2007

Robustness and information processing constraints in economic models

Kurt Frederick Lewis
University of Iowa

ROBUSTNESS AND INFORMATION PROCESSING CONSTRAINTS IN ECONOMIC MODELS

by

Kurt Frederick Lewis

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Economics
in the Graduate College of
The University of Iowa

July 2007

Thesis Supervisor: Professor Charles H. Whiteman

ABSTRACT

In this dissertation, I examine the impact of uncertainty and information processing restrictions on standard economic models. Chapter 1 examines a reevaluation of the excess volatility puzzle in asset prices by assessing the impact of a shift in the agent's focus from minimizing average loss to minimizing maximum loss. Chapters 2 and 3 extend and clarify the newly developing arena of economic models in which the agent's capacity for information processing is systematically limited, as in the recent rational inattention literature.

Chapter 1, which represents joint work with Charles Whiteman, studies the consequences changing the present value formula for stock prices. In place of the squared-error-loss minimizing *expected* present value of future dividends, we use a predictor optimal for the min-max preference relationship appropriate in cases of ambiguity. With "robust" predictions, the well-known variance bound is reversed in that prices are predicted to be far more volatile than what is observed. We also investigate an intermediate "partially robust" case in which the degree of ambiguity is limited, and discover that such an intermediate model cannot be rejected in favor of an unrestricted time series model.

Chapter 2 demonstrates the properties and solutions for the more general two-period rational inattention model. We show that the problem is convex, can be solved in seconds, and highlights several important features of information-processing-capacity-constrained models. Additionally, we show the importance of deriving,

rather than assuming, the form of the final solution in rational inattention models.

Chapter 3 extends the work of Chapter 2 to a finite-horizon dynamic setting by creating a structure in which distributional state and control variables interact under information-processing constraints. Limited information processing capacity is used optimally, and agents have the opportunity to trade processing capacity for higher expected future income. The framework is applied to the canonical life-cycle model of consumption and saving, and an analysis of the impact of preference parameters on optimal attention allocation is conducted. The model produces a distinct hump-shaped profile in expected consumption.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

ROBUSTNESS AND INFORMATION PROCESSING CONSTRAINTS IN ECONOMIC MODELS

by

Kurt Frederick Lewis

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Economics
in the Graduate College of
The University of Iowa

July 2007

Thesis Supervisor: Professor Charles H. Whiteman

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Kurt Frederick Lewis

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Economics at the July 2007 graduation.

Thesis Committee: _____
Charles H. Whiteman, Thesis Supervisor

_____
John F. Geweke

_____
M. Beth Ingram

_____
N. Eugene Savin

_____
Kurt M. Anstreicher

For my girls.

On Prioritizing Information Processing

Bruno: He might get asked about Title IX.
C.J.: What are your thoughts?
Bruno: On Title IX?
C.J.: Yeah.
Bruno: I have none. I'm indifferent.
C.J.: You can't be indifferent.
Bruno: I have to be. I have only so much RAM in my head. I have to prioritize. I have to throw some things overboard, so, I've chosen, for instance, not to care whether or not Purdue has a fencing team.

— Aaron Sorkin ("College Kids", *The West Wing*)

ACKNOWLEDGEMENTS

I would like to thank my committee members: Kurt Anstreicher, John Geweke, Beth Ingram, Gene Savin and especially my advisor Charles Whiteman for their helpful comments and encouragement. Like many of his students, much of what I know about economics I owe to working on joint projects with Charles Whiteman, an example being the first chapter of this dissertation.

Learning economics at the University of Iowa has been not just the culmination of my formal education, but also a personal high-point — this process has simultaneously sated and increased my desire for the pursuit of knowledge. The faculty, students and staff have each been responsible for a large component of my education and while I'm grateful to everyone, there are several I would like to thank by name. From the faculty beyond my committee I would like to thank B. Ravikumar, Paul Muhly and Paolo Ghirardato. From among my fellow graduate students (current and past) I would like to thank Todd Walker, Mark Kurt, James Chapman and Ryan Haley. Additionally, I would like to thank Renea Jay, the Ph.D. program's "mom", who works tirelessly as an advocate for each and every one of her charges, including those of us who rarely turned things in on time.

Finally, this dissertation would not have been possible without the love and support of my wife, Tara, our two girls, Olaiya and Naomi, and our families.

ABSTRACT

In this dissertation, I examine the impact of uncertainty and information processing restrictions on standard economic models. Chapter 1 examines a reevaluation of the excess volatility puzzle in asset prices by assessing the impact of a shift in the agent's focus from minimizing average loss to minimizing maximum loss. Chapters 2 and 3 extend and clarify the newly developing arena of economic models in which the agent's capacity for information processing is systematically limited, as in the recent rational inattention literature.

Chapter 1, which represents joint work with Charles Whiteman, studies the consequences changing the present value formula for stock prices. In place of the squared-error-loss minimizing *expected* present value of future dividends, we use a predictor optimal for the min-max preference relationship appropriate in cases of ambiguity. With "robust" predictions, the well-known variance bound is reversed in that prices are predicted to be far more volatile than what is observed. We also investigate an intermediate "partially robust" case in which the degree of ambiguity is limited, and discover that such an intermediate model cannot be rejected in favor of an unrestricted time series model.

Chapter 2 demonstrates the properties and solutions for the more general two-period rational inattention model. We show that the problem is convex, can be solved in seconds, and highlights several important features of information-processing-capacity-constrained models. Additionally, we show the importance of deriving,

rather than assuming, the form of the final solution in rational inattention models.

Chapter 3 extends the work of Chapter 2 to a finite-horizon dynamic setting by creating a structure in which distributional state and control variables interact under information-processing constraints. Limited information processing capacity is used optimally, and agents have the opportunity to trade processing capacity for higher expected future income. The framework is applied to the canonical life-cycle model of consumption and saving, and an analysis of the impact of preference parameters on optimal attention allocation is conducted. The model produces a distinct hump-shaped profile in expected consumption.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
ROBUSTIFYING SHILLER: DO STOCK PRICES MOVE *ENOUGH* TO BE JUSTIFIED BY SUBSEQUENT CHANGES IN DIVIDENDS?

1.1 Introduction

In 1981, Shiller (1981) and LeRoy and Porter (1981) sparked a controversy whose legacy even today continues to occupy the attention of researchers in economics and finance. Their focus was on the simplest present value model of stock prices: that a stock price corresponds to the expected present value of the stock's dividends discounted at a constant rate. Using the orthogonality property of least squares projections (that projection errors are uncorrelated with information used in constructing the projection), they showed that the *actual* present value must be no less variable than its expectation, which according to the present value model is the current stock price. In the data, this variance bound is violated in dramatic fashion–stock prices are far more variable than subsequent realizations of dividend would appear to permit.

The controversy involves *why* prices are so volatile–are stock prices influenced by fads or "irrational exuberance"? Or, is something amiss with the volatility calculations, the treatment of dividends, or the assumption of a constant discount factor? Much work has been undertaken to pursue each of these latter three resolutions of the puzzle. For example, Flavin (1983) focussed on whether the sampling variability in the volatility calculations could be sufficient to generate the result even when the simple pricing model is true. Regarding dividends, Marsh and Merton (1986)

argued that if the dividend process is difference stationary (rather than trend stationary, as Shiller had assumed), the variance bound is reversed. Time series work suggested, however, that the trend-stationary assumption was more plausible than the difference-stationary one (e.g. DeJong and Whiteman (1991)). Subsequent efforts included development of volatility tests that loosen stationarity assumptions (e.g., Campbell and Shiller (1987) and West (1988)).

Still, the controversy was unresolved. Perhaps if the nature of the econometric procedures or the assumption regarding the dividend process were not the culprits, the fault might lie with the economic environment. Indeed, the assumption of a constant discount factor was challenged quickly by LeRoy and LaCivita (1981) and Michener (1982), who showed that with a stochastic discount factor (the intertemporal marginal rate of substitution) appropriate for a representative risk-averse agent, some (but alas not enough) extra volatility would be generated in stock prices.

In 1985, the controversy morphed into the more general but closely related "equity premium" puzzle (Mehra and Prescott (1985)): the "excessively" volatile stock prices are associated with a return to stocks that is too much larger than the very low return to bonds to be explained in the context of the simplest asset pricing model unless the representative consumer is unrealistically risk averse. Hansen and Jagannathan (1991) developed a version of Shiller's variance bound for more general economies which has facilitated the study of more general specifications of preferences involving time-nonseparabilities (Constantinides (1990)) and state-nonseparabilities (Epstein and Zin (1989)) These modifications have not resolved the controversy either

(Otrok et al. (2002)).

We illustrate a different approach: it isn't sampling variability in volatility estimates, it isn't the dividend assumption, it isn't the discount factor, *it's the expectation*. We study the simplest, Shiller-style present value model, and imagine that the predictions of future dividends are not those of a forecaster minimizing squared error loss, but rather those of a forecaster facing and dealing with ambiguity regarding the economic environment. That is, the forecaster admits the possibility that the correct prediction model is unknown. To accommodate such an unpleasant but realistic situation, as much decision-theoretic literature suggests, the forecaster might behave in such a way as to mitigate the worst outcome that could conceivably occur.

To motivate this minimax loss function, consider the following experimental example of an agent's aversion to uncertainty and of actions taken to minimize downside risk that was originally proffered in Ellsberg (1961):

Experiment 1

> There are two urns, labeled Urn_I and Urn_{II}, each containing a total of 100 chips painted either black or white. The subject is allowed to examine the contents of Urn_I and discovers that it contains exactly 50 white chips and 50 black chips. The subject is not permitted to examine Urn_{II} other than to be told it contains 100 chips colored either black or white, but in an unknown ratio. The subject is given a list of four possible gambles:
>
> (A) The chip drawn from Urn_I is black.

(B) The chip drawn from Urn_I is white.

(C) The chip drawn from Urn_{II} is black.

(D) The chip drawn from Urn_{II} is white.

Winning a bet entitles the agent to a prize of value, for example \$100.

In studies replicating this experiment, the following preference orderings over gambles A through D have been observed: $A \sim B \succ C \sim D$.[1] As noted by Gilboa and Schmeidler (1989), who cite a version of this experiment in their seminal minimax decision theory work, it is easy to see that there is no probability measure supporting these preference orderings under expected utility maximization. Minimax decision theory is based on the following explanation of this preference ordering: Aversion to the uncertainty regarding the contents of Urn_{II} causes the agent to treat the worst case scenario among a set of possibilities as the foundation of his prior probabilities on Urn_{II}. Thus, the agent's preferences represent a desire to minimize potential downside risk.

In the present value prediction problem, this means replacing ordinary expectations of future dividends with robust predictions. The standard reference for this in the economics literature is Hansen and Sargent (2005), who for the most part employ state space methods for calculating robust procedures. Such methods are convenient for numerical calculations, but make analytic derivations difficult. Because the envi-

[1]One example of a classroom experiment in which this result is exhibited is found in Study 1 of Fox and Tversky (1995), which used 141 undergraduates at Stanford University, responding to a questionnaire consisting of this and several other unrelated items that subjects completed for class credit.

ronment is simple and we wish to illustrate the effect of robust decisions on economic outcomes, we utilize related "frequency domain" procedures that build on Whiteman (1985), Whiteman (1986), and Kasa (2001). As these procedures are arcane but not deep, we present details of the derivations to make the paper self-contained. Briefly, we exploit the robust decision-maker's aversion to serially correlated errors to derive the robust present value stock price analytically.

Our results indicate that robust predictions can be quite different from ordinary ("least squares") predictions, and the robust expected present value of dividends can be quite variable–so variable, in fact, that our initial univariate calculations (robust analogues of Shiller's) *reverse* the volatility relationship in dramatic fashion. It turns out, however, that this result implies that the robust predictor is hedging against a potential degree of model misspecification that, while possible, is extremely unlikely.

To obtain a reasonable limitation on the degree of robustness the predictor adopts, we employ a version of the "evil agent" game of Hansen and Sargent (2005). In this game, the predictor uses least squares methods, but plays a dynamic game against an evil "nature" who may deliver dividend processes different from those the forecaster believes to characterize the data. Nature is constrained in how much noise she can add to the situation. With this formulation, the question becomes one of how much freedom nature would have to possess to cause the predictor agent to make present value predictions consistent with actual stock prices, and whether this would be implausible. Our estimates of a simple evil agent game suggest that the required

freedom is not that great and that the other implications of the model might not be implausible. That is, the "moderately robust" present value model generates prices that *are* consistent with subsequent changes in dividends.

1.2 The Setup and the Variance Bound

Shiller (1981) begins with the simple present value model

$$p_t = E_t \sum_{k=0}^{\infty} \gamma^k d_{t+k}. \tag{1.1}$$

where $\gamma = 1/(1+r)$ with r the (assumed constant) real rate of interest, p_t and d_t are the real stock price and divident at time t, and E_t denotes conditional expectation given information available at t. Shiller writes the model in terms of the ex post rational price series p_t^*, which is defined as the present value of the *actual* subsequent real dividends:

$$p_t^* = \sum_{k=0}^{\infty} \gamma^k d_{t+k}. \tag{1.2}$$

Clearly, as this requires us to have the *actual* dividend sequence out into the infinite future, it is impossible to observe p_t^* without error. Shiller notes that with a long enough dividend sequence, we can observe an approximate p_t^* by choosing a terminal date, making some assumptions about the dividend series after that terminal date and then constructing the p_t^* series via backwards recursion using[2]

[2]The assumptions that Shiller makes about dividends after the terminal date are that they are smooth and grow at the exponential growth rate consisten with of the original data.

$$p_t^* = \gamma(p_{t+1}^* + d_{t+1}). \tag{1.3}$$

In any case, from (1.1) and (1.2), we have

$$p_t = E_t(p_t^*) \tag{1.4}$$

which implies

$$p_t^* = p_t + \epsilon_t$$

where ϵ_t is orthogonal to information available at time t (including p_t). Then $var(p_t^*) = var(p_t) + var(\epsilon_t)$, and we have

$$\sigma(p) \leq \sigma(p^*). \tag{1.5}$$

That actual p is vastly more variable than measured p^* is demonstrated in Shiller's Figure 1 for the SP Composite Index from 1871-1979, as replicated below. The p^* in the figure was generated by the backwards recursion on (1.3) and an estimated $r = 0.048$, implying $\gamma = 0.943$. For the data in the figure, the estimated standard deviation of dividends $\sigma(d)$ is 1.12 while $\sigma(p)$ is 42.74 and $\sigma(p^*)$ is 7.24. Note that price volatility is about 34 times that of dividends.[3]

[3]An updated dataset has been examined and found to yield qualitatively the same results as what is presented throughout this paper. Future work will include results with updated data, but our results remain largely unchanged, see footnote 6 on page 38 for more details.

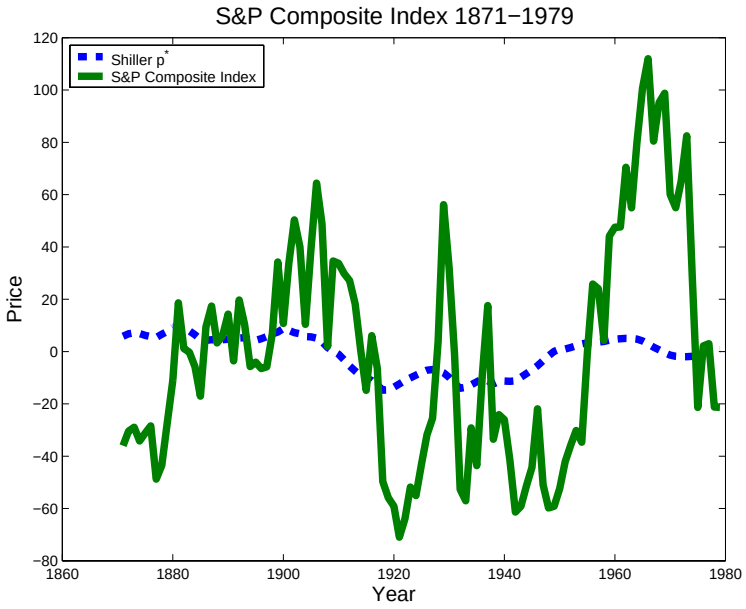


Figure 1.1: Shiller's Figure 1

1.3 The Least Squares Prediction

To provide a benchmark for the robust present value calculation to be presented below, in this section we present a first-principles derivation of the stochastic process for stock prices implied by the present value model and the assumption that the dividend process is stationary after detrending. This benchmark case involves computation of the so-called "Hansen and Sargent (1980) formula".[4] The robust calculation of the ensuing section is an example of the "robust Hansen-Sargent formula" of Kasa (2001).

We will assume that the agent's information set includes current and past observations on a trend-stationary dividend process; in fact, we will assume that the

[4]A survey of these methods can be found in Whiteman (1983).

market forecasters know that the data generating process for the detrended dividends (i.e., the "model") coincides with the Wold representation, which we write

$$d_t = \sum_{j=0}^{\infty} q_j \varepsilon_{t-j} = q(L)\varepsilon_t \qquad\qquad E(\varepsilon_t) = 0\,, \;\; E(\varepsilon_t^2) = 1 \tag{1.6}$$

using the lag operator L and defining $q(L)$ implicitly. Using (1.6) and (1.1), we have

$$p_t = E_t \sum_{j=0}^{\infty} \gamma^j d_{t+j} = E_t \left(\frac{q(L)}{1-\gamma L^{-1}} \right) \varepsilon_t = E_t(p_t^*). \tag{1.7}$$

The coincidence of the Wold representation and the data generation process means that ε_t is both statistically and economically "fundamental" for d_t. Thus (1.6) not only *represents* dividends via the Wold representation, but also *generates* dividends: (1.6) is the "correct" model. We defer to the next section the case in which the representative investor is not so sure about this.

The least-squares minimization problem the representative agent faces is to find a stochastic process p_t to minimize the expected squared prediction error $E(p_t - p_t^*)^2$. In terms of the information known at date t, the agent's task is to find a linear combination of current and past dividends, or, equivalently, of current and past dividend innovations ε_t that is "close" to p_t^*. Writing $p_t = f(L)\varepsilon_t$, the problem becomes one of finding the coefficients f_j in $f(L) = f_0 + f_1 L + f_2 L^2 + ...$ to minimize $E(f(L)\varepsilon_t - p_t^*)^2$. Using the technique in Whiteman (1985), this problem has an equivalent, frequency-domain representation

$$\min_{f(z)\in H^2} \frac{1}{2\pi i} \oint \left| \frac{q(z)}{1-\gamma z^{-1}} - f(z) \right|^2 \frac{dz}{z} \tag{1.8}$$

where H^2 denotes the Hardy space of square-integrable analytic functions on the unit disk, and $\oint$ denotes (counterclockwise) integration about the unit circle. The requirement that $f(z) \in H^2$ ensures that the forecast is causal, and contains no future values of the ε's. Expression (1.8) may seem exotic, but is quite simple: noting that on the unit circle $z = e^{i\omega}$, making the substitution and integrating with respect to ω from 0 to 2π, the expression calls for minimizing the average value (or area under) the spectral density of the prediction error $p_t - p_t^*$. Of course this follows from the fact that the variance of a process is equal to the integral of its spectral density. In the next section, we shall study a forecaster who seeks a function $f(z)$ to minimize not the average value on the unit circle, but the *maximum* value on the unit circle.

The first-order conditions for choosing f_j are, for $j = 0, 1, 2, ...,$

$$\begin{aligned}
&\frac{\partial}{\partial f_j}\left[\frac{1}{2\pi i}\oint\left|\frac{q(z)}{1-\gamma z^{-1}} - f(z)\right|^2 \frac{dz}{z}\right] \\
&\quad = -\frac{1}{2\pi i}\oint z^j\left[\frac{q(z^{-1})}{1-\gamma z} - f(z^{-1})\right] + z^{-j}\left[\frac{q(z^{-1})}{1-\gamma z^{-1}} - f(z)\right]\frac{dz}{z} \\
&\quad = -\frac{1}{2\pi i}\oint z^j\left[\frac{q(z^{-1})}{1-\gamma z} - f(z^{-1})\right]\frac{dz}{z} - \frac{1}{2\pi i}\oint z^{-j}\left[\frac{q(z^{-1})}{1-\gamma z^{-1}} - f(z)\right]\frac{dz}{z} = (1.9)
\end{aligned}$$

Now change variables in the second contour integral to $w = z^{-1}$, implying $dw = -1(z^{-2})$ and $dz/z = -dw/w$, with integration with respect to w being clockwise. Upon changing variables in the first integral and multiplying by -1 to reverse integration back to clockwise, the the two integrals in (1.9) become identical and the equality collapses to

$$-\frac{2}{2\pi i}\oint z^{-j}\left[\frac{q(z)}{1-\gamma z^{-1}}-f(z)\right]\frac{dz}{z}=0. \tag{1.10}$$

Now define

$$H(z)=\frac{q(z)}{1-\gamma z^{-1}}-f(z) \tag{1.11}$$

so that (1.10) becomes

$$-\frac{2}{2\pi i}\oint z^{-j}H(z)\frac{dz}{z}=0. \tag{1.12}$$

This equality requires that for $j=0,1,2,\ldots,$ twice the coefficient H_j in the Laurent expansion of $H(z)$ valid for $|z|=1$ equal zero. Multiplying by z^j and summing over *all* $j=0,\pm1,\pm2,...,$ we find that

$$H(z)=\sum_{-\infty}^{-1}$$

where $\sum_{-\infty}^{-1}$ denotes an unknown function involving only negative powers of z. Then by recalling the definition of $H(z)$, we have

$$\frac{q(z^{-1})}{1-\gamma z^{-1}}-f(z)=\sum_{j=-\infty}^{-1}$$

which is known as a "Wiener-Hopf" equation. Application of the "plussing" operator to both sides of the equation yields:[5]

[5]The plussing operator is a linear annihilator operator that means "ignore negative powers of z."

$$\left[\frac{q(z)}{1-\gamma z^{-1}}\right]_+ - [f(z)]_+ = 0$$

implying

$$f(z) = \left[\frac{q(z)}{1-\gamma z^{-1}}\right]_+ = \left[\frac{zq(z)}{z-\gamma}\right]_+ \tag{1.13}$$

which points to the fact that $f(z)$ is, by construction, one-sided in non-negative powers of z. We now use a method highlighted in Appendix A of Hansen and Sargent (1980) to determine the form of the function in (1.13). First, note that the function being "plussed" in (1.13) is well-behaved $|z| < 1$ except for a single simple pole at γ. By examining the Laurent expansion of $q(z)$ around γ, we are able to determine that the principle part (that is, the part of the Laurent expansion containing negative powers of z) is $P(z) = \gamma q(\gamma)/(z-\gamma)$. Second, we note that "plussing" involves simply subtracting off those parts. That is:

$$[A(z)]_+ = A(z) - P(z)$$

where $P(z)$ is the principle part of the Laurent series expansion of $A(z)$. This implies

$$f(z) = \left[\frac{q(z)}{1-\gamma z^{-1}}\right]_+ = \left[\frac{zq(z)}{z-\gamma}\right]_+ = \frac{zq(z)-\gamma q(\gamma)}{z-\gamma}.$$

To illustrate how the formula works, suppose detrended dividends are described by a first-order autoregression; i.e., that $q(L) = (1-\rho L)^{-1}$. Then

$$p_t = f(L)\varepsilon_t = \frac{Lq(L) - \gamma q(\gamma)}{L - \gamma}\varepsilon_t = \left(\frac{1}{1 - \rho\gamma}\right) d_t.$$

In this simple first order case,

$$\sigma(p) = \left(\frac{1}{1 - \rho\gamma}\right) \sigma(d). \tag{1.14}$$

With $\gamma = 0.943$, as estimated from the S&P data, the largest $\sigma(p)$ can be for stationary dividends ($|\rho| < 1$) is about 22 times the standard deviation of dividends. Estimating ρ from the same data, the ratio is about 11, far short of the factor of 34 needed to match the observed volatility.

It is instructive to note that while the pricing formula (1.13) makes p_t the best least squares predictor of p_t^*, the prediction errors $p_t - p_t^*$ will not be serially uncorrelated. Indeed

$$\begin{aligned} p_t - p_t^* &= \gamma \left\{ \frac{Lq(L) - \gamma q(\gamma)}{L - \gamma} - \frac{q(L)}{1 - \gamma L^{-1}} \right\} \varepsilon_t \\ &= \frac{-\gamma^2 q(\gamma)}{L - \gamma}\varepsilon_t \;=\; -\gamma^2 q(\gamma) \frac{L^{-1}}{1 - \gamma L^{-1}}\varepsilon_t \\ &= -\gamma^2 q(\gamma) \left\{ \varepsilon_{t+1} + \gamma\varepsilon_{t+2} + \gamma^2\varepsilon_{t+3} + \dots \right\}. \end{aligned}$$

Thus the prediction errors will be described by a highly persistent (γ is close to unity) stochastic process having a Wold representation that is a first-order autoregression. But because this autoregression involves *future* ε_t's, the serial correlation structure of the errors cannot be exploited to improve the quality of the prediction of p_t^*. The reason is that the predictor "knows" the *model* for price setting (the present value

formula) and the dividend process; the best predictor $p_t = E_t p_t^*$ of p_t^* "tolerates" the serial correlation because the (correct) model implies that it involves *future* ε_t's and therefore cannot be predicted. If one only had data on the errors (and did not know the model that generated them), they would appear (rightly) to be characterized by a first-order autoregression; fitting an AR(1) (i.e., the best *linear* model) and using it to "adjust" p_t by accounting for the serial correlation in the errors $p_t - p_t^*$ would decrease the quality of the estimate of p_t^*. The reason is the usual one that the Wold representation for $p_t - p_t^*$ is not the economic model of $p_t - p_t^*$, and (correct) models always beat Wold representations. This also serves as a reminder of circumstances under which one should be willing to tolerate serially correlated errors: when one knows the model that generated them, and the model implies that they are as small as they can be made.

