Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

The Two-Period Rational Inattention Model: Accelerations and Analyses

Kurt F. Lewis

2008-22

The Two-Period Rational Inattention Model: Accelerations and Analyses [1]

Kurt F. Lewis

Board of Governors of the Federal Reserve System, Washington, DC 20551, USA

Abstract

This paper demonstrates the properties of and a solution method for the more general two-period Rational Inattention model of Sims (2006). It is shown that the corresponding optimization problem is convex and can be solved very quickly. This paper also demonstrates a computational tool well-suited to solving Rational Inattention models and further illustrates a critique raised in Sims (2006) regarding Rational Inattention models whose solutions assume parametric formulations rather than solve for their optimally-derived, non-parametric counterparts.

Key words: Rational Inattention, Information-Processing Constraints, Numerical Optimization

1 Introduction

The Rational Inattention (RI) paradigm introduced in Sims (2003) began the examination of information-processing-constrained economic agents with a model in which agents have quadratic utility and face linear constraints. Sims (2006) extended his earlier work to more general environments by demonstrating how one could think about the information-processing-constrained agent's decision outside the linear-quadratic framework and by providing a numerical solution to the two-period consumption-choice problem of a rationally inattentive agent facing an information-processing-capacity constraint. Such an agent, unlike the capacity unconstrained agent who knows a specific value for the state variable and chooses a corresponding value for the choice variable, knows only a distribution for the state

Email address: kurt.f.lewis@frb.gov (Kurt F. Lewis).

[1] I would like to thank Kurt Anstreicher, John Geweke, Ryan Haley, Beth Ingram, Gene Savin, Chris Sims, Todd Walker, and the organizers of the 2006 Institute on Computational Economics at Argonne National Laboratory. Special thanks to Charles Whiteman for guidance and many helpful conversations throughout this project. All remaining errors are my own. The views expressed here are those of the author and do not necessarily reflect the views of the Federal Reserve Board, its members, or its staff.

variable and chooses the joint distribution of state and choice variables. This paper seeks to better enable the development of more sophisticated RI problems by giving some guidance on the properties of the models Sims has introduced. It is the author's hope that this paper encourages others to implement RI models and move toward fulfilling the titular goal of Sims (2006), "Rational Inattention: A Research Agenda." It is important to note that Sims' central conclusions are robust to this new formulation of the RI problem: the consumption choices of information-processing-capacity-constrained individuals have a discrete nature even when the wealth distribution is continuous, and more risk averse individuals choose distributions for consumption (given the wealth distribution) that are more disperse at high wealth and more precise at low wealth.

The remainder of the paper is organized as follows: Section 2 examines the two-period problem and a generalization of the standard solution is presented that leads to implementation of the RI version of the two period problem. Section 3 demonstrates that the optimization problem of the rationally inattentive agent in the two-period model is convex and contains a discussion of formulation of the problem relative to that of Sims (2006). Section 4 discusses the use of the `AMPL/KNITRO` software suite and why it is particularly well-suited to these types of problems; this section also qualitatively replicates the results found in Sims (2006). Section 5 illustrates Sims' critique of RI models that *assume* the form of the optimal distribution of states and decisions, and shows that the most common parametric approximation not only misrepresents the agent's optimal behavior, but does so by yielding less "stickiness" (one supposed goal of implementing the RI framework) than the true optimal decision does. Section 6 concludes.

2 The Two-Period Model

The two-period model of Sims (2006) highlights the central difference between rational inattention and other information frictions. The choice variable in the model is the *form of the joint distribution* of consumption and wealth, and the informational "shortage" is one of processing capacity, rather than information availability.

Absent information-processing constraints, Sims' model is a two-period choice of consumption, with an undiscounted, two-period utility function that divides a pool of resources into those consumed now with some probability, and expected consumption in the subsequent time period. This is an undiscounted "cake-eating" problem in which the agent takes a given amount of wealth, w, and divides it optimally between consuming c in period one and $w - c$ in period two. That is, for CRRA preferences, the agent solves

$$\max_{c \leq w} \frac{c^{1-\gamma} + (w-c)^{1-\gamma}}{1-\gamma}.$$

The solution to this problem is an optimal decision rule, denoted f, that describes the optimal plan for the choice variable, c, given a value for the state variable, w. That is, the solution

is a one-to-one mapping from the state-space to the choice-space, described by $c^* = f(w)$. The solution to the agent's maximization problem here is given by:

$$c^* = f(w) = \frac{w}{2},$$

that is, the agent should consume half his wealth in each of the two periods. For a given value of w, this describes a corresponding value for c. Even when wealth is characterized by a probability distribution, the optimal f describes a mapping from each potential value of w to a single corresponding value for c.

2.1 A Generalization

To set the stage for the information-constrained problem, consider a generalization of the cake-eating problem in which the cake (wealth) and bites of the cake (consumption) only come in a finite set of discrete values $c_1, c_2, \ldots, c_{N_c}$ and $w_1, w_2, \ldots, w_{N_w}$. Suppose further that wealth is characterized by a discrete probability distribution $g(w)$. The decision rule, $c^* = f(w) = w/2$, becomes the method for generating a set of conditional distributions – one for each wealth value. Each of these conditional distributions for consumption is degenerate, that is, the joint distribution $f(c, w)$ describes the same thing as the $c^* = f(w) = w/2$: a one-to-one mapping from state space to choice space. The discretized version of the two period model is written:

$$\max_{\{f(c_i,w_j)\}} \sum_{i=1}^{N_c} \sum_{j=1}^{N_w} \frac{c_i^{1-\gamma} + (w_j - c_i)^{1-\gamma}}{1-\gamma} f(c_i, w_j) \tag{1}$$

subject to:

$$f(c_i, w_j) \geq 0 \tag{2}$$

$$\sum_{i=1}^{N_c} f(c_i, w_j) = g(w_j) \quad \text{for } j = 1, \ldots, N_w \tag{3}$$

$$f(c_i, w_j) = 0\,, \quad \forall (i,j) \text{ such that } c_i > w_j. \tag{4}$$

$f(c, w)$ is the joint distribution of consumption and wealth, and is the choice variable of this optimization problem, while $g(w)$ is the marginal distribution of wealth in the problem (taken as given).

