Heterogeneous Costs and Demand*

David Walker-Jones†

July 31, 2026

Abstract

Studying how demand changes with price is a standard comparative static and tool for analysis in economics, and yet an understanding of what demand looks like in a revealed preference environment with standard costly learning models has been lacking. This paper provides necessary and sufficient conditions for demand being rationalized by several natural costly learning models. It also shows that heterogeneous costs for learning can produce demand patterns that cannot be rationalized by a representative agent version of the standard Posterior Separable model, which is relevant when observations are aggregated across different individuals or points in time.

Keywords: Posterior Separable, Rational Inattention, Heterogeneous Information Costs

1 Introduction

In our modern world decision makers (DMs) are increasingly inundated with information. Incorporating this information into choices requires the costly investment

*I would like to thank Andrew Caplin, Mark Dean, Rahul Deb, Tommaso Denti, Henrique de Oliveira, Yoram Halevy, Johannes Hoelzemann, Daniel Martin, and Colin Stewart, for their helpful advice and support.

†University of Surrey, d.walker-jones@surrey.ac.uk, ORCID: 0000-0001-6229-9958.

of time and effort. As a result, DMs often do not acquire all the information that may be relevant to their decision. This scarcity of attention that results from an over abundance of information motivates the study of rational inattention (Sims, 2003), a model in which DMs acquire information "as if" they solve an optimization problem, weighing the benefits of better choices against the costs of acquiring more detailed information. The work of Sims (2003) has inspired a rapidly growing literature that demonstrates the significance of costly learning in a number of economics' sub-fields (Maćkowiak, Matějka, & Wiederholt, 2023).

Understanding and accurately modelling the cost of information is important for answering standard economic questions because costly learning changes what can be expected when conducting counterfactual analysis. If the information a DM acquires changes when the price of a good changes, knowing this is crucial for accurately predicting the probability of the good being selected at a new price.

Downward sloping demand, for instance, is perhaps as foundational of a prediction as there is in economics, dating back to King and Davenant in the 17th century (Heberton Jr, 1967), and under standard assumptions it produces an elementary tool for welfare analysis as it identifies the distribution of values for the good. Costly learning challenges this intuition, however, as it can generate locally upward sloping demand for some versions of a good.[1] To see this, suppose a DM wants to purchase a good if it is of high quality and does not want to purchase the good if it is of low quality. But, if the costly learning of the DM results in a symmetric "accuracy" q, meaning the DM only purchases the good when a good signal is realized, which happens with probability q if the good is of high quality and probability $1-q$ otherwise, then if their accuracy changes with price there is locally upward sloping demand for one of the two types of good. If the accuracy q goes up with price then demand is upward sloping locally for a high quality good, while if accuracy goes down with price then demand is upward sloping locally for a low quality good.

[1] See Example 1 in the work of Walker-Jones (2026) and the "q accuracy" model characterized in that paper.

While consequential, costly learning is difficult to study because it depends on inputs that are difficult to measure such as time, effort, and cognitive resources, and it is thus difficult to quantify the relationship between these inputs and outputs such as the quality or accuracy of decisions. As a result, it can be hard to know what mathematical structure should be imposed onto the costs of information, and different cost structures can lead to starkly different behavioral predictions. While the impact of a changing price on demand is one of the most basic comparative statics in economics, the implications of standard costly learning models are still not known in revealed preference settings where utility is unobserved.

This paper thus studies demand in costly learning environments and provides necessary and sufficient conditions for the "state dependent stochastic choice data" (Caplin & Dean, 2015; Caplin & Martin, 2015) produced by several standard classes of cost functions in settings where the utility function is not known. This paper shows that a basic assumption, that more information is more costly, creates starkly different predictions for demand depending on the domain of the assumption as state dependent signals and posterior beliefs are both natural domains for the assumption.

This paper is also the first to study the testable implications of a "random cost" version of the standard "Posterior Separable" model (Caplin & Dean, 2013),[2] and allows for heterogeneous beliefs and costs for information, which complements standard "random utility" models.[3] The heterogeneity of DMs is empirically relevant because studying costly learning typically necessitates the aggregation of data. Incomplete information acquisition usually results in behavior that is stochastic; what

[2]In the Posterior Separable model the cost of reaching different posterior beliefs is assumed to be additively separable. This assumption is conducive to analysis, has been provided with microfoundation from dynamic learning models (e.g., Morris & Strack, 2019; Bloedel & Zhong, 2021; Hébert & Woodford, 2023), and axiomatic models of costly learning often produce Posterior Separable models (e.g., Mensch, 2018; Pomatto, Strack, & Tamuz, 2023).

[3]Notably, Brown and Jeon (2024) study the subset of Posterior Separable cost functions that measure reduction in Shannon Entropy and allow the cost parameter to vary with the observable characteristics of the agent, while this paper studies the broader class of Posterior Separable cost functions and allows the non-parametric costs to vary even when agent observables do not. This modelling choice mirrors the way the utility for an option can vary in a random utility model even when agent observables do not.

a researcher wishes to study is the probabilities of different options being selected in different states of the world. Collecting this type of data requires aggregation of some kind across decision problems, yet the impact of aggregation has not received much attention in the theory literature. This paper shows that heterogeneity is important to the theory because aggregating the behavior of DMs that each pay for information according to the standard Posterior Separable model can rationalize changes in demand that are inconsistent with a representative DM version of the Posterior Separable model. This result may be surprising to some readers whose intuition is based on random utility models, as in such models, if the value distribution of DMs is not restricted to be in a parametric class, aggregating over heterogeneous DMs always produces behavior that is consistent with some representative DM. As a result, the heterogeneity of DMs has important implications for the behavior that can be rationalized by standard costly learning models and the ways in which costly information acquisition should be modelled.

Denti (2022), for instance, provides an important characterization of the behavior that can be rationalized by a representative agent with Posterior Separable learning costs when utility is known, but further shows that subjects in an experiment conducted by Dean and Neligh (2023) frequently violate the standard and quite flexible Posterior Separable model. One conclusion that could be drawn from this exercise is that Posterior Separable costs of information are not flexible enough to capture common behavioral patterns, and generalizations of Posterior Separable models are required. When the specifics of the experiment conducted by Dean and Neligh (2023) are considered, however, the aggregation of data seems to provide a natural explanation for the shortfalls of the Posterior Separable model in their context.[4] If subjects were initially willing to expend effort to acquire information and pick options that are risky at their prior, but eventually began to fatigue, became less willing to exert effort to learn, and reverted to selecting an option that is best at their prior,

[4]See Example 1 in Section 2.3.

this would generate violations of the representative agent version of the Posterior Separable model.

If there are more than two potential states of the world then a Posterior Separable model is capable of rationalizing an increase in demand in a subset of the states of the world when price increases.[5] If this seems problematic in a certain environment, then distributions over subsets of Posterior Separable cost functions, not generalizations of Posterior Separable cost functions, may be a desirable tool for studying the commonly observed behavioral patterns. For another example of how significant the heterogeneity of learning costs is for modelling subject behavior in the context of an experiment see the work of Walker-Jones (2026).

2 Model

Consider a DM that is deciding if they should purchase a good Y, or not, which is denoted $\neg$Y. The DM's payoff from Y, which is $u(\omega) - p \in \mathbb{R}$, is determined by the **state** $\omega \in \Omega$ as $u : \Omega \to \mathbb{R}$, and the **price** of selecting Y denoted p, which the DM must pay to obtain Y. The payoff from $\neg$Y is normalized to be 0. Assume that there are a finite number of states in Ω and that in particular there are states $\underline{\omega}, \overline{\omega} \in \Omega$ with $u(\underline{\omega}) < u(\overline{\omega})$. Let $\sigma(\Omega)$ denote the set of events generated by Ω, which is the algebra generated by Ω. The DM knows the distribution $\mu \in \Delta(\Omega)$ over states, which is referred to as their **prior belief**.[6]

Given prior $\mu \in \Delta(\Omega)$, the **behavior** at a finite set of prices $\mathcal{P} \subset \mathbb{R}$, is a mapping $s : \mathcal{P} \to S \equiv [0, 1]^{|\Omega|}$. Namely, at each price $p \in \mathcal{P}$ the researcher observes the **information outcome** $s(p) \equiv (s(\omega_1, p), s(\omega_2, p), \ldots, s(\omega_N, p)) \equiv (\Pr(\text{Y}|\omega_1, p), \ldots, \Pr(\text{Y}|\omega_N, p))$, that describes the probability of the DM selecting Y at each $\omega \in \Omega$, denoted $\Pr(\text{Y}|\omega, p)$. If there are exactly two states of the world, as a minor abuse of notation, $s(p) \equiv (\underline{s}(p), \overline{s}(p)) \equiv (\Pr(\text{Y}|\underline{\omega}, p), \Pr(\text{Y}|\overline{\omega}, p))$. A

[5] See Example 2 in Section 2.3.

[6] The belief μ assigns a strictly positive probability to each state unless stated otherwise.

generic member $s \in S$ describes a probability of the DM selecting Y at each $\omega \in \Omega$: $s = (s(\omega_1, \cdot), s(\omega_2, \cdot), \ldots, s(\omega_N, \cdot))$. If there are exactly two states of the world a generic information outcome is then denoted $s = (\underline{s}, \overline{s}) \in S$.

The DM is said to **learn** at a price p if there are states ω_i and ω_j in Ω such that $\Pr(\mathrm{Y}|\omega_i, p) \neq \Pr(\mathrm{Y}|\omega_j, p)$. In other words, the DM is learning if their probability of selecting Y changes with the realization of ω, because this is indicative of the DM at least partially differentiating between two states ω_i and ω_j. Let $S^L \subset S$ denote the set of information outcomes where learning occurs: $S^L = \{s \in S | \exists \omega_i, \omega_j \in \Omega \text{ with } s(\omega_i, \cdot) \neq s(\omega_j, \cdot)\}$.

This paper's models focus on information outcomes because this is the observable behavior that is measured in a typical dataset on "state dependent stochastic choice data" (Caplin & Dean, 2015; Caplin & Martin, 2015). Assuming that the prior μ is known by the DM and the researcher, and that the state dependent demand is observed by the researcher, are admittedly a demanding set of assumptions, though not out of line with other recent papers in the literature (e.g., Matějka & McKay, 2015; Denti, 2022). Further, this paper does relax the common assumption that the utility function is observed, and difficulties with measuring state dependent stochastic choice data and the prior belief of a DM actually help to motivate the heterogeneous models introduced in Section 2.3 and the Appendix.

There could also be some contexts in which it is problematic to assume that the prior μ does not change with price, because the price may contain information about the realized value of Y. The simplifying assumption employed in this paper, that μ does not change with p, is meant to mirror the assumption in classic demand analysis that the demand curve generated by a distribution over values does not change with price. There are contexts where such a comparative static is natural, including ones in which changes in price are generated on the "supply side" due to changes in, for instance, costs of production of tariffs, and do not provide information about the quality of a good.

2.1 Costly Learning

Given a prior belief μ, a cost function (for information) $C_\mu : S \to \mathbb{R}_+$ describes the minimal cost of achieving each potential information outcome.[7] The cost function does not depend on price, which is to say this paper's model assumes that a price change does not impact the cost to the DM of achieving a given information outcome. The realized cost $C_\mu(s(p))$ might change when p changes, but the function C_μ does not change if p changes. For the sake of robustness I also allow for a fixed **cost of accessing information**, $\overline{C_\mu} \in \mathbb{R}_+$, though it ends up being rather inconsequential for the results. This fixed cost is paid iff some learning occurs, namely iff $\chi(s \in S^L) = 1$, where χ is the indicator function.

If the DM can achieve an outcome without doing any learning, then it is natural to think the outcome should be costless. So, the DM should be able to select Y, not select Y, or randomize over these options, all without incurring any learning costs. **Assumption 1:** $\forall x \in [0,\, 1]$, $C_\mu(x,\, \ldots,\, x) = 0$. Further, it is natural to assume that the DM can randomize over information outcomes without cost. So, if an information outcome is a convex combination of other information outcomes, then the cost of the information outcome should be weakly lower than the appropriate convex combination of the costs of the other information outcomes, which implies that more information in a Blackwell (1951, 1953) sense is weakly more costly.[8] **Assumption 2:** C_μ is a weakly convex function.

Assumption 1 and Assumption 2 represent the minimal structure that is assumed of all cost functions for information in this paper. Given the belief of the DM, μ, the **cost function** for information outcomes is defined as a mapping from S onto the positive reals, $C_\mu : S \to \mathbb{R}_+$, which satisfies Assumption 1 and Assumption 2.

[7] The cost function has a subscript μ because I do not impose that it is constant in μ, as would be the case if I instead, for instance, assumed a "uniformly" Posterior Separable cost function (Caplin, Dean, & Leahy, 2022).

[8] Given Assumption 1, Lemma 1 from the work of Cheng and Kim (2025) implies that Assumption 2 imposes that more information in a Blackwell (1951, 1953) sense is weakly more costly.

