Review of Economic Studies (2020) **01**, 1–44 0034-6527/20/00000001$02.00

© 2020 The Review of Economic Studies Limited

E-mail: junehyuk@math.tamu.edu E-mail: jkim230@emory.edu E-mail: filip.matejka@cerge-ei.c

E-mail: sims@princeton.edu [1]

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 "channel capacity". Modeling economic agents as constrained by Shannon capacity as they process freely available information 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 an investor choosing portfolio allocations, as well as to some mathematically simpler "tracking" problems that illustrate the mechanism. Trying to use costs of adjustment to explain "stickiness" of actions when interpreting the behavior in our economic examples would lead to mistaken conclusions.

Discrete actions in information-constrained decision problems

JUNEHYUK JUNG

Department of Mathematics, Texas A&M University

JEONG HO (JOHN) KIM

Department of Economics
Emory University

FILIP MATEJKA

CERGE-EI
A joint workplace of the Center for Economic Research and Graduate Education,
Charles University, and the Economics Institute of the Academy of Sciences of the Czech Republic.
Politickych veznu 7
Prague 11121, Czech Republic.

CHRISTOPHER A. SIMS

Department of Economics
Princeton University

1. This work was partially supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370, by NSF grant SES-0719055, by the grant Agency of the Czech Republic,

We begin in section I with a discussion briefly motivating the use of Shannon's information measure to study the effects of limited attention on optimization under uncertainty. More detailed motivation for the Shannon measure, and its relation to other measures of information cost, appears later, in section VIII. In section II we relate this paper to some previous literature. In sections III and IV we specify the general static, information-constrained decision problem that is our focus and some intermediate results useful for solving it. In section V we characterize some classes of problems where we can guarantee the solutions make the distribution of actions discrete. In that section we also provide simple examples that may guide intuition about the nature of the discreteness result and how it arises. In section VII we show analytically that a problem of portfolio choice under uncertainty, with costly information, gives rise to actions concentrated on a discrete set of portfolios and also provide numerical solutions to a version of that problem. The results show that an empirically observed pattern, in which investors pay attention to relative yields in normal times, but occasionally switch to a "risk off" or "risk on" mode with less attention to individual asset yields, can emerge from rational inattention. An online appendix includes more numerical examples, including one that considers a non-Shannon information cost measure and one that illustrates how Shannon's coding theorem applies in practice, in a context where our discreteness results apply.

Section VII, on portfolio choice, has the most interesting economic insights, and can be read for its results, independently of the sections preceding it that use simpler or more abstract models to explain how discrete behavior arises.

I. STICKINESS FROM COSTLY INFORMATION PROCESSING

Many economic variables are constant for spans of time, then jump to new values. Prices of some products and portfolio allocations of individual investors are examples. We have simple theories that imply such behavior is optimal (explicit menu or adjustment costs models) or that treat such behavior as a constraint (Calvo pricing). As fine-grained microdata on individual product prices have 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, but when they do change, they sometimes move back and forth across a finite array of values.[1] These simple models explain the fixity for spans of time, but on the face of it imply that when a price change occurs, the change should be continuously distributed.

project P402/12/G097 DYME, and by the European Research Council under the EU's Horizon 2020 research and innovation programme (grant agreement No. 67801). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF)

1. See Eichenbaum et al. (2011); Stevens (2015).

They do not explain why the price should change, then come back to *exactly the same* price as before the change, for example.

Rational inattention is the idea that a rational person, confronted with the need to take actions in response to a stream of randomly varying observations, must recognize that she is a finite capacity Shannon channel. Roughly speaking, this means she recognizes that there is a bound on how precisely she can make her actions respond to a rapid flow of information, and that allocating some of her information-processing "capacity" to an economic decision problem is therefore costly. The theory examines the nature of optimal behavior subject to this cost. In section VIII below we discuss in more detail the nature of a Shannon channel and the associated measure of information processing costs. We also discuss how it relates to other measures of information processing costs that have been proposed.

