$. We conclude that $\log\left(q^{\epsilon}_{a_U,i}(\theta)/\pi_i(\theta)\right)\left(p^1_{a_U}(\theta) - \pi(\theta)\right) \geq 0$.

This proves the assertion. It follows that, for every $\epsilon \in (0,1)$,

$$\begin{aligned}
&\nabla H(q^{\epsilon}_{a_U}) \cdot \left(p^1_{a_U} - \pi\right) + \nabla H(q^{\epsilon}_{a_D}) \cdot \left(p^1_{a_D} - \pi\right) \\
&\qquad \geq \kappa_k \sum_{\theta \in B_k} \log\left(\frac{q^{\epsilon}_{a_U,k}(\theta)}{\pi_k(\theta)}\right)\left(p^1_{a_U}(\theta) - \pi(\theta)\right) \\
&\qquad \geq \kappa_k \sum_{\theta \in B_k \setminus U} \log\left(\frac{q^{\epsilon}_{a_U,k}(\theta)}{\pi_k(\theta)}\right)\left(p^1_{a_U}(\theta) - \pi(\theta)\right),
\end{aligned} \tag{90}$$

where $k \in I$ is the index that, by supposition, satisfies $\kappa_k > 0$, $B_k \cap U \neq \emptyset$, and $B_k \setminus U \neq \emptyset$. Plugging the definition of $p^1_{a_U}$ and (89) into the final expression, we obtain

$$\begin{aligned}
\kappa_k \sum_{\theta \in B_k \setminus U} \log\left(\frac{q^{\epsilon}_{a_U,k}(\theta)}{\pi_k(\theta)}\right)\left(p^1_{a_U}(\theta) - \pi(\theta)\right) &= \frac{\kappa_k}{4} \sum_{\theta \in B_k \setminus U} \log\left(\frac{\epsilon\,|B_k| + (1-\epsilon)\,|B_k \cap U|}{\epsilon\,|B_k|}\right) \\
&\to +\infty \quad \text{as } \epsilon \to 0,
\end{aligned}$$

where the limit is infinite because $\kappa_k > 0$, $B_k \setminus U \neq \emptyset$, and $|B_k \cap U| \geq 1$. Plugging this into (90) then establishes (87), and hence the desired contradiction. We conclude that $\kappa_k = 0$. □

Claim 26 delivers the first conclusion of the proposition. It remains to show that $\kappa^* = \mathbf{0}$ implies that $\lim_{n\to\infty} P^{3,n}_{\theta}(a) = \mathbf{1}(a(\theta) = 1)$ for all $\theta \in \Theta$ and $a \in A_3$. Suppose, towards a contradiction, that there exist some $\tau \in \Theta$ and $a \in A_3$ such that $a(\tau) = 1$ and yet $\liminf_{n\to\infty} P^{3,n}_{\tau}(a) < 1$. Then there is a subsequence $(P^{3,n_k})_{k\in\mathbb{N}}$ such that

$$\limsup_{k\to\infty} \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A_3} P^{3,n_k}_{\theta}(a)a(\theta) - C(P^{3,n_k};\kappa^{n_k}) \leq \limsup_{k\to\infty} \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A_3} P^{3,n_k}_{\theta}(a)a(\theta) < 1.$$

Define $P^3 \in \Delta(A_3)^{\Theta}$ as $P^3_{\theta}(a) = \mathbf{1}(a(\theta) = 1)$ for all $\theta \in \Theta$ and $a \in A_3$. Since $\kappa^* = 0$,

$$\lim_{k\to\infty} \sum_{\theta\in\Theta} \pi(\theta) \sum_{a\in A_3} P^3_{\theta}(a)a(\theta) - C(P^3;\kappa^{n_k}) = 1 - C(P^3;\kappa^*) = 1.$$

This implies that, for sufficiently large k, P^{3,n_k} is not optimal in decision problem 3. This delivers the desired contradiction, and thereby completes the proof.

References

S. Acharya and S. L. Wee. Rational inattention in hiring decisions. *American Economic Journal: Macroeconomics*, 12(1):1–40, 2020.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1): 131–142, 1966.

S. Ambuehl, A. Ockenfels, and C. Stewart. Who opts in? Composition effects and disappointment from participation payments. *Review of Economics and Statistics*, 107(1):78–94, 2025.