Definition: Given a prior belief μ and a set of prices $\mathcal{P}$, the behavior is **rationalized by a costly learning model** if there are payoffs for Y, $u : \Omega \to \mathbb{R}$, a cost function for information outcomes C_μ, and a cost of accessing information $\overline{C_\mu} \geq 0$, such that $\forall\, p \in \mathcal{P}$:

$$s(p) \in \arg\max_{s \in S} \left(\sum_{\omega \in \Omega} \mu(\omega) s(\omega, \cdot)(u(\omega) - p) - C_\mu(s) - \overline{C_\mu} \chi(s \in S^L) \right).$$

Given p and μ, the expected **payoff** the DM receives when they choose an information outcome $s \in S$ is defined by:

$$\sum_{\omega \in \Omega} \mu(\omega) s(\omega, p)(u(\omega) - p) - C_\mu(s) - \overline{C_\mu} \chi(s \in S^L).$$

When p increases, the payoff from an information outcome s decreases in proportion to the **unconditional probability** of selecting Y that the information outcome results in, which is:

$$\Pr(\mathrm{Y}|p) \equiv \sum_{\omega \in \Omega} \mu(\omega) s(\omega, p).$$

This suggests the DM should choose an information outcome with a lower unconditional probability of selecting Y when price increases, as is shown by Theorem 1. Before introducing the theorem, I introduce some helpful notation. For each event $B \in \sigma(\Omega)$, define its complement as: $B^c \equiv \{\omega \in \Omega | \omega \notin B\}$. Then, for each event $B \in \sigma(\Omega)$, define the probability of Y being selected given the price and the realization of the event as:

$$\Pr(\mathrm{Y}|B, p) \equiv \sum_{\omega \in \Omega} \mu(\omega|B) s(\omega, p).$$

Theorem 1. Given a prior belief μ and a finite and non-empty set of prices $\mathcal{P}$, the behavior is rationalized by a costly learning model if it satisfies the following three properties, and only if it satisfies properties **(ii)** and **(iii)**. Further, these conditions are all necessary for the behavior being rationalized by a costly learning model if

there are only two states of the world.

(i) There is an event in which Y is more likely to be selected if some learning is done: $\exists B \in \sigma(\Omega)\backslash\{\emptyset, \Omega\}$ such that $\forall\, p \in \mathcal{P}$ with $s(p) \in S^L$: $\Pr(\mathrm{Y}|B,\, p) > \Pr(\mathrm{Y}|B^c,\, p)$.

(ii) Y is weakly less likely to be selected when its price increases: $\Pr(\mathrm{Y}|p)$ is weakly decreasing in p.

(iii) If there is a price at which the DM randomizes over selecting Y and not selecting Y without learning, then there is no learning at any price, and Y is either selected or not selected with probability one at all other prices: if $\exists p \in \mathcal{P}$ such that $\Pr(\mathrm{Y}|\omega_1,\, p) = \cdots = \Pr(\mathrm{Y}|\omega_N,\, p) \in (0,\, 1)$, then $\forall \tilde{p} \in \mathcal{P}\backslash p$: $\Pr(\mathrm{Y}|\tilde{p}) \in \{0,\, 1\}$.

Proof. See the Appendix.

Figure 1: Implications of Theorem 1 ($p_1 < p_2$)

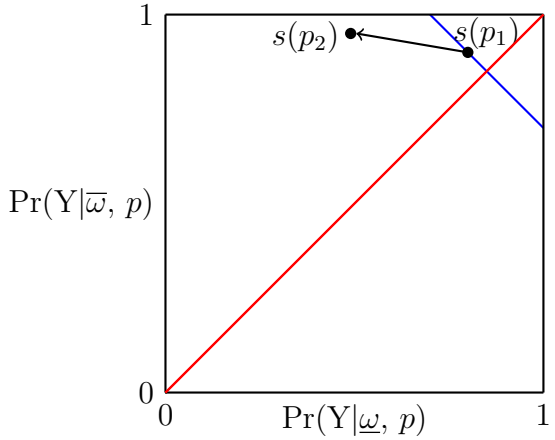


Figure 1 depicts S and the implications of Theorem 1 in the case where there are exactly two states of the world $\underline{\omega}$, $\overline{\omega} \in \Omega$ with $u(\underline{\omega}) < u(\overline{\omega})$. The red line is the information outcomes that result in the same probability of selecting Y given either realization of the state (these are the information outcomes Assumption 1 requires to have a cost of zero). In Theorem 1, part **(i)** requires that each $s(p)$ be on or above the red line. The blue line, defined by $\underline{s}\mu(\underline{\omega}) + \overline{s}\mu(\overline{\omega}) = \underline{s}(p_1)\mu(\underline{\omega}) + \overline{s}(p_1)\mu(\overline{\omega})$, is the collection of points that create the same unconditional probability of selecting Y. Part **(ii)** requires that $s(p_2)$ be on or below the blue line. Part **(iii)** requires that $s(p)$ not be in the interior of S and on the red line unless for all other $\tilde{p} \in \mathcal{P}$ either $s(\tilde{p}) = (0,\, 0)$ or $s(\tilde{p}) = (1,\, 1)$.

If no structure is imposed on the learning costs of the DM beyond Assumption 1 and Assumption 2, then Theorem 1 provides information on behavioral patterns that are consistent with this most general model of costly learning. To see why

(i) in Theorem 1 is not a necessary condition for rationalization, see Example 2. Randomization is significant in property **(iii)** because it indicates that the DM could instead optimally pick Y or not pick Y, which, if they learned at some other price, could be used to create a violation of **(ii)**. This characterization is robust to concerns that a DM might receive a private signal or have a distribution over cost functions, as is shown by Theorem 5 in the Appendix.

Theorem 1 implies that an increase in price can cause an increase in the probability of Y being selected in a particular state of the world as long as the unconditional probability of Y decreases. See Figure 1 for more details.

2.2 The Posterior Separable (PS) Learning Model

The standard in the field of rational inattention is to model the DM paying for information according to a Posterior Separable (PS) cost function (Caplin & Dean, 2013; Caplin et al., 2022). This means the DM selects what to learn by choosing how confident to be in their choices after they finish learning.

To apply a PS model I need a function that measures how "informed" posteriors are. Denote this weakly convex function $c : \Delta(\Omega) \to \mathbb{R}$. When the DM learns they pay for the information based on the change in c it creates. Again, weak convexity of c ensures that more information (in a Blackwell (1951, 1953) sense) is weakly more costly. An example of a commonly used measure of informativeness of posteriors is Shannon Entropy (Shannon, 1948), e.g. in the work of Matějka and McKay (2015). For any information strategy $s \in S^L$, given belief μ, define the two posteriors it is "revealed" to induce, $\mu(\cdot|\text{Y}, s)$, $\mu(\cdot|\neg\text{Y}, s) \in \Delta(\Omega)$, in the following way:

$$\forall \omega \in \Omega, \text{ let } \mu(\omega|\text{Y}, s) \equiv \frac{s(\omega, \cdot)\mu(\omega)}{\sum_{\tilde{\omega}\in\Omega} s(\tilde{\omega}, \cdot)\mu(\tilde{\omega})},$$

$$\text{and let } \mu(\omega|\neg\text{Y}, s) \equiv \frac{(1 - s(\omega, \cdot))\mu(\omega)}{\sum_{\tilde{\omega}\in\Omega} (1 - s(\tilde{\omega}, \cdot))\mu(\tilde{\omega})}.$$

Definition: Given a prior belief μ, the cost function for information outcomes C_μ is a **PS cost function** for information outcomes if there is **measure of informativeness**, a weakly convex $c : \Delta(\Omega) \to \mathbb{R}$, such that $\forall s \in S^L$:

$$C_\mu(s) = \left(\sum_{\omega \in \Omega} \mu(\omega) s(\omega, \cdot) \right) \Big(c(\mu(\cdot|\mathrm{Y}, s)) - c(\mu) \Big)$$

$$+ \left(\sum_{\omega \in \Omega} \mu(\omega)(1 - s(\omega, \cdot)) \right) \Big(c(\mu(\cdot|\neg\mathrm{Y}, s)) - c(\mu) \Big).$$

Definition: Given a prior belief μ and a set of prices $\mathcal{P}$, the behavior is **rationalized by a PS model** if it is rationalized by a costly learning model with a PS cost function for information outcomes C_μ.

Theorem 2. Given a prior belief μ and a finite and non-empty set of prices $\mathcal{P}$, the behavior is rationalized by a PS model if it is rationalized by a costly learning model, there is an event $B \in \sigma(\Omega) \backslash \{\emptyset, \Omega\}$ such that $\forall p \in \mathcal{P}$ with $s(p) \in S^L : \Pr(\mathrm{Y}|B, p) > \Pr(\mathrm{Y}|B^c, p)$, and $\Pr(B|\neg\mathrm{Y}, p)$ and $\Pr(B|\mathrm{Y}, p)$ are both weakly increasing over the set of p where the DM learns. Further, these conditions are necessary for the behavior being rationalized by PS model if there are exactly two states of the world.

Proof. See the Appendix.[9]

If there are exactly two states of the world, Theorem 2 says that if p increases in a PS model the DM is more confident they have made the right decision when selecting Y and less confident they made the right decision when selecting $\neg$Y. Bayes' Rule and Theorem 2 together imply that only behavior with downward sloping demand (state dependent probabilities of Y being selected that decrease with price) can be rationalized by a Posterior Separable model (see Figure 2). The analogous changes in posteriors and downward sloping demand are no longer necessary when there are

[9] Ambuehl (2017) shows necessity of the conditions for a smooth and strictly convex c and a binary state space.

Figure 2: Implications of Theorem 2 (PS model, $p_1 < p_2$)

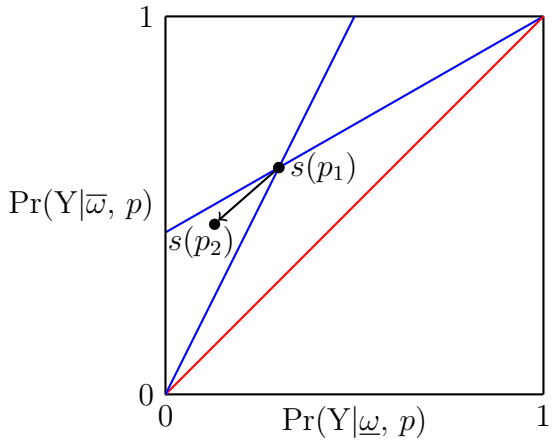


Figure 2 depicts S and the implications of Theorem 2 in the case where there are exactly two states of the world $\underline{\omega}$, $\overline{\omega} \in \Omega$ with $u(\underline{\omega}) < u(\overline{\omega})$. The blue lines are the lines through $s(p_1)$ from (0, 0) and (1, 1). When the DM is learning and price increases, $\Pr(\mathrm{Y}|\underline{\omega},\, p)$ and $\Pr(\mathrm{Y}|\overline{\omega},\, p)$ must both weakly *decrease*, but the proportional reduction in $\Pr(\mathrm{Y}|\underline{\omega},\, p)$ must be weakly larger than the proportional reduction in $\Pr(\mathrm{Y}|\overline{\omega},\, p)$, so $s(p_2)$ must remain weakly above the blue line from (0, 0) through $s(p_1)$. Further, $\Pr(\neg\mathrm{Y}|\underline{\omega},\, p)$ and $\Pr(\neg\mathrm{Y}|\overline{\omega},\, p)$ must both weakly *increase*, but the proportional increase in $\Pr(\neg\mathrm{Y}|\underline{\omega},\, p)$ must be weakly smaller than the proportional increase in $\Pr(\neg\mathrm{Y}|\overline{\omega},\, p)$, so $s(p_2)$ must remain weakly below the blue line from (1, 1) through $s(p_1)$ (assuming $s(p_1) \neq (1,\, 1)$).

more states, as is shown by Example 2.

2.3 An Aggregate PS Model: Fatigue and Novelty

Even if the structure imposed by a particular class of costly learning models is compelling in a setting, datasets used for analysis do not typically focus on an individual DM. Typically datasets aggregate behavior over DMs with potentially heterogeneous (accurate) prior beliefs[10] and costs for information outcomes. Even when a dataset does feature observations from an individual, that individual might have varying private pieces of information (that change their prior belief) and a varying information cost function due to either something akin to fatigue or variation in the choice environment that is not observable to the researcher, e.g. variation in the number of store clerks on the floor when the individual makes the decision. This variation in costs of information has important implications for what types of behavior

[10] As would be generated if DMs get unobserved private information before they start learning.

can be rationalized by the PS model, as demonstrated by the following example.

Example 1. Suppose that $\mu(\underline{\omega}) = \mu(\overline{\omega}) = \frac{1}{2}$, and that in addition to being able to pick $\neg$Y and Y, the DM has the ability to select an option Z. Suppose that $p = \frac{1}{2}u(\overline{\omega}) + \frac{1}{2}u(\underline{\omega}) + \epsilon$ for some small $\epsilon > 0$, and that the payoff from Z is $u(\overline{\omega})$ in state $\underline{\omega}$ and $u(\underline{\omega})$ in state $\overline{\omega}$. Denti (2022) shows with that paper's Proposition 4 that, generically, the Posterior Separable model predicts that there should not be more options with a positive probability of being selected than there are possible states of the world. The experiment of Dean and Neligh (2023), in contrast, indicates that one should expect a subject will select all three of the options in this paragraph's example with positive probabilities. If a non-constant cost function for information is introduced, as is done in this subsection, then subjects selecting all three options with positive probabilities begins to make sense. Suppose that initially a subject is putting effort into learning and they always convince themselves that one of the states is likely enough that they select Y or Z. But, eventually, the repeated choices that are necessary for aggregating data cause the subject to fatigue (which is analogous to an increase in the cost of information), and eventually the subject stops investing effort into learning and simply picks the option that is best at their prior belief; $\neg$Y. Then, all three options would thus be selected with a positive probability.

To better understand the impact of heterogeneous DMs on demand, this subsection characterizes an aggregate version of the PS model from Section 2.2, and shows that aggregating choice data from heterogeneous PS DMs (either different individuals or varying versions of the same individual) can result in behavior that cannot be rationalized with a PS model. This is in contrast with the implication of heterogeneous DMs for the flexible costly learning model from Section 2.1 that, as is shown by Theorem 5 in the Appendix, produces the same characterization of behavior if heterogeneous DMs are introduced.

Informally, behavior is rationalized by an aggregate PS model if there are differ-

ent types of DM, each with their own probability of occurring, prior belief,[11] and PS cost function for information outcomes, such that each type has rationalized behavior, and if behavior is averaged over the different types then the observed behavior is obtained. This is akin to the way DMs are modeled in Random Utility models.