The properties of the problem and the optimum are qualitatively unchanged under this generalization; that is, the agent's behavoir is not different in expectation from what it would be under the original problem. Suppose that the marginal distribution of wealth is discrete triangular, meaning higher levels of wealth have higher probability.[2] The optimal decision rule is the joint distribution $f(c, w)$ that describes the same one-to-one mapping that divides wealth in two. Under the generalization, however, this is accomplished by assigning probability to specific (c_i, w_j) pairs. That is, given a distribution for wealth, the agent disperses the probability weight $g(w_j)$ across the possible values $\{c_i\}_{i=1}^{N_c}$ [equation (3)] such that weight is only allowed where $c_i \leq w_j$ [equation (4)]. The optimal choice, shown in figure 1, is to place all of the probability of being at wealth node w_j on the pair $(c_i = w_j/2, w_j)$, that is, $f(c_i = w_j/2, w_j) = g(w_j)$.

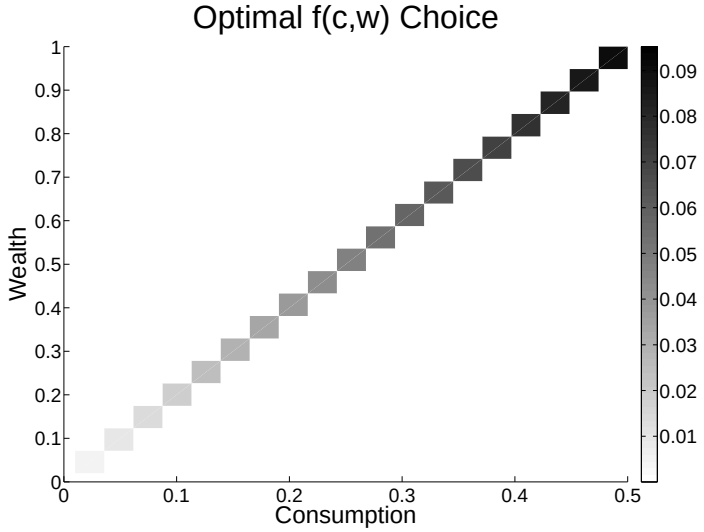

Fig. 1. A One-to-One Mapping via the Joint Distribution $f(c, w)$.

Figure 1 represents the joint distribution of c and w over the $[0, 1]$ interval when the (c, w) space is discretized. The darkness of the boxes indicates the weight of probability on that specific (c, w) pair. The darker the box, the higher the probability of the agent realizing that consumption-wealth pair. The boxes get darker as they progress "northeast" because the marginal distribution of wealth, $g(w)$, is triangular. The solution, $f(c_i = w_j/2, w_j) = g(w_j)$, demonstrates that within this generalization the one-to-one mapping takes the form of creating a set of conditional distributions of c given w that are degenerate at $c_i = w_j/2$.

2.2 The Processing-Constrained Problem

The rational inattention framework uses the metric of mutual information (MI) to quantify the amount of information-processing capacity the agent is using to solve his optimization

[2] This distributional choice was made only to follow the setup of Sims (2006).

problem.[3] By placing a constraint on mutual information, the framework limits the strength of the relationship between c and w thus limiting the precision with which either variable can be understood by the agent. As the amount of information the agent can process is reduced from the amount required to produce the one-to-one relationship described in figure 1, the agent must decide how best to allocate the finite resource of processing capacity across the space of his choice variable.

The agent's optimization problem in the information-processing constrained universe is the same as the one detailed in equations (1) through (4), with the addition of the following constraint on the amount of mutual information in the model:[4]

$$\sum_{j=1}^{N_w}\sum_{i=1}^{N_c}\log[f(c_i,w_j)]\cdot f(c_i,w_j) - \sum_{i=1}^{N_c}\left(\log\left(\sum_{j=1}^{N_w} f(c_i,w_j)\right)\cdot\sum_{j=1}^{N_w} f(c_i,w_j)\right) - \sum_{j=1}^{N_w}\log(g(w_j))\cdot g(w_j) \leq \kappa. \tag{5}$$

As the amount of information-processing capacity (κ) decreases, Sims (2003) notes that the effect on the agent is similar to that of increasing the noise in a signal-extraction version of the same problem. In the past, economic models have tried to explain the difference between theory and empirical observation in many models by assuming the existence of an exogenous noise that complicates the understanding of the model's state. The rational inattention framework does something similar to this by describing an environment in which the "noise" is endogenously determined rather than exogenously given: it arises from the agent's inability to accurately assess the state because he does not have the information-processing resources to do so.[5]

[3] Mutual Information (MI) is an information-theoretic metric showing the strength of the relationship of two jointly distributed random variables. It measures the amount of information contained in the joint distribution that is unavailable in the marginal distributions. For example, if x and y are jointly distributed, $MI(x,y)$ would tell the observer how much information about x could be gained from an observation of y and vice-versa. For more information on mutual information and its role in the RI framework, see Sims (2006).

[4] Mutual Information, as defined in equation (5), is given (within the information theory literature) in *bits* (*bi*nary digi*ts*), as a result of using base 2 logarithms. The units of κ in this model are known as "nats," because of the use of natural logarithms.

[5] See Sims (2003) for a full discussion of the signal-extraction/RI link and an example using a linear-quadratic permanent income model.

3 A Convex Problem

RI models represent a potentially large burden on numerical optimization algorithms. Rather than choosing specific values for choice variables given state variables, the optimizer is asked to choose large joint distributions of state and decision variables. It is beneficial to know that this model, though large, is still numerically tractable. It will be shown later that Sims' log-transformation imposes an additional burden in terms of numerical optimization. As a first step, it will be shown that his original, un-transformed problem is, in fact, convex and well suited to numerical optimization.

The agent chooses $f(c_i, w_j)$ (hereafter: $f_{i,j}$). The nodes for consumption and wealth are fixed by the model-designer, rather than the agent, and it is the probabilities $f_{i,j}$ that are chosen by the agent. Thus, the objective function [equation (1)] is a weighted sum and linear. Constraints (2), (3) and (4) are also linear; therefore, in order for the problem to be convex, it must be shown that:

Theorem 3.1 *For $f_{i,j} > 0$,*

$$\begin{aligned} MI(f_{i,j}) = &\sum_{i=1}^{N_c}\sum_{j=1}^{N_w} f_{i,j}\cdot\log(f_{i,j}) \\ &-\sum_{i=1}^{N_c}\left\{\left[\sum_{j=1} f_{i,j}\right]\cdot\left[\log\left(\sum_{j=1} f_{i,j}\right)\right]\right\} \\ &-\sum_{j=1}^{N_w} g(w_j)\cdot\log(g(w_j)) \leq \kappa. \end{aligned}$$

is convex in $f_{i,j}$.

