Federal Reserve Bank of Philadelphia

Working Papers
Research Department

WP 21-30
September 2021
https://doi.org/10.21799/frbp.wp.2021.30

Geometric Methods for Finite Rational Inattention

Roc Armenter
Federal Reserve Bank of Philadelphia Research Department

Michèle Müller-Itten
University of Notre Dame

Zachary R. Stangebye
University of Notre Dame

ISSN: 1962-5361

Geometric Methods for Finite Rational Inattention

Roc Armenter*, Michèle Müller-Itten†, Zachary R. Stangebye‡

September 8, 2021

Abstract

We present a geometric approach to the finite Rational Inattention (RI) model, recasting it as a convex optimization problem with reduced dimensionality that is well-suited to numerical methods. We provide an algorithm that outperforms existing RI computation techniques in terms of both speed and accuracy. We also introduce methods to quantify the impact of numerical inaccuracy on the behavioral predictions and to produce robust predictions regarding the most frequently implemented actions.

Keywords: Rational inattention, Shannon entropy, information acquisition, learning, consideration sets.

JEL: D81, D83, C63

*Federal Reserve Bank of Philadelphia: roc.armenter@phil.frb.org

†University of Notre Dame: michele.muller@nd.edu

‡University of Notre Dame: zstangeb@nd.edu

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 or the Federal Reserve System. The authors would like to thank Isaac Baley, Thomas Gresik, Maciej Kotowski, John Leahy, Jun Nie, and participants at various conferences and seminars for constructive feedback and comments. Financial support from the Notre Dame Institute for Scholarship in the Liberal Arts is gratefully acknowledged. Philadelphia Fed working papers are free to download at https://philadelphiafed.org/research-and-data/publications/working-papers.

1 Introduction

Introduced by Sims [2003], Rational Inattention (RI) has been gaining increased acceptance as a model of information acquisition and processing, particularly in macroeconomics and finance. However, by and large we still do not know how effective RI models are at explaining real-world phenomena. Maćkowiak et al. [2018, p.27] states that "[t]he model of RI is well suited for a boom in empirical work, which has not yet occurred," and Gabaix [2014] notes that the limited scope of existing applied work is partly due to the conceptual and computational complexity of RI optimization problems. Outside of a handful of special cases, RI models do not admit a closed-form solution. Existing numerical methods are often computationally intensive and may suffer from accuracy problems.

We introduce a geometric approach to finite RI models with the most common information cost specification, the average reduction in Shannon entropy [Caplin et al., 2018, Matějka and McKay, 2015, Sims, 2003]. Our approach is very well suited for numerical computation, simplifying the original problem and providing novel methods to quantify the accuracy of its behavioral predictions.

To obtain the geometric approach, we transform each action's payoffs across states into what we term an "attention vector" — a simple mapping that accentuates payoff differences when information is cheap and attenuates them when it is expensive. The convex hull spanned by the action payoff vectors forms a convex polytope, the "attention possibilities set." The original RI problem is equivalent to finding the optimal attention vector from this set, resulting in a strictly convex optimization problem. This geometric approach is markedly simpler, reducing the dimensionality of the original problem, providing two sets of convenient optimality conditions, and separating the roles of prior beliefs and payoffs. Yet, it is straightforward to recover the joint probability distribution of states and actions from the optimal attention vector.

There are plenty of known algorithms for convex problems that can be readily used for our geometric approach. We provide one such algorithm, adapted from standard sequential quadratic programming with active set methods.[1] Our algorithm performs favorably compared with other approaches, delivering substantial gains in both speed and accuracy. The attention vector transformation informs three noteworthy design

[1]Source codes are available on GitHub at https://github.com/mmulleri/GAP-SQP.

features: Our choice of active sets, the stopping criterion, and a scaling routine to avoid floating point errors.

We also provide novel methods designed to improve the accuracy and robustness of the behavioral predictions of the model's numerical solutions. We first propose a precision metric that relies on the similarity of the implied attention vectors and ensures that the joint distribution of states and actions converges to the true solution. It does so by balancing tolerance errors on two margins: The extensive margin that identifies *which* actions are chosen with positive probability, and the intensive margin that specifies the relative frequency of each chosen action. The set of actions chosen with positive probability, also known as the "consideration set," is often an object of interest by itself but is notoriously difficult to identify numerically.

We show how to obtain robust predictions regarding the most and least frequently implemented actions despite numerical inaccuracy. Manipulating the optimality conditions from the geometric approach, we derive "cover" sets of actions that, altogether, are played with an overall probability of our choice. These covers can be used either to approximate the consideration set or, conversely, to identify actions that are rarely, if ever, chosen. The latter is helpful for hypothesis testing: Identifying groups of actions that jointly have a low choice probability ensures that a rejection of the underlying RI specification is not clouded by noise from the computation.

Finally, we provide three applications intended to illustrate some of the advantages of our approach. The first application is based on the price-setting problem of Matějka [2016]. We document that our algorithm is orders of magnitude faster than the well-known Blahut-Arimoto algorithm [see Cover and Thomas, 2012] and scales well with the size of the action and state spaces. Precision, rather than speed, is the focus of our second application, based on the two-dimensional portfolio design problem of Jung et al. [2019]. Our algorithm unveils some behavioral differences of importance. In both applications, we obtain robust predictions by deploying our geometric methods, which provide very tight estimates of the consideration set. Our third application is a novel task-assignment problem, a complex but naturally finite RI problem that is the ideal scenario for the geometric approach. We find that RI's flexible and costly learning predicts subtle adjustments in management strategy and information acquisition — patterns that would be easy to miss without a high degree of numerical accuracy.

Related literature. Rational inattention models have rapidly found their way into a variety of fields, from finance to monetary economics. We do not aim to properly review what is by now a large literature — see Maćkowiak et al. [2018] for an excellent survey of both theoretical and applied work with RI models. Instead we briefly discuss previous key developments in solution methods for RI models.

Early work on RI models restricted analysis to Linear-Quadratic Gaussian (LQG) frameworks, or assumed that the solution was Gaussian as an approximation, to obtain analytic results that can, in turn, be embedded in an equilibrium model and have led to many insights for aggregate phenomena.[2]

Sims [2006] exhorted researchers to go beyond the LQG case, and more recent research has addressed some of the shortcomings of the LQG framework: Luo et al. [2017] apply Gaussian techniques with constant absolute risk aversion preferences, allowing them to study the dynamics of consumption and wealth in general equilibrium. Mondria [2010] allows signals to be linear combinations of the underlying state of the economy, an approach that is also followed in Kacperczyk et al. [2016], among others. Miao et al. [2019] make further progress in multi-variate LQG environments. In settings where choices are discrete, some researchers use the Cardell distribution to obtain closed forms that are amenable to empirical analysis [Bertoli et al., 2020, Brown and Jeon, 2020, Dasgupta and Mondria, 2018].

Our methods offer an alternative to these distributional form assumptions, for problems with finitely many states and actions. While many of the RI problems in finance and macroeconomics feature a continuum of actions and states, the computational gains of our approach make it feasible to use very fine grids. Our applications in Sections 5.1 and 5.2 draw on two leading examples that apply such discretization to continuous problems, Matějka [2016] and Jung et al. [2019].

Paper structure. We formally describe the classic Rational Inattention approach in Section 2 and introduce our geometric approach in Section 3. In Section 4, we develop a toolkit for finite RI problems that improve the accuracy and robustness of numerical methods. Section 5 illustrates the relevance of these new tools in three specific applications, and Section 6 concludes.

[2] A necessarily incomplete list of examples is: Peng [2005], Peng and Xiong [2006], and Huang and Liu [2007] for asset pricing; Maćkowiak and Wiederholt [2009] for monetary shocks; Gaglianone et al. [2020] for forecasting; Van Nieuwerburgh and Veldkamp [2009] and Van Nieuwerburgh and Veldkamp [2010] for home bias and under-diversification in asset portfolios.

Notation. We use vector notation throughout the paper and rely on the following conventions: Boldface letters such as $\boldsymbol{x}$ denote I-dimensional vectors. To describe its component-wise construction, we also refer to $\boldsymbol{x}$ as $[x_i]$. For example, $[x_i/y_i]$ describes the vector $\boldsymbol{z}$ with i-th component $z_i = x_i/y_i$. When comparing vectors, we write $\boldsymbol{v} \geq \boldsymbol{w}$ if and only if $v_i \geq w_i$ for all i, $\boldsymbol{v} > \boldsymbol{w}$ if and only if $\boldsymbol{v} \geq \boldsymbol{w}$ and $\boldsymbol{v} \neq \boldsymbol{w}$, and $\boldsymbol{v} \gg \boldsymbol{w}$ if and only if $v_i > w_i$ for all i.

2 Rational Inattention Problem

We consider the standard RI problem where an agent faces a finite menu of options with state-dependent payoffs and can condition her choice on arbitrary but costly signals. More accurate signals are more costly, and we follow the literature [Caplin et al., 2018, Matějka and McKay, 2015, Sims, 2003, 2006] in focusing on information-processing costs that are proportional to Shannon entropy.

Formally, an agent faces an unknown state of the world $i \in \mathcal{I} = \{1, ..., I\}$, each occurring with positive *prior* probability $\pi_i > 0$. The agent has to implement a single *action* $\boldsymbol{a}$ from the finite *menu* $\mathcal{A}$, each identified by its state-dependent consumption payoffs $\boldsymbol{a} = (a_1, ..., a_I) \in \mathbb{R}^I$. We denote the set of probability mass functions over the menu as $\Delta\mathcal{A}$. Before implementing an action, the agent can acquire information about the state of the world, but more accurate information is more costly. By the obedience principle, it is without loss of generality to assume that the decision maker relies on a signal that directly recommends a specific action. We denote the resulting conditional implementation probabilities by $\boldsymbol{P} \in (\Delta\mathcal{A})^I$, where $P_i(\boldsymbol{a})$ denotes the probability of implementing action $\boldsymbol{a}$ conditional on state i.

The optimal conditionals $\boldsymbol{P}$ maximize expected consumption utility net of information processing costs, measured as the average reduction in Shannon entropy between prior and posterior. These costs are also known as the mutual information and are equal to the expected Kullback-Leibler divergence between conditional and marginal choice probabilities [Cover and Thomas, 2012],

$$\mathsf{MI}(\boldsymbol{P}, p \,|\, \boldsymbol{\pi}) = \sum_{i \in \mathcal{I}} \sum_{\boldsymbol{a} \in A} \pi_i P_i(\boldsymbol{a}) \ln\left(\frac{P_i(\boldsymbol{a})}{p(\boldsymbol{a})}\right),$$

where $p(\boldsymbol{a}) = \boldsymbol{\pi} \cdot \boldsymbol{P}(\boldsymbol{a})$ refers to the marginal implementation probability of $\boldsymbol{a}$.[3] A proportionality constant $\lambda > 0$ translates the informational burden from nats into utils. Mathematically, the choice problem is parametrized by a triplet $(\mathcal{A}, \boldsymbol{\pi}, \lambda)$,

$$W(\mathcal{A}, \boldsymbol{\pi}, \lambda) = \begin{cases} \max\limits_{\boldsymbol{P} \in (\Delta\mathcal{A})^I,\, p \in \Delta\mathcal{A}} & \sum\limits_{i \in \mathcal{I}} \sum\limits_{\boldsymbol{a} \in \mathcal{A}} \pi_i P_i(\boldsymbol{a}) a_i - \lambda\, \mathsf{MI}(\boldsymbol{P}, p \,|\, \boldsymbol{\pi}) \\ \text{s.t.} & p(\boldsymbol{a}) = \boldsymbol{\pi} \cdot \boldsymbol{P}(\boldsymbol{a}) \qquad \forall \boldsymbol{a} \in \mathcal{A}. \end{cases} \tag{RI}$$

The value function is nondecreasing under menu expansion, $W(\mathcal{A}, \boldsymbol{\pi}, \lambda) \leq W(\mathcal{A}', \boldsymbol{\pi}, \lambda)$ whenever $\mathcal{A} \subseteq \mathcal{A}'$, since the agent can always restrict the support of $\boldsymbol{P}$ to a subset of available actions at no cost. RI agents frequently implement only a subset of actions with positive probability, and we follow Caplin et al. [2018] in referring to $\mathsf{support}(\boldsymbol{P})$ as the agent's consideration set.