We show that, with rational inattention, in broad classes of cases, the action is not a function of the underlying state, but is random instead, even conditional on the underlying state. Furthermore, in broad classes of cases the distribution of the action is concentrated on a set of lower dimension than would emerge without the information costs. Where action and state are both scalars, the continuous one-dimensional distribution of actions that would emerge without information costs can become a zero-dimensional one, i.e. one whose support is a finite, or in some cases countable, set of points.

The question of whether "stickiness" of actions reflects something like menu costs, or instead optimal allocation of limited Shannon capacity, is important for macroeconomic policy modeling. If the stickiness has mistakenly been modeled as being due to a stable adjustment cost in a standard rational expectations model, the adjustment costs will appear to change as the stochastic process driving the economy changes. This 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 importantly, models that explain stickiness via adjustment costs of some sort imply that rapid change in sticky choice variables is costly or distasteful. Models with rational inattention, which explain stickiness as reflecting information processing costs, do not imply that rapid change is in itself costly or distasteful — response to incentives may be slow or damped only because agents are not fully aware of changes in their environment. On the other hand, information processing costs may be high even when decision variables are not changing, and such costs are not captured in the behavior of direct arguments of production and utility functions. Rational inattention therefore does not necessarily imply that cyclical fluctuations are more or less important for welfare than they are in adjustment-cost models, but it could imply quite different estimates of welfare costs and, in turn, different conclusions about optimal policy.

II. RELATED PREVIOUS LITERATURE

In Sims (2003a) the effects of finite capacity on behavior were examined in dynamic linear models with quadratic objective functions and Gaussian disturbances, similar to those appearing in many macroeconomic models. Several other researchers have also worked with Gaussian prior uncertainty (Moscarini, 2004; Mackowiak and Wiederholt, 2009; Maćkowiak and Wiederholt, 2015; Luo, 2008; Hellwig and Veldkamp, 2009; Van Nieuwerburgh and Veldkamp, 2010; Mondria, 2010; Myatt and Wallace, 2011). In most of these papers the objective is quadratic, for which Gaussian ex post uncertainty about the state is optimal. In such models the distribution of actions also often emerges as continuous.

In the present paper we allow for non-Gaussian distributions of fundamental uncertainty and non-quadratic objective functions (see also Sims (2006); Woodford (2009); Tutino (2013); Matějka (2015); Yang (2015); Matějka and McKay (2015); Caplin and Dean (2015).) Our findings are closely related to some earlier results in rate-distortion theory in engineering (Fix, 1978; Rose, 1994). That literature[2] studies a version of our static, information-constrained decision problem. Both Fix and Rose obtained versions of our results for one-dimensional cases and quadratic utility. The main technical differences between these earlier results and our work are that our results extend to more general cases, beyond quadratic loss functions, and to multivariate cases.

The recent paper most closely related to ours are Caplin et al. (2018) and Matějka and McKay (2015). They consider a one-period decision problem with Shannon information costs as we do. They restrict both the state space and the action space to be finite sets of a priori fixed points. They find that solutions frequently put probability zero on many of the elements of the action space, which corresponds to our result that with continuously distributed state and subsets of $\mathbb{R}^n$ as action spaces, probability frequently concentrates on finite sets of points or low dimensional subsets of the action space. Of course their framework has no notion of the dimensionality of the support of the distribution of actions. We discuss some further parallels and contrasts between their results and ours in section VI.

The general model we consider includes as special cases the decision problems of rationally inattentive price-setters studied in Matějka (2016) and Stevens (2015). In those papers, where monopolistic sellers process information about several types of shocks (demand, input cost, etc.), numerical solutions exhibit discreteness. Our analytical results provide the reassurance that the apparent discreteness in the computational solutions in previous works was not an anomaly, and may also give us insight into the conditions under which solutions with rational inattention are likely to emerge as discrete. Moreover, because this paper's setup covers multivariate actions, it applies to

2. See Berger (1971) and Cover and Thomas (2006)

interesting economic models not considered in earlier papers, including a portfolio choice problem that we consider in some detail.