Proof. See Appendix A. □

The problem specified in equations (1) through (5) has the requirement that $f_{i,j} \geq 0$, rather than $f_{i,j} > 0$. It should also be noted that some $f_{i,j}$'s will be zero by the feasibility constraint (4), thus it is important that $f_{i,j} = 0$ be considered. What has been demonstrated in the proof of theorem 3.1 is that the MI constraint is convex on the interior of the feasible set. However, because $\lim_{p\to 0} p\log(p) = 0$, the MI constraint is continuous on the set $[0, 1]$. Therefore, since the function is continuous on the closed set $[0, 1]$ and convex on the interior, the function is convex on the closed set. Therefore the problem specified in equations (1) through (5) is a convex programming problem.

3.1 Differences from Sims (2006)

Three differences exist between what has been done here regarding the numerical optimization and what was done in Sims (2006): First, Sims uses a normalization to eliminate (3), where here it is left explicit. Second, rather than pick a value for λ (the LaGrange multiplier on the capacity constraint) and maximize the LaGrangian for a given multiplier value, a value for the capacity κ is chosen, and the constraint remains intact.[6] Third, I optimize directly over the values of f rather than their logarithm.

The first two differences are minor in comparison to the third. The third difference is counter-intuitive, but represents a large element of the difference between the optimization results presented here and the ones in Sims (2006). Sims' reasons for optimizing over $\log(f_{i,j})$ is that because logarithms are undefined at zero, we can use $\log(f)$ to make sure that the problem stays in the region of $f > 0$ values which are well behaved (in terms of the gradient).[7] When $\log(f)$ is very large and negative, it is taken to be zero. While the optimization problem is the same theoretically, it has become much more difficult for numerical optimizers to solve. This transformation is responsible for a large part of the difference in computational times.

4 The Solution Procedure

RI problems are inherently large, in terms of the number of variables, relative to their unconstrained counterparts. `AMPL` (literally: `A` `M`athematical `P`rogramming `L`anguage) was chosen because it can accommodate problems of a very large size and includes a differentiation feature that aids in accurately finding the optimum in such a large variable space. `AMPL` is not an optimizer in itself, but rather a front end, that is, a piece of software designed to allow the user to interact with other software, for a large number of potential optimization algorithms, each of which has properties suited to specific problems.[8,9]

[6] The optimization software being used deals directly with the constraints and takes derivatives through the LaGrangian automatically, allowing it to use the fastest optimization methods.

[7] In general, there are additional reasons for this practice. Usually, this monotone transformation smooths the objective function, making the optimum easier to find without changing its location. Here, the linear objective function and convex constraint set are only complicated as a result.

[8] This problem has several hundred variables, and it is a fairly small RI problem (Lewis (2007) has close to 6000). Moving from an environment in which univariate state variables are mapped to univariate choice variables to one in which the optimization chooses a non-parametric joint distribution of state and choice variables dramatically increases the computational burden. The author was introduced to `AMPL` at the ICE 2006 and it has proven to be a good replacement when other optimizers, such as MATLAB, are unable to handle the problem.

[9] Appendix B demonstrates a quasi-analytical solution procedure for this problem which works from the F.O.C.'s of the LaGrangian of the problem and requires that $f_{i,j} > 0$. The properties of the problem allow that solution to be qualitatively identical to the one found here and illustrate the role of the information-processing constraint and the agent's preferences in determining the

The key to effective numerical optimization lies in the derivatives, meaning that gradients and Hessians provide the data required to complete the task of the optimizer. Here, these are generated by means of *automatic* (or *algorithmic*) *differentiation*. The speed and accuracy of the optimizer depend on the information available about the hill being climbed. Automatic differentiation (AD) generates the gradients without truncation errors (unlike divided differencing) or the excessive memory usage of symbolic differentiation. AD is best thought of as a close cousin of symbolic differentiation in that both are the result of systematic application of the chain rule. However, in the case of AD, the chain rule is applied not to symbolic expressions but to actual numerical values. [10]

4.1 Interior-Point Optimization

The optimizer used for this model is called `KNITRO`. `KNITRO` implements an interior point optimization algorithm that is exceptionally well suited to the current problem and to RI models in general. Interior point methods approach the boundaries of variable-space in an organized way, without taking derivatives or evaluating the function *at* the boundaries. This aspect of the algorithm is important because the derivatives of this problem are infinite at some of the boundaries (that is, derivatives yield $1/x$ where $x \in [0, 1]$), but the optimization problem is continuous on the closed set, meaning that a solution in which a variable would be optimally set to zero can be represented by the optimization algorithm stopping when a value is within a certain tolerance of zero. [11]

The optimization scheme will guarantee that, to the tolerances set by the user, a local optimum is found. The convexity of the problem demonstrated above guarantees that the optimum will be global. The time that this takes is dramatically shorter than the time (11 minutes) listed in Sims (2006): The computational time for the problem is slightly less than one second for the grid size suggested in Sims (2006) on a 3 GHz Pentium 4 machine with 4 GB of RAM. While 11 minutes is not a long time to wait for a solution, the grid in this model is fairly small. Controlling for all the constraints, this problem has less than 400 $f_{i,j}$ nodes to optimize. The computational time increases nonlinearly in number of nodes and

optimal choice $f_{i,j}$.

[10] For a discussion on this and further exposition of AD, see Griewank (2000) and Rall (1981). For a discussion specific to its application within `AMPL`, see Gay (1991).

[11] Both the objective function, $\sum_{i,j} U(c_i, w_j) f_{i,j}$, and the Mutual Information constraint (5) are continuous on $f_{i,j} \in [0, 1]$, so while optimization algorithms that jump to the boundary and could get "stuck" where the derivative is undefined, the interior point method will drive the value of variables that are optimally zero toward zero without getting there, but the behavior of the agent would be accurately described as a result of the optimization. Furthermore, following the interior-point optimization, values below a certain tolerance were set to zero and the problem was re-entered into an optimizer that did not use interior-point methods, and the zeros were left alone with the remaining infintesimal weight reallocated to other $f_{i,j}$'s with no qualitative change to the optimum (e.g. $f_{i,j} \leq 10^{-8}$ are set to zero, problem is reoptimized starting at this point using `CONOPT`, no qualitative change is observed, no zeros are changed back to positive probabilities).

this time savings opens the door to more sophisticated models that have more variables, finer grids, or more time periods, such as Lewis (2007).

