Rational Inattention and the Ignorance Equivalent[*] PRELIMINARY AND INCOMPLETE

Roc Armenter[†]
Michèle Müller-Itten[‡]
Zachary R. Stangebye[§]

February 15, 2019

Abstract

We demonstrate that a wide class of Rational Inattention (RI) problems can be re-cast as a set of geometric orthogonality conditions. This approach allows us to separate the role of information costs and payoffs from that of prior beliefs. It also allows us to characterize the optimal conditional choice by what we call the ignorance equivalent action (IEA) of a menu. We use the IEA to derive novel comparative statics for changes in the menu. Finally, it serves as the foundation for our new, highly efficient numerical algorithm for the solution of these models: Our algorithm provides accurate solutions orders of magnitudes faster than current techniques.

1 Introduction

Rational Inattention (RI) has gained a lot of steam in recent years as a means to consolidate the predictions of rational choice theory with actual

[*] The views expressed in this paper are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia, the Federal Reserve Bank of New York, or the Federal Reserve System.

[†] Federal Reserve Bank of Philadelphia
[‡] University of Notre Dame
[§] University of Notre Dame

human decision making. The main idea is that costly information acquisition and/or processing may cause an agent to behave *as if* she weren't aware of all menu options. Shannon entropy is by far the dominant way of modeling these information costs (Sims (2003), . . .). In this paper, we show that this particular cost structure admits a geometric representation that offers new insights into RI behavior and admits a fast algorithm that can numerically solve complex instances, expanding the horizons for applied work.

At the heart of our paper is a geometric transformation that recasts the utility maximization of a rationally inattentive decision maker as a simple orthogonality condition. The transformation proceeds in two basic steps: In the first, we reduce the dimensionality of the RI problem via a substitution of the necessary first-order conditions. In the second, we demonstrate that this problem is isomorphic to a convex maximization problem whose solution can be interpreted as the construction of a particular (and unique) separating hyperplane.

Our formulation reveals a number of interesting results. First, it corroborates the finding of Caplin and Dean (REF) that there will always be fewer actions chosen with positive probability than there are states. Second, it cleanly separates the role played by prior beliefs over states from the state-dependent payoffs of individual actions: For instance, we find that marginal changes in prior beliefs will only affect the distribution over actions with positive mass rather than altering the set of such actions. Changes in payoff or information costs, though, will generally affect both.

Most interestingly, our approach allows us to construct an 'ignorance equivalent' action for any given menu of actions, much in the way risk aversion admits the construction of a 'certainty equivalent' bundle for a monetary gamble. The ignorance equivalent action (IEA) is an action with state-dependent payoffs that, if added to the existing menu, makes the decision maker no better off, but renders her exactly indifferent between blindly selecting that action and conditioning her choice on the costly signal she designed in the RI problem.

To see this concretely, consider a rationally inattentive agent shopping for health insurance. She can choose from many different plans that vary with regard to her payoffs across different states. These payoffs depend both on her own health and the esoteric details of the plan. She can take costly steps to acquire more information both about her own health and the details of these plans to inform her decision, which is the RI problem. But our formulation suggests that there is an alternative plan that she would choose immediately without being any worse off and for which she would forego all of this costly information gathering. This alternate plan is the IEA, and while it is generally not part of the original consideration set, it is unique and has useful properties.

We discuss comparative statics that advocate for the usefulness of this measure as a more parsimonious way to compare across menus. We show that the ignorance equivalent is always unique, and that this uniqueness translates to the the optimal conditional choice over actions for almost all menus.

Another main contribution is that the idea of the ignorance equivalent translates readily into a distance notion across conditional choices. To our knowledge, such a notion is currently absent from the literature. This makes it hard to compare solutions or evaluate the numerical precision of an algorithmic output. Convergence in our distance notion relies solely on the marginal probabilities, but implies that the state-dependent utility converges in distribution.

In the second part of the paper, we derive a numerical algorithm from the geometric intuition, and show that it outperforms the existing methods both in terms of speed and precision. In the third part of the paper, we illustrate the applicability of our theory with some practical examples. First, we point out how a researcher can use knowledge of the full menu (rather than just the subset of chosen actions) to learn about the agent's preference. We do so by showing that the confidence interval is reduced if the researcher is aware that he can learn from chosen *as well* as unchosen actions, an approach that could strengthen the takeaways from empirical studies. Second, we revisit some famous results from the theoretical literature with simpler, geometric proofs. Third, we focus on interval menus and show additional implications that hold for the most commonly used parametrized families of utility functions.

2 Geometric Representation

2.1 Model Setup

An agent implements an action from the (possibly uncountable) **menu** $\mathcal{A}$. She faces uncertainty regarding the (payoff-relevant) state of the world, modeled as a random variable $\mathcal{S} : \Omega \to \{1, ..., I\}$ with finite support.[1] We denote

[1] The set Ω refers to an underlying probability space that allows us to express correlations between the agent's choice and the state realization.

the probability mass function of s in vector form, $\boldsymbol{\pi} \in \Delta^I$, $\pi_i = \mathsf{Pr}(\mathcal{S} = i)$, and refer to it as the **prior**. The realized payoff from action a in state i is written $u_i(a) = u(a,i)$. We assume that the set of vectors $\boldsymbol{u}(\mathcal{A})$ is closed and bounded, and that no two actions are payoff equivalent.[2] The agent can condition the actions on a costly signal, and thus partially tailor her choice to the realized state. More accurate signal structures are more costly, and we follow the rational inattention literature (Sims, 2003, 2006) in focusing on entropy-based information-processing costs. The agent's joint choice over the signal structure and the conditionally implemented action can be recast as the choice of a random variable $\mathcal{C} : \Omega \to \mathcal{A}$ that is correlated with $\mathcal{S}$ (Matějka and McKay, 2015, Corollary 1). The choice $\mathcal{C}$ is optimal if and only if it maximizes expected utility net of information processing costs, measured as the average reduction in entropy between prior and posterior. These costs are also known as the mutual information $\mathcal{I}(\mathcal{C}, \mathcal{S})$ (Cover and Thomas, 2012) and capture the idea that it is costly to tailor the choice $\mathcal{C}$ closely to the realized state $\mathcal{S}$. The relative importance of this **information cost** is governed by a proportionality constant $\lambda > 0$. Formally,

$$V(\mathcal{A}) = \begin{cases} \max_{\mathcal{C}} & \mathsf{E}\left[u(\mathcal{C}, \mathcal{S})\right] - \lambda \mathcal{I}(\mathcal{C}, \mathcal{S}) \\ \text{s.t.} & \mathcal{C} : \Omega \to \mathcal{A}. \end{cases} \tag{P1}$$

Since we assume a finite state space, Jung, Kim, Matejka, and Sims (2015) show that the agent will only implement actions from a finite **consideration set** $A = \mathsf{support}(\mathcal{C}) \subset \mathcal{A}$. In this case, both $\mathcal{C}$ and $\mathcal{S}$ are discrete, and mutual information can be written as

$$\mathcal{I}(\mathcal{C}, \mathcal{S}) := \sum_{i=1}^{I} \sum_{a \in A} \mathsf{Pr}(\mathcal{C} = a, \mathcal{S} = i) \ln\left(\frac{\mathsf{Pr}(\mathcal{C} = a, \mathcal{S} = i)}{\mathsf{Pr}(\mathcal{C} = a)\, \mathsf{Pr}(\mathcal{S} = i)} \right).$$

Borrowing from utility theory, we refer to the optimal objective value in Problem (P1) as the agent's 'indirect utility' from an menu $\mathcal{A}$. We denote the induce preference over menus as $\succsim_{\mathcal{A}}$.

The first order conditions of (P1) imply an explicit relationship between the marginal distribution of $\mathcal{C}$ and the joint distribution of $\mathcal{S}$ and $\mathcal{C}$ (Matějka and McKay, 2015, Lemma 2), given by

$$\mathsf{Pr}(\mathcal{C} = a, \mathcal{S} = i) = \frac{\mathsf{Pr}(\mathcal{S} = i)\, \mathsf{Pr}(\mathcal{C} = a)\, e^{u_i(a)/\lambda}}{\sum_{\tilde{a} \in A} \mathsf{Pr}(\mathcal{C} = \tilde{a})\, e^{u_i(\tilde{a})/\lambda}} \quad \forall a \in \mathcal{A}, \forall i = 1, ..., I. \tag{1}$$

[2] Formally, $\boldsymbol{u}(a) = \boldsymbol{u}(a')$ if and only if $a = a'$.

It is thus possible to simplify Problem (P1) by referring only to the marginal distribution of $\mathcal{C}$ and $\mathcal{S}$,[3]

$$V(\mathcal{A}) = \begin{cases} \max_{\mathcal{C}} & \lambda \mathsf{E}_{\mathcal{S}} \left[\ln \left(\mathsf{E}_{\mathcal{C}} \left[e^{u(\mathcal{C},\mathcal{S})/\lambda} \right] \right) \right] \\ \text{s.t.} & \mathcal{C} : \Omega \to \mathcal{A}. \end{cases} \tag{P2}$$

Our innovation lies in further transforming problem (P2) to allow for a simple geometric interpretation, while maintaining an explicit characterization of the agent's choice. To do so, we make extensive use of the mapping $\boldsymbol{\beta} : \mathcal{A} \to \mathbb{R}^I_+ := (0, \infty)^I$ defined coordinate-wise as $\beta_i(a) = e^{u_i(a)/\lambda}$, and study the optimization problem

$$W(\mathcal{A}) = \begin{cases} \max_{\boldsymbol{b}} & w(\boldsymbol{b}) := \sum_{i=1}^{I} \pi_i \ln(b_i) \\ \text{s.t.} & \boldsymbol{b} \in \mathcal{B} := \mathsf{conv.hull}(\{\boldsymbol{\beta}(a) \mid a \in \mathcal{A}\}). \end{cases} \tag{P3}$$

The following theorem shows that there is a one-to-one mapping between the optimal solutions to Problems (P2) and (P3): The weights in the convex representation of $\boldsymbol{b}^*$ over $\boldsymbol{\beta}(\mathcal{A})$ represent the optimal marginal probabilities $\mathsf{Pr}(\mathcal{C}^* = a)$. Moreover, $W = \frac{1}{\lambda} V$ is a positive affine transformation of the indirect utility V, and hence represent the same preference over menus.