Definition: Given a prior belief μ, the behavior is **rationalized by an aggregate costly learning model** if there are $T \in \mathbb{N}$ types of DM each of which has probability of occurring $\pi_t > 0$, an accurate belief about the state of the world $\mu_t \in \Delta(\Omega)$, and behavior s_t, which is $s_t(\omega, p) \equiv \Pr_t(\mathrm{Y}|\omega, p) \in [0, 1]$ for each state and price, such that each type's behavior is rationalized by a costly learning model[12] with payoffs for Y of $u : \Omega \Rightarrow \mathbb{R}$,[13] and:

(i) The probabilities of the different types sum to one, and their mean belief is the distribution over states observed by the researcher:

$$\sum_{t=1}^{T} \pi_t = 1, \quad \text{and for each } \omega \in \Omega : \sum_{t=1}^{T} \mu_t(\omega)\pi_t = \mu(\omega).$$

(ii) Given any pair of price p and state ω, the mean behavior is the behavior observed by the researcher:

$$\forall\, p \in \mathcal{P},\, \forall\, \omega \in \Omega : \sum_{t=1}^{T} \Pr_t(\mathrm{Y}|\omega, p)\frac{\mu_t(\omega)\pi_t}{\mu(\omega)} = \Pr(\mathrm{Y}|\omega, p).$$

Definition: Given a prior belief μ, the behavior is **rationalized by an aggregate PS model** if it is rationalized by an aggregate costly learning model where each type has behavior that is rationalized by a PS model.

[11] Allowing for heterogeneity of prior beliefs is not required for achieving a characterization of behavior that differs from the representative agent characterization, as is shown by Theorem 4.

[12] Note that different types can have their behavior rationalized by different costly learning models, namely I do not impose that different types have the same cost function for information outcomes.

[13] I assume that u is the same for each type as this is more challenging. Given the result in Theorem 5, allowing utility to differ across types would not substantially change the characterization of behavior rationalized by an aggregate costly learning model, the exception being that property **(iii)** from Theorem 5 could be violated.

Theorem 3. Given a prior belief μ and a finite and non-empty set of prices $\mathcal{P}$, the behavior is rationalized by an aggregate PS model if it is rationalized by a costly learning model, there is event $B \in \sigma(\Omega)\backslash\{\emptyset, \Omega\}$ such that $\forall p \in \mathcal{P}$ with $s(p) \in S^L$: $\Pr(\mathrm{Y}|B,\, p) > \Pr(\mathrm{Y}|B^c,\, p)$, and $\Pr(\mathrm{Y}|\omega,\, p)$ is weakly decreasing in p for each $\omega \in \Omega$. Further, these conditions are necessary for the behavior being rationalized by an aggregate PS model if there are exactly two states of the world.

Proof. See the Appendix.

Figure 3: Implications of Theorem 3 (aggregate PS model, $p_1 < p_2$)

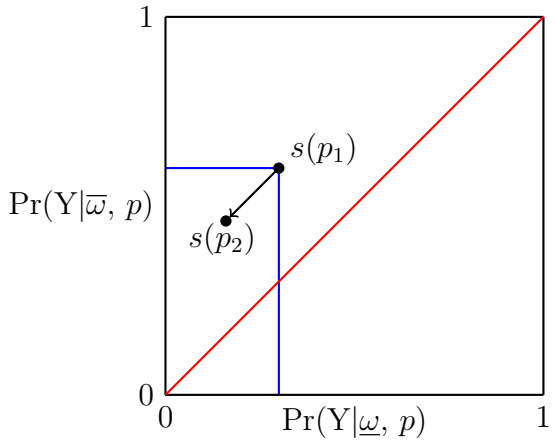


Figure 3 depicts S and the implications of Theorem 3 in the case where there are exactly two states of the world $\underline{\omega}, \overline{\omega} \in \Omega$ with $u(\underline{\omega}) < u(\overline{\omega})$, which requires that $\Pr(\mathrm{Y}|\underline{\omega},\, p)$ and $\Pr(\mathrm{Y}|\overline{\omega},\, p)$ are weakly decreasing in p. In Figure 3, this means that $s(p_2)$ is below and to the left of the old information outcome $s(p_1)$, weakly between the two blue lines but above the red line.

If there are exactly two states of the world and behavior is rationalized by a costly learning model, Theorem 3 says that the behavior can be rationalized with heterogeneous DMs that all behave in line with a PS model if and only if when price increases the probability of selecting Y weakly decreases in both states. Using Bayes' Rule, it is easy to show that such behavior can violate the predictions of Theorem 2, and one can visually see the difference in predicted behavior when PS DMs are aggregated by comparing figures 2 and 3. As a result, aggregating over DMs that behave in line with a PS model can produce behavior that cannot be rationalized by a PS model. Aggregating over PS DMs still produces more specific predictions than

the general model of costly learning introduced in Section 2.1, however, as can be seen by comparing figures 1 and 3. To see why the conditions are no longer necessary when there are more than two states of the world, see Example 2 below.

If the aggregate model did not allow for heterogeneity in beliefs, and only allowed for heterogeneity in the PS costs of information, then the necessary and sufficient conditions in Theorem 3 still characterize the behavior that can be rationalized if there are exactly two states of the world and some learning is done at each price, as is shown by Theorem 4 in the Appendix.

Example 2. Consider a setting with three states, $\Omega = \{\omega_1, \omega_2, \omega_3\}$, a prior belief such that $\mu(\omega_1) = \frac{1}{2}$ and $\mu(\omega_2) = \mu(\omega_3) = \frac{1}{4}$, and payoffs such that $u(\omega_1) = 0$, $u(\omega_2) = 2$, and $u(\omega_3) = 3$. Also assume that there is a PS cost function for information outcomes generated by a measure of informativeness c such that $c(\mu) = 0$ for all $\mu = (\mu(\omega_1), \mu(\omega_2), \mu(\omega_3))$ inside of the convex hull generated by the four distributions $\mu_1 = (0.4, 0.4, 0.2)$, $\mu_2 = (0.525, 0.025, 0.45)$, $\mu_3 = (0.6, 0.1, 0.3)$, $\mu_4 = (0.475, 0.475, 0.05)$, and that outside of this convex hull c is strictly increasing in an arbitrarily steep fashion. Notice that μ_1 and μ_2 both produce the same expected value of 1.4 for selecting Y, which is more than the expected value at the prior of 1.25. Further, the prior is exactly halfway along the cord from μ_1 to μ_3, and the prior is also exactly halfway along the cord from μ_2 to μ_4. Further, $s_1 = (0.4, 0.8, 0.4)$ generates a posterior of μ_1 after Y is selected and μ_3 after $\neg$Y is selected, while $s_2 = (0.525, 0.05, 0.9)$ generates a posterior of μ_2 after Y is selected and μ_4 after $\neg$Y is selected. This all means that for all $p \in [1.1, 1.4]$, both information outcomes s_1 and s_2 are optimal. This is significant as it means that if $\mathcal{P} = \{p_1 = 1.1, p_2 = 1.2\}$, then $s(p_1) = s_1$ and $s(p_2) = s_2$ is rationalized by the constructed PS model, and yet there is no event $B \in \sigma(\Omega)\backslash\{\emptyset, \Omega\}$ such that $\forall p \in \mathcal{P}$ with $s(p) \in S^L : \Pr(\text{Y}|B, p) > \Pr(\text{Y}|B^c, p)$. Further, if $\mathcal{P} = \{p_1 = 1.1, p_2 = 1.2, p_3 = 1.3\}$, then $s(p_1) = s_1$, $s(p_2) = s_2$, and $s(p_3) = s_1$, is rationalized by the constructed PS model, and yet there is no event $B \in \sigma(\Omega)\backslash\{\emptyset, \Omega\}$ such that $\Pr(B|\text{Y}, p)$ is weakly increasing with

p. It should also be noted that while this example uses indifference between two information outcomes over a range of prices to make these two points as simply as is possible, this feature is not required. It is possible to make these same points in a more complicated setting where there is a unique optimal information outcome at each price considered.

3 Literature Review

There is a growing literature that demonstrates the importance of inattention and choice mistakes in standard economic settings, even in ones where it seems acquiring information should be costless. Chetty, Looney, and Kroft (2009), for instance, show that including sales tax in posted prices reduces demand, presumably because consumers were underestimating sales tax when making purchasing decisions, and Taubinsky and Rees-Jones (2018) actually show that in such situations accounting for the heterogeneity of subjects' under-reaction to taxes is crucial for accurately estimating the efficiency loss from taxation. Papers have demonstrated the significance of inattention in a wide variety of fields such as finance (Huberman, 2001), labor search (Acharya & Wee, 2020), trade (Dasgupta & Mondria, 2018), and voting behavior (Shue & Luttmer, 2009). For a recent review of the costly learning literature see the work of Maćkowiak et al. (2023).

Like this paper, Caplin and Dean (2015), Denti (2022), and Lipnowski and Ravid (2023), are also interested in determining if choice data can be rationalized with inattention. Caplin and Dean (2015) provide two testable conditions that, given a belief and a utility function, are satisfied if and only if the data can be rationalized with a costly learning model. So, given behavior s and a utility function u, their result allows one to test whether or not the behavior is rationalized by a costly learning model. But, if the utility function is not known and one wishes to determine the set of information outcomes that are rationalizable for some utility function at a new price,

then each combination of a utility function and an information outcome at the new price must be tested separately. In a general context (Denti, 2022) characterizes the behavior that can rationalized with a Posterior Separable model, but, again, assumes the utility function is know. Lipnowski and Ravid (2023) show that even if the cost of information is well understood, predicting behavior is essentially impossible if the utility function is unknown, but that structure can be imposed across choice problems. The results from my paper also study a setting where the utility function is unknown, and provide conditions that characterize the set of observed behaviors that can be rationalized by a given costly learning model when appropriate payoffs are selected in the special case of a price change, but, further, introduces heterogeneous learning costs.

4 Conclusion

This paper studies the implications of costly learning for demand. While welfare and counterfactual analysis with demand represent some of the most foundational tools in economics, an understanding of what demand looks like in a revealed preference environment with standard models of costly learning has been lacking. This paper provides necessary and sufficient conditions for demand being rationalized by several compelling costly learning models. Costly learning changes the nature of demand substantially and some classes of costs for information can actually create upward sloping demand in some states of the world.

This paper is also the first to study heterogeneous agent versions of several standard costly learning models. It is shown that the introduction of heterogeneous decision makers into a Posterior Separable model can rationalize choice behavior that cannot be rationalized by a representative decision maker version of the Posterior Separable model. However, introducing heterogeneous decision makers into a Posterior Separable model still provides more precise predictions than the most general

representative agent model that is studied.

Further, it is argued that variation in the cost function for information explains violations of the predictions of the Posterior Separable model that have been identified by other papers. This is true because fatigue over the course of an experiment could cause subjects to stop learning and simply pick an option that is best at their prior belief, which can create behavioral patterns that contradict the Posterior Separable model even though the Posterior Separable model has a number of compelling foundations and is the standard in the field of rational inattention. If heterogeneity is not considered then generalizations of Posterior Separable costs are required to rationalize commonly occurring patterns found in other experiments, but then upward sloping demand in some states of the world cannot be ruled out. In contexts where downward sloping demand seems appropriate, heterogeneity over a set of Posterior Separable cost functions that maintain downward sloping demand provides an appealing alternative for explaining observed behavior.

Appendix: Proofs and Additional Results

Many of the proofs consider the case with exactly two states of the world separately. Remember that we have assumed that there are states $\underline{\omega}$, $\overline{\omega} \in \Omega$ with $u(\underline{\omega}) < u(\overline{\omega})$, so when there are exactly two states these are the two states being considered, and to ease exposition in such cases $s(p) \equiv (\underline{s}(p), \overline{s}(p)) \equiv (\Pr(\mathrm{Y}|\underline{\omega}, p), \Pr(\mathrm{Y}|\overline{\omega}, p))$.

Proof of Theorem 1. I begin with necessity of points **(ii)** and **(iii)**. Suppose **(ii)** does not hold: so $p_1 < p_2$ and $\Pr(\mathrm{Y}|p_1) < \Pr(\mathrm{Y}|p_2)$. I have a contradiction since if $s(p_1)$ is optimal at p_1 then it provides a weakly higher payoff at p_1 than $s(p_2)$ does, and, when price increases to p_2 the decrease in payoff from $s(p_1)$ is strictly smaller than the decrease in payoff from $s(p_2)$: $(p_2 - p_1)\Pr(\mathrm{Y}|p_1) < (p_2 - p_1)\Pr(\mathrm{Y}|p_2)$. So, **(ii)** is necessary.

(iii): Again, I proceed with a proof by contradiction. Suppose $\exists p_1 \in \mathcal{P}$

such that: $\Pr(\text{Y}|\omega_1, p_1) = \cdots = \Pr(\text{Y}|\omega_N, p_1) \in (0, 1)$, and $\exists p_2 \in \mathcal{P}\backslash p_1$ such that $\Pr(\text{Y}|p_2) \in (0, 1)$. Notice that optimality of the DM's behavior at p_1 implies $s = (0, \ldots, 0)$ and $s = (1, \ldots, 1)$ are both optimal when price is p_1 since all information outcomes $s = (x, \ldots, x)$ have zero cost for $x \in [0, 1]$. If $p_2 < p_1$, then I have a contradiction because the DM could optimally pick $s = (1, \ldots, 1)$ at p_1, so $\Pr(\text{Y}|p_1) = 1$, which combined with **(ii)** implies the DM is not behaving optimally at p_2. If $p_2 > p_1$, then I have a contradiction because the DM could optimally pick $s = (0, \ldots, 0)$ at p_1, so $\Pr(\text{Y}|p_1) = 0$, which combined with **(ii)** implies the DM is not behaving optimally at p_2.