1.4 The Robust Prediction Case

What happens in case the individual making the prediction of future dividends does not know for certain that dividends are generated as in (1.6)? This notion of ambiguity was introduced in the linear, time-invariant context we are studying in the engineering literature by Zames (1981), and has been studied more generally in the economics literature by Gilboa and Schmeidler (1989), Hansen and Sargent (2005) and others. In our setup, the ambiguity would be manifested in possible departures from the moving average representation (1.6). Following the development Kasa (2001) used in a related context, suppose the dividend process is given by

$$d_t = [q(L) + \Delta(L)]\,\varepsilon_t \tag{1.15}$$

where $\Delta(L)\varepsilon_t$ is a "perturbation" of the original dividend process. Then if the forecaster uses (1.13), the actual squared error loss $\mathcal{L}^A = E[p_t - p_t^*]^2$ will be given by

$$\begin{aligned}
\mathcal{L}^A &= \mathcal{L}^q + \|\Delta(z)\|_2^2 + \frac{2}{2\pi i}\oint \Delta(z)\left[\frac{q(z)}{1-\gamma z^{-1}}\right]_- \frac{dz}{z} \\
&= \mathcal{L}^q + \|\Delta(z)\|_2^2 + \frac{2}{2\pi i}\oint \Delta(z)\left(\frac{\gamma q(\gamma)}{z-\gamma}\right)\frac{dz}{z} \\
&= \mathcal{L}^q + \|\Delta(z)\|_2^2 + 2q(\gamma)\left[\Delta(\gamma) - \Delta(0)\right]
\end{aligned}$$

where $\mathcal{L}^q$ is the loss when dividends are indeed generated by $d_t = q(L)\varepsilon_t$. The second line follows from the linear annihilator operator $[\cdot]_-$ which means "ignore positive powers of z." This leaves only the principle part of the element in the brackets, which was shown earlier to be $\gamma q(\gamma)/(z-\gamma)$. The third equality is a result of the application of Cauchy Residue Theorem. The expression for $\mathcal{L}^A$ indicates that the actual loss could be much larger than $\mathcal{L}^q$ even for a small perturbation provided $q(\gamma)$ is large. This result, combined with the knowledge that the true dividend process is hard to come by, suggests that the forecast should be constructed with greater robustness to model misspecification.

The problem with a misspecified model is that the "wrong" sequence of "errors" ε_t could "excite" $\Delta(L)$ in such a way that very large prediction errors occur. To guard against this, the predictor might wish to make forecasts that minimize the maximum possible squared error loss rather than the average or expected squared

error loss.

The robust predictor solves

$$\min_{f(z)\in H^\infty} \max_{|z|=1} \left| \frac{q(z)}{1-\gamma z^{-1}} - f(z) \right|^2 \Leftrightarrow \min_{f(z)\in H^\infty} \max_{|z|=1} \left| \frac{zq(z)}{z-\gamma} - f(z) \right|^2 . \tag{1.16}$$

Unlike in the least squares case (1.8), where $f(z)$ was restricted to the class H^2 functions finitely square integrable on the unit circle, the restriction now is to the class of functions with finite maximum modulus on the unit circle, and the H^2 norm has been replaced by H^∞ norm.

To begin the solution process, note that there is considerable freedom in designing the minimizing function $f(z)$: it must be well-behaved (i.e., must have a convergent power series in nonnegative powers of z on the unit disk), but is otherwise unrestricted. Further note that $zq(z)/(z-\gamma)$ can be thought of as the associated Laurent expansion, which is of the form

$$\frac{zq(z)}{z-\gamma} = \frac{b_{-1}}{z-\gamma} + b_0 + b_1(z-\gamma) + b_2(z-\gamma)^2 + ...$$

Intuitively, while in the least squares case $f(z)$ is set to "cancel" all the terms of this series except the first, here the object is to set $f(z)$ to minimize a different function of the prediction errors. Now define the "Blaschke factor" $B_\gamma(z) = (z-\gamma)/(1-\gamma z)$ and note that

$$\left| \frac{z-\gamma}{1-\gamma z} \right|^2 = \frac{(z-\gamma)(z^{-1}-\gamma)}{(1-\gamma z)(1-\gamma z^{-1})} = \frac{(z-\gamma)z^{-1}(1-\gamma z)}{(1-\gamma z)z^{-1}(z-\gamma)} = 1.$$

Multiplying the objective by the Blaschke factor thus does not alter its value on the unit circle, but the factor does cancel the pole at γ, yielding

$$\min_{\{f(z)\}} \sup_{|z|=1} \left| \frac{zq(z)}{1-\gamma z} - \frac{z-\gamma}{1-\gamma z} f(z) \right|^2 .$$

Defining

$$T(z) = \frac{zq(z)}{1-\gamma z} \tag{1.17}$$

we have

$$\min_{f \in H^\infty} \sup_{|z|=1} |T(z) - B_\gamma(z) f(z)| \Leftrightarrow \min_{f \in H^\infty} \|T(z) - B_\gamma(z) f(z)\|_\infty. \tag{1.18}$$

Define the function inside the $\|$'s as

$$\phi(z) = T(z) - B_\gamma(z) f(z) \tag{1.19}$$

and note that $\phi(\gamma) = T(\gamma)$. Thus the problem of finding $f(z)$ reduces to the problem of finding the smallest $\phi(z)$ satisfying $\phi(\gamma) = T(\gamma)$:

$$\min_{\phi \in H^\infty} \|\phi(z)\|_\infty \text{s.t. } \phi(\gamma) = T(\gamma)$$

Theorem 1.1. *(Kasa, 2001) The solution to* (1.20) *is the constant function* $\phi(z) = T(\gamma)$.

Proof. To see this, first note that the norm of a constant function is the modulus of the constant itself. This is written as

$$\|\phi(z)\|_\infty = \|T(\gamma)\|_\infty = |T(\gamma)|^2. \tag{1.20}$$

Next, suppose that there exists another function $\Psi(z) \in H^\infty$, with $\Psi(\gamma) = T(\gamma)$ and also

$$\|\Psi(z)\|_\infty < \|\phi(z)\|_\infty. \tag{1.21}$$

Recall the definition of the L^∞ norm, and using equations (1.20) and (1.21):

$$\|\Psi(z)\|_\infty = \sup_{|z|=1} |\Psi(z)|^2 < |T(\gamma)|^2. \tag{1.22}$$

The maximum modulus theorem states that a function f which is analytic on the disk U achieves its maximum on the boundary of the disk. That is

$$\sup_{z\in U} |f(z)|^2 \leq \sup_{z\in \partial U} |f(z)|^2. \tag{1.23}$$

Therefore, we can see that

$$\sup_{|z|<1} |\Psi(z)|^2 \leq \sup_{|z|=1} |\Psi(z)|^2 < |T(\gamma)|^2. \tag{1.24}$$

However, one of the values on the interior of the unit disk is $z = \gamma$, which can be plugged in to the far LHS of equation (1.24) to get the result

$$|\Psi(\gamma)|^2 \leq \sup_{|z|=1} |\Psi(z)|^2 < |T(\gamma)|^2 \implies |\Psi(\gamma)|^2 < |T(\gamma)|^2. \tag{1.25}$$

This contradicts the requirement that $\Psi(\gamma) = T(\gamma)$. Therefore, we have verified that there does not exist another function $\Psi(z) \in H^\infty$ such that $\Psi(\gamma) = T(\gamma)$ and $\|\Psi(z)\|_\infty < \|\phi(z)\|_\infty$. □

Now that we have a form for $\phi(z)$, we can use it to find a formula for $f(z)$. Recalling the form of $f(z)$ and completing some tedious algebra, we obtain

$$f(z) = \frac{T(z) - \phi(z)}{B_\gamma(z)} = \frac{zq(z) - \gamma q(\gamma)}{z - \gamma} + \frac{\gamma^2}{1 - \gamma^2} q(\gamma)$$

which is the least-squares solution plus a constant. This means that after the initial period, the impulse response function for the robust predictor is identical to that of the least squares predictor. In the initial period, the least squares impulse response is $q(\gamma)$, while the robust impulse response is larger: $q(\gamma)/(1-\gamma^2)$. Recalling that γ is

the discount factor, and therefore close to unity, the robust impulse response can be considerably larger than that of the least squares response. Relatedly, the volatility of prices in the robust case will be larger as well. For example, in the first-order autoregressive case studied above,

$$p_t = f(L)\varepsilon_t = \frac{1}{1-\rho\gamma}d_t + \frac{\gamma^2}{(1-\gamma^2)(1-\rho\gamma)}\varepsilon_t \tag{1.26}$$

from which the variance can be calculated as

$$\sigma^2(p_t) = \left(\frac{1}{1-\rho\gamma}\right)^2 \sigma^2(d_t) + \frac{2\gamma^2-\gamma^4}{(1-\rho\gamma)^2(1-\gamma^2)^2}. \tag{1.27}$$

Using the data from Figure 1, and the values for $\sigma(d)$, r, ρ and γ that we have been using throughout this example, equation (1.27) gives us the result that $\sigma(p) = 89.52$.

The standard deviation of the actual price sequence in the SP dataset is 42.74. Thus when the agent is robust to the most misspecification possible, the resulting price volatility will have a standard deviation over twice as high as would be needed to exhibit the excess volatility seen in the data: *the robust present value model predicts prices that are substantially more volatile than those seen in the data.* The "robust puzzle" is therefore why prices are so *smooth*. The reversal of the volatility relationship is apparent in the figure below.

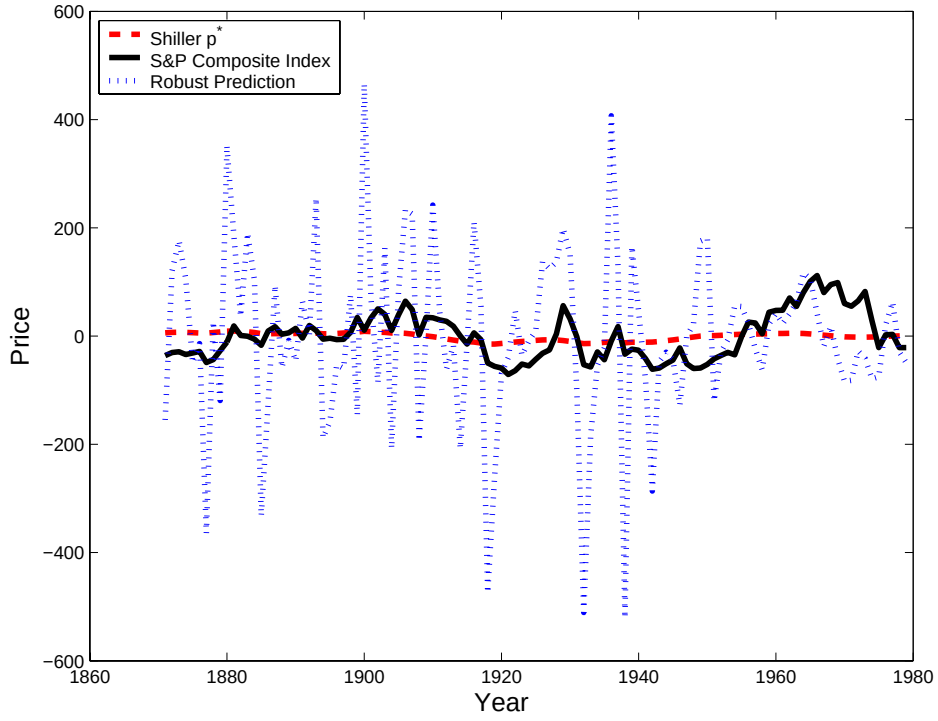


Figure 1.2: Robust Price Result

1.5 The Evil Agent Game

In this section we investigate an intermediate case in which the investor-predictor behaves robustly relative to a restricted set of possible models for dividends. In particular, following Hansen and Sargent (2005), we consider a a game played between the predicting agent and a malicious nature in which the degree of nature's malevolence is restricted. This restriction comes in the form of a cost associated with delivering excessively "noisy" dividends to the agent. The solution procedure, taken from Whiteman (1986), requires only slight modification from the development in sections 1.3 and 1.4.

1.5.1 The Investing Agent's Problem

Suppose that the dividend process investing agents perceive is written as

$$d_t = [q(L) + m(L)]\,\varepsilon_t \equiv C(L)\varepsilon_t.$$

Thus, the investor agent's problem becomes to choose an analytic function $f(z)$ in order to

$$\min_{f(z)} \frac{1}{2\pi i} \oint \left| \frac{C(z)}{1-\gamma z^{-1}} - f(z) \right|^2 \frac{dz}{z}. \tag{1.28}$$

As above, the optimization in equation (1.28) leads to the following Weiner-Hopf equation:

$$\frac{C(z)}{1-\gamma z^{-1}} - f(z) = \sum_{-\infty}^{-1}. \tag{1.29}$$

1.5.2 The Evil Agent's Problem

The Evil Agent's problem in this game is to make it as difficult as possible for the investing agent to make this prediction. However, there are restrictions on how much power the Evil Agent (EA) has to exercise. The EA has control over $m(z)$, which, as seen above, is a component of the dividend process that Investing Agents

(IA) take as given. The EA's problem is thus to add noise to maximize the investor's prediction error variance subject to a penalty of θ per unit of added variance.

$$\max_{C(z)} \frac{1}{2\pi i} \oint \left| \frac{C(z)}{1-\gamma z^{-1}} - f(z) \right|^2 - \theta \left| \frac{C(z)-q(z)}{1-\gamma z^{-1}} \right|^2 \frac{dz}{z} \tag{1.30}$$

1.5.3 Solving the Evil Agent Game

First, we find the Weiner-Hopf equation that results from the EA's optimization problem.

$$\begin{aligned} 0 = \frac{1}{2\pi i} \oint \Bigg\{ & \frac{z^j}{1-\gamma z^{-1}} \left[\frac{C(z^{-1})}{1-\gamma z} - f(z^{-1}) \right] - \frac{\theta z^j}{1-\gamma z^{-1}} \left[\frac{C(z^{-1})-q(z^{-1})}{1-\gamma z} \right] \\ & + \frac{z^{-j}}{1-\gamma z} \left[\frac{C(z)}{1-\gamma z^{-1}} - f(z) \right] - \frac{\theta z^{-j}}{1-\gamma z} \left[\frac{C(z)-q(z)}{1-\gamma z^{-1}} \right] \Bigg\} \frac{dz}{z}, \end{aligned}$$

which implies

$$\frac{(1-\theta)C(z)}{1-\gamma z} - \left[\frac{(1-\gamma z^{-1})}{1-\gamma z} f(z) \right] + \frac{\theta q(z)}{1-\gamma z} = \sum_{-\infty}^{-1}. \tag{1.31}$$

We recall from (1.29) that the Investor Agent's Weiner-Hopf equation can be written

$$C(z) = (1-\gamma z^{-1}) f(z) + \sum_{-\infty}^{-1}.$$

After applying the plussing operator, this leaves

$$C(z) = \left[(1-\gamma z^{-1}) f(z) \right]_+. \tag{1.32}$$

Using (1.32) in (1.31) and applying the plussing operator again, we have

$$\frac{1-\theta}{1-\gamma z}\left[(1-\gamma z^{-1})f(z)\right]_+ - \left[\frac{(1-\gamma z^{-1})}{1-\gamma z}f(z)\right]_+ + \frac{\theta q(z)}{1-\gamma z} = 0. \tag{1.33}$$

In order to proceed, we need to solve for the "plussed" elements of equation (1.33). The first "plussed" term is equivalent to

$$\begin{aligned}\left[(1-\gamma z^{-1})f(z)\right]_+ &= f(z) - \gamma\left[\frac{f(z)}{z}\right]_+ \\ &= f(z) - \gamma\left[\frac{f(z)-f_0}{z}\right] \\ &= (1-\gamma z^{-1})f(z) + \frac{\gamma f_0}{z}.\end{aligned} \tag{1.34}$$

Similarly,

$$\left[\frac{(1-\gamma z^{-1})}{1-\gamma z}f(z)\right]_+ = \frac{(1-\gamma z^{-1})}{1-\gamma z}f(z) + \frac{\gamma f_0}{z}. \tag{1.35}$$

Using (1.34) and (1.35) in (1.33), we obtain (via somewhat length algebraic manipulation):

$$f(z) = \frac{zq(z)-\gamma q(\gamma)}{z-\gamma} + \frac{\gamma^2}{\theta-\gamma^2}q(\gamma). \tag{1.36}$$

It is seen from (1.36) that $f(z)$ takes a recognizable form. The first term in $f(z)$ is the solution to the standard prediction problem in the event that the forecaster is attempting to minimize MSE, rather than playing the game against an Evil Agent.

Therefore, we can see that a forecaster using robust methods will end up using a function $f(z)$ that looks like an MSE forecast plus an extra term having to do with what he is trying to be robust against. Note that the value of θ will control just how large a part the second term in (1.36) will play in the forecast. Recall from (1.30) that θ is the Lagrange multiplier on the constraint the EA faces. Changing the value for θ is interpreted as loosening or tightening the constraint faced by the EA. As we increase θ, we make it more costly (in terms of the trade-off within his optimization) for the EA to add noise to the system. As we decrease θ, we make it less costly.

It turns out that there are two values for θ which make $f(z)$ immediately recognizable. The following relationship is clear from (1.36):

$$\begin{cases} \theta \to \infty & f(z) \to \text{MSE Result} \\ \theta \downarrow 1 & f(z) \to H^{\infty}\text{-norm Result} \end{cases}$$

A natural question at this point would be to ask why the worst-case scenario does not occur at $\theta = 0$, as that is the value for θ that people naturally associate with a non-binding constraint and therefore total freedom for the EA. The answer to this mystery lies in the saddle-point nature of this optimization problem. The second order conditions for finding a maximum are violated for values of $\theta \in (0, 1)$. Therefore, the lower limit on θ is equal to 1.

Now that a form for $f(z)$ has been uncovered, the next question involves the final form of $C(z)$, that is: what do dividends look like to the investor agent? Further,

we would like to know how much noise is being added by the EA in equilibrium. We pursue the form of $C(z)$ first by recalling equations (1.32) and (1.34), which gave us

$$C(z) = \left[(1-\gamma z^{-1})f(z)\right]_+ . \tag{1.37}$$

Substituting from (1.36) into (1.37) results in another "plussing" problem:

$$C(z) = \left[(1-\gamma z^{-1})\left\{\frac{q(z)-\gamma z^{-1}q(\gamma)}{1-\gamma z^{-1}}\right\} + (1-\gamma z^{-1})\left\{\frac{\gamma^2}{\theta-\gamma^2}q(\gamma)\right\}\right]_+$$

which when solved using the methods shown above yields

$$C(z) = q(z) + \frac{\gamma^2}{\theta-\gamma^2}q(\gamma). \tag{1.38}$$

1.5.4 Additional Noise from the Evil Agent

Now that we have the form of $C(z)$, we can use the Cauchy Integral formula to solve for the present value of the noise added in by the Evil Agent. We will denote the present value of the noise by η.

$$\begin{aligned}
\eta &= \frac{1}{2\pi i}\oint \left|\frac{C(z)-q(z)}{1-\gamma z^{-1}}\right|^2 \frac{dz}{z} \\
&= \frac{\gamma^4 q^2(\gamma)}{(\theta-\gamma^2)^2}\frac{1}{2\pi i}\oint \frac{dz}{(1-\gamma z^{-1})(1-\gamma z)z} \\
&= \frac{\gamma^4 q^2(\gamma)}{(\theta-\gamma^2)^2(1-\gamma^2)}
\end{aligned}$$

1.6 Empirical Results

Do parameterizations of the Evil Agent game do a good job of fitting the data? We begin with the simplest possible univariate setup and then proceed to more realistic and complex environments. In the univariate case, we use the dividend process that was used in Shiller (1981). Thus

$$q(L) = \frac{1}{1 - \rho L} \tag{1.39}$$

where $\rho = 0.95$. The analytical results allow us to use any dividend process; this particular choice reflects our desire to keep the initial model comparison between the results of Shiller (1981) and this work as easy as possible.

1.6.1 Univariate θ Estimation

As a first pass, we take the function for price prediction and ask, "given the data, what value of θ would create price volatility in our model equal to that in the data?" This is done assuming the dividends were generated by the AR(1) process in Shiller (1981); the required $\theta = 1.62$. To put this into context, we examine what this implies the investor must be thinking about dividends.

1.6.2 Detectability of θ

In the EA game, the final dividend process being targeted by the IA is expressed by equation (1.38). This means that the variance calculation of the resulting dividend process can be written:

$$
\begin{aligned}
var\left[C(L)\varepsilon_t\right] &= \sigma_\varepsilon^2 \left\{ \left(1 + \frac{\gamma^2}{\theta - \gamma^2} q(\gamma)\right)^2 + c_1^2 + c_2^2 + c_3^2 + \dots \right\} \\
&= \sigma_\varepsilon^2 \left\{ \left(1 + \frac{\gamma^2}{\theta - \gamma^2} q(\gamma)\right)^2 + \rho^2 + \rho^4 + \rho^6 + \dots \right\} \\
&= \sigma_\varepsilon^2 \left\{ \left(1 + \frac{\gamma^2}{\theta - \gamma^2} q(\gamma)\right)^2 + \frac{\rho^2}{1 - \rho^2} \right\} \qquad (1.40)
\end{aligned}
$$

Now, compare the result in equation (1.40) with the variance of dividends if they were produced by the AR(1) process given above:

$$
var[q(L)\varepsilon_t] = \frac{\sigma_\varepsilon^2}{1 - \rho^2}. \qquad (1.41)
$$

To compare the two, we calculate

$$
\begin{aligned}
\frac{var[C(L)\varepsilon_t]}{var[q(L)\varepsilon_t]} &= \frac{\sigma_\varepsilon^2 \left\{ \left(1 + \frac{\gamma^2}{\theta - \gamma^2} q(\gamma)\right)^2 + \frac{\rho^2}{1 - \rho^2} \right\}}{\frac{\sigma_\varepsilon}{1 - \rho^2}} \\
&= (1 - \rho^2) \left\{ \left(1 + \frac{\gamma^2}{\theta - \gamma^2} q(\gamma)\right)^2 + \frac{\rho^2}{1 - \rho^2} \right\} \\
&= (1 - \rho^2) \left(1 + \frac{\gamma^2}{\theta - \gamma^2} q(\gamma)\right)^2 + \rho^2 \\
&= 606.55,
\end{aligned}
$$

which indicates that the investing agent is guarding against a dividend process that is *quite* different from what has been observed. The problem with this comparison is that it does not permit compromise: is there a θ that gets price variability "close"

without making dividends too variable? To study this, we will need to estimate a system which simultaneously fits both the price and dividend processes.

1.6.3 Bivariate Estimation

We begin by specifying a general bivariate moving average representation for d_t and p_t.:

$$\begin{pmatrix} d_t \\ p_t \end{pmatrix} = \begin{pmatrix} A(L) & B(L) \\ C(L) & D(L) \end{pmatrix} \begin{pmatrix} \varepsilon_{d_t} \\ \varepsilon_{p_t} \end{pmatrix}.$$

In order to accommodate the cross-equation restrictions implied by the present value relationship, we identify the system by restricting the innovations to be uncorrelated and unit variance Guassian processes. It is possible to describe both the least-squares prediction problem and the Evil Agent game within this structure. In fact, the calculations of the previous sections are now applied column by column, so that the elements of $C(L)$ are functions of the elements of $A(L)$, and $D(L)$ is a function of $B(L)$. We choose to specify dividends as the sum of a persistent component (like Shiller's AR(1) process) and a transitory component. Then the restricted moving average representation in the least squares case is

$$\begin{pmatrix} A_{LS}(L) & B_{LS}(L) \\ C_{LS}(L) & D_{LS}(L) \end{pmatrix} = \begin{pmatrix} A_0 & \frac{B_0}{1-bL} \\ \frac{LA(L)-\gamma A(L)}{L-\gamma} & \frac{LB(L)-\gamma B(\gamma)}{L-\gamma} \end{pmatrix} = \begin{pmatrix} A_0 & \frac{B_0}{1-bL} \\ A_0 & \frac{B_0}{1-b\gamma}\left(\frac{1}{1-bL}\right) \end{pmatrix}$$

while in the Evil Agent game the MA is

$$\begin{pmatrix} A_{EA}(L) & B_{EA}(L) \\ C_{EA}(L) & D_{EA}(L) \end{pmatrix} = \begin{pmatrix} A_0 + \frac{\gamma^2}{\theta-\gamma^2}A_0 & \frac{B_0}{1-bL} + \frac{\gamma^2}{\theta-\gamma^2}\left(\frac{B_0}{1-b\gamma}\right) \\ \frac{LA(L)-\gamma A(L)}{L-\gamma} + \frac{\gamma^2}{\theta-\gamma^2}A(\gamma) & \frac{LB(L)-\gamma B(\gamma)}{L-\gamma} + \frac{\gamma^2}{\theta-\gamma^2}B(\gamma) \end{pmatrix}$$
$$= \begin{pmatrix} \frac{A_0\theta}{\theta-\gamma^2} & \frac{B_0}{1-bL} + \frac{\gamma^2}{\theta-\gamma^2}\left(\frac{B_0}{1-b\gamma}\right) \\ \frac{A_0\theta^2}{(\theta-\gamma^2)^2} & \frac{B_0\theta-\gamma^2 B_0 bL}{(1-bL)(\theta-\gamma^2)(1-b\gamma)} \end{pmatrix}.$$

1.6.4 State Space Formulations

By shifting into the state space, we can make use of the powerful set of tools available via Kalman filtering. This requires that each system be written in terms of a state and observer system. In order to estimate the system, we construct a state-space formulation of each of these vector MA systems which will allow us to use the Kalman filter to perform maximum likelihood estimation. The estimation will find parameter values for A_0, B_0 and b in the least-squares system, and A_0, B_0, b and θ in the Evil Agent game system.

1.6.4.1 Least-Squares State Space

The natural state of the system is the persistent component of dividends:

$$s_t = bs_{t-1} + \varepsilon_{p_t} \tag{1.42}$$

$$\begin{pmatrix} d_t \\ p_t \end{pmatrix} = \begin{pmatrix} B_0 \\ \frac{B_0}{1-b\gamma} \end{pmatrix} s_t + \begin{pmatrix} A_0 & 0 \\ A_0 & 0 \end{pmatrix} \begin{pmatrix} \varepsilon_{d_t} \\ \varepsilon_{p_t} \end{pmatrix} \tag{1.43}$$

where equation (1.42) is the state equation and (1.43) is the observer equation. With this forumlation, the Kalman filter can be used to evaluate the likelihood using a standard procedure (see, e.g., Hamilton (1994)). We find the following estimates of the parameters:

$$A_0 = 1.9894\,, \quad B_0 = 1.2478\,, \quad b = 0.9998.$$

The log-likelihood of this model is -718.48. This will provide a benchmark by which to evaluate the Evil Agent game model.

