DISCRETE ACTIONS IN INFORMATION-CONSTRAINED DECISION PROBLEMS

JUNEHYUK JUNG, JEONG-HO KIM, FILIP MATEJKA, AND CHRISTOPHER A. SIMS

ABSTRACT. Changes in economic behavior often appear to be delayed and discontinuous, even in contexts where rational behavior seems to imply immediate and continuously distributed reactions to market signals. One possible explanation is the presence of information-processing costs. Individuals are constantly processing external information and translating it into actions. This draws on limited resources of attention and requires economizing on attention devoted to signals related to economic behavior. A natural measure of such costs is based on Shannon's measure of "channel capacity". Introducing information costs based on Shannon's measure into a standard framework of decision-making under uncertainty turns out to imply that discretely distributed actions, and thus actions that persist across repetitions of the same decision problem, are very likely to emerge in settings that without information costs would imply continuously distributed behavior. We show how these results apply to the behavior of a risk-averse monopoly price setter and to an investor choosing portfolio allocations, as well as to some mathematically simpler "tracking" problems that illustrate the mechanism. Interpreting the behavior in our examples ignoring information costs and postulating fixed ("menu") costs of adjustment would lead to mistaken conclusions.

I. INTRODUCTION

Rational inattention (RI) theory models individuals as having finite information-processing "capacity" in the sense of Shannon (see MacKay (2003); Cover and Thomas (1991) for textbook treatments). It is a theory about why, when information appears to be available at little or no cost, individuals may not respond to it, or may respond erratically. It has intuitive appeal — most of us, most days, do not look up the term structure of interest rates and make corresponding fine adjustments in our checking account balances, even though, with an internet connection this could be done very easily. RI also has qualitative implications about delay and noise in reactions to information that roughly match empirical relationships among macro variables.[1]

Prices of individual products in most markets do not change continually, but instead stay fixed for spans of time, then jump to new values. We have simple theories that imply such behavior of prices is optimal (menu costs models) or treat such behavior as a constraint (Calvo pricing). As fine-grained microdata on individual product prices has become available, however, we can see that in at least some markets (e.g. grocery stores) prices not only stay fixed for spans of time, when they do change they sometimes move back and forth across a finite array of values. Our simple models explain the fixity for spans of time, but on their face imply that when a price change occurs, the change should be continuously distributed. We don't explain why the price should change, then come back to *exactly the same* price as before the change, for example.

Though stickiness of prices has received special attention in the literature, inertial economic behavior appears in many other areas. Macroeconomic modelers find

---

[1]Rational inattention was introduced in Sims (2003). It has mostly been applied to model pricing and consumption choices (Maćkowiak and Wiederholt, 2009b,a; Woodford, 2009; Luo, 2008; Matějka, 2008; Tutino, 2009), or portfolio allocation (Van Nieuwerburgh and Veldkamp, 2010; Mondria, 2010). Other applications are in theory (Moscarini, 2004; Yang, 2015; Matějka and McKay, 2015; Caplin and Dean, 2015).

they need to postulate "habit" in consumption to account for slow responses of consumption spending and costs of adjustment in investment spending to account for slow movements in investment. As Sims (2003) pointed out, most macroeconomic variables do not have time paths as smooth as would be implied by the strong adjustment costs needed to explain their slow responses to the state of the economy. Rational inattention accounts for this stylized fact.[2] Bonaparte and Cooper (2009) observe that 71% of households do not adjust their common stock portfolios over the course of a year, and that this is difficult to explain with adjustment cost models. In this paper we show that such stickiness in portfolios can emerge from information- processing costs.

In a simple model he interpreted as a two-period savings problem, Sims (2006a) found that numerical solutions for optimal behavior, even when exogenous randomness was continuously distributed, implied discretely distributed behavior. Matějka (2008) explored a model he interpreted as describing a Shannon-capacity constrained monopolistic seller with random costs and showed there again that numerical solution tended to imply discretely distributed behavior. In fact, the time paths of prices emerging from his simple model matched many of the qualitative features of individual product time paths shown in, e.g., Eichenbaum, Jaimovich, and Rebelo (2008).

Matějka (2009) considered the behavior of a monopolistic price setter that is not information-constrained, facing consumers who have finite capacity. The consumers will choose discretely distributed behavior, and this turns out to imply that it is optimal for the seller to set discretely distributed prices.

The question of whether "stickiness" reflects something like menu costs, or instead rational inattention, is important for economic policy modeling. If stickiness reflects RI, its form will change systematically if the stochastic process followed by

---

[2]The intuition and this qualitative match are discussed at more length in Sims (2003).

the environment changes. If the stickiness has mistakenly been modeled as due to a stable adjustment cost in a standard rational expectations model, the adjustment costs will appear to change as the stochastic process followed by the economy changes. RI thus implies that rational expectations models, developed to explain how a change in policy behavior could change the non-policy part of a model, are themselves subject to a similar critique.

Perhaps more important, models that explain stickiness via adjustment costs of one sort or another imply that rapid change in the sticky choice variables is costly or distasteful. RI models, which explain stickiness as reflecting information processing costs, do not imply that rapid change is in itself costly or distasteful. On the other hand, they imply that there is a cost to processing information that existing theories do not take into account. If, for example, an environment of high inflation requires individual consumers and producers to devote more attention to tracking prices, there may be a cost that is not captured in the behavior of direct arguments of production and utility functions. RI therefore does not necessarily generically imply that cyclical fluctuations are more or less important than in adjustment-cost models, but it could imply quite different estimates of welfare costs — and thus different conclusions about optimal policy.

In this paper we consider a generic problem of decision-making under uncertainty. There is an objective function that depends on a vector of actions and on an exogenous random vector. Part of the random vector is observable, part is unobservable. In the standard version of this problem, when the exogenous randomness is continuously distributed and the objective function is concave (assuming it is being maximized), we expect the solution to make the action a smooth function of the observable randomness. We add to this framework an assumption that translating the observable randomness into actions has an information cost, in units of

Shannon's measure. We show that in broad classes of cases, the action is not a function of the observable data, but instead, because the observable data is imperfectly assimilated, is random, even conditional on the observable data. Furthermore, in broad classes of cases the distribution of the action is concentrated on a set of lower dimension than what would emerge without the information costs. Where action and observation are both scalars, the continuous one-dimensional distribution of actions that emerges without information costs becomes a zero-dimensional one — i.e. one whose support is a finite, or in some cases countable, set of points.

Our results show that the apparently discrete numerical solutions emerging in the earlier papers really are discrete, and are representative. Because this paper's setup covers multivariate actions, it applies to interesting economic models not considered in earlier papers, including a portfolio choice problem that we consider in some detail.

II. COMPARISON TO OTHER FORMS OF BOUNDED RATIONALITY AND COSTLY INFORMATION

Rational inattention is a form of bounded rationality, an idea that has been part of economics (and psychology) since at least the work of Herbert Simon 1976; 1979 and is still current, as for example in Todd and Gigerenzer (2000) and Gabaix (2011). Bounded rationality includes both the idea that economic agents do not use all available information and the idea that they respond to information in ways that are "simpler" than would be expected if they were perfect dynamic optimizers. Both of these deviations from perfect dynamic optimization are no doubt present in reality, but there is a reason to isolate the effects of economizing on information. If we can come up with a notion of the cost of information, we can go back to a model of an optimizing agent, just adding the cost of information-processing, or a constraint on information-processing, to the optimization problem. Of course this pretends that people have no limitations in rationally optimizing subject to

information processing costs,and in this sense the approach is unrealistic. But as elsewhere in economics, it may be useful to study optimizing behavior even when we know that actual behavior will be at best approximately rational. People can use trial and error, formal education, imitation of others, even cell-phone apps to arrive at near-optimal patterns of behavior even if unable to calculate them with explicit mathematics. To instead fully allow for sub-rationality requires confronting all the problems of psychology and neurological models of human decision-making, an area where, despite recent progress, there is still no standard model to guide us.

How, then, to specify costs of information? To see how rational inattention differs from other approaches, it might be best to sketch a formal decision problem with information costs. An agent must make a decision δ which can be a function only of a random vector X that she has observed. Her utility depends on $\delta(X)$ and on another random variable, Y, via a function $U(\delta(X), Y)$, and she maximizes $E[U(\delta(X), Y)]$. Y and X of course have some joint distribution. This is the standard setup of decision theory in the presence of uncertainty. One can ask the question, what is the cost in utility of failing to use the observation X in setting δ? That is, what is the difference $E[U(\bar{\delta}, Y)] - E[U(\delta(X), Y)]$ when $\bar{\delta}$ is the optimal non-random choice of δ and $\delta(X)$ is the optimally chosen mapping from X to δ. This is one way to define a cost of information. One could also specify that X is a noisy observation of the vector Y, $X = Y + \varepsilon$, with ε i.i.d. normal and with a variance that can be reduced according to a cost schedule. Such approaches to specifying information costs can be appropriate in applied work where there are measurable costs to information acquisition. Drilling an exploratory oil well or commissioning a marketing survey are quantifiably costly actions that generate information, for example.

But even when such direct costs of generating information are present, decision-makers may not make full use of available information because of limits on human

abilities to translate a stream of information into a stream of actions. Rational inattention studies this type of information cost. It may be difficult to cleanly separate the two kinds of cost, so studying information costs without making this distinction, as do Caplin and Dean (2015), may be useful, but studying the effects of the internal costs to an economic agent of processing freely available information is nonetheless interesting.

Mankiw and Reis (2002), and Reis (2006) proposed an approach to modeling information frictions that assumes a cost to frequent updating of information . Agents in this theory process no information until the time when they acquire perfect information about the state of the world and change current and planned future actions. Therefore at the time an action plan is formed, it reflects complete information, while between these dates no information is acquired. Rational inattention in contrast implies continual collection of imperfect information. When the underlying state is continuously distributed, rational inattention implies that it is never exactly known. In many contexts the two approaches make starkly different predictions about behavior.

III. WHY USE THE SHANNON MEASURE OF INFORMATION PROCESSING COST?

Economics studies how people behave as they interact with each other, nature, and market signals. Our theories describe how changes in the environment, including market signals, are translated into changes in people's behavior. Shannon's theory measures information flow as the reduction in uncertainty about one random quantity when some other random quantity is observed. If we think of the price of an asset traded in a thick market as a signal, to which in principle economic behavior should respond, continual adjustment of economic behavior to the constant, unpredictable changes in price would imply an implausibly high rate of flow of information in Shannon's sense. Recognizing this, and its broader implications for the way we model economic behavior, seems like a good idea.

One way to justify the Shannon measure is to start by accepting the idea that the cost of information flow between two random quantities should be a function of their joint distribution, then look for a measure that has appealing properties. The Shannon measure of mutual information between two random variables X and Y whose joint distribution is defined by a density function relative to a product measure $\mu_X \times \mu_Y$ is

$$I(Y,X) = E[-\log(p(Y))] - E[-\log(q(Y \mid X))]$$
$$= \int -\log(p(y))p(y)\,d\mu_Y(y) - \int -\log(q(y \mid x))q(y \mid x)d\mu_Y(y)r(x)d\mu_x(x), \quad (1)$$

where the expectations are over both X and Y, p is the unconditional pdf of Y, q is the conditional pdf of Y given X, and r is the unconditional pdf of X. $E[-\log(p(Y)]$ is called the entropy of Y, so the measure can be interpreted as the expected reduction in entropy of Y from observing X. The base of the logarithm in this formula is conventionally taken to be two, but using a different base just rescales the measure. When base 2 is used, the unit of measure is bits, and when base e is used, the unit of measure is nats.

This measure has the property that if X has two components, X_1 and X_2, The mutual information between X and Y is the mutual information between X_1 and Y plus the expected value (over the distribution of X_1) of the mutual information between X_2 and Y when we condition their joint distribution on X_1. That is, when we break available information into pieces defined by distinct random variables, the information in the pieces can be added up.

It also has the property that it is unit-free. If we make a monotone transformation of X or Y, there is no effect on $I(X,Y)$.