3 Geometric Approach

Key to our results is the equivalence between (RI) and a simpler optimization problem, which we call the Geometric Approach (G). To get there, we show that it is without loss of generality to relax the constraint in (RI), as the optimal marginals are always consistent with the conditionals (see Lemma 3 in the appendix). Using standard optimization techniques, we find that optimal conditionals are equal to

$$P_i(\boldsymbol{a}) = \frac{p(\boldsymbol{a}) e^{a_i/\lambda}}{\sum_{\boldsymbol{a}' \in \mathcal{A}} p(\boldsymbol{a}') e^{a_i'/\lambda}}, \tag{1}$$

based on the same first-order conditions as previously reported Matějka and McKay [2015]. As observed by Caplin et al. [2018], most terms in (RI) cancel out when we substitute in these conditionals,

$$W(\mathcal{A}, \boldsymbol{\pi}, \lambda) = \lambda \max_{p \in D} \sum_{i \in \mathcal{I}} \pi_i \ln \left(\sum_{\boldsymbol{a} \in \mathcal{A}} p(\boldsymbol{a}) e^{a_i/\lambda} \right), \tag{2}$$

but only a finite set of marginals $p \in D \subseteq \Delta\mathcal{A}$ is consistent with Equation (1). In our relaxed version of (RI), this restriction falls away and any marginal in $\Delta\mathcal{A}$ is feasible. The resulting maximization problem is at the heart of our approach, but for

[3] In line with conventional notation, we assume that $0 \ln 0 = 0$.

convenience we divide by the constant λ and apply a change of variables.

Specifically, we assign to each action $\boldsymbol{a}$ an *attention vector* $\boldsymbol{\beta}(\boldsymbol{a}) := [e^{a_i/\lambda}] \in (0,\infty)^I$. This mapping accentuates differences in payoffs when the information is cheap and attenuates them when it is costly. In particular, if action $\tilde{\boldsymbol{a}}$ has a lower payoff in state i than action $\boldsymbol{a}$, the relative size of its attention vector,

$$\frac{\beta_i(\tilde{\boldsymbol{a}})}{\beta_i(\boldsymbol{a})} = e^{(\tilde{a}_i - a_i)/\lambda},$$

converges to 0 as $\lambda \to 0^+$ and 1 as $\lambda \to \infty$, so that action $\tilde{\boldsymbol{a}}$ attracts almost no attention relative to $\boldsymbol{a}$ in state i when attention costs are small, and largely equal attention when they are large.

The convex hull over all attention vectors spanned by $\mathcal{A}$,

$$\mathcal{B} := \left\{ \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{a})\boldsymbol{\beta}(\boldsymbol{a}) \;\middle|\; p \in \Delta\mathcal{A} \right\} \subset \mathbb{R}^I_+,$$

forms an 'attention possibilities set'. We assign utility $w(\boldsymbol{b}) := \boldsymbol{\pi} \cdot \ln(\boldsymbol{b})$ to each attention vector $\boldsymbol{b}$ and then solve for the most attractive, feasible attention vector by restating the optimization in Equation (2) as

$$\max_{\boldsymbol{b}\in\mathcal{B}} w(\boldsymbol{b}). \tag{G}$$

This strictly convex optimization problem is what we call the *Geometric Approach*, or (G) for short. Figure 1 attests to why we call the approach 'geometric'. The mapping $\boldsymbol{\beta}$ transforms each available action $\boldsymbol{a}$ into an attention vector $\boldsymbol{\beta}(\boldsymbol{a})$ (drawn as black dots). Their convex hull describes the set of feasible attention vectors $\mathcal{B}$ (shaded in gray), and the indifference curves from the strictly convex utility function w (dashed lines) indicate the unique optimal attention vector $\boldsymbol{b}^*$, which always lies on the upper boundary $\partial^+\mathcal{B} = \{\boldsymbol{b} \in \mathcal{B} \mid \nexists \boldsymbol{b}' \in \mathcal{B} : \boldsymbol{b}' > \boldsymbol{b}\}$ of the feasible set (drawn in blue) due to monotonicity of w.

The geometric representation yields an immediate bound on the number of actions that the RI agent implements with positive probability.[4] The new formulation also

[4]Carathéodory's Theorem states that at most I points are required to span any point in a $I-1$-dimensional convex hull [e.g. Eggleston, 1958, Theorem 18]. Since $\boldsymbol{b}^*$ is part of the $I-1$-dimensional upper boundary $\partial^+\mathcal{B}$, this implies that there always exists an optimal solution to (RI) that uses no more than I actions, no matter the cardinality of the menu. For a more general treatment regarding

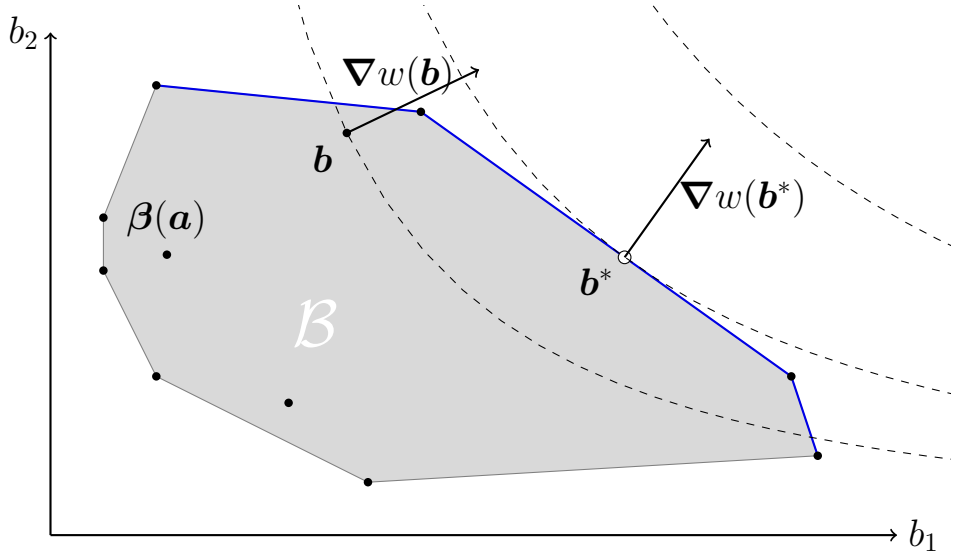


Figure 1: Visual representation of the geometric approach.

separates the role of prior beliefs from that of payoffs: Prior $\boldsymbol{\pi}$ enters the objective function through utility w — akin to preferences — but the attention possibilities set $\mathcal{B}$ is entirely determined by the action payoffs $\boldsymbol{a}$ as well as the attention cost parameter λ. This stark split proves useful to understand how the optimal attention vector responds to changes in external parameters.

The weights that describe $\boldsymbol{b}^*$ as a convex combination $\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{a})\boldsymbol{\beta}(\boldsymbol{a})$ over attention vectors indicate the optimal marginals for the original (RI) problem. Conditionals can then be derived by Equation (1). These weights may not be unique, which happens when there are multiple optimal learning strategies in the (RI) problem.

The optimal attention vector can also be of interest by itself. The payoffs from the pre-image $\boldsymbol{\beta}^{-1}(\boldsymbol{b}^*)$ describe a fictitious action that, whether added to the original menu or offered in replacement of it, leaves the decision maker no better or worse off. It is precisely the *ignorance equivalent* of the RI problem, as defined in Müller-Itten et al. [2021], and can play a strategic role in contract games with RI agents.

The similarity of the attention vectors that span $\boldsymbol{b}^*$ captures the amount of learning that the decision maker undertakes. In the extreme case where the optimum is spanned by a single action, $\boldsymbol{b}^* \in \boldsymbol{\beta}(\mathcal{A})$, the decision maker forgoes learning altogether and blindly implements a single action. In all other cases, the optimal choice always involves learning. Learning is largest when $\boldsymbol{b}^*$ is spanned by wildly different attention

the cardinality of the consideration set in infinite state spaces, see Jung et al. [2019].

vectors, in which case the agent will closely tailor his action to the realized state. The following example with a classic transport theme illustrates the link between the geometric representation and the agent's learning and choice behavior.

Example 1. Bill is choosing a transport option for an upcoming trip. He is familiar with the train route, so this choice yields a certain payoff of zero. Upon consulting the bus map, he sees that there are two bus routes that go to his desired destination: One is direct and yields payoff 0.1, the other includes a long detour and yields payoff -0.9. Unfortunately, Bill is colorblind and cannot tell whether the red or green bus will take the detour. Regardless of trip length, the green bus is a double-decker bus, to which Bill associates a payoff bonus of $1/3$.

We can formalize this as a two-state RI problem, where the red bus takes the detour in state 1 and the green bus in state 2:

$$\boldsymbol{a}^{\text{train}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{a}^{\text{red bus}} = \begin{bmatrix} \text{-0.9} \\ 0.1 \end{bmatrix}, \quad \boldsymbol{a}^{\text{green bus}} = \begin{bmatrix} 0.1{+}1/3 \\ \text{-0.9}{+}1/3 \end{bmatrix}, \text{ and } \boldsymbol{\pi} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}.$$

In the no-information benchmark where learning is impossible, Bill maximizes expected utility and decides to take the train. Conversely, if learning is free, Bill acquires full information and takes the bus with the direct route. Costly learning forms an intermediate scenario where Bill bases his choice on some, but not all, information about the state. Figure 2 illustrates how the attention vectors $\boldsymbol{\beta}(\boldsymbol{a}^x)$ accentuate payoff differences for small information costs λ and attenuate them for large costs.

When information costs are low ($\lambda = 0.25$), the optimal attention vector is spanned by those of the busses and so Bill never takes the train. But contrary to the full-information solution, he still errs occasionally by taking the detour. We can determine his proclivity for detours by comparing the coordinates along each axis, since Equation (1) pins down the conditional implementation probabilities for each chosen transport option x as $p(\boldsymbol{a}^x)\boldsymbol{\beta}(\boldsymbol{a}^x)/\boldsymbol{b}^*$. As both busses are close to an axis, Bill's error rates are small but not zero: He avoids almost all detours on the red bus, $P_1(\boldsymbol{a}^{\text{red bus}}) \approx 0.04\%$, but accepts some on the green double-decker bus, $P_2(\boldsymbol{a}^{\text{green bus}}) \approx 7.4\%$.

Under moderate information costs ($\lambda = 0.7$), the optimal attention vector is spanned by those of the train and green bus. This indicates that Bill still acquires some information but also hedges by taking the riskless transport option with probability $p(\boldsymbol{a}^{\text{train}}) \approx 68\%$. As the spread of the attention vectors is smaller, Bill acquires

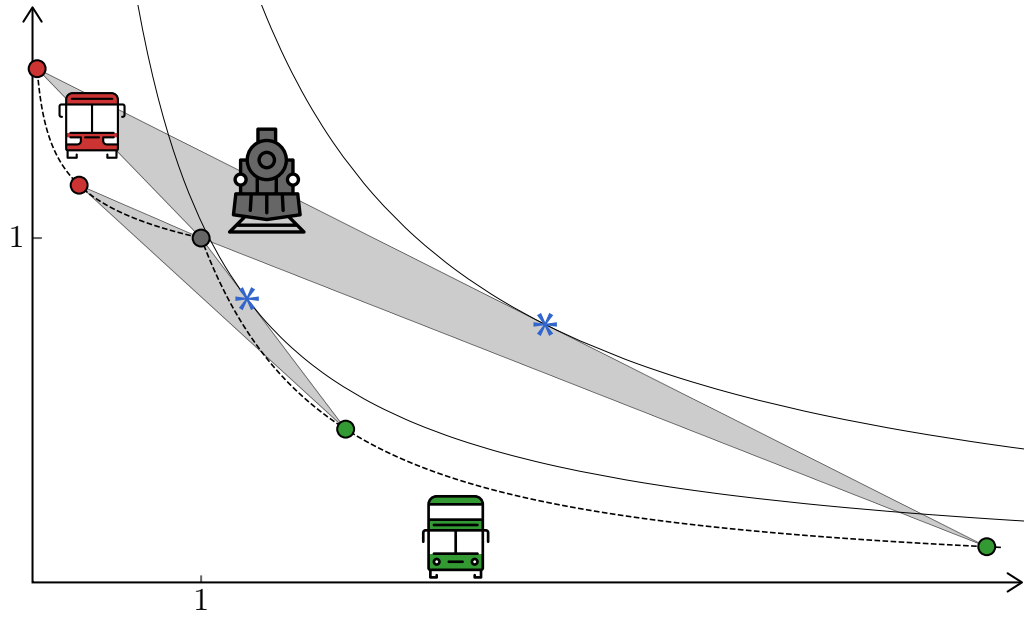


Figure 2: Visual representation of Example 1. As learning becomes more costly, the attention vectors of the busses move along the dashed lines, towards that of the train. Dots indicate their position for cost parameters $\lambda = 0.25$ (top and right) and $\lambda = 0.7$ (closer to origin). The feasible set $\mathcal{B}$ is drawn in gray and the optimal attention vector is indicated as a blue asterisk. Solid lines represent indifference curves.

less information. He misses out on many direct green bus trips, $P_1(\boldsymbol{a}^{\text{green bus}}) \approx 46\%$, and takes quite a few detours, $P_2(\boldsymbol{a}^{\text{green bus}}) \approx 17\%$.

If information costs increase even further, the attention vectors for the busses move closer to that of the train along the dashed lines. As a result, the spread decreases, information acquisition goes down and the error rate goes up, until for some (finite) attention costs he abandons learning altogether and sticks to train rides.