1.6.4.2 Evil Agent State Space

The state space for the Evil Agent game system can be written in the following way:

$$s_t = bs_{t-1} + \varepsilon_{p_t}$$

$$\begin{pmatrix} d_t \\ p_t \end{pmatrix} = \begin{pmatrix} B_0 \\ \frac{B_0}{1-b\gamma} \end{pmatrix} s_t + \begin{pmatrix} \frac{A_0\theta}{\theta-\gamma^2} & \frac{\gamma^2 B_0}{(\theta-\gamma^2)(1-b\gamma)} \\ \frac{A_0\theta^2}{(\theta-\gamma^2)^2} & \frac{\gamma^2 B_0}{(\theta-\gamma^2)(1-b\gamma)} \end{pmatrix} \begin{pmatrix} \varepsilon_{d_t} \\ \varepsilon_{p_t} \end{pmatrix}.$$

This system is estimated in exactly the same way as the least-squares system, thus demonstrating the flexibility of state space methods for problems such as these. The

estimation of the system results in the following maximum likelihood estimates:

$$A_0 = 1.1693 * 10^{-5}\,, \quad B_0 = 1.1922\,, \quad b = 0.9998\,, \quad \theta = 10.3037.$$

The log-likelihood of this model is -712.03. The improvement is significant over the least squares model. Furthermore, when compared to the univariate estimate of θ, we that these parameter values do not make dividends too variable. Because the estimated A_0 is so small, both dividends and prices are dominated by the persistent component, and thus the relevant ratio in this case is

$$\frac{var[B_{EA}(L)\varepsilon_t]}{var[B_{LS}(L)\varepsilon_t]} = (1-\rho^2)\left(1 + \frac{\gamma^2}{\theta - \gamma^2}q(\gamma)\right)^2 + \rho^2 = 1.2375.$$

Unlike the univariate case, this value for θ is very promising in terms of generating prices and dividends which are simultaneously "close" to those in the data. The dividends in the model would be roughly only 12% more volatile than those in the data.

The problem, however, lies not in the volatility ratio, but in the overall fit of the model. While the evil agent model beats the least squares model, neither fares very well against a slightly less restrictive model. In particular, the system

$$s_t = bs_{t-1} + \varepsilon_{p_t} \tag{1.44}$$

$$\begin{pmatrix} d_t \\ p_t \end{pmatrix} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} s_t + \begin{pmatrix} B_3 & 0 \\ B_4 & 0 \end{pmatrix} \begin{pmatrix} \varepsilon_{d_t} \\ \varepsilon_{p_t} \end{pmatrix} \tag{1.45}$$

achieves a value of the log likelihood about 100 higher than the evil agent system, indicating that there is a long way to go in fitting d_t and p_t jointly.

1.6.5 More Sophisticated Data-Generating Processes

The answer to this challenge lies in the use of more sophisticated processes for dividends and prices. Consider the following bivariate ARMA(3,1) process.

$$\begin{pmatrix} d_t \\ p_t \end{pmatrix} = \begin{pmatrix} A(L) & B(L) \\ C(L) & D(L) \end{pmatrix} \begin{pmatrix} \varepsilon_{d_t} \\ \varepsilon_{p_t} \end{pmatrix}$$

where

$$A(L) = \frac{\rho_0}{(1-\rho_3 L)(1-\rho_4 L)}\,, \quad B(L) = \frac{\mu_0(1-\mu_1 L)}{(1-L)(1-\mu_3 L)(1-\mu_4 L)}. \tag{1.46}$$

This results in the following processes for $C(L)$ and $D(L)$:

$$C(L) = \frac{LA(L) - \gamma A(\gamma)}{L-\gamma} + \frac{\gamma^2}{\theta - \gamma^2} A(\gamma)\,, \quad D(L) = \frac{LB(L) - \gamma B(\gamma)}{L-\gamma} + \frac{\gamma^2}{\theta - \gamma^2} B(\gamma). \tag{1.47}$$

We were led to this specification of $A(L)$ and $B(L)$ by an exploration of the likelihood prompted by difficulties in estimating the model with general ARMA(3,1) specifications for $A(L)$ and $B(L)$. We suspect that these difficulties were caused by near cancellations of roots of the numerator and denominator polynomials of our specified ARMAs, together with the presence of a highly persistent autoregressive component. Both maximum likelihood estimation and exploration of a diffuse-prior Bayesian posterior by Markov Chain Monte Carlo methods were much better behaved with the more parsimonious specification.

To determine the relevant unrestricted alternative to (1.46) and (1.47), note that for the given $A(L)$ and $B(L)$, the cross-equation restrictions of the evil agent setup cause $C(L)$ to be ARMA(2,2) and $D(L)$ to be ARMA(3,3). Thus the unrestricted model specification has

$$A(L) = \frac{\alpha_0}{(1-\alpha_1 L)(1-\alpha_2 L)}, \quad B(L) = \frac{\beta_0(1-\beta_1 L)}{(1-L)(1-\beta_3 L)(1-\beta_4 L)}. \tag{1.48}$$

$$C(L) = \frac{\chi_0(1+\chi_1 L)(1+\chi_2 L)}{(1-\chi_3 L)(1-\chi_4 L)}, \quad D(L) = \frac{\delta_0(1-\delta_1 L)(1-\delta_2 L)(1-\delta_3 L)}{(1-L)(1-\delta_5 L)(1-\delta_6 L)}. \tag{1.49}$$

It is helpful to note the way in which the models (the game model, the unrestricted model, and the least-squares model) are nested. The least-squares model

Parameter	**Estimate**	**Standard Error**
ρ_0	0.6694	0.046
ρ_3	0.9181	0.046
ρ_4	0.1676	0.104
μ_0	0.1667	0.035
μ_1	-0.4785	0.033
μ_3	-0.0390	0.064
μ_4	0.9784	0.016

Table 1.1: Least-Squares Model Parameter Estimates. Log-Likelihood Value: -589.63

is nested within the game model, by placing a restriction on one parameter: the LS model restricts θ to be ∞. The unrestricted model nests the game model – the cross equation restrictions represent specific restrictions on the values of the χ's and δ's in $C(L)$ and $D(L)$. Because of these nesting relationships, comparison between models can be accomplished with a simple likelihood ratio test.

1.6.6 The Model Estimates

1.6.6.1 The Least-Squares Model

The estimation of the least-squares model is summarized below.

1.6.6.2 The Game Model

The estimation of the game model is described below.

Parameter	**Estimate**	**Standard Error**
ρ_0	0.5844	0.039
ρ_3	0.6561	0.112
ρ_4	0.1505	0.162
μ_0	0.2704	0.063
μ_1	-0.6767	0.038
μ_3	0.5249	0.157
μ_4	0.7639	0.122
θ	2.1645	0.135

Table 1.2: Game Model Parameter Estimates. Log-Likelihood Value: -577.77

As can be seen by looking at the estimates of the maximum likelihood of both models, the likelihood ratio test produces a rejection of the "restricted" model, the LS model, in favor of the game model. This rejection is at the 99% level.

1.6.6.3 The Unrestricted Time-Series Model

The estimation of the unrestricted model is described below.

Due to the fact that the unrestricted model has a total of ten fewer restrictions than the game model, the likelihood ratio test critical value (at 95%) is approximately 18.3. The test statistic in this case is 16.9, well inside the region in which we fail to reject the more restrictive game model in favor of the pure time-series model listed

Parameter	**Estimate**	**Standard Error**
α_0	0.5924	0.040
α_1	0.5862	0.225
α_2	0.3091	0.262
β_0	0.1448	0.059
β_1	0.9988	0.981
β_3	0.0451	0.364
β_4	0.0425	0.410
χ_0	4.9994	1.001
χ_1	0.1782	0.077
χ_2	0.9999	0.996
χ_3	-0.4469	0.090
χ_4	0.5127	0.044
δ_0	19.0235	1.042
δ_1	0.6666	0.064
δ_2	-0.6250	0.216
δ_3	-0.6227	0.200
δ_5	-0.2505	0.025
δ_6	0.8350	0.084

Table 1.3: The Unrestricted Model Parameter Estimates. Log-Likelihood Value: -569.32

above.[6]

1.6.6.4 Analysis of Results

The results show that we reject the LS model in favor of the Game model. This is significant, but not totally unexpected, given that we use an additional parameter. However, we later see that untying the rational expectation cross-equation restrictions, creating the unrestricted model — giving the flexibility of *ten* additional free parameters — this much less restrictive framework generates less of a gain over the Game model than the Game model achieved over the LS model using a single additional parameter.

1.7 Conclusion

In most modern economic models, agents deal in risk rather than uncertainty. In reality, economic decision-makers are forced to account for both. This paper has placed the agents in the model on the same footing as the authors of the model: the real world contains data generating processes (DGP) for which we have estimates, but not certainties. Through the mechanism of robust prediction and control, agents deal with this uncertainty by making decisions that are less sensitive to misspecifications of the DGP.

For the present value model of stock prices, the application of robust decision-

[6]The test statistic with the updated data is 19.0, which has a marginal significance level of 0.0401 rather than the original dataset's marginal significance of approximately 0.0766. Tests for marginal significance levels are under way to determine if the change in p-level is attributable to a change in the performance of the model, or just to the fact that the model is being examined with 30% more data.

making yields a model whose behavior more closely mimics that of the actual data. With robust predictions, the simple present-value model produces stock prices that display the "excess volatility" seen in the data. Thus not only is uncertainty regarding the DGP realistic, it also suggests the resolution of a long-standing economic puzzle in a plausible manner. The resolution of the excess volatility puzzle by such a simple modification in such a simple model suggests that the modification might bear fruit in other, more complex settings.

CHAPTER 2
THE TWO-PERIOD RATIONAL INATTENTION MODEL: ACCELERATIONS, ADDITIONS AND ANALYSES

2.1 Inattention and the Two-Period Model

The Rational Inattention (RI) paradigm introduced in Sims (2003) initiated analysis of economic models of agents with limited information-processing capacity who have quadratic objectives and face linear constraints. Sims (2006) demonstrated the specific approach to be taken in pursuing these types of models in more general environments than those of the earlier work. This paper demonstrates that there is a formulation of Sims (2006) problem that renders it into a convex programming problem, demonstrates a powerful solution tool for these more general RI models, and provides analysis of why using these procedures can be important.

Sims (2006) provides 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 variable and chooses the joint distribution of state and choice variables. This paper addresses solution methods for these types of problems and demonstrates that RI need not impose large computational burdens. We will see that a new software suite, combined with a well-posed version of the model, can generate solutions in seconds rather than minutes, opening up new horizons for RI problems. The solutions we derive for the two-period problem

are similar but not identical to those found in Sims (2006), and we will address the reasons for these differences. Despite the difference, Sims' central conclusions are robust: 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 new computational approaches are generalizable to richer environments; one generalization undertaken here is the addition of production to Sims' model; his central themes are preserved in this context as well.

2.2 The Two-Period Model

The two-period model of Sims (2006) highlights the central difference between rational inattention (RI) 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.[1]

To begin, absent information-processing constraints, Sims' model is a simple 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

[1]In an effort to make each essay self-contained, much of this discussion is repeated in Lewis (2007b).

it optimally between consuming c in period one and $w - c$ in period two. That is, for CRRA preferences,

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

for a given w. 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 if 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.

2.2.1 A Generalization

To set the stage for the information-constrained problem to come, 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 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. To clarify what is being solved, the discretized version of the two period model is written:

$$\max_{\{f(c_i,w_j)\}} \sum_{i=1}^{N_c} \sum_{j=1}^{N_w} \frac{c_i^{1-\gamma} + (w_j - c_i)^{1-\gamma}}{1-\gamma} f(c_i, w_j) \tag{2.1}$$

subject to:

$$f(c_i, w_j) \geq 0 \tag{2.2}$$

$$\sum_{i=1}^{N_c} f(c_i, w_j) = g(w_j) \quad \text{for } j = 1, \ldots, N_w \tag{2.3}$$

$$f(c_i, w_j) = 0\,, \quad \forall (i,j) \text{ such that } c_i > w_j \tag{2.4}$$

Regarding the formulae above, the domains of the consumption and wealth distributions have been discretized into two sets of size N_c and N_w, respectively, where i is the index over c and j is the index over w (this convention will be maintained throughout). $f(c, w)$ is the joint distribution of consumption and wealth, and is the choice variable of this optimization problem. $g(w)$ 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. Suppose that the marginal distribution of wealth is triangular, meaning higher levels of wealth have higher probability. The optimal decision rule is the joint distribution $f(c, w)$ that describes the same one-to-one mapping that divides wealth into two halves and consumes one in each time period. 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 $\{c_i\}_{i=1}^{N_c}$ such that weight is only allowed where $c_i \leq w_j$. The optimal choice, as seen in figure 3.3, 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, $f(c_i = w_j/2, w_j) = g(w_j)$.

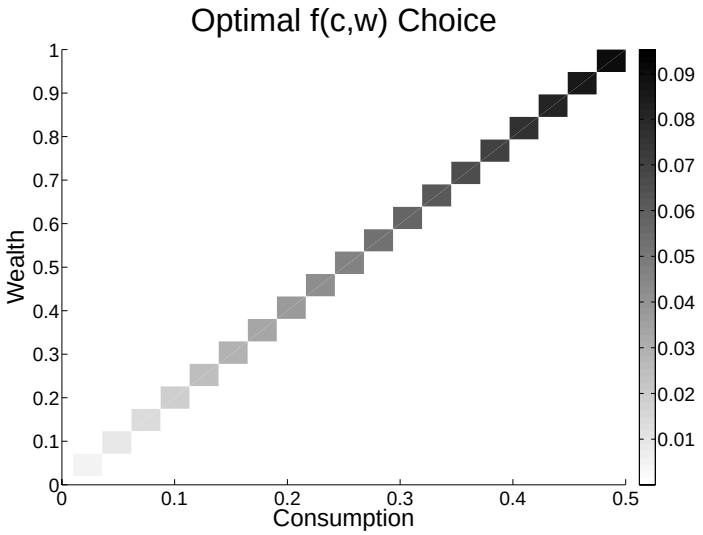


Figure 2.1: A One-to-One Mapping via the Joint Distribution $f(c, w)$.

Figure 3.3 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 weight, as indicated by the legend on the right-hand side of the figure. The boxes get darker as they progress "northwest" 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 $c_i = w_j/2$.

2.2.2 The Processing-Constrained Problem

The rational inattention framework uses the metric of mutual information to quantify the amount of information-processing capacity the agent is using to solve his optimization problem. By placing a constraint on mutual information, the framework limits the strength of the relationship between c and w by 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 3.3, 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 (2.1) through (2.4) above, with the addition of the following constraint on the amount of mutual information in the model:

$$\sum_{j=1}^{N_w}\sum_{i=1}^{N_c}\log[f_t(c_i,w_j)]\cdot f_t(c_i,w_j)-\sum_{i=1}^{N_c}\left(\log\left(\sum_{j=1}^{N_w}f_t(c_i,w_j)\right)\cdot\sum_{j=1}^{N_w}f_t(c_i,w_j)\right) \quad (2.5)$$
$$-\sum_{j=1}^{N_w}\log(g_t(w_j))\cdot g_t(w_j)\leq\kappa.$$

As the amount of information-processing capacity (κ) decreases, the effect on the agent is similar to that of increasing the noise in a signal-extraction version of the same problem.[2] 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 state of the model. 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 asses this state of the model because he does not have the information-processing resources to do so.

2.3 A Convex Problem

RI models represent a potentially large burden on numerical optimizers. 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 far from numerically intractable. We show here that, contrary to what is stated in Sims (2006), the constraint set and therefore the problem as a whole *is* convex.

[2]Mutual Information, as defined in equation (2.5), 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 our models are known as "nats", because we us natural logarithms.

The choice in the problem is $f(c_i, w_j)$ (from now on: $f_{i,j}$). Note that the nodes for consumption and wealth are fixed and it is the probabilities $f_{i,j}$ that are chosen. Thus, the objective function is a weighted sum and linear. Constraints (2.2), (2.3) and (2.4) are linear. In order for the problem to be convex, we must demonstrate:

Theorem 2.1. *For $f_{i,j} > 0$,*

$$\begin{aligned} MI(f_{i,j}) = &\sum_{i=1}^{N_c}\sum_{j=1}^{N_w} f_{i,j}\cdot\log(f_{i,j}) \\ &-\sum_{i=1}^{N_c}\left\{\left[\sum_{j=1} f_{i,j}\right]\cdot\left[\log\left(\sum_{j=1} f_{i,j}\right)\right]\right\} \\ &-\sum_{j=1}^{N_w} g(w_j)\cdot\log(g(w_j)) \leq \kappa. \end{aligned} \tag{2.6}$$

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

Proof. Because $g(w)$ is fixed, we can limit our attention to demonstrating that

$$MI'(f_{i,j}) = \sum_{i=1}^{N_c}\sum_{j=1}^{N_w} f_{i,j}\log(f_{i,j}) - \sum_{i=1}^{N_c}\left\{\left[\sum_{j=1}^{N_w} f_{i,j}\right]\cdot\left[\log\left(\sum_{j=1}^{N_w} f_{i,j}\right)\right]\right\} \tag{2.7}$$

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

$$MI'(f_{i,j}) = \sum_{i=1}^{N_c}\left\{\sum_{j=1}^{N_w} f_{i,j}\log\left(f_{i,j}\right) - \left[\sum_{j=1}^{N_w} f_{i,j}\right]\cdot\left[\log\left(\sum_{j=1}^{N_w} f_{i,j}\right)\right]\right\} \tag{2.8}$$

The outermost summation in equation (2.8) is the only summation over the i index, which means that we consider only the piece inside that sum, because if it is convex,

the summation over i's will be the sum of convex functions, which is also convex. Now, the goal becomes to prove that

$$MI''(f_{i,j}) = \sum_{j=1}^{N_w} f_{i,j} \log(f_{i,j}) - \left[\sum_{j=1}^{N_w} f_{i,j}\right] \cdot \left[\log\left(\sum_{j=1}^{N_w} f_{i,j}\right)\right] \tag{2.9}$$

is convex for a given i. To this end, we clean the notation up with the following substitutions:

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

Also, we define

$$e = [1, 1, \ldots, 1]'\,, \qquad E = ee' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ 1 & \cdots & \cdots & 1 \end{bmatrix}$$

where e and E are $N \times 1$ and $N \times N$, respectively. With this new notation, the problem reduces to demonstrating that

$$h(x) = x^T \log(x) - (e^T x) \log(e^T x) \tag{2.10}$$

is convex. To this end, we will show that the hessian of h is positive semi-definite.

$$\nabla h(x) = \begin{bmatrix} \log(x_1) \\ \log(x_2) \\ \vdots \\ \log(x_N) \end{bmatrix} - \begin{bmatrix} \log(e^T x) \\ \log(e^T x) \\ \vdots \\ \log(e^T x) \end{bmatrix} \tag{2.11}$$

$$\nabla^2 h(x) = \begin{bmatrix} \frac{1}{x_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{x_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{x_N} \end{bmatrix} - \begin{bmatrix} \frac{1}{e^T x} \cdot 1 & \frac{1}{e^T x} \cdot 1 & \cdots & \frac{1}{e^T x} \cdot 1 \\ \frac{1}{e^T x} \cdot 1 & \frac{1}{e^T x} \cdot 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{1}{e^T x} \cdot 1 & \cdots & \cdots & \frac{1}{e^T x} \cdot 1 \end{bmatrix} = X^{-1} - \frac{1}{e^T x} E \tag{2.12}$$

For the hessian to be positive semi-definite, we need to show that, for all non-zero $y \in \mathbb{R}^N$,

$$y^T \nabla^2 h(x) y \geq 0. \tag{2.13}$$

Some preliminary algebra yields:

$$y^T \nabla^2 h(x) y = y^T \left(X^{-1} - \frac{1}{e^T x} E \right) y = y^T X^{-1} y - \frac{1}{e^T x} y^T E y. \tag{2.14}$$

Breaking this into two pieces, we address the rightmost element first:

$$y^T E y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 1 & \cdots & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = (e^T y)^2 \tag{2.15}$$

Therefore,

$$\frac{1}{e^T x} y^T E y = \frac{(e^T y)^2}{e^T x}. \tag{2.16}$$

The remaining part of the equation can be simplified to:

$$y^T X^{-1} y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix} \begin{bmatrix} \frac{1}{x_1} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{x_2} & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \frac{1}{x_N} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_N \end{bmatrix} = \sum_{j=1}^{N} \frac{y_j^2}{x_j}. \tag{2.17}$$

Therefore, we need to show that

$$\sum_{j=1}^{N} \frac{y_j^2}{x_j} \geq \frac{(e^T y)^2}{e^T x}. \tag{2.18}$$

In order to demonstrate (2.18), we will make two additional assumptions:

1. 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 $e^T y$ while leaving y_j^2 unchanged.

Implicitly, we are making use of 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.

2. Assume, without loss of generality, $\sum y_j = e^T y = 1$.

   - 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 $e^T y = 0$ is impossible when $y_i \geq 0$ and $y \neq \mathbf{0}$.

Before taking advantage of the two assumptions, note that

$$\sum_{j=1}^{N} \frac{y_j^2}{x_j} = \sum_{j=1}^{N} y_j \left(\frac{y_j}{x_j}\right) = \sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j}\right)^{-1}. \tag{2.19}$$

Therefore, we need to show:

$$\sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j}\right)^{-1} \geq \frac{(e^T y)^2}{e^T x}. \tag{2.20}$$

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

$$\sum_{j=1}^{N} y_j f(q_j) \geq f\left(\sum_{j=1}^{N} y_j q_j\right) \tag{2.21}$$

$$\implies \sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j}\right)^{-1} \geq \left[\sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j}\right)\right]^{-1} \tag{2.22}$$

$$= \left(\sum_{j=1}^{N} x_j\right)^{-1} \tag{2.23}$$

$$= \frac{1}{e^T x} \tag{2.24}$$

$$= \frac{(e^T y)^2}{e^T x}. \tag{2.25}$$

Meaning

$$\sum_{j=1}^{N} y_j \left(\frac{x_j}{y_j}\right)^{-1} = \sum_{j=1}^{N} \frac{y_j^2}{x_j} \geq \frac{(e^T y)^2}{e^T x}. \tag{2.26}$$

Therefore, $\nabla^2 h(x)$ is positive semi-definite. This means that equation (2.9) is convex for a given i, and thus that the sum over i in equation (2.8) is the sum over convex functions making equation (2.7) convex, meaning that the mutual information constraint, equation (2.6), is convex for $f_{i,j} > 0$. □

The problem specified in equations (2.1) through (2.5) 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 restricted to be zero by the feasibility constraint [equation (2.4)], thus it is important that we consider $f_{i,j} = 0$. What we have demonstrated is that the MI constraint is convex on the interior of the feasible set. We know, however, that because $\lim_{p \to 0} p \log(p) = 0$, that the MI constraint is continuous on the whole feasible set

$f_{i,j} \in [0,1]$, including the boundary. 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 (2.1) through (2.5) is a convex programming problem.

2.4 The Solution Procedure

Because we know that the problem is convex, any local optimum will be a global optimum. The numerical optimizing method we implement is *very* similar to the one proposed in Sims (2006), but makes use of a sophisticated software package, `AMPL` (literally: `A` Mathematical `P`rogramming `L`anguage), that deserves recognition within this branch of the economics literature. `AMPL` is the ideal front-end for a powerful numerical optimizer, and the clarity of coding afforded by `AMPL` allows easy access to complicated problems, making RI models an excellent forum in which to demonstrate the software.

Some features of the software suite should be illustrated. First, `AMPL` is not an optimizer in itself, but rather a front end for a large number of potential optimization algorithms, each of which has properties suiting it to a specific set of problems. `AMPL` has several features that bolster its capabilities as an optimization front-end. First, `AMPL` is an intuitive language for economists: the problems are written down almost exactly as they appear in the literature. Second, `AMPL` has a function called `presolve` that examines the problem and eliminates variables based on constraints and makes calculations that can simplify the problem. Third, and most important, is the way

that AMPL deals with differentiation.

The key to the efficacy of a large class of numerical optimization scheme lies in derivatives, meaning that gradients and hessians must be supplied. In AMPL, 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) provides the gradients without truncation errors of a procedure like divided differencing or the excessive memory usage (storage and retrieval) 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.[3]

While AMPL's features allow the problem to be stated intuitively, pre-solved, and differentiated; the object of the exercise is to find the optimum. The optimizer used for this model is called KNITRO. KNITRO implements an interior point optimization algorithm that makes it exceptionally well suited to the problem we are working on here, and to RI models in general. Interior point methods approach the boundaries of variable-space in an organized way, without taking derivatives *at* the boundaries. This is important because the derivatives of this problem are not defined at the boundaries.

The combination of AMPL and KNITRO will guarantee that, to the tolerances set

[3]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).

by the user, a local optimum is found. The convexity of the problem, demonstrated above guarantees that the optimum found by the optimizer will be a global optimum. 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 size suggested in Sims (2006) on a 3 GHz Pentium 4 machine with 4 GB of RAM.[4]

2.5 Computational Issues

Three important differences exist between what has been done here using numerical optimization and what was done in Sims (2006): First, Sims uses a normalization to eliminate (2.3), while I leave it explicit. Second, rather than pick a value for λ and maximize the LaGrangian for a given multiplier value, I choose a capacity value, κ, and leave the constraint intact. `AMPL` deals directly with the constraints and takes derivatives through the LaGrangian automatically, allowing it to use the fastest optimization methods. Third, I optimize directly over the values of f rather than their logarithm.