To see why **(i)** is necessary if there are only two states of the world: if $\exists p \in \mathcal{P}$ with $\Pr(\text{Y}|\underline{\omega}, p) > \Pr(\text{Y}|\overline{\omega}, p)$, then behavior is not rationalized because the DM would have achieved a strictly higher payoff at p by choosing either (depending on the value of $u(\underline{\omega})$ relative to p) $s = (\Pr(\text{Y}|\overline{\omega}, p), \Pr(\text{Y}|\overline{\omega}, p))$ or $s = (1, 1)$).

To show **(i)**, **(ii)**, and **(iii)** are together sufficient, I assume they are all satisfied and order the prices in $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, so $p_1 < p_2 < \cdots < p_n$, and assume $\overline{C_\mu} = 0$. Further, I can assume that there is a price in $\mathcal{P}$ where the DM learns, because if not I can easily rationalize the behavior by making learning costly enough. **(iii)** thus tells me that there is not behavior at a price $p_i \in \mathcal{P}$ with $s(p_i) = (x, \ldots, x)$ and $x \in (0, 1)$. It is without loss to also assume that $s(p_1) = (1, \ldots, 1)$ and $s(p_n) = (0, \ldots, 0)$, since if this is not the case I can add prices to $\mathcal{P}$, and data to the behavior, so that this is the case, and then rationalizing the richer dataset rationalizes the original dataset. There is then a highest $p \in \mathcal{P}$ such that $\Pr(\text{Y}|p) = 1$, denote it p_h, and a lowest $p \in \mathcal{P}$ such that $\Pr(\text{Y}|p) = 0$, denote it p_l. Notice $p_h < p_l$. **(i)** tells me I can pick $B \in \sigma(\Omega)\backslash\{\emptyset, \Omega\}$ so that $\forall\, p \in \mathcal{P}$ with $s(p) \in S^L :\ \Pr(\text{Y}|B, p) > \Pr(\text{Y}|B^c, p)$, with representative states $\overline{\omega} \in B$ and $\underline{\omega} \in B^c$. Pick a mean $m = p_h$, which is a reference point for conducting mean preserving spreads on $u(\underline{\omega})$ and $u(\overline{\omega})$ and is shifted at the end of the argument. Then pick $u(\underline{\omega}) < p_1$ and $u(\overline{\omega}) > p_n$ so that $\mu(B^c)u(\underline{\omega}) + \mu(B)u(\overline{\omega}) = m$, and make it so that $u(\omega) = u(\overline{\omega})$ for all $\omega \in B$ and

$u(\omega) = u(\underline{\omega})$ for all $\omega \in B^c$. For each $s \in S$, define $\underline{s}(s)$ and $\overline{s}(s)$ so that:

$$\underline{s}(s) \equiv \sum_{\omega \in B^c} \mu(\omega|B^c)s(\omega, \cdot) \text{ and } \overline{s}(s) \equiv \sum_{\omega \in B} \mu(\omega|B)s(\omega, \cdot),$$

and for each price $p \in \mathbb{P}$, define $\underline{s}(p) \equiv \underline{s}(s(p))$ and $\overline{s}(p) \equiv \overline{s}(s(p))$.

Remember, for all $x \in [0, 1] : C_\mu(x, \ldots, x) = 0$. I next define a function $C_B : [0, 1] \times [0, 1] \to \mathbb{R}_+$ recursively and simply set $C_\mu(s) = C_B(\underline{s}(s), \overline{s}(s))$. Define:

$$C_B(\underline{s}(p_{l-1}), \overline{s}(p_{l-1})) = \mu(B^c)\Big(\underline{s}(p_{l-1})(u(\underline{\omega}) - p_l)\Big) + \mu(B)\Big(\overline{s}(p_{l-1})(u(\overline{\omega}) - p_l)\Big)$$

$$= \underline{s}(p_{l-1})(m - p_l) + \mu(B)\Big(\overline{s}(p_{l-1}) - \underline{s}(p_{l-1})\Big)(u(\overline{\omega}) - p_l).$$

This means the DM is indifferent between $s(p_{l-1})$ and $s(p_l)$ when price is p_l, and thus strictly prefers $s(p_{l-1})$ to $s(p_l)$ when price is p_{l-1} because when price decreases the payoff from $s(p_{l-1})$ strictly increases since there is a strictly positive probability of the DM selecting Y when they choose $s(p_{l-1})$ (by construction) while the payoff from $s(p_l)$ is zero (by construction). If this $C_B(\underline{s}(p_{l-1}), \overline{s}(p_{l-1}))$ is strictly positive I continue, and if it is not I do a mean preserving spread on $u(\underline{\omega})$ and $u(\overline{\omega})$ so it is. Next, if $l - 2 > h$, I let:

$$C_B(\underline{s}(p_{l-2}), \overline{s}(p_{l-2})) = \underline{s}(p_{l-2})(m - p_{l-1}) + \mu(B)\Big(\overline{s}(p_{l-2}) - \underline{s}(p_{l-2})\Big)(u(\overline{\omega}) - p_{l-1})$$

$$-\underline{s}(p_{l-1})(m - p_{l-1}) - \mu(B)\Big(\overline{s}(p_{l-1}) - \underline{s}(p_{l-1})\Big)(u(\overline{\omega}) - p_{l-1}) + C_B(\underline{s}(p_{l-1}), \overline{s}(p_{l-1})).$$

This means the DM is indifferent between $s(p_{l-2})$ and $s(p_{l-1})$ when price is p_{l-1}, and thus weakly prefers $s(p_{l-2})$ to $s(p_{l-1})$ when price is p_{l-2}. If this $C_B(\underline{s}(p_{l-2}), \overline{s}(p_{l-2}))$ is strictly positive I continue, if it is not I do a mean preserving spread on $u(\underline{\omega})$ and $u(\overline{\omega})$ (updating $C_B(\underline{s}(p_{l-1}), \overline{s}(p_{l-1}))$ accordingly) so $C_B(\underline{s}(p_{l-2}), \overline{s}(p_{l-2}))$ is strictly positive, which works since the value of $-\mu(B)\Big(\overline{s}(p_{l-1}) - \underline{s}(p_{l-1})\Big)(u(\overline{\omega}) - p_{l-1}) + C_B(\underline{s}(p_{l-1}), \overline{s}(p_{l-1}))$ does not change. I continue in this fashion until I have set

$C_B(\underline{s}(p_{h+1}), \overline{s}(p_{h+1}))$. If I keep the mean m the same (equal to p_h), then the DM strictly prefers $s(p_{h+1})$ to $s(p_h)$ when price is p_{h+1}, since they prefer $s(p_{h+1})$ to $s(p_l)$, which they strictly prefer to $s(p_h)$. I now increase $u(\underline{\omega})$ and $u(\overline{\omega})$ by the same amount so that the mean m increases, and all $C_B(\underline{s}(p_i), \overline{s}(p_i)) > 0$ so that the equations I used to define costs are still satisfied, until the DM is indifferent between $s(p_{h+1})$ and $s(p_h)$ when price is p_{h+1}. As a result, the DM strictly prefers $s(p_h)$ to $s(p_{h+1})$ when price is p_h since there is a strictly higher unconditional chance the DM selects Y when they pick $s(p_h)$ compared to $s(p_{h+1})$ by construction. Finally, I can assume $C_B(\underline{s}, \overline{s}) = 0$ for all $\underline{s}$ and $\overline{s}$ with $\underline{s} > \overline{s}$.

At each price p_i I assigned a cost to the information outcome so that the DM is indifferent at p_i between $s(p_i)$, and $s(p_{i-1})$. This implies, out of the set of information outcomes I observe in the behavior, the DM is selecting their strategy optimally at each price, since as price decreases the payoff from a strategy increase by the unconditional probability of choosing Y, and as price decreases the unconditional probability of selecting Y increases.

Now, I assign an arbitrarily high value to $C_B(0, 1)$ (so that $(0, 1)$ is strictly worse than $s(p)$ for all $p \in \mathcal{P}$), if I have not already assigned a value to it, and then define C_B on $s \in S$ such that $\underline{s}(s) < \overline{s}(s)$ to be the the maximal convex function that is equal to the $C_B(\underline{s}(p), \overline{s}(p),)$ I assigned for all $p \in \mathcal{P}$.

Why is this possible? Is it instead possible I assigned a $C_B(\underline{s}(p_i), \overline{s}(p_i))$ that was strictly above the relevant convex combination of other $C_B(\underline{s}(p_j), \overline{s}(p_j))$'s I assigned? This is not possible, as I can show with a quick proof by contradiction. Assume there is a set of prices $\tilde{\mathcal{P}} = \{p_m, \dots p_k\} \subseteq \mathcal{P}$, and a price $p_i \in \mathcal{P}$ such that there are positive weights α_j that sum to one such that $\alpha_m s(p_m) + \cdots + \alpha_k s(p_k) = s(p_i)$, but at the same time $\alpha_m C_B(\underline{s}(p_m), \overline{s}(p_m)) + \cdots + \alpha_k C_B(\underline{s}(p_k), \overline{s}(p_k)) < C_B(\underline{s}(p_i), \overline{s}(p_i))$. Since $\underline{s}(u(\underline{\omega}) - p)\mu(B^c) + \overline{s}(u(\overline{\omega}) - p)\mu(B)$ is linear in the probabilities $\underline{s}$ and $\overline{s}$, this implies when price is p_i the DM strictly prefers randomizing over their strategies from the $\tilde{\mathcal{P}} = \{p_m, \dots p_k\}$ prices compared to selecting $s(p_i)$, but this implies there is a

$p_j \in \tilde{\mathcal{P}} = \{p_m, \ldots p_k\}$ such that the DM strictly prefers $s(p_j)$ to $s(p_i)$ at p_i, which was ruled out by my recursive definition for the costs of information outcomes.

Similarly, at each p_i the information outcome of the DM is optimal given choice from the entire set of S since at each s with $\underline{s} \leq \overline{s}$ the cost is a convex combination of costs from information outcomes used for prices in $\mathcal{P}$ and the corner (0, 1), and if the DM would prefer to switch to a different s at p_i, then they strictly prefer randomizing over the set of information outcomes used to generate the cost of s, and again one of the information outcomes used to generate the cost at s would then have been strictly preferred at p_i, and my recursive definition for the cost of information outcomes (and choice of arbitrarily high $C_B(0, 1)$) rules this out. ■

Proof of Theorem 2. To begin with (first part of the proof), assume there are two states of the world. I assume there is a $p \in \mathcal{P}$ such that some learning is done, otherwise the necessary conditions are trivially established, and sufficiency is easy to establish by making learning costly enough (since belief is fixed).

When the DM pays for information according to a measure of informativeness (weakly convex c), the way for them to maximize their expected payoff is to maximize the weighted average over option specific net gains in utility $V(\text{Y}|p, \cdot)$ and $V(\neg\text{Y}|\cdot)$, defined in the next paragraph. Each option specific net gain in utility takes into account the expected payoff of the relevant option, $\neg$Y or Y, given the DM's posterior, and the cost of the posterior reached by the DM when they choose said option, where the posterior can be described by the resultant probability of $\overline{\omega}$ being realized, $\Pr(\overline{\omega}|\text{Y}, p)$ or $\Pr(\overline{\omega}|\neg\text{Y}, p)$ respectively.

If the DM does no learning they choose $\neg$Y or Y depending on the expected payoff from selecting Y, $m - p = \mu(\underline{\omega})u(\underline{\omega}) + \mu(\overline{\omega})u(\overline{\omega}) - p$, and which is larger as a result, $V(\text{Y}|p, \mu(\overline{\omega})) = m - p$ or $V(\neg\text{Y}|\mu(\overline{\omega})) = 0$. If the DM does some learning, then when they choose $\neg$Y their payoff is $V(\neg\text{Y}|\Pr(\overline{\omega}|\neg\text{Y}, p)) = -c(\Pr(\overline{\omega}|\neg\text{Y}, p)) + c(\mu(\overline{\omega}))$, and when they select Y their payoff is $V(\text{Y}|p, \Pr(\overline{\omega}|\text{Y}, p)) = \Pr(\overline{\omega}|\text{Y}, p)(u(\overline{\omega}) - p) + (1 - \Pr(\overline{\omega}|\text{Y}, p))(u(\underline{\omega}) - p) - c(\Pr(\overline{\omega}|\text{Y}, p)) + c(\mu(\overline{\omega}))$. Both $V(\text{Y}|p, \cdot)$ and $V(\neg\text{Y}|\cdot)$

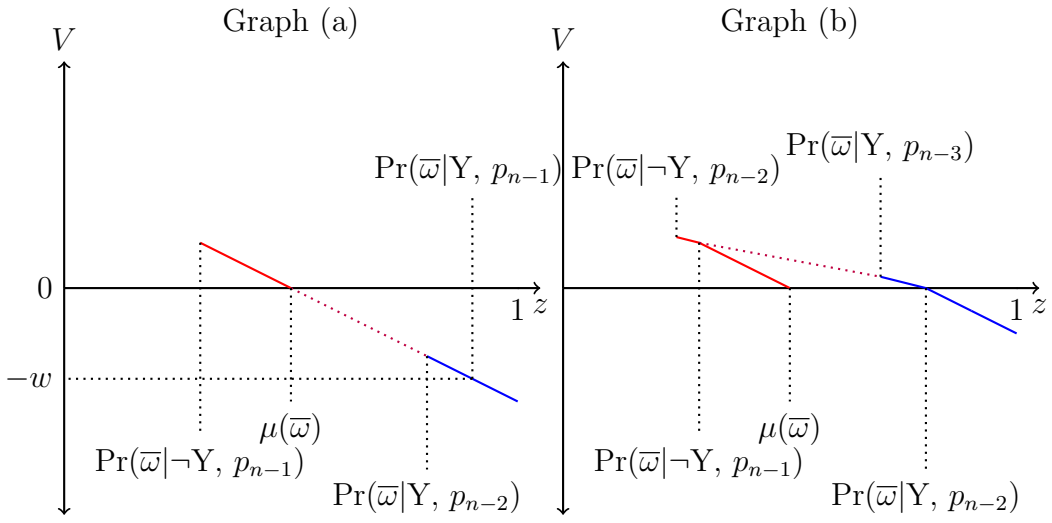
are weakly concave functions since c is weakly convex. The DM is maximizing the weighted average of the two. Notice that in any optimal solution, if the DM is learning, $\Pr(\overline{\omega}|\neg \mathrm{Y},\, p) < \mu(\overline{\omega}) < \Pr(\overline{\omega}|\mathrm{Y},\, p)$. As a result, to find the optimal solution of the DM I must find the cord from the top of $V(\neg \mathrm{Y}|\cdot)$ on the left of $\mu(\overline{\omega})$ to the top of $V(\mathrm{Y}|p,\, \cdot)$ on the right, so that I have a weakly concave closure.