Figure 2 also illustrates how Bill's transport choice changes if he discerns the colors on the map with partial accuracy. If the direct route appears to be painted more green than red, $\pi_1 > \pi_2$, this tilts all the indifference curves towards the right but leaves the feasible set $\mathcal{B}$ unchanged. For small changes, Bill still uses the same modes of transport, albeit in different contingencies. ◇

3.1 Optimality conditions

The simple structure of (G) allows us to draw upon a vast literature within convex geometry when it comes to locating the optimal attention vector. For instance, the optimal attention vector can be succinctly characterized via linear inequality conditions.

Theorem 1 (Optimality conditions)**.** *The solution $\boldsymbol{b}^* \in \mathcal{B}$ to (G) is unique and fully identified by either of the following two optimality conditions:*

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

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

Proof. See Appendix A.1. □

Figure 1 captures the geometric intuition for this result: Condition (a) says that $\mathcal{B}$ lies weakly below the hyperplane that is tangent to the indifference curve at $\boldsymbol{b}^*$. Condition (b) says that $\boldsymbol{b}^*$ lies above all hyperplanes that are tangent to the indifference curves at any suboptimal $\boldsymbol{b} \in \mathcal{B}$. Both hold thanks to the convexity of (G).

Both sets of optimality conditions are central to our paper. Constraints (a) are linear over the points in the convex hull $\mathcal{B}$. These constraints have been stated before in terms of action payoffs and form the backbone for Caplin et al. [2018].[5] One key observation is that the conditions jointly imply that the inequality $\boldsymbol{\nabla} w(\boldsymbol{b}^*) \cdot \boldsymbol{b} \leq 1$ binds at the optimal attention vector $\boldsymbol{b} = \boldsymbol{b}^*$. The same must then be true for all attention vectors that span $\boldsymbol{b}^*$, including $\boldsymbol{\beta}(\boldsymbol{a})$ for all actions $\boldsymbol{a}$ that are part of an optimal consideration set.

To our knowledge, the second set of optimality conditions is new to the RI literature. Constraints (b) are constructive in the sense that *any* feasible point $\boldsymbol{b} \in \mathcal{B}$ restricts the potential location of $\boldsymbol{b}^*$ to a linear half-space dictated by the vector $\boldsymbol{\nabla} w(\boldsymbol{b})$. A successive choice of feasible points $\boldsymbol{b}^n \in \mathcal{B}$ then allows us to "close in" on the optimum and make precise statements about the true optimum based on numerical estimates.

[5] Necessity was highlighted previously by Matějka and McKay [2015].

3.2 Scaling and Posteriors

The functional form of (G) has another separability feature that is particularly helpful for numerical evaluation: Scaling the feasible set $\mathcal{B}$ with a positive constant along any dimension maintains optimality, even if the objective function is left intact. Mathematically, component-wise scaling $\boldsymbol{b} \mapsto [k_i b_i]$ merely offsets the objective value by a constant factor,

$$w([k_i b_i]) = \sum_{i \in \mathcal{I}} \pi_i \cdot \ln(k_i b_i) = \boldsymbol{\pi} \cdot \ln(\boldsymbol{b}) + \boldsymbol{\pi} \cdot \ln(\boldsymbol{k}) = w(\boldsymbol{b}) + w(\boldsymbol{k}).$$

As a consequence, the optimum scales by the same vector as the feasible set.

Lemma 1 (Axis Scaling)**.** *Consider any scaling vector $\boldsymbol{k} \in \mathbb{R}^I_+$. Attention vector $\boldsymbol{b}$ solves (G) if and only if $[k_i b_i]$ solves $\max_{\boldsymbol{b}' \in \{[k_i b_i] \,|\, \boldsymbol{b} \in \mathcal{B}\}} w(\boldsymbol{b}')$.*

Proof. Since $w([k_i b_i]) = w(\boldsymbol{k}) + w(\boldsymbol{b})$ for all $\boldsymbol{b} \in \mathcal{B}$, $w(\boldsymbol{b}^*) \geq w(\boldsymbol{b})$ for all $\boldsymbol{b} \in \mathcal{B}$ if and only if $w([k_i b_i^*]) \geq w(\boldsymbol{b}')$ for all $\boldsymbol{b}' \in \{[k_i b_i] \,|\, \boldsymbol{b} \in \mathcal{B}\}$. □

Scalability greatly helps reduce floating point imprecision in our numerical algorithm. It also captures the fact that shifting all action payoffs by a constant vector $\boldsymbol{u} \in \mathbb{R}^I$ does not affect the agent's optimal learning strategy — after all, the payoff boost is independent from the action choice.

Lemma 1 also allows us to identify the agent's posterior beliefs under optimal learning. Indeed, scaling attention vectors statewise by $\boldsymbol{k}^* = \boldsymbol{\nabla} w(\boldsymbol{b}^*) = [\pi_i / b_i^*]$ moves the optimum to the prior $\boldsymbol{\pi}$. Similarly, it moves the attention vector for each action to $[k_i^* \beta_i(\boldsymbol{a})]$, which corresponds to the agent's posterior for all actions in the consideration set by Bayes' rule,

$$\frac{P_i(\boldsymbol{a})}{p(\boldsymbol{a})} \pi_i \overset{(1)}{=} \frac{\beta_i(\boldsymbol{a})}{\sum_{\boldsymbol{a}' \in \mathcal{A}} p(\boldsymbol{a}') \beta_i(\boldsymbol{a}')} \pi_i = \frac{\beta_i(\boldsymbol{a})}{b_i^*} \pi_i = k_i^* \beta_i(\boldsymbol{a}).$$

Although expressed differently, a similar transformation is present in previous concavification procedures [Caplin et al., 2018, Gentzkow and Kamenica, 2014, Kamenica and Gentzkow, 2011] that express the optimal (RI) solution in terms of the agent's optimal posterior beliefs.

While scaling to the simplex offers 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 vector $\boldsymbol{k}^*$. For conceptualization, this loop is not tragic – for computation, it is fatal. One fundamental advantage of our approach is that it yields a convex optimization problem where both the set of candidate points $\mathcal{B}$ and the objective function w are explicitly defined.

4 Practical Implications

While there exist powerful algorithms to approximate convex optimization problems like (G),[6] some numerical noise is inevitable. One strength of the geometric approach is that it allows us to narrow down the true optimum based on noisy estimates.

4.1 Consideration set approximations

While agent behavior is ultimately described by the conditional probabilities $\boldsymbol{P}$, there are instances where the primary focus is on identifying which actions are chosen with positive probability. This consideration set reveals which products are on a consumer's radar, predicts possible price jumps in markets with sticky prices, and uncovers correlations in portfolio investments. By classifying some actions as 'never chosen,' it also yields a testable hypothesis for choice data that is too sparse to yield reliable estimates of conditional (or even marginal) implementation probabilities.

We use the optimality conditions from Theorem 1 to approximate the optimal consideration set despite numerical imprecision. To do so, we find it useful to assign scores to each action $\boldsymbol{a} \in \mathcal{A}$ based on any feasible attention vector $\boldsymbol{b} \in \mathcal{B}$. The "$\boldsymbol{b}$-score" of action $\boldsymbol{a}$, written $s(\boldsymbol{a}|\boldsymbol{b})$, captures the location of $\boldsymbol{\beta}(\boldsymbol{a})$ relative to the hyperplane tangent to the indifference curve at $\boldsymbol{b}$,

$$\begin{cases} s : \mathcal{A} \times \mathcal{B} \to \mathbb{R} \\ s(\boldsymbol{a}|\boldsymbol{b}) = \boldsymbol{\nabla} w(\boldsymbol{b}) \cdot \boldsymbol{\beta}(\boldsymbol{a}) - 1. \end{cases}$$

Actions with positive scores lie above the tangent hyperplane, those with negative scores lie below. Rewriting Theorem 1 in this way, condition (a) states that all actions have non-positive $\boldsymbol{b}^*$-scores, and condition (b) states that the optimal choice

[6] We outline one such method in Section 4.3 and compare its effectiveness against other approaches in Section 5.

has a non-negative expected $\boldsymbol{b}$-score for any $\boldsymbol{b} \in \mathcal{B}$ since

$$\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{a})s(\boldsymbol{a}|\boldsymbol{b}) = \boldsymbol{\nabla} w(\boldsymbol{b}) \cdot \left[\sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{a})\boldsymbol{\beta}(\boldsymbol{a})\right] - 1 = \boldsymbol{\nabla} w(\boldsymbol{b}) \cdot \boldsymbol{b}^* - 1 \geq 0. \tag{3}$$

This expression is useful because it allows us to identify the most frequently chosen actions and disregard the rest.

Partial Cover. To account for numeric noise, we generalize the notion of a consideration set to describe any subset of the menu that contains the most frequently chosen actions.

Definition 1. A set $A \subseteq \mathcal{A}$ is a "q-cover" of the (RI) problem $(\mathcal{A}, \boldsymbol{\pi}, \lambda)$ if

$$p(A) = \sum_{i\in\mathcal{I}} \pi_i \sum_{\boldsymbol{a}\in A} P_i(\boldsymbol{a}) \geq q \in [0, 1]$$

for all optimal conditionals $\boldsymbol{P}$.

The consideration set is always a 1-cover, as are any of its supersets. Our goal is to identify q-covers with high probability q and small cardinality $|A|$, in order to isolate the most frequently chosen actions.

Starting with any feasible attention vector $\boldsymbol{b}$, Equation (3) implies that the agent can choose actions with negative $\boldsymbol{b}$-scores only if she compensates by often enough choosing actions with sufficiently positive $\boldsymbol{b}$-scores — and when the maximal score $\bar{s}(\boldsymbol{b}) := \max_{\boldsymbol{a}\in\mathcal{A}} s(\boldsymbol{a}|\boldsymbol{b})$ is close to zero, actions with very low scores just cannot be chosen often. This yields a threshold rule that can be used to generate a q-cover for the (unknown) optimal choice based on a numerical approximation.

Corollary 1. *For any $\boldsymbol{b} \in \mathcal{B}$ and any $q \in (0, 1)$, the set*

$$A = \left\{ \boldsymbol{a} \in \mathcal{A} \;\middle|\; s(\boldsymbol{a}|\boldsymbol{b}) \geq -\frac{q\bar{s}(\boldsymbol{b})}{1-q} \right\} \subseteq \mathcal{A}$$

is a q-cover.

Proof. If $\bar{s}(\boldsymbol{b}) = 0$, all $\boldsymbol{b}$-scores are non-positive and the attention vector $\boldsymbol{b}$ is optimal. The set A then contains only actions with a $\boldsymbol{b}$-score of zero, and is equal to the

agent's consideration set (or union of consideration sets, if there are multiple optimal solutions). Otherwise, $\bar{s}(\boldsymbol{b}) > 0$ and we proceed by bounding $\boldsymbol{b}$-scores above,

$$0 \overset{(3)}{\leq} \sum_{\boldsymbol{a}\in\mathcal{A}} p(\boldsymbol{a})s(\boldsymbol{a}|\boldsymbol{b}) \leq \sum_{\boldsymbol{a}\in\mathcal{A}\setminus A} p(\boldsymbol{a})\left(-\frac{q\bar{s}(\boldsymbol{b})}{1-q}\right) + \sum_{\boldsymbol{a}\in A} p(\boldsymbol{a})\bar{s}(\boldsymbol{b})$$
$$= \bar{s}(\boldsymbol{b})\left[-\frac{q}{1-q}(1-p(A)) + p(A)\right] = \frac{\bar{s}(\boldsymbol{b})}{1-q}\left[p(A) - q\right]$$

Dividing by $\bar{s}(\boldsymbol{b})/(1-q)$ yields $p(A) \geq q$. □

Corollary 1 can be readily deployed to assess computation accuracy. Let $\boldsymbol{b}$ be the researcher's numerical estimate of the optimum and let A be a partial cover with high q, say 95%. The researcher may find that A is pretty large, perhaps close to the full menu $\mathcal{A}$. This may indicate that computational error is substantial if the numerical solution had a small support or, simply, the researcher had reason to expect a sparse consideration set. Alternatively, the researcher may find that A has very few actions or that, from the perspective of the specific application, actions in A are clustered around only a handful of relevant values. In this case, the researcher has effectively identified the key features of the consideration set under the true optimum.

Partial covers are also useful while searching for the right parameters to replicate a salient fact, say, a particular action $\boldsymbol{a}$ being observed with a frequency higher than 10%. The researcher does not need a very precise estimate $\boldsymbol{b}$ of the optimum for each parameter value: As soon as the 90%-cover excludes the aforementioned action $\boldsymbol{a}$, the parameter value can be rejected.

In practice (see Section 5), we find that accurate algorithms yield estimates $\boldsymbol{b}$ that are very close to the (G) solution. The resulting q-covers typically have small cardinality even when q is close to one, making this approach very attractive for empirical research.