The third difference is one that seems minor, but is potentially the source of the difference between the optimization results presented here and the ones in Sims (2006). The solution procedure presented in this paper optimizes directly over the values of f rather than their logarithm, while Sims uses the log-transform. His reason for this is that since logarithms are undefined at zero, we can use $\log(f)$

[4] A quasi-analytical solution to this model, demonstrated in appendix A, is also possible, relatively fast, and achieves the same solutions as the method using `AMPL`.

to make sure that the problem stays in the region of $f > 0$ values which are well behaved (in terms of the gradient). When $\log(f)$ is very large and negative, we will take that to be zero. The problem with the log-transform is that it vastly increases the difficulty of the numerical optimization problem. Examination of the original problem in the previous section illustrates the well-posed nature of the problem from a numerical optimization standpoint. Log-transformation of the problem gives us a non-linear, concave objective to maximize and a non-convex constraint set. While the optimization problem is the same theoretically, it has become much harder for numerical optimizers to solve. This is likely partially responsible for the difference in computation times. The transformation is not the whole story, though, as we use automatic derivatives from `AMPL` which will certainly increase the speed of the optimizer as well.

2.6 Results

The results are qualitatively the same as Sims (2006). First, note that the algorithm described above using `AMPL` allows us to set κ to values that no longer bind the information-processing constraint. We see in figure 2.2 the progressive tightening of the capacity constraint and its effect on the choice of the consumption-wealth joint distribution. 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 "tight" or "loose" depend on the size of the grid and the "complexity" of the wealth distribution. The value $\kappa = 4$, in the case of figure 2.2, 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 3.3. This means that the constraint is ineffective for values of $\kappa > 4$. Four nats of information processing may seem comically 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 2.2 increases rapidly.[5]

Next, we make the comparison that Sims makes regarding risk aversion. The differences between figure 2.3 below and its counterparts in Sims (2006) are curious, but in the final analysis, minimal. The fact that both solution procedures outlined about match in results gives reassurance that they both work, while further analysis shows that the underlying message of section IV in Sims (2006) remains intact. 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

[5]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 2.2, $\kappa = 4$, 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.

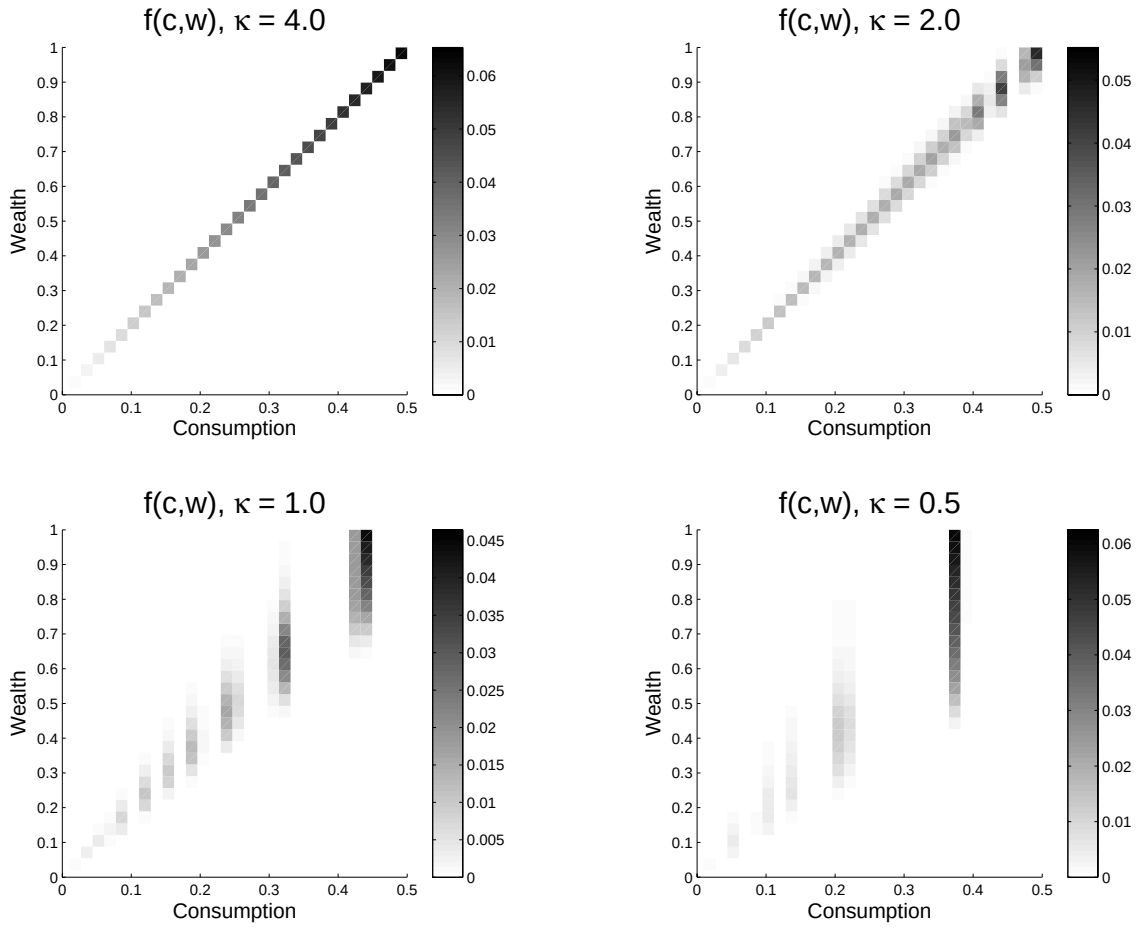


Figure 2.2: Comparison of Different Levels of Information-Processing Capacity

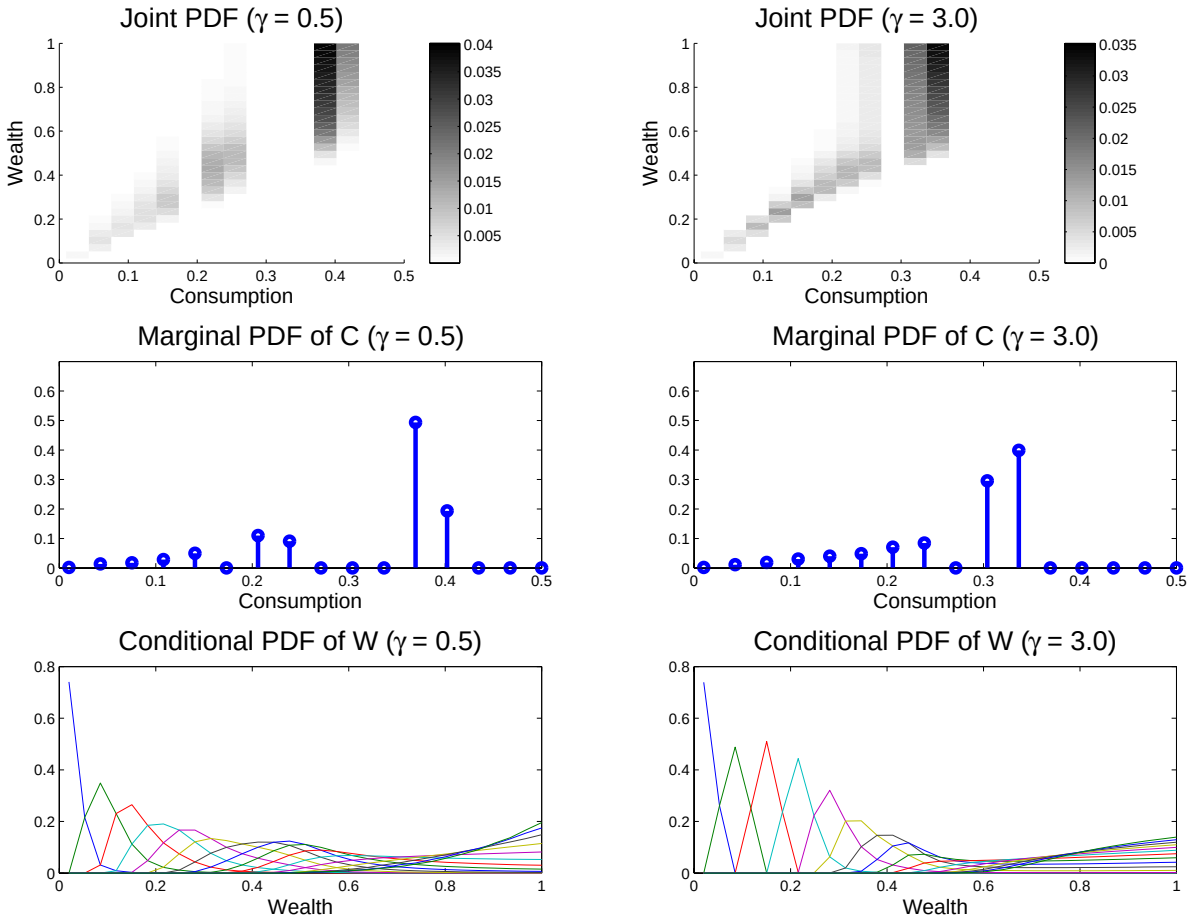


Figure 2.3: Comparison of Two Levels of Risk Aversion with Information Processing Capacity of $\kappa = 0.85$ bits.

parameter, γ, is changed.

The 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, we use different optimization algorithms. However, the biggest difference is in the log-normalization, so I have conducted an optimization using the log-transformation and an optimiza-

tion algorithm called CONOPT, which is more similar to the optimizer Sims used. 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 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 (basically guaranteeing a solution will be found), this 200-fold increase in speed allows us to examine a richer class of models using the current computing technology, as in Lewis (2007b).[6]

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

Following on the heels of Sims (2003), a number of other information-processing-constrained-agent models were introduced into the economics and finance literature. The problem, as pointed out in Sims (2006), is that most all of these models included an inappropriate feature. By misusing the Gaussian-in, Gaussian-out framework of the linear-quadratic setup of Sims (2003), model designers mistakenly used a parametric version of the joint distribution of state variables and choice variables (in our model, $f(c, w)$). That is, in order to simplify the structure of the model, the model

[6]Two additional results-based appendices are available from the author upon request. The first is a short document 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.

designers *assumed* that the utility-maximizing form of $f(c, w)$ would be Gaussian, and optimized over the parameters of the corresponding distribution. What Sims (2006) points out is that by limiting examination of the solution-space to parametric distributions (even more – to a specific parametric distribution) is that this is keeping the agent from *actually optimizing*: "In a model of an optimizing agent, the agents objective function will therefore determine the stochastic process for the joint behavior of actions and external signals" (Sims, 2006, pg. 3).

In order to demonstrate the problem that Sims brings up, we demonstrate the effect on the two-period model of (ceteris paribus) requiring that $f(c, w)$ be a bivariate Gaussian distribution. To this end, we will use a Gaussian wealth distribution ($\mu_w = 0.5$, $\sigma_w = 0.25$), and ask the optimizer 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 2.4 below. 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 we see 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, we see 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 smooth "down" from there. This means that the agent's risk preferences cannot be taken into account as well as they can in

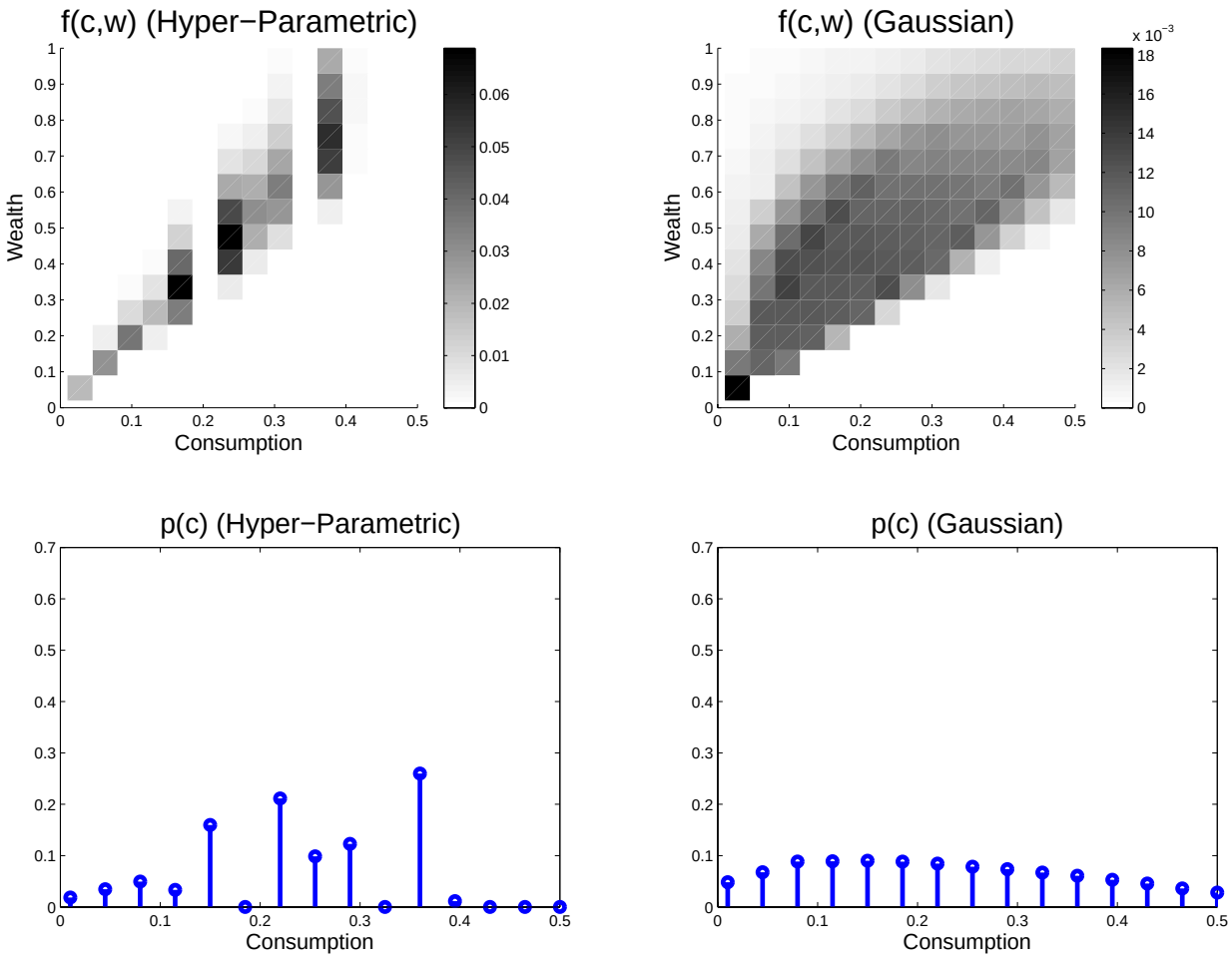


Figure 2.4: The Assumption of Gaussianity

the non- (or hyper-) 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. Ironically, the Gaussian decision rule is 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. To illustrate this, examine the scales of the colorbars on the sides of the plots in figure 2.4. As an example, 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, we have said that the agent is 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. Further, it should be noted that while the non-parametric model solves in under 2 seconds, the restricted model restricted to be parametric takes 3-4 minutes.

2.8 Conclusion

RI models represent exciting potential in economic modeling. These tools can be extended to be used to examine models where agents are constrained in processing capability without constraining *how* the agents use that capacity. This paper

demonstrates that these problems can have tractable, convex representations, how to find them, and what tools exist to deal with them quickly. We confirm the results in Sims (2006) with regard to the effects of information-processing constraints, and demonstrate Sims' point regarding assumption vs. derivation of optimal behavior in an RI environment. Our semi-analytical approach to the problem brings to light a few numerical issues that confront all RI models and demonstrate that care needs to be taken in translating analytical work of this nature to the computer. We demonstrated that extensions can be easily examined using the tools presented, and look into a future of larger and richer models of information-processing capacity constrained agents.

CHAPTER 3
THE LIFE-CYCLE EFFECTS OF INFORMATION-PROCESSING CONSTRAINTS

3.1 Introduction

Mounting evidence suggests that the hump-shaped profile of consumption over the life-cycle is a consequence of individual choice, and not an artifact of other factors such as growth in family size or accumulation of durable goods or even general economic growth during an agent's life [Gourinchas and Parker (2002), Fernández-Villaverde and Krueger (2004)]. The hump-shape runs counter to intuition regarding the smoothing behavior of economic agents: one would expect consumers to save and dissave so as to keep consumption relatively constant from year to year. This paper seeks to demonstrate that the phenomenon can be explained within a framework that includes uncertainty and information-processing constraints popularized recently in the Rational Inattention (RI) paradigm of Sims (2003, 2006). The paper takes seriously the notion of optimal use of information-processing capacity and in doing so, derives a hump-shaped profile of consumption over the life-cycle, as well as circumstances under which agents have a non-trivial probability of leaving behind wealth when they die, despite exact knowledge of the timing of their mortality.

The RI paradigm has been recently advanced to a more general framework in Sims (2006). Prior to this work, models including information-processing constraints bypassed the central issue of RI by appealing (in some cases, incorrectly) to a Gaussian-in, Gaussian-out framework that side-stepped the question of *how* agents

should allocate their information-processing resources. This paper seeks to examine this resource-allocation question in a simple life-cycle setting, using much of the same technology of the Sims (2006) two-period model, but with the addition of a trade-off allowing agents to give up processing power for something of economic value. It focuses on the dynamics of the information problem and the value of information-processing capacity, and demonstrates anew that simple applications of uncertainty and information-processing constraints can produce substantial changes in otherwise standard models.

Many papers in the life-cycle literature share the presumption that uncertainty regarding the agent's environment likely plays a role in the profile of consumption over time [Caballero (1990), Carroll (1992, 2004), Deaton (1992), Sims (2003), Luo (2004)]. How the agent confronts this uncertainty is at the center of several attempts to resolve the consumption puzzle. Most of these (excluding those of Sims and Luo) assume that the agent is aware of all information in the system at all times, even in cases where "all the information" includes knowing complicated distributions over many variables. This represents a large abstraction from reality that is accepted in the name of model tractability. On the other hand, the purpose of the RI framework is to take seriously the notion that agents do not possess infinite computational power. As demonstrated in Sims (2003), among others, information-processing constraints produce results which look more like observed data; a similar result characterizes this extension. By providing an analysis of the dynamics of optimal information-processing allocation, this paper takes a step toward including these types of frictions

in more general economic models.

The remainder of the paper is structured as follows: Section 3.2 contains a discussion of the rational inattention problem both in general and as applied specifically in this model. Section 3.3 describes the life-cycle consumption problem and a brief description of its history. Section 3.4 lays out the specifics of the model, while section 3.5 provides an analysis of the model results and discusses the optimal allocation of information processing that leads to the hump-shaped behavior. Section 3.6 concludes.

3.2 Rational Inattention

The idea behind rational inattention is not new. It can traced back at least as far as a 1978 address to the AEA meetings by Herbert Simon, who titled a portion of his talk "Attention As The Scarce Resource" [(Simon, 1978, p. 13)]. Indeed, we observe in our own lives that we do not possess unlimited information processing ability. As Sims points out: "...modeling agents as finite-capacity channels...fits well with intuition; most people every day encounter, or could very easily encounter, much more information that is in principle relevant to their economic behavior than they actually respond to" [(Sims, 2006, p. 2)]. What rational inattention provides is a comprehensive framework for stepping away from superhuman performance assumptions about economic agents, and doing so without behavioral assumptions that eliminate optimal use of resources, and without introducing arbitrary frictions.

Rational inattention is more than just limiting the information used by an

agent in his or her decision-making. It does not correspond to delaying or disguising the information either, as with most "information frictions." The central theme is one of optimal choice regarding how to reduce uncertainty. Agents make decisions that affect how the uncertainty in their world is reduced in a given time period by choosing what to pay attention to. A fundamental tenet of rational inattention is that the agent has *all the current information* at his or her disposal, but chooses optimally what to pay attention to – thus the monicker. This is different from the "inattentiveness" of Reis (2006) and Reis (forthcoming), wherein price-setting producers and consumers update their information only occasionally, but *completely*. Reis states in both papers that the applications of the two concepts are in theory quite similar, but practically speaking they approach the issue from two distinct sides: *rational inattention* focuses "on the information problem facing agents, at the cost of simplifying the study of their real actions;" while *inattentiveness* "focuses on these real decisions. . . at the cost of simplifying the information acquisition problem" [(Reis, forthcoming, p. 3)]. Thus while inattention is the optimal response to a finite capacity to process, inattentiveness results from an inability to acquire information frequently.

3.2.1 Information Theory and the Linear-Quadratic Model

The usefulness of information theory in the attention-allocation problem comes from its well-defined accounting of uncertainty. The measure of uncertainty is Shannon entropy, to be defined presently.[1] Using this metric, anything that reduces the

[1]Information theory is not the only scientific field to make use of the word *entropy*. Among them, physics – specifically the physics of heat – also uses this term. The word

amount of uncertainty (entropy) in the system can be accurately measured in terms of its information content, where information is defined as anything that changes the level of entropy in the system.

In economic models, uncertainty generally takes the form of variables that are characterized by probability distributions rather than being known explicitly. In the case of a Gaussian random variable, entropy is determined by the variance.[2] In Sims (2003), the use of information theory to assess the information being processed took advantage of this variance-entropy link, as well as other properties of the Gaussian p.d.f., in order to examine the permanent income hypothesis (PIH) under information constraints.[3]

In a simple PIH model with linear constraints and a quadratic utility function, maximization of utility subject to an additional information-processing constraint implies a Gaussian distribution over the state variable, wealth, conditional on currently processed information. Further analysis by Sims showed that in steady state, the agent's actions would be those of an individual who receives a noisy signal regarding wealth in the form of the true value plus some idiosyncratic noise. The covariance structure of this idiosyncratic noise is then a function of the information-processing

---

entropy in this and all rational inattention literature refers specifically to Shannon Entropy, which is defined below.

[2]By plugging in the Gaussian p.d.f. for $f(x)$ in the continuous case, we can see that the entropy will be equal to one-half times the log of the variance plus a constant term, meaning all the uncertainty in a Gaussian distribution is summarized by the variance parameter.

[3]Luo (2004, 2005) also make use of the LQG results of Sims (2003) in order to gain tractability in the RI paradigm.

constraint and the variance properties of transitory income. This type of result, that the information-processing constrained agent will behave as if he observes a noisy signal whose noise is an i.i.d. Gaussian random variable, is a result specific to the LQG framework and is not generally true outside of these very special circumstances. The result that the agent's optimal ex-post uncertainty was Gaussian combined with the linearity of the model due to the variance-entropy relationship of the Gaussian distribution allows this very special framework to be examined using tools designed for use in dynamic signal-extraction models with rational expectations. That is, the ability to make use of the Kalman filtering technology is a direct result of the fact that optimal attention allocation yields Gaussianity in the ex-post uncertainty. Sims (2006) emphasizes that *assumptions* about the form of optimal ex-post uncertainty are almost assuredly wrong (except in a few special cases), and that models that assume an arbitrary functional form for ex-post uncertainty are implicitly assuming that the agents in the model are not optimizers.

3.2.2 Information Theory and the General Model

Once we are outside the LQG universe, the behavior of an agent allocating attention to resolve uncertainty in the way that will generate the highest lifetime utility cannot be represented as a Kalman-filtering problem. As demonstrated by Sims (2006), the choice of attention that provides the highest expected utility does not result in ex-post uncertainty that appears to take any known functional form except in a few isolated cases, meaning new avenues will need to be pursued.

To facilitate a better understanding of how information is accounted for in these types of models, a brief digression on uncertainty and information in more general terms is warranted. For the random variable $x \sim f(x)$; the entropy of x is written:

$$S(x) = \begin{cases} -\int f(x) \log[f(x)] dx & \text{for continuously distributed } x, \\ -\sum_{i=1}^{N} f(x_i) \log[f(x_i)] & \text{for discretely distributed } x. \end{cases} \tag{3.1}$$

Shannon's entropy formula in (3.1) is designed to measure the amount of uncertainty in a random variable. Shannon's formula is derived from the following requirements. For a discrete random variable, suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \ldots, p_n$. These probabilities are known but that is all that is known concerning which event will occur. Shannon argued that a measure $[S(p_1, p_2, \ldots, p_n)]$ of how uncertain one is of the outcome of this random process should satisfy three properties:

1. S should be continuous in the p_i's.

2. If all the p_i are equal, $p_i = 1/n$, then S should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.

3. If a choice be broken down into two successive choices, the original S should be the weighted sum of the individual values of S. The meaning of this is illustrated in figure 3.1. At the left we have three

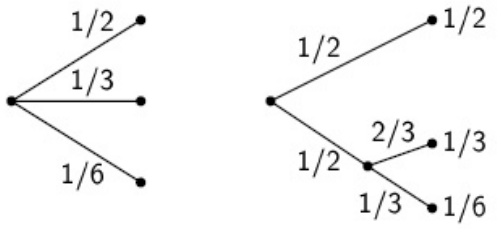


Figure 3.1: Decomposition of a Random Process

possibilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. On the right we first choose between two possibilities each with probability $\frac{1}{2}$, and if the second occurs make another choice with probabilities $\frac{2}{3}$, $\frac{1}{3}$. The final results have the same probabilities as before. We require, in this special case, that

$$S\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = S\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} S\left(\frac{2}{3}, \frac{1}{3}\right)$$

The coefficient $\frac{1}{2}$ is because this second choice only occurs half the time.

Appendix 2 of Shannon (1948) proves:

Theorem 3.1. *The only S satisfying the three properties is of the form:*

$$S = -K \sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant only concerned with the units of the uncertainty measure.

To see how entropy is a measure of uncertainty, imagine a two-point discrete distribution over the values $x = a$, with probability p, and $x = b$, with probability

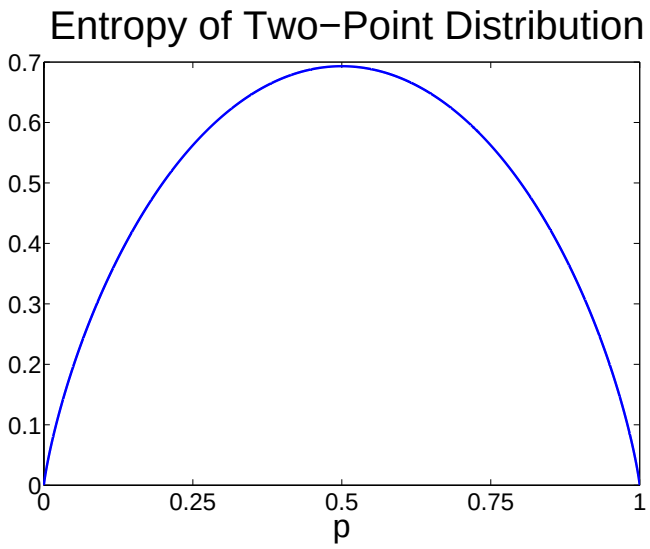


Figure 3.2: The Entropy of the Two-Point Distribution for Potential Values of p.