When p increases, either the DM stops learning, which means the concave closure at $\mu(\overline{\omega})$ is $V(\neg \mathrm{Y}|\mu(\overline{\omega}))$, or the DM continues to learn, in which case, $V(\mathrm{Y}|p,\, \cdot)$ shifts downward at every point by the change in price, and the point where the cord that creates the weakly concave closure hits $V(\mathrm{Y}|p,\, \cdot)$ and $V(\neg \mathrm{Y}|\cdot)$ must then both weakly move to the right, which means $\Pr(\overline{\omega}|\neg \mathrm{Y},\, p)$ and $\Pr(\overline{\omega}|\mathrm{Y},\, p)$ are both weakly increasing, which establishes necessity.

Next I show sufficiency by constructing a weakly convex c function that generates the observed behavior, and assume $\overline{C_\mu} = 0$. It is without loss to assume $\mathcal{P} = \{p_1,\, p_2,\, \ldots,\, p_n\}$, with $p_1 < \cdots < p_n$, and $\Pr(\mathrm{Y}|\Omega,\, \mathcal{P})$ are such that $n \geq 4$ (in the example graphs, $n = 6$), with $\Pr(\mathrm{Y}|p_1) = 1$, $\Pr(\mathrm{Y}|p_n) = 0$, and $\Pr(\mathrm{Y}|p_i) \in (0,\, 1)$ for $p_i \in \mathcal{P}\backslash\{p_1,\, p_n\}$. I draw the graphs step by step and the different components are going to help me pick $u(\underline{\omega})$ and $u(\overline{\omega})$, and tell me a suitable c. The horizontal axis goes from zero to one, I call the horizontal coordinate z (since Y is already being used for something other than height I do not want to use x). The vertical axis may take positive and negative values, I call this coordinate height, whether it be positive or negative, since Y is being used.

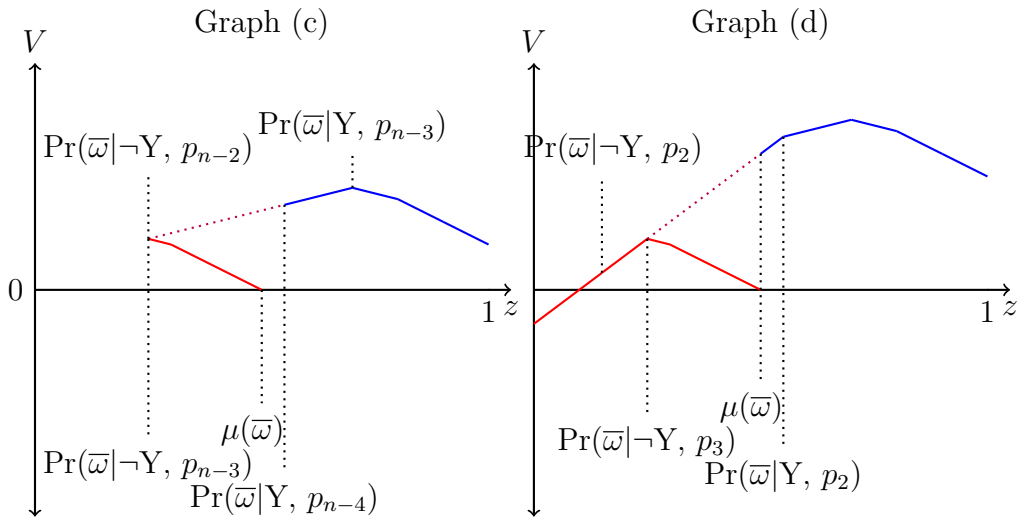
First (graph a): I draw a line segment from $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-1})$ and a positive height, to z=1 and a negative height, so that when $z = \mu(\overline{\omega})$, the height is 0, and so that the height at $\Pr(\overline{\omega}|\mathrm{Y},\, p_{n-1})$ is $-w$ such that $p_{n-1} - p_2 - w > 0$. The part of the segment between $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-1})$ and a $z = \mu(\overline{\omega})$ I make red, the segment between $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_{n-2})$ and $z = 1$ I make blue, and the rest I erase. Second (graph b): I take the blue segment and increase the height by $p_{n-1} - p_{n-2}$, then I draw a new line segment from the blue line at $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_{n-2})$, through the red segment at $z =$

$\Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-1})$ and continue on to $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-2})$. The part of the new segment between $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-1})$ and $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-2})$ I make red, the part of the new segment between $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_{n-2})$ and $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_{n-3})$ I make blue, and the rest I erase. Third (graph c): I take the blue segment and increase the height by $p_{n-2} - p_{n-3}$, then I draw a new line segment from the blue line at $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_{n-3})$, through the red segment at $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-2})$ and continue on to $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-3})$. The part of the new segment between $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-2})$ and $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-3})$ I make red, the part of the new segment between $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_{n-3})$ and $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_{n-4})$ I make blue, and the rest I erase.[14] Eventually (graph d), after continuing in the above fashion, I take the blue segment and increase the height by $p_3 - p_2$, then I draw a new line segment from the blue line at $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_2)$, through the red segment at $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_3)$ and continue on to $z = 0$. The part of the new segment between $z = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_3)$ and $z = 0$ I make red, the part of the new segment between $z = \mu(\overline{\omega})$ and $z = \Pr(\overline{\omega}|\mathrm{Y},\, p_2)$ I make blue, and the rest I erase.



The red line segments I take to be $-c(z) + c(\mu(\overline{\omega}))$ for $z \in [0,\, \mu(\overline{\omega})]$. Next, let $b(z)$ denote the height of the blue segment for $z \in [\mu(\overline{\omega}),\, 1]$ (graph d). $b(\mu(\overline{\omega})) > 0$

[14] In the graphs $\Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-2}) = \Pr(\overline{\omega}|\neg \mathrm{Y},\, p_{n-3})$, which is meant to provide some insight into what happens when $\Pr(\overline{\omega}|\neg \mathrm{Y},\, p)$ or $\Pr(\overline{\omega}|\mathrm{Y},\, p)$ are only weakly decreasing in p.



because either the slope of the blue segments is negative (where it is defined), in which case $b(\mu(\overline{\omega}))$ is more than $b(\Pr(\overline{\omega}|\text{Y}, p_{n-1}))$, which is strictly positive based on how I chose $-w$ at the beginning, or the slope of the blue segments is positive or zero immediately to the right of $z = \mu(\overline{\omega})$, in which case $b(\mu(\overline{\omega}))$ is greater or equal to the height of the red segments at $\Pr(\overline{\omega}|\neg\text{Y}, p_{n-1})$, which is strictly positive by construction. Further, $b(\mu(\overline{\omega})) - p_{n-1} + p_2 < 0$, again based on how I picked w. Next I pick the mean quality to be $m = p_2 + b(\mu(\overline{\omega})) \in (p_2, p_{n-1})$ so that when $p = m$ the DM is indifferent between choosing ¬Y without learning and choosing Y without learning. Next, I pick $u(\underline{\omega})$ and $u(\overline{\omega})$ so that $\mu(\underline{\omega})u(\underline{\omega}) + \mu(\overline{\omega})u(\overline{\omega}) = m$, and so that (to help ensure convexity of c):

$$u(\overline{\omega}) - u(\underline{\omega}) + \frac{c(\mu(\overline{\omega})) - c(\Pr(\overline{\omega}|\neg\text{Y}, p_{n-1})}{\mu(\overline{\omega}) - \Pr(\overline{\omega}|\neg\text{Y}, p_{n-1})} \geq \frac{b(\Pr(\overline{\omega}|\text{Y}, p_2)) - b(\mu(\overline{\omega}))}{\Pr(\overline{\omega}|\text{Y}, p_2) - \mu(\overline{\omega})}.$$

Next, for $z \in [\mu(\overline{\omega}), 1]$ I let $-c(z)+c(\mu(\overline{\omega})) = b(z)-b(\mu(\overline{\omega}))-(z-\mu(\overline{\omega}))(u(\overline{\omega})-u(\underline{\omega}))$. Finally, I fix $c(\mu(\overline{\omega}))$ so that $\min_{z\in[0,1]} c(z) = 0$.

Next, the second part of the proof (what follows) establishes the sufficiency of the conditions when there are more than two states of the world. Pick $B \in \sigma(\Omega)\backslash\{\emptyset, \Omega\}$ such that $\Pr(B|\neg\text{Y}, p)$ and $\Pr(B|\text{Y}, p)$ are both weakly increasing over the set of p

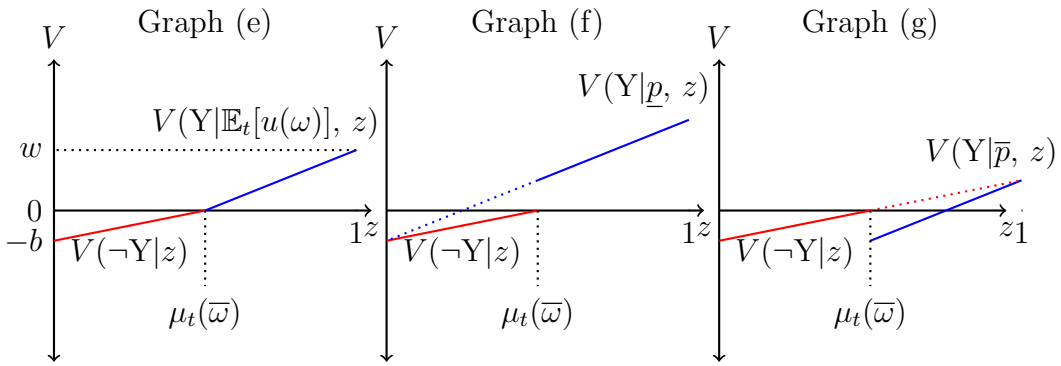
where the DM learns and $\forall p \in \mathcal{P}$ with $s(p) \in S^L : \Pr(\text{Y}|B, p) > \Pr(\text{Y}|B^c, p)$, with representative states $\overline{\omega} \in B$ and $\underline{\omega} \in B^c$. Then pick $u(\underline{\omega})$ and $u(\overline{\omega})$ with $u(\underline{\omega}) < u(\overline{\omega})$ so that $\mu(B^c)u(\underline{\omega}) + \mu(B)u(\overline{\omega}) = m$ (which will be shifted), and make it so that $u(\omega) = u(\overline{\omega})$ for all $\omega \in B$ and $u(\omega) = u(\underline{\omega})$ for all $\omega \in B^c$. Next, assume c is only a function of $\mu(B)$, and that $\overline{C_\mu} = 0$. I can then follow the same proof strategy as in the first part of the proof, except I replace each $\mu(\overline{\omega})$ with $\mu(B)$, each $\mu(\underline{\omega})$ with $\mu(B^c)$, each $\Pr(\overline{\omega}|\neg\text{Y}, p)$ with $\Pr(B|\neg\text{Y}, p)$, and each $\Pr(\overline{\omega}|\text{Y}, p)$ with $\Pr(B|\text{Y}, p)$. ■

Proof of Theorem 3. In the first part of the proof assume there are two states of the world. Behavior being rationalized by a costly learning model is necessary for it being rationalized by an aggregate costly learning model, as established by comparing the results in Theorem 1 and Theorem 5. I will now introduce a helpful lemma before continuing.

Lemma 1. Given prices $\underline{p}$ and $\overline{p} > \underline{p}$, payoffs for Y $u(\underline{\omega})$ and $u(\overline{\omega})$ such that $u(\underline{\omega}) \leq \underline{p}$ and $u(\overline{\omega}) \geq \overline{p}$, and a type t with a prior belief such that $\mu_t(\overline{\omega}) \in (0, 1)$ and $\mu_t(\underline{\omega}) \equiv 1 - \mu_t(\overline{\omega})$, I can create a measure of informativeness $c_t : [0, 1] \to \mathbb{R}_+$ such that if DMs of type t have a PS cost function for information outcomes C_μ^t that uses c_t as its measure of informativeness, then they are indifferent between doing no learning ($\Pr(\text{Y}|\underline{\omega}, p) = \Pr_t(\text{Y}|\overline{\omega}_t, p)$) and perfectly observing the payoff from Y ($\Pr_t(\text{Y}|\underline{\omega}, p) = 0$, $\Pr_t(\text{Y}|\overline{\omega}, p) = 1$) when the price is $p \in \{\underline{p}, \overline{p}\}$, strictly prefer perfectly observing the payoff from Y over all other learning strategies (any other probabilities of selecting Y in each state) when the price is $p \in (\underline{p}, \overline{p})$, and strictly prefer doing no learning over all other learning strategies when $p \notin [\underline{p}, \overline{p}]$, if the payoffs from Y are a mean preserving spread of the prices: $\mu_t(\underline{\omega})u(\underline{\omega}) + \mu_t(\overline{\omega})u(\overline{\omega}) = \mu_t(\underline{\omega})\underline{p} + \mu_t(\overline{\omega})\overline{p}$.