4.2 Results

The results are qualitatively the same as Sims (2006). In figure 2, the progressive tightening of the capacity constraint and its effect on the choice of the consumption-wealth joint distribution is seen. First, note that because the information processing constraint is left explicit, values for κ that do not bind the information-processing constraint can be chosen. In fact, this enables the discovery of the value of κ for which the information-processing constraint does not bind. Here, the darkness of the box within the joint distribution indicates the level of probability of being at that particular consumption-wealth pair. The values for κ that are small or large depend on the size of the grid and the "complexity" of the wealth distribution. The value $\kappa = 4$, in the case of figure 2, is the level of information processing capacity required to make a one-to-one decision, making it a value that produces the same decision as the unrestricted case shown in figure 1. This means that the constraint is ineffective for values of $\kappa > 4$. Four nats of information processing may seem small, but the reality is that the level of κ required to make one-to-one decisions can be made to be arbitrarily high by the model designer. As the number of nodes increases, the amount of possible combinations of consumption and wealth increase and the value of κ required to get the result in the upper-left-hand corner of figure 2 increases rapidly. An area of potential benefit for this literature would be to adopt a new constraint convention that states everything in terms of the percentage of "one-to-one-decision-making capacity". That is, in figure 2, $\kappa = 4$, 2, 1 and 0.5 would be replaced with $\bar{\kappa} = 1$, 0.5, 0.25, and 0.125. This convention could avoid future conversations about the reasonableness the size of κ when comparing across models.

Next, a comparison regarding risk aversion is made. The differences between figure 3 below and its counterparts in Sims (2006) are curious, but in the final analysis, minimal. The point of Sims' figures was to demonstrate how risk aversion impacts the choice of the joint density of consumption and wealth. This impact – that as risk aversion increases the agent prefers to give up some precision in decision making over a larger range of consumption and wealth in favor of more accuracy where it matters most, at low wealth levels, and less where it matter less, at higher wealth levels – is still seen here, where the information processing capacity is fixed at 0.85 bits, and the risk aversion parameter, γ, is changed.

The small differences in results between Sims (2006) and here are potentially attributable to several factors. First, any replication that involves optimization depends on tolerances, and that is certainly possible here. Second, different optimization algorithms are used. However, the biggest difference is in the log-normalization, so optimization was performed over the log-transformed model using an optimization algorithm called `CONOPT`, which is more similar to the optimizer implemented by Sims. The results are largely the same as previous figures, but simply take more time. For example, the `CONOPT`/$\log(f)$ model produced qualitatively identical results to those of the interior-point `KNITRO` algorithm working on the

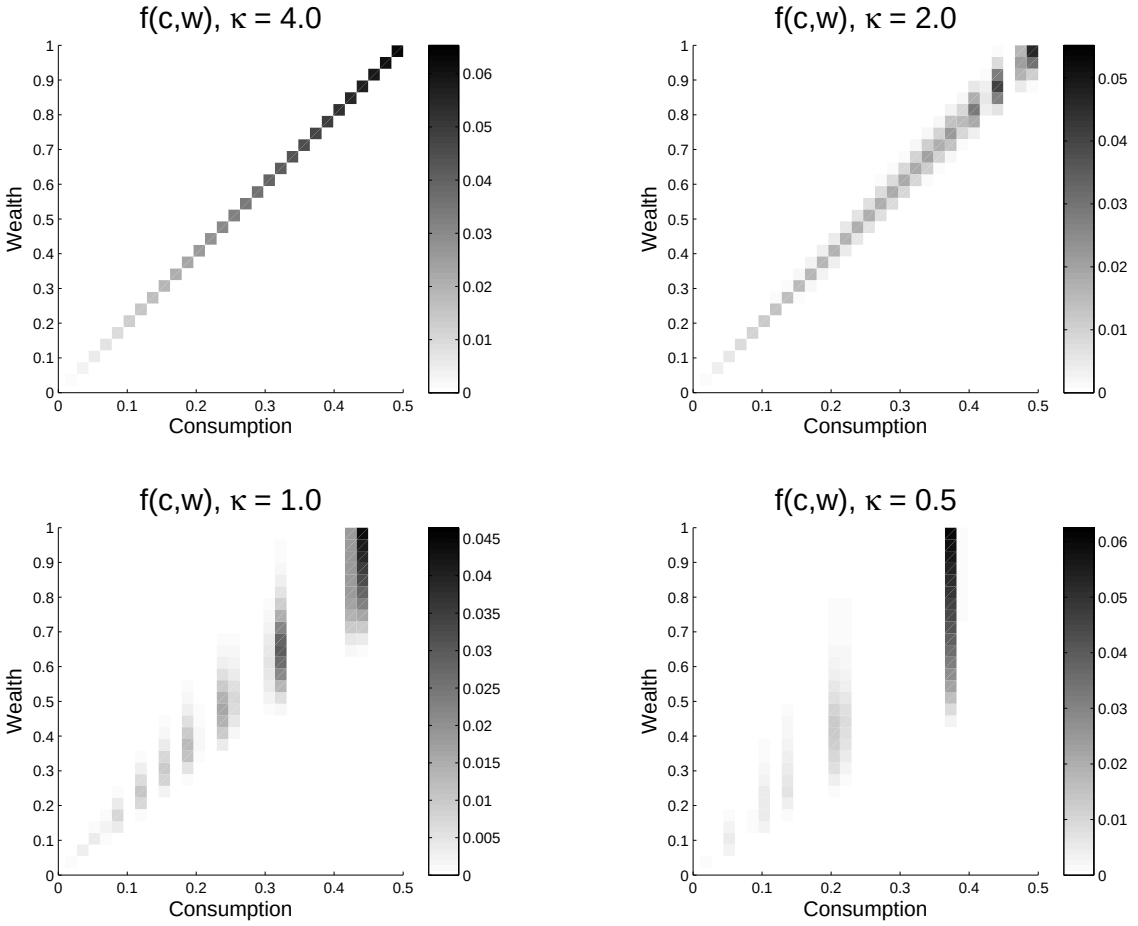

Fig. 2. Comparison of Different Levels of Information-Processing Capacity

f's themselves, but instead of taking one to two seconds, the `CONOPT`/$\log(f)$ version takes almost 7 minutes. This serves to illustrate the reality that the problem is theoretically the same, but much harder numerically. In addition to the excellent numerical qualities of the un-transformed model, this 200-fold increase in speed allows us to examine a richer class of models using the current computing technology, as in Lewis (2007). [12]

[12] Two additional results-based appendices are available from the author upon request. The first is a short document (in the spirit of McCullough and Vinod (2003)) containing the results for f under various optimization schemes, where all the parameters of the model are held constant. The second (also quite brief) appendix is a demonstration that the results shown here in endowment economies hold up in production economies as well.