Under mild regularity conditions, these two properties uniquely define Shannon's measure. Other measures of mutual information based on the joint distribution have been proposed, but they lose one of these properties, usually the first (Csiszár, 2008).

Notice that if X and Y are independent, $I(X, Y) = 0$. Also, though the formula in (1) looks asymmetric, in fact it can be easily checked that $I(X, Y) = I(Y, X)$, that is that information in X about Y is the same as information in Y about X.

Aside from the possible appeal of Shannon's measure based on its abstract properties, its usefulness is apparent in its ubiquity in modern electronic communication. We measure file sizes in bits and the speed of our internet connections in bits per second. These bits are units of Shannon's measure. Internet connections can be based on fiber optics, electricity sent over copper wires, or radio waves. The capacity of these connections to transmit information can be measured in bits per second, regardless of the physics underlying their operation. The idea of rational inattention in economics is to apply this same bits per time unit measure to the transmission of observed external information, through a person's sensory system and brain, to an action.

This idea builds on Shannon's notion of a "channel" and its "capacity". A channel takes as input an element of an "alphabet". In the simplest case the alphabet might consist of zero and one. Or it might be A to Z, or an interval of real numbers. The channel's definition specifies the conditional distribution of an output for every possible value of the input alphabet. Since the mutual information between the channel's input and output depends on their joint distribution, while the channel's definition specifies only the conditional distribution of output given input, the rate of information flow through the channel depends on the distribution of the input. The channel's capacity is the rate of information flow through it when the distribution of inputs is chosen so that the rate of information flow is maximized.

Shannon's coding theorem shows that regardless of the distribution of the random variables we want to send through the channel, we can map them into sequences of values of the channel's alphabet in such a way that the rate of information flow is as close as desired to the channel's capacity.

In applying the idea of channel to human behavior, we assume that even though market prices, advertising, weather, etc. are not measured in the "alphabet" of the human nervous system, it may nonetheless be reasonable to suppose that people map these inputs into actions efficiently, so that a common information-processing shadow cost applies across reactions to all signals a person is reacting to.

IV. THE GENERAL INFORMATION-CONSTRAINED DECISION PROBLEM

While many of the most interesting potential applications of rational inattention in economics are to dynamic decision problems, in this paper we focus on static problems. There are previous papers discussing linear-quadratic dynamic problems (Sims, 2010, 2003), and Tutino (2009) has considered a finite state-space dynamic problem. Here we are interested in the emergence of discrete or reduced-dimension behavior, and how it depends on the nature of uncertainty. To make the analysis tractable, we stick to the static case, though it is most natural to think of the problems here as repeated frequently over time.

All the examples we will discuss below are versions of a general problem, which we can state as

$$
\max_{f, \mu_x} \int U(x,y) f(x,y)\, \mu_x(dx)\, \mu_y(dy)
$$

$$
- \lambda \left(\int \log(f(x,y)) f(x,y)\, \mu_x(dx)\, \mu_y(dy) \right. \tag{2}
$$

$$
\left. - \int \log\left(\int f(x,y')\, \mu_y(dy') \right) f(x,y)\, \mu_x(dx)\, \mu_y(dy) \right)
$$

$$
\text{subject to } \int f(x,y)\, \mu_x(dx) = g(y)\, , \text{ a.s. } \mu_y \tag{3}
$$

$$f(x,y) \geq 0 \text{, all } x,y \,, \tag{4}$$

where $x \in \mathbb{R}^k$ and $y \in \mathbb{R}^n$, μ_x and μ_y are σ-finite Borel measures, possibly but not necessarily Lebesgue measure, f is the joint pdf of the choice x and the target y, g is the given pdf for y, before information collection, $U(x,y)$ is the objective function being maximized, and λ is the the cost of information.

The first term in (2) is the expectation of U, and the second is the cost of information. (3) requires consistency of prior and posterior beliefs, and (4) requires non-negativity of the pdf f. The formulation of the model as here, via the joint distribution of x and y, is equivalent to a two-step formulation where the agent chooses a signal Z with some joint distribution with X, then optimally chooses a function δ that maps Z to a decision $X = \delta(Z)$, with the information cost applied to the mutual information $I(Z,Y)$ rather than $I(X,Y)$. This should be clear, because $I(\delta(Z),Y) \leq I(Z,Y)$ for any function δ, and with the signal Z freely chosen we could always just choose the signal to be the optimal $\delta(Z)$ itself. Choosing anything else as signal that delivers the same $\delta(Z)$ can at best leave the information cost unchanged, and certainly leaves the expected utility unchanged.

The objective function is concave in the measure on xy space defined by f, μ_y and $\mu_x{}^3$, and the constraints are linear, so we can be sure that a solution to the first order conditions (FOC's) is a solution to the problem. However the non-negativity constraints can be binding, so that exploration of which constraints are binding may make solution difficult.

Related problems have been studied before in the engineering literature. When $U(x,y)$ depends only on $x - y$ and is maximized at $x = y$, the problem is the static special case of what that literature calls rate-distortion theory. Fix (1978) obtained

---

[3]Expected utility is linear in this probability measure, and mutual information between two random variables is a convex function of their joint distribution, so expected utility minus λ times the mutual information is concave in the measure.

for the quadratic-U case a version of our result in section VII.2, that bounded support for Y implies finitely many points of support for X. Rose (1994) uses different methods, also for the quadratic-U case, and shows discreteness for a somewhat broader class of specifications of the exogenous uncertainty than Fix. We believe that our results for more general forms of objective function and for the multivariate case are new.

The FOC's of the problem with respect to f imply that at all values of x, y with $f(x, y) > 0$ and $g(y) > 0$

$$U(x, y) = \theta(y) + \lambda \log \left(\frac{f(x, y)}{\int f(x, y)\, dy} \right) \tag{5}$$

$$\therefore f(x, y) = p(x) e^{U/\lambda} h(y) \tag{6}$$

$$\therefore \int e^{U(x,y)/\lambda} h(y)\, dy = 1 \,, \text{ all } x \text{ with } p(x) > 0 \tag{7}$$

$$\therefore \quad \int p(x) e^{U(x,y)/\lambda}\, dx \cdot h(y) = g(y) \,, \tag{8}$$

where $\theta(y)$ is the Lagrange multiplier on the constraint (3), p is the pdf of the action x and $h(y) = \exp(-\theta(y))$ is a function that is non-zero where g is non-zero, zero otherwise. At points x where $f(x, y) = 0$, the FOC's require that the left hand side of (5) be less than or equal to the right hand side. Note that if $p(x) = \int f(x, y)\, dy > 0$, the right hand side of (5) is minus infinity wherever $f(x, y) = 0$, so with U bounded above we can conclude that $f(x, y) = 0$ for a particular x, y only if $f(x, y) = 0$ for all y, i.e. $p(x) = 0$.

At points x with $p(x) = 0$, the right-hand side of (5) is undefined. However we can reparameterize $f(x, y)$ as $p(x) q(y \mid x)$ and take the first order condition with respect to p. Since at points with $p(x) = 0$ the value of $q(\cdot \mid x)$ makes no marginal contribution to the objective function or the constraints, the first order condition with respect to p at points with $p(x) = 0$ becomes

$$\max_q E[U(x, y) - \lambda \log(q(y \mid x)) - \theta(y) \mid x] \leq 0 \,. \tag{9}$$

These FOC's do not take explicit account of the possibility of varying μ_x. Adding or deleting a point x with non-zero μ_x probability is accounted for, since that can be treated as setting $p(x)$ to a zero or non-zero value at a point where $\mu_x(x) > 0$. If μ_x puts discrete probability π on a point x_0, we can, though, derive an additional FOC by considering changing the location of x_0. If we change the location of x_0 to a nearby $x^* \neq x_0$ that initially had probability zero (though possibly a non-zero density value w.r.t. Lebesgue measure), while keeping the pdf of $y \mid x_0$ and of $y \mid x^*$ the same, we leave the mutual information between x and y the same and continue to satisfy the boundary condition (3), but we change the expected value of U. The derivative of the expected value of U w.r.t. x_0 when x_0 is changed in this way is

$$\mu_x(x_0) \int \frac{\partial U(x_0, y)}{\partial x_0} f(x_0, y) \, dy \,, \tag{10}$$

which then implies that $E[\partial U / \partial x \mid x]$ must be zero at every point x_0 that has positive probability. For the case where U is quadratic in $X - Y$, this is the familiar requirement that the optimal X is the conditional mean of $Y \mid X$.

We discuss below in section VI details of how to determine conditions under which the distribution of X in this problem's solution are continuously distributed. With the general problem statement and first-order conditions in hand, though, we can proceed to discuss specific examples and their solutions.

V. EXAMPLES

In this section we present example decision problems and discuss their solutions, reserving for the section VII more technical parts of the arguments supporting our results. Some of our solutions are numerical, and the methods used to compute them are described in appendix A.

V.1. **Linear-quadratic tracking in one dimension.** To provide some intuition about the nature of these results, we consider first an example where the mathematics are relatively simple, so we can see how discreteness emerges. This is a one-dimensional tracking problem with quadratic loss. It is the canonical framework for Gaussian rate-distortion theory in the engineering literature. It is also similar to many empirical models in economics, where an agent is modeled as trying to keep a choice variable close to its optimal value, except that here rather than postulating physical adjustment costs, we postulate information processing costs.

Formally the problem is to choose the joint distribution of the decision variable X with the exogenous uncertainty Y so as to maximize

$$-\tfrac{1}{2}E[(Y-X)^2] - \lambda I(X,Y),$$

where λ is the cost of information in utility units and $I(X,Y)$ is the mutual information between X and Y. It is well known that if the fixed marginal distribution of Y is Gaussian in this problem it is optimal to make the joint distribution of X, Y Gaussian[4]. Furthermore if σ_y^2 is the unconditional variance of Y and ω^2 is the conditional variance of $Y \mid X$ (which is of course constant for a joint normal distribution), $I(X,Y) = \tfrac{1}{2}\log(\sigma_y^2/\omega^2)$. It is also optimal to make $E[Y \mid X] = X$.

With no information cost, the solution is trivially to set $Y = X$, so that X and Y both have full support on $\mathbb{R}$. With non-zero information cost, we can use the characteristics of the solution we have laid out above to see that the problem becomes

$$\max_{\omega} -\tfrac{1}{2}\omega^2 - \lambda\left(\log \sigma_y^2 - \log \omega^2\right). \tag{11}$$

Taking first-order conditions it is easy to see that the solution is $\omega^2 = \lambda$. This is reasonable; it implies that with higher information costs uncertainty about y increases, and that as information costs go to zero, uncertainty about Y disappears. It makes the marginal distribution of X normal and gives it full support on the real

---

[4]See Sims (2003) or Cover and Thomas (1991).

line. However, the solution only applies when $\omega^2 < \sigma_y^2$. It is not possible to specify a joint distribution for X, Y in which the conditional variance of $Y \mid X$ is larger than the unconditional variance of Y. So for $\lambda > \sigma_y^2$, we instead have the trivial solution $X \equiv E[Y]$.

In this example we have just one, trivial, possible form of discrete distribution for X: a discrete lump of probability 1 on $E[Y]$. Otherwise, X is normally distributed, with variance smaller than that of Y. We can think of this solution as implemented by the decision-maker observing a noisy measure of Y, $Y^* = Y + \varepsilon$, where ε is itself normal and independent of Y. X is then a linear function of Y^*, and λ determines the variance of ε.

Now consider a problem that seems nearly the same. We alter it only by specifying that the given marginal distribution of Y is not $N(0, \sigma_y^2)$, but that normal distribution truncated at $\pm 3\sigma_y$. The probability of observations in the $3\sigma_y$ tail of a normal distribution is .0027, so this problem is in some sense very close to the one without truncation. We could consider just using the previous solution as an approximation — observe $Y^* = Y + \varepsilon$ and set X to be the same linear function of Y^* as in the un-truncated problem. And indeed this would give results very close to those of the optimal solution.