Proof of Lemma 1. In the fashion of the proof of Theorem 2, I construct the functions $V(\neg\text{Y}|z)$ in red on the left of $\mu_t(\overline{\omega})$ and $V(\text{Y}|p, z)$ in blue on the right of $\mu_t(\overline{\omega})$ (graphs below), and assume $\overline{C_\mu} = 0$ for each type. When $p = \mu_t(\underline{\omega})u(\underline{\omega}) + \mu_t(\overline{\omega})u(\overline{\omega}) = \mathbb{E}_t[u(\omega)]$, these functions must both be equal to zero. So, I am going to draw $V(\neg\text{Y}|z)$ and $V(\text{Y}|\mathbb{E}_t[u(\omega)], z)$ as line segments on either side of $\mu_t(\overline{\omega})$, so

that $V(\neg \mathrm{Y}|\mu_t(\overline{\omega})) = V(\mathrm{Y}|\mathbb{E}_t[u(\omega)],\, \mu_t(\overline{\omega})) = 0$ (graph e), and so that they satisfy two other properties. **(i)**: First, when the blue line segment is shifted up by $\mathbb{E}_t[u(\omega)] - \underline{p}$, it must be that if I were to extend the blue line segment so that it reaches $z = 0$, it hits the red line segment at $z = 0$ (graph f), so that when $p = \underline{p}$ the DM is indifferent between learning nothing and everything, and at all lower prices they strictly prefer learning nothing. **(ii)**: Second, it must be that when the blue line segment is shifted down by $\overline{p} - \mathbb{E}_t[u(\omega)]$, it must be that if I were to extend the red line segment so that it reaches $z = 1$, it hits the blue line segment at $z = 1$ (graph g), so that when $p = \overline{p}$ the DM is indifferent between learning nothing and everything, and at all higher prices the DM strictly prefers learning nothing. Together, these properties imply that the slope of the blue line segment is strictly greater than the slope of the red line segment, and the DM strictly prefers learning everything over all other learning strategies when $p \in (\underline{p},\, \overline{p})$.



How do I find line segments that satisfy the two properties? I select the height of the red segment at zero ($-b$ in graph e) and the blue segment at one (w in graph e) so that:

$$\textbf{(i)}: \ \frac{w}{\mu_t(\underline{\omega})} = b + w + (\mathbb{E}_t[u(\omega)] - \underline{p}), \ \textbf{(ii)}: \ \frac{b}{\mu_t(\overline{\omega})} = b + w - (\overline{p} - \mathbb{E}_t[u(\omega)])$$

$$\Leftrightarrow b = \frac{\mu_t(\overline{\omega})w}{\mu_t(\underline{\omega})} - (\mathbb{E}_t[u(\omega)] - \underline{p}),\ b = \frac{\mu_t(\overline{\omega})w}{\mu_t(\underline{\omega})} - \frac{\mu_t(\overline{\omega})}{\mu_t(\underline{\omega})}(\overline{p} - \mathbb{E}_t[u(\omega)]).$$

$$\text{But: } \mathbb{E}_t[u(\omega)] = \mu_t(\underline{\omega})\underline{p} + \mu_t(\overline{\omega})\overline{p} \Rightarrow -(\mathbb{E}_t[u(\omega)] - \underline{p}) = -\frac{\mu_t(\overline{\omega})}{\mu_t(\underline{\omega})}(\overline{p} - \mathbb{E}_t[u(\omega)]),$$

$$\text{so: } b = \frac{\mu_t(\overline{\omega})w}{\mu_t(\underline{\omega})} - (\mathbb{E}_t[u(\omega)] - \underline{p}) \Leftrightarrow b = \frac{\mu_t(\overline{\omega})w}{\mu_t(\underline{\omega})} - \frac{\mu_t(\overline{\omega})}{\mu_t(\underline{\omega})}(\overline{p} - \mathbb{E}_t[u(\omega)]).$$

Thus, given any b, I simply make $w = \frac{\mu_t(\underline{\omega})}{\mu_t(\overline{\omega})}b + (\overline{p} - \mathbb{E}_t[u(\omega)])$.

For $z \in [0,\, \mu_t(\overline{\omega})]$, I let $-c(z) + c(\mu_t(\overline{\omega})) = V(\neg \mathrm{Y}|z)$. For $z \in [\mu_t(\overline{\omega}),\, 1]$, I let $-c(z) + c(\mu_t(\overline{\omega})) = V(\mathrm{Y}|\mathbb{E}_t[u(\omega)],\, z) - (z - \mu_t(\overline{\omega}))(u(\overline{\omega}) - u(\underline{\omega}))$. Finally, fix $c(\mu(\overline{\omega}))$ so that $\min_{z\in[0,\,1]} c(z) = 0$. It is easy to show that, given how I picked w based on b, this procedure produces a weakly convex c. ■

I now return to the **proof of Theorem 3.** I next show that for any type t it must be that $\Pr_t(\mathrm{Y}|\underline{\omega},\, p)$ and $\Pr_t(\mathrm{Y}|\overline{\omega},\, p)$ are weakly decreasing in p. If $\Pr_t(\mathrm{Y}|\underline{\omega},\, p) = \Pr_t(\mathrm{Y}|\overline{\omega},\, p) = 1$ clearly both need to decrease. If $\Pr_t(\mathrm{Y}|\underline{\omega},\, p) < \Pr_t(\mathrm{Y}|\overline{\omega},\, p) \leq 1$, then if $\Pr_t(\mathrm{Y}|\underline{\omega},\, p)$ increases, Theorem 1 requires $\Pr_t(\mathrm{Y}|\overline{\omega},\, p)$ decreases, and Bayes' Rule tells us:

$$\Pr_t(\overline{\omega}|\mathrm{Y},\, p) = \frac{\Pr_t(\mathrm{Y}|\overline{\omega},\, p)\mu_t(\overline{\omega})}{\Pr_t(\mathrm{Y}|\underline{\omega},\, p)\mu_t(\underline{\omega}) + \Pr_t(\mathrm{Y}|\overline{\omega},\, p)\mu_t(\overline{\omega})} = \frac{1}{1 + \dfrac{\Pr_t(\mathrm{Y}|\underline{\omega},\, p)\mu_t(\underline{\omega})}{\Pr_t(\mathrm{Y}|\overline{\omega},\, p)\mu_t(\overline{\omega})}},$$

so $\Pr_t(\overline{\omega}|\mathrm{Y},\, p)$ decreases, which violates Theorem 2. If $\Pr_t(\mathrm{Y}|\underline{\omega},\, p) \leq \Pr_t(\mathrm{Y}|\overline{\omega},\, p) < 1$, and $\Pr_t(\mathrm{Y}|\overline{\omega},\, p)$ increases, then Theorem 1 tells me $\Pr_t(\mathrm{Y}|\underline{\omega},\, p)$ decreases, and Bayes' Rule tells us:

$$\Pr_t(\overline{\omega}|\neg\mathrm{Y},\, p) = \frac{\Pr_t(\neg\mathrm{Y}|\overline{\omega},\, p)\mu_t(\overline{\omega})}{\Pr_t(\neg\mathrm{Y}|\underline{\omega},\, p)\mu_t(\underline{\omega}) + \Pr_t(\neg\mathrm{Y}|\overline{\omega},\, p)\mu_t(\overline{\omega})} = \frac{1}{1 + \dfrac{\Pr_t(\neg\mathrm{Y}|\underline{\omega},\, p)\mu_t(\underline{\omega})}{\Pr_t(\neg\mathrm{Y}|\overline{\omega},\, p)\mu_t(\overline{\omega})}},$$

so $\Pr_t(\overline{\omega}|\neg\mathrm{Y},\, p)$ decreases, which violates Theorem 2. As a result $\Pr(\mathrm{Y}|\underline{\omega},\, p)$ and $\Pr(\mathrm{Y}|\overline{\omega},\, p)$ both being weakly decreasing is necessary for the behavior to be rationalized by an aggregate PS model.

Sufficiency is the challenge. I assume $\overline{C_\mu} = 0$ (for all types) and that there is a $p \in \mathcal{P}$ such that $\underline{s}(p) < \overline{s}(p)$, otherwise sufficiency is easy to establish by

making learning costly enough (since belief is fixed). It is without loss to assume $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$, with $p_1 < \cdots < p_n$. Notice that it is without loss to also assume that between any two adjacent prices p_i and p_{i+1}, if $\Pr(\mathrm{Y}|p)$ decreases, then only one of $\Pr(\mathrm{Y}|\underline{\omega}, p)$ and $\Pr(\mathrm{Y}|\overline{\omega}, p)$ decreases, since otherwise I can always enrich the behavior in a way so that behavior is still rationalized by a costly learning model, and so that this is true. I can then rationalize the richer dataset, which in turn rationalizes the less rich dataset. Similarly, I can assume without loss that $\Pr(\mathrm{Y}|p_1) = 1$ and $\Pr(\mathrm{Y}|p_n) = 0$, and that there are at least two pairs of prices between which $\Pr(\mathrm{Y}|\overline{\omega}, p)$ strictly decreases. Further, to ease exposition, I can assume $\Pr(\mathrm{Y}|p_2) < 1$ and $\Pr(\mathrm{Y}|p_{n-1}) > 0$, because otherwise I can focus on a simpler dataset with less prices in it, and rationalizing this simpler dataset rationalizes the more complicated one. Notice that the change between p_1 and p_2 is then a reduction in $\Pr(\mathrm{Y}|\underline{\omega}, p)$, while the change between p_{n-1} and p_n is then a reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)$.

I am now going to start pairing reductions in $\Pr(\mathrm{Y}|\underline{\omega}, p)\mu(\underline{\omega})$ with reductions in $\Pr(\mathrm{Y}|\overline{\omega}, p)\mu(\overline{\omega})$. The goal is to rationalize each pair with a type using Lemma 1, but I have to be careful about how I construct the pairs, and I start off with an initial pairing that intentionally fails in an eventually fixable way. Initially (henceforth call the "initial pairing"), pair reductions $\delta_{\underline{\omega}}$ in $\Pr(\mathrm{Y}|\underline{\omega}, p)\mu(\underline{\omega})$ between two adjacent prices, with reductions $\delta_{\overline{\omega}}$ in $\Pr(\mathrm{Y}|\overline{\omega}, p)\mu(\overline{\omega})$ between higher adjacent prices that are not p_{n-1} and p_n (none of the reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)\mu(\overline{\omega})$ between p_{n-1} and p_n is paired in the initial pairing), so that the following four properties are satisfied. First, all reduction in $\Pr(\mathrm{Y}|\underline{\omega}, p)$ is paired, all reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)$ except that between p_{n-1} and p_n is paired. Second, each pairing has $\frac{\delta_{\underline{\omega}}}{\mu(\underline{\omega})} > \frac{\delta_{\overline{\omega}}}{\mu(\overline{\omega})}$. Third, the pair that has the strictly lowest $\delta_{\underline{\omega}}/(\delta_{\underline{\omega}} + \delta_{\overline{\omega}})$ is a pairing (denoted $(\tilde{\delta_{\underline{\omega}}}, \tilde{\delta_{\overline{\omega}}})$) between reduction in $\Pr(\mathrm{Y}|\underline{\omega}, p)\mu(\underline{\omega})$ between p_1 and p_2, and reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)\mu(\overline{\omega})$ between the lowest pair of prices that have a reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)\mu(\overline{\omega})$, call them p_j, p_{j+1}.

Fourth, if I pick $u(\overline{\omega}) = p_n$, and pick $u(\underline{\omega}) < p_1$ so that:

$$\frac{\tilde{\delta_{\underline{\omega}}}u(\underline{\omega}) + \tilde{\delta_{\overline{\omega}}}u(\overline{\omega})}{\tilde{\delta_{\underline{\omega}}} + \tilde{\delta_{\overline{\omega}}}} = \frac{\tilde{\delta_{\underline{\omega}}}p_1 + \tilde{\delta_{\overline{\omega}}}p_j}{\tilde{\delta_{\underline{\omega}}} + \tilde{\delta_{\overline{\omega}}}}, \tag{1}$$

$$\frac{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}})u(\underline{\omega}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})u(\overline{\omega})}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})} < \frac{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}})p_2 + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})p_{j+1}}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})}. \tag{2}$$

Why is it possible to satisfy all four properties simultaneously? The second property is possible because of the first property, and ensures that a reduction in $\Pr(\mathrm{Y}|\underline{\omega}, p)$ is always paired with a smaller reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)$. Then, when I pick the pair with the strictly lowest $\delta_{\underline{\omega}}/(\delta_{\underline{\omega}} + \delta_{\overline{\omega}})$ in the third property, I can make sure the reduction in $\Pr(\mathrm{Y}|\underline{\omega}, p)$ is paired with an arbitrarily close in size reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)$, which makes the two values:

$$\frac{\tilde{\delta_{\underline{\omega}}}u(\underline{\omega}) + \tilde{\delta_{\overline{\omega}}}u(\overline{\omega})}{\tilde{\delta_{\underline{\omega}}} + \tilde{\delta_{\overline{\omega}}}} \text{ and } \frac{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}})u(\underline{\omega}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})u(\overline{\omega})}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})}$$

arbitrarily close together, which ensures the fourth property can be satisfied since satisfying (1) and continuity then implies (2).