Theorem 1 (Optimality conditions)**.** *The conditional choice $\mathcal{C}$ solves Problem (P2) if and only if $\boldsymbol{b}^* = \sum_{a \in \mathsf{support}(\mathcal{C})} \mathsf{Pr}(\mathcal{C} = a)\, \beta(a)$ solves (P3).*

Moreover, the solution to Problem (P3) is unique and fully characterized by either of the following two equivalent optimality conditions:

(a) $\boldsymbol{\nabla} w(\boldsymbol{b}^*) \cdot \boldsymbol{b} \leq 1$ for all $\boldsymbol{b} \in \mathcal{B}$.

(b) $\boldsymbol{\nabla} w(\boldsymbol{b}) \cdot \boldsymbol{b}^* \geq 1$ for all $\boldsymbol{b} \in \mathcal{B}$.

Proof. See `OptimalityConditions.tex` for details. □

There are two important advantages of our approach: First, (P3) is convex thanks to its strictly concave objective function $w : \mathbb{R}^I_+ \to \mathbb{R}^I_+$ over the convex domain $\mathcal{B}$. As such, (P3) always admits a unique solution $\boldsymbol{b}^*$ as depicted in Fig. 1. Its weights over the spanning points in $\boldsymbol{\beta}(\mathcal{A})$ identify the (not necessarily unique) optimal marginals of $\mathcal{C}$. Second, this formulation separates the role of prior beliefs ($\boldsymbol{\pi}$) and utility parameters (u, λ). The

[3] For clarity, note that $\mathsf{E}_{\mathcal{C}}[f(\mathcal{C}, s)]$ is the *marginal* expectation $\sum_{a \in A} \mathsf{Pr}(\mathcal{C} = a)\, f(\mathcal{C}, s)$ rather than the *conditional* expectation $\sum_{a \in A} \mathsf{Pr}(\mathcal{C} = a | \mathcal{S} = s)\, f(\mathcal{C}, s)$.

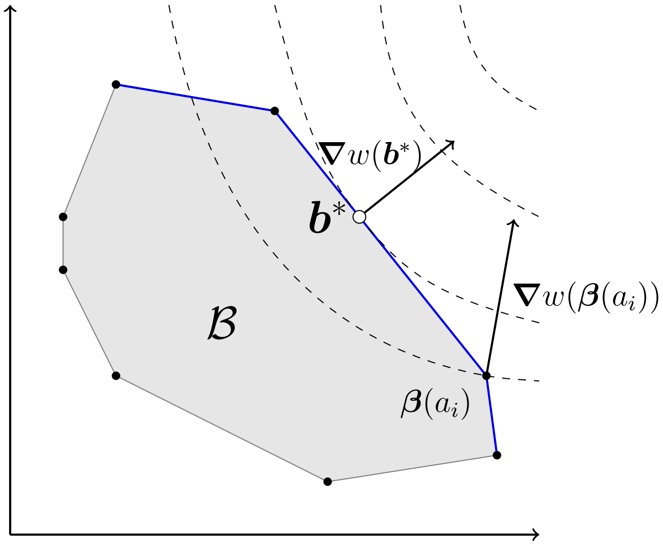

$\nabla w(\boldsymbol{b}^*)$
$\boldsymbol{b}^*$
$\nabla w(\boldsymbol{\beta}(a_i))$
$\mathcal{B}$
$\boldsymbol{\beta}(a_i)$

Figure 1: Geometric mapping. The set $\boldsymbol{\beta}(\mathcal{A})$ is the union of the black dots, the set $\mathcal{B} = \mathsf{conv.hull}(\boldsymbol{\beta}(\mathcal{A})$ is the gray shape. $\mathcal{B}$'s upper boundary $\partial^+\mathcal{B}$ is indicated in blue. Indifference curves for w are drawn as dashed lines.

former affect the value w of different points $\boldsymbol{b} \in \mathbb{R}^I_+$, the latter affect the feasible set $\mathcal{B}$. This separation facilitates intuition regarding changes in the choice environment and simplifies subsequent proofs.

Problem (P3) readily generates an upper bound on the size of the minimal consideration set. Indeed, Carathéodory's Theorem (Eggleston, 1958, Theorem 18) states that any point $\boldsymbol{b} \in \mathsf{conv.hull}(\boldsymbol{\beta}(\mathcal{A})) \subset \mathbb{R}^I$ can be written as a convex combination of at most $I + 1$ points in $\boldsymbol{\beta}(\mathcal{A})$. The minimal consideration set therefore contains at most $|A| \leq I + 1$ actions.[4] This complements Jung, Kim, Matejka, and Sims (2015)'s treatment of more general state distributions with a simple intuition for finite state spaces. For brevity, we will henceforth write convex combinations over $\mathcal{B}$ as $\sum_{a\in\mathcal{A}} w_a\boldsymbol{\beta}(a)$ even in the case of uncountable $\mathcal{A}$; with the understanding that only a finite number of the terms are positive. If $\boldsymbol{\beta}(\tilde{a}) = \sum_a w_a\boldsymbol{\beta}(a)$, we say that $\tilde{a}$ is a **$\boldsymbol{\beta}$-transformed convex combination** over actions in $\mathcal{A}$. We call a solution $\mathcal{C}$ to Problem (P2) **unique** if the weights w_a, and hence the marginals $\mathsf{Pr}(\mathcal{C} = a)$ and conditionals $\mathsf{Pr}(\mathcal{C} = a, \mathcal{S} = i)$, are uniquely determined.

[4] The bound can be tightened to $|A| \leq I$ by making reference to the hyperplane $\nabla w(\boldsymbol{b}^*)\cdot \boldsymbol{b} = 1$ (see Lemma 2 in the appendix).

Problem (P3) is robust to arbitrary scaling of the axes. Formally, we call an optimization problem **scale-invariant** if and only if scaling all feasible points component-wise by a positive vector $\boldsymbol{k} \in \mathbb{R}^I_+$ also scales the optimum by $\boldsymbol{k}$. In our situation, this boils down to scaling the spanning points of the convex hull by replacing $\boldsymbol{\beta}$ with the component-wise product $\boldsymbol{k} \circ \boldsymbol{\beta}$. Doing so merely adds a constant term $\sum_i \pi_i \ln(k_i)$ to the value of any convex combination, and hence does not affect the location of the optimum.

Corollary 1 (Axis Scaling)**.** *Problem (P3) is scale-invariant.*

The fact that the optimum scales one-to-one with the spanning points is a direct consequence of the entropy-based information costs. This scalability greatly helps avoid numerical imprecision (see Section 3). It is also what unites our approach with previous convexification procedures (Kamenica and Gentzkow, 2011; Gentzkow and Kamenica, 2014; Caplin, Dean, and Leahy, 2018). Indeed, scaling by $\hat{\boldsymbol{k}} = \boldsymbol{\nabla} w(\boldsymbol{b}^*) \in \mathbb{R}^I_+$ moves the optimum to $\boldsymbol{\pi}$ by the optimality conditions Theorem 1(a). To see why, note that $\boldsymbol{\nabla} w(\boldsymbol{\pi}) = \mathbf{1}$ and hence

$$\nabla_i w(\boldsymbol{b}^*) \cdot \boldsymbol{b} \leq 1 \qquad \text{for all } \boldsymbol{b} \in \mathcal{B} = \mathsf{conv.hull}(\boldsymbol{\beta}(\mathcal{A}))$$

is equivalent to

$$\boldsymbol{\nabla} w(\boldsymbol{\pi}) \cdot \tilde{\boldsymbol{b}} = \underbrace{\boldsymbol{\nabla} w(\boldsymbol{b}^*)}_{\hat{\boldsymbol{k}}} \cdot \boldsymbol{b} \leq 1 \qquad \text{for all } \tilde{\boldsymbol{b}} = \hat{\boldsymbol{k}} \circ \boldsymbol{b} \in \mathsf{conv.hull}(\hat{\boldsymbol{k}} \circ \boldsymbol{\beta}(\mathcal{A})).$$

In other words, it is possible to scale the axes so that the optimum lies in the simplex Δ^I. The convex representation of $\boldsymbol{\pi} = \sum_{a \in A} p(a) \hat{\boldsymbol{k}} \circ \boldsymbol{\beta}(a)$ partially characterizes the optimal posterior distribution: With probability $p(a)$, the expected utility under the posterior is maximized at action $a \in A$, and the distribution is Bayes-plausible since the expected posterior belief equals the prior. While scaling to the simplex offers more intuition regarding the possible posterior beliefs, it has one obvious draw-back: One needs to know the optimal $\boldsymbol{b}^*$ in order to determine the scaling of the implicit menu (and hence the posteriors). For conceptualization, this loop is not tragic – for computation, it is fatal. Our approach yields a convex optimization problem where both the set of candidate points $\mathcal{B}$ and the objective function w are explicitly defined. This clarifies at once the effect of changes in the menu (see Theorem 2 and Corollary 3), and readily lends itself to efficient algorithmic implementation (see Section 3).