We show below, though, (and this result appeared earlier in the engineering literature for this model) that whenever the support of Y is bounded in this problem with quadratic loss, the support of X is a finite set of points. How can this be?

If information costs are small, the truncated-Y solution gives the X distribution finitely many points of support, but a large number of them. The weights in this fine-grained finite distribution have a Gaussian shape. The distribution, though discrete, is close in the metric of convergence in distribution to the distribution for X in the untruncated solution.

If information costs are moderately large, though, so that the untruncated solution is not reduced to $X \equiv EY$, the distribution for X can have a small number of points of support, so that it looks quite different from the distribution of X in the untruncated solution. For example, suppose $EY = 0$, $\sigma_y^2 = 1$ and $\lambda = .5$. Then the untruncated solution makes $\omega^2 = .5$ and gives X a $N(0, .5)$ distribution. If we set $X = .5(Y + \varepsilon)$, with $\varepsilon \sim N(0, 1)$ and independent of Y, we would be using the same formulas as in the untruncated case, and would achieve almost the same result, $E[(Y - X)^2] = .49932$ instead of .5 and information costs no higher than in the untruncated case. X would be continuously distributed, not exactly normal, but still with the whole of $\mathbb{R}$ as support.

But the optimal solution with this truncation and this cost of information has just four points of support for the X distribution: -1.0880, -.1739, .1739, 1.0880. It achieves almost exactly the same $E[(Y - X)^2]$ as in the untruncated case[5], while using about 4% less information. The naive use of the untruncated solution wastes information capacity, because in the rare cases where the observed noisy signal Y^* is much above 3, it is giving us extremely precise information in the truncated case: the true value of Y must be very near 3 if Y^* is much greater than 3. There is no point in achieving such low conditional variance for this particular type of rare event, so the optimal solution uses that information capacity elsewhere. This explains why the distribution of X has smaller support than Y's in the optimal solution for the truncated problem.

The conditional density functions for Y at each of the four points in the X distribution, i.e. the four possible distributions of posterior beliefs about Y, are shown in Figure 1. They are weighted by the probability of the corresponding X value, so that at any point on the x axis the sum of the heights of these weighted pdf's

---

[5]These numbers are based on solving the problem numerically with a grid of one thousand points between -3 and 3. They may not be accurate to more than about 3 decimal places as approximations to the continuously distributed problem.
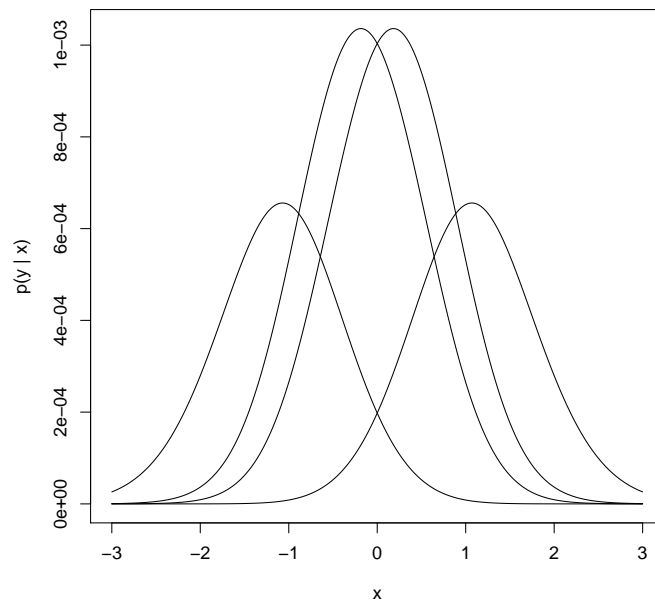
FIGURE 1. Weighted conditional pdf's for Y in tracking problem, $\lambda = .5$

is exactly the unconditional truncated normal density for Y. While the four densities are of course not exactly normal, they are of the same general shape as normal densities, and have roughly the same .5 variances as do the conditionally normal densities for the untruncated problem. We might ask what would happen if the decision-maker mistakenly used the results of the truncated problem when in fact the distribution of Y is not truncated. For example, we can suppose that the signal is one of the four points of support in the optimal truncated solution when $|Y| < 3$, but 0, or "ERROR", when $|Y| > 3$. In these rare cases, then the decision-maker would simply set $X = 0$.[6] Losses would be slightly higher than in the smooth solution for the untruncated case, but by less than 0.3%.

---

[6]More formally, we would make the joint distribution of X and Y conditional on $|Y| < 3$ match the solution of the truncated problem, while making that joint distribution simply a lump of probability on $X = 0$, independent of Y, conditional on $|Y| > 3$.

These two problems are very similar in objective function and initial distribution of uncertainty. Their solutions are very similar in terms of objective function value and conditional distribution of the unknown Y. The solutions are also close in the sense that either problem's solution can be applied to the other problem with little deterioration in objective function value. Nonetheless the solutions differ discontinuously in the marginal distribution of the choice variable X. This kind of result reappears in the other, more complex setups we examine below, and is inherent in the structure of these problems. In some examples we find there are no solutions with continuously distributed X, but in others we find there is a family of densities for Y and corresponding densities for X for which the optimal solutions makes X continuously distributed. Where this is true, though, small perturbations of the problem again make the optimal distribution for X discrete.

The message of the paper, then, is not that continuously distributed behavior cannot emerge from decision problems with an information constraint. Indeed, standard frameworks like linear-quadratic problems imply continuously distributed solutions in many cases. But discretely distributed behavior is optimal in a wide class of cases, and often will be close to optimal even in problems where the exact optimum is continuously distributed.

V.2. **A risk-averse monopolist.** This next example has more interesting economic content and is outside the range of cases considered in the engineering rate-distortion literature.

Suppose a risk-averse monopolist faces a demand curve $Q = q(X)$ and a constant returns production technology with unit cost C, where C is a random variable. We use X instead of P as notation for price, to avoid confusion with probabilities and probability densities. Suppose the monopolist has logarithmic utility for profits. With a utility cost λ per nat (the unit of measurement for mutual information when log base e is used in defining it) of transmitting the information in C to a choice of

X, the problem becomes

$$\max E\Big[\log\big((X - C)q(X)\big)\Big] - \lambda I(X, C), \tag{12}$$

where the maximization is over the joint distribution of X and C. We assume C is non-negative and continuously distributed with pdf g. This is a special case of our generic decision problem with information cost as described in section IV. To proceed further we need to assume an explicit form for the demand curve $q()$. Consider $q(x) = x^{-\theta}$ for $x > 0$.[7]

In section VII.3 we provide an analysis of this problem and show that there is a class of continuously distributed solutions for X when the density $g(\cdot)$ of C is a certain mixture of scaled beta distributions. This lets us reach some general conclusions that can be summarized as asserting that any kind of distribution for X can emerge if we can freely vary the distribution of C, but distributions for X whose support contains intervals of the form $(0, T)$ can emerge only with a restricted class of C distributions. More specifically:

(i) For any combination of distribution of price X on $\mathbb{R}^+$, demand elasticity $\theta > 0$, and information cost $\lambda > 0$, there is a density function $g(\cdot)$ for the exogenous cost variable C that makes the given X-distribution the optimal distribution for that combination of θ, λ, and g. In other words, every kind of distribution of X is possible as a solution to the problem, as we vary the distribution of C over all continuous distributions on $\mathbb{R}^+$.

(ii) For any given $g(\cdot)$ and θ, even if the problem admits a solution with continuously distributed X for some value of λ, as $\lambda \to 0$ eventually the solution does not have full support on any interval of the form $(0, T)$, for any $0 < T \leq \infty$.

_____

[7]Matějka (2008) studies this problem with a risk neutral monopolist and the same demand curve. The mathematics of this example with the logarithmic utility is close to that of the two-period savings problem in Sims (2006b).

(iii) When there is an a priori known upper bound $\bar{c}$ on cost C, any solution in which X is continuously distributed over some interval of the form $(0, \bar{x})$ must have $\bar{x} = \bar{c}$. This might seem counterintuitive, since with $\lambda = 0$ we know that the solution is $x \equiv \theta c/(\theta - 1) > c$, and indeed solutions often make $P[X > \bar{c}] > 0$, but when they do so the support of X never contains an interval of the form $(0, \bar{c})$.

(iv) If there is a known upper bound $\bar{c}$ on C and $E[C] > \bar{c}(\theta - 1)/\theta$, there is no solution in which the support of X contains $(0, \bar{c})$.

We have not been able to show that when X is not continuously distributed, it must necessarily have countable discrete support, but in Proposition VII.5 we do show that the support must be discrete and finite when the support of the cost distribution is contained within an interval of the form $(c^*, \theta c^*/(theta - 1)$. With such a tight bound, the support of the X distribution lies outside the support of the C distribution, and it turns out this implies discreteness.

We solve this problem numerically with the distribution of C a Beta(4,4) distribution scaled to cover the interval (0,10). The prior distribution for costs is thus symmetric around $c = 5$. It is in the family of distributions shown in section VII.3 to be consistent with continuously distributed optimal X, but only for $\lambda = 1/3$, $\theta = 7/3$, and the distribution of X concentrated on the single point $x = 10$, which obviously itself has a discrete distribution for X. When $\theta = 1.5$, and $\lambda = .05$, full support on a $(0, T)$ interval is impossible because the mean of the $B(4, 4)$ distribution scaled to $(0, 10)$ is 5. Since this exceeds $10 \cdot (\theta - 1)/\theta = 3\frac{1}{3}$, a solution with full support on $(0, 10)$ is impossible by iv above.

We find numerically that X is distributed on a support of 6 points: 0.845, 1.790, 3.254, 5.648, 9.796, and 17.061, with probabilities .000028, .000385, ,00334, .025203 .170879, and .800163. The weighted and unweighted conditional densities for C given these 6 values for the decision variables are shown in Figures 2 and 3. When

the densities are weighted by the probabilities of the corresponding X values, two of the four are essentially invisible because weights on them are so small. But they imply very precise knowledge that c is small and in each case certainty that C does not exceed the corresponding value of X.

Because we necessarily approximate the Beta(4,4) distribution with a discrete grid, it is possible, even likely, that the fully optimal solution for the continuous version of the problem has a countable infinity of points of support with a limit point at $x = 0$ and probabilities converging rapidly to zero as the support point approaches zero, but because these additional support points would have very low probability, our six point solution (which is indeed optimal for our grid of 1000 equi-spaced points of support for approximating the Beta(4,4)) will be very close to optimal even in the truly continuous case.