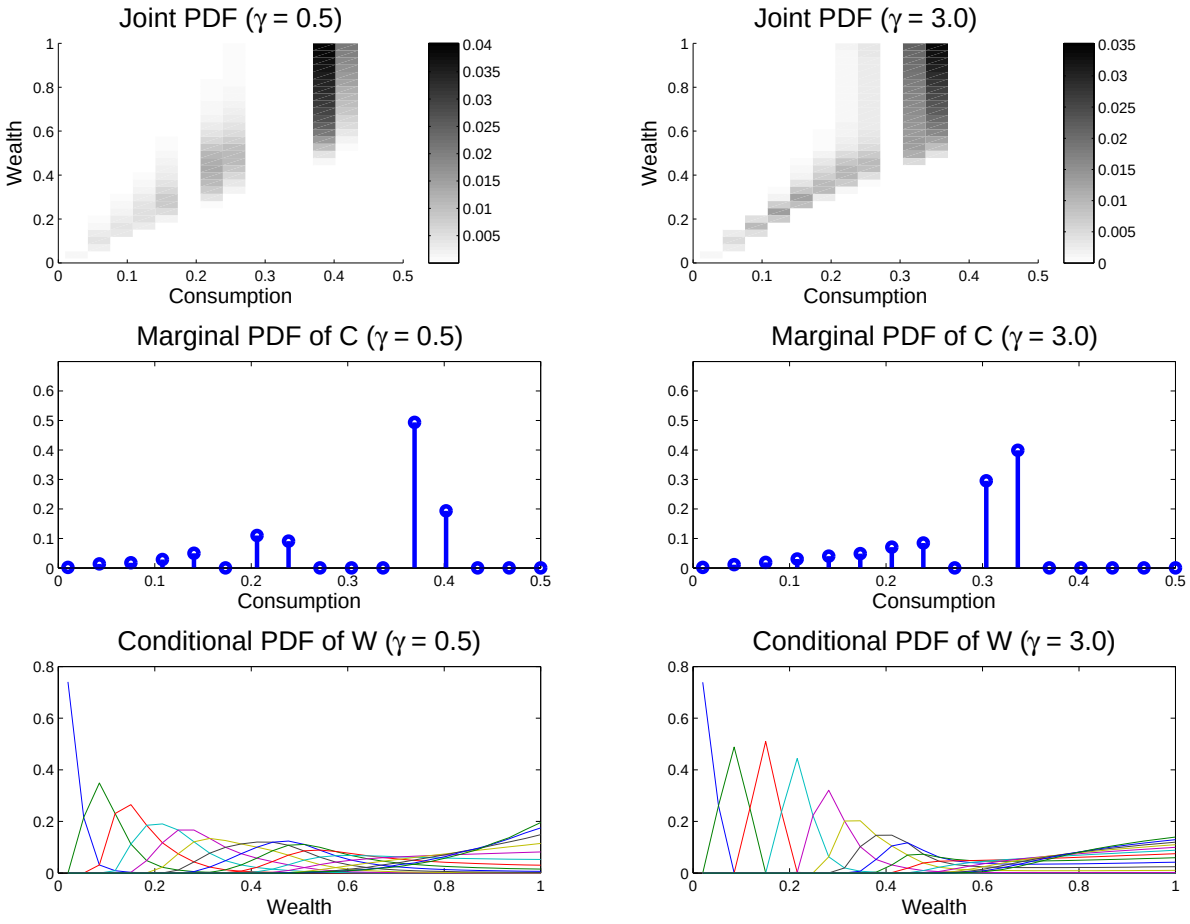

Fig. 3. Comparison of Two Levels of Risk Aversion with Information Processing Capacity of $\kappa = 0.85$ bits.

5 The Importance of Non-Parametric Choices for $f(c_i, w_j)$

Following Sims (2003), a number of information-processing-constrained-agent models were introduced into the economics and finance literature. The issue, as per Sims (2006), is that most of these models included an approximate solution method. By incorporating the "Gaussian-in, Gaussian-out" framework of the linear-quadratic setup of Sims (2003), model designers used a parametric version of the joint distribution of state and choice variables (in this model, $f(c, w)$). That is, in order to simplify the structure of the model, the designers approximated the utility-maximizing form of the joint distribution of states and choices with a joint Gaussian process, and optimized over the parameters (means and covariances) of the corresponding distribution.

A famous G.E.P. Box quote notes that: "All models are wrong but some are useful." While all models are approximations, this model can be used to examine the implications of assum-