2.2 Ignorance Equivalent

The convexity of Problem (P3) allows us to define what we call the **ignorance equivalent** of a menu $\mathcal{A}$, which is a (possibly new) action with state-dependent payoffs for which the agent would forgo all learning.

Definition 1. *The* ***ignorance equivalent*** *of a menu $\mathcal{A}$ is defined as an action α with state-dependent payoffs such that the decision maker is indifferent between the three menus $\mathcal{A}$, $\mathcal{A} \cup \{\alpha\}$ and $\{\alpha\}$.*

We can rephrase this definition equivalently in terms of the indirect utilities $V(\mathcal{A}) = V(\mathcal{A} \cup \{\alpha\}) = V(\{\alpha\})$, or $W(\mathcal{A}) = W(\mathcal{A} \cup \{\alpha\}) = W(\{\alpha\})$.[5] The notion is analogous to the concept of a certainty equivalent for a lottery π over money, defined as the amount $X \in \mathbb{R}$ of a riskless payment that renders the decision maker exactly indifferent to π. The certainty equivalent for each lottery is unique, and it decreases with the agent's risk-aversion.[6] Similarly, the ignorance equivalent α is a single action with state-dependent payoffs which renders the decision maker indifferent between the original menu $\mathcal{A}$, the augmented set $\mathcal{A} \cup \{\alpha\}$, and the singleton $\{\alpha\}$. We call it 'ignorance equivalent' because it compares menus where the agent can condition her action on a costly signal, with one where action α is implemented blindly. The ignorance equivalent is unique for any menu $\mathcal{A}$. Its expected payoff decreases with the agent's information cost and weakly increases as new actions are added to $\mathcal{A}$. While its state-wise payoffs do not always respond monotonically to changes in the prior, a weaker condition holds on marginal changes.

Theorem 2 (Ignorance equivalent)**.** *The unique ignorance equivalent for menu $\mathcal{A}$ yields payoff $u_i(\alpha) = \beta_i^{-1}(b_i^*) = \lambda \ln b_i^*$ in state i, where $\boldsymbol{b}^*$ is the unique solution to Problem (P3). It has the following comparative statics:*

(i) The expected payoff $\boldsymbol{\pi} \cdot \boldsymbol{u}(\alpha)$ is decreasing in the information cost λ and weakly increasing in the addition of new actions.

(ii) For a generic problem, a marginal change in prior $d\boldsymbol{\pi}$ can be decomposed into two orthogonal components $d\boldsymbol{\pi} = d\boldsymbol{\pi}^{\perp} + d\boldsymbol{\pi}^{\parallel}$ with

$$d\boldsymbol{\pi}^{\perp} \cdot d\boldsymbol{\pi}^{\parallel} = 0 \quad \textit{and} \quad \boldsymbol{\pi}^{\perp} \cdot \mathbf{1} = d\boldsymbol{\pi}^{\parallel} \cdot \mathbf{1} = 0, \tag{2}$$

[5] Since a maximization problem is non-decreasing in the set of available options, an equivalent but more parsimonious definition requires only $\mathcal{A} \succsim_{\mathcal{A}} \{\alpha\} \succsim_{A} \mathcal{A} \cup \{\alpha\}$.

[6] Assuming, as is standard, that the agent strictly prefers lotteries that unconditionally return higher amounts of money.

such that the state-wise payoffs $\boldsymbol{u}(\alpha)$ don't respond to $d\boldsymbol{\pi}^{\perp}$ and respond monotonically to $d\boldsymbol{\pi}^{\parallel}$ for almost all $\boldsymbol{\pi}$, meaning

$$u_i(\alpha^{\boldsymbol{\pi}}) \geq u_i(\alpha^{\boldsymbol{\pi}+d\boldsymbol{\pi}}) \quad \Longleftrightarrow \quad d\boldsymbol{\pi}_i^{\parallel} \geq 0,$$

where $\alpha^{\boldsymbol{\pi}}$ denotes the ignorance equivalent under prior $\boldsymbol{\pi}$.

Proof. To be completed. □

It is important to point out that the optimal choice always involves learning unless α is payoff-equivalent to an action in $\mathcal{A}$. The mere fact that $\boldsymbol{b}^*$ can be written as a convex combination $\sum_{a\in A} p(a)\boldsymbol{\beta}(a)$ over a finite subset $A \subseteq \mathcal{A}$ does *not* mean that the decision maker implements an unconditional lottery over the consideration set A.[7] Instead, the decision maker generates indirect utility $W(\{\alpha\})$ by implementing the optimal joint distribution between $\mathcal{C}$ and $\mathcal{S}$ according to Eq. (1). The amount of learning is dictated by the geometric distance between $\boldsymbol{\beta}(\alpha)$ and its spanning points on the convex hull. In colloquial terms, if the implemented actions in $\mathcal{A}^1$ are 'more similar' than those in $\mathcal{A}^2$, then the choice of actions in the former is more informative regarding the underlying state of the world than in the latter.

Corollary 2. *Assume menus $\mathcal{A}^1$ and $\mathcal{A}^2$ have the same ignorance equivalent α under information cost λ, and that the optimal choices $\mathcal{C}^1$ and $\mathcal{C}^2$ are unique. If $\boldsymbol{\beta}(\mathsf{support}(\mathcal{C}^1)) \subseteq \mathsf{conv.hull}(\boldsymbol{\beta}(\mathcal{A}^2))$, then $\mathcal{I}(\mathcal{C}^1, \mathcal{S}) \leq \mathcal{I}(\mathcal{C}^2, \mathcal{S})$.*

Proof. See Appendix A.1. □

Even when the ignorance equivalent action α is not part of the underlying menu $\mathcal{A}$, this 'synthetic' action still has intuitive meaning, as shows the following stylized example:

Example. Anna is shopping for health insurance. The various plans $a \in \mathcal{A}$ differ across several attributes: premiums, deductibles, coverage, lifetime maxima, provider networks. By letting $\mathcal{A} \subset \mathbb{R}^I$, we let a_i denote the intensity of attribute i in action a. Anna's utility from a health plan will depend on her (as of yet unknown) health status during the coming year. For simplicity, we assume that in each state of the world i, her utility depends on a distinct attribute $u_i(a) = a_i$. The prior π_i describes the ex-ante likelihood of health

[7] To emphasize, note that the state-wise payoffs of any such non-degenerate lottery would be strictly lower than those of ignorance equivalent, since $\sum_{a\in A} p(a)\boldsymbol{u}_i(a) < u_i(\alpha)$ by concavity of the logarithm.

status i. Anna can learn about her future health status through checkups and tests, but doing so is costly. Optimally, she thus engages in limited learning and purchases a plan based on her findings.

The ignorance equivalent describes a unique (and possibly inexistent) insurance plan $\alpha \in \mathbb{R}^I$. Adding this plan to the lineup does not make Anna any better off, but she would now buy α without any further learning. This in turn is attractive for insurers. Indeed, note that in a zero-sum game between clients and insurers,[8] an insurer makes maximal profits by offering Anna plan α. In essence, this allows the firm to capture a "learning rent", which is profitable as long as designing plan α is cheap compared to Anna's own health inquiry costs. $\diamondsuit$

For non-marginal changes, the comparative statics often depend on the exact geometry of $\mathcal{B}$. Two observations follow readily from Theorem 1 however: First, the full menu $\mathcal{A}$ has additional 'option value' relative to the ignorance equivalent α, in so far as it leaves more opportunities for learning when a new action a^+ is added to the menu. This is similar to how a decision maker may be indifferent between an isolated lottery and its certainty equivalent, but weakly prefers the lottery in more complex portfolio management problems because it may cancel out some aggregate uncertainty. Second, the decision maker benefits from the addition of a new action to a menu $\mathcal{A}$ if and only if she also benefits when that action is added to the ignorance equivalent of $\mathcal{A}$. In other words, the ignorance equivalent can answer the question of whether a^+ improves the complex menu $\mathcal{A}$. It can however only give a lower bound for the absolute strength of that improvement.

Corollary 3. *Let α denote the ignorance equivalent of menu $\mathcal{A}$. The addition of a new action a^+ enhances menu $\mathcal{A}$ weakly more than menu $\{\alpha\}$, but it either enhances both menus or neither. Formally,*

(i) $W(\mathcal{A} \cup \{a^+\}) \geq W(\{\alpha\} \cup \{a^+\})$,

(ii) $W(\mathcal{A} \cup \{a^+\}) > W(\mathcal{A})$ *if and only if* $W(\{\alpha\} \cup \{a^+\}) > W(\{\alpha\})$.

Proof. The first part is immediate since $\boldsymbol{\beta}(\{\alpha, a^+\})$ is included in the convex hull generated by $\boldsymbol{\beta}(\mathcal{A} \cup \{a^+\})$. Second, the addition of a^+ has no impact on the indirect utility if and only if $\boldsymbol{\beta}(\alpha)$ remains optimal in the menus $\mathcal{A} \cup \{a^+\}$ or $\{\alpha, a^+\}$ respectively. The optimality condition for either menu is the same by Theorem 1(a), namely $\boldsymbol{\nabla} w(\boldsymbol{\beta}(\alpha)) \cdot \boldsymbol{\beta}(a^+) \leq 1$. $\square$

[8] In a zero-sum game, Anna's enrollment in plan a costs the insurer $\boldsymbol{\pi} \cdot a$ in expectation.