Dominated Actions. In many situations, it is even possible to rule out some dominated actions altogether — effectively finding a 1-cover that is significantly smaller than the menu. Sometimes, this is trivial: If an action delivers less payoff in each state than a blind lottery over other actions, the action is suboptimal at any posterior and thus would never be chosen. Tighter bounds are possible, since the optimality conditions (a) and (b) imply that any action in the consideration set has a $\boldsymbol{b}^*$-score

of zero. Using numerical approximations, it is often possible to limit the location of the possible optimal attention vector $\boldsymbol{b}^*$, which in turn restricts the possible gradient $\boldsymbol{\nabla} w(\boldsymbol{b}^*) = [b_i/\pi_i]$ to some subset $\Psi \subset \mathbb{R}^I_+$. If we can bound the feasible $\boldsymbol{b}^*$-scores for some action $\boldsymbol{a}$ below zero,

$$s(\boldsymbol{a}|\boldsymbol{b}^*) \leq \sup_{\boldsymbol{\psi}\in\Psi} \boldsymbol{\psi} \cdot \boldsymbol{\beta}(\boldsymbol{a}) - 1 < 0, \tag{4}$$

we can rule it out for good. The RI agent never implements this action in an optimal solution.

Computationally, finding dominated actions is significantly slower than finding a partial cover. A valid q-cover can be obtained from any feasible attention vector based solely on explicit score computations, while testing for dominance requires determining first a feasible set of gradients Ψ, and then solving maximization problem (4) for each individual action. Still, the dominated actions approach is useful in situations where accuracy is paramount.

4.2 Precision Metric

Even though (G) is computationally simpler, the primary object of interest is ultimately the conditional choice P that predicts behavior, not the optimal attention vector. Be it to compare model predictions to empirical data, or to write a stopping criterion for numerical methods, the researcher eventually needs to decide when two conditional choices are "similar." We now show that the distance between implied attention vectors offers a parsimonious metric to gauge similarity of the implied conditional choice.

Definition 2. The *attention distance* between choices $\boldsymbol{P}, \boldsymbol{P}' \in (\Delta\mathcal{A})^I$ is defined as

$$d_{(\mathrm{G})}(\boldsymbol{P}, \boldsymbol{P}') := \sqrt{\sum_{i\in\mathcal{I}} \pi_i \Big(\beta_i^{-1}\big(\mathsf{E}[\beta_i(\boldsymbol{a})|\boldsymbol{a}\sim P_i]\big) - \beta_i^{-1}\big(\mathsf{E}[\beta_i(\boldsymbol{a})|\boldsymbol{a}\sim P_i']\big)\Big)^2}.$$

The attention distance is obtained in three steps: First, we compute the expected attention vector $[\mathsf{E}[\beta_i(\boldsymbol{a})|\boldsymbol{a}\sim P_i]]$ under each conditional $\boldsymbol{P}$. Then, we transform the attention vector back into the original payoff space using the inverse mapping $\boldsymbol{\beta}^{-1}$. This transformation gets rid of any distortions introduced by axis scaling. Finally, we

apply the standard Euclidean metric to the resulting payoff vectors, weighted by the prior probability for each state. The weights ensure that the distance is unaffected by a payoff-irrelevant splitting of states. Nonnegativity, symmetry, and the triangle inequality are directly inherited from the standard Euclidean distance.

A potential shortcoming of the attention distance is its failure to distinguish between choices that imply the same attention vector. From a computational perspective, this may be a satisfactory compromise in the interest of parsimony. Indeed, we now show that distance $d_{(\mathrm{G})}$ provides a suitable convergence criterion whenever (RI) admits a unique solution. From any sequence of conditionals $\boldsymbol{P}^n$ that converge to the solution $\boldsymbol{P}$ under $d_{(\mathrm{G})}$, we can construct conditionals that converge in terms of both marginal and conditional probabilities. So although the attention distance only considers I-dimensional vectors, it is sufficient to generate convergence for all $I \times |\mathcal{A}|$ conditional probabilities thanks to continuity of the first-order conditions in Equation (1).

Lemma 2. *Suppose (RI) admits a unique solution $\boldsymbol{P}$. For any sequence of conditionals $\{\boldsymbol{P}^n\} \subset (\Delta\mathcal{A})^I$ that converges to $\boldsymbol{P}$ according to $d_{(G)}$, the conditionals*

$$\left[\frac{(\boldsymbol{\pi} \cdot \boldsymbol{P}^n(\boldsymbol{a}))\beta_i(\boldsymbol{a})}{\sum_{\boldsymbol{a}' \in \mathcal{A}}(\boldsymbol{\pi} \cdot \boldsymbol{P}^n(\boldsymbol{a}'))\beta_i(\boldsymbol{a}')}\right] \tag{5}$$

converge to $\boldsymbol{P}$ under the standard Euclidean metric over $(\Delta\mathcal{A})^I$.

Proof. See Appendix A.1. □

The attention distance $d_{(\mathrm{G})}$ is ideally suited to serve as a stopping criterion for numerical solution methods since it penalizes numerical noise when it leads to substantial payoff differences and ensures that conditional choices converge to the actual optimum. In this sense, the attention distance strikes a balance between other common stopping criteria: Methods that rely on objective values alone can lead to noisy estimates of the model's behavioral implications, since several conditional choices may — and often do [Jung et al., 2019] — share very similar objective values. At the other end, a straightforward comparison based on the Euclidean distance between the conditional matrices $\boldsymbol{P}$ and $\boldsymbol{P}'$ (or the vectors of marginals p and p') treats all actions as equally distinct. However, numerical (RI) estimates over large menus err along both the extensive and intensive margin: They typically misidentify both the support of the optimal choice (the consideration set) as well as the relative frequency

of chosen actions. The matrix distance disproportionately penalizes errors on the extensive margin, while $d_{(\mathrm{G})}$ recognizes when a candidate consideration set contains near-optimal actions.

4.3 Algorithm Design

The simple geometry of (G) offers fertile ground for numerical methods that solve finite RI problems. Standard algorithmic techniques for convex problems perform well, and the reduced dimensionality of (G) brings obvious gains in performance.

We provide an algorithm based on Sequential Quadratic Programming (SQP) and active set methods (see, e.g. Judd [1998]), using $d_{(\mathrm{G})}$ as a stopping criterion. The codes are available at `https://github.com/mmulleri/GAP-SQP` and a detailed explanation of the code is provided in Appendix A.2. Although the optimization methods we use are relatively unsophisticated, we find that our algorithm performs favorably when compared to other state-of-the-art techniques that are typically used to estimate RI models, in terms of both speed and accuracy. The methods and ideas can also be combined with other solution methods to yield further gains in performance.[7]

A few practical challenges arise when dealing with large state and action spaces, for example when approximating continuous RI problems. This can routinely lead to memory issues when storing and accessing the payoff matrix,[8] and we briefly elaborate on them here.

Dealing with large menus. Large menus can often be partitioned into clusters of actions with similar payoff vectors. Sometimes, this clustering is explicit since the menu represents a discrete approximation to a continuous choice variable, and so the granularity of the discretization grid determines which actions are 'lumped together'. Even if the goal is to characterize the optimal choice over a very fine grid, active set methods can significantly speed up the algorithm and reduce memory usage.

[7] For instance, one may use the Blahut-Arimoto algorithm with a very high tolerance error to create a starting guess for the GAP-SQL algorithm. Or one may replace the SQP approach with more advanced convex optimization algorithms.

[8] In the portfolio optimization (Section 5.2) the associated $300^2 \times 300^2$ payoff matrix would require 64.8Gb of memory.

Practically, we start with a coarse grid over actions and increasing the grid precision stepwise. At each step, we compute the optimal attention vector $\boldsymbol{b}^k$ and then include K actions from the finer grid with the highest $\boldsymbol{b}^k$-score $s(\boldsymbol{a}|\boldsymbol{b}^k)$. We increase grid precision once the 99% cover stabilizes. When K is large relative to the optimal consideration set, this approach can approximate large action spaces without running into memory management issues. And while the numerical estimates of the algorithm depend on the path of subgrids, any partial covers computed in the last round accurately describe the optimal choice over the *entire* menu $\mathcal{A}$.

Large state spaces. Although this step-wise optimization effectively avoids the load of large menus, each step still relies on the entire payoff vectors for the currently considered actions. As such, large state spaces inherently pose a bigger challenge for our algorithm than large menus. When memory constraints are binding, a lower threshold K may offer some relief: By reducing the number of actions that are added to the candidate consideration set at each step, fewer payoff vectors are considered simultaneously, but more iterations may be necessary to achieve convergence.

Approximating continuous problems. One caveat is in order when using any discrete approximations to continuous problems: Even with a fine grid $\mathcal{A} \times \mathcal{I}$ over actions and states, it is possible that unmodeled actions generate learning opportunities that increase the attractiveness of actions $\boldsymbol{a} \in \mathcal{A}$, and that the approximation thus dismisses $\boldsymbol{a}$ in error. Similarly, unmodeled states may affect the relative appeal of actions in ways that are hard to predict, and our methods are not well suited to judge the accuracy of our estimates for continuous state spaces. That said, it appears in practice that the geometric approach also provides sensible numerical estimates when applied to a fine discretization of a continuous (RI) problem, as we show in the applications below.

5 Applications

In this section, we illustrate by way of example that both the conceptual framework and the computationally tractable algorithm have the potential to expand the purview of further research. We consider three applications: The first is a monopolist problem with uncertain demand as proposed by Matějka [2016]. We use this well-

known application to benchmark the GAP-SQP algorithm described in Section 4.3 against existing methods, focusing primarily on speed. The second is a portfolio choice problem with a massive state and action space proposed by Jung et al. [2019]. We primarily use it to highlight the precision of GAP-SQP and showcase the more robust behavioral predictions that we develop in Section 4.1. The third is a task assignment problem that is novel to the RI literature. It represents the ideal scenario for the (G) approach – finite state spaces coupled with rich action spaces – and illustrates that this combination arises naturally in economically relevant problems.

5.1 Sticky Prices [Matějka, 2016]

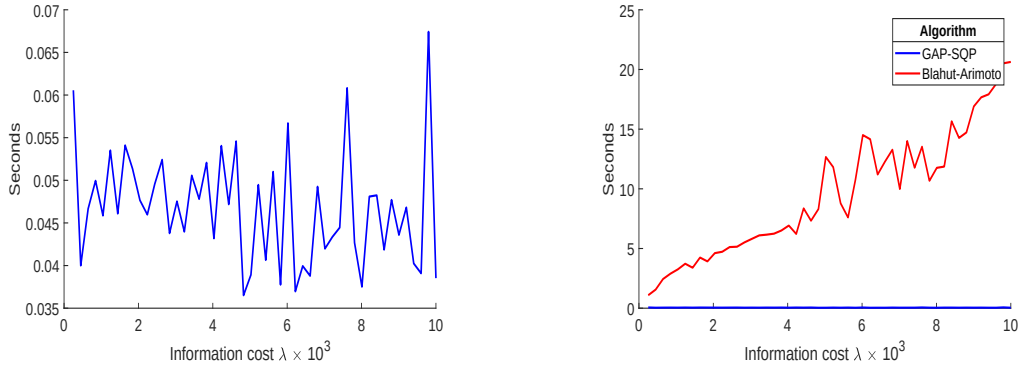
Our first application is based on the "rationally inattentive seller" model of Matějka [2016]. A monopolistic seller has a per unit input cost of 1 and sets the price p facing an isoelastic demand function whose elasticity, $\frac{d+1}{d}$, is a random variable uniformly distributed. Profits are given by $\Pi(d,p) = p^{-\frac{d+1}{d}}(p-1)$, where the demand variable d is the ex-ante unknown state and the price p corresponds to the seller's action.[9] As in Matějka [2016], actions and states are discretized. As a benchmark we use a grid of 200×200 points, and we will improve the grid precision to increase the computational demands of the problem without introducing any further complexity in the model.

For comparison to our base routine described in Section 4.3, we also solve the model using the Blahut-Arimoto (BA) algorithm, a solution method that originated in Rate Distortion theory and has recently gained some usage in RI problems. As with our GAP-SQP algorithm, the BA algorithm is guaranteed to converge to the optimum and operates with a reduced dimensionality, updating the marginal distribution over actions. We implement both algorithms in MATLAB.[10]

Figure 3 documents the running times in seconds across a range of information costs λ for our benchmark case with a grid of 200×200 points. As shown in panel (a), the GAP-SQP algorithm terminates in less than 0.05 seconds in all runs, with

[9] The demand variable d is uniformly distributed in $(\frac{1}{9}, \frac{1}{2})$, following Matějka [2016], Section 4.2. Matějka [2016] assumes a channel capacity constraint rather than information being acquired at a cost. To match the channel capacity constraint of half a bit, we find that we need to set $\lambda = 0.0053$.

[10] We use a desktop computer with 16 GB of RAM and an Intel(R) 3.20 GHz Core i7-8700 processor on a Windows 10 Enterprise 64-bit operating system. For further discussion of the BA algorithm, see Caplin et al. [2018] and Cover and Thomas [2012]. Matějka [2016] instead used proprietary software AMPL/LOQO to solve for the joint probability distribution. Due to licensing restrictions, we were not able to use AMPL/LOQO to document running times. A comparable, freely available solver (IPOPT) was substantially slower and less precise than both the GAP-SQP and BA algorithms.