This pattern of results emerges because profits are unbounded above as costs approach zero, while utility is minus infinity if profits become negative. Very precise information about C, including an upper bound on it, is therefore extremely valuable when C is in fact very low. The information-constrained monopolist simply sets a high price, enough to ensure a profit even if C is at its maximum possible level, 80% of the time, with his beliefs about C in this case spread broadly over the interval 3 to 10. But on the rare occasions when C is low, he collects precise information about it, including a firm upper bound. Observing his behavior over time, we would see extended periods of constant prices, with occasional isolated instances of sharply lower prices. Of course since without information costs price would be $\theta C/(\theta - 1)$, it would in that case have a distribution that mimicked the form of the cost distribution, a density centered centered at 15 and spread symmetrically between zero and 30. Someone observing the information-constrained behavior of the monopolist would, not accounting for the effects of information costs, draw

mistaken conclusions about the distribution of costs or the elasticity of demand, as well of course about the size of "menu costs" of changing prices.
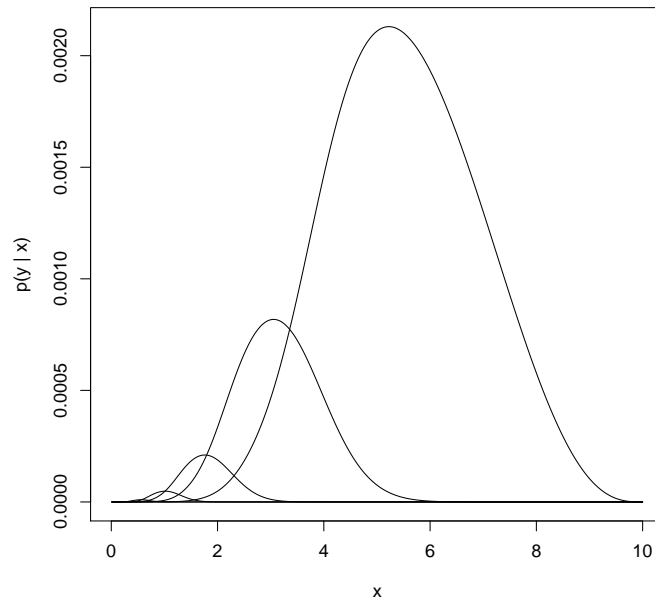

FIGURE 2. Weighted conditional pdf's for Y in risk-averse monopolist problem, $\lambda = .05, \theta = 1.5$

V.3. **Multivariate linear-quadratic tracking.** This again is a classic problem from the engineering rate-distortion literature. We consider it here because it is the simplest case where in a multivariate problem, the support of the marginal distribution of the decision variable becomes measure-zero for some range of information costs, while not being a countable set of points. It is also close to models with an economic interpretation. For example, a consumer trying to choose a consumption bundle close to an optimal bundle, when the optimal bundle is varying because of changing prices and incomes. Or a monopoly price setter producing multiple products with stochastically varying costs, trying to keep the prices close to an optimal target defined by the costs and demand.
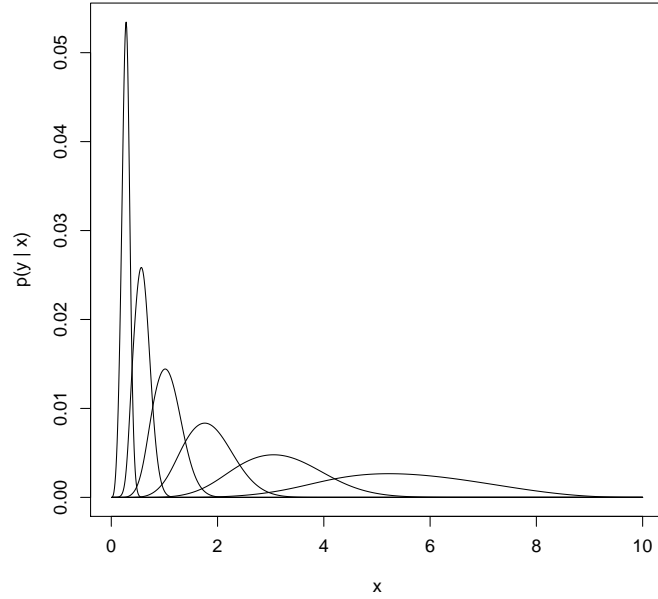
FIGURE 3. Unweighted conditional pdf's for Y in risk-averse monop-
olist problem, $\lambda = .05, \theta = 1.5$

The problem is to choose the joint distribution of X, Y to maximize

$$-\tfrac{1}{2}E[\|X - Y\|^2] - \lambda I(X, Y),\tag{13}$$

where X and Y are n-dimensional vectors and Y has a given $N(0, \Sigma)$ distribution. Of course again here the solution when $\lambda = 0$ simply sets $X \equiv Y$ and the distribution of X is also $N(0, \Sigma)$ and thus has full support on $\mathbb{R}^n$. If Σ is a scalar matrix $\sigma^2 I$, the solution is just to apply the solution for the one-dimensional problem of section V.1, one dimension at a time, allocating information capacity equally to all components of Y.

But if Σ is diagonal, but with unequal variances on the diagonal, the support of X becomes measure-zero in $\mathbb{R}^n$ for a certain range of values of λ, and does so without being reduced to a point. This result is the simplest form of what is known in the engineering literature as the "water-filling" result. Using σ_i^2 to denote the variance

of the marginal distribution of Y_i and ω_i^2 to denote the conditional variance of Y_i given information, the objective function is

$$-\tfrac{1}{2}E\left[\sum_{i=1}^{n}\omega_i^2 - \lambda(\log(\sigma_i^2/\omega_i^2)\right]. \tag{14}$$

The interior solution would make ω_i^2 constant across i, and for low information costs this is indeed the solution, with low values of ω_i^2 corresponding to low value of information costs λ. But as in the one-dimensional case, $\omega_i^2 > \sigma_i^2$ is impossible. If this constraint is binding for all i, we are back to the trivial solution with no information collected and $X \equiv E[Y]$. But for intermediate values of λ the solution will set ω_i^2 equal to a constant $\bar{\omega}^2$ for those values of i with $\sigma_i^2 > \bar{\omega}^2$, and leave $\omega_i^2 = \sigma_i^2$ where $\bar{\omega}^2 > \sigma_i^2$. In other words, it is optimal to collect information in these cases only about the Y_i variables with the largest variance. For very high λ, the support of X is the single point $X = E[Y]$. As λ falls, initially information capacity is used only to reduce uncertainty about the Y_i with the largest variance. For such a solution, the support of X is on a one-dimensional subspace of $\mathbb{R}^n$. As λ drops further, information is collected about additional dimensions of Y, giving the distribution of X higher dimension, but still measure-zero, support, until finally λ falls far enough that information is collected about all dimensions and the distribution of X has full support.

Low-dimensional support for X can take other forms. If Σ is scalar, and we alter the problem by making Y truncated normal rather than normal, the solution must give X support that contains no open sets, as was the case in the one-dimensional version of this problem. But in this case the nature of the support of X depends on the shape of the truncation boundary.

If Y is bivariate $N(0, I)$, truncated at $\|Y\| \le 3$, the support of X is likely to be one or more circles in $\mathbb{R}^2$. We show in proposition VII.3 that if there are any solutions to the problem that concentrate on a countable collection of points, then there are also

solutions for which the support of X is a countable collection of circles centered at the origin. On the other hand, this result depends on the fact that the truncation is itself at a circle around the origin, so the problem is rotationally symmetric. If instead the $N(0, I)$ distribution for Y is truncated at a rectangle, numerical calculations show that the solution for X is likely to be supported at a finite set of isolated points in $\mathbb{R}^2$.

A numerical solution for the optimal X, Y joint distribution constrained to finitely many points of support cannot of course deliver the continuously distributed solution we know exists in the case of a symmetric truncation. But we can see the nature of the solution by calculating, for the case $\lambda = 2/3$, a solution with 10 points of support. All the X values in this solution turn out to lie on the circle of radius .7604 about the origin, and the points are equally spaced around that circle and have equal probabilities. Figure 4 shows the 10 conditional densities for $\theta \mid X$, the angle in the polar coordinates for $Y \mid X = x$, for the 10 optimal X values. A solution allowing 12 points of support concentrated X on the same circle, and produced the same value of the objective function to 12 digits.[8] The conditional pdf of the length of the X vector (the first polar coordinate) is identical over all 10 points of support, and thus also equal to the unconditional pdf for the vector length. In other words, it is optimal here to collect information only about the relative size of the two components of Y, not on their absolute size. Furthermore, the fact that the 10 and 12 point solutions produce the same objective function value strongly suggests that here the solution with X continuously distributed would simply be a mixture of the solutions with 10 or 12 points of support, with the same conditional distribution of $Y \mid X$ for every value of X in its support.

With the same objective function and the same value of λ, but with the truncation at the square bounded by ± 3 for both components of Y, the solution concentrates

---

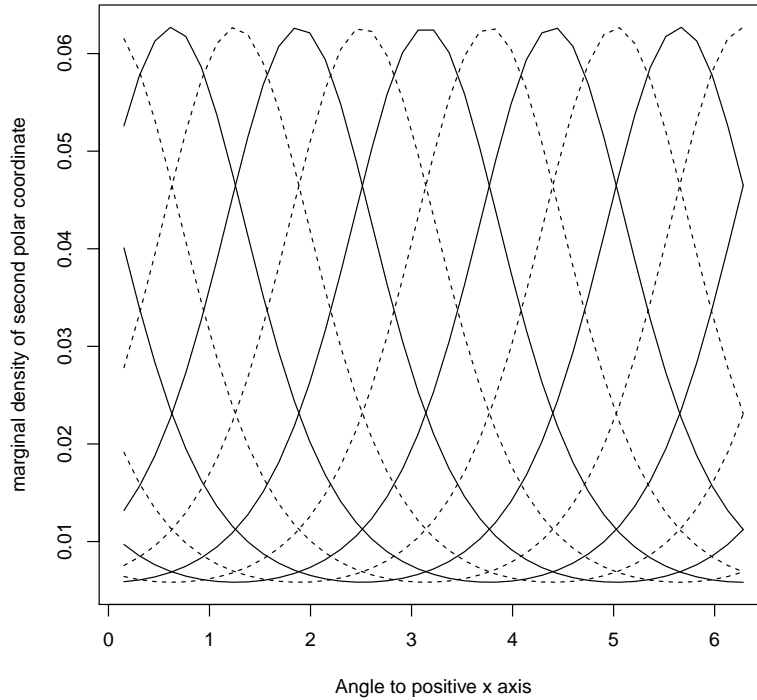[8]Details about the numerical methods are in appendix A.

FIGURE 4. Conditional pdf's of polar angle, 2d tracking with symmetric truncation

on 9 points, arranged in an equally space grid on a square centered at zero and with side 1.59, with probability .25 at the center, .125 at the centers of the four sides, and .0625 at the four corners. Figure 5 shows the points of support for the two truncations. The dark squares are the points of support when the truncation is to the square, and their areas are proportional to the probability weights. The small circles are the 10 equi-probable points of support with the circular truncation. Note that the truncations boundaries, $-3 < x < 3$, $-3 < y < 3$ for the rectangular case and $\|x\| < 3$ in the circular case, lie far outside the boundaries of the graph in both cases.

In this problem, arising from just a slight truncation applied to a problem that we know has a solution with continuously distributed X, there is, as might be expected,
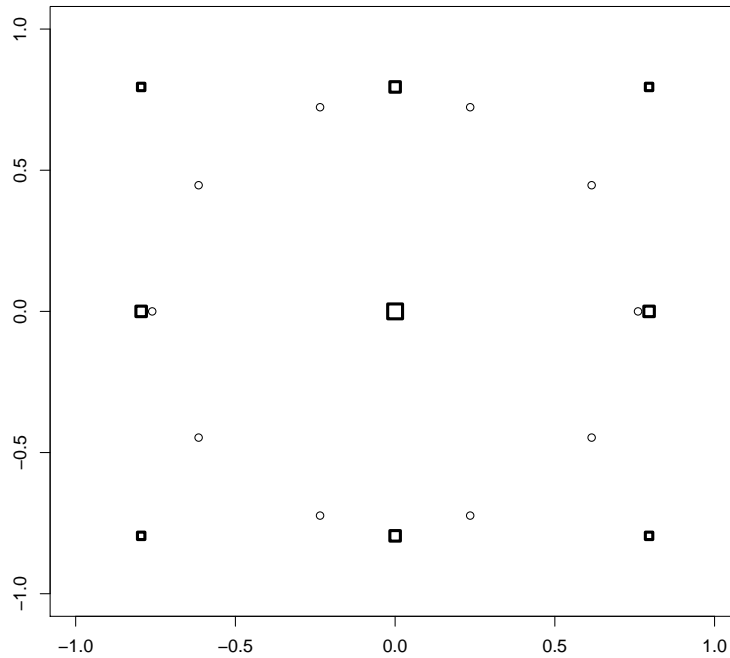
FIGURE 5. Support of x for 2-D tracking with square and circle truncation

Note: The squares show the points of support with truncation at ± 3 in both directions. Their areas are proportional to the probability weights on the points. The circles are the 10 equal-weighted points of support for the solution with truncation at $\|x\| < 3$. In both cases information cost λ is set at $2/3$ and the exogenous uncertainty to be tracked is $N(0, I)$.

only a fairly narrow range of values of λ that produce neither a trivial solution with all probability concentrated at $X = EY = 0$ nor a solution with very many points of support that is close in distribution to the solution of the untruncated problem. With $\lambda = 1$, the solution to either form of truncated problem reduces to $X \equiv 0$. As λ decreases to .5 or less the number of points of support in the solution rapidly increases.