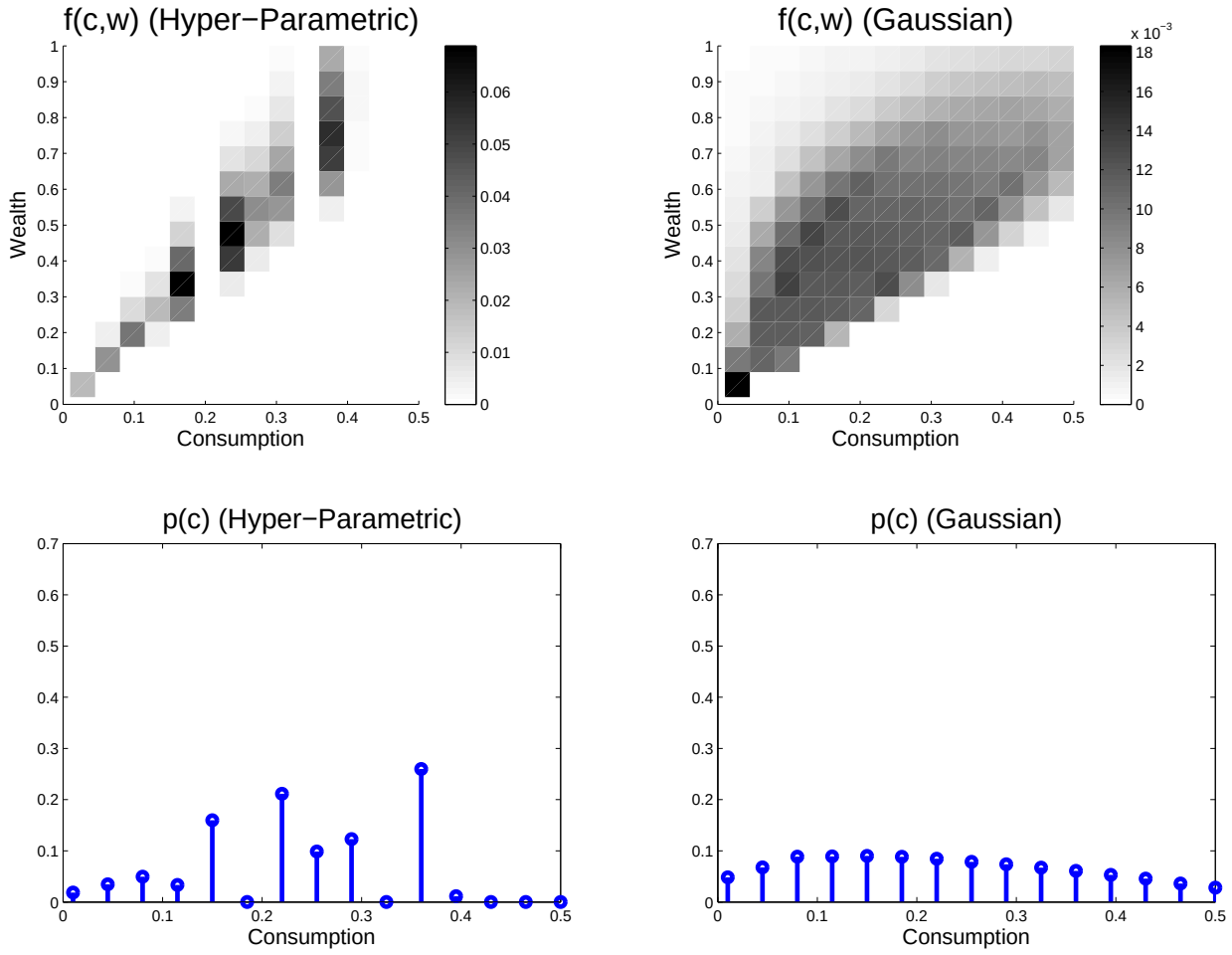

Fig. 4. The Assumption of Gaussianity

ing Gaussianity as an approximation to the truly optimal (non-parametric) choice.[13] This will indicate whether the assumption of Gaussianity, which has been demonstrated to dramatically ease analysis of more dynamic models, damages the model's ability to accurately predict agent behavior relative to the "true model." What is discovered is that the model produces drastically different implications for agent behavior when the nature of the ex-post uncertainty is assumed, rather than derived.

This effect is demonstrated in the two-period model by requiring (ceteris paribus) that $f(c, w)$ be a bivariate Gaussian distribution. To this end, a Gaussian wealth distribution ($\mu_w = 0.5$,

[13] In many literatures models are simplified before being analyzed because the true model is too difficult or impossible to solve. Here, while the "true model" has many more free parameters, we have already demonstrated that it is solvable and well-behaved, though large. Therefore, the true model can be examined relative to the approximation and analyzed to see if the behavior of the agent is similar enough to suggest that use of the Gaussian approximation is unimportant to the analysis.

$\sigma_w = 0.25$) is used, and the optimizer is required to respect the processing constraints ($\kappa = 1$) while choosing μ_c, σ_c and ρ to form $f(c, w)$. The results of the choice can be seen in figure 4. Clearly, the consumption behavior depicted by this restricted model is different from its unrestricted partner. The utility of the agent is approximately 4% lower when restricted to making Gaussian choices, and it is seen that this is clearly the result of being forced to use a smoother dispersion of probability across the consumption-wealth grid.

Aside from the lower utility achieved by restricting the form of f, it is clear that the consumption behavior itself has been changed quite dramatically. By insisting on the smooth form of the Gaussian distribution, note that the agent is forced to choose a single highest probability point and smoothly decrease the probability away from that mode. This means that the agent's risk preferences cannot be taken into account as well as they can in the non-parametric version. The agent is able to "discretize" his or her consumption in the unrestricted model. That is, they are able to choose more than one mode for the resulting marginal distribution of cosumption and they can surround consumption values they would like the give high probability with consumption levels they can give very little probability. The Gaussian approximation rule is actually likely to produce less stickiness than the more appropriate non-parametric specification. By forcing the consumption marginal to take such a smooth shape, the model designer is forcing the agent to place probability where they would prefer not to, and choose a much smoother reaction function of consumption choices to states than is suggested by the non-parametric model that fully incorporates the agent's risk attitude into this reaction function. To see this, examine the scales of the colorbars on the sides of the plots in figure 4. Note that the agent would like to have essentially no probability weight on $c = 0.18$, while the nodes directly to the right and left of that value are among the most heavily weighted. By enforcing Gaussianity, the agent must choose to be essentially indifferent across those three nodes. The tractable nature of the Gaussian-In-Gaussian-Out assumption is a siren-call that leads not just to a different result, but simply incorrect predictions about the consumer's optimal behavior. Thus, it is a poor approximation to the full RI framework as it fails to accurately incorporate the agent's preferences.

6 Conclusion

The rational inattention framework is unique in that it is presently the only paradigm with the capability to quantify, constrain, and optimally allocate the scarce resource of information-processing capacity. While other information frictions restrict *when* attention is paid and information is acquired, the RI framework allows the agent to optimally allocate his or her pool of attention thus giving control over not only when, but *how much* attention is paid.

The Sims (2006) model, while simplified, demonstrates the power of the framework by showing how the agent's preferences, combined with their information-processing constraint, result in the optimal allocation of the "attention resource." This paper demonstrates that this problem is convex and can be solved very quickly using certain tools, thus demonstrating

that the computational intensity of this smaller problem is far less than originally perceived and opening the door to more complex and dynamic problems of interest in both macro- and microeconomic literatures. This paper also illustrates the importance of optimally deriving the non-parametric decision rule $f_{i,j}$, rather than assuming a Gaussian form. Gaussian assumptions, which lend considerable tractability in certain dynamic frameworks, are shown to produce behavior on the part of the agent that is dramatically at odds with the optimally derived behavior—in fact lessening inertial effects of information-processing constraints by reducing the "discreteness" of the behavior of the RI agent. It is hoped that by illustrating the importance of the fully-optimal approach to attention-allocation problems, and by demonstrating that solutions can be quickly and accurately found, that further research in this paradigm will be encouraged.