$(1-p)$. It is instructive to consider how the measure of uncertainty varies with p. In this case,[4]

$$S(x) = -p\log(p) - (1-p)\log(1-p) \tag{3.2}$$

The first order conditions of either the maximization or minimization of equation (3.2) yield the result

$$\log(1-p) - \log(p) = 0, \quad \text{for } p > 0$$

[4]The value of entropy is expressed in specific units. The common nomenclature of information theory expresses these values in binary units, or *bits*. In order for this to be the units of the entropy calculation, the logarithms in equation (3.1) would need to be logarithms of base two. In most modern programming languages, and in nearly all economic examples, the logarithms are base e (natural logarithms) and the proper units for these calculations are named natural units, or *nats*. In this paper, we will use nats, however, units of information can be easily translated. For example, to convert from logarithms of base a to logarithms of base b, simply multiply the former by $\log_b(a)$.

The second order condition is negative for all values of $p > 0$, and it is clear from the FOC and figure 3.2 that the value of p that creates the most uncertainty about potential values of x (thus, the maximum entropy for this distribution) is $p = 1/2$. It is also clear from figure (3.2) that the distribution with minimum entropy is the one where either $x = a$ or $x = b$ with probability one. That is, there is zero entropy – meaning zero uncertainty – when we know that every realization of x will be a.[5] Uncertainty is reduced for randomly distributed variables as weight is concentrated on a relatively small number of points. As weight is dispersed evenly across all possibilities, uncertainty – and therefore entropy – increases, to a maximum when distributed uniformly of $\log n$ when all n of the p_i are equal. Additionally,

$$S(x, y) \leq S(x) + S(y)$$

with equality only when x and y are independent. Further interesting properties of $S(x)$ that justify its use as a reasonable measure of uncertainty are given in Shannon (1948).

In rational inattention, *relative* entropy is formed in relation to a concept called mutual information: a measurement by which one can understand the amount of uncertainty about one variable that is reduced by observation of another. In the model to be laid out below, in each period we will constrain the amount of mutual information agents can process between consumption and wealth. In particular, there

[5]It should be noted that while $\log(0)$ is undefined, $x \log(x)$ is a continuous function on $x \in [0, \infty)$ and it can be demonstrated via L'Hôpital's Rule that $\lim_{x \to 0} x \log(x) = 0$.

will be a limit to the expected reduction in the entropy of consumption that can be achieved by observing draws from wealth. As this paper will be working in discrete distributions, the mutual information of two discrete random variables x and y having joint distribution f, identical N-point support, and marginal distributions p and g is given by

$$MI(x,y) = \sum_{j=1}^{N}\sum_{i=1}^{N} f(x_i,y_j)\cdot\log[f(x_i,y_j)] - \sum_{i=1}^{N} p(x_i)\cdot\log[p(x_i)] - \sum_{j=1}^{N} g(y_j)\cdot\log[g(y_j)]. \tag{3.3}$$

The value of $MI(x,y)$ is equal to the sum of three components in this two-variable case: beginning at the far right, the $MI(x,y)$ is the sum of the entropy of the marginal distribution of x and the the marginal distribution of y, minus the entropy of the joint distribution of x and y. Mutual information is a generalization of correlation in that as the conditional distributions of x given y (or y given x) tighten around a single x (or y), the amount of mutual information increases. Note that if x and y were independent, then $f(x_i,y_j) = p(x_i)g(y_j)$, and it is easy to see that $MI(x,y) = 0$. That is, observing a random variable x independent of y provides no information about y and vice-versa.

3.2.3 Optimal Allocation

Accounting for information flows is the groundwork for the theory of rational inattention. Shannon (1948) also addresses what is referred to in the information theory literature as the optimal coding problem. The optimal coding problem *is*

the problem solved by agents in the rational inattention model. The goal is to take the scarce resource, attention capacity, and allocate it in such a way as to maximize lifetime utility. As the economic version of the optimal coding problem is presented as the model below, we will introduce the idea of it with a quick non-economic example.[6]

> Suppose you communicate with another individual using a system of only zeroes and ones. You need to send the person on the other end of the line some data represented as sequence of zeroes and ones of length 10,000. The series, it turns out, is simply 9,999 zeroes followed by a single one.

The optimal coding problem is: What is the best way to transmit these data? Clearly, the brute force method would be adequate, but would require 10,000 "characters" to be transmitted. Suppose, however, that you could not transmit that many characters in the time you are allotted. One intuitive way to get the information across could be using Morse code (0 = dot, 1 = dash) to transmit the sentence: "The stream is nine-thousand, nine-hundred, ninety-nine zeroes followed by a single one." You could also shorten it further by using blocks of zeroes and ones to stand for letters and numbers, as is done in modern computers. This would require you to send only 4 blocks: "9,999", "0", "1", "1". While this example is overly simple, it illustrates the requisite point: agents can format the information they need in order to deal with information constraints.

[6]This example comes from the theory of data compression, and has been slightly oversimplified for clarity.

The optimal coding of information is the task of distilling data to the form that best represents the original data-stream while meeting the requirements of a finite-capacity transmission channel. Information-processing constraints on economic agents force them to distill the information in their world in such a way as to consider only the most useful elements of their complete dataset. That is, agents must solve the problem of determining what to consider and what to ignore when faced with their inability to consider everything.

3.2.4 Optimal Allocation and the MI Constraint

The amount of information-processing required to solve economic problems depends on the complexity of the problem, but even the simplest problems have processing requirements. To see how information-processing constrained models differ from their unconstrained counterparts, consider the two-period model of Sims (2006). 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,

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

for a given w. 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. The use of f to denote a decision rule here after using it in an earlier subsection to denote a probability distribution is intentional: under rational

inattention, decision rules are probability distributions. 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 if 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.

3.2.4.1 A Generalization

To set the stage for the information-constrained problem to come, 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 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.

Under this "generalization," the agent's optimization problem is to choose the joint distribution $f(c, w)$ to:

$$\max_{\{f(c_i,w_j)\}} \sum_{i=1}^{N_c} \sum_{j=1}^{N_w} \frac{c_i^{1-\gamma} + (w_j - c_i)^{1-\gamma}}{1-\gamma} f(c_i, w_j) \tag{3.4}$$

subject to:

$$\sum_{i=1}^{N_c} f(c_i, w_j) = g(w_j) \qquad \forall \;\; j = 1, \ldots, N_w \tag{3.5}$$

$$f(c_i, w_j) \in [0, 1] \qquad \forall \;\; i, j \tag{3.6}$$

$$f(c_i, w_j) = 0 \qquad \text{for } c_i > w_j. \tag{3.7}$$

The properties of the problem and the optimum are qualitatively unchanged under this generalization. Suppose that the marginal distribution of wealth is triangular, meaning higher levels of wealth have higher probability. The optimal decision rule is the joint distribution $f(c, w)$ that describes the same one-to-one mapping that divides wealth into two halves and consumes one in each time period. 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 $\{c_i\}_{i=1}^{N_c}$ such that weight is only allowed where $c_i \leq w_j$. The optimal choice, as seen in figure 3.3, 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, $f(c_i = w_j/2, w_j) = g(w_j)$.

Figure 3.3 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 weight, as indicated by the legend on the right-hand side of the figure. The boxes get

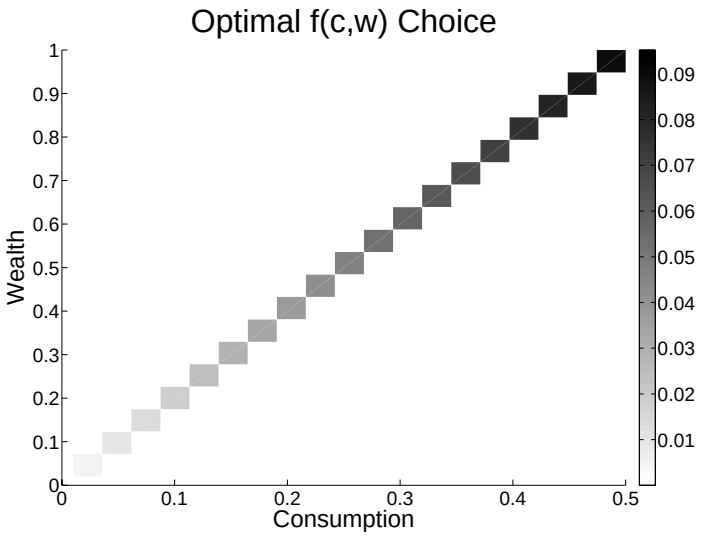


Figure 3.3: A One-to-One Mapping Via the Joint Distribution $f(c, w)$.

darker as they progress "northwest" 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 $c_i = w_j/2$.

3.2.4.2 Correlation and Mutual Information

The joint distribution of c and w in figure 3.3 shows a strong relationship between the variables. Indeed, as the set of conditional distributions becomes degenerate, the amount of mutual information increases toward a maximum. Mutual information is as large as it can be in figure 3.3 because for each draw of w, a *specific* c value is determined, and vice versa.

3.2.4.3 Mutual Information and Information-Processing Capacity

The heart of the rational inattention framework is the idea that agents are incapable of processing all the information related to their economic decisions. In models in which wealth is the state variable, if the agent does not *exactly* know his wealth in each period, he is modeled as having a distribution over wealth which stems from a noisy signal he has received. In the rational inattention framework, the agent is also modeled as choosing conditional distributions over consumption given wealth. This can be thought of as stemming from the amount of attention that must be paid to prices in order to purchase a specific value of consumption, rather than a specific bundle of goods. An exact level of consumption spending *can* be met, but in order to do so a large investment of information-processing capacity must be made: prices must be compared and calculated, expenditures within the time period must be accounted for precisely and updated continually. All of this is possible, but it requires processing many small pieces of information. Processing this information tightens the conditional distributions of consumption given wealth, moving the agent closer to a one-to-one mapping between consumption and wealth.

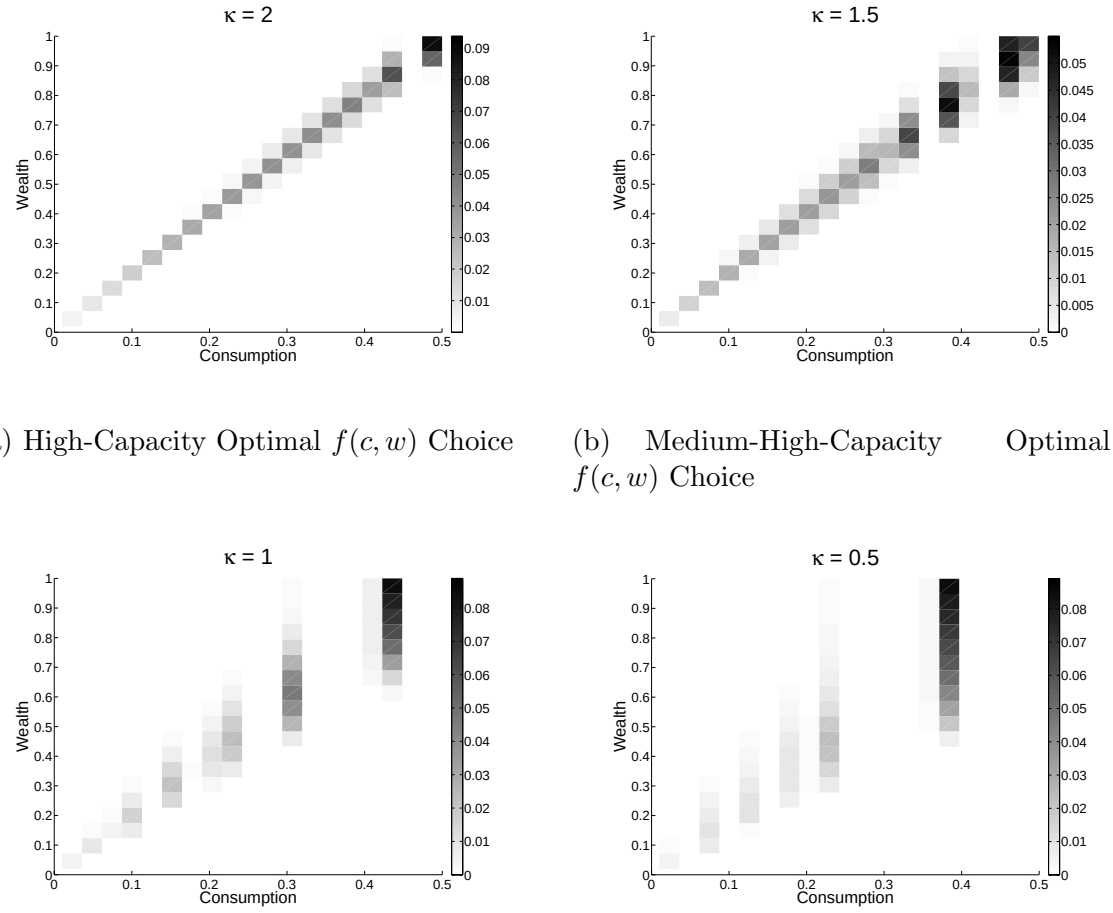
The rational inattention framework uses the metric of mutual information to quantify the amount of information-processing capacity the agent is using to solve his optimization problem. By placing a constraint on mutual information, the framework limits the strength of the relationship between c and w by 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 3.3, the agent must decide how best to allocate the finite resource of processing capacity across the space of his choice variable. An important result of Shannon (1948) should be noted: the information-processing capacity constraint will always bind when the amount of mutual information allowed is less than or equal to the amount of mutual information required for an optimal unconstrained mapping. The agent will use all available capacity to process the information in his environment, thus capacity equals the amount of information processed.

The agent's optimization problem in the information-processing constrained universe is the same as the one detailed in equations (3.4) through (3.7) above, with the addition of the following constraint on the amount of mutual information in the model:

$$\sum_{j=1}^{N_w}\sum_{i=1}^{N_c}\log[f_t(c_i,w_j)]\cdot f_t(c_i,w_j) - \sum_{i=1}^{N_c}\left(\log\left(\sum_{j=1}^{N_w} f_t(c_i,w_j)\right)\cdot\sum_{j=1}^{N_w} f_t(c_i,w_j)\right) \quad (3.8)$$
$$-\sum_{j=1}^{N_w}\log(g_t(w_j))\cdot g_t(w_j) \leq \kappa.$$

As the amount of information-processing capacity (κ) decreases, 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 state of the model. The rational inattention framework does something similar to this by describing an environment in which the "noise" is endogenously determined rather than exogenously



(a) High-Capacity Optimal $f(c, w)$ Choice

(b) Medium-High-Capacity Optimal $f(c, w)$ Choice

(c) Medium-Low-Capacity Optimal $f(c, w)$ Choice

(d) Low-Capacity Optimal $f(c, w)$ Choice

Figure 3.4: Optimal Choice of Joint Distribution $f(c, w)$ Given Triangular $g(w)$ for Various Levels of Information-Processing Capacity.

given: it arises from the agent's inability to accurately asses this state of the model because he does not have the information-processing resources to do so.

Figure 3.4 demonstrates the effect of lowering the information-processing capacity: a discretization of the consumption choices by the agent. In panel 3.4(c), the agent chooses to place no probability weight on consuming at the level of, for example, 0.275. The agent's noisy signal for wealth generates consumption behavior that does

not as closely track potential differences in wealth as the information-processing *unconstrained* model would. That is, there is a range of possible wealth levels that could result in a level of consumption, and also a range of possible consumption levels that correspond to a given potential wealth level. This is a many-to-many mapping rather than a one-to-one mapping, but still describes how the agent chooses consumption based on his information regarding wealth.

3.3 The Life-Cycle Model

The finite life-cycle model is a convenient vehicle for studying the dynamics of general RI behavior. The effect of planning over a multi-period time horizon with a per-period processing constraint can be articulated clearly without the further complexity induced by an infinitely-lived agent. Even in this simple model, the effects of RI-style uncertainty are significant.

3.3.1 The Canonical Life-Cycle Model

The canonical life-cycle model has no growth in income or household size, no borrowing, no shocks to the income stream or preferences, and a certain ending period. The agent has a constant income, no initial savings, discounts the future geometrically, and has no uncertainty regarding the future. That is, a consumer with period utility of consumption, c_t, given by $U(c_t)$, discount rate β, initial wealth w_0, and facing a constant interest rate R, chooses $c_1, c_2, \ldots$ to

$$\max_{\{c_t\}_{t=1}^{T}} \quad \sum_{t=1}^{T} \beta^t U(c_t)$$

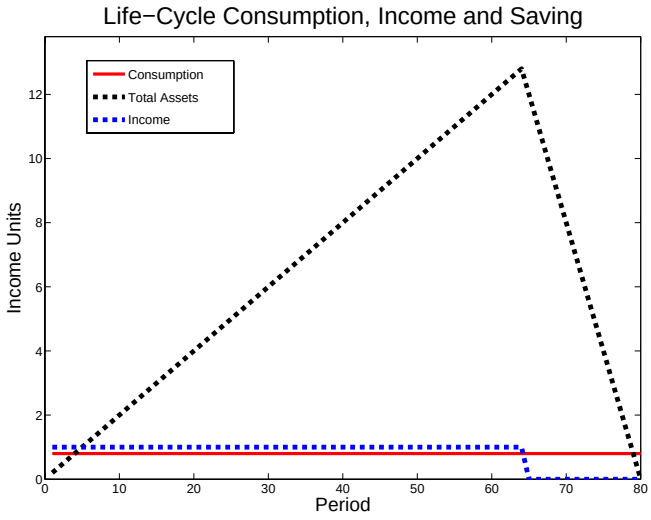


Figure 3.5: The Stripped Down Life-Cycle Model

subject to:

$$W_{t+1} = (1+r)W_t + y_{t+1} - c_{t+1}, \qquad t = 1, \ldots, T-1$$

$$S_t \geq 0, \qquad t = 1, \ldots, T$$

This "stripped-down" life-cycle model [demonstrated in Modigliani (1986), and dating back to the mid-50's] is a textbook staple and produces the well-known result in figure 3.5,[7] of consuming a fixed fraction of life-time wealth in each period.

What is observed in the data is clearly different from figure 3.5; one recent study shows: "...total consumption expenditures, as well as expenditures for non-durables and durables display a significant hump over the life cycle, even after

[7]This is a picture of the result for $\beta(1+r) = 1$. Additionally, U needs to be monotone increasing and concave, in this case, it is CRRA. In general, the path of consumption will increase or decrease depending on the value of βR; see e.g. Yaari (1964).

accounting for changes in family size," [(Fernández-Villaverde and Krueger, 2004, pg. 2)]. Attempts to resurrect the canonical model have demonstrated the effectiveness of introducing income or wealth uncertainty (Aiyagari (1994), Carroll (1994), Deaton (1991)). Gourinchas and Parker (2002) note the case for long-term consumption smoothing weakens substantially in light of recent empirical studies which "often find that [household level] consumption responds to predictable changes in income." Similarly, so-called buffer-stock behavior ([Carroll (1997)], also called precautionary savings and prudence elsewhere) is witnessed in the presence of uncertainty and borrowing constraints, creating non-smooth consumption as well.

The RI framework can produce the observed hump-shape without the use of income growth, changes in household size, non-geometric discounting, adjustment costs, or many of the other common mechanisms imposed to generate these results. This study finds that with very few free parameters, RI uncertainty can replicate the observed hump. The only changes to the canonical model are the introduction of uncertainty to income and wealth, and the RI mechanism by which agents can choose how and, in this model: *if and when*, they would like to work to resolve this uncertainty.

3.4 The RI Life-Cycle Model

In the canonical model, agents know their wealth and income and make precise choices regarding consumption in each period. The RI framework postulates an environment in which this amount of attention to detail is infeasible. Agents both

observe the world and make decisions imprecisely as a result of an inability to pay attention to everything. That is, precision itself, in both understanding and action, is a resource being allocated in the information-processing constrained model.

Why can't agents choose precise levels of consumption given wealth? As mentioned earlier, the idea of attention as a scarce resource in the life-cycle model can be thought of in the context of purchasing everyday goods and services. Consider the purchase of groceries or household items that are obtained at regular intervals. Small fluctuations in the prices of individual goods within the bundle are commonly absorbed or ignored. Agents have a sense of (or a distribution for) their net worth (wealth), and when wealth changes greatly, so will day-to-day consumption behavior. How much wealth has to change in order to affect consumption behavior is a function of how much attention is paid to the distribution over wealth, and how complex this distribution is. Similarly, total wealth is a function of income and consumption. Small changes in consumption, due to small price fluctuations for example, are difficult to track. Having very tight conditional distributions for consumption at a specific wealth level requires that agents very, very closely monitor small changes in prices, an action that requires a large amount of attention. This attention might be better well spent (in terms of lifetime utility) by less accurately determining each conditional distribution of consumption, but determining these conditional distributions over more possible levels of wealth. Thus, the higher the entropy of the wealth distribution, the higher the information-processing requirement of forming optimal strategies, and conversely, the more precisely one wishes to determine behavior at a given wealth level,

the less precision will be "available" for consumption determinations at each additional level of wealth. Agents are simultaneously allocating attention across wealth and consumption.

3.4.1 Income vs. Processing Capacity

In an attempt to think about the value of a unit of information-processing capacity, we give the agent the opportunity to trade this capacity for potential future income. One possible way to do this is to consider an environment in which the agent has the opportunity to divide his time in the current period between two activities: time devoted to processing information related to the current consumption-wealth decision, and time devoted to increasing future income.

The division of time spent in the current period on the two activities is represented by the parameter, $\alpha_t \in [0, 1]$, which is chosen by the agent. Current-period information processing activities can be thought of as time spent balancing one's checkbook or checking one's debit card balance, as well as clipping coupons, checking internet sites for sales, comparison shopping within and across similar stores, determining where the lowest gas prices are locally, etc. The maximum fraction of time that can be spent in period t is $\alpha_t = 1$. The formula for period t processing capacity is written:

$$\kappa_t = \alpha_t \kappa^M, \quad t = 1, \ldots, T \tag{3.9}$$

where κ^M is the instantaneous information-processing capacity of the agent. As α_t

increases, the agent spends more time processing information.

The other side of the tradeoff is the expected income of the agent in the following period. One can think about the intuition of this side of the tradeoff in the following way: this is time spent working overtime, or time spent reorganizing an investment portfolio, or time spent playing golf with the boss, or anything that is likely to increase the expected value of future income. Here, income is modeled in a fairly simplistic way: period income arrives as a distribution over K nodes that are fixed for the entire lifetime. There is no growth in income in this model. By spending time to improve income prospects, the agent shifts probability from lower income states to higher income states. In each period, income takes one of K values, e_r, $r = 1, \ldots, K$. If in the previous period all effort was devoted to information processing ($\alpha_{t-1} = 1$), the income distribution at time t is uniform. As α_{t-1} decreases, probability is shifted toward values as depicted in figure 3.6, according to the following formula for the distribution of per-period income:

$$b_t(e_r|\alpha_{t-1}) = \frac{r^{2(1-\alpha_{t-1})}}{\sum_{s=1}^{K} s^{2(1-\alpha_{t-1})}}. \tag{3.10}$$

This distribution has the property that lower-α distributions stochastically dominate (to first order) higher-α distributions. This model choice (the specific distribution of per-period income and the corresponding effects of the tradeoff) is the focus of ongoing research.

In each time period, the agent will have the opportunity to vary his total pool of attention, as well as how it is used. This is important because to date, work in ra-

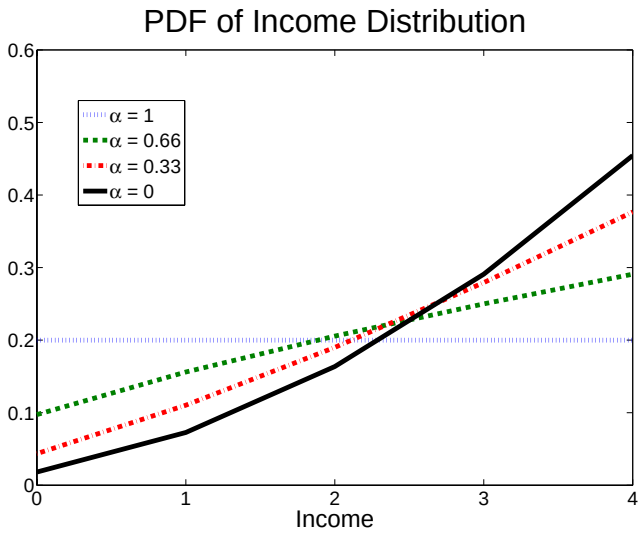


Figure 3.6: An Example of the Income Process When $e_1 = 0$, $e_2 = 1$, etc.

tional inattention assumes only an exogenous information processing constraint that binds. The intuition behind the binding constraint is obvious: no one can process all the information available to them. However, one could consider devoting more time to information processing than he or she does currently. The flexibility must be capped such that no one can process all information, thus the κ^M parameter; but an important addition to the RI literature is the concept of "paying" for additional processing capacity, which is the central idea behind the model components described in equations (3.9) and (3.10). By beginning to analyze the value of a unit of information-processing capacity, we can move toward models that consider the ways in which the market helps agents solve their optimal attention-allocation problem.

3.4.2 On the Entropy Effects of the Income Process

In equation (3.10), the mean of the time t income process is controlled by the value of α_{t-1}. What is seen in figure 3.6 is that in addition to the mean effect, there is also an effect on the entropy of the income distribution. As α_{t-1} gets closer to zero, more and more probability weight is taken from low income values and moved to high income values. This has the effect of reducing the entropy of the income process (concentrating more weight on fewer points reduces entropy). This, in turn, influences the entropy of the wealth distribution in the subsequent (and therefore every future) period. Thus, while the original intent of the α_t tradeoff was to represent the optimal mixing of two goals — increasing future income and increasing current processing capacity — the entropy-reducing side-effect of this particular income process potentially muddles the interpretation of the tradeoff represented by α_t.

The ideal solution to this problem would be to use an entropy-neutral (with respect to α_{t-1}) income process. This solution, however, is problematic. The existing income process removes probability from lower-income and places it on higher income values. This monotonically increases the mean and decreases the entropy of the income distribution. What is required is an income process whose shape does not change *at all* as the mean is increased by the α_t parameter. For example, a univariate Gaussian distribution with mean μ and standard deviation σ has a theoretical entropy value of $1/2\log(2\pi e)+\log(\sigma)$, which only depends on σ and therefore, any α_{t-1} scheme that shifts the distribution by making μ a simple function of α_{t-1} should achieve this entropy-neutral property.