(a) GAP-SQP algorithm (b) Algorithm comparison

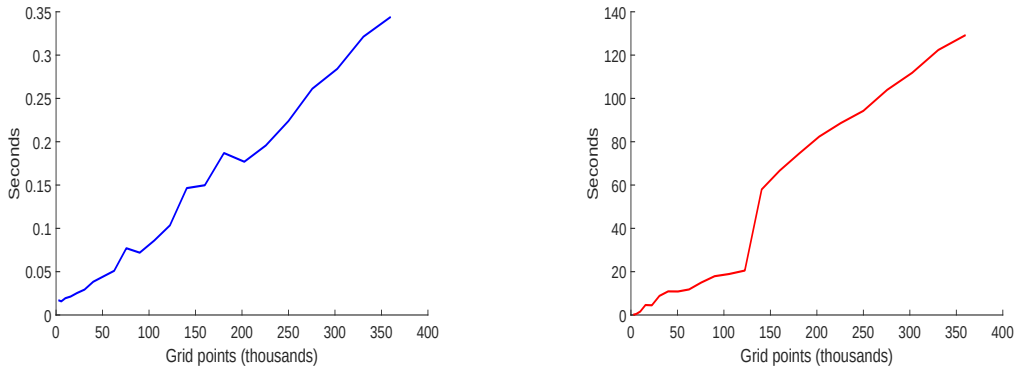
Figure 3: Running times across information costs

minimal differences across information costs. The BA algorithm, reported in panel (b), runs in about one second when information costs are very low — and thus the solution is very close to the full information benchmark — but is substantially slower for higher information costs, up to 20 seconds. Both algorithms achieve very similar objective values, with the GAP-SQP algorithm outperforming by about 10^{-8}.[11]

Next we ratchet up the computation burden by increasing each grid precision up to 600 points, adding up to $600^2 = 360,000$ total grid points. As shown in Figure 4(a), running times scale roughly linearly for the GAP-SQP algorithm. Even at a 600×600 grid running times stay well below half a second. Figure 4(b) shows that the BA algorithm also scales well; though this means that computing times approach two minutes for the largest grids.

We also compute the set of dominated actions as well as the 95% cover using the output from our GAP-SQP algorithm. Figure 5 displays the GAP-SQP numerical solution as solid bars over the full price grid — with insets at two points of the full support of prices for visibility. The 99% cover is indicated with a dark blue background. It is identical to the consideration set of the numerical solution. Thus, even if we had low confidence in the accuracy of the algorithm, we would be able to conclude that the total probability of observing any price outside of this set is at most 1%. Indeed, the main point in [Matějka, 2016] is that optimal pricing behavior is discrete, clustering mass on a comparatively small number of points. This observation can also be made by looking at the set of non-dominated actions (light blue background): It

[11] Numerical solutions for $\lambda = 0.0053$ are also very similar to those reported in Matějka [2016]. See online Appendix B.1. We thank Filip Matějka for sharing his numerical output.



(a) GAP-SQP algorithm (b) Blahut-Arimoto algorithm

Figure 4: Running times across grid precision

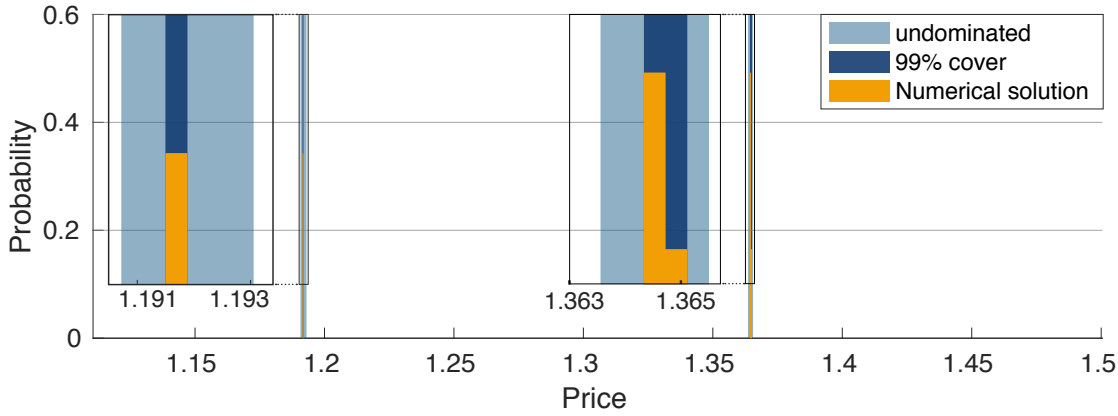


Figure 5: Partial cover and undominated actions

shows that the vast majority of prices are not used in any contingency. The validity of the claim does not rely on having found the true optimum of the (RI) problem, which makes it significantly more robust.

5.2 Portfolio Choice [Jung et al., 2019]

Our second application considers the portfolio choice problem of Jung et al. [2019], who illustrate that RI can explain low rates of household portfolio rebalancing. In this problem, an investor with unit wealth designs a portfolio composed of three uncorrelated assets, without restrictions on short sales or overall leverage. The investor has constant absolute risk aversion (CARA) utility $u(x) = -e^{-\alpha x}$ with risk aversion parameter α. Asset zero is a safe asset with constant return 1.03. The returns from

risky assets $j = 1$ and $j = 2$ are each modeled as the sum of two independent random variables around a slightly higher mean return, $1.04 + Z_j + Y_j$. The random variable $Z_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_z^2)$ reflects factors that are inherently unforeseeable. The random variable Y_j reflects factors that are not known at the outset but can be learned at a cost. Each portfolio $(\theta_1, \theta_2) \in \mathbb{R}^2$ describes an available action, where θ_j is the position in risky asset j and $\theta_0 := 1 - \theta_1 - \theta_2$ is the position in the safe asset. The expected utility from the portfolio conditional on state $\boldsymbol{Y} = (Y_1, Y_2)$ is

$$U(\boldsymbol{\theta}, \boldsymbol{Y}) = \mathsf{E}\left[u\left(1.03\theta_0 + \sum_{j=1}^{2} (1.04 + Z_j + Y_j)\theta_j \right) \middle| \boldsymbol{Y} \right]. \tag{6}$$

We follow Jung et al. [2019] and assume that $\boldsymbol{Y}$ follows a discrete distribution over a 300×300 grid that is obtained from a normal distribution $\mathcal{N}(\mathbf{0}, 0.02^2 I)$ truncated at three standard deviations. We report results for parameter values $\alpha = 1$, $\lambda = 0.1$, and $\sigma_z = 0.0173$.[12]

We approximate the continuous menu $(\theta_1, \theta_2) \in \mathbb{R}^2$ by first (without loss of generality) imposing the upper and lower bounds given by the full-information solution, and then iteratively doubling the grid resolution using 99% covers until we reach $513 \times 513 = (2^9 + 1)^2$ points.[13] Jung et al. [2019] instead use a variant of the Blahut-Arimoto algorithm that optimizes the points of support at each step of the algorithm. We refer to this algorithm as JKMS. The algorithms reach a comparable objective value, with GAP-SQP mildly outperforming JKMS.[14] Both algorithms perform significantly better than approximating the objective with a second-order polynomial to obtain a Linear-Quadratic Gaussian (LQG) problem (for details, see Online Appendix B.2),[15] an approach that is common in the applied literature.[16]

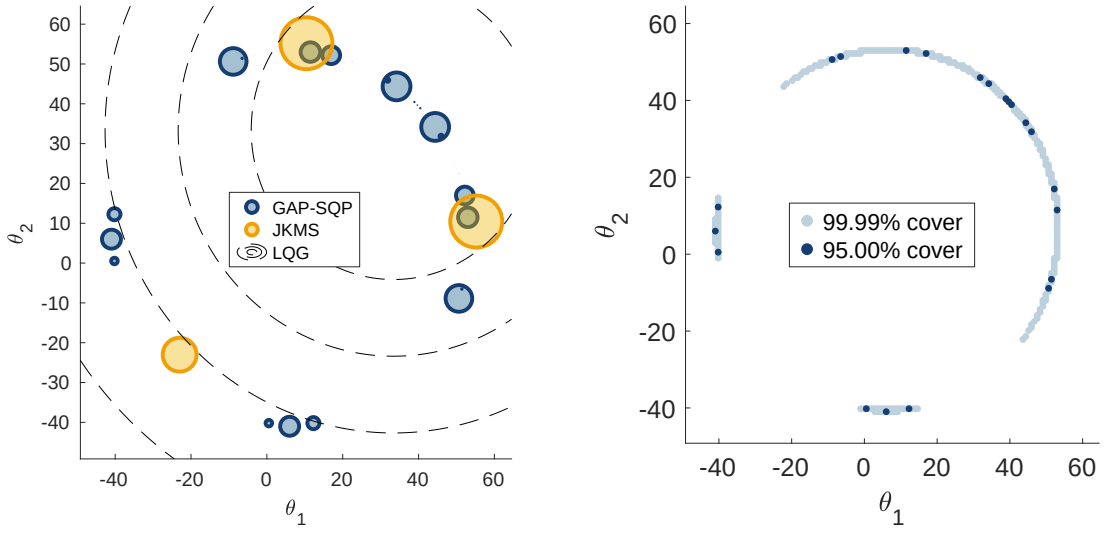
[12] These parameters correspond to scenario B in Jung et al. [2019]. Although not shown, the GAP-SQP solution has larger support and achieves a higher objective value than the JKMS solution in all four parameter scenarios.

[13] The iterative approach (see Section 4.3 and Appendix A.2) allows us to handle a large action grid. However, it does not reduce the memory demands imposed by the large state space. In order to compute the solution at the same state grid resolution as Jung et al. [2019], we opted to run the algorithm on a computational cluster.

[14] The solution published in Jung et al. [2019] closes 0.607153 of the payoff gap between no and full information, while GAP-SQP closes 0.614877 of the gap. For a comparison of the statewise payoff distribution across algorithms, see Online Appendix B.2.

[15] The (continuous) LQG solution closes roughly 0.4843 of the payoff gap between no and full information.

[16] Examples of LQG models include Kacperczyk et al. [2016], Luo et al. [2017], Mondria [2010], Van Nieuwerburgh and Veldkamp [2009, 2010].



(a) estimated portfolio choice

(b) partial covers

Figure 6: Portfolio distributions under GAP-SQP (blue), JMKS (orange), and LQG (black). In panel (a), the circle size of each portfolio $\boldsymbol{\theta}$ is proportional to the probability weight $p(\boldsymbol{\theta})$, and the probability that the LQG solution falls between any two dashed contour lines is equal to 0.2.

Turning to the behavioral implications, Figure 6(a) shows the estimated portfolio choice probabilities across all three algorithms. The LQG solution stands out as the only continuous solution, but even the two discrete solutions are measurably different. Jung et al. [2019] caution that their solution method may miss solutions with a larger support, and this is exactly what we find with GAP-SQP. The main point in Jung et al. [2019], that portfolio rebalancing is relatively rare, remains valid — and indeed the consideration set shrinks for higher information costs. Quantitatively, though, GAP-SQP finds that portfolio rebalancing is substantially more common and, occasionally, the investor makes small adjustments.

The partial covers displayed in Figure 6(b) allow more robust statements regarding the true optimum given the state and action grid that we use. Contrary to the JKMS estimates, it appears that the investor actually rarely takes large short positions simultaneously on both risky assets. This may suggest that the investor looks for good news rather than bad. Another pattern that arises from Figure 6(b) is that the RI investor selects only portfolios from a circle, hinting that some further analytic results are possible — and may help elucidate the relative role of risk aversion and

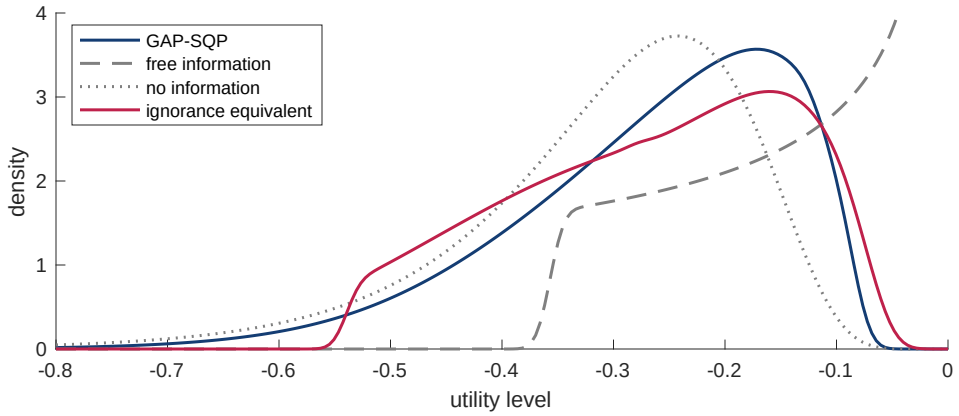


Figure 7: Payoff distribution across choices, smoothed with a kernel density estimate.

information processing costs, for instance.[17] Overall, the example illustrates the need for more robust estimation techniques that not only deliver a "better estimate" but allow a valid characterization of the true optimal choice.