V.4. **Portfolio choice.** Here we consider a problem that has an interesting economic interpretation, a non-quadratic objective function, and a multi-dimensional

choice vector. We show in VII that the solution necessarily gives X support containing no open sets, but have not been able to rule out solutions that are continuously distributed on lower-dimensional sets. We display numerical solutions that turn out to give X support on a finite set of points.

The problem is

$$\max E[\theta'(Y + Z) - \tfrac{1}{2}(\theta'(Y + Z)^2] - I(\theta, Y)] \tag{15}$$

subject to

$$\theta'\mathbf{1} = 1, \tag{16}$$

where θ is a vector of portfolio weights, summing to one and $Y + Z$ is a vector of random yields, of the same dimension as θ. Before information collection, $Y \sim N(\mu_y, \Sigma_y)$ and $Z \sim N(0, \Sigma_z)$, with Y and Z independent. Z is intrinsic uncertainty that cannot be reduced. Y is freely available information that can be processed and translated into a choice of θ only at an information cost $I(\theta, Y)$. There is no restriction on short sales or borrowing, so the elements of θ can be positive or negative. This framework might describe an individual investor who does not respond to every available increment of information about available investments, or at a very different time scale, to a professional high-frequency trader receiving market information over a costly data connection billed in bits per second.

We display two numerical solutions. Both are for the case of three available securities, with the first security risk free. In the first, the two risky securities have a higher expected yield, while in the second the risky securities have the same expected yield as the riskless one. When the risky securities have a higher unconditional expected yield, it will be optimal to hold some of both even when information about Y makes conditional expected yield higher in one of the two. Furthermore, holding both in roughly equal amounts reduces variance. As a result, in this case it is optimal to collect no information, simply borrowing to invest equally in both

risky securities, unless information cost becomes quite low. And when information cost does become just low enough, the optimal solution still has near-equal positive weights on the two risky securities about 95% of the time, using information capacity mainly to identify rare cases where one of the risky securities has a high enough relative yield to justify shorting the other (which occurs about 4% of the time in our example) and even more rarely (about 0.7% of the time in our example) shorting both risky assets while holding large amounts of the riskless asset. Investors behaving this way would, over time, appear usually to be making little or no change in their portfolios in response to available information about relative yield, but would in rare conditions make large portfolio changes, shorting risky investments to hold large amounts of a safe asset.

When we make the expected yield on both types of securities the same, information starts being used at higher levels of the cost of information, and it is used to invest long or short in the three assets more or less symmetrically. In repetitions of this decision process over time, investors would be seen to be more frequently changing their portfolios than in the first case. This pattern of behavior might be the starting point of a model predicting different market volatility when, as at the zero lower bound on nominal rates, expected returns on risky and riskless assets converge.

In both cases we set $\mu_y = .03$ for the risky securities. We make Σ_z diagonal, with 0 at the upper left (making the first security risk free) and .0003, implying a standard deviation of .0173, as the value of the other two diagonal elements of Σ_z. The difference between the two cases is that in the first we set the first element of μ_y, the return on the risk-free security, to .02, implying a higher expected return on the risky investments. while in the second we set it to .03, matching the expected return on the risky securities.

Results for two numerical solutions are displayed in Figures 6 and 7. The information cost is set lower for the case with riskless yield below the risky yield, because at the higher cost used for the equal-expected-yield case the solution with lower riskless yield is the trivial solution with no information collected. On the other hand at the lower information cost used for the other case, the solution with expected yields equal has 8 points of support, with weights on the two risky securities nearly, but not exactly, lying on a circle.
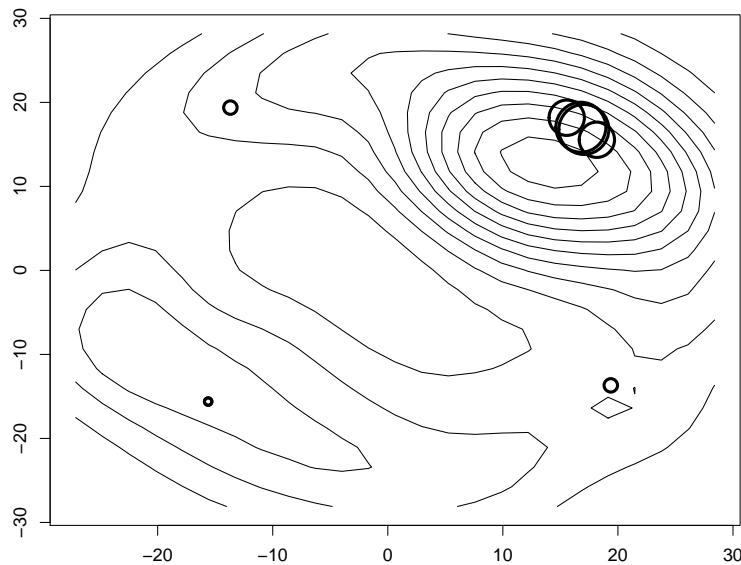

FIGURE 6. Portfolio weight distribution: risky yield higher

Note: Both the irreducible component and the reducible components of uncertainty for the risky assets have standard deviation .0173 in all cases. Expected yields are .02 for the riskless asset and .03 for the risky assets. Contour lines are for a smoothed density formed from 400 simulated draws of optimal portfolios with zero information costs. Dark circles are centered at the 7 points of support of the distribution of optimal portfolios with information cost λ of 0.1. Circle areas are proportional to the probability weights of the portfolios.

Figure 8 shows the level curves of the joint pdf's of the two risky asset yields for each of the 7 points of support in the distribution of portfolio weights shown in Figure 6. The first four portfolios in the table have high probability and show small
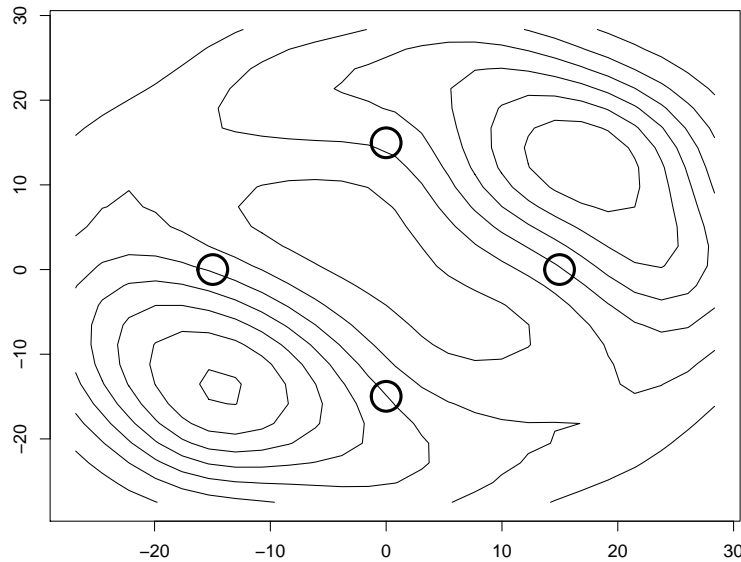
FIGURE 7. Portfolio weight distribution: equal expected yields

Note: See note to Figure 6. Here the distribution with information costs has equal weights on the four points of support. The expected yields are .03 on all three assets. Information cost λ is .167.

variations in portfolio weights. They correspond to the four sets of level curves drawn with dotted lines in the figure. The other three portfolio allocations occur with much lower probability, and they correspond to the three pdf's in the plot drawn with solid lines. As one would expect, these three low-probability portfolios, since they imply shorting one or both risky assets, imply relatively low yields on one or both assets. It is important to recognize, though, that the pdf's all show some overlap. That is, the investor does not choose to acquire information that partitions the space of possible yields. Also note that the pdf's shown here must sum to deliver the given marginal pdf, which is that of two independent normal distribution with mean .03 and standard deviation .01732. The high-probability conditional densities therefore have nearly circular level curves, nearly matching
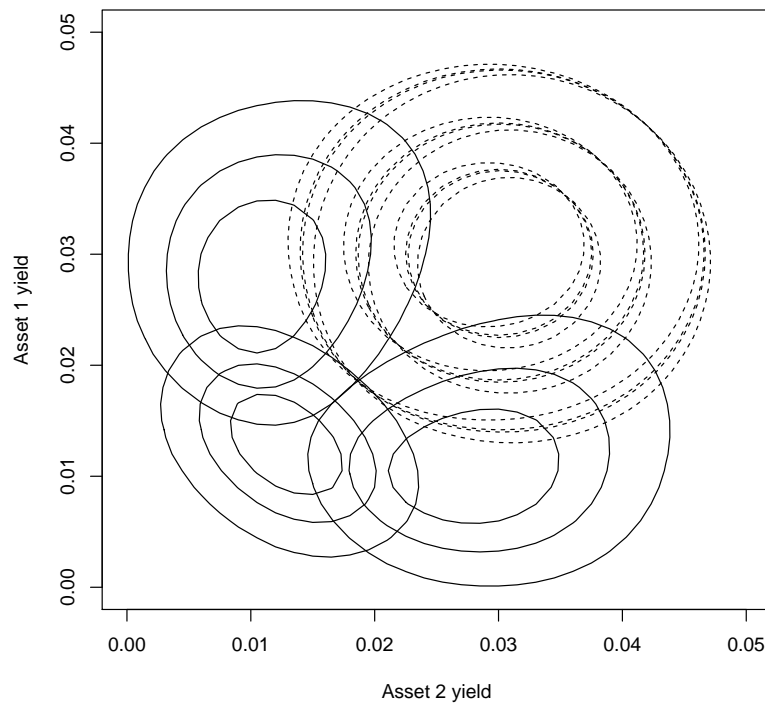
FIGURE 8. Conditional pdf contours, $\lambda = .1$, riskless return lower

those of the marginal density in shape, while the low-probability conditional densities are twisted away from Gaussian elliptical forms.

Note that investors behaving according to the Figure 6 solution would mislead researchers trying to interpret their behavior with a "sticky information" or an "menu cost" approach. The four larger circles have probabilities .32, .32, .16 and .16. Thus in i.i.d. repetitions of the decision problem, the decision-maker would in most periods make a portfolio change, but very rarely make a large portfolio change. The small shifts in portfolio are reactions to weakly informative signals, which therefore contribute little to information costs. Menu cost or sticky information models would interpret the portfolio shifts that occur as reflecting full information at the time of change. Since the changes are small (because in fact the

information the changes respond to is weak), the menu cost or sticky information model would need to suppose high risk aversion to explain the small size of the portfolio shifts. But then on the rare occasions when the investors get a signal of very low returns on one or both risky assets, they heavily short one or both risky assets, making their overall portfolio quite risky. This would seem to contradict high risk aversion. Of course in the rational inattention model, the frequent small changes do not reflect low risk aversion, just small amounts of information being collected most of the time.

VI. SOLVING THE STATIC INFORMATION-CONSTRAINED DECISION PROBLEM

From (6) we can see that $e^{\alpha U(x,y)}h(y)$ must be the conditional pdf of y given x, and therefore for every x with $p(x) > 0$,

$$C(x) = \int e^{\lambda^{-1}U(x,y)}h(y)\,dy = 1\,. \tag{17}$$

If $U(x,y)$ is analytic in x on some open set S, then there are various combinations of assumptions on g, the marginal density of y, and on the behavior of U as a function of y that will allow us to conclude that $C(x)$ is analytic in x. In an economic application, putting together the regularity conditions is the central concern, but it is worthwhile understanding first how analyticity of $C(x)$ leads to the dimension-reduction results.