2.3 Generic Uniqueness

An important open question is whether the optimal conditional choice $\mathcal{C}$ is typically unique when the menu is large. To our knowledge, there is a sizable gap between known necessary and sufficient conditions (Matějka and McKay, 2015): The necessary condition is that the ignorance equivalent α can be written as a unique $\boldsymbol{\beta}$-transformed convex combination over actions in $\mathcal{A}$. While the intuition is clear from Theorem 1, it is difficult to assess the class of menus $\mathcal{A}$ that satisfy this implicit property. The sufficient condition requires that all points in $\boldsymbol{\beta}(\mathcal{A})$ are linearly independent. Obviously, this immediately rules out all menus containing more than I actions. So, should we expect unique choices also from larger menus, or do these typically "overload" the decision maker with options until she is indifferent across several optimal choices?

We show that even rich finite menus typically result in a unique optimal choice. We call a finite problem $(\boldsymbol{u}(\mathcal{A}), \lambda, \boldsymbol{\pi}) \in \mathfrak{P} := \left(\bigcup_{n\in\mathbb{N}}(\mathbb{R}^I)^n\right) \times \mathbb{R}_+ \times \Delta^{I-1}$ **generic** if and only if the optimal choice $\mathcal{C}^*$ is unique and puts positive weight on all actions where the condition of Theorem 1(a) is binding. We choose that name because the notion is not very restrictive.

Theorem 3. *There exists a set $E \subseteq \mathfrak{P}$ of Lebesgue measure zero such that all problems $(\boldsymbol{u}(\mathcal{A}), \lambda, \boldsymbol{\pi}) \in \mathfrak{P} \setminus E$ are generic.*

Proof. We first show that if the weights for any point in the upper boundary $\partial^+\mathcal{B} = \{\boldsymbol{b} \in \mathcal{B} \mid \nexists\, \boldsymbol{b}' \in \mathcal{B} : \boldsymbol{b}' > \boldsymbol{b}\}$ are not unique, then $\boldsymbol{\beta}(\mathcal{A})$ contains an affinely dependent subset of cardinality at most $I+1$. Indeed, suppose there exists such a $\tilde{\boldsymbol{b}}$ along with its supporting hyperplane $\tilde{\boldsymbol{\psi}} \cdot \boldsymbol{b} = 1$. Let $\tilde{A} = \{a \in \mathcal{A} \mid \boldsymbol{\psi} \cdot \boldsymbol{\beta}(a) = 1\}$ denote all actions that map to this hyperplane. Clearly, $\tilde{\boldsymbol{b}} = \sum_{a\in\mathcal{A}} w(a)\boldsymbol{\beta}(a)$ can put weight only on actions in A^0. Furthermore, if A^0 admits a subset of cardinality $I+1$, this subset lives in an $I-1$ dimensional hyperplane and is therefore affinely dependent. For smaller sets, we rearrange $\tilde{\boldsymbol{b}} = \sum_{a\in A^0} w(a)\boldsymbol{\beta}(a) = \sum_{a\in A^0} \tilde{w}(a)\boldsymbol{\beta}(a)$ to state that $\sum_{a\in A^0}(w(a) - \tilde{w}(a))\boldsymbol{\beta}(a) = 0$. This establishes that A^0 itself is affinely dependent whenever there exist two distinct weights w and $\tilde{w}$.

For any fixed information cost λ (and hence a specific $\boldsymbol{\beta}$ mapping), the union of finitely many hyperplanes

$$\bigcup_{\substack{A\subset\mathcal{A}\\|A|\leq I}} \left\{ \sum_{a\in A} m(a)\boldsymbol{\beta}(a) \;\middle|\; m : A \to \mathbb{R}, \sum_{a\in A} m(a) = 1 \right\}$$

has Lebesgue measure zero for any finite menu $\mathcal{A}$. By induction, it therefore follows that no finite menu will have this property. Consequently, for any cost λ, almost all finite menus admit a unique solution no matter the prior $\boldsymbol{\pi}$.

□

2.4 Distance between conditional choices

Standard optimization methods typically judge convergence based on changes in objective value. Matlab solvers for instance rely on tolerances for first order conditions or differences in function value between iterations. However, since utility is a fundamentally relative measure, it is difficult to interpret these tolerance parameters. Two conditional choices may – and often do – share similar objective values even though their behavioral implications differ significantly. We therefore think that the (applied) Rational Inattention literature could benefit from a notion of 'similarity' across conditional choices, which lends itself to a more meaningful measure of numerical accuracy, and eventually could be incorporated into confidence intervals of estimated parameters.

Our geometric mapping offers an intuitive distance metric. Indeed, for any two conditional choices $\mathcal{C}, \mathcal{C}' : \Omega \to \mathcal{A}$, the Euclidean distance of the corresponding vectors in $\mathcal{B}$ is well defined as

$$d_{\boldsymbol{\beta}}(\mathcal{C}, \mathcal{C}') := \| \mathsf{E}\left[\boldsymbol{\beta}(\mathcal{C})\right] - \mathsf{E}\left[\boldsymbol{\beta}(\mathcal{C}')\right] \| = \left\| \sum_{a \in \mathcal{A}} [\mathsf{Pr}(\mathcal{C} = a) - \mathsf{Pr}(\mathcal{C}' = a)] \boldsymbol{\beta}(a) \right\| .$$

To ensure that the mapping $d_{\boldsymbol{\beta}}$ is a proper distance, we therefore restrict attention to choices that are optimal under *some* prior belief, which we call **rationalizable** and write as $\mathcal{C} \in C_r$.

Lemma 1. *In a generic problem, the set of rationalizable choices C_r, along with distance $d_{\boldsymbol{\beta}}$, describes a complete metric space.*

Proof. Still needs to be spelled out (and moved to appendix). The rationalizable choices map one-to-one into $\partial^+ \mathcal{B}$, and then the distance properties are essentially inherited directly from $\| \cdot \|$ over $\mathbb{R}^I$. □

Three key characteristics make $d_{\boldsymbol{\beta}}$ an appealing metric for conditional choices: First, it is parsimonious because it only considers marginal choice probabilities rather than the full joint distribution of $\mathcal{C}$ and $\mathcal{S}$. Second, it does not merely compare the selection probabilities among actions in a way that $d_{\mathsf{Pr}} := \|(\mathsf{Pr}(\mathcal{C} = a) - \mathsf{Pr}(\mathcal{C}' = a))_{a\in\mathcal{A}}\|$ would. This is important since numerical solutions over large menus typically identify both the consideration set *and* the marginal choice probabilities with some noise. The distance d_{Pr} fails to account for similarity across actions, but $d_{\boldsymbol{\beta}}$ weighs both types of errors appropriately. By this we refer to a third characteristic of $d_{\boldsymbol{\beta}}$, and the main result of this section: Convergence according to $d_{\boldsymbol{\beta}}$ implies not only that the expected payoffs *across* states converge, but also that the conditional choice converges in distribution *for each* individual state.

Theorem 4. *Consider a sequence of conditional choices $\{\mathcal{C}^n\}$ that satisfy the first order condition (1) over a generic problem. If there exists $C \in C_r$ such that $d_{\boldsymbol{\beta}}(\mathcal{C}^n, \mathcal{C}) \to 0$, then $(\mathcal{C}^n, \mathcal{S}^n) \xrightarrow{d} (\mathcal{C}, \mathcal{S})$.*

Proof. See Lemma 4 in the appendix. □

This is particularly relevant for numerical solution methods: Exact optimality cannot be guaranteed, but since (1) yields an explicit formula for the conditionals, we expect that any numerical noise will manifest itself in incorrect marginals. Theorem 4 implies that even in large menus with many similar actions, the state-wise payoff across two choices are close exactly when their $\boldsymbol{\beta}$ images are close.

Conjecture 1. *Given a point $\boldsymbol{b}^0 \in \partial^+\mathcal{B}$ and a supporting hyperplane satisfying $\boldsymbol{\psi} \cdot (\boldsymbol{b} - \boldsymbol{b}^0) = 0$, we can bound the $d_{\boldsymbol{\beta}}$ distance above by ....*

Proof. We may be able to compute in closed form the maximal distance between $\boldsymbol{b}^0$ and any point in $\{\boldsymbol{b} | w(\boldsymbol{b}) \geq w(\boldsymbol{b}^0) \text{ and } \boldsymbol{\psi} \cdot (\boldsymbol{b} - \boldsymbol{b}^0) \leq 0\}$. That problem is max $\|\boldsymbol{b} - \boldsymbol{b}^0\|$ □

The bound becomes zero at the optimum, so although it may be somewhat slack, it's a useful measure of accuracy. It is preferable to one that is merely based on successive iterations, because it is able to bound the distance between $\boldsymbol{b}^0$ and the true optimum $\boldsymbol{b}^*$, and thus does not fake accuracy when an algorithm merely stalls at a suboptimal level.

3 Algorithm Design

The optimality conditions of Theorem 1 lend themselves to a useful algorithm for solving general discrete RI problems or fine discrete approximations to continuous RI problems. The goal is to find the separating hyperplane that lies tangent to both $\mathcal{B}$ and the extremal $w(b)$. From this hyperplane, we can deduce immediately both which actions have positive weight and what those marginals are.

There are many possible algorithms that would work, but we will outline one that we have found to work efficiently and reliably across a wide class of problems. The algorithm proceeds iteratively in a two-step procedure. The first directs the solution to the boundary of $\mathcal{B}$, and the second ensures that it is the correct boundary.

In the first step, given any conjectured set of b's (actions), we project onto $\mathcal{B}$ using the $\tilde{b}$ that maximizes $w(b)$ in the convex hull of the conjectured b's. Call this b^*. Our updated set of b's include those that span b^*.