Figure 7 plots the statewise payoff distribution $U(\boldsymbol{\theta}, \boldsymbol{Y}) - \lambda \mathsf{MI}$, assuming $(\boldsymbol{\theta}, \boldsymbol{Y})$ is distributed according to the numeric solution of GAP-SQP (blue) and the information cost MI is borne unconditionally. For comparison, we also show the payoff distribution under no information ($\lambda \to \infty$, grey dotted) and free information ($\lambda \to 0$, grey dashed).

We can compute the ignorance equivalent of this RI investment problem, obtained as the pre-image $\boldsymbol{\beta}^{-1}(\boldsymbol{b}^*)$ of the optimal attention vector. This fictitious asset is uniquely designed to ensure that the investor would abandon learning and select it unconditionally, but without gaining a payoff boost in doing so [Müller-Itten et al., 2021]. As can be seen from the payoff distribution in Figure 7, the ignorance equivalent yields the same expected utility as the optimal portfolio choice with learning, but – to dissuade learning – it avoids the lowest payoffs in a way that mimics the full-information distribution.

5.3 Task Assignment

Our last application is designed to illustrate how the GAP geometry is particularly helpful in RI problems when the action space is naturally large and discrete. A manager has to assign N workers across three tasks $\{0, 1, 2\}$. Either task one or task

[17] A more thorough investigation of this conjecture is outside the scope of this research.

two is critical; task zero is never critical and represents dismissal. All but one of the workers are skilled. The unskilled worker is not productive and prevents a skilled worker from contributing (if there are any assigned to the same task). Output is thus determined by the number of skilled workers assigned to the critical task, minus the unskilled worker if he is also assigned to the critical task, n^*. We assume a simple form of decreasing returns to scale, letting output be given by the production function $\Phi(n^*) = \sum_{n=1}^{n^*} \delta^n$ for $\delta = 0.9$.

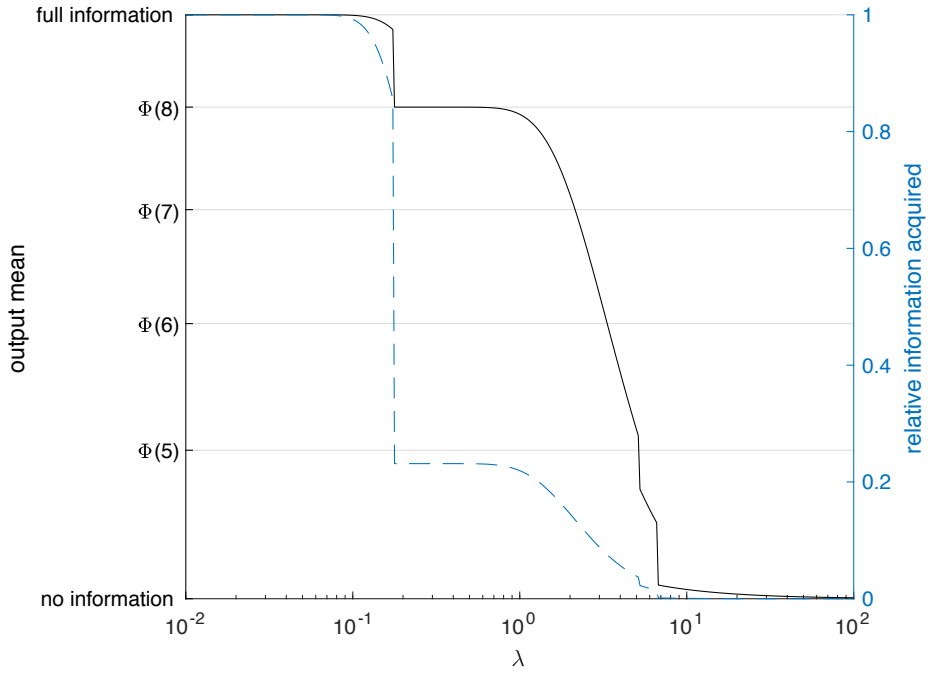
Despite its simple description, the task assignment problem generates a complex optimization problem. There are $2N$ possible states, indicating which task is critical $c \in \{1, 2\}$ and the identity of the unskilled worker $w \in \{1, ..., N\}$. An action is a task assignment a_w for each worker w that can be summarized as a vector $\boldsymbol{a} \in \{0, 1, 2\}^N$. There are 3^N such vectors, and at least 2^N that are optimal under some information structure. We consider a fully symmetric setup with $N = 10$ workers, resulting in 20 states and $3^{10} = 59,049$ potential assignments. Figure 8 summarizes expected output, information flow, and optimal assignment strategies for a range of information costs $\lambda \in [.01, 100]$. As information costs increase, the manager uses four distinct allocation strategies (indicated by letters A to D).

When information costs are low, the manager aims for the full-information solution, dismissing the unskilled worker and assigning everyone else to the critical task (Strategy A). Initially, she acquires nearly all the information and consistently achieves the full-information benchmark output $\Phi(9)$. As information costs go up, the manager occasionally misidentifies the unskilled worker or, with much lower probability, the critical task.[18]

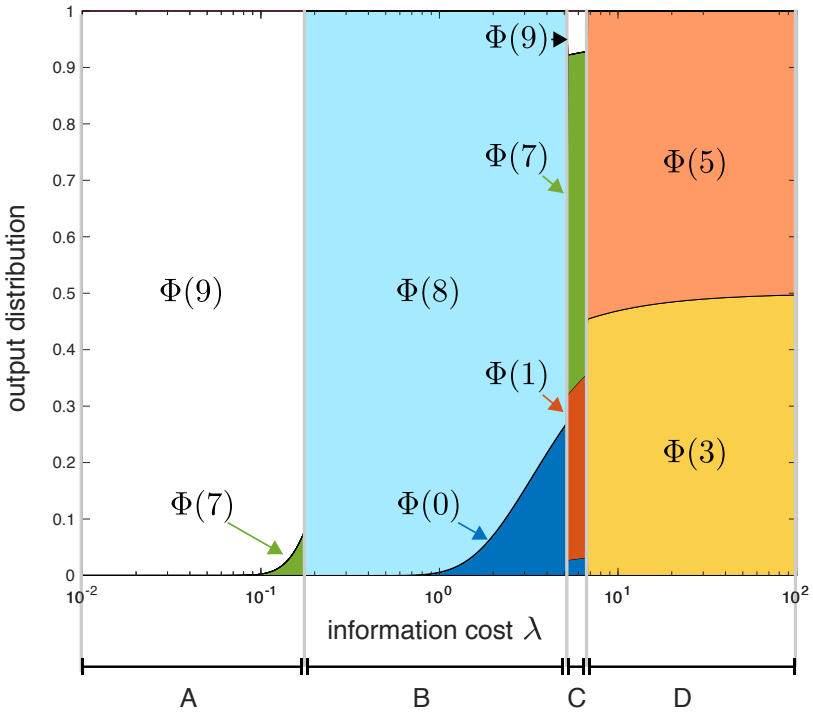
When λ reaches a certain threshold, the manager changes tack, sending all workers to the task that she identifies as critical (Strategy B). Because all learning on workers is forgone, we see a discrete drop in information acquisition that compensates the manager for the reduction in expected output due to the unskilled worker.[19] Because the manager is initially very accurate at identifying the critical task, output once again is near constant at $\Phi(8)$—slightly lower than in the full-information benchmark output $\Phi(9)$. As information costs increase, so does the likelihood of an extreme zero-output assignment resulting from sending all workers to the misidentified critical task.

[18] Although not visible in Figure 8(b), there is a small but positive probability of output $\Phi(0)$ or $\Phi(1)$ resulting from the critical task having been misidentified.

[19] Strategy A's expected output decreases with the probability of misidentifying tasks or workers, while output under Strategy B is affected only by task misidentification, by construction.



(a) Mean output (solid black) and information acquisition (dashed blue).



(b) Output distributions

Figure 8: Optimal management strategies under varying information costs.

Output volatility peaks under this strategy.

Once information costs are high enough, the manager aims to hedge and acquires little information. The first hedging strategy is to send all but one of her workers to the task she believes to be critical, but do so with hardly any information on their skills (Strategy C). This strategy is advantageous because the unskilled worker does no harm when he is by himself, while the presence of a single skilled worker is useful if the critical task is misidentified.[20] As information costs increase even further, the manager assigns an equal number of workers to either task (Strategy D), hedging output as much as possible. What little information she still gathers concerns both the worker and the task, but expected output quickly approaches the no-information benchmark.

6 Conclusion

We have introduced a novel geometric approach to finite RI models and developed a computational toolkit with a focus on robust methods for behavioral predictions. Our hope is that our contribution enables further substantial progress in more complex, quantitatively-relevant RI models — the kind necessary for applied work.

We conclude with some observations regarding RI models with a continuum of actions or states. These are commonplace in macroeconomics and finance, where researchers typically either focus on specific functional forms that can be solved analytically or approximate the solution with a tractable distribution. Our methods provide both an alternative and a complement to these distributional assumptions through discretization of the action and state spaces. Given our computational gains, it is possible to use very fine grids to minimize precision loss. Researchers can thus now assess how general the analytic functional forms are or whether distributional approximations are satisfactory. If the behavioral implications prove to be robust, then the elegance and tractability of the analytic or approximate solutions justify their use. If they do not, then discretization may be preferable — with the advantage that it grants the researcher more flexibility in tailoring the payoffs and beliefs to data.

[20] To an outside observer, firm output is most unpredictable over this range, as output entropy peaks under this strategy.

A Appendix

A.1 Additional Proofs

We start by showing that it is without loss of generality possible to relax the (RI) optimization such that the decision maker can separately choose the marginals $p \in \Delta\mathcal{A}$ and conditionals $\boldsymbol{P} \in (\Delta\mathcal{A})^I$,

$$\max_{p\in\Delta\mathcal{A},\boldsymbol{P}\in(\Delta\mathcal{A})^I} \sum_{i\in\mathcal{I}} \pi_i \left[\sum_{\boldsymbol{a}\in\mathcal{A}} P_i(\boldsymbol{a})a_i - \lambda D_{\mathsf{KL}}(P_i \,\|\, p)\right], \tag{7}$$

where $D_{\mathsf{KL}}(P_i \,\|\, p) = \sum_{\boldsymbol{a}\in A} P_i(\boldsymbol{a}) \ln\left(\frac{P_i(\boldsymbol{a})}{p(\boldsymbol{a})}\right)$ denotes the Kullback-Leibler divergence between P_i and p.

Lemma 3. *The optimal conditionals $\boldsymbol{P}$ in (RI) and the relaxed problem Equation (7) are equal, and the optimal marginals in Equation (7) satisfy $p(\boldsymbol{a}) = \boldsymbol{\pi}\cdot\boldsymbol{P}(\boldsymbol{a})$ for all $\boldsymbol{a}\in\mathcal{A}$.*

Proof. It is an exercise in pure algebra to show that

$$\begin{aligned}
&\sum_{i\in\mathcal{I}} \pi_i \left[D_{\mathsf{KL}}(P_i \,\|\, p) - D_{\mathsf{KL}}(P_i \,\|\, \boldsymbol{\pi}\cdot\boldsymbol{P})\right] \\
&\qquad = \sum_{i\in\mathcal{I}} \pi_i \sum_{\boldsymbol{a}\in\mathcal{A}} P_i(\boldsymbol{a}) \left[\ln(P_i(\boldsymbol{a})) - \ln(p(\boldsymbol{a})) - \ln(P_i(\boldsymbol{a})) + \ln(\boldsymbol{\pi}\cdot\boldsymbol{P}(\boldsymbol{a}))\right] \\
&\qquad = \sum_{\boldsymbol{a}\in A} \boldsymbol{\pi}\cdot\boldsymbol{P}(\boldsymbol{a}) \left(\ln(\boldsymbol{\pi}\cdot\boldsymbol{P}(\boldsymbol{a})) - \ln p(\boldsymbol{a})\right) = D_{\mathsf{KL}}(\boldsymbol{\pi}\cdot\boldsymbol{P} \,\|\, p) \geq 0,
\end{aligned}$$

with strict inequality whenever $p(\boldsymbol{a})$ differs from $\boldsymbol{P}$'s marginals $\boldsymbol{\pi}\cdot\boldsymbol{P}(\boldsymbol{a})$. Consequently, any other choice of p would increase costs without any consumption benefits. By optimality, the agent avoids all unnecessary costs by setting $p = \boldsymbol{\pi}\cdot\boldsymbol{P}$. □

Next, we formally establish the validity of the optimality conditions that are central to our results.