However, employing a discretization of the Gaussian distribution on the wealth grid eliminates the tidiness of the theoretical entropy calculation. Consider the following income process: a discretized Gaussian distribution where $\mu = (1-\alpha_{t-1})\mu_{HIGH} + \alpha_{t-1}\mu_{LOW}$. Thus, as $\alpha_{t-1} \to 0$, the agent is making the mean higher and higher, just as in the original process. The standard deviation of the discrete distribution is σ (fixed), which on a continuum, fixes the entropy of the income process. The discretization is accomplished by finding the value of the kernel of the $N(\mu, \sigma)$ distribution at each income node and then dividing by the sum to normalize the income process making it sum to one. Suppose that $\mu_{HIGH} = 14$, $\mu_{LOW} = 7$, $\sigma = 1$ and the income grid is the integers from 0 to 21. Figure 3.7 shows the mean and entropy of the income distribution for $\alpha_{t-1} \in [0, 1]$.

What is seen in figure 3.7 is the result of a continuous choice for α_{t-1} and its effect on a discrete grid. Note that the entropy values in this example (given on the right y-axis) differ in the seventh decimal place. This is a very small variation in the entropy, but from a numerical optimization point of view, catastrophic. The sine-wave pattern of entropy creates an effect that makes determination of the optimal choice of α_{t-1} extremely difficult, as indicated by the optimizer's inability to find an optimum under this income formulation, despite such a tiny change in entropy. Therefore, while the original model has an income process in which entropy is monotonically increasing in α_{t-1}, it is at least optimizable. The peaks of figure 3.7 are associated with the integer values between μ_{HIGH} and μ_{LOW}. *This sine-wave pattern exists for any continuous choice of α_{t-1} on a discrete grid.* The problem is a result of small

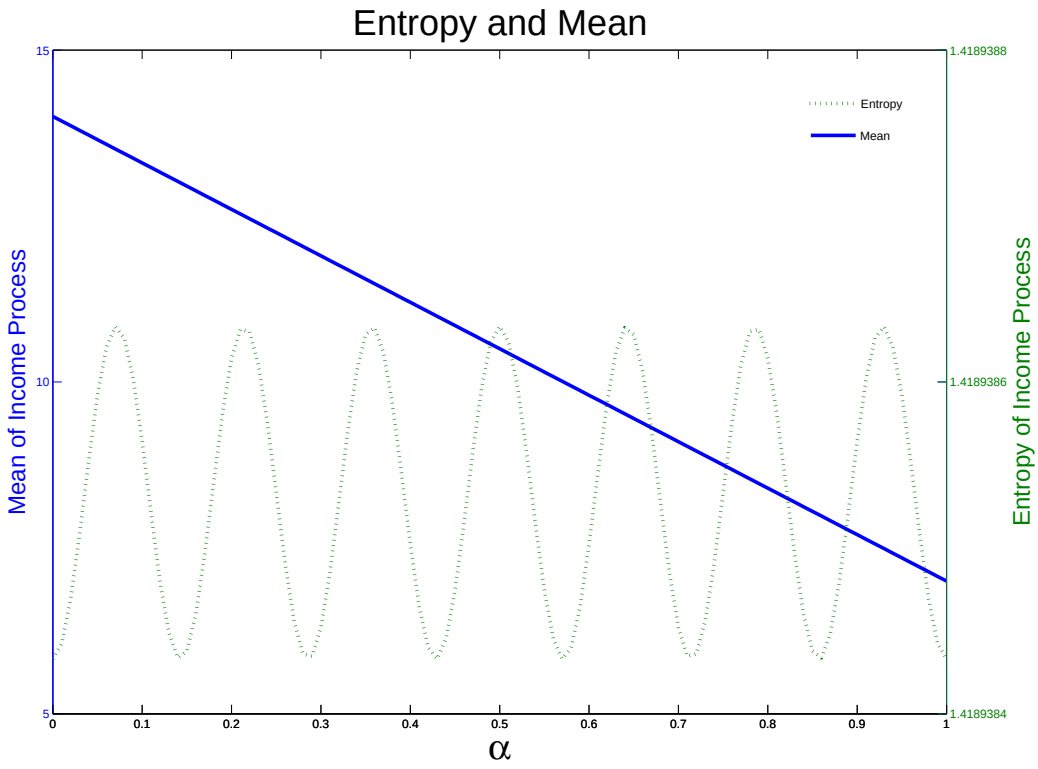


Figure 3.7: The Entropy and Mean of an Income Process for $\alpha_{t-1} \in [0, 1]$.

changes in entropy due to moving tiny probability weights from lower portions of the support to higher portions of the support, as a result of changes to α_{t-1}.

The goal remains unsatisfied: a discrete distribution for income in which α_{t-1} has no effect on entropy while controlling the mean. One way of having two distributions over the same support p and q with the same entropy and different resulting means would be to have p be a simple re-ordering of q. That is, $p_i = q_j$ for some i and j. Imagine a scenario where q is a discretized Gaussian distribution with the conventional bell shape. It is possible to have p be simply the probabilities from q lined up in ascending order, over the same support. Thus, $H(p) = H(q)$ but $E(p) > E(q)$. Another option would be to have a small uniform distribution that shifts (e.g. a three-node pdf with 1/3 weight on each node), where α_{t-1} controlled *which* three nodes received probability. Either of the two potential solutions mentioned here require some kind of discretization of the choice of α_{t-1}, which *dramatically* complicates the problem. Allowing the choice of α_{t-1} to be from some discrete list (e.g. $\alpha_{t-1} \in \{0, 0.1, 0.2, \ldots, 1\}$), transforms the optimization problem into what is known as a mixed-integer nonlinear programming (MINLP) problem. These problems are the subject of the edge of optimization theory, are ill-tempered even under the best of circumstances and are not, at this time, a practical research avenue for this problem.[8]

A potential solution to the entropy-neutrality problem within the context of

[8]The method commonly used to solve these problems (known colloquially as "branch and cut" or "branch and bound") is unreliable to implement on a problem with this structure, given the high dimensionality and complexity of the problem that remains after stipulating a vector of α_t's. For a quick overview of MINLP's, see Bussieck and Pruessner (2003) and the references therein.

the above discussion using a continuous α_t choice would be to modify two probability nodes from the discretized Gaussian distribution generated by α_{t-1}. One could theoretically alter two of the probability nodes to achieve two simultaneous goals:

- The modified distribution has a specific entropy, fixed to eliminate the wave in figure 3.7.
- The modified distribution sums to one.

This process could theoretically be used to zero out the effect seen in figure 3.7, but the process would essentially involve finding the solution to the equation $p\log(p) = C$, and therefore mean an internal optimization designed to push the objective of the optimization $\min_p p\log(p) - C$ to zero, from the point of view of the computer, not just to an arbitrary tolerance chosen by the user. If the computer does not zero out the entropy difference completely, the problem illustrated in figure 3.7 would persist. The internal optimization described here is *very* time-consuming (asking a computer to iterate on a problem, using smaller and smaller steps until "machine zero" is reached), if not computationally infeasible. Before this is pursued, it is reasonable to ask if the effect we are attempting to eliminate is important to the model results. The "side effect" we are trying to eliminate would allow the agent to get both higher expected income and lower uncertainty about that income. Is the agent interested in this side effect, or is the increase in the expected income the sole reason for the agent's choice of α_t?

3.4.2.1 Is the entropy feedback effect being used by the agent?

While it is potentially infeasible to use a continuous α_{t-1} to change the mean of the discrete income process in an entropy-neutral way, it *is* possible to use a continuous choice of α_{t-1} to change the entropy of the income process while leaving the mean of the distribution intact. Consider the following income process:

$$b'_t(e_r|\alpha_{t-1}) = \frac{(K^2 - (r-M)^2)^{(1-\alpha_{t-1})Z}}{\sum_{s=1}^{K}(K^2 - (s-M)^2)^{(1-\alpha_{t-1})Z}} \tag{3.11}$$

where M is the middle node of the income grid. This process is uniform when $\alpha_{t-1} = 1$, and as $\alpha_{t-1} \to 0$, weight is moved from the tails of the distribution to the center. Thus, the mean of the income distribution never changes, but the entropy monotonically increases in α_{t-1}.

Optimizing using this process will let us examine the tradeoff purely between information-processing capacity and the entropy of the future income process. The agent will have the ability to process more information "today", or have a more certain income "tomorrow." The expected income for "tomorrow" will not change, regardless of the agent's choice of α. What we see in figure 3.9 is that the choice is clear: the agent prefers processing power to lowering the entropy of the income process, regardless of time period. That is, the agent will devote all his "energy" (as evidenced in α_t being pushed to the boundary $\alpha_t = 1,\ \forall t$) to one activity: information-processing; and completely ignore the alternative activity of reducing the entropy of the income process. This tradeoff is of no interest to the agent.

Figure 3.9 demonstrates that the agent is unwilling to give up any processing

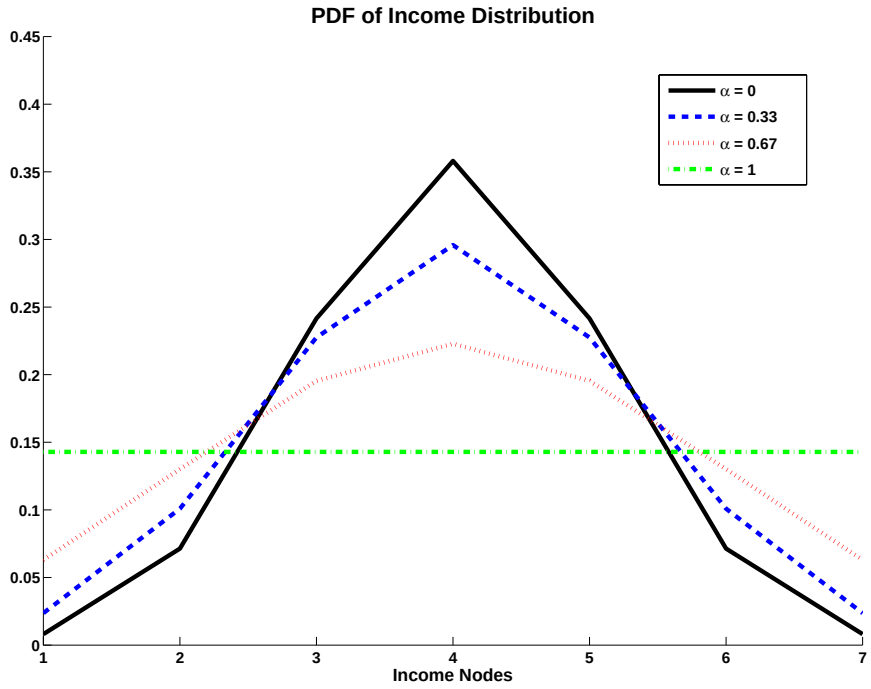


Figure 3.8: Mean-Neutral Income Process for Several α_t Choices, $K = 7$, $M = 4$, $Z = 20$

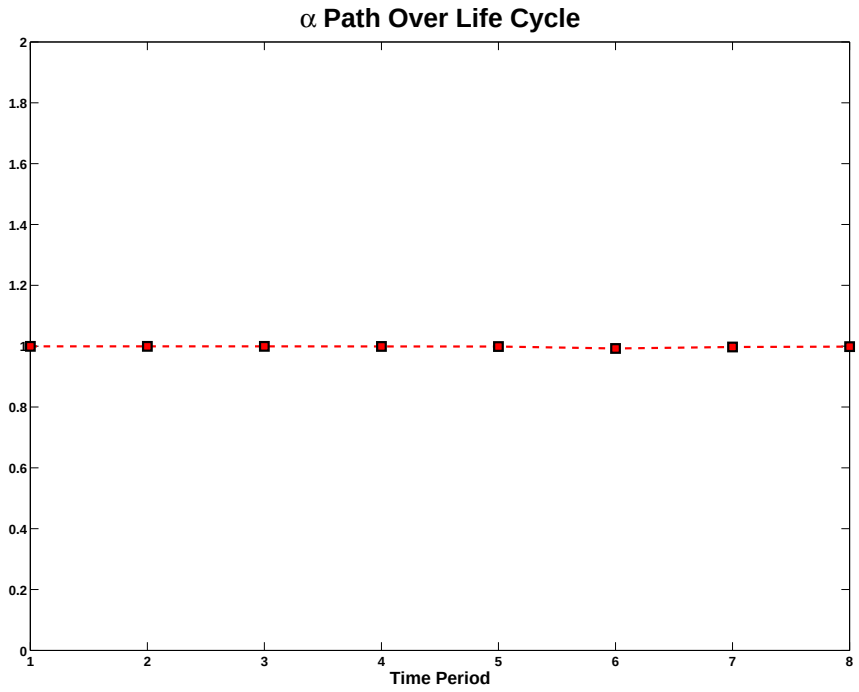


Figure 3.9: Choice of α_t, Mean-Neutral Income Process

power currently to reduce the entropy of the future problem. Later, we will look at the optimal choices for α_{t-1} given the original income process (the one specified in equation (3.10)). The role of Z is to control how entropy is affected by changes in α_{t-1}. The value for Z used to create figures 3.8 and 3.9 changes entropy at 10 times the rate of the entropy change in the original income process (the process in equation 3.10 which is the mean-shifting process), and values up to 50 times the rate of change in the original income process were tested with no effect on the results seen in figure 3.9. The exercise performed here for much more extreme parameterizations indicate that we can look at the optimal choices for α_{t-1} under the original specification with reasonable expectation that the agent is making use of the feedback effect in order to

get future and current benefits out of the income/processing-capacity tradeoff.

Therefore, while the current income process includes a feature that could be seen as a potential wrinkle in the model results, the important feature of monotonicity compensates for the inconvenience of the change in entropy, and the entropy effect appears to be inconsequential from the perspective of maximizing expected utility.

3.4.3 The Agent's Problem

The agent has the same period utility function as in the canonical model:

$$U(c) = \frac{c^{1-\gamma}}{1-\gamma}.$$

The objective function, however, is the generalization described in section 3.2.4.1: the choice variable is the joint distribution of consumption and wealth in each time period and the objective is to maximize lifetime *expected* utility. The choice variable, the joint distribution over consumption and wealth in each period, is a choice of probability weights on a fixed domain of (c, w) pairs. That is, the agent is choosing a set of probabilities for c_i and w_j, $i, j = 1, \ldots N$: $f(c_i, w_j)$'s, which means that agents are placing weight on points, not choosing the points themselves. *The grids of support for consumption and wealth are identical, and the points are evenly spaced.* The problem is to choose α and the probabilities at each date to

$$\max_{\{f_t(c_i,w_j),\alpha_t\}_{t=1}^{T}} \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N} \beta^t U(c_i) f_t(c_i, w_j). \tag{3.12}$$

In choosing the weights, the agent has a standard budget/borrowing constraint

in that he is unable to consume more than his available wealth. That is, in placing weight on consumption-wealth pairs, he is unable to assign positive probability to situations where consumption exceeds wealth:

$$f_t(c_i, w_j) = 0 \text{ for } c_i > w_j, \quad t = 1, \ldots, T. \tag{3.13}$$

Also, the standard constraint regarding probability weights must hold

$$0 \leq f_t(c_i, w_j) \leq 1\,, \quad i = 1, \ldots, N;\ j = 1, \ldots, N;\ t = 1, \ldots, T. \tag{3.14}$$

Additionally, recall that the model is one of having a distribution over wealth and making choices that are sets of conditional distributions of consumption given wealth. This is modeled as the choice of a joint distribution that is restricted to agree with the marginal distribution of wealth. Therefore,

$$\sum_{i=1}^{N} f_t(c_i, w_j) = g_t(w_j)\,, \quad j = 1, \ldots, N;\ t = 1, \ldots, T. \tag{3.15}$$

3.4.4 The Information-Processing Constraint

The information-processing constraint is a per-period constraint. That is, the mutual information between consumption and wealth is restricted to be less than the total capacity κ_t in each time period. The mutual information calculation in this discrete-distribution case is identical in form to that of (3.3):

$$\sum_{j=1}^{N}\sum_{i=1}^{N}\log[f_t(c_i,w_j)]\cdot f_t(c_i,w_j)-\sum_{i=1}^{N}\left(\log\left(\sum_{j=1}^{N}f_t(c_i,w_j)\right)\cdot\sum_{j=1}^{N}f_t(c_i,w_j)\right) \quad (3.16)$$
$$-\sum_{j=1}^{N}\log(g_t(w_j))\cdot g_t(w_j)\leq\kappa_t\,,\quad t=1,\ldots,T$$

where the second term is the entropy of the marginal distribution of consumption that results from choices in the joint.

3.4.5 The Wealth Transition

The challenge in the dynamic RI framework is the transition of the state variable. Because the state variable in the general framework is a distribution, as is the choice variable $f_t(c_i, w_j)$, determining the next period's distribution of wealth is a matter of determining the probability of being at each potential wealth node.

The wealth distribution is fully determined by the joint distribution of consumption and wealth in the past period and the distribution of per-period income in the current period. Current income is independent of prior wealth and is received prior to any consumption decision. That is, the timing works as follows: The choice variable in period $t-1$, $f_{t-1}(c_i, w_j)$, is combined with the current period's income distribution, $b_t(e)$ during the working portion of the agent's life, to determine his current marginal distribution of wealth. In the retired portion of the agent's lifetime, the consumption-wealth joint distribution in the previous period fully determines the wealth marginal distribution in the current period. The equation for the transition of the marginal distribution of wealth during the employed portion of the lifetime is

$$g_t(w_j) = \sum_{r=1}^{\min(K,j)} \left[b_t(e_r) \cdot \sum_{p=j-r+1}^{D_{t-1}} f_{t-1}(c_{p-j+r}, w_p) \right], \quad t = 2, \ldots, R-1;\ j = 1, \ldots, N, \tag{3.17}$$

with the transition during retirement being given by

$$g_t(w_j) = \sum_{p=j}^{D_{t-1}} f_{t-1}(c_{p-j+1}, w_p), \quad t = R, \ldots, T\,; \quad j = 1, \ldots, N, \tag{3.18}$$

where K is the number of nodes in the income distribution, R is the first period in which the agent is retired, N is the number of grid-points in the support of both the consumption and wealth distributions. The wealth transition represented in equations (3.17) and (3.18) are facilitated by the choice of a specific "gridding" of the supports for the discrete distributions involved in the model.

The support for the wealth and consumption distributions in this model is a pair of identical, N-point grids that begin at zero and increase in equally spaced increments. The support for the income distribution is the first K nodes of the support of the wealth and consumption distributions. By using the same support for all the distributions, we keep the state space as small as possible for a given set of distributions. The transition from the current period wealth distribution ($g_t(w)$) to next period's wealth distribution ($g_{t+1}(w)$), given the choice of the joint distribution $f_t(c, w)$ and per-period income distribution $b_{t+1}(e|\alpha_t)$ is outlined in figure 3.10. The process is as follows: for a given node within the future wealth distribution (for example $w = 1$, in figure 3.10's example), we find the probability by looking for all the possible combinations of current wealth, consumption, and income that could

bring us to that point, and sum the probabilities of those events.

Figure 3.10 outlines the interaction between each of the three distributions involved in forming the distribution of wealth in a given time period: the state variable $g_t(w)$, the current income variable $b_{t+1}(e)$, and the choice variable $f_{t+1}(c, w)$. Note first how the period t distribution of wealth, $g_t(w)$, restricts possible forms of the period t joint distribution. The columns in the f matrix composed entirely of zeroes and surrounded by dashed boxes represent the restriction in the joint distribution due to the fact that the marginal distribution of wealth has no probability weight on those wealth values. Therefore, the choices available to the agent regarding $f_t(c, w)$ are as follows: for each wealth level w_j, the conditional distribution for consumption divides the weight from $g_t(w_j)$ among the feasible elements of $f_t(c_i, w_j)$. This is done in period t under the processing constraint that the mutual information of c and w not exceed κ_t, while also choosing α_t in order to balance the benefits of increased processing capacity with increases in expected future income. The choice of α_t fully determines the weights in $b_{t+1}(e)$ as well. Once $f_t(c, w)$ and $b_{t+1}(e)$ have been determined, we have determined the marginal distribution of wealth in period $t + 1$. The probability weights of period-t based decisions are combined thusly in order to determine the weights in $g_{t+1}(w)$: Take, for example, $g_{t+1}(w = 1)$. There are two and only two ways to get to $w_{t+1} = 1$ in a borrowing-constrained model: The first is to consume everything in period t and receive an income of one at the beginning of period $t + 1$. The second is to consume all but one unit in period t and receive no income at the beginning of period $t + 1$. Figure 3.10 focuses on the first of these two possibilities

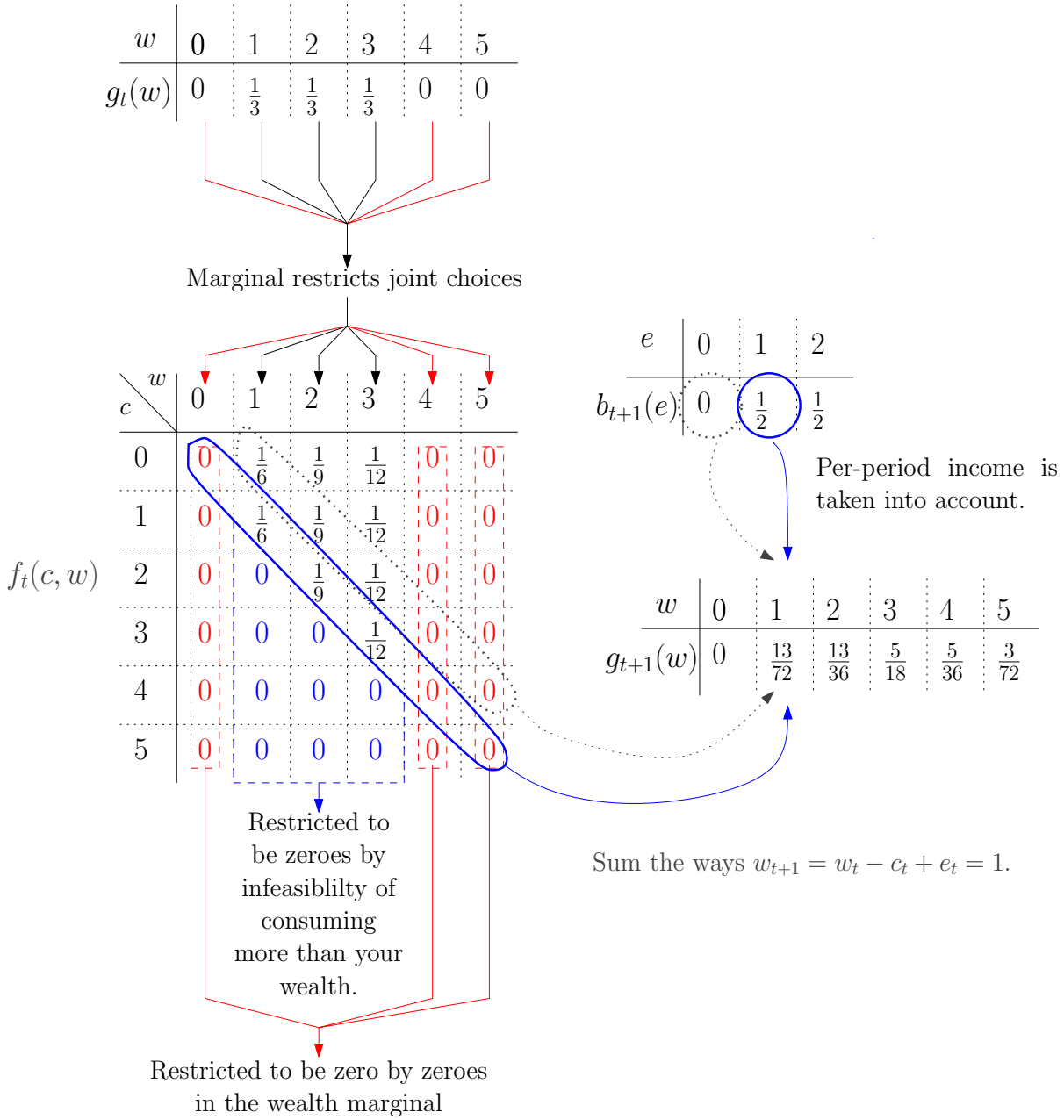


Figure 3.10: An Example of the Transition of Wealth From One Period to the Next.

by summing the probability of all the consumption-wealth combinations that leave behind zero post-consumption wealth in period t (the solid outline around the main diagonal), and multiplying that by the probability of receiving one unit of income in period $t+1$. A similar calculation characterizes post-consumption wealth of one and income of zero (the dotted outline around the first off-diagonal). These two calculations are then summed to arrive at the total probability of being at $w_{t+1} = 1$. It should be noted that the specific values in the f_t matrix of figure 3.10 in no way represent an optimal choice and are simply for illustrative purposes.

Regarding information concerning the next period's wealth, note that "...the agent must allow some noise to affect the choice of c in the current period, but can use the noisy observation that entered determination of c to update beliefs about next period's w" [(Sims, 2006, p. 18)]. This is accounted for in equations (3.17) and (3.18) and clearly reflected in figure 3.10. That is, the restriction on information processing in the current period constrains decisions regarding consumption, but also effects the information known about the following period's wealth. Where attention is focused by the agent (in the wealth-consumption space) in the current period will have an impact on the precision of the agent's wealth distribution in the following period(s).

3.4.6 Parameters and Initial Conditions

In the analysis that follows, the life-cycle is divided into $T = 8$ periods, where the agent is employed for six periods and retires in period $R = 7$. The life-cycle is assumed to begin during the working portion of life and we assume that it takes

place during ages 25-80, meaning that a model period is just under seven years and $\beta = 0.96^{\frac{55}{8}} \approx 0.76$. The initial state, $g_1(w)$, is assumed to be a flat one-period income distribution, with the exception that there is no weight placed on the $w_1 = 0$ node. We want to be able to analyze preferences with higher risk aversion, and therefore do not want to force agents to absorb zero consumption in the initial state. In future periods, there can be probability on a per-period income of zero, but agents with higher-γ values choose never to allow this to become a problem.