Finally, note that other measures of information costs, or formulations of information processing constraints, have been used in macroeconomics, for example in Mankiw and Reis (2002); Alvarez et al. (2011). Veldkamp (2011) provides a review. Some of our discussion in section VIII compares these other approaches to ours, and in section Appendix C.4 of the online appendix we discuss a particular non-Shannon information measure in some detail.

III. THE GENERAL STATIC DECISION PROBLEM WITH COSTLY SHANNON CAPACITY

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, 2003b), 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 independently repeated frequently over time, or as part of a large collection of similar, independent problems being solved simultaneously and drawing on a common Shannon capacity resource.

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

$$\max E[U(X,Y)] - \lambda I(X,Y)\,, \tag{III.1}$$

where the maximization is over the joint distribution of the random variables (or vectors) X, Y. X is the maximizing agent's action, and Y is the state. $U(\ ,\ )$ is the agent's objective function, and $I(\ ,\ )$ is the Shannon mutual information between X and Y. The distribution of Y before information is collected is given: we assume it is defined by a density $g(y)$ over a base measure μ_y, which in most cases will be ordinary Lebesgue measure. Without an information constraint, the optimal x will simply be a function of y, so the joint distribution is singular, but with an information cost generally x must be chosen with Y still uncertain.

To make it more explicit what is being chosen and what the information cost is, we can rewrite this as

$$\begin{aligned}\max_{f,\mu_x} \quad & \int U(x,y) f(x,y)\, \mu_x(dx)\, \mu_y(dy) \\ - \quad & \lambda \left(\int \log(f(x,y)) f(x,y)\, \mu_x(dx)\, \mu_y(dy) \right. \\ & \left. - \int \log\Big(\int f(x,y')\, \mu_y(dy') \Big) f(x,y)\, \mu_x(dx)\, \mu_y(dy) \right)\end{aligned} \tag{III.2}$$

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

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

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 (III.2) is the expectation of U, and the second is the cost of channel capacity. (III.3) requires consistency of prior and posterior beliefs, and (III.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 Y, 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 μ_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 for the case of univariate X and Y and $U(x,y) = -(x-y)^2$ a version of our result in section V, 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 first order conditions (FOC's) of the problem with respect to f imply that at all values of x,y

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.

with $f(x,y) > 0$ and $g(y) > 0$

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

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

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

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

where $\theta(y)$ is the Lagrange multiplier on the constraint (III.3), p is the density with respect to μ_x 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 (III.5) be less than or equal to the right hand side. Note that if $p(x) = \int f(x,y)\,\mu_y(dy) > 0$, the right hand side of (III.5) is minus infinity wherever $f(x,y) = 0$, so with $U(x,y) > -\infty$ everywhere, we can conclude that $f(x,y) = 0$ for a particular x, y only if $f(x,y) = 0$ for all y, i.e. $p(x) = 0$. This implies immediately

Lemma 1. *If $U(x,y) > -\infty$ everywhere and $\lambda > 0$, then at all points x in the action space that have positive density $p(x)$, the support of the conditional density of y, $q(\mid x)$, is the same as the support of the marginal density of y, $g()$.*

Of course this result means that even when the optimal distribution of x is discrete, the joint distribution of y with x is never quantized. That is, it is never optimal to partition the support of $g(y)$ and make the information collected about y simply which set of the partition contains y.

At points x with $p(x) = 0$, the right-hand side of (III.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

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

Since the information cost depends on $q(y \mid x)$ only through its shape as a function of y, it is intuitively clear that x will always optimally maximize $E[U(x,y) \mid x]$. That is

$$x = \text{argmax}_{x'} \int U(x',y)q(y \mid x)\,\mu_y(dy)\,. \tag{III.10}$$

IV. SOLVING THE STATIC INFORMATION-CONSTRAINED DECISION PROBLEM

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

$$C(x) = \int e^{\lambda^{-1}U(x,y)}h(y)\mu_y(dy) = 1\,. \tag{IV.11}$$

The support of the distribution of x in the optimal solution is therefore either equal to, or contained within, the set

$$B = \{x \mid C(x) = 1\} \ . \tag{IV.12}$$

If we can show that $C()$ is analytic, we can use the properties of analytic functions to determine properties of the set B, and hence of the support of x in the solution.