The behavior of the $(\tilde{\delta_{\underline{\omega}}}, \tilde{\delta_{\overline{\omega}}})$ pairing can be rationalized by a single type according to Lemma 1, and this continues to be true as long as I spread $u(\underline{\omega})$ and $u(\overline{\omega})$ away from each other so (1) is satisfied. Further, the second property then implies such spreads that maintain the equality in (1) increase the mean $m = \mu(\underline{\omega})u(\underline{\omega}) + \mu(\overline{\omega})u(\overline{\omega})$. Then, looking at how the ratios change when the weights change for reductions $\delta_{\underline{\omega}}$ between p_i and p_{i+1} and $\delta_{\overline{\omega}}$ between p_k and p_{k+1} with $i + 1 > 1$ and $k + 1 > j$:

$$\frac{\partial \dfrac{\delta_{\underline{\omega}}u(\underline{\omega}) + \delta_{\overline{\omega}}u(\overline{\omega})}{\delta_{\underline{\omega}} + \delta_{\overline{\omega}}}}{\partial \delta_{\underline{\omega}}} = \frac{\delta_{\overline{\omega}}(u(\underline{\omega}) - u(\overline{\omega}))}{(\delta_{\underline{\omega}} + \delta_{\overline{\omega}})^2} < \frac{\delta_{\overline{\omega}}(p_{i+1} - p_{k+1})}{(\delta_{\underline{\omega}} + \delta_{\overline{\omega}})^2} = \frac{\partial \dfrac{\delta_{\underline{\omega}}p_{i+1} + \delta_{\overline{\omega}}p_{k+1}}{\delta_{\underline{\omega}} + \delta_{\overline{\omega}}}}{\partial \delta_{\underline{\omega}}}, \tag{3}$$

$$\text{so } \frac{\delta_{\underline{\omega}}u(\underline{\omega}) + \delta_{\overline{\omega}}u(\overline{\omega})}{\delta_{\underline{\omega}} + \delta_{\overline{\omega}}} < \frac{\delta_{\underline{\omega}}p_{i+1} + \delta_{\overline{\omega}}p_{k+1}}{\delta_{\underline{\omega}} + \delta_{\overline{\omega}}}.$$

What I do next is I reduce each $\delta_{\underline{\omega}}$ to $\hat{\delta_{\underline{\omega}}}$ (leaving the excess unpaired), so:

$$\frac{\hat{\delta_{\underline{\omega}}}u(\underline{\omega}) + \delta_{\overline{\omega}}u(\overline{\omega})}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}} = \frac{\hat{\delta_{\underline{\omega}}}p_{i+1} + \delta_{\overline{\omega}}p_{k+1}}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}} \geq \frac{\hat{\delta_{\underline{\omega}}}p_2 + \delta_{\overline{\omega}}p_{j+1}}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}} > \frac{\hat{\delta_{\underline{\omega}}}p_1 + \delta_{\overline{\omega}}p_j}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}}. \tag{4}$$

Then (3) tells me there is a unique $\hat{\delta_{\underline{\omega}}}$ that satisfies the equality in (4). Further, the inequalities from (4) and (1) and (2) imply that for each resultant pairing:

$$\frac{\hat{\delta_{\underline{\omega}}}}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}} < \frac{\tilde{\delta_{\underline{\omega}}}}{\tilde{\delta_{\underline{\omega}}} + \tilde{\delta_{\overline{\omega}}}}, \text{ and } \frac{\hat{\delta_{\underline{\omega}}}}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}} < \frac{\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})}. \tag{5}$$

So (4), (5) and (2) say: $\frac{\hat{\delta_{\underline{\omega}}}u(\underline{\omega}) + \delta_{\overline{\omega}}u(\overline{\omega})}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}} = \frac{\hat{\delta_{\underline{\omega}}}p_{i+1} + \delta_{\overline{\omega}}p_{k+1}}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}} \geq \frac{\hat{\delta_{\underline{\omega}}}p_2 + \delta_{\overline{\omega}}p_{j+1}}{\hat{\delta_{\underline{\omega}}} + \delta_{\overline{\omega}}}$

$$> \frac{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}})p_2 + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})p_{j+1}}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})} > \frac{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}})u(\underline{\omega}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})u(\overline{\omega})}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})},$$

which means that each pair $(\hat{\delta_{\underline{\omega}}}, \delta_{\overline{\omega}})$, not including $(\tilde{\delta_{\underline{\omega}}}, \tilde{\delta_{\overline{\omega}}})$, has a higher mean than the mean that is left after removing only $(\tilde{\delta_{\underline{\omega}}}, \tilde{\delta_{\overline{\omega}}})$, so the pairings with $\hat{\delta_{\underline{\omega}}}$'s have reduced the mean of what is left over.

But, I need to match the reductions in $\Pr(Y|\underline{\omega}, p)\mu(\underline{\omega})$ that are now unpaired with the reduction in $\Pr(Y|\overline{\omega}, p)\mu(\overline{\omega})$ between p_{n-1} and p_n. So, beginning with the lowest pair of prices with unmatched change and working my way up I take all of the unmatched reduction in $\Pr(Y|\underline{\omega}, p)\mu(\underline{\omega})$ between p_m and p_{m+1}, denoted $\bar{\delta_{\underline{\omega}}}$, and match it with enough reduction in $\Pr(Y|\overline{\omega}, p)\mu(\overline{\omega})$ from between p_{n-1} and p_n, denoted $\bar{\delta_{\overline{\omega}}}$, so that:

$$\frac{\bar{\delta_{\underline{\omega}}}u(\underline{\omega}) + \bar{\delta_{\overline{\omega}}}u(\overline{\omega})}{\bar{\delta_{\underline{\omega}}} + \bar{\delta_{\overline{\omega}}}} = \frac{\bar{\delta_{\underline{\omega}}}p_{m+1} + \bar{\delta_{\overline{\omega}}}p_n}{\bar{\delta_{\underline{\omega}}} + \bar{\delta_{\overline{\omega}}}},$$

$$\frac{\partial \frac{\delta_{\underline{\omega}}u(\underline{\omega}) + \delta_{\overline{\omega}}u(\overline{\omega})}{\delta_{\underline{\omega}} + \delta_{\overline{\omega}}}}{\partial \delta_{\overline{\omega}}} = \frac{\delta_{\underline{\omega}}(u(\overline{\omega}) - u(\underline{\omega}))}{(\delta_{\underline{\omega}} + \delta_{\overline{\omega}})^2} > \frac{\delta_{\underline{\omega}}(p_n - p_{m+1})}{(\delta_{\underline{\omega}} + \delta_{\overline{\omega}})^2} = \frac{\partial \frac{\delta_{\underline{\omega}}p_{m+1} + \delta_{\overline{\omega}}p_n}{\delta_{\underline{\omega}} + \delta_{\overline{\omega}}}}{\partial \delta_{\overline{\omega}}}, \tag{6}$$

so there is a unique $\bar{\delta_{\overline{\omega}}}$ for each $\bar{\delta_{\underline{\omega}}}$. But,

$$\frac{\bar{\delta_{\underline{\omega}}}p_{m+1} + \bar{\delta_{\overline{\omega}}}p_n}{\bar{\delta_{\underline{\omega}}} + \bar{\delta_{\overline{\omega}}}} > \frac{\bar{\delta_{\underline{\omega}}}p_2 + \bar{\delta_{\overline{\omega}}}p_{j+1}}{\bar{\delta_{\underline{\omega}}} + \bar{\delta_{\overline{\omega}}}} > \frac{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}})p_2 + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})p_{j+1}}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})}$$

$$> \frac{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}})u(\underline{\omega}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})u(\overline{\omega})}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})},$$

where the second inequality is true due to (3) or (6), and third is due to (2), which in a fashion similar to how (4) implies (5), implies:

$$\frac{\bar{\delta_{\underline{\omega}}}}{\bar{\delta_{\underline{\omega}}} + \bar{\delta_{\overline{\omega}}}} < \frac{\tilde{\delta_{\underline{\omega}}}}{\tilde{\delta_{\underline{\omega}}} + \tilde{\delta_{\overline{\omega}}}}, \text{ and } \frac{\bar{\delta_{\underline{\omega}}}}{\bar{\delta_{\underline{\omega}}} + \bar{\delta_{\overline{\omega}}}} < \frac{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}})}{(\mu(\underline{\omega}) - \tilde{\delta_{\underline{\omega}}}) + (\mu(\overline{\omega}) - \tilde{\delta_{\overline{\omega}}})}. \tag{7}$$

This all means I am doomed to failure since the mean required by each remaining pair is strictly higher than the mean of the unmatched reductions I have left and I run out of reduction in $\Pr(Y|\overline{\omega}, p)\mu(\overline{\omega})$ from between p_{n-1} and p_n before I have paired all the unpaired reduction in $\Pr(Y|\underline{\omega}, p)\mu(\underline{\omega})$.

How do I do better? I return to the initial pairing, but increase $u(\overline{\omega})$ and decrease $u(\underline{\omega})$ so that (1) is satisfied, which increases the mean of what is left over after the $(\tilde{\delta_{\underline{\omega}}}, \tilde{\delta_{\overline{\omega}}})$ pairing (eventually (2) is violated as a result, but I do not need (2) to be satisfied by the final solution, and its eventual violation makes this strategy work). Then, (5) tells me that the $\hat{\delta_{\underline{\omega}}}$ I had previously picked to satisfy the equality in (4) were too low, so this iteration I reduce each $\delta_{\underline{\omega}}$ less (but still reduce given how I picked $(\tilde{\delta_{\underline{\omega}}}, \tilde{\delta_{\overline{\omega}}})$ to satisfy the second property), which means I have less unmatched reduction in $\Pr(Y|\underline{\omega}, p)\mu(\underline{\omega})$ to pair after producing the $\hat{\delta_{\underline{\omega}}}$'s. Further, for any amount of unmatched $\Pr(Y|\underline{\omega}, p)\mu(\underline{\omega})$, the previous amount of reduction in $\Pr(Y|\overline{\omega}, p)\mu(\overline{\omega})$ from between p_{n-1} and p_n I would have matched it with is too large given (7), since I spread $u(\underline{\omega})$ and $u(\overline{\omega})$ but (1) is satisfied, so the unmatched $\Pr(Y|\overline{\omega}, p)\mu(\overline{\omega})$ goes farther, and eventually this strategy works.

The second part of the proof (which follows) deals with more than two states of the world. Achieving the desired result is close to trivial given the work done in

the first part of the proof. Picking $B \in \sigma(\Omega)\backslash\{\emptyset, \Omega\}$ such that $\forall p \in \mathcal{P}$ with $s(p) \in S^L$: $\Pr(\mathrm{Y}|B,\, p) > \Pr(\mathrm{Y}|B^c,\, p)$, with representative states $\overline{\omega} \in B$ and $\underline{\omega} \in B^c$, assume $u(\omega) = u(\overline{\omega})$ for all $\omega \in B$ and $u(\omega) = u(\underline{\omega})$ for all $\omega \in B^c$. Also assume each c_t is only a function of $\mu(B)$, and that $\overline{C_\mu} = 0$ (for all types). Reductions in $\Pr(\mathrm{Y}|\omega,\, p)\mu(\omega)$ for $\omega \in B^c$ can be paired with reductions in $\Pr(\mathrm{Y}|\omega,\, p)\mu(\omega)$ for $\omega \in B$, denoting the two states with reductions for a type t by $\underline{\omega}_t \in B^c$ and $\overline{\omega}_t \in B$, and noticing that Lemma 1 still holds if I replace each $\underline{\omega}$ with $\underline{\omega}_t$ and replace each $\overline{\omega}$ with $\overline{\omega}_t$ in the proof and statement of Lemma 1. Then, an argument analogous to the one in the first part of the proof achieves the desired result. ■

Definition: Given a prior belief μ, the behavior is **rationalized by an aggregate PS model with consistent beliefs** if it is rationalized by an aggregate PS model where each type has the same belief $\mu_t(\overline{\omega}) = \mu(\overline{\omega})$.

Theorem 4. Given a prior belief μ and a finite and non-empty set of prices $\mathcal{P}$, and assuming that there are exactly two states of the world and for all $p \in \mathcal{P} : 0 < \Pr(\mathrm{Y}|\, p) < 1$: the behavior is rationalized by an aggregate PS model with consistent beliefs iff it is rationalized by an aggregate PS model.

Proof of Theorem 4. Behavior being rationalized by an aggregate PS model is clearly necessary for behavior being rationalized by an aggregate PS model with consistent beliefs. Behavior being rationalized by any aggregate costly learning model necessitates that it is rationalized by a costly learning model (see Theorem 1 and Theorem 5), and thus Theorem 1 tells us that for all $p \in \mathcal{P}$ I have $\Pr(\mathrm{Y}|\underline{\omega},\, p) \leq \Pr(\mathrm{Y}|\overline{\omega},\, p)$ and if there is a p such that $\Pr(\mathrm{Y}|\underline{\omega},\, p) = \Pr(\mathrm{Y}|\overline{\omega},\, p)$ then there is no price with learning and the behavior is rationalized by an aggregate PS model with consistent beliefs as it is rationalized by a single type by Theorem 2.

I thus assume that for all $p \in \mathcal{P}$ I have $\Pr(\mathrm{Y}|\underline{\omega},\, p) < \Pr(\mathrm{Y}|\overline{\omega},\, p)$ and that behavior is rationalized by an aggregate PS model, and show that behavior is rationalized by an aggregate PS model with consistent beliefs. I can pick an $\epsilon > 0$ such that for all $p \in \mathcal{P}$ I have $\Pr(\mathrm{Y}|\underline{\omega},\, p) + \epsilon < \Pr(\mathrm{Y}|\overline{\omega},\, p)$. I now add some additional prices (and

associated behavior) to $\mathcal{P}$, turning it into $\tilde{\mathcal{P}}$, so that the behavior is still rationalized by an aggregate PS costly learning model and so that at the highest price $\overline{p} \in \tilde{\mathcal{P}}$: $\Pr(\mathrm{Y}|\underline{\omega}, \overline{p}) = 0$ and $\Pr(\mathrm{Y}|\overline{\omega}, \overline{p}) = \epsilon$, and at the lowest price $\underline{p} \in \tilde{\mathcal{P}}$: $\Pr(\mathrm{Y}|\underline{\omega}, \underline{p}) = 1-\epsilon$ and $\Pr(\mathrm{Y}|\overline{\omega}, \underline{p}) = 1$. As I do in the proof of Theorem 3, I can assume without loss that between any two adjacent prices p_i and p_{i+1}, if $\Pr(\mathrm{Y}|p)$ decreases, then only one of $\Pr(\mathrm{Y}|\underline{\omega}, p)$ and $\Pr(\mathrm{Y}|\overline{\omega}, p)$ decreases, since otherwise I can always enrich the behavior in a way so that behavior is still rationalized by an aggregate PS costly learning model, and so that this is true. I can then rationalize the richer dataset, which in turn rationalizes the less rich dataset.