Second, in order to avoid being trapped on a sub-optimal facet of $\mathcal{B}$, we compute the hyperplane implied by the gradient of $\nabla w(\tilde{b})$; if there are points in $\mathcal{B}$ that lie outside of it then the solution cannot be optimal, and in fact can be improved upon by including those points. Consequently, we add at least one of those points and return to step one, repeating until convergence. It is not hard to see that this sort of procedure necessarily converges to the optimal solution, but we demonstrate this explicitly in Appendix ...

More carefully specified, the algorithm proceeds as follows. Before we begin the iterative procedure, we make use of Corollary 1 by scaling $\mathcal{B}$ until it is in the I-dimensional cube. We do this by choosing the positive vector $\boldsymbol{k}$ as follows: $\boldsymbol{k}_i = 1/(\max_{a\in\mathcal{A}} \beta_i(a))$. Denote this scaled set by $\hat{\mathcal{B}}$ and the image from actions into this set by $\hat{\beta}(a)$.

Begin the algorithm with a conjectured set of positive probability actions, $A_0 \subset \mathcal{A}$. In any step i, we proceed as follows

1. Solve for $p_i^* = \arg\max_{p\in\Delta_{|A_0|}} \; w(p\hat{\beta}(A_i))$. Define $\tilde{b}_i = p_i^*\hat{\beta}(A_i)$.

2. Solve for b_i^*, which is the largest such b along $\tilde{b}_i$ that is also in $\hat{\mathcal{B}}$. This can accomplished via the following linear program:

$$\max_{K\geq 0} \; K\tilde{b}_i$$
$$\text{s.t.} \quad K\tilde{b}_i \in \hat{\mathcal{B}}$$

where $b_i^* = K_i^* \tilde{b}_i$. Define A_i^- as the set of actions whose image, $\hat{\beta}(A_i^-)$, spans b_i^*.

3. Let $a_i^* = \arg\max_{a \in \mathcal{A}} \nabla w(\tilde{b}_i)' \hat{\beta}(a)$. Define $A_{i+1} = a_i^* \cup A_i^-$.

4. Return to Step (1), repeating until $|\tilde{b}_i - \tilde{b}_{i-1}| < \epsilon$ for some small epsilon, i.e., until there is convergence in the IEA.

Upon convergence, this algorithm will deliver a valid solution by construction. Though any initial set A_0 typically works, we have found in practice that it yields faster convergence when given an initial set with a large span: For instance, using the set of full-information actions as the initial set typically generates the best results.

The IEA offers a natural convergence criterion since it is unique for any given set of actions and is independent of the cardinality of the conjectured set A_i.

As we will show in the next section, this algorithm offers a vast improvement in terms of speed and accuracy over current techniques for a couple of reasons. Current techniques that handle RI in its general form rely on the application of a non-linear solver on a fairly high-dimensional problem. In our algorithm, the high-dimensional problem is completely linear and thus orders of magnitude faster. The algorithm also employs a non-linear solver but on a much lower-dimensional problem: Rather than solving the problem over all possible actions, which is a large grid, we only ever solve the non-linear optimization over the conjectured set, A_i, which typically has far fewer elements.

4 Applications

In this section, we illustrate by way of example that both the conceptual framework and the computationally tractable algorithm have an important potential to enhance further research. In particular, we consider two key examples: The first demonstrates the efficiency of our algorithm relative to existing methods; the second highlights the behavioral implications that are lost when Rational Inattention is simplified for tractability rather than considered in the general form we have presented here.

4.1 Revisiting Matějka (2015)

We begin by revisiting the example of Matějka (2015), who considers a monopolistic seller facing uncertain input costs. We choose this example because the author considers rational inattention in its general form and implements his solution on a fairly large scale.

The basic environment is as follows: A monopolistic seller faces isoelastic demand with an elasticity parameter θ. He also faces an per-unit input cost, c, which is stochastic. He suffers from information processing costs, indicated by a per-unit cost of λ per bit of mutual information between the input cost and the price. This implies the following payoff structure

$$u(p, c) = p^{-\theta}(p - c) \tag{3}$$

The benchmark model is calibrated as follows: $\theta = 3.0$ and c is distributed uniformly on $[0.8, 1.2]$. Matějka (2015) assumes that information is bounded by a channel capacity rather than being acquired at a cost, but our problem can be thought of as the Lagrangian of that problem and we will present it as such for simplicity. We will set $\lambda = 0.003$ because, as we will show, this will imply that roughly 1 bit of information is used, which is the capacity of his channel.

Matějka (2015) solves his problem by discretizing over the joint distribution of prices and input costs and using a gradient-based constrained non-linear optimizer. He assumes a uniformly spaced grid with 70 points along each dimension. The input cost grid ranges from .8 to 1.2 like its continuous counterpart, while the price grid ranges from 1.2 to 1.8, which correspond to the range of prices that would be optimal in the absence of information frictions.

His central result is that even though input costs follow a very disperse distribution, optimal pricing behavior is discrete, clustering mass on a comparatively small number of points.

We solve this model a number of different ways under a couple of different assumptions. The results are given in Table 1 and Figure 2. Columns I and II solve the model in Matlab using Matejka's approach with two different function-difference tolerances for Matlab's non-linear solver (*fmincon*): $1e-6$ and $1e-12$ respectively.

Columns III and IV use the same discretizations over the prior as I and II but two different grids over candidate actions: In Column III, we assume the same price grid as Columns I and II, i.e., Matejka's grid, but we solve the model with our algorithm; Column IV increases the price grid to 1000 uniformly spaced points over the same interval. In both models, the convergence tolerance is $1e - 6$.

For comparison, we also solve the model using the Blahut-Arimoto algorithm proposed by Caplin, Dean, and Leahy (2018) (CHECK REFERENCE). This is an iterative algorithm taken from Rate Distortion Theory that bypasses optimization entirely, instead achieving the optimum by iterating on Equation 1 in a particular way. These results are given in Columns V and VI. The former employs the same grid as III and the latter that of IV. The tolerance for both of these models is $1e - 6$.

Runtimes are given in seconds and are computed on a simple laptop with 8.00 GB of RAM and an Intel(R) 2.60 GHz Core i5 processor on a Windows 10 64-bit operating system.

	I	II	III	IV	V	VI
Solution Time	341.2	12,072.5	1.3	3.4	36.0	461.9
Objective Value	0.1503	0.1510	0.1511	0.1511	0.1511	0.1511

Table 1: Matějka (2015) Replication Results

Table 1 tells us two important things: First, using standard Matlab machinery for all approaches, our approach yields higher objective values than the standard approach as well as solving it orders of magnitude more quickly. Second, the solution time scales very little with the discretization over the prior: Column IV implements a solution on a grid more than ten times the size of Column III but the solution time does not even triple.

Columns V and VI reveal that the Blahut-Arimoto algorithm does offer an advantage over the benchmark non-linear solver approach in terms of speed and accuracy.[9] Nevertheless, our algorithm outperforms as well by at least an order of magnitude. It also scales up more efficiently. Though it cannot be seen in Table 1, our algorithm further delivers higher objectives than the Blahut-Arimoto approach.

Figure 2 further reveals that our algorithm does a better job at getting the right solution shape.[10] For clarity of exposition, the solutions in Figure

[9] The primary reason for this is the following: While the Blahut-Arimoto algorithm does not require optimization, its convergence speed slows substantially with the fineness of the grid.

[10] Note that from these marginals we can instantly derive the conditional structure via

2 cluster together points in a .01 neighborhood. Matejka's central finding that rationally inattentive agents tend to cluster actions discretely manifests itself clearly in both implementations of our algorithm. Further, it clusters points in the immediate neighborhood as and with roughly the same mass that Matejka finds. In contrast, using Matlab's standard solver with the old method, the price distribution is far too disperse.[11] It's clear that as we reduce the tolerance significantly (Solution II, which is roughly machine precision) the solution begins to approach the discrete solution, but it is still far too disperse. This dispersion manifests itself in the lower objectives in Table 1.

4.2 Linear-Quadratic-Gaussian Consumption-Saving

We now consider a different example: A simple, two-period consumption-saving problem. An agent has no income today and uncertain income tomorrow, $\tilde{y}$, against which she may borrow an amount $\tilde{b}$. She faces information processing costs of λ per nat of mutual information between $\tilde{y}$ and $\tilde{b}$. We normalize the interest rate to zero.

Following Sims (2003), she has quadratic flow utility of the form $u(c) = c - \frac{1}{2}c^2$. She weights utility today with $\alpha_0 > 0$ and utility tomorrow with $\alpha_0 > 0$. $\tilde{y}$ is assumed to be distributed normally. This problem can be written as

$$\begin{aligned} \max_{\tilde{b}} \; & E\left[\alpha_0\left(\tilde{c}_0 - .5\tilde{c}_0^2\right) + \alpha_1\left(\tilde{c}_1 - .5\tilde{c}_1^2\right)\right] - \lambda I(\tilde{b};\tilde{y}) \\ \text{s.t.} \quad & c_0 = b \\ & c_1 = y - b \end{aligned} \tag{4}$$

Since this problem has a simple Linear-Quadratic-Gaussian (LQG) structure, it has a well-known and closed-form solution. In particular, if the solution is non-degenerate then $\tilde{b}$ and $\tilde{y}$ follow a multivariate normal distri-

Equation 1. For clarity of exposition we do not include the Blahut-Arimoto results, but they closely resemble those generated by our algorithm.

[11] The reason for the discrepancy between Solution I and the original solution in the third panel is the software. We use Matlab's standard solvers, which are a workhorse framework for applied researchers, whereas Matejka uses AMPL, a language specifically designed to solve large-scale multidimensional optimization problems.