Proof of Theorem 1: Since w is strictly concave over a convex domain $\mathcal{B}$, it admits a unique maximum. We first show that the optimum $\boldsymbol{b}^*$ necessarily satisfies both conditions. Indeed, consider any $\boldsymbol{b} \in \mathcal{B}\setminus\{\boldsymbol{b}^*\}$ and let $\boldsymbol{\eta} : [0,1] \to \mathcal{B}$ be defined as $\boldsymbol{\eta}(t) = t\boldsymbol{b}^* + (1-t)\boldsymbol{b}$. The function $w\circ\boldsymbol{\eta}$ represents the gain in utility as the attention

vector moves from $\boldsymbol{b}$ towards $\boldsymbol{b}^*$. The function is strictly increasing: If $t < t'$, then

$$(w \circ \boldsymbol{\eta})(t') > \frac{t'-t}{1-t} w(\boldsymbol{\eta}(1)) + \frac{1-t'}{1-t} w(\boldsymbol{\eta}(t)) \geq (w \circ \boldsymbol{\eta})(t),$$

where the first inequality follows from strict concavity of w and the second from optimality of $\boldsymbol{b}^*$, as $w(\boldsymbol{\eta}(1)) = w(\boldsymbol{b}^*) \geq w(\boldsymbol{\eta}(t))$. Since the derivative of $(w \circ \boldsymbol{\eta})$ is equal to $\boldsymbol{\nabla} w(\boldsymbol{\eta}(t)) \cdot (\boldsymbol{b}^* - \boldsymbol{b})$, and $\boldsymbol{\nabla} w(\boldsymbol{b}) \cdot \boldsymbol{b} = \sum_{i \in \mathcal{I}} \lambda \frac{\pi_i}{b_i} b_i \equiv 1$ by construction, the nonnegativity of $(w \circ \boldsymbol{\eta})'(1)$ yields condition (a) and the nonnegativity of $(w \circ \boldsymbol{\eta})'(0)$ yields (b).

We show sufficiency through the contrapositive: If $\boldsymbol{b}^*$ is not optimal, then it satisfies neither condition. Let $\boldsymbol{b}$ equal the true optimum and define $\boldsymbol{\eta}(t)$ as above. The function $(w \circ \boldsymbol{\eta})$ is now strictly decreasing since $w(\boldsymbol{\eta}(0)) > w(\boldsymbol{\eta}(t'))$ by uniqueness of the optimum and hence for any $t < t'$,

$$(w \circ \boldsymbol{\eta})(t) \geq \frac{t'-t}{t'} w(\boldsymbol{\eta}(0)) + \frac{t}{t'} w(\boldsymbol{\eta}(t')) > (w \circ \boldsymbol{\eta})(t').$$

At $t = 1$, the condition $(w \circ \boldsymbol{\eta})'(t) < 0$ violates (a) and at $t = 0$ it violates (b). □

Proof of Lemma 2: Consider any subsequence $\{\boldsymbol{P}^{n_k}\}$ that converges to some conditionals $\bar{\boldsymbol{P}}$ under the standard Euclidean metric over $(\Delta\mathcal{A})^{\mathcal{I}}$. Since $\boldsymbol{P}^n$ converges to $\boldsymbol{P}$ under $d_{(\mathrm{G})}$ and continuity of the bijective mapping $\boldsymbol{\beta}$, we know that $\boldsymbol{\beta}(\boldsymbol{\alpha}^{\bar{\boldsymbol{P}}}) = \boldsymbol{\beta}(\boldsymbol{\alpha}^{\boldsymbol{P}})$. As long as the solution to (RI) is unique, this point can be written in a unique way as a convex combination over $\boldsymbol{\beta}(\mathcal{A})$. This implies that the marginals $p^{n_k} = \sum_{i \in \mathcal{I}} \pi_i P_i^k$ converge to $p = \sum_{i \in \mathcal{I}} \pi_i P_i$.

As such, all convergent subsequences of the bounded sequence p^n converge to the same limit p^*. The Bolzano-Weierstrass theorem thus implies that p^n itself converges to p^*. By continuity of the first-order conditions (1), the convergence translates to the conditional choice defined in Equation (5). □

A.2 Algorithm description

Our base routine for small to moderate menus works as follows: We make use of the scaling property (Lemma 1) to avoid floating point imprecision and store the normalized attention vectors in a I-by-$|\mathcal{A}|$ matrix with entries $B_{i\boldsymbol{a}} = \beta_i(\boldsymbol{a}) / \max_{\tilde{\boldsymbol{a}} \in \mathcal{A}} \beta_i(\tilde{\boldsymbol{a}})$.

Starting with an initial guess for the marginals,[21] $\boldsymbol{p}^0$, we iteratively solve a second-order Taylor approximation to (G),[22] which after dropping constant terms yields

$$\boldsymbol{q}^k := \arg\max_{\boldsymbol{p}\in\Delta^{|\mathcal{A}|-1}} \frac{1}{2}\boldsymbol{p}^\mathsf{T} B^\mathsf{T} H B\boldsymbol{p} - 2\boldsymbol{\nabla} w(B\boldsymbol{p}^k)^\mathsf{T} B,$$

where H refers to the diagonal matrix with entries $H_{ii} = \pi_i/(B\boldsymbol{p}^k)_i^2$. When the objective function w is particularly flat, it can happen that this process 'overshoots' and results in attention vectors $\boldsymbol{b}^k = B\boldsymbol{p}^k$ and $\boldsymbol{b}' = B\boldsymbol{q}^k$ that are on opposite sides of the optimum, which can lead to long cycles. We can partially avoid this by making sure that $\boldsymbol{\nabla} w(\boldsymbol{b}') \cdot \boldsymbol{b}^k \leq 1$, which ensures that the bounding hyperplane imposed by the candidate $\boldsymbol{b}'$ points away from $\boldsymbol{b}^k$. If the inequality holds, we set $\boldsymbol{p}^{k+1} := \boldsymbol{q}^k$ and move to the next iteration. Otherwise, we know that by strict convexity of w, there exists a point along the segment between $[\boldsymbol{b}^k, \boldsymbol{b}']$ that does better than both. In that case, we determine the candidate marginals $\boldsymbol{p}^{k+1} = t\boldsymbol{p}^k + (1-t)\boldsymbol{q}^k$ by identifying the root of the monotone function $\boldsymbol{\nabla} w(t\boldsymbol{p}^k + (1-t)\boldsymbol{q}^k) \cdot (\boldsymbol{q}^k - \boldsymbol{p}^k)$ where the indifference curve lies tangent to the segment.

We repeat this quadratic approximation until the implied IE converges, i.e. until $d_{(\mathrm{G})}(\boldsymbol{P}^k, \boldsymbol{P}^{k+1}) < \varepsilon$, where $\boldsymbol{P}^k$ is defined from $\boldsymbol{p}^k$ according to (1). As a default, and in all our applications, we use tolerance parameter $\varepsilon = 10^{-12}$. By construction, our approach ensures that the objective value $w(B\boldsymbol{p}^k)$ increases with each iteration.

When the action space is rich, the attention matrix B can require a lot of memory. To avoid this limitation, we first apply the base routine to a coarse subgrid of the menu, $\bar{A}^0 \subset \mathcal{A}$. Upon convergence, we denote the estimated marginals by $\boldsymbol{q}^0$, its associated attention vector as $\boldsymbol{b}^0 = B^0\boldsymbol{q}^0$, and the tentative consideration set as $A^0 = \mathsf{support}(\boldsymbol{q}^0)$. We then compute the $\boldsymbol{b}^0$-scores over a finer subgrid $\bar{A}^1 \supseteq \bar{A}^0$. We add the actions with the highest score to A^0 until the menu reaches some maximum cardinality K or contains all actions in some p-cover of grid $\bar{A}^1$. We repeatedly apply the base routine to obtain updated estimates $\boldsymbol{q}^{1,m}$, $\boldsymbol{b}^{1,m}$ and $\mathcal{A}^{1,m}$, and after each step we augment $\mathcal{A}^{1,m}$ with the actions in $\bar{A}^1$ that have the highest $\boldsymbol{b}^{1,m}$-scores. As long as K is large enough, the p-cover eventually stabilizes, and we move to the next finer

[21] Practically, we use the full-information marginals by placing weight π_i on $\arg\max_{\boldsymbol{a}\in\mathcal{A}} a_i$.

[22] We implement this code using MATLAB's built-in quadprog solver (Version 2019b). Since the solver does not accept an initial guess, we use an equivalent centered problem by solving for $\boldsymbol{dp} = \boldsymbol{p} - \boldsymbol{p}^k$ instead.

subgrid $\bar{\mathcal{A}}^2$. We continue this process until the grid encompasses all of $\mathcal{A}$ and the p-cover stabilizes.

The provided code returns the estimated choice probabilities and computes partial covers at any desired probability level p, as described in Section 4.1. To identify dominated actions, we restrict the set of feasible optimal gradients Ψ as follows: First, we take the final numerical estimate $\boldsymbol{b}^0$, perturb it slightly along each dimension, and consider the largest feasible attention vector along the perturbed ray.[23] Together, this yields a finite set of near-optimal solutions $\widehat{B} = \{\boldsymbol{b}^0, ..., \boldsymbol{b}^I\} \subset \mathcal{B}$. The optimality conditions in Theorem 1(b) imply that $\boldsymbol{\nabla} w(\boldsymbol{b}^k) \cdot \boldsymbol{b} \geq 1$ for all k, and thus restricts $\boldsymbol{b}^*$ to a small sub-region $\widehat{B} \subset \mathcal{B}$ that is obtained by imposing these additional inequality constraints. Since $\nabla_i w(\boldsymbol{b}) = \pi_i / b_i$ is strictly decreasing in b_i, we obtain the linear bounds

$$\nabla_i w(\boldsymbol{b}) \in \left[\frac{\pi_i}{\max_{\widehat{B}} b_i}, \frac{\pi_i}{\min_{\widehat{B}} b_i} \right]. \tag{8}$$

Moreover, the optimality conditions in Theorem 1(a) require that the gradient $\boldsymbol{\nabla} w(\boldsymbol{b}^*)$ satisfies the additional linear inequality constraints

$$\boldsymbol{v} \cdot \boldsymbol{\beta}(\boldsymbol{a}) \leq 1 \qquad \forall \boldsymbol{a} \in \mathcal{A}. \tag{9}$$

Together Equations (8) and (9) form a feasible polytope Ψ for the optimal gradient $\boldsymbol{\nabla} w(\boldsymbol{b}^*)$. Dominated actions can thus be identified as those with a negative maximal $\boldsymbol{b}^*$-score, i.e. those with $\max_{\boldsymbol{\psi} \in \Psi} \boldsymbol{\psi} \cdot \boldsymbol{\beta}(\boldsymbol{a}) - 1 < 0$.

References

Simone Bertoli, Jesús Fernández-Huertas Moraga, and Lucas Guichard. Rational inattention and migration decisions. *Journal of International Economics*, 126:103364, 2020. ISSN 0022-1996. doi: https://doi.org/10.1016/j.jinteco.2020.103364. URL https://www.sciencedirect.com/science/article/pii/S0022199620300805.

Zach Y. Brown and Jihye Jeon. Endogenous information and simplifying insurance choice. *Unpublished working paper*, 2020.

[23] Formally, we solve $\max_{k \in \mathbb{R}, \boldsymbol{b} \in \mathcal{B}} k$ subject to $\boldsymbol{b} \geq k(\boldsymbol{b}^0 + \varepsilon \boldsymbol{e}^i)$, where $\varepsilon > 0$ is a small perturbation scalar and $\boldsymbol{e}^i$ denotes the unit vector in dimension i.

Andrew Caplin, Mark Dean, and John Leahy. Rational inattention, optimal consideration sets, and stochastic choice. *Review of Economic Studies*, 86(3):1061–1094, 07 2018. ISSN 0034-6527. doi: 10.1093/restud/rdy037. URL https://doi.org/10.1093/restud/rdy037.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Kunal Dasgupta and Jordi Mondria. Inattentive importers. *Journal of International Economics*, 112(C):150–165, 2018. doi: 10.1016/j.jinteco.2018.03. URL https://ideas.repec.org/a/eee/inecon/v112y2018icp150-165.html.

H. G. Eggleston. *Convexity*. Cambridge Tracts in Mathematics. Cambridge University Press, 1958. doi: 10.1017/CBO9780511566172.

Xavier Gabaix. A sparsity-based model of bounded rationality. *Quarterly Journal of Economics*, 129(4):1661–1710, 2014.

Wagner Piazza Gaglianone, Raffaella Giacomini, João Victor Issler, and Vasiliki Skreta. Incentive-driven inattention. *Journal of Econometrics*, 2020.

Matthew Gentzkow and Emir Kamenica. Costly persuasion. *American Economic Review*, 104(5):457–462, 2014.

Lixin Huang and Hong Liu. Rational inattention and portfolio selection. *Journal of Finance*, 62(4):1999–2040, 2007. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/4622323.

Kenneth Judd. *Numerical Methods in Economics*, volume 1. The MIT Press, 1 edition, 1998. URL https://EconPapers.repec.org/RePEc:mtp:titles:0262100711.