Proposition VI.1. *Suppose $C : \mathbb{R}^n \to \mathbb{R}$ is analytic on a connected open set $S \subset \mathbb{R}^n$ and $B = \{x \mid C(x) = 1\}$. If $f : \mathbb{R}^k \to S$ is analytic on $\mathbb{R}^k$, then either $f(\mathbb{R}^k)$ lies entirely within B, or $f^{-1}(f(\mathbb{R}^k) \cap B)$ contains no non-empty open sets. If $k = 1$ and $f(\mathbb{R})$ is not contained in B, $f^{-1}(f(\mathbb{R}^k) \cap B)$ consists of a countable set of points with no limit points.*

Proof. A function analytic on an open set S can be constant on a connected open subset of S only if it is constant on all of S. The mapping $C(f(\cdot))$ from $\mathbb{R}^k$ to S is itself an analytic function, and $f(\mathbb{R}^k) \cap B$ is a set on which this mapping is constant.

Thus that set either is included in B (so the mapping is constant on $\mathbb{R}^k$) or contains no open sets in its inverse image. If x is one-dimensional, $C(f(\cdot))$ is an analytic mapping from $\mathbb{R}$ to $\mathbb{R}$. Such a mapping cannot be constant on any infinite set that contains a limit point, unless it is constant on the whole of its domain. So in this case $f(\mathbb{R}) \cap B$ is either contained in B or consists of a countable set of points. □

The interpretation of this proposition for our decision problem runs as follows. Assuming the problem makes $C()$ analytic, the set $B = \{x \mid C(x) = 1\}$ contains the support of $p(\cdot)$, the density function of the decision variable x. Therefore if x is one-dimensional and unconstrained, then either $B = S$, in which case the support of the distribution of x in the optimal solution can be any subset of $\mathbb{R}$, including possibly the entire real line, or B is not the whole of S, in which case it is a countable set of points with no limit points in S. If B's closure is a subset of S, then it, and hence the support of $p(x)$, consists of a finite set of points. In higher dimensional problems, where $x \in \mathbb{R}^n$, B will either be all of $\mathbb{R}^n$ or a set containing no open subsets. Even where B is bounded, this does not necessarily mean that the support consists of discrete points. For example, $C(x) = x^2$ defines an analytic function on $\mathbb{R}$, and $C(x) = a$ for each a defines a set with 0, 1, or 2 discrete points as elements. But $C(x, y) = x^2 + y^2$ is an analytic function on $\mathbb{R}^2$, and for it $C(x, y) = a$ defines a circle for positive values of a — a set that contains no open subsets of $\mathbb{R}^2$. Note that any straight line in $\mathbb{R}^2$ intersects a circle in at most two points. This fits the proposition's conclusion that any analytic function f mapping $\mathbb{R}$ to S maps $\mathbb{R}$ to a set that either is inside B or intersects B in a countable number of points. Here B is the circle and the straight line is f.

This means that, for example, when $C(\cdot)$ is not constant, while a circle is a possible shape for the support of x in $\mathbb{R}^2$, a square is not. A square, or any polygon, has sides that are segments of unbounded straight lines in $\mathbb{R}^2$. $C()$ evaluated along a straight line passing through that segment would be an analytic mapping from $\mathbb{R}$

to S whose inverse image contained an open set on which it was constant. But by construction the line is in B only along the segment, and hence the mapping would not be constant, which is impossible.

When considering whether solutions to particular economically interesting decision problems must give the distribution of the decision variable finite support, one can proceed by checking two conditions: Is $C(\cdot)$ in (17) analytic? and can it be constant? In proving that it cannot be constant, sometimes it may be possible to show that the function $h(\cdot)$ in (7) is uniquely defined by that equation. That equation is a Fredholm integral equation of type 1, which is well known[9] to require a case by case approach to solution. In some cases the equation has a unique solution for a wide class of $g(\cdot)$ functions given some side conditions on U. Then marginal densities that cannot be generated as

$$g(y) = \int p(x)e^{U(x,y)/\lambda}\,dx h(y)$$

must, if $C(\cdot)$ is analytic, imply support for X containing no open sets. In other cases there may be no $h(\cdot)$ that satisfies (7) for all X in S and a direct argument that $C(\cdot)$ cannot be constant is required.

We apply these ideas in the proofs in section VII below.

VII. DETAILED ARGUMENTS

VII.1. **Sufficient conditions for analyticity of** $C()$. A result that is useful in proving analyticity of $C()$ in many cases is

Proposition VII.1. *Suppose*

(i) *for every y, $v(x,y) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is analytic in x over the domain (i.e., open connected set) $S \in \mathbb{R}^n$ (with S not varying with y);*

---

[9]See, e.g, (Kythe and Puri, 2002)

(ii) *for every y v(x, y) as a function of x can be extended to the same complex domain $S^* \supset S$;*

(iii) *and v(x, y)h(y) is integrable jointly in x and y.*

Then $\int v(x, y)h(y)\, dy$ is analytic in x on S.

Proof. Functions on a complex domain in $\mathbb{C}$ that integrate to zero around circles are the analytic functions, and this extends coordinate by coordinate to multivariate analytic functions on domains in $\mathbb{C}^n$. (See Krantz (1992, definition IV, p.5). So in our case $v(x, y)h(y)$ integrates to zero, around circles in S^*, for each x-co-ordinate for each value of y. Fubini's theorem lets us interchange orders of integration for integrable functions, so $\int v(x, y)h(y)\, dy$ is analytic on $\mathbb{C}^n$, and an analytic function on $\mathbb{C}^n$ is real analytic when restricted to $\mathbb{R}^n$. □

VII.2. **Tracking.** When the objective function can be written as $U(x, y) = V(x - y)$ the integral equations in the first order conditions can be studied with Fourier transforms, because they become convolutions. That is, we can write (7) and (8) as

$$\int e^{V(x-y)/\lambda} h(y) \mu_y(dy) = (e^{V/\lambda} * h)(x) = 1 \tag{18}$$

$$\int p(x) e^{V(x-y)/\lambda} \mu_x(dx) h(y) = (p * e^{V/\lambda})(y) h(y) = g(y). \tag{19}$$

The first of these, (18) always has as one solution a constant $h(y) \equiv 1/\kappa$, where $\kappa = \int e^{V(z)} dz$. The only assumption needed is that e^V is integrable. Furthermore, if the Fourier transform of e^V is non-zero on the whole of $\mathbb{R}^n$, this is the only h that satisfies the equation for all $x \in \mathbb{R}^n$. But if this h is going to provide the solution for the decision problem, it must be, from the second equation (19), that $p * e^{V/\lambda} = g/\kappa$. Thus full-support solutions must admit the interpretation that the random variable Y is the sum of X and a "noise" random variable that is independent of X and has a pdf proportional to $e^{V/\lambda}$. We can see immediately then that if V is bounded below, full-support solutions are impossible because $e^{V/\lambda}$ is not a pdf. Furthermore, if $V(x - y) > -\infty$ for all finite arguments and $E^{V(x-y)/\lambda}$ is integrable,

the "noise" variable in the solution has all of $\mathbb{R}^k$ as support, and therefore no full-support solution can exist for cases where the support of Y is bounded.

Both our one-dimensional (section V.1) and our multivariate (section V.3) tracking problems fit the assumptions here, because with quadratic V, $e^{V/\lambda}$ has the shape of a normal density, and its Fourier transform therefore also has the shape of a normal density. This then implies that any full-support solution with $g()$ a normal pdf must make X and Y jointly normal. Whether this full-support solution is possible then depends on whether $\kappa \tilde{g}/e^{\tilde{V}/\lambda}$, which would have to be equal to $\tilde{p}$ in the solution, is the Fourier transform of a probability measure.

In the one-dimensional case, when Y is normal, $\kappa \tilde{g}/e^{\tilde{V}/\lambda}$ is the Fourier transform of a non-singular normal density if and only if the variance of Y exceeds the variance of the normal density $e^{V/\lambda}/\kappa$. In the multivariate case the corresponding condition is that the covariance matrix of Y, less the covariance matrix of the $e^{V/\lambda}/\kappa$ density, is positive definite.

When the support of Y is bounded, we can conclude immediately in these linear-quadratic tracking problems that solutions with full support are impossible, because any solution that makes Y have the distribution of the sum of X and a normally distributed error with density $e^{V/\lambda}/\kappa$ necessarily implies that Y has full support. This result holds more generally for tracking problems, and we can also conclude that when full-support solutions do not exist for these problems, the solution gives X a distribution with finitely many points of support. These results are summarized in the following proposition.

Proposition VII.2 (Discreteness for tracking with bounded support)**.** *Suppose*

    (a) $V() : \mathbb{R}^k \to \mathbb{R}$ *is analytic on all of* $\mathbb{R}^k$*;*
    (b) $M(a) = \max_{\|x\| \geq a} V(x)$ *satisfies* $M(a) \to -\infty$ *as* $a \to \infty$*; and*
    (c) Y *has bounded support in* $\mathbb{R}^k$*.*

Then the solution to the information-constrained decision problem in 2 with $U(x, y) = V(x - y)$ gives X a distribution whose support contains no open sets. When $k = 1$, this set consists of finitely many points.

Proof. g, as a probability density, must be integrable. $\int p(x)e^{V(x-y)/\lambda}\mu_x(dx)$ is continuous in y and therefore bounded below away from zero on the bounded support of Y. From 19, therefore, we can conclude that h is integrable. Therefore $C(x) = (e^{V/\lambda} * h)(x)$ is analytic, because the convolution of an analytic function with any integrable function (indeed any finite measure) with bounded support is analytic. Furthermore, for large enough values of x we have from the convolution formula and the boundedness of the support of h that $C(x)$ is a weighted average of values of $V(z)$ with $\|z\|$ arbitrarily large. $C(x)$ therefore goes to zero as $x \to \infty$ and thus cannot be constant. The support of X, on which we must have $C(x) = 1$ is a set on which an analytic function is constant. Since $C(x)$ is not constant on its whole domain, it can be constant only on a set containing no open sets. If $k = 1$, the set on which $C(x)$ is constant cannot contain any sequences with limit points, and since it must lie inside a bounded set, it consists of finitely many points. $\square$

Consider the multivariate tracking problem, when the distribution of Y is rotationally symmetric around some point (say 0) in $\mathbb{R}^n$ and the objective function is $E[-\|X - Y\|^2]$. The rotational symmetry implies of course that the support of Y is itself rotationally symmetric. Suppose that there is an optimal solution that concentrates probability on k values of X, $\{x_1, \ldots, x_k\}$ with probabilities $\{p_j\}$ and corresponding conditional distributions of Y, $\{p(y \mid x_j)\}$. Because of the rotational symmetry, a solution that rotated each of the k values of X around the center of symmetry through the same angle, while at the same time rotating the conditional pdf's of y through the same angle, would deliver the same value of the objective function. But then so would any mixture of these two solutions. That is, we could specify a probability π for the first solution and $1 - \pi$ for the second, and the result

would give the same objective function value and (because the expected reduction in entropy of the Y distribution is still the same) the same information cost. But then we can also construct arbitrary continuous mixtures of such rotated solutions and again achieve the same objective function values. A continuous mixture of rotated versions of a solution with finitely many points of support would have support concentrated on a finite collection of circles or (in higher dimensions) spheres. Here is our conclusion as a proposition.

Proposition VII.3. *In a rotationally symmetric multivariate tracking problem, if there is any solution that gives X finitely many points of support, there are also solutions with support a finite collection of spheres or circles.*

VII.3. **The risk-averse monopolist's problem.** Adapting the first order conditions for the general decision problem discussed in section VI to this case gives us

$$p(c \mid x) = ((x - c)q(x))^{1/\lambda} h(c) \tag{20}$$

$$\therefore \int_0^x ((x - c)q(x))^{1/\lambda} h(c) \, dc = 1 \quad \text{for any } x \text{ such that } p(x) > 0 \tag{21}$$

$$\int p(x) ((x - c)q(x))^{1/\lambda} h(c) \, d\mu_x(x) = g(c) \quad \text{all } c, \tag{22}$$

where h is some non-negative function and we are using the convention that $p()$ is a pdf, with its arguments showing what it is the pdf for.