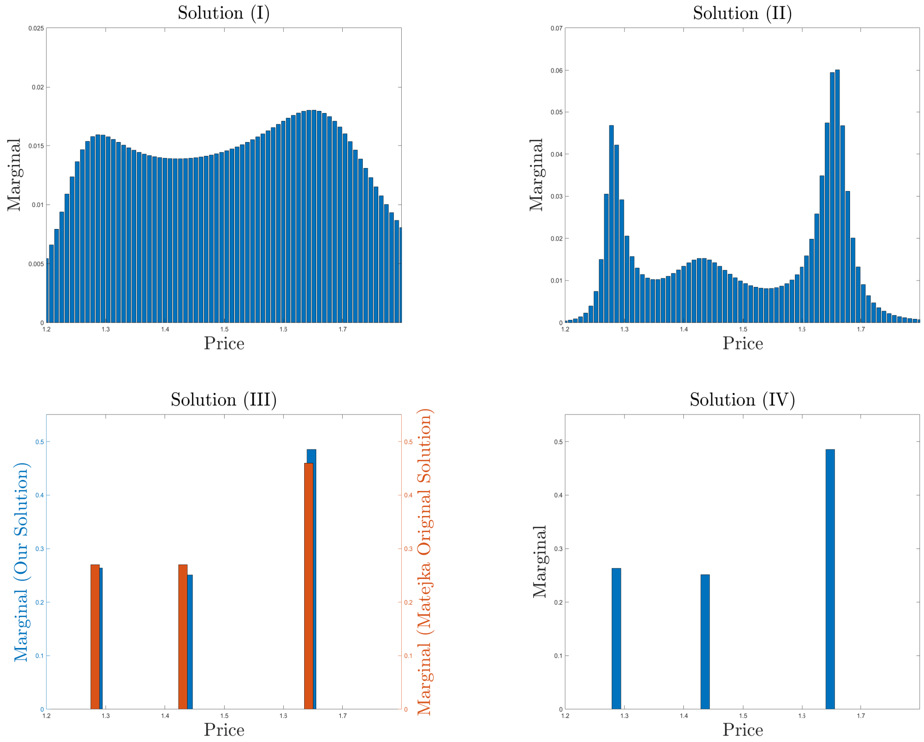


Figure 2: Matějka (2015) Replication Solutions

bution (Sims (2003), Cover and Thomas (2012)). As information costs rise, the variance in $\tilde{b}$ shrinks, as does its correlation with $\tilde{y}$.

This result arises from the fact that, conditional on a given variance, the Gaussian distribution maximizes entropy. Its convenient structure has given rise to a large applied literature (LIST LIT HERE). We show here that, while not incorrect, the LQG approach is fragile in the following sense: Very fine discrete approximations to Gaussian distributions, solved with our algorithm, yield highly non-Gaussian solutions.

To see this, we solve the model as follows. We assume that $\bar{y} = 1$ and that $\sigma_y^2 = .05$. We then discretize $\tilde{y}$ equidistantly over a 1000 grid points, truncating it at four standard deviations from the mean in both directions. We then employ our algorithm.

The results can be found in Figure 3 for a variety of different information costs. Again, for clarity of exposition we cluster together points in a .01

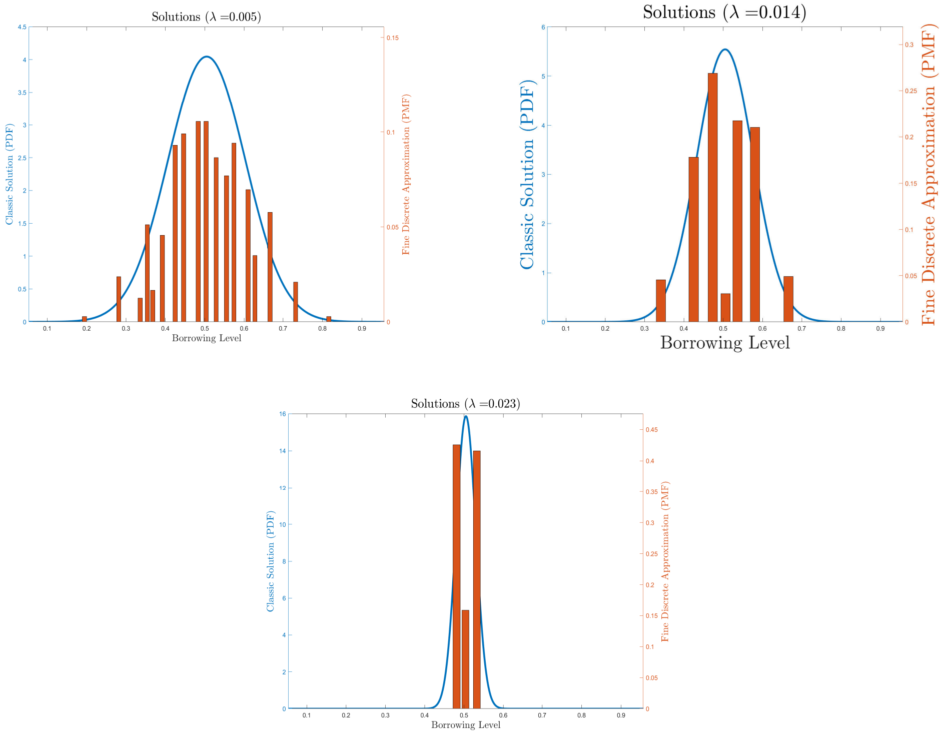


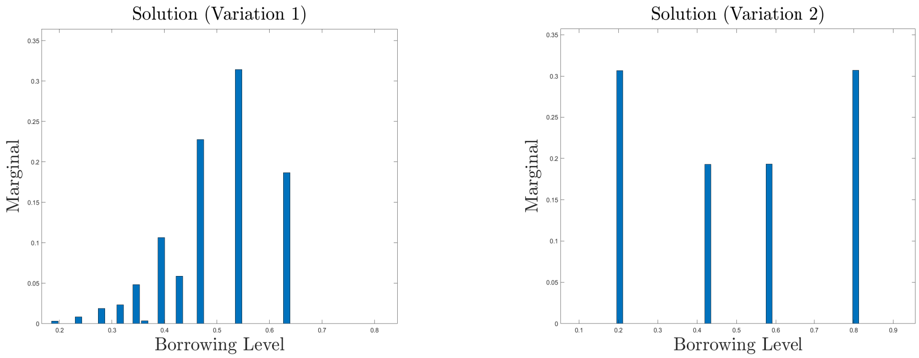
Figure 3: Consumption-Saving Solutions

neighborhood of each other. In all three cases, the objectives are nearly identical across the two classes of solution, but it is also the case that our discrete approximation is mildly higher. Since the problems are not exactly the same, this does not mean that the LQG solution is wrong (it's not), but only that it is quite fragile and small changes can cause substantial variation in behavior without.

It is apparent that despite the deep similarity of these problems, the solutions exhibit striking differences. Two in particular stand out. First, while the variance of borrowing shrinks in both solutions, in our approximating model the cardinality of the support set shrinks as well. More information gets 'packaged' into discrete chunks as information gets more expensive. This feature is absent in the oft-employed LQG solution. This is noteworthy since it is a documented feature in non-LQG RI problems (Sims (2006) or Matějka (2015)). Here we show that this logic generalizes to problems arbitrarily close to LQG.

Second, the optimal discretized distribution does not appear to mimic a Gaussian in important ways. For instance it seems that less weight is placed on median actions and more on extremes than a Gaussian. When $\lambda = .005$, for instance, the distribution is relatively flat compared to a Gaussian. More strikingly, when $\lambda = .014$ or $.023$ the median borrowing level is *less likely* than its neighboring extremal levels, not more likely. This suggests that LQG solutions *as approximations to higher-order problems* (LIST EXAMPLES HERE) may actually be missing fundamental elements of RI behavior, such as the tendency to hollow out median actions.

We explore this further with a pair of deviations from the Problem 4. These results can be seen in Figure 4. In Variation 1, setting $\lambda = .02$, we consider the impact of changing the flow utility from linear-quadratic to log. In this solution, we can see that the curvature in the utility impacts the solution: Miscalculated overborrowing results in an asymmetric utility loss in the second period, so the agent does not do this.


Solution (Variation 1)

Solution (Variation 2)

Figure 4: Consumption-Saving Solutions: Variations

In Variation 2, we instead change the distribution of $\tilde{y}$ from truncated Gaussian to uniform on the same domain. Here, we retain the symmetry, but the solution is far from Gaussian; in fact, much like the LQG approximation at higher information costs, less weight is placed on the median borrowing levels and more weight is placed at the extremes.