I then repeatedly group (each such grouping of four reductions, to be described, becomes a type that is going to be rationalized by a PS type t) a reduction in $\Pr(\mathrm{Y}|\underline{\omega}, p)$ between two adjacent prices p_i^t and p_{i+1}^t (with $p_i^t < p_{i+1}^t$) with an equal size reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)$ between two higher adjacent prices p_j^t and p_{j+1}^t (with $p_{i+1} < p_j < p_{j+1}$) and a tiny reduction in $\Pr(\mathrm{Y}|\underline{\omega}, p)$ before $\underline{p}$ (making sure to leave some reduction in $\Pr(\mathrm{Y}|\underline{\omega}, p)$ below $\underline{p}$ remaining unless it is the last grouping) and an equal (tiny) reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)$ above $\overline{p}$ (making sure to leave some reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)$ above $\overline{p}$ remaining unless it is the last grouping). I continue to make these groupings until all reductions between $\underline{p}$ and $\overline{p}$ have been grouped. For each grouping I then have that a proportion z^t of their reductions in both $\Pr(\mathrm{Y}|\underline{\omega}, p)$ and $\Pr(\mathrm{Y}|\overline{\omega}, p)$ happen outside of $\underline{p}$ and $\overline{p}$. I assume that for each type t that their reductions in $\Pr(\mathrm{Y}|\underline{\omega}, p)$ occur at p_{i+1}^t and some (yet to be chosen) $\underline{p}^t < \underline{p}$ and their reduction in $\Pr(\mathrm{Y}|\overline{\omega}, p)$ occur at p_{j+1}^t and some (yet to be chosen) $\overline{p}^t > \overline{p}$.

I begin rationalizing the behavior of each type by selecting some $u(\underline{\omega}) < \underline{p}$ and $u(\overline{\omega}) > \overline{p}$ with some mean $m \equiv \mu(\underline{\omega})u(\underline{\omega}) + \mu(\overline{\omega})u(\overline{\omega})$. I then conduct mean preserving spreads on $u(\underline{\omega})$ and $u(\overline{\omega})$ until, for each type t:

$$\mu(\underline{\omega})\Big((1-z_t)(p_{i+1}^t - u(\underline{\omega})) + z_t(\underline{p} - u(\underline{\omega}))\Big) - \mu(\overline{\omega})\Big((1-z_t)(u(\overline{\omega}) - p_{j+1}^t)\Big)$$

$$= (1-z_t)(\mu(\underline{\omega})p_{i+1}^t + \mu(\overline{\omega})p_{j+1}^t - m) + \mu(\underline{\omega})z_t(\underline{p} - u(\underline{\omega})) > 0,$$

and

$$\mu(\overline{\omega})\Big((1-z_t)(u(\overline{\omega})-p^t_{j+1})+z_t(u(\overline{\omega})-\overline{p})\Big)-\mu(\underline{\omega})\Big((1-z_t)(p^t_{i+1}-u(\underline{\omega}))\Big)$$

$$=(1-z_t)(m-\mu(\underline{\omega})p^t_{i+1}-\mu(\overline{\omega})p^t_{j+1})+\mu(\overline{\omega})z_t(u(\overline{\omega})-\overline{p})>0.$$

This means I can, for each type t, pick $\underline{p}^t \in (u(\underline{\omega}),\, \underline{p})$ and $\overline{p}^t \in (\overline{p},\, u(\overline{\omega}))$ so that:

$$\mu(\overline{\omega})\Big((1-z_t)(u(\overline{\omega})-p^t_{j+1})+z_t(u(\overline{\omega})-\overline{p}^t)\Big)=\mu(\underline{\omega})\Big((1-z_t)(p^t_{i+1}-u(\underline{\omega}))+z_t(\underline{p}^t-u(\underline{\omega}))\Big).$$

Then, for each type t, I assume $\overline{C_\mu}=0$ and I can assign costs to the information outcomes (with $\underline{s}<\overline{s}$) by picking the maximal cost function for information outcomes such that
$C^t_\mu(\Pr(\mathrm{Y}|\underline{\omega},\, \underline{p}^t),\, \Pr(\mathrm{Y}|\overline{\omega},\, \underline{p}^t))=\mu(\underline{\omega})z_t(\underline{p}^t-u(\underline{\omega}))$, $C^t_\mu(0,\,1)=\mu(\underline{\omega})z_t(\underline{p}^t-u(\underline{\omega}))+\mu(\underline{\omega})(1-z_t)(p^t_{i+1}-u(\underline{\omega}))$, and $C^t_\mu(\Pr(\mathrm{Y}|\underline{\omega},\, p^t_{j+1}),\, \Pr(\mathrm{Y}|\overline{\omega},\, p^t_{j+1}))=\mu(\overline{\omega})z_t(u(\overline{\omega})-\overline{p}^t)$. This ensures that these information outcomes are optimal at each price. I can then create a measure of informativeness c_t that generates these costs with $c_t(\mu(\overline{\omega}))=0$, and I can do so with only kinks at the prior and the two interior posteriors that are reached, so its convexity is implied by the optimality of the posteriors reached by the type (notice that each type has behavior that satisfies the sufficient conditions from Theorem 2). ■

For each type t, define the expected **gross payoff for type** t when they choose an information outcome $s_t(p)\in S$ at price p to be:

$$\sum_{\omega\in\Omega}\mu_t(\omega)s_t(\omega,\,p)(u(\omega)-p).$$

Define the **gross payoff of the average information outcome** at a price p to be:

$$\sum_{\omega\in\Omega}\mu(\omega)\Pr(\mathrm{Y}|\omega,\,p)(u(\omega)-p).$$

Then, notice that, if behavior is rationalized by an aggregate costly learning model, the average (across types) gross payoff at a price is equal to the gross payoff of the average information outcome at the price given **(ii)** in the definition of behavior being rationalized by an aggregate costly learning model:

$$\sum_{t=1}^{T} \pi_t \Big(\sum_{\omega \in \Omega} \mu_t(\omega) s_t(\omega,\, p)(u(\omega) - p) \Big) = \sum_{\omega \in \Omega} \mu(\omega) \Pr(\mathrm{Y}|\omega,\, p)(u(\omega) - p).$$

Theorem 5. Given a prior belief μ and a finite and non-empty set of prices $\mathcal{P}$, the behavior is rationalized by an aggregate costly learning model if it satisfies the following three properties, and only if it satisfies properties **(ii)** and **(iii)**. Further, these conditions are all necessary for the behavior being rationalized by an aggregate costly learning model if there are only two states of the world.

(i) There is an event in which Y is more likely to be selected if some learning is done: $\exists B \in \sigma(\Omega) \backslash \{\emptyset, \Omega\}$ such that $\forall\, p \in \mathcal{P}$ with $s(p) \in S^L$: $\Pr(\mathrm{Y}|B,\, p) > \Pr(\mathrm{Y}|B^c,\, p)$.

(ii) Y is weakly less likely to be selected when its price increases: $\Pr(\mathrm{Y}|p)$ is weakly decreasing in p.

(iii) If there is a price at which $\neg$Y and Y are randomized over without learning, then there is no learning at any price, and $\neg$Y or Y are selected with probability one at all other prices: if $\exists p \in \mathcal{P}$ such that $\Pr(\mathrm{Y}|\omega_1,\, p) = \cdots = \Pr(\mathrm{Y}|\omega_N,\, p) \in (0,\, 1)$, then $\forall \tilde{p} \in \mathcal{P} \backslash p$: $\Pr(\mathrm{Y}|\tilde{p}) \in \{0,\, 1\}$.

Proof of Theorem 5. To see why **(i)** is necessary if there are only two states of the world: if $\exists p \in \mathcal{P}$ with $\Pr(\mathrm{Y}|\underline{\omega},\, p) > \Pr(\mathrm{Y}|\overline{\omega},\, p)$, then all types could employ an alternate information outcome $s_t(p) = (\Pr(\mathrm{Y}|p),\, \Pr(\mathrm{Y}|p))$, but this then raises the gross payoff of the average information outcome at p, which is not possible as this then must strictly increase the gross payoff of a type t while weakly reducing the cost of their learning. The necessity of **(ii)** follows from Theorem 1 as it holds for each type. To show the necessity of **(iii)**: suppose $\exists p \in \mathcal{P}$ such that: $\Pr(\mathrm{Y}|\omega,\, p) = z \in (0,\, 1)$ for all $\omega \in \Omega$. Notice that the gross payoff of the average information outcome at price

p must be zero:

$$\sum_{\omega\in\Omega}\mu(\omega)\Pr(\mathrm{Y}|\omega,\, p)(u(\omega)-p)=0\Rightarrow\sum_{\omega\in\Omega}\mu(\omega)u(\omega)=p,$$

because otherwise I could strictly increase the gross payoff of the average information outcome by replacing the information outcome of each type with (depending on the sign of the inequality between 0 and the gross payoff of the average information outcome) either $s_t(p)=(0,\ldots,0)$ or $s_t(p)=(1,\ldots,1)$, which again is not possible as this then must strictly increase the gross payoff of a type t while weakly reducing the cost of their learning. But, then replacing the information outcome of each type with $s_t(p)=(0,\ldots,0)$ does not reduce the gross payoff of the average information outcome, and since the change cannot be strictly increasing the gross payoff of any type, it means the change does not change the gross payoff of any type. Further, replacing the information outcome of each type with $s_t(p)=(1,\ldots,1)$ does not reduce the gross payoff of the average information outcome, and since the change cannot be strictly increasing the gross payoff of any type, it means the change does not change the gross payoff of any type. The optimality of their initial information outcome then implies that each type t has $\mu_t(\omega)=\mu(\omega)$ for each $\omega\in\Omega$, and for each $z_t\in[0,\,1]$ picking $\Pr_t(\mathrm{Y}|\omega,\,p)=z_t$ for each $\omega\in\Omega$ is optimal since the gross payoff of the average information outcome at price p is zero, and Theorem 1 then establishes that at all other prices no type learns and they all pick the same of (depending on the price) Y or $\neg$Y with probability one.

The fact that **(i)**, **(ii)**, and **(iii)** are sufficient together is established by Theorem 1, as a single type can rationalize the data when the three points are satisfied. ■

References

Acharya, S., & Wee, S. L. (2020). Rational inattention in hiring decisions. *American Economic Journal: Macroeconomics*, *12*(1), 1–40.

Ambuehl, S. (2017). An offer you can't refuse? incentives change how we inform ourselves and what we believe. *CESifo Working Papers*(6296).

Blackwell, D. (1951). Comparison of experiments. In *Proceedings of the second berkeley symposium on mathematical statistics and probability* (Vol. 2, pp. 93–103).

Blackwell, D. (1953). Equivalent comparisons of experiments. *The annals of mathematical statistics*, 265–272.

Bloedel, A. W., & Zhong, W. (2021). The cost of optimally-acquired information. *Unpublished Manuscript, November*.

Brown, Z. Y., & Jeon, J. (2024). Endogenous information and simplifying insurance choice. *Econometrica*, *92*(3), 881–911.

Caplin, A., & Dean, M. (2013). *Behavioral implications of rational inattention with shannon entropy* (Tech. Rep.). National Bureau of Economic Research.

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

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

Caplin, A., & Martin, D. (2015). A testable theory of imperfect perception. *The Economic Journal*, *125*(582), 184–202.

Cheng, X., & Kim, Y. (2025). On the monotonicity of information costs.

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *American economic review*, *99*(4), 1145–77.

Dasgupta, K., & Mondria, J. (2018). Inattentive importers. *Journal of International Economics*, *112*, 150–165.

Dean, M., & Neligh, N. (2023). Experimental tests of rational inattention. *Journal of Political Economy*, *131*(12), 3415–3461.

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

Hébert, B., & Woodford, M. (2023). Rational inattention when decisions take time. *Journal of Economic Theory*, *208*, 105612.

Heberton Jr, E. (1967). The law of demand–the roles of gregory king and charles davenant. *The Quarterly Journal of Economics*, *81*(3), 483–492.

Huberman, G. (2001). Familiarity breeds investment. *The Review of Financial Studies*, *14*(3), 659–680.

Lipnowski, E., & Ravid, D. (2023). Predicting choice from information costs. *arXiv preprint arXiv:2205.10434*.

Maćkowiak, B., Matějka, F., & Wiederholt, M. (2023). Rational inattention: A review. *Journal of Economic Literature*, *61*(1), 226–273.

Matějka, F., & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, *105*(1), 272–98.

Mensch, J. (2018). Cardinal representations of information. *Available at SSRN 3148954*.

Morris, S., & Strack, P. (2019). The wald problem and the relation of sequential sampling and ex-ante information costs.

Pomatto, L., Strack, P., & Tamuz, O. (2023). The cost of information: The case of constant marginal costs. *American Economic Review*, *113*(5), 1360–1393.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shue, K., & Luttmer, E. F. (2009). Who misvotes? the effect of differential cognition costs on election outcomes. *American Economic Journal: Economic Policy*, *1*(1), 229–57.

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

Taubinsky, D., & Rees-Jones, A. (2018). Attention variation and welfare: theory and evidence from a tax salience experiment. *The Review of Economic Studies*, *85*(4), 2462–2496.

Walker-Jones, D. (2026, July). *Heterogeneous costs and the decision to learn.*