With this pair of functional forms for utility and demand, it turns out to be possible to find a function $h(c)$ that satisfies (21) for all positive prices x. If we guess $h(c) = Kc^{\gamma}$, we find

$$\int_0^x \left((x - c)x^{-\theta} \right)^{1/\lambda} Kc^{\gamma} \, dc = Kx^{\frac{1-\theta}{\lambda} + \gamma + 1} B\left(\gamma + 1, \frac{1}{\lambda} + 1 \right), \tag{23}$$

where $B(\cdot, \cdot)$ is the beta function. Thus by choosing $K = 1/B\left(\gamma + 1, \frac{1}{\lambda} + 1 \right)$ and $\gamma = (\theta - 1)/\lambda - 1$, we can make the integral 1 for all values of $x > 0$. For the beta function to be well defined, we require $\gamma > -1$, but this is guaranteed by

the condition $\theta > 1$, which is required in any case for the monopolist's problem to have a non-trivial solution. We show in proposition VII.4 that this $h(c)$ function is unique; no other $h(c)$ can satisfy (23) over any interval of the form $(0, \bar{c})$ with $0 < \bar{c} \le \infty$.

This implies that for any distribution of x on $x > 0$, defined by a density function $\pi()$ and a measure μ_x, we can construct a function

$$g(c) = \frac{\int_c^\infty \pi(x) \left((x - c) x^{-\theta} c^{\theta - 1 - \lambda} \right)^{1/\lambda} \mu_x(dx)}{B\left(\frac{\theta - 1}{\lambda}, \frac{1}{\lambda} + 1 \right)}, \tag{24}$$

and this distribution defined by π, μ_x will be the marginal distribution of x in the monopolist's problem if this $g(c)$ is taken as the given marginal distribution on costs c. So, as we vary the choice of g, every kind of marginal distribution for x can emerge as solution to the problem — purely discrete with finite support, mixed continuous/discrete, purely continuous, bounded support or unbounded support.

The result and its proof is in the following proposition.

Proposition VII.4. *In the risk-averse monopolist problem of section V.2 where the utility of profits is $\log\left(x^{-\theta}(x - c) \right)$, if $g(y) > 0$ on an interval $(0, T)$ with $T > 0$, and the marginal density $p(x)$ of X is positive on an interval $(0, S)$ with $0 < S \le \infty$, the $h()$ function that solves (21) for all x in $(0, S)$ is $h(c) = Kc^\gamma$, with $\gamma = (\theta - 1)/\lambda - 1$ and $K = 1/\operatorname{Beta}(\gamma + 1, 1 + \lambda^{-1})$. This solution is unique.*

Proof. That this form of h does solve the equation is directly verifiable by substituting it into the equation and evaluating integrals. That it is unique follows because if there were another function $h^*()$ that also solved the equation, we could set $\tilde{h} = h - h^*$ and use (21) to derive the conclusion

$$x^{-\theta/\lambda} \int_0^x (x - c)^{1/\lambda} \tilde{h}(c) \mu_c(dc) = 0 \quad \text{all } x \in (0, S). \tag{25}$$

Denote by α the least positive real number such that $1/\lambda + \alpha = m$, where m is an integer. Then we have

$$0 = \int_0^y (y-x)^\alpha \int_0^x (x-c)^{1/\lambda} \tilde{h}(c) \mu_c(dc) dx$$

$$= \int_0^y \int_c^y (y-x)^\alpha (x-c)^{1/\lambda} dx \tilde{h}(c) \mu_c(dc)$$

$$\int_0^y \int_0^{y-c} (y-c-x)^\alpha x^{1/\lambda} dx \tilde{h}(c) \mu_c(dc)$$

$$= \int_0^y (y-c)^m \tilde{h}(c) \mu_c(dc) \int_0^1 (1-x)^\alpha x^{1/\lambda} dx$$

$$= \mathrm{Beta}(\alpha+1, 1/\lambda+1) \int_0^y (y-c)^m \tilde{h}(c) \mu_c(dc), \quad (26)$$

for all $y \in (0, S)$, and by taking the mth derivative, we conclude that

$$0 = \int_0^y \tilde{h}(c) \mu_c(dc) \quad (27)$$

for all $y \in (0, S)$, which implies that $\tilde{h}(c) = 0$ except on a set of μ_c-measure zero. $\square$

However, (24) shows us that any solution corresponding to this choice of Kc^γ as the form for $h(c)$ must be a weighted average of scaled beta densities. For each price x, the conditional distribution of costs has the shape of a beta density, scaled to the interval $(0, x)$, but with the same shape parameters for each x. This means, for example, that no density g for C with $g(c) = 0$ in some neighborhood of zero (i.e. no density that implies a known lower bound on C) has this form. If C has support with upper limit $\bar{c}$, an X distribution that satisfies (24) must have the same upper limit to its support. This might seem surprising, since without information costs it is always optimal to make $x = \theta c/(\theta - 1)$, which is larger than c. And indeed it often will be optimal to make $P[X > \bar{c}] > 0$ — but in this case, because (24) cannot hold, X cannot have full support from zero to its upper limit. A distribution for C that admits a solution in the family described by (24) and has an upper bound $\bar{c}$ to its support must imply that C has a finite, but positive, expectation. For any given θ, $E[C/X \mid X] \to 0$ as $\lambda \to \infty$. Therefore even if g does have the form (24) for some

θ and λ, with upper limit to its support $\bar{c}$, the optimal distribution of X must cease to have full support on $0, \bar{c}$ for large enough λ.

Whenever the solution fails to fall in the mixture-of-scaled-betas class, it will necessarily not have full support on a single interval. It may be that it always has discrete support, but the strongest result we have obtained is that the solution gives X support on a finite number of points whenever the support of the distribution of C is contained in an interval (c_1, c_2) with $c_2 < \theta c_1 / (\theta - 1)$, as shown in this proposition.

Proposition VII.5. *In the risk-averse monopolist problem of section V.2, if cost C has a distribution whose support is contained within an interval (c_1, c_2) with $\theta c_1 / (\theta - 1) > c_2$, the distribution of the optimal X is concentrated on a finite set of points.*

Proof. The optimal choice of X if C were known would be $\theta c / (\theta - 1)$. With an a priori known lower bound c_1 on C, it can never be optimal to let X go below $\theta c_1 / (\theta - 1)$. So if there is an upper bound c_2 on C with $c_2 < \theta c_1 / (\theta - 1)$ the support of the distribution of X will necessarily lie entirely above the support of C, and in fact within the interval $\theta c_1 / (\theta - 1), \theta c_2 / (\theta - 1)$. Because in this case X always exceeds C, the conditions i and ii of proposition VII.1 are satisfied, with the role of $v(x, y)$ in that proposition played by $x^{-\theta/\lambda} \cdot (x - c)^{1/\lambda}$. The $h(y)$ from that proposition is, from the first-order conditions, the function

$$h(c) = \frac{g(c)}{\int p(x) x^{-\theta/\lambda} (x - c)^{1/\lambda} \mu_x(dx)}. \tag{28}$$

The numerator of this expression is a probability density, hence integrable, and the denominator is bounded away from zero for $c \in (c_1, c_2)$, because the support of X lies above that interval. The $v(x, y)$ function is itself bounded and the domain of integration over c is bounded, so $v(x, y) h(y)$ is indeed jointly integrable as required

in iii. Therefore

$$\int_0^x x^{-\theta/\lambda} \cdot (x - c)^{1/\lambda} h(c) dc = \int_{c_1}^{c_2} x^{-\theta/\lambda} \cdot (x - c)^{1/\lambda} h(c) dc \tag{29}$$

is an analytic function of x on the interval containing the support of X. Note that the interval over which this function is analytic is (c_2, ∞), even though X has support bounded by a smaller interval. But this function must take the value 1 at every point of support of X. If the support of X contains any of its limit points, the function would have to be one over the whole (c_2, ∞) interval. But since $\theta > 1$, it is also clear that the function goes to zero as $x \to \infty$ and is therefore not constant. Since the support of X is inside a finite interval and contains no limit points, it must be a finite set of points. $\qquad\square$

VII.4. **Singular support in the portfolio allocation problem.** We require that $\Sigma_z + \mu_y \mu_y'$ be strictly positive definite, which implies that there is at most one risk-free asset and that that asset not produce a zero (i.e. -100%) return. There are no limits on short selling, so θ is unbounded despite $\theta' \mathbf{1} = 1$. The observable (with error) randomness Y is, before information collection, distributed with pdf $g(y)$, and has bounded support[10]. The $U(\theta, y)$ for our canonical problem then becomes here

$$U(\theta, y) = \theta'y - \tfrac{1}{2}(\theta'y)^2 - \tfrac{1}{2}\theta'\Sigma_z\theta. \tag{30}$$

We obtain this expression from the expected utility term in (15) by taking the expectation over Z, conditional on Y and θ. Then our $C()$ function from Proposition VI.1 is given by

$$C(\theta) = \int \exp\left(\lambda^{-1}\left(\theta'y - \tfrac{1}{2}(\theta'y)^2 - \tfrac{1}{2}\theta'\Sigma_z\theta\right)\right) h(y)\, dy, \tag{31}$$

---

[10]The bounded support assumption could be weakened to a requirement that $g(y)$ decline rapidly enough for large $\|y\|$ that $e^{b\|y\|^2} g(y) \to 0$ is integrable in y.

where λ as before is the utility cost of information. The first order conditions also give us

$$h(y) = \frac{g(y)}{\int p(\theta) \exp\left(\lambda^{-1}(\theta'y - \frac{1}{2}(\theta'(y))^2 - \frac{1}{2}\theta'\Sigma_z\theta)\right) d\mu_\theta(\theta)} . \tag{32}$$

Note that the denominator on the right side of (32) is bounded away from zero on the bounded domain of g, so $h()$ is integrable.

The function $\exp(U(\theta, y))h(y)$ is clearly analytic in θ for any fixed y, over the whole $n - 1$-dimensional hyperplane $\theta'\mathbf{1} = 1$. It is jointly integrable in y and θ because

$$\int \exp\left(-\frac{1}{2\lambda}(\theta'(\Sigma_z)\theta)\right) \exp\left(\lambda^{-1}(\theta'(y) - \frac{1}{2}(\theta'y)^2)\right) h(y)\, dy\, d\theta <$$

$$\int \exp\left(-\frac{1}{2\lambda}(\theta'(\Sigma_z)\theta)\right) \exp(\tfrac{1}{2\lambda}) \left(\int h(y)\, dy\right) d\theta . \tag{33}$$

The integral in y in this expression is a finite constant. The Σ_z matrix is singular, but θ ranges only over the submanifold where $\theta'\mathbf{1} = 1$, and on that set $\theta'\Sigma_z\theta = \sigma_z^2(\theta_2^2 + \theta_3^2)$, where θ_2 and θ_3 are the portfolio allocations to the two risky assets. $-\frac{1}{2}\theta'\Sigma_z\theta$ therefore goes to $-\infty$ as $\|\theta\|^2 \to \infty$, and the term in θ therefore also has a finite integral.

Now we can apply Proposition VII.1 to conclude that $C()$ is analytic.

Then because as we have already observed $C(\theta) \to 0$ as $\|\theta\|^2 \to \infty$, $C()$ is not constant. We can conclude that it satisfies $C(\theta) = 1$ only on a subset B of the hyperplane $\theta'\mathbf{1} = 1$ that contains no open sets and intersects straight lines in only a countable set of points with no limit points.

VIII. CONCLUSION

The kind of model explored here, in which a decision-maker reacts to external information subject to a tight information constraint, seems a natural one to apply to the behavior of individuals reacting to fairly frequent and/or numerous economic signals for which the consequences of imprecise responses are modest. This

might be true, for example, of price-setters in retail establishments that must set hundreds of prices every day in response to fluctuations in demand and costs for all the items. It might also be true of day-to-day or month-to-month savings and spending decisions of individuals, facing a potentially vast array of information about asset markets.

We show that reduced dimensionality of the support of the decision variable emerges in a large class of problems. Propositions VI.1 and VII.1 provide general tools for assessing it, and we apply these tools in our examples. Proposition VII.2 shows reduced dimensionality in a broad class of cases where exogenous uncertainty has bounded support and the objective is to keep a decision vector close to the exogenous random vector. The reduced dimensionality results imply discrete distributions in models with one-dimensional decision variables, and our numerical examples show that discreteness also commonly emerges in multivariate examples.