References

Gay, D. M. (1991): "Automatic Differentiation of Nonlinear AMPL Models," *AT&T Bell Labratories Numerical Analysis Manuscript*, 91-05.

Griewank, A. (2000): *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, no. 19 in Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics.

Lewis, K. F. (2007): "The Life-Cycle Effects of Information-Processing Constraints," Working Paper, Currently available at http://myweb.uiowa.edu/kflewis/research.htm.

McCullough, B., and H. D. Vinod (2003): "Verifying the Solution from a Nonlinear Solver: A Case Study," *American Economic Review*.

Rall, L. B. (1981): *Automatic Differentiation: Techniques and Applications*, vol. 120 of *Lecture Notes in Computer Science*. Springer-Verlag.

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

——— (2006): "Rational Inattention: A Research Agenda," Working paper, Princeton University.

A Proof of the Convexity of the Mutual Information Constraint

Theorem A.1 *For $f_{i,j} > 0$,*

$$\begin{aligned} MI(f_{i,j}) = &\sum_{i=1}^{N_c}\sum_{j=1}^{N_w} f_{i,j}\cdot\log(f_{i,j}) \\ &-\sum_{i=1}^{N_c}\left\{\left[\sum_{j=1} f_{i,j}\right]\cdot\left[\log\left(\sum_{j=1} f_{i,j}\right)\right]\right\} \\ &-\sum_{j=1}^{N_w} g(w_j)\cdot\log(g(w_j)) \leq \kappa. \end{aligned} \tag{A.1}$$

is convex in $f_{i,j}$.

Proof. Because $g(w)$ is fixed, attention can be limited to

$$MI_1(f_{i,j}) = \sum_{i=1}^{N_c}\sum_{j=1}^{N_w} f_{i,j}\cdot\log(f_{i,j}) - \sum_{i=1}^{N_c}\left\{\left[\sum_{j=1}^{N_w} f_{i,j}\right]\cdot\left[\log\left(\sum_{j=1}^{N_w} f_{i,j}\right)\right]\right\}. \tag{A.2}$$

To begin, simplify the remaining problem by separating the i and j summations.

$$MI_1(f_{i,j}) = \sum_{i=1}^{N_c}\left\{\sum_{j=1}^{N_w} f_{i,j}\log\left(f_{i,j}\right) - \left[\sum_{j=1}^{N_w} f_{i,j}\right]\cdot\left[\log\left(\sum_{j=1}^{N_w} f_{i,j}\right)\right]\right\} \tag{A.3}$$

The outermost summation in equation (A.3) is over the i index, meaning that concentration can be focused only on the inner summations, over j (due the convexity of the sum of convex functions). Now, the goal becomes to prove that

$$MI_2(f_{i,j}) = \sum_{j=1}^{N_w} f_{i,j}\log(f_{i,j}) - \left[\sum_{j=1}^{N_w} f_{i,j}\right]\cdot\left[\log\left(\sum_{j=1}^{N_w} f_{i,j}\right)\right] \tag{A.4}$$

is convex for a given i. To this end, the following notational substitutions are made:

$$f_{i,j} = x_k\,, \qquad x = [x_1, x_2, \ldots, x_N]^T\,, \qquad X = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & x_N \end{bmatrix}$$

where $N = N_W$. Also, define u to be a column vector of ones of length N, and $U = uu^T$. With this new notation, the problem reduces to demonstrating that

$$h(x) = x^T \log(x) - (u^T x) \log(u^T x)$$

is convex. To this end, it will be shown that the Hessian of h is positive semi-definite.

$$\nabla h(x) = \begin{bmatrix} \log(x_1) \\ \log(x_2) \\ \vdots \\ \log(x_N) \end{bmatrix} - \begin{bmatrix} \log(u^T x) \\ \log(u^T x) \\ \vdots \\ \log(u^T x) \end{bmatrix}\,,\ \nabla^2 h(x) = X^{-1} - \frac{1}{u^T x} U$$

For the Hessian to be positive semi-definite, it needs to be shown that, for all non-zero $y \in \mathbb{R}^N$,

$$y^T \nabla^2 H(x) y = y^T X^{-1} y - \frac{1}{u^T x} y^T U y \geq 0.$$

Breaking this into two pieces, address the right-most element first:

$$y^T U y = (u^T y)^2 \implies \frac{1}{u^T x} y^T U y = \frac{(u^T y)^2}{u^T x}.$$

The remaining part of the equation can be simplified to:

$$y^T X^{-1} y = \sum_{j=1}^{N} \frac{y_j^2}{x_j},$$

and thus, it remains to be shown that

$$\sum_{j=1}^{N} \frac{y_j^2}{x_j} \geq \frac{(u^T y)^2}{u^T x}. \tag{A.5}$$

In order to demonstrate (A.5), two additional assumptions will be made:

Assumption A.2 *Assume, without loss of generality,* $y_j \geq 0 \; \forall \, j$.

The reason that this assumption can be made without loss of generality is that replacing y with $|y|$ will only increase $u^T y$ while leaving y_j^2 unchanged. Implicitly, this uses the fact that requiring $|y_1 + y_2 + \cdots + y_N| \leq |y_1| + |y_2| + \cdots + |y_N|$ does nothing to aid in the proof.

Assumption A.3 *Assume, without loss of generality,* $\sum y_j = u^T y = 1$.

This assumption is allowed because the sign of $y^T \nabla^2 h(x) y$ is invariant with respect to a scaling of y and $u^T y = 0$ is impossible when $y_i \geq 0$ and $y \neq 0$.

Before taking advantage of the two assumptions, note that

$$\sum_{j=1}^{N} \frac{y_j^2}{x_j} = \sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j} \right)^{-1}.$$

Therefore, show:

$$\sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j} \right)^{-1} \geq \frac{(e^T y)^2}{e^T x}.$$

To this end,

(1) Define $q_j = \frac{x_j}{y_j}$,
(2) Recall that $y_j \geq 0$ for $i = 1, \ldots, N$ and $\sum y_j = e^T y = 1$,
(3) Note that $f(q) = 1/q$ is a convex function.

Therefore

$$\sum_{j=1}^{N} y_j f(q_j) \geq f\left(\sum_{j=1}^{N} y_j q_j \right)$$

$$\implies \sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j} \right)^{-1} \geq \left[\sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j} \right) \right]^{-1} = \frac{1}{u^T x} = \frac{(u^T y)^2}{u^T x}.$$

Therefore, $\nabla^2 h(x)$ is positive semi-definite. This means that equation (A.4) is convex for a given i, and thus that the sum over i in equation (A.3) is convex making equation (A.2)

convex, meaning that the mutual information constraint, equation (A.1), is convex for $f_{i,j} > 0$.