G.-M. Angeletos and K. A. Sastry. Inattentive economies. *Journal of Political Economy*, 133 (7):2265–2319, 2025.

R. Armenter, M. Muller-Itten, and Z. Stangebye. Geometric methods for finite rational inattention. *Quantitative Economics*, 15:115–144, 2024.

A. Ben-Tal and A. Ben-Israel. A recourse certainty equivalent for decisions under uncertainty. *Annals of Operations Research*, 30(1):1–44, 1991.

A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

A. W. Bloedel and I. Segal. Persuading a rationally inattentive agent. 2021.

A. W. Bloedel and I. Segal. The proper (scoring rule) approach to incentivizing information acquisition. 2025.

A. W. Bloedel and W. Zhong. The cost of optimally acquired information. 2024.

P. Bordalo, N. Gennaioli, and A. Shleifer. Salience. *Annual Review of Economics*, 14(1): 521–544, 2022.

D. Bordoli and R. Iijima. Convex cost of information via statistical divergence. *arXiv preprint arXiv:2509.00229*, 2025.

T. Boyacı and Y. Akçay. Pricing when customers have limited attention. *Management Science*, 64(7):2995–3014, 2018.

D. E. Broadbent. *Perception and communication*. Pergamon Press, 1958.

Z. Y. Brown and J. Jeon. Endogenous information and simplifying insurance choice. *Econometrica*, 92(3):881–911, 2024.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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

A. Caplin and D. Martin. A testable theory of imperfect perception. *The Economic Journal*, 125(582):184–202, 2015.

A. Caplin, M. Dean, and J. Leahy. Rational inattention, optimal consideration sets, and stochastic choice. *Review of Economic Studies*, 86(3):1061–1094, 2019.

A. Caplin, D. Csaba, J. Leahy, and O. Nov. Rational inattention, competitive supply, and psychometrics. *Quarterly Journal of Economics*, 135(3):1681–1724, 2020.

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

S. Cerreia-Vioglio, F. Maccheroni, M. Marinacci, and A. Rustichini. Multinomial logit processes and preference discovery: inside and outside the black box. *Review of Economic Studies*, 90(3):1155–1194, 2023.

C. P. Chambers, C. Liu, and J. Rehbeck. Costly information acquisition. *Journal of Economic Theory*, 186:104979, 2020.

C. P. Chambers, Y. Masatlioglu, P. Natenzon, and C. Raymond. Weighted linear discrete choice. *American Economic Review*, 115(4):1226–1257, 2025.

X. Cheng and Y. Kim. On the monotonicity of information costs. *arXiv preprint arXiv:2404.15158*, 2025.

P. L. Combettes. Perspective functions: Properties, constructions, and examples. *Set-Valued and Variational Analysis*, 26(2):247–264, 2018.

T. Cover and M. Thomas. *Elements of information theory*. John Wiley & Sons, second edition, 2006.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

I. Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2:191–213, 1972.

C. M. Cusumano, F. Fabbri, and F. Pieroth. Competing to commit: Markets with rational inattention. *American Economic Review*, 114(1):285–306, 2024.

H. de Oliveira. Axiomatic foundations for entropic costs of attention. 2019.

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

M. Dean and N. Neligh. Experimental tests of rational inattention. *Journal of Political Economy*, 131(12):3415–3461, 2023.

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

T. Denti. Unrestricted information acquisition. *Theoretical Economics*, 18(3):1101–1140, 2023.

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

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

J. Duchi, K. Khosravi, and F. Ruan. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

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

F. Fabbri. Attention holdup. 2024.

J. P. Flynn and K. A. Sastry. Strategic mistakes. *Journal of Economic Theory*, 212:105704, 2023.

M. Fosgerau, E. Melo, A. De Palma, and M. Shum. Discrete choice and rational inattention: A general equivalence result. *International Economic Review*, 61(4):1569–1589, 2020.

D. Fudenberg, R. Iijima, and T. Strzalecki. Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409, 2015.

D. García-García and R. C. Williamson. Divergences and risks for multiclass experiments. In *Proceedings of the 25th Annual Conference on Learning Theory*. PMLR, 2012.