Junehyuk Jung, Jeong Ho (John) Kim, Filip Matějka, and Christopher A. Sims. Discrete actions in information-constrained decision problems. *Review of Economic Studies*, 03 2019. ISSN 0034-6527. doi: 10.1093/restud/rdz011. URL https://doi.org/10.1093/restud/rdz011. rdz011.

Marcin Kacperczyk, Stijn Van Nieuwerburgh, and Laura Veldkamp. A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626, 2016.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Yulei Luo, Jun Nie, Gaowang Wang, and Eric R. Young. Rational inattention and the dynamics of consumption and wealth in general equilibrium. *Journal of Economic Theory*, 172:55 – 87, 2017. ISSN 0022-0531. doi: https://doi.org/10.1016/j.jet.2017.08.005. URL http://www.sciencedirect.com/science/article/pii/S0022053117300832.

Bartosz Maćkowiak and Mirko Wiederholt. Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803, 2009. ISSN 00028282, 19447981. URL http://www.jstor.org/stable/25592482.

Bartosz Maćkowiak, Filip Matějka, and Mirko Wiederholt. Survey: Rational inattention, a disciplined behavioral model. Working paper, Center for Economic and Policy Research, October 2018.

Filip Matějka. Rationally inattentive seller: Sales and discrete pricing. *Review of Economic Studies*, 83(3):1125–1155, 2016.

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1): 272–98, January 2015. doi: 10.1257/aer.20130047. URL http://www.aeaweb.org/articles?id=10.1257/aer.20130047.

Jianjun Miao, Jieran Wu, and Eric Young. Multivariate rational inattention. Working paper, Boston University, January 2019.

Jordi Mondria. Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory*, 145(5):1837–1864, 2010.

Michèle Müller-Itten, Roc Armenter, and Zachary Stangebye. Rational inattention via ignorance equivalence. *Unpublished working paper*, 2021. Available at https://www.philadelphiafed.org/the-economy/macroeconomics/rational-inattention-ignorance-equivalence.

Lin Peng. Learning with information capacity constraints. *Journal of Financial and Quantitative Analysis*, 40(2):307–329, 2005. ISSN 00221090, 17566916. URL http://www.jstor.org/stable/27647199.

Lin Peng and Wei Xiong. Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80(3):563 – 602, 2006. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2005.05.003. URL http://www.sciencedirect.com/science/article/pii/S0304405X05002138.

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

Christopher A. Sims. Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, 96(2):158–163, 2006.

Stijn Van Nieuwerburgh and Laura Veldkamp. Information immobility and the home bias puzzle. *Journal of Finance*, 64(3):1187–1215, 2009. doi: 10.1111/j.1540-6261.2009.01462.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2009.01462.x.

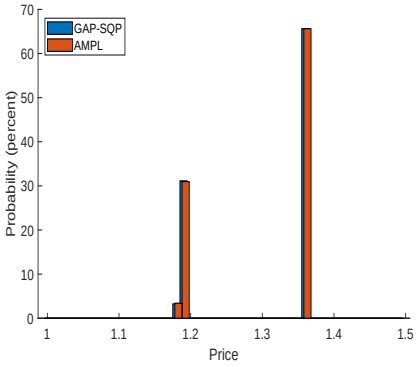
Stijn Van Nieuwerburgh and Laura Veldkamp. Information acquisition and underdiversification. *Review of Economic Studies*, 77(2):779–805, 2010.

B Online Appendix

B.1 Sticky Prices [Matějka, 2016]

Additional Figures. Both the GAP-SQP and BA algorithms replicate the results in Matějka [2016] very closely. Figure 9 shows the marginal distributions over prices for the GAP-SQP algorithm (panel (a)) and BA algorithm (panel(b)), together with the numerical solutions from AMPL provided by Filip Matějka. Solutions are so close that we had to offset the histograms for visibility. We find that increasing grid precision for actions does not meaningfully alter the solution.

(a) GAP-SQP algorithm



(b) Blahut-Arimoto algorithm

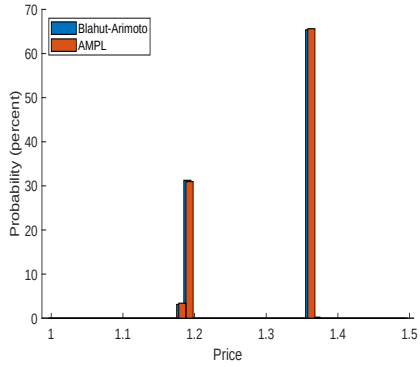


Figure 9: Replication of Matějka [2016]

Figure 10 reports the differences in the objective function value, at the computed maximum, between the GAP-SQP and BA algorithms, for the benchmark case. The difference is positively thorough for all the information values, indicating that the GAP-SQL algorithm achieves greater precision despite running on a fraction of the time of the BA algorithm. The difference, though, is very small by our choice of stopping values.

B.2 Portfolio Choice [Jung et al., 2019]

Derivation of the LQG solution. Because of the properties of the CARA utility function, it is possible to rewrite Equation (6) as

$$U(\boldsymbol{\theta}, \boldsymbol{Y}) = -\exp\left(-\alpha\left(1.03 + \sum_{j=1}^{2}(0.01 + Y_j)\theta_j\right) + \frac{\alpha^2}{2}(\theta_1^2 + \theta_2^2)\sigma_z^2\right).$$

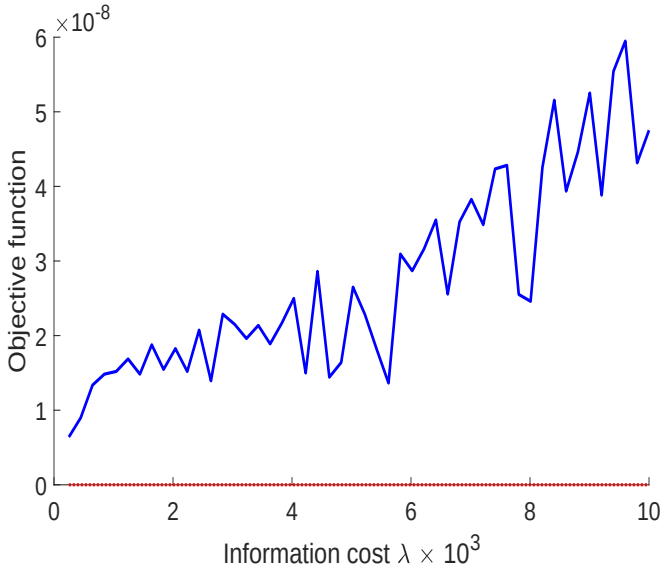


Figure 10: Objective function: GAP-SQP minus Blahut-Arimoto algorithm

We now construct a second-order approximation of this objective function around $\hat{\boldsymbol{\theta}}$ such that

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} U(\hat{\boldsymbol{\theta}}, \mathbf{0}) = \mathbf{0} \qquad \Longleftrightarrow \qquad \hat{\boldsymbol{\theta}} \approx (33.4124, 33.4124),$$

i.e., those portfolio shares that would be optimal if evaluated at the ex-post realization $\boldsymbol{Y} = \mathbf{0}$. Note this is not the same as the no-information solution because it does not take into account the risk associated with $\boldsymbol{Y}$.

Because $\mathsf{E}[\boldsymbol{Y}] = \mathbf{0}$, the second-order Taylor approximation around $(\hat{\boldsymbol{\theta}}, \mathsf{E}[\boldsymbol{Y}])$ equals

$$\tilde{U}(\boldsymbol{\theta}, \boldsymbol{Y}) = U(\hat{\boldsymbol{\theta}}, \mathbf{0}) + \begin{bmatrix} \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \\ \boldsymbol{Y} - \mathbf{0} \end{bmatrix}^{\mathsf{T}} \boldsymbol{\nabla} U(\hat{\boldsymbol{\theta}}, \mathbf{0}) + \frac{1}{2} \begin{bmatrix} \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \\ \boldsymbol{Y} - \mathbf{0} \end{bmatrix}^{\mathsf{T}} \boldsymbol{\nabla}^2 U(\hat{\boldsymbol{\theta}}, \mathbf{0}) \begin{bmatrix} \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \\ \boldsymbol{Y} - \mathbf{0} \end{bmatrix}.$$

The LQG approximation seeks to design a random variable, $\boldsymbol{\theta}$, to maximize

$$\max_{\boldsymbol{\theta}} \; E_{\boldsymbol{\theta}, \boldsymbol{Y}} \left[\tilde{U}(\boldsymbol{\theta}, \boldsymbol{Y}) \right] - \lambda \mathsf{MI}(\boldsymbol{\theta}; \boldsymbol{Y}).$$

Cover and Thomas [2012] document a well-known solution to this problem: We simply set $\boldsymbol{\theta}$ to be jointly normal with $\boldsymbol{Y}$. This follows from the fact that a Gaussian distribution maximizes entropy for a fixed variance. It is not hard to see that, given the choice of approximating point, the optimal mean is just $\hat{\boldsymbol{\theta}}$. Given this, we

need only solve for covariance matrix $\boldsymbol{\Sigma}$ that optimally balances smaller conditional dispersion against information costs.

We can simplify the objective by dropping the linear terms, since they do not depend on the covariance matrix. We can then simplify further using a couple of well-known facts: First is the functional form for the mutual information of a pair of multivariate normals, which has a simple closed-form expression; second is for any random variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$,

$$\mathsf{E}[\boldsymbol{X}_1^{\mathsf{T}} A \boldsymbol{X}_2] = \mathsf{tr}(A\mathsf{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2)) + \mathsf{E}[\boldsymbol{X}_1]^{\mathsf{T}} A \mathsf{E}[\boldsymbol{X}_2].$$

Plugging these in implies that solving the RI problem is tantamount to selecting a positive-definite $\boldsymbol{\Sigma}$ that is consistent with the marginal distribution over $\boldsymbol{Y}$ so as to maximize

$$\frac{1}{2}\mathsf{tr}\left(\left[\boldsymbol{\nabla}^2 U(\hat{\boldsymbol{\theta}}, \mathbf{0})\right]\boldsymbol{\Sigma}\right) - \lambda\frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| \times |\boldsymbol{\Sigma}_{\boldsymbol{Y}}|}{|\boldsymbol{\Sigma}|}\right),$$

where $|\cdot|$ denotes the matrix determinant and $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ denotes the marginal covariance of the $\boldsymbol{X}$.

Plugging in the optimal covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta Y}} \\ \boldsymbol{\Sigma}_{\boldsymbol{Y\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{Y}} \end{bmatrix} = \begin{bmatrix} 3158.4 & 0 & 0.9453 & 0 \\ 0 & 3158.4 & 0 & 0.9453 \\ 0.9453 & 0 & 0.0004 & 0 \\ 0 & 0.9453 & 0 & 0.0004 \end{bmatrix}$$

yields the distribution found in Figure 6(a). The objective function net of information costs is derived using Monte-Carlo methods. We take 10 million draws from the optimal distribution and compute the sample average utility. We repeat this 100 times and take sample statistics of the estimates. This yields an average payoff, net of information costs, of $-.3220$ with a 95% confidence band of $[-.3221, -.3219]$.

Additional Figures. For comparison purposes, Figure 11 plots the statewise payoff distribution $U(\boldsymbol{\theta}, \boldsymbol{Y}) - \lambda\mathsf{MI}$, assuming $(\boldsymbol{\theta}, \boldsymbol{Y})$ is distributed according to the numeric solution of GAP-SQP (blue) or JKMS (orange), and the information cost MI is borne unconditionally.

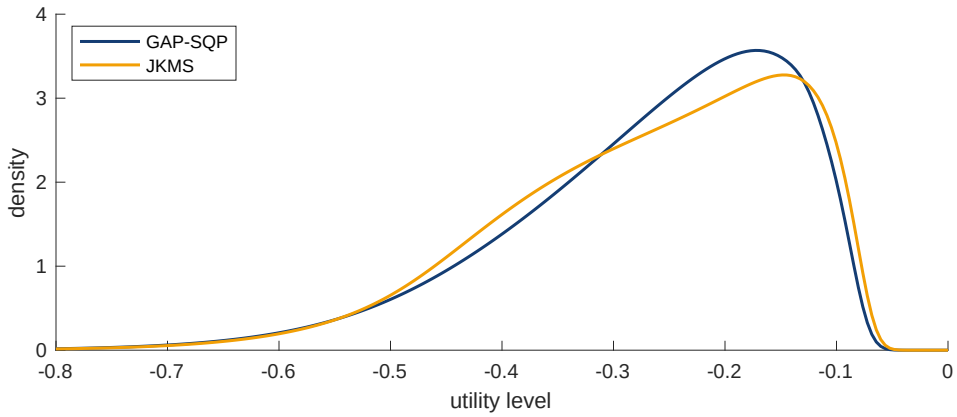


Figure 11: Payoff distribution across algorithm estimates, smoothed with a kernel density estimate.