The value for N (the number of nodes in the wealth, and therefore also consumption, grid) is determined by the value for K (the number of nodes in the income grid) and the retirement age R in this model. In each period, the income received adds to potential existing wealth and the maximum possible value of wealth increases. Therefore, the total size of the wealth and consumption grids is given by $N = (R-2)(K-1)+K$, and in the figures that follow, $K = 9$, making $N = 49$.

3.4.7 The Solution Method

As in Lewis (2007a), this problem is handed over to a numerical optimizer. The optimization is performed using a combination of programs known as `AMPL` and `KNITRO`. `AMPL` is a front end for many powerful optimizers, one of which being `KNITRO`. By front end we mean the following: Problems are entered into `AMPL` via a very explicit system which essentially requires nothing more than copying the objective function in equation (3.12) as well as the constraints in equations (3.9), (3.10), (3.13)-(3.18) into a file exactly as they appear above. Once the problem has been described to `AMPL` it

performs what is called pre-solve, which looks at the problem and does what it can to eliminate complexity from a hill-climbing perspective by performing basic exercises such as solving for equality-constrained variables and so forth. Finally, `AMPL` performs what is known as *automatic* or *algorithmic* differentiation. The speed and accuracy of any optimizer depend on the information available about the hill being climbed. Automatic differentiation (AD) provides the gradients without the truncation errors of a procedure like 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.[9]

Given the specifications for K, R, and N above, the number of free parameters (5882, after accounting for adding-up and zero-restriction constraints) seems *very* large. Sims had 345 free parameters and needed 11 minutes. However, using `AMPL/KNITRO` on his problem [See section 3.2.4.3 above; see Lewis (2007a) for more detail regarding the numerical optimization issues.] required 1-2 seconds. The 5882-parameter problem of this section requires about 2 minutes. The problem itself is straightforward from a numerical optimization standpoint except for the pre-retirement transition of wealth probabilities, specifically elements pertaining to per-period income. With the exception of α_t, the problem is a very well-posed optimiza-

[9]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).

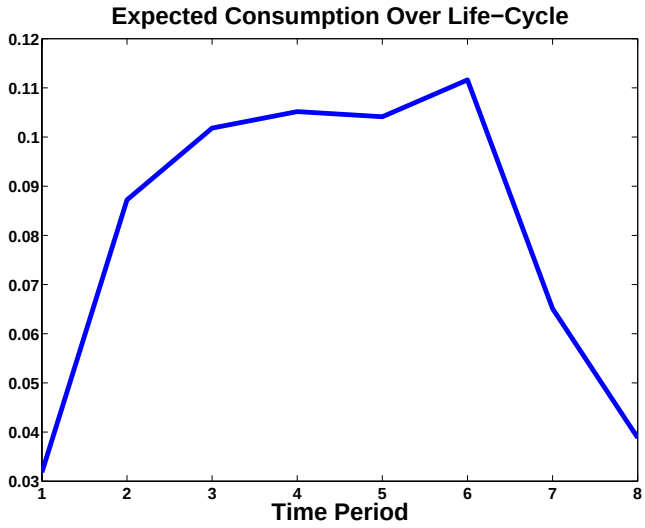


Figure 3.11: The Path of the Expectation of Consumption Over the Life-Cycle.

tion problem. The objective is linear, and the constraints would be convex if not for the effect of α on per-period income (that is, the trade-off variable accounts for eight of the 5800+ choice variables). Thus, the problem does not appear to be badly behaved. Several specifications of the model have been tested with dozens of random starting points for both $f_t(c, w)$ and α_t, and always within each specification, optima were identical across starting values.

3.5 Analysis and Results

The addition of uncertainty and information-processing constraints to the canonical model results in a clear hump-shape in the expected consumption path, as indicated in figure 3.11. The initial slope of the consumption hump is the result of "buffer-stock" style savings early in the life-cycle designed to protect again low wealth states in the future. The downward slope at the end is the result of a struggle

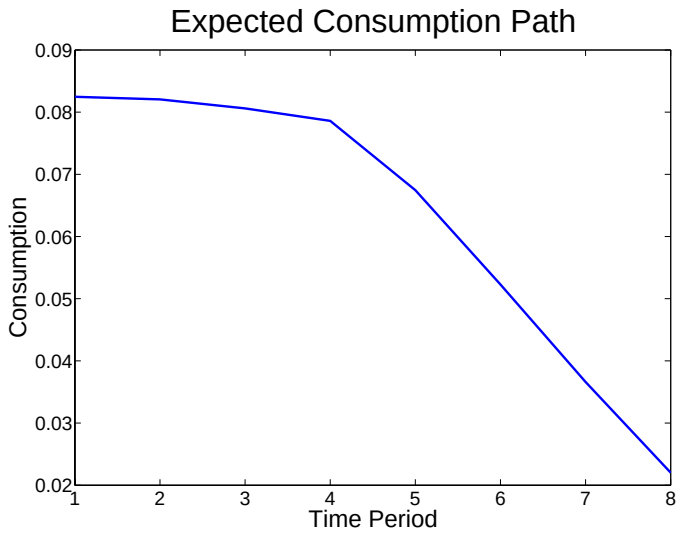
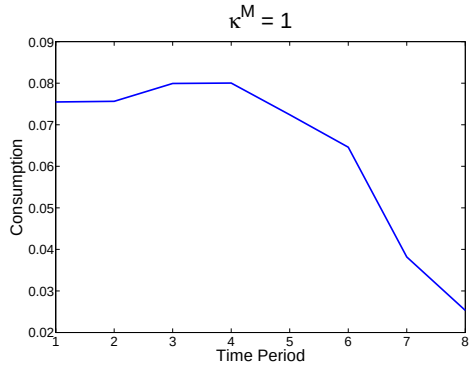


Figure 3.12: The Expected Path of Consumption for $\alpha_t = 1$, $\forall\, t$, and $\kappa^M = 10$.

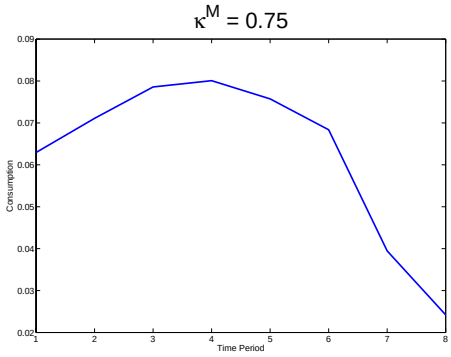
between the desire to consume as much as possible in each period and the desire to avoid a large probability of being left with zero wealth in the final period. This behavior is the result of unresolvable uncertainty on the part of the agent.

To understand where this hump-shape comes from, we examine the impact of decreasing the maximum information-processing capacity level, κ^M. A note about discounting and returns: in this model the gross rate of return is one making the value of $\beta R \approx 0.76$. As has been documented in the life-cycle literature [see, e.g. Yaari (1964)], the path of consumption in the canonical model will only be flat for $\beta R = 1$, while $\beta R > 1$ leads to growth and $\beta R < 1$ leads to decline. Therefore, when the α_t trade-off is eliminated and the infomation-processing capacity is made very large, we would expect the consumption path to decay, as indicated in figure 3.12.

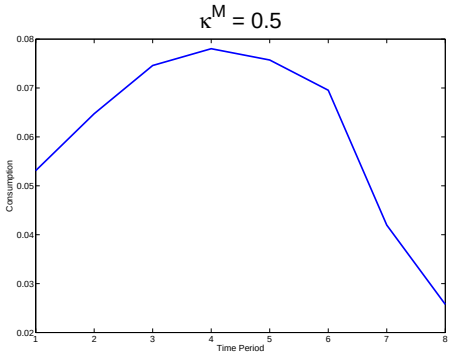
As the information-processing constraint is tightened, we see a clear hump-shape emerge in the expected consumption path, as seen in figure 3.13.



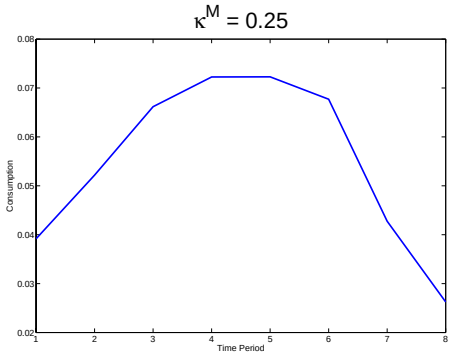
(a) Exp. Consumption Path, $\kappa^M = 1.0$



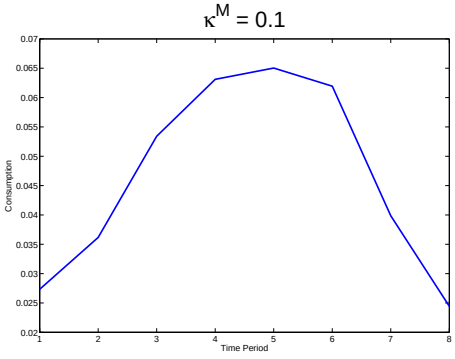
(b) Exp. Consumption Path, $\kappa^M = 0.75$



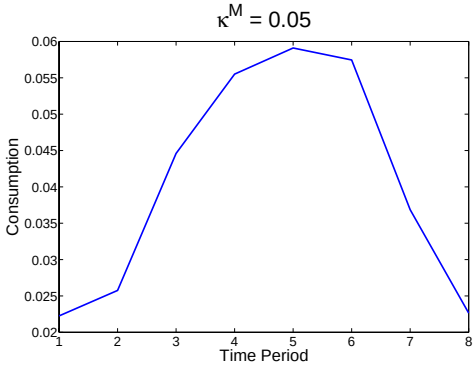
(c) Exp. Consumption Path, $\kappa^M = 0.5$



(d) Exp. Consumption Path, $\kappa^M = 0.25$



(e) Exp. Consumption Path, $\kappa^M = 0.1$



(f) Exp. Consumption Path, $\kappa^M = 0.05$

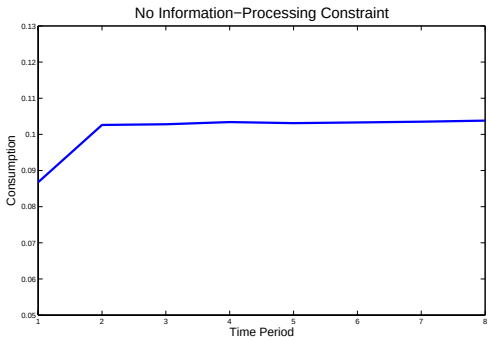
Figure 3.13: Expected Consumption Path for Decreasing Levels of Information-Processing Capacity, Given a Fixed $\alpha_t = 1$ for All Time Periods.

It is important to remember that the agent never has exact knowledge of his wealth or income. This unresolvable uncertainty due to information-processing constraints is what gives rise to the hump. Careful examination of the differences in the decay of consumption in closing time periods shows that while the unconstrained model (figure 3.12) has a decrease in consumption, the information-processing constraint clearly plays a role in the profile of the decrease (note the change in the profile of the decrease in panel 3.13(a) compared to figure 3.12).

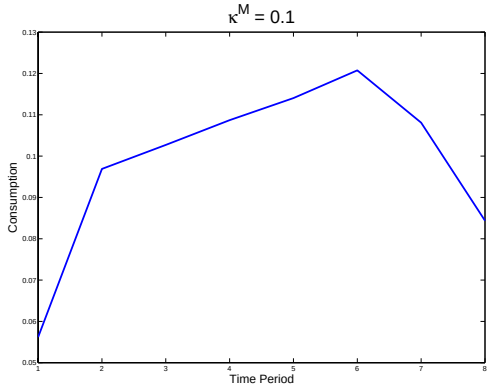
The downside of the hump is partially created by the fact that the no-processing-constraint model in which $\beta R < 1$ has a natural downward slope. However, we can see that the both sides of the hump clearly are produced by the information processing constraints even in the more standard universe of $\beta R = 1$, in figure 3.14.

The feature of the model that is important is that information-processing constraints create both precautionary saving and dis-saving. The agent in the model is scaling back his or her consumption more dramatically than they would in the event that they knew their wealth *exactly* and could control their consumption *precisely*. Figure 3.14 demonstrates that this "precautionary dis-savings" generates a downside to the consumption hump without the aid of the normal consumption decay implied by $\beta R < 1$.

The particular consumption path in figure 3.11 is based on an agent who has fairly low risk aversion: CRRA utility with $\gamma = 0.5$. As would be expected, the consumption behavior of the agent changes as risk aversion rises, as demonstrated in figure 3.15: more risk averse agents are more frugal early in life and consume more



(a) Expected Consumption Path, No Restrictions on Processing Capacity



(b) Expected Consumption Path, $\kappa^M = 0.1$

Figure 3.14: Expected Consumption Path for Decreasing Levels of Information-Processing Capacity, Given a Fixed $\alpha_t = 1$.

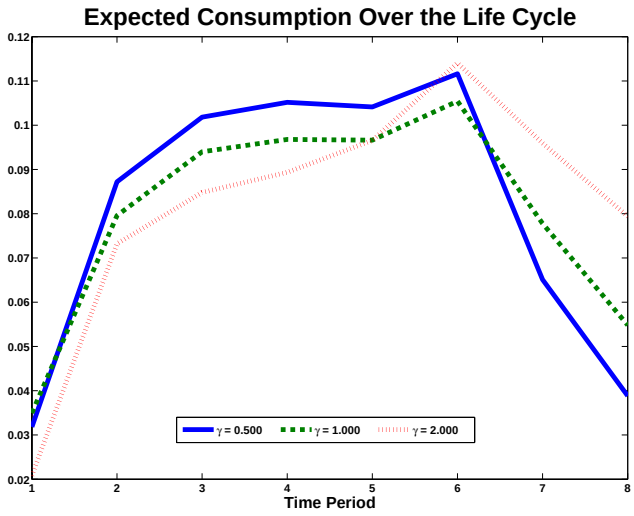


Figure 3.15: The Path of Expected Consumption Over the Life Cycle for Different Risk Preferences.

in retirement as a result.

From figure 3.16 it is seen that risk aversion changes result in very different behavior during the "employed" portion of the life cycle. Careful examination of the first-period marginal distribution of consumption reveals that the more risk averse agent ($\gamma = 2$) chooses to place probability on only the lowest positive consumption point, regardless of the level of wealth. The other two parameterizations, $\gamma = 0.5$ and $\gamma = 1$, spread probability across multiple potential consumption levels. The three initial consumption strategies summarized in the first panel of figure 3.16 require different levels of information-processing capacity. While the consumption strategy of the agent when $\gamma = 0.5$ requires some processing capacity (he keeps roughly a third of his potential capacity), the strategy of comsuming the lowest level possible requires almost no information-processing capacity, and the $\gamma = 2$ agent chooses α_1

	$\gamma = 0.5$	$\gamma = 2.0$
α_1	0.3	10^{-3}

Table 3.1: Values of Tradeoff Parameter, α, in the First Period.

accordingly, (see Table 3.1).

Figure 3.17 displays the joint distribution of consumption and wealth in the first period. Panel 3.17(a) is analogous to previous RI treatments in which the information-processing capacity of the agent is fixed exogenously. Panel 3.17(b) depicts the more general case in which the agent optimally chooses α_t. In each figure, as in the joint distribution figures of section 3.2, darker boxes indicate heavier probability weight, with the key to the value of the joint pdf given in the legend to the right of each distribution.

To understand the figure, consider the $\gamma = 2$ case in panel 3.17(a). This risk averse agent is choosing to save in the first period, regardless of wealth level. This is seen by the solid bar on the lowest positive consumption node, indicating that he is placing all his probability on consuming at the lowest non-zero level. With lower risk aversion ($\gamma = 0.5$), the agent in panel 3.17(a) chooses to place probability on higher levels of consumption for higher levels of wealth. For example, the consumption distribution conditional on the wealth level just above 0.1 places probability on five possible consumption points, with the majority of the weight being placed on 0.1. It can be seen in this panel that given lower risk aversion ($\gamma = 0.5$), the agent will assign

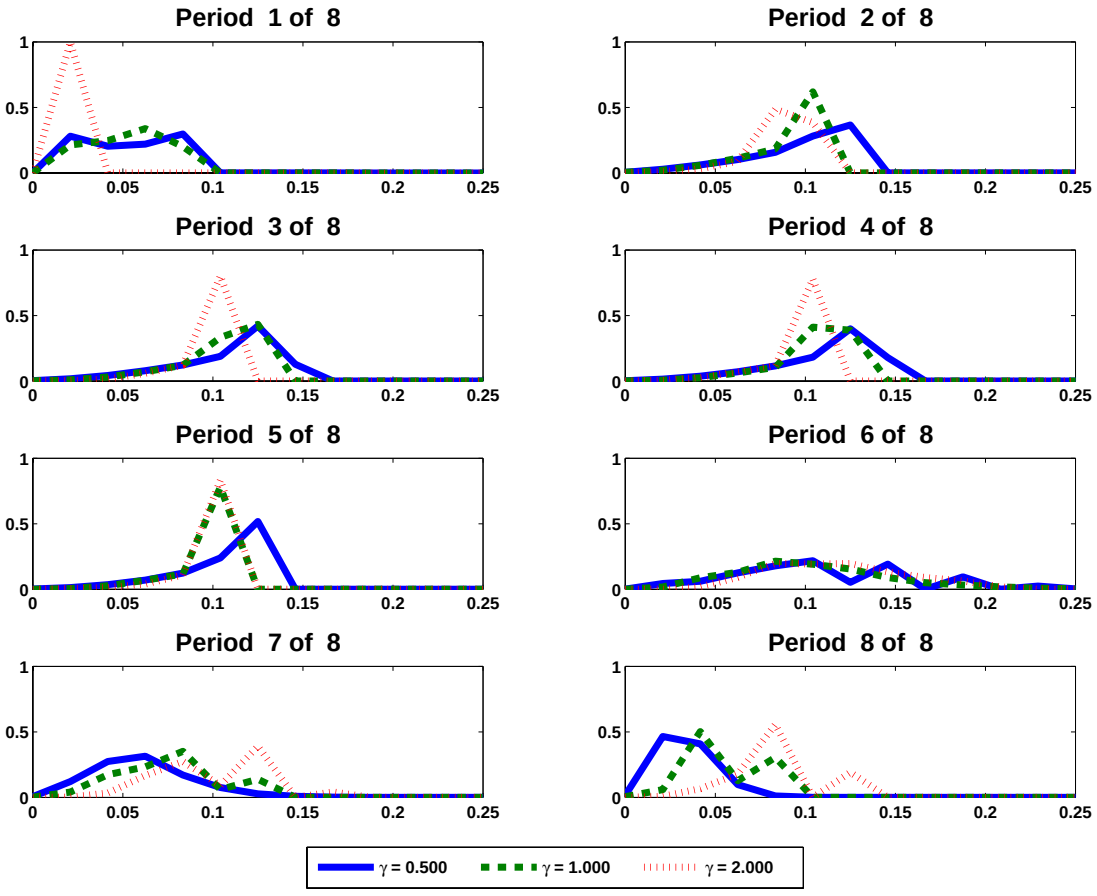
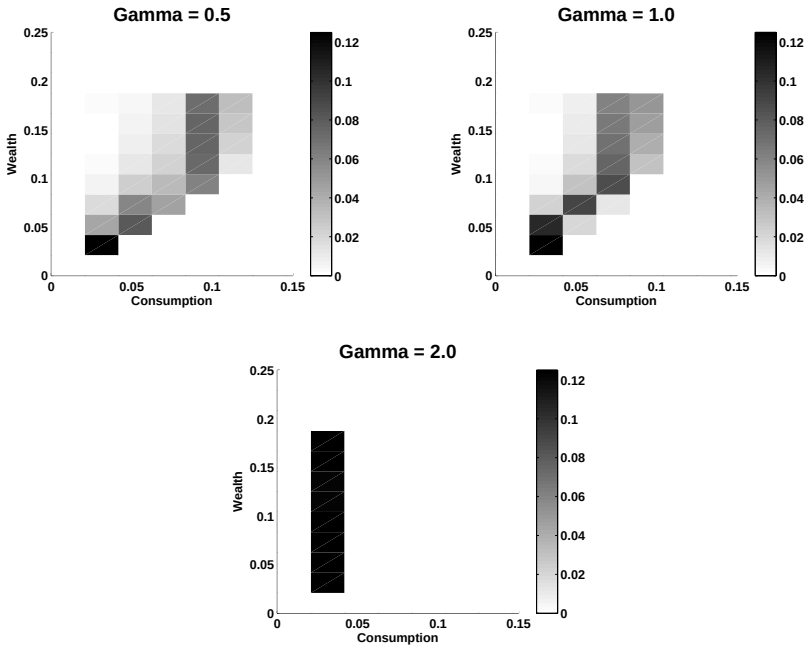
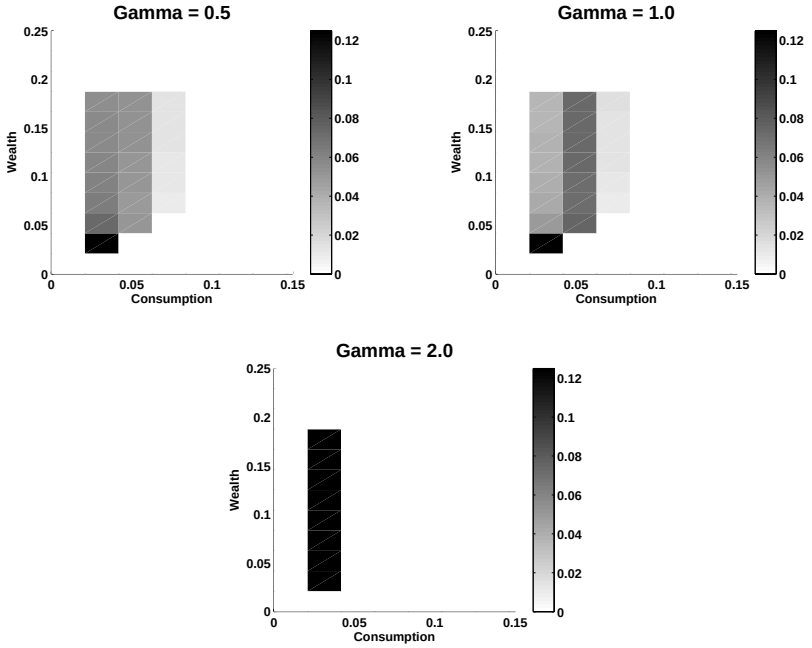


Figure 3.16: The Marginal Distributions of Consumption Throughout the Life-Cycle for Different Levels of γ.



(a) The fixed-α choice



(b) The flexible-α choice

Figure 3.17: The Choice of Joint Distribution in the First Period for the Fixed- and Flexible-α Cases.

positive probability to several consumption points that are on the feasibility boundary. This is because unlike the higher risk aversion parameterizations, consumption of zero is not penalized nearly as heavily.

The two panels in figure 3.17 allow analysis of the information-processing capacity and income tradeoff. By fixing $\alpha_1 = 1$ we can examine how the agent's behavior changes when he does not have the ability to trade information-processing capacity for expected income in period two. By looking at the $\gamma = 2$ panels of both figure 3.17(a) and 3.17(b), we see that the agent does not wish to use any of his available processing capacity. As a result, when offered the trade of information-processing capacity (something he does not need) for future income (something he wants very much), it is obvious that he will give up much of his processing capacity except for enough to know where the lowest positive consumption node is located. That is, the choice made by the risk averse agent when $\alpha_1 = 1$ is a choice that requires almost no information-processing capacity, so two cases are identical.[10]

The $\gamma = 0.5$ parameterization, on the other hand, requires the agent to balance current processing needs with future income desires. When α_1 is fixed, the agent places weight on consumption possibilities he would not consider when he is offered the choice of α_1. That is, when $\gamma = 0.5$, giving the agent the ability to reduce his processing capacity in exchange for increases in expected future income results in a

[10]Note that the only mutual information "connection" between c and w implied by the joint distribution in panel 3.17(a) concerns feasibility of the lowest positive consumption node. Beyond that, the distribution implies independence, thus the very low mutual information content and therefore information-processing requirement.

change in the behavior of the agent. From Table 3.1, we see that when given the opportunity, the agent will give up 70% of his information-processing capacity in exchange for improvements in his future income. The optimal allocation of "precision" changes when the total amount of precision to be allocated changes. First, note that the levels of wealth and their associated probabilities are identical in panels 3.17(a) and 3.17(b). That is, for example, the probability of $w = 0.025$ is the same in both panels. Next, note that the agent assigns probability to consumption possibilities when α_1 is exogenously-fixed at 1 that he ignores when α_1 is endogenously set to 0.3. Additionally, note that the probability is placed more heavily on the feasible boundary when α_1 is fixed, and that consumption appears to be more strongly correlated with wealth. That is, when α_1 is determined endogenously, the probability mass of consumption at each of the three possibilities is spread more uniformly across wealth levels when compared to the exogenously-fixed α_1 choice. These differences are all a result of using less information-processing capacity. First, when the amount of information-processing capacity decreases, the agent can compensate by paying attention to fewer things. This is accomplished by considering a smaller set of consumption possibilities. By eliminating higher consumption levels, the agent is able to "spend more" of his information-processing capacity on the remaining three levels. In addition to this, the remaining consumption-wealth pairings will have less precise conditional distributions. The problem can be thought of as one of allocating a total pool of precision, or attention. First, the agent can reduce the number of (c, w) pairings over which he is trying to be precise, and then lower the precision of pairings

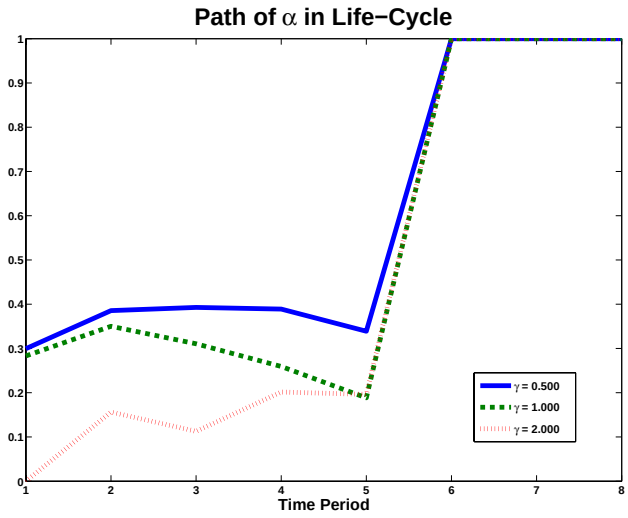


Figure 3.18: The Path of the Choice Variable α_t Over Time for Different Risk Preferences.

still to be considered. The mathematics of the problem boil down to how much "correlation" (more precisely mutual information) can be represented in the optimally chosen joint distribution. As the agent reduces information-processing capacity, the optimal choice of the joint distribution must imply a weaker relationship between c and w. It is important to remember that the agent is choosing to give up the information processing capacity that represents the difference between panels 3.17(a) and 3.17(b). The information-processing capacity is being traded for future income prospects, meaning that the agent is acting optimally when he chooses to move from panel 3.17(a) to 3.17(b).