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

F. Gul, P. Natenzon, and W. Pesendorfer. Random choice as behavioral optimization. *Econometrica*, 82(5):1873–1912, 2014.

L. Györfi and T. Nemetz. f-dissimilarity: A generalization of the affinity of several distributions. *Ann. Inst. Statist. Math*, 30(Part A):105–113, 1978.

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

J. He and P. Natenzon. Moderate utility. *American Economic Review: Insights*, 6(2):176–195, 2024.

B. Hébert and J. La'O. Information acquisition, efficiency, and nonfundamental volatility. *Journal of Political Economy*, 131(10):2666–2723, 2023.

B. Hébert and M. Woodford. Neighborhood-based information costs. *American Economic Review*, 111(10):3225–55, 2021.

J.-B. Hiriart-Urruty. A note on the Legendre-Fenchel transform of convex composite functions. In *Nonsmooth Mechanics and Analysis: Theoretical and Numerical Advances*. Springer, 2006.

J. Hofbauer and W. H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

S. Karlin and Y. Rinott. Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467–498, 1980.

M. W. Khaw, Z. Li, and M. Woodford. Cognitive imprecision and small-stakes risk aversion. *Review of Economic Studies*, 88(4):1979–2013, 2021.

M. S. Kimball. Precautionary saving in the small and in the large. *Econometrica*, 58(1): 589–611, 1990.

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

Y.-H. Lin. Stochastic choice and rational inattention. *Journal of Economic Theory*, 202: 105450, 2022.

E. Lipnowski and D. Ravid. Predicting choice from information costs. *arXiv preprint arXiv:2205.10434*, 2023.

E. Lipnowski, L. Mathevet, and D. Wei. Attention management. *American Economic Review: Insights*, 2(1):17–32, 2020.

F. Maccheroni, M. Marinacci, and A. Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498, 2006.

B. Maćkowiak, F. Matějka, and M. Wiederholt. Rational inattention: A review. *Journal of Economic Literature*, 61(1):226–273, 2023.

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

L.-G. Mattsson and J. W. Weibull. Probabilistic choice and procedurally bounded rationality. *Games and Economic Behavior*, 41(1):61–78, 2002.

J. Mensch. Rational inattention and the monotone likelihood ratio property. *Journal of Economic Theory*, 196:105284, 2021.

J. Mensch. Screening inattentive buyers. *American Economic Review*, 112(6):1949–1984, 2022.

J. Mensch and D. Ravid. Monopoly, product quality, and flexible learning. *arXiv preprint arXiv:2202.09985*, 2022.

P. R. Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12:380–391, 1981.

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

S. Morris and M. Yang. Coordination and continuous stochastic choice. *Review of Economic Studies*, 89(5):2687–2722, 2022.

X. Mu, L. Pomatto, P. Strack, and O. Tamuz. From Blackwell dominance in large samples to Rényi divergences and back again. *Econometrica*, 89(1):475–506, 2021.

M. Muller-Itten, R. Armenter, and Z. Stangebye. Rational inattention via ignorance equivalence. 2024.

H. Pashler. *The psychology of attention*. MIT Press, 1998.

L. Pomatto, P. Strack, and O. Tamuz. The cost of information: The case of constant marginal costs. *American Economic Review*, 113(5):1360–1393, 2023.

D. Ravid. Ultimatum bargaining with rational inattention. *American Economic Review*, 110 (9):2948–2963, 2020.

R. T. Rockafellar. *Convex analysis*. Princeton University Press, 1970.

C. Shubatt and J. Yang. Tradeoffs and comparison complexity. *arXiv preprint arXiv:2401.17578*, 2024.

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

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

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

T. Strzalecki. *Stochastic choice theory*. Cambridge University Press, 2025.

J. Thereze. Screening costly information. 2025.

A. Tversky and J. E. Russo. Substitutability and similarity in binary choices. *Journal of Mathematical Psychology*, 6(1):1–12, 1969.

D. Walker-Jones. Rational inattention with multiple attributes. *Journal of Economic Theory*, page 105688, 2023.

C.-H. Wen and F. S. Koppelman. The generalized nested logit model. *Transportation Research Part B: Methodological*, 35(7):627–641, 2001.

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

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

N. Yoder. Designing incentives for heterogeneous researchers. *Journal of Political Economy*, 130(8):2018–2054, 2022.