A Additional proofs

A.1 Ignorance Equivalent Action

Lemma 2. *Any point $\boldsymbol{b} \in \partial^{+}\mathcal{B}$ can be written as a convex combination of at most I points in $\boldsymbol{\beta}(\mathcal{A})$.*

Proof. Since $\boldsymbol{b}$ is on the boundary of $\mathcal{B}$, there exists a supporting hyperplane such that $\boldsymbol{\psi} \cdot \boldsymbol{b} = 1$ and $\boldsymbol{\psi} \cdot \boldsymbol{b}' \leq 1$ for all $\boldsymbol{b}' \in \mathcal{B}$. The convex combination $\boldsymbol{b}$ achieves this upper bound if and only if $\boldsymbol{\psi} \cdot \boldsymbol{\beta}(a) = 1$ for all actions a in its support. The set $\boldsymbol{\beta}(\mathcal{A}_{\boldsymbol{\psi}}) := \{\boldsymbol{\beta}(a) \mid a \in \mathcal{A}, \boldsymbol{\psi} \cdot \boldsymbol{\beta}(a) = 1\}$ is contained in a $I-1$-dimensional vector space, and thus Carathéodory's Theorem implies that $\boldsymbol{b} \in \mathsf{conv.hull}(\boldsymbol{\beta}(\mathcal{A}_{\boldsymbol{\psi}})$ can be written as a convex combination over at most I points. □

Proof of Corollary 2: By Theorem 1, the consideration set $A^k = \mathsf{support}(\mathcal{C}^k)$ maps into the spanning points of $\boldsymbol{b}^*$, while the marginal probabilities $\mathsf{Pr}\big(\mathcal{C}^k = a\big)$ equal the corresponding weights,

$$\boldsymbol{\beta}(\alpha) = \sum_{a \in A^k} \mathsf{Pr}\big(\mathcal{C}^k = a\big)\, \boldsymbol{\beta}(a) \qquad \forall k = 1, 2. \tag{5}$$

As a side-product, this simplifies the optimal joint probabilities in Eq. (1) to $\mathsf{Pr}\big(\mathcal{C}^k = a, \mathcal{S} = i\big) = \mathsf{Pr}\big(\mathcal{C}^k = a\big)\, \mathsf{Pr}(\mathcal{S} = i)\, \frac{\beta_i(a)}{\beta_i(\alpha)}$ and allows us to write the mutual information of $\mathcal{C}^k$ and $\mathcal{S}$ as

$$\begin{aligned}\mathcal{I}(\mathcal{C}^k, \mathcal{S}) &= \sum_{i=1}^{I} \sum_{a \in A^k} \mathsf{Pr}\big(\mathcal{C}^k = a, \mathcal{S} = i\big) \ln\!\left(\frac{\mathsf{Pr}\big(\mathcal{C}^k = a, \mathcal{S} = i\big)}{\mathsf{Pr}\big(\mathcal{C}^k = a\big)\, \mathsf{Pr}(\mathcal{S} = i)}\right) \\ &= \sum_{i=1}^{I} \pi_i \sum_{a \in A^k} \mathsf{Pr}\big(\mathcal{C}^k = a\big)\, f_i(a),\end{aligned} \tag{6}$$

where $f_i(a) = \frac{\beta_i(a)}{\beta_i(\alpha)} \ln\left(\frac{\beta_i(a)}{\beta_i(\alpha)}\right)$ for all $a \in \mathcal{A}^1 \cup \mathcal{A}^2$.

The condition $\boldsymbol{\beta}(A^1) \subseteq \mathsf{conv.hull}(\boldsymbol{\beta}(\mathcal{A}^2))$ states that each action $a \in A^1$ can be written as a $\boldsymbol{\beta}$-transformed convex combination over finitely many points in $\mathcal{A}^2$, allowing us to write $\boldsymbol{\beta}(a) = \sum_{\tilde{a} \in \mathcal{A}^2} q(a, \tilde{a}) \boldsymbol{\beta}(\tilde{a})$ and hence

$$\boldsymbol{\beta}(\alpha) = \sum_{a \in A^1} \mathsf{Pr}\big(\mathcal{C}^1 = a\big)\, \boldsymbol{\beta}(a) = \sum_{\tilde{a} \in \mathcal{A}^2} \underbrace{\sum_{a \in A^1} \mathsf{Pr}\big(\mathcal{C}^1 = a\big)\, q(a, \tilde{a})}_{w(\tilde{a})}\, \boldsymbol{\beta}(\tilde{a}).$$

Since the optimal choice $\mathcal{C}^2$ is unique, so are the weights w, implying that $w(\tilde{a}) = \mathsf{Pr}(\mathcal{C}^2 = \tilde{a})$ for each $a \in \mathcal{A}^2$.

Continuing from above, the decomposition of $\boldsymbol{\beta}(a)$ and the convexity of each f_i imply by Jensen's inequality that

$$\mathcal{I}(\mathcal{C}^1, \mathcal{S}) \leq \sum_{i=1}^{I} \pi_i \sum_{a \in A^1} \mathsf{Pr}\big(\mathcal{C}^1 = a\big) \sum_{\tilde{a} \in \mathcal{A}^2} q(a, \tilde{a}) f_i(\tilde{\alpha}).$$

Reshuffling the right side yields $\sum_i \pi_i \sum_{\tilde{a} \in \mathcal{A}^2} w(\tilde{a}) f_i(\tilde{\alpha}) \overset{(6)}{=} \mathcal{I}(\mathcal{C}^2, \mathcal{S})$. □

A.2 Convergence of the statewise payoff distributions

Consider a compact set $V \in \mathbb{R}^I$. The sequence $(p^n)_{n=0}^{\infty}$ consists of probability mass functions over V whose support $V^n = \{\boldsymbol{v} \in V | p(\boldsymbol{v}) > 0\}$ is of bounded cardinality $|V^n| \leq K < \infty$. Assume that the expectations $\boldsymbol{w}^n = \sum_{\boldsymbol{v} \in V^n} p^n(\boldsymbol{v})\boldsymbol{v}$ converge to a limit $\boldsymbol{w}^0 \in \mathsf{conv.hull}(V)$, and that the representation of the limit is unique. Specifically, assume that there exists a unique probability mass function p^0 over $V_0 \subset V$ such that $\boldsymbol{w}^0 = \sum_{\boldsymbol{v} \in V^0} p^n(\boldsymbol{v})\boldsymbol{v}$. Note that uniqueness requires all vectors in V^0 to be affinely independent, and hence $|V^0| \leq I + 1$.

Let $B_\epsilon(\boldsymbol{v}) = \big\{\boldsymbol{v'} \in \mathbb{R}^I \,\big|\, \|\boldsymbol{v} - \boldsymbol{v'}\| < \epsilon\big\}$ denote the ϵ-ball around $\boldsymbol{v} \in V$, and $B_\epsilon(V^0) = \bigcup_{\boldsymbol{v} \in V} B_\epsilon(\boldsymbol{v})$ their union for all points in V^0. For $\epsilon < \bar{\epsilon} := \frac{1}{2} \min_{\boldsymbol{v}, \boldsymbol{v'} \in V^0} \|\boldsymbol{v} - \boldsymbol{v'}\|$ small enough, no two balls intersect. The total probability over a subset $V' \subset \mathbb{R}^I$ is written as $p^n(V') := \sum_{\boldsymbol{v} \in V^n \cap V'} p^n(\boldsymbol{v})$.

Lemma 3. *For any $\boldsymbol{v}^0 \in V^0$, the limit $\bar{p}(\boldsymbol{v}^0) := \lim_{n \to \infty} p^n(B_\epsilon(\boldsymbol{v}^0))$ is independent of $\epsilon \in (0, \bar{\epsilon})$ and equal to $p^0(\boldsymbol{v}^0)$.*

Proof. The proof proceeds in two steps. First, we show that p^n concentrates all weight close to V^0 as n grows. Second, we show that the total mass near each $\boldsymbol{v}^0 \in V^0$ approaches $p^0(\boldsymbol{v}^0)$. The main challenge is to rule out small but persistent imprecisions in the support that counterbalance small imprecisions in the weights.

Claim 1: As $n \to \infty$, $p^n(\mathbb{R}^I \setminus B_\epsilon(V^0)) \to 0$ for all $\epsilon \in (0, \bar{\epsilon})$.

By contradiction, assume that there exists $\epsilon \in (0, \bar{\epsilon})$ and $\delta > 0$ along with a subsequence p^{n_k} such that $p^{n_k}(\mathbb{R}^I \setminus B_\epsilon(V^0)) \geq \delta$ for all k. For each k, the support of p^{n_k} is of cardinality at most K, and hence at least one vector $\tilde{\boldsymbol{v}}^k \in V^{n_k} \setminus B_\epsilon(V^0)$ has weight $q^k := p^{n_k}(\tilde{\boldsymbol{v}}^k) \geq \delta/K$. The bounded sequence

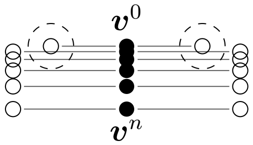


Figure 5: How Lemma 3 can fail if V is not closed.

$(q^k, \tilde{\boldsymbol{v}}^k) \subset [0,1] \times V$ admits a convergent subsequence $(q^{k_m}, \tilde{\boldsymbol{v}}^{k_m}) \to (\bar{q}, \bar{\boldsymbol{v}})$ by Bolzano Weierstrass, with two important limit properties: $\bar{\boldsymbol{v}} \in V \setminus B_\epsilon(V^0)$ since V is closed and weak inequalities are maintained in the limit. For the same reason, $\bar{q} \geq \delta/K$. The fact that $\bar{\boldsymbol{v}}$ is a member of V is crucial,[12] for it ensures that the limit of the expectations

$$\bar{\boldsymbol{w}} := \lim_{m\to\infty} \sum_{\boldsymbol{v}\in V^{n_{k_m}}} p^{n_{k_m}}(\boldsymbol{v})\boldsymbol{v} = \bar{q}\bar{\boldsymbol{v}} + (1-\bar{q}) \underbrace{\lim_{m\to\infty} \sum_{\substack{\boldsymbol{v}\in V^{n_{k_m}} \\ \boldsymbol{v}\neq\tilde{\boldsymbol{v}}^{k_m}}} \frac{p^{n_{k_m}}(\boldsymbol{v})}{1-\bar{q}}\boldsymbol{v}}_{\in \mathsf{conv.hull}(V)}$$

can be represented as a convex combination that puts at least weight $\bar{q} > 0$ on $\bar{\boldsymbol{v}} \in V$. Since the limit of any subsequence is equal to overall limit $\boldsymbol{b}^0$, the representation of $\boldsymbol{b}^0$ is not unique, as $\bar{\boldsymbol{v}} \notin V^0$.