□

B The Iterative Procedure

There is an alternative solution procedure for the two-period RI problem. This problem lends itself to a semi-analytical approach based on iteration on the first-order conditions of the optimization problem. As was noted earlier, the use of solution methods where $f_{i,j} > 0$ will produce qualitatively identical results because $\lim_{x \to 0} x \log(x) = 0$, meaning the that solution has no discontinuities at $f_{i,j} = 0$ and therefore the results where some $f_{i,j}$'s $< \varepsilon$ where ε is very small (say 10^{-8}) will incorporate the properties of the truly optimal result. The first order conditions for $f_{i,j} > 0$ are:

$$U_{i,j} - \lambda \left[\log(f_{i,j}) - \log \left(\sum_{j=1}^{N_w} f_{i,j} \right) \right] - \omega_j = 0$$

where

$$U_{i,j} = \frac{c_i^{1-\gamma} + (w_j - c_i)^{1-\gamma}}{1-\gamma},$$

Specifying a value for λ (the multiplier on the information-processing constraint, equation (5)) is identical to choosing a κ (as expected, different parameterizations result in different λ's for the same value of κ). The utility function and λ are known, so elimination of ω_j (the multiplier on the constraint requiring the joint distribution to conform to the properties of the exogenous wealth distribution, equation (3), for a specific j) is all that is required before solving for the probabilities. This is done by taking advantage of the log-properties inherited from the entropic constraint.

$$\begin{aligned} \frac{1}{\lambda}\left(U_{i,j} - \omega_j\right) &= \log \left(\frac{f_{i,j}}{\sum_{j=1}^{N_w} f_{i,j}} \right) \\ \implies \frac{f_{i,j}}{\sum_{j=1}^{N_w} f_{i,j}} &= \exp\left[(1/\lambda)(U_{i,j} - \omega_j)\right] \end{aligned} \tag{B.1}$$

At this point, it should be noted that the denominator of the LHS of (B.1) represents the marginal probability of c_i, which is called $p(c_i)$. Thus,

$$f_{i,j} = \exp\left[(1/\lambda)(U_{i,j} - \omega_j)\right] p(c_i) = \frac{\exp[(1/\lambda)U_{i,j}]}{\exp[\omega_j/\lambda]} p(c_i).$$

Equation (3) yields that:

$$\begin{aligned} \sum_{i=1}^{N_c} f_{i,j} = g(w_j) &= \frac{1}{\exp[\omega_j/\lambda]} \sum_{i=1}^{N_c} \exp[(1/\lambda)U_{i,j}]p(c_i) \\ \Longrightarrow \quad \frac{1}{\exp[\omega_j/\lambda]} &= \frac{g(w_j)}{\sum\limits_{i=1}^{N_c} \exp[(1/\lambda)U_{i,j}]p(c_i)}. \end{aligned}$$

Therefore, a solution (almost) for the probabilities is given as:

$$f_{i,j} = \frac{\exp[(1/\lambda)U_{i,j}]p(c_i)g(w_j)}{\sum\limits_{i=1}^{N_c} \exp[(1/\lambda)U_{i,j}]p(c_i)}. \tag{B.2}$$

Equation (B.2) is the solution, but recall that $p(c_i) = \sum_{j=1}^{N_w} f_{i,j}$.[14] The procedure is completed via iteration on f. Starting with a random f matrix and generating a marginal distribution over c by summing, one can use these values to construct the solutions for $\{f_{i,j}\}$, then sum the rows of the f matrix to form the next iteration of $p(c)$, and continue until subsequent f distributions are arbitrarily close to each other, giving us $\{f_{i,j}\}$ values which satisfy (B.2). This procedure appears to converge on the same distribution for any starting f distribution. Successive iterations are within 10^{-7} of each other within 60 seconds when done using `MATLAB` on a 3 GHz Pentium 4 running Windows XP.

Equation (B.2) shows that the heart of the RI framework is interaction between the information-processing constraint (recall that λ is the LaGrange multiplier on that constraint) and the agent's preferences. The agent chooses how much attention to pay, and where, based on that interaction. That the agent's risk tolerance affects what he or she observes has the potential to open up several new avenues of reseach in the future.

It should be noted that this procedure derives identical solutions to the procedure outlined above using `AMPL`, and without using sophisticated optimizers, but also that it is slower.

[14] That is, equation (B.2) must hold for all feasible (c_i, w_j) pairs at the optimum, but much like value-function iteration, the solution lies in the answer to the question "what $\{f_{i,j}\}$ matrix makes (B.2) true?"

Additionally, this problem has an undiscounted utility function and it's static nature make FOC-based analysis possible.

B.1 Theory vs. Computation

The quasi-analytical approach of this appendix yields an equation for the probability of consuming c_i at wealth level w_j driven by $\exp[(1/\lambda)U_{i,j}]$. While this theoretical result is central to RI theory (that the probability is driven by the interaction of the utility of that (c, w) pair and the processing capacity), this equation represents a potential numerical pitfall. The problem is one of computer accuracy: when the absolute value of $U_{i,j}/\lambda$ is large for all (i, j) combinations, the theory will predict a smooth, descriptive function of $U_{i,j}$ while the computer will return either zero or ∞ for all (c_i, w_j) pairs (If the utility is negative, zero; if positive, ∞). The $U_{i,j}/\lambda$ problem is purely an artifact of the computer's inability to deal with *very* large or small numbers, but it is dramatically exacerbated in this exponential situation. [15] The bad news is that $\exp[(1/\lambda)U_{i,j}]$ *is* central to RI theory and therefore present in analytically represented RI model-solutions in general. The good news is that it is easy to identify: utility and λ values can be determined, and model-designers can plan to work around, or find model-specific solutions to, this issue.

B.2 Results Compared to `AMPL/KNITRO`

Because of the first-order-conditions-based approach of the iterative method, this technique requires that $f_{i,j} > 0$, which, as seen above, is not reasonable. Due to the interior-point nature of the `KNITRO` solver, the results are identical up to the error introduced by the stopping tolerances of the iterative scheme.

[15] The severity of this problem is a function of the software and architecture of the computer being used for the calculations. As computers grow in sophistication, this problem will be alleviated but never eliminated.