These models suggest that it is a mistake to identify times at which decision makers' choices change as times at which they fully optimize in reaction to the current state. Rationally inattentive decision makers as modeled here may change their choices randomly, even when the state of the world (the draw from the g distribution) is unchanged. As modeled here, they acquire information about the true state every period, but may nonetheless not change their behavior, even though the true state is changing.

Rationally inattentive decision makers as in our portfolio allocation or risk-averse monopolist examples may go for long periods making no change or small changes in their behavior, then make a large and temporary change. Long periods of unchanged behavior are not an indication that changing behavior is "costly", frequent small changes are not an indication that big changes are costly, and the large, rare changes are not an indication that changing behavior is not costly. There is no cost

of change at all in these models. Apparent inertia in the face of changing circumstances simply reflects the fact that with information flows valuable, it can make sense to concentrate attention on the rare extreme draws of y, reacting little or not at all to the usual small fluctuations about the central value.

If behavior like this explains even part of observed inertia and stickiness in economic behavior, conclusions from models that use adjustment costs, menu costs, or an ad hoc assumption of infrequent but complete information updates, could give misleading conclusions. Formal solution of optimization problems with an information constraint is challenging, even in the most manageable linear-quadratic, Gaussian uncertainty case, and extending the solution methods this paper has used for non-quadratic, non-Gaussian settings to dynamic models is even more challenging. Furthermore, as our examples have shown, the nature of solutions to these problems is in some dimensions sensitive to small changes in the problem.

It is therefore not realistic to expect that formal models incorporating rational inattention can soon simply replace standard rational expectations models. On the other hand, that rational inattention leads to muted, and sometimes to "sticky" responses to changes in the state of the world is a robust result. We should recognize, therefore, that structural models of rational behavior under uncertainty that ignore information processing costs, need to be taken with a grain of salt. Their implied "costs of adjustment" may not correspond to actual components of technology or utility functions.

APPENDIX A. NUMERICAL METHODS

There is a well-known algorithm in the rate-distortion literature, known as the Blahut algorithm (Blahut, 1972), for calculating the optimal distribution of inputs in a completely discrete version of the problem. In the notation of our (2)-(4), the Blahut algorithm works with the case where Y has a given discrete distribution and the points of support of the X distribution are given. The $U(x, y)$ function can then

be characterized as a fixed matrix, and the first-order conditions (7) and (8) are used to generate a fixed-point algorithm that solves the optimization problem.

For the problems that interest us in this paper, though, both X and Y are best thought of as having arbitrary distributions in finite-dimensional Euclidean spaces. Even if the solution for X gives it finitely many points of support, as in many of our examples, the location of those points of support is not known a priori. One possibility is to choose a fine grid in both X and Y spaces. One can then either solve the problem as a constrained maximization (constrained because of the requirement that $p(X) > 0$) or use the Blahut fixed point algorithm. With either approach, the need for a fine grid makes the problem high-dimensional (or inaccurately approximated). Furthermore, when the true solution has support on a fairly small number of points, the solution with this grid approach is unlikely to make the discreteness of the support cleanly apparent, as the points of support in the underlying continuous problem will not usually lie exactly on the grid.

When Y is given a dense grid, to approximate a continuous distribution, but X emerges as having a modest number (say 2-20) points of support, we have found it efficient to solve by initially fixing the number of points of support of X and iterating to a fixed point by methods analogous to the Blahut algorithm, but with an added step of optimizing the location of the points of support at each round of the iteration. The solution to the optimization problem, subject to the given bound on the number of points of support, is a fixed point of this algorithm. Experimenting with different numbers of points of support can lead us to be fairly confident that we have found the optimum number.

One could also take an optimization approach, optimizing jointly over the probabilities on the points of support and their locations, and we experimented with this approach. However, in this problem when the number of points of support is larger than required, the objective function is flat in some dimensions. Even when

the number of points of support is correctly chosen, the objective function can be close to flat in some dimensions, which makes gradient-based optimization slow, erratic, and often inaccurate.

Our fixed-point algorithm can also be slow in terms of iteration count, but each iteration is quick. When the number of points of support is larger than necessary, the algorithm converges either with $p(x_i) = 0$ on some points of support or with some points of support showing the same values of x_i and the same conditional distributions for $Y \mid X = x_i$, possibly with non-zero probabilities on several of the equivalent x_i's. The probabilities from these repeated values of x in the solution are then added up to obtain the implied solution with fewer points of support.

Like the Blahut algorithm, our algorithm does not guarantee a global solution. Suppose we solve to find the optimum distribution of X over N_0 points, but then expand the model by adding extra points of support up to $N > N_0$. If we start any of these algorithms from the solution for N_0 points, either by putting 0 probability initially on the added points, or by giving them positive probability but repeating them, none of these algorithms will leave the starting value. Furthermore, they can converge to a solution with fewer points of support than the optimum. We therefore always check apparently converged solutions by randomizing starting values.

Our algorithm is implemented in four R functions, two of which are specific to the particular problem being solved. One of the problem-specific functions takes a vector of x values and a vector of y values and computes the corresponding value of $U(x, y)$. The other takes as one argument an $nx \times ny$ matrix of conditional probabilities for $y \mid x$, where nx is the number of points of support of x (generally 2-20) and ny is the number of points of support of y (usually on the order of 1000-2000 in our examples). Its second argument is the $ny \times my$ matrix giving the my-dimensional vector of y values at each point of support in the distribution of y. This function

returns the optimal values of x at its nx points of support. It may be easiest to understand what these functions do by looking at them in the simplest case, our multivariate tracking problem, where they take easily understood form:

```
U2drec <- function(x,y) {
  -.5 * sum((x-y)^2)
}


xfcn2drec <- function(ygivenx, y) {
  return(ygivenx %*% y)
}
```

The first of these returns the squared Euclidean distance between y and x, and the second returns the conditional expectation of Y at each of the points of support of the X distribution, which is the optimal choice for these values with the given conditional distributions.

These two functions, one giving the value of U, the other giving the optimal choices of x for given conditional distributions on Y, are passed as arguments to `DiscFP()`, which monitors the iterations, and in turn to `DiscPObjXmv()`, which does the computation for each iteration.

All the examples we consider are ones where there is a closed form solution to the problem of choosing the optimal x value given the conditional distribution of $y \mid x$, or else the solution is a well-behaved one-dimensional zero-finding problem. If this optimization itself required a time-consuming iterative solution, our approach would require modification.

The full set of programs to compute the solutions in our examples is available in the on-line appendix.

REFERENCES

BLAHUT, R. (1972): "Computation of Channel Capacity and Rate Distortion Functions," *IEEE Transctions in Information Theory*, IT-18, 460–473.

BONAPARTE, Y., AND R. COOPER (2009): "Costly Portfolio Adjustment," working paper 15227, NBER.

CAPLIN, A., AND M. DEAN (2015): "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, 105(7), 2183–2203.

COVER, T. M., AND J. A. THOMAS (1991): *Elements of Information Theory*. Wiley-Interscience.

CSISZÁR, I. (2008): "Axiomatic Characterizations of Information Measures," *Entropy*, 10(2008), 261–73.

DOSS, R. (1988): "An Elementary Proof of Titchmarsh's Convolution Theorem," *Proceedings of the American Mathematical Society*, 104(1), 181–184.

EICHENBAUM, M., N. JAIMOVICH, AND S. REBELO (2008): "Reference Prices and Nominal Rigidities," Discussion paper, Northwestern University and Stanford University, NBER Working paper 13829.

FIX, S. L. (1978): "Rate Distortion Functions for Squared Error Distortion Measures," Mimeo.

GABAIX, X. (2011): "A sparsity-based model of bounded rationality," Discussion paper, National Bureau of Economic Research.

KRANTZ, S. G. (1992): *Function Theory of Several Complex Variables*. AMS Chelsea Publishing, 2nd edn.

KYTHE, P. K., AND P. PURI (2002): *Computational Methods for Linear Integral Eauations*. Birk, Boston, Basel, Berlin.

LUO, Y. (2008): "Consumption dynamics under information processing constraints," *Review of Economic Dynamics*, 11(2), 366 – 385.

MACKAY, D. J. C. (2003): *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

MAĆKOWIAK, B., AND M. WIEDERHOLT (2009a): "Business Cycle Dynamics under Rational Inattention," Discussion paper, European Central Bank and Northwestern University, http://faculty.wcas.northwestern.edu/ mwi774/RationalInattentionDSGE.pdf.

——— (2009b): "Optimal Sticky Prices under Rational Inattention," *American Economic Review*, 99(3), 769–803.

MANKIW, N. G., AND R. REIS (2002): "Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve*," *Quarterly Journal of Economics*, 117(4), 1295–1328.

MATĚJKA, F. (2008): "Rationally Inattentive Seller: Sales and Discrete Pricing," Discussion paper, PACM, Princeton University, `http://home.cerge-ei.cz/matejka/RI_seller.pdf`.

——— (2009): "Rigid Pricing and Rationally Inattentive Consumer," Discussion paper, Princeton University.

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

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

MOSCARINI, G. (2004): "Limited information capacity as a source of inertia," *Journal of Economic Dynamics and Control*, 28(10), 2003–2035.

REIS, R. (2006): "Inattentive producers," *The Review of Economic Studies*, 73(3), 793–821.

ROSE, K. (1994): "A Mapping Approach to Rate-Distortion Computation and Analysis," *IEEE Transactions on Information Theory*, 40(6), 1939–1952.

SIMON, H. A. (1976): *Administrative Behavior*. Cambridg.

——— (1979): "Information processing models of cognition," *Annual review of psychology*, 30(1), 363–396.

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

——— (2006a): "Rational Inattention: A Research Agenda," Discussion paper, Princeton University, `http://sims.princeton.edu/yftp/RIplus/`.

——— (2006b): "Rational Inattention: Beyond the Linear-Quadratic Case," *American Economic Review*, 96(2), 158–163.

——— (2010): "Rational Inattention and Monetary Economics," in *Handbook of Monetary Economics*. Elsevier, http://sims.princeton.edu/yftp/RIMP.

TODD, P. M., AND G. GIGERENZER (2000): "Précis of Simple heuristics that make us smart," *Behavioral and Brain Sciences*, 23, 727–780.

TUTINO, A. (2009): "The Rigidity of Choice: Lifetime Savings under Information-Processing Constraints," Ph.D. thesis, Princeton University, `http://docs.google.com/fileview?id=` `0B7CdO9AORsjcNWYwZmM1MWEtNDZiNi00NzQzLTgzOTItZmNiM2IzOWQ3MDhh&` `hl=en`.

VAN NIEUWERBURGH, S., AND L. VELDKAMP (2010): "Information Acquisition and Under-Diversification," *Review of Economic Studies*, 77(2), 779–805.

WOODFORD, M. (2009): "Information-Constrained State-Dependent Pricing," *Journal of Monetary Economics 56(S): 100-124 (2009)*, 56(S), 100–124.

YANG, M. (2015): "Coordination with flexible information acquisition," *Journal of Economic Theory*.

DEPARTMENT OF MATHEMATICAL SCIENCE, KAIST, DAEJEON 305-701, SOUTH KOREA

E-mail address: junehyuk@kaist.ac.kr

DEPARTMENT OF ECONOMICS, PRINCETON UNIVERSITY

E-mail address: jk@princeton.edu

CERGE-EI, A JOINT WORKPLACE OF THE CENTER FOR ECONOMIC RESEARCH AND GRADU-ATE EDUCATION, CHARLES UNIVERSITY, AND THE ECONOMICS INSTITUTE OF THE ACADEMY OF SCIENCES OF THE CZECH REPUBLIC., POLITICKYCH VEZNU 7, PRAGUE 11121, CZECH REPUBLIC.

E-mail address: filip.matejka@cerge-ei.cz

DEPARTMENT OF ECONOMICS, PRINCETON UNIVERSITY

E-mail address: sims@princeton.edu