Claim 1 implies that $\bar{p}$ is a well-defined probability mass function over V^0 and is independent of $\epsilon \in (0, \bar{\epsilon})$, since the support of p^n converges to V^0. It remains to show that it is equal to p^0.

Claim 2: $\bar{p}(\boldsymbol{v}^0) = p^0(\boldsymbol{v}^0)$ for all $\boldsymbol{v}^0 \in V^0$.

By contradiction, assume that there exists $\tilde{\boldsymbol{v}} \in V^0$ such that $\delta := |\bar{p}(\tilde{\boldsymbol{v}}) - p^0(\tilde{\boldsymbol{v}})| > 0$. Obviously, $\delta \leq 1$. Let $\bar{\boldsymbol{w}} := \sum_{\boldsymbol{v}^0\in V^0} \bar{p}(\boldsymbol{v}^0)\boldsymbol{v}^0$ denote the expectation of $\bar{p}$, and $d := \|\boldsymbol{w}^0 - \bar{\boldsymbol{w}}\|$ its distance from $\boldsymbol{w}^0$. Since $\boldsymbol{w}^0$ admits a unique representation, $d > 0$. Also, let $\bar{v} = \max_{\boldsymbol{v}\in V} \|\boldsymbol{v}\|$ denote an upper bound on the norms inside the compact set V.

By appropriately choosing $\epsilon = \min\left\{\frac{1}{6}d\delta, \bar{\epsilon}\right\}$, we will show that this con-

[12] Figure 5 contains an illustration that shows the result does not hold if V is not closed.

tradicts convergence of $\boldsymbol{w}^n \to \boldsymbol{w}^0$. Indeed, note that

$$\begin{aligned}
\|\boldsymbol{w}^0 - \boldsymbol{w}^n\| = &\Big\|\boldsymbol{w}^0 - \bar{\boldsymbol{w}} + \textstyle\sum_{\boldsymbol{v}^0\in V^0} \big(\bar{p}(\boldsymbol{v}^0) - p^n(B_\epsilon(\boldsymbol{v}^0))\big)\boldsymbol{v}^0 \\
&+ \textstyle\sum_{\boldsymbol{v}^0\in V^0} \sum_{\boldsymbol{v}\in V^n \cap B_\epsilon(\boldsymbol{v}^0)} p^n(\boldsymbol{v})\big(\boldsymbol{v}^0 - \boldsymbol{v}\big) \\
&- \textstyle\sum_{\boldsymbol{v}\in V^n \setminus B_\epsilon(V^0)} p^n(\boldsymbol{v})\boldsymbol{v}\Big\| \\
\geq &\Big\|\boldsymbol{w}^0 - \bar{\boldsymbol{w}}\Big\| - \textstyle\sum_{\boldsymbol{v}^0\in V^0} |\bar{p}(\boldsymbol{v}^0) - p^n(B_\epsilon(\boldsymbol{v}^0))| \, \|\boldsymbol{v}^0\| \\
&- \textstyle\sum_{\boldsymbol{v}^0\in V^0} \sum_{\boldsymbol{v}\in V^n \cap B_\epsilon(\boldsymbol{v}^0)} p^n(\boldsymbol{v}) \, \|\boldsymbol{v}^0 - \boldsymbol{v}\| \\
&- \textstyle\sum_{\boldsymbol{v}\in V^n \setminus B_\epsilon(V^0)} p^n(\boldsymbol{v}) \, \|\boldsymbol{v}\| ,
\end{aligned} \tag{7}$$

where the equality is obtained by adding and subtracting $\sum_{\boldsymbol{v}^0\in V^0} \big(\bar{p}(\boldsymbol{v}^0) - p^n(B_\epsilon(\boldsymbol{v}^0))\big)\boldsymbol{v}^0$ and regrouping the terms in V^n according to the ϵ-balls. The triangle inequalities $\|\boldsymbol{w}\| - \|\boldsymbol{w}'\| \leq \|\boldsymbol{w} + \boldsymbol{w}'\| \leq \|\boldsymbol{w}\| + \|\boldsymbol{w}'\|$ generate a lower bound.

Each of the terms in Eq. (7) is bounded for n large enough. First, $\|\boldsymbol{w}^0 - \bar{\boldsymbol{w}}\| = d$ by definition. The norms $\|\boldsymbol{v}^0\|$ and $\|\boldsymbol{v}\|$ are bounded above by $\bar{v}$, while $\|\boldsymbol{v}^0 - \boldsymbol{v}\| < \epsilon$ for any $\boldsymbol{v} \in B_\epsilon(\boldsymbol{v}^0)$. The resulting lower bound

$$d - \left(\sum_{\boldsymbol{v}^0\in V^0} |\bar{p}(\boldsymbol{v}^0) - p^n(B_\epsilon(\boldsymbol{v}^0))|\right)\bar{v} - p^n(B_\epsilon(V^0))\epsilon - p^n(\mathbb{R}^I \setminus B_\epsilon(V^0))\bar{v}$$

depends only on the p^n distributions and their limit. By the definition of $\bar{p}$, the expression in brackets is smaller than $\epsilon/\bar{v}$ for n large enough. Because p^n is a probability function, $p^n(B_\epsilon(V^0)) \leq 1$ for any n. And finally, by Claim 1, $p^n(\mathbb{R}^I \setminus B_\epsilon(V^0)) < \epsilon/\bar{v}$ for n large enough. This however contradicts the limit assumption $\boldsymbol{w}^n \to \boldsymbol{w}^0$ since

$$\left\|\boldsymbol{w}^0 - \boldsymbol{w}^n\right\| > d - 3\epsilon \geq d\left(1 - \frac{\delta}{2}\right) \geq \frac{d}{2} > 0$$

for all n large enough. □

Lemma 4. *Consider a sequence of conditional choices (A^n, P^n) with bounded support $|\bigcup_{n=1}^\infty A^n| \leq K \in \mathbb{N}$. If the associated $\boldsymbol{b}^n = \sum_{a\in A^n} \sum_{i=1}^I \pi_i P^n_{a|i} \boldsymbol{\beta}(a)$ converges to some $\boldsymbol{b} \in \partial^+ \boldsymbol{\mathcal{B}}$ with unique representation $\sum_{a\in A} q(a)\boldsymbol{\beta}(a)$, then $P^n \xrightarrow{d} P$, where P are the optimal conditionals resulting from q and Eq. (1).*

Proof. Mathematically, each element of the sequence $\boldsymbol{b}^n$ is a convex combination over at most K vectors from a compact set $\boldsymbol{\beta}(\mathcal{A}) \in \mathbb{R}^I$. This sequence converges to a limit $\boldsymbol{\beta}(\alpha)$ that admits a unique representation

$p^\alpha : \mathcal{A} \to [0,1]$ as a convex combination of vectors in $\boldsymbol{\beta}(\mathcal{A})$, since $\mathcal{A}$ is generic. Lemma 3 in Appendix A.2 establishes that in the limit, p^n concentrates weight $p^\alpha(a)$ on actions arbitrarily close to $\boldsymbol{\beta}(a)$ for n large, for each $a \in A^0 := \{a \in \mathcal{A} | p^\alpha(a) > 0\}$. Formally, we show that $\lim_{\epsilon\to 0}\lim_{n\to 0} p^n(B_\epsilon(a^0)) = p^\alpha(a^0)$ for any $a^0 \in A^0$, where $p^n(\mathcal{B}_\epsilon(a^0))$ denotes the total weight that p^n places on actions a with $||\boldsymbol{\beta}(a) - \boldsymbol{\beta}(a^0)|| < \epsilon$. Since the expression of conditional probabilities **??** is continuous in $\beta_i(a)$ and in b_i^n, the same holds for the conditional distributions,

$$\lim_{\epsilon\to 0}\lim_{n\to\infty} p_i^n(B_\epsilon(a^0)) = \frac{\beta_i(a^0)}{b_i^0}\lim_{\epsilon\to 0}\lim_{n\to\infty} p^n(B_\epsilon(a^0)) = p_i^\alpha(a^0) \quad \forall a^0 \in A^0.$$

In other words, as $n \to \infty$, the jumps in the distribution of $u_i(\boldsymbol{a}_{p_i^n})$ converge to those of $u_i(\boldsymbol{a}_{p_i^0})$ both in location and in magnitude, implying convergence in distribution. □

B Algorithmic Implementation

B.1 Pseudocode

B.2 Proof of Convergence

References

Caplin, A., M. Dean, and J. Leahy (2018): “Rational Inattention, Optimal Consideration Sets and Stochastic Choice,” *The Review of Economic Studies*, p. rdy037.

Cover, T. M., and J. A. Thomas (2012): *Elements of information theory.* John Wiley & Sons.

Eggleston, H. G. (1958): *Convexity*, Cambridge Tracts in Mathematics. Cambridge University Press.

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

Jung, J., J.-H. Kim, F. Matejka, and C. A. Sims (2015): “Discrete actions in information-constrained decision problems,” Discussion paper, working paper.

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

Matějka, F. (2015): “Rationally inattentive seller: Sales and discrete pricing,” *The Review of Economic Studies*, 83(3), 1125–1155.

Matějka, F., and A. McKay (2015): “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model,” *American Economic Review*, 105(1), 272–98.

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

——— (2006): “Rational inattention: Beyond the linear-quadratic case,” *American Economic Review*, 96(2), 158–163.