The optimal path of α is given in figure 3.18. Returning to the higher risk aversion specification ($\gamma = 2$), the amount of information-processing required for the first period is nearly zero, but the second period makes use of a significant amount.

This is done for two reasons, the first of which is that it is no longer optimal to save nearly all income in the second period. The "buffer-stock" accumulation of the first period gives way to a lower marginal expected savings rate. Second, the space of potential wealth levels doubles from period 1 to period 2.[11] These two effects combine to make the optimal choice for the high risk aversion parameterization essentially a coin-flip over two higher levels of consumption than he consider in the first period (see figure 3.16). It should also be noted that period 2 represents the period in which the agent places probability on the highest consumption node he will consider during periods 1-5. This statement must be differentiated from describing the expected consumption path, which clearly continues to rise in subsequent periods for the high risk aversion agent. What is meant here is that the agent solves his attention allocation problem in such as way as to place weight on higher levels of consumption in period 2 than he did in period 1, and that he does not continue to consider even higher levels again in period 3. That is, the agent has reached the levels of consumption sustainable given his lifetime expected income by period 2. The next few working periods (3-5) are a process of fine-tuning the consumption choice, as seen by the fact that the $\gamma = 2$ parameterization places the majority of the weight on a single consumption level in periods 3 through 5.

The low risk aversion parameterization ($\gamma = 0.5$) also needs more processing

[11]The agent begins with an initial wealth level equal to a flat one-period income distribution with no weight on zero. When the agent moves to the second period he gets his K-node period income distribution whose lowest node is zero. As a result, the first period had K wealth levels while the second period has $K + (K - 1)$ wealth levels.

power in the second time period due to the increase in the size of the wealth space. However, when $\gamma = 0.5$, the agent's consumption choices do not need to grow as much relative to the first period as the higher risk aversion parameterizations, so his need for information-processing capacity growth is much less sharp. Still, he continues to use more processing power than his high risk aversion counterpart because he wants to be able to consider more points than the high risk aversion parameterization.

To explain why the low risk aversion agent wishes to consider a broader range of consumption nodes than the higher risk aversion agent for a given wealth distribution, we examine figure 3.19, the joint distributions in period 6 – the period just before retirement. Before progressing, it must be clearly understood that the α-path for all parameterizations goes to one for periods 6-8 because of the tradeoff used in the model. Agents trade information processing capacity for benefits in their next period's income distribution. At retirement, the agent stops receiving income, so because there is no income in period 7 or 8, there is no reason to spend time trying to improve it in periods 6 and 7 (there is no "future" following period 8, so similarly there is no incentive to give up any processing capacity). Beginning in the period directly before retirement, the agent no longer has anything to trade for his processing capacity, and as a result, the amount of information processed in period six ($\alpha_6 = 1$) is nearly three times that of previous periods. One aspect of this model is that the agent anticipates this increase in processing power and is waiting to increase consumption once it arrives.

To return to the question of how risk aversion impacts the range of consump-

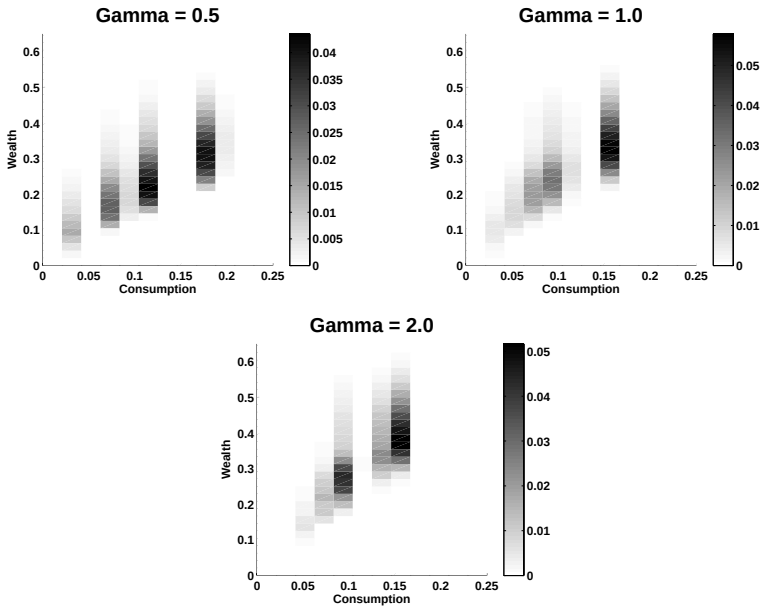


Figure 3.19: The Choice of the Joint Distribution of Consumption and Wealth in the Sixth Time Period.

tion levels considered given a wealth distribution, figure 3.19 demonstrates that while the anticipated explosion of processing power does, in fact, cause agents with each parameterization to expand their attention considerably, a central difference between high and low risk aversion can be seen. Low-γ agents spread their attention more broadly, strategically spacing consumption allocation so that they can cover more of the consumption-wealth space with some accuracy, while high-γ agents focus more intently on a smaller number of points, paying special attention to the lower levels to guarantee precise behavior there. For example, in Sims's undiscounted two-period case discussed earlier, given a marginal distribution for wealth, it is clear that the optimal unconstrained choice would be to choose $c = w/2$ for every w. It was demon-

strated by Sims that, in the information-processing constrained world (his figures 5 and 6), the less risk averse agents will spread their attention over a larger region of the (c, w) space, discretely, so as to generate *adequate* consumption over a broader range of wealth. That is, the consumption conditionals are centered around the $c = w/2$ optimum but give up tightness around the optimum and careful examination of lower consumption nodes for the ability to focus attention on consuming at higher levels when wealth is, in fact, high. The agent has a total amount of "preciseness" that can be used. He could, for example, be very precise around a few levels of wealth by forming very tight conditional distributions around $c_i = w_j/2$ for several w_j's. Or, he could be very imprecise around every w_j. Where and how the agents wish to be precise is a function of their preferences. More risk averse agents give up higher consumption in the high wealth state for the ability to consume more accurately at lower wealth levels, meaning tighter distributions around the optimum and fewer gaps in attention overall at lower consumption levels. They do this because they are concerned about consumption-wealth mismatches at lower wealth levels and want to be able to consume everything up to their boundary in these cases. Further, they are willing to pay the price of moderate consumption in high wealth states in order to do so.

This result, that lower risk aversion agents will spread their attention across more consumption possibilities, combined with the fact that they reach a sustainable level of consumption by period 2, accounts for the relatively flat nature of the α-path for $\gamma = 0.5$ in figure 3.18. Similarly, the high risk aversion agent has more spread-out

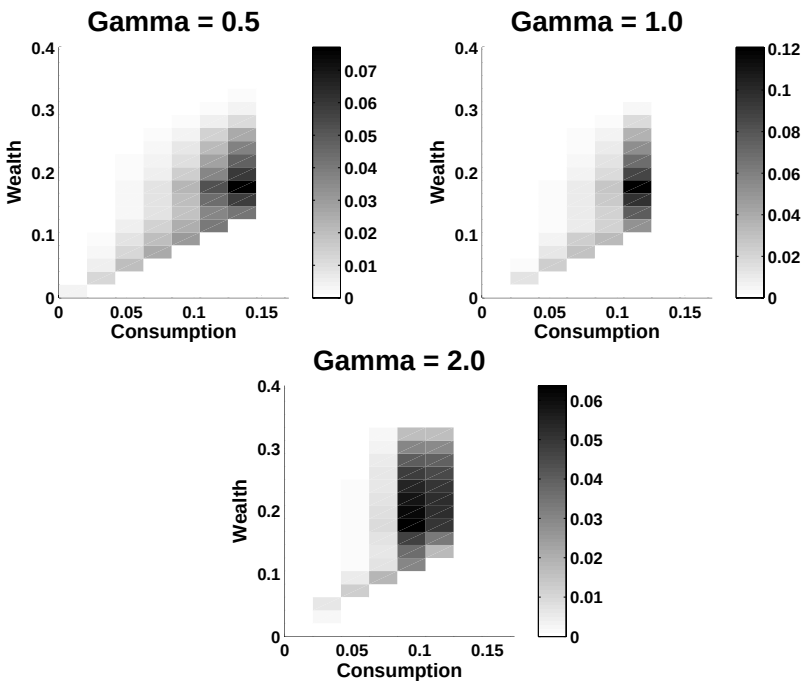


Figure 3.20: The Choice of the Joint Distribution of Consumption and Wealth in the Second Time Period.

consumption behavior in period 2 because by saving heavily in the first period, he has increased his potential wealth to a point where there is very little probability of a low wealth state, therefore it does not take a large amount of information-processing capacity to have tight conditionals at the few wealth levels he considers to be "low." (Note the tight conditional distributions for consumption given the lower three wealth levels in figure 3.20.) Because he is able to behave cautiously at low wealth levels at a fairly low "attention" cost, he is able to consider a consumption lottery that is essentially a coin flip over two higher consumption levels.

In exchange for giving up processing capacity, the agent receives the same thing in each time period: improvement in the next period's income distribution.

Therefore, the variations in α must represent variations in the value of information-processing over time. Just as consumption-smoothing is the optimal result of the canonical model, there could be informational smoothing in inattention models as well. Future research will include an examination of differences in the marginal value of processing capacity that could cause a sort of information smoothing over time that could explain the type of "over-shooting" observed in the α-path of the log-preferences agent in figure 3.18.

Figure 3.21 indicates that the wealth distributions are similar across the three γ specifications. As wealth grows, the distributions spread out and as agents dissave, their wealth distributions collapse on a small number of points. When an agent reaches the point of retirement, he wants to eat from his savings as much as possible, but there is a struggle between not wanting to leave anything on the table and not wanting to have a high probability of zero consumption in the final period. As can be seen in the final pane of figure 3.21, each wealth distribution collapses on a point, though the point is different for each parameterization. The distance of that "collapsing point" from the origin and the shape around the point is a function of the agent's preferences regarding zero consumption: more highly-risk-averse agents bring their final wealth distribution to a sharper point than their less cautious counterparts, and that point is shifted to the right to ensure safety regarding zero-consumption. The "sharpness" of the final wealth distribution is increasing in γ because of the effect described earlier regarding how agents allocate their attention. Because of the focus of the higher risk aversion parameterization on a smaller number of points,

these agents tend to eliminate probability weight from certain regions in the wealth distribution while leaving other regions untouched. This behavior is different from lower risk aversion parameterizations which tend to eliminate some probability from a large number of nodes rather than all probability from a small number of nodes. As a result, highest probability weight in the high risk aversion parameterization is higher than it's counterpart in the lower risk aversion parameterizations.

As a result of the inability to eliminate uncertainty, we observe in this model what have been called "accidental bequests" [see e.g. Hendricks (2002)]. These bequests result not from uncertainty regarding time of death, but from uncertainty arising from an inability to process all the information available. As is seen in figure 3.22, the expected bequest is increasing in γ. Note that according to figure 3.22, an agent with CRRA preferences and $\gamma = 2$ has a 20% probability of leaving behind a bequest at least as large as a full period's consumption (just under seven years' worth).

3.6 Conclusions

Following the two-period model of Sims (2006), this paper presents a simple life-cycle framework for addressing the optimal allocation of attention over time. The framework is fully scalable in a finite-horizon model and could be used to study optimal behavior under processing constraints in more general economic environments. In the framework studied here, the value of information-processing capacity varies over time, and the agent's degree of risk aversion plays a significant role in determining

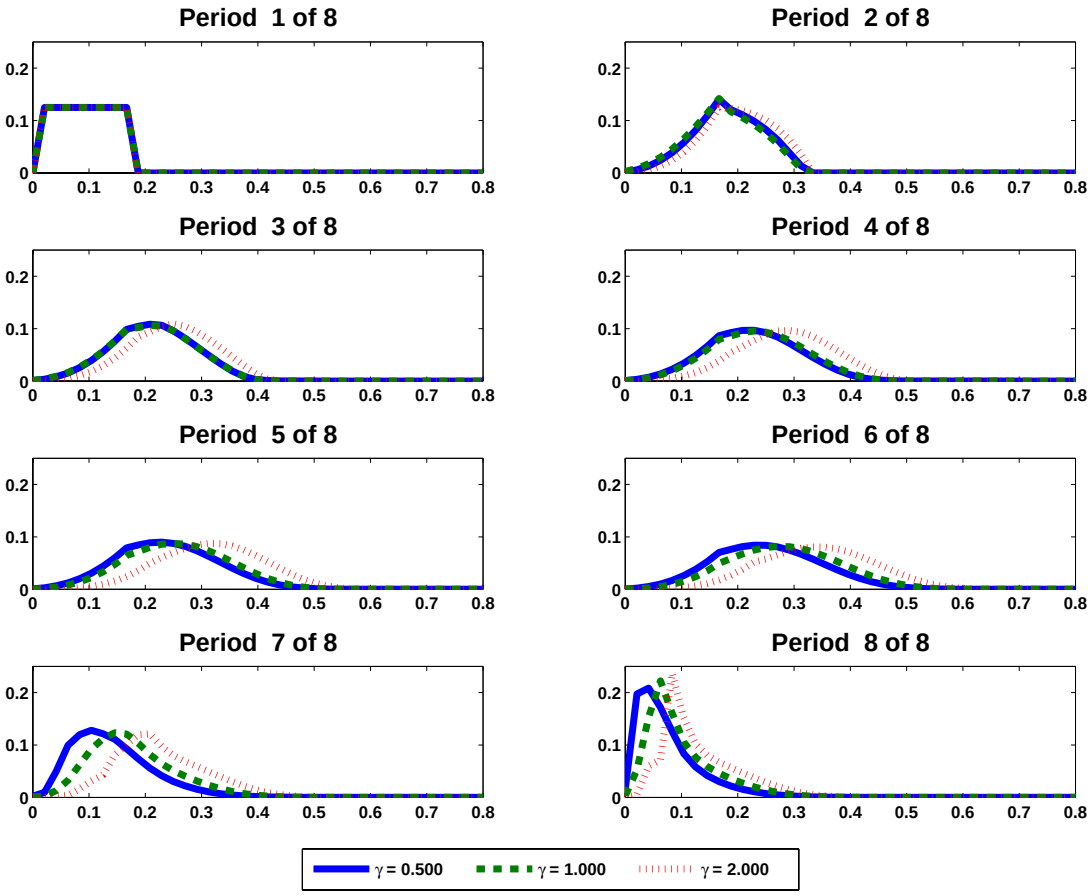


Figure 3.21: The Marginal Distributions of Wealth Throughout the Life-Cycle for Different Levels of γ.

that value. Life cycle agents with finite information-processing capacity display the hump-shape pattern of consumption observed so frequently in the data. Additionally, the struggle between wanting to consume as much as possible and wishing to avoid zero consumption can lead to a high probability that an agent will leave behind non-trivial wealth at death, thus generating an "accidental bequest" that is solely an artifact of imprecise knowledge.

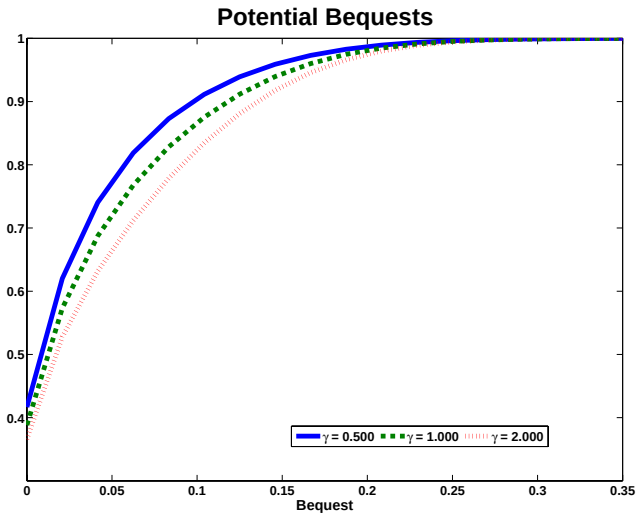


Figure 3.22: The CDF of the Bequest Distribution for Different Levels of γ.

APPENDIX A
AN ALTERNATIVE SOLUTION PROCEDURE FOR THE TWO-PERIOD PROBLEM

A.1 The Iterative Procedure

It should be noted that there is an alternative solution procedure for this specific 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, we can assume $f_{i,j} > 0$ because $\lim_{x \to 0} x \log(x) = 0$, and therefore the first-order conditions for this problem can be written:

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

where

$$U_{i,j} = \frac{c_i^{1-\gamma} + (w_j - c_i)^{1-\gamma}}{1-\gamma}, \tag{A.2}$$

Specifying a value for λ (the multiplier on (2.5)) is identical to choosing a κ (as you would expect, 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 (2.3) for a specific j) is all that is required before we are able to solve for the probabilities. This is done by taking advantage of the log-properties inherited from the entropic constraint.

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

$$\implies \frac{f_{i,j}}{\sum_{j=1}^{N_w} f_{i,j}} = \exp\left[(1/\lambda)(U_{i,j} - \omega_j)\right] \tag{A.3}$$

At this point, it should be noted that the denominator of the LHS of (A.3) represents the marginal probability of c_i, which we will call $p(c_i)$. Thus,

$$f_{i,j} = \exp\left[(1/\lambda)(U_{i,j} - \omega_j)\right] p(c_i) = \frac{\exp[(1/\lambda)U_{i,j}]}{\exp[\omega_j/\lambda]} p(c_i). \tag{A.4}$$

Equation (2.3) yields that:

$$\sum_{i=1}^{N_c} f_{i,j} = g(w_j) = \frac{1}{\exp[\omega_j/\lambda]} \sum_{i=1}^{N_c} \exp[(1/\lambda)U_{i,j}]p(c_i)$$

$$\implies \frac{1}{\exp[\omega_j/\lambda]} = \frac{g(w_j)}{\sum_{i=1}^{N_c} \exp[(1/\lambda)U_{i,j}]p(c_i)}. \tag{A.5}$$

Therefore, we (almost) have a solution for the probabilities:

$$f_{i,j} = \frac{\exp[(1/\lambda)U_{i,j}]p(c_i)g(w_j)}{\sum_{i=1}^{N_c} \exp[(1/\lambda)U_{i,j}]p(c_i)}. \tag{A.6}$$

Equation (A.6) is the solution, but recall that $p(c_i) = \sum_{j=1}^{N_w} f_{i,j}$.[1] The procedure is completed via iteration on f. Starting with a random f matrix and generating a

[1]That is, equation (A.6) 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 (A.6) true?"

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 (A.6). 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.

This procedure derives identical solutions to the procedure outlined above using `AMPL`, and without using sophisticated optimizers. That being said, it is slower and it cannot be implemented in general. This problem has an undiscounted utility function and it's static nature make FOC-based analysis possible.

A.2 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 (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, infinite). The low-λ, high-U 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.[2] The bad news is that $\exp[(1/\lambda)U_{i,j}]$ *is* central to RI theory and therefore present in RI models 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, the problem.

[2]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.

REFERENCES

S. Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving. *The Quarterly Journal of Economics*, 109(3):659–684, August 1994.

Michael R. Bussieck and Armin Pruessner. Mixed-integer nonlinear programming. *SIAG/OPT Newsletter: Views & News*, 14(1), February 2003. Available at: http://www.gamsworld.eu/minlp/siagopt.pdf.

Richard Caballero. Consumption puzzles and precautionary savings. *Journal of Monetary Economics*, 25(1):113–136, January 1990.

John Y. Campbell and Robert J. Shiller. Cointegration and tests of present value models. *Journal of Political Economy*, 95:1062–1088, 1987.

Christopher D. Carroll. The buffer-stock theory of savings: Some macroeconomic evidence. *Brookings Papers on Economic Activity*, 23:61–156, 1992.

Christopher D. Carroll. How does future income affect current consumption? *The Quarterly Journal of Economics*, 109(1):111–147, February 1994.

Christopher D. Carroll. Buffer-stock saving and the life-cycle/permanent income hypothesis. *The Quarterly Journal of Economics*, 112(1):1–55, February 1997.

Christopher D. Carroll. Theoretical foundations of buffer-stock savings. Johns Hopkins University, Department of Economics Working Paper, November 2004.

George M. Constantinides. Habit formation: A resolution of the equity premium puzzle. *The Journal of Political Economy*, 98(3):519–543, June 1990.

Angus S. Deaton. Saving and liquidity constraints. *Econometrica*, 59(5):1221–1248, September 1991.

Angus S. Deaton. *Understanding Consumption*. Clarendon / Oxford University Press, 1 edition, 1992.

Douglas N. DeJong and Charles H. Whiteman. The temporal stability of dividends and stock prices: Evidence from the likelihood function. *American Economic Review*, 1991.

Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, November 1961.

Larry G. Epstein and Stanley Zin. Substitution, risk aversion, and the temporal behaviorof consumption and asset returns: A theoretical framework. *Econometrica*, 57(4):937–969, 1989.

Jesús Fernández-Villaverde and Dirk Krueger. Consumption over the life-cycle: Facts from consumer expenditure survey data. Working Paper, University of Pennsylvania, September 2004.

Marjorie Flavin. Excess volatility in the financial markets: A reassessment of the empirical evidence. *The Journal of Political Economy*, 91:929–956, 1983.

Craig R. Fox and Amos Tversky. Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, 11(3):585–603, August 1995.

David M. Gay. Automatic differentiation of nonlinear ampl models. *AT&T Bell Labratories Numerical Analysis Manuscript*, 91-05, August 1991.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *The Journal of Mathematical Economics*, 18:141–153, 1989.

Pierre-Olivier Gourinchas and Jonathan A. Parker. Consumption over the life cycle. *Econometrica*, 70(1):47–89, January 2002.

Andreas Griewank. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Number 19 in Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, 2000.

James D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

Lars P. Hansen and Ravi Jagannathan. Implications of security market data for models of dynamic economies. *Journal of Political Economy*, 99:225–262, 1991.

Lars P. Hansen and Thomas J. Sargent. Formulating and estimating dynamic linear rational expectations models. *The Journal of Economic Dynamics and Control*, 2: 7–46, 1980.

Lars P. Hansen and Thomas J. Sargent. *Misspecification in Recursive Macroeconomic Theory*. Princeton University Press, 2005.

Lutz Hendricks. Intended and accidental bequests in a life-cycle economy. Arizona State University Working Paper, February 2002.

Kenneth Kasa. A robust hansen-sargent prediction formula. *Economic Letters*, 71: 43–48, 2001.

Stephen F. LeRoy and C. J. LaCivita. Risk aversion and the dispersion of asset prices. *The Journal of Business, Vol. 54, No. 4. (Oct., 1981), pp. 535-547*, 54(4):535–547, October 1981.

Stephen F. LeRoy and Richard D. Porter. The present value relation: Tests based on implied variance bounds. *Econometrica*, 64:555–574, 1981.

Kurt F. Lewis. The two-period rational inattention model: Accelerations, additions and analyses. Working Paper, University of Iowa, June 2007a.

Kurt F. Lewis. The life-cycle effects of information-processing constraints. Working Paper, University of Iowa, June 2007b.

Yulei Luo. Consumption dynamics, asset pricing, and welfare effects under information processing constraints. Discussion paper, Princeton University, 2004.

Yulei Luo. Consumption dynamics under information processing constraints. Princeton University Working Paper, May 2005.

Terry A. Marsh and Robert C. Merton. Dividend variability and variance bounds tests for the rationality of stock market prices. *American Economic Review*, 76: 483–498, 1986.

Rajnish Mehra and Edward C. Prescott. The equity premium: A puzzle. *The Journal of Monetary Economics*, 15(2):145–161, March 1985.

Ronald W. Michener. Variance bounds in a simple model of asset pricing. *The Journal of Political Economy*, 90(1):166–175, Feb. 1982.

Franco Modigliani. Life cycle, individual thrift, and the wealth of nations. *American Economic Review*, 76(3):297–313, June 1986.

Christopher Otrok, B. Ravikumar, and Charles H. Whiteman. Habit formation: A resolution to the equity premium puzzle? *The Journal of Monetary Economics*, 49:1261–1288, 2002.

Louis B. Rall. *Automatic Differentiation: Techniques and Applications*, volume 120 of *Lecture Notes in Computer Science*. Springer-Verlag, 1981.

Ricardo Reis. Inattentive producers. *Review of Economic Studies*, 73(3):793–821, 2006.

Ricardo Reis. Inattentive consumers. *Journal of Monetary Economics*, forthcoming.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

Robert J. Shiller. Do stock prices move too much to be justified by subsequent changes in dividends? *The American Economic Review*, 71(3):421–436, June 1981.

Herbert A. Simon. Rationality as process and as product of thought. *The American Economic Review*, 68(2):1–16, May 1978. Papers and Proceedings of Ninetieth Annual Meeting of the American Economics Association.

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

Christopher A. Sims. Rational inattention: A research agenda. Working paper, Princeton University., March 2006.

Kenneth D. West. Dividend innovations and stock price volatility. *Econometrica*, 56 (1):37–61, January 1988.

Charles H. Whiteman. *Linear Rational Expectations: A User's Guide*. University of Minnesota Press, Minneapolis, MN, 1983.

Charles H. Whiteman. Spectral utility, wiener-hopf techniques, and rational expectations. *The Journal of Economic Dynamics and Control*, 9:225–240, 1985.

Charles H. Whiteman. Analytical policy design under rational expectations. *Econometrica*, 54(6):1387–1405, November 1986.

Menahem E. Yaari. On the consumer's lifetime allocation process. *International Economic Review*, 5(3):304–317, September 1964.

G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 16:301, 1981.