The following proposition states well known properties of analytic functions that we need for our main results.

Lemma 2. *An analytic function defined on on a connected open set S in $\mathbb{R}^n$ that is constant on any open subset of S is constant on all of S. If $n = 1$, the function is constant on all of S if it is constant on any sequence of x values with an accumulation point in x.*

It follows immediately that if in the information-constrained decision problem we can show that $C()$ is analytic on an open set S and that C cannot be constant on S, the support of x contains no open sets. If furthermore x is one-dimensional, the support of x contains no accumulation points.

If U is analytic, it will often be true that we can show that $C()$ is also analytic, but because the definition of $C(x)$ involves $h(y)$, which is part of the problem solution, not a priori given, showing that C is analytic generally involves problem-specific arguments.

A result[4] that is useful in proving analyticity of $C()$ in many cases is

Proposition 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);

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 $v(x,y)$ is integrable jointly in x and y over $S^* \times \mathbb{R}^m$, with respect to the product of Lebesgue measure on S^* with a measure μ_y on $\mathbb{R}^m$.

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

Corollary 1.1. *If*

i π is a probability measure on $\mathbb{R}$;

ii γ is an integrable, positive analytic function on $\mathbb{R}$;

iii and the radius of convergence of γ's Taylor expansion is bounded away from zero on $\mathbb{R}$;

4. The proof of this proposition, and that of several other theorems, corollaries and propositions below, appear in appendix Appendix A.

then the convolution $\pi * \gamma$ is analytic.

V. TRACKING

Tracking problems have $U(x, y) = V(x - y)$, with V maximized when its argument is zero. That is they are problems in which, without an information cost, the solution simply sets $y = x$. We discuss tracking separately in this section because many economically interesting examples take this form, because we can state a simple and fairly general set of assumptions under which in these problems discretely distributed x must arise, and because for these problems we can easily construct examples that give insight into when discreteness does and does not arise.

Here is the theorem.

Theorem 1 (Discreteness for tracking with bounded support). *Suppose*

a $V() : \mathbb{R}^m \to \mathbb{R}$ *is analytic on all of* $\mathbb{R}^m$*, with the radius of convergence of the Taylor series of* e^V *bounded away from zero on* $\mathbb{R}$*;*

b $V(a)$ *is maximized at* $a = 0$*;*

c $M(a) = \max_{\|z\| \geq a} V(z)$ *satisfies* $M(a) \to -\infty$ *as* $a \to \infty$*; and*

d Y *has bounded support in* $\mathbb{R}^m$*.*

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

Note that when $V(z) = -\|z\|^2$, e^V has no singularities, so its radius of convergence is infinite everywhere.

In tracking problems the integral equations in the first order conditions can be studied with Fourier transforms, because they become convolutions. Using this fact allows us to easily check whether solutions must be discrete in a large class of problems. We can write (III.7) and (III.8) for tracking problems as

$$\int e^{V(x-y)/\lambda} h(y) \mu_y(dy) = (e^{V/\lambda} * h)(x) = 1 \tag{V.13}$$

$$\int p(x) e^{V(x-y)/\lambda} \mu_x(dx) h(y) = (p * e^{V/\lambda})(y) h(y) = g(y)\,. \tag{V.14}$$

The first of these, (V.13) always has as one solution a constant $h(y) \equiv 1/\kappa$, where $\kappa = \int e^{V(z)/\lambda} dz$. 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 equation (V.14), that

$p * e^{V/\lambda} = g/\kappa$. We can calculate the implied p from this equation using Fourier transforms, and if the result is a probability distribution, it is the solution. Otherwise, the support of x cannot be the whole of the space over which y and x are defined. In that case, if V and g are analytic, it is likely that the solution is discrete, though additional problem-specific arguments are generally required to prove this.

This leads us to

Proposition 2. *In a tracking problem with $e^{V/\lambda}$ integrable, defined on the whole of $\mathbb{R}^n$, and having a Fourier transform everywhere non-zero,*

i when $\tilde{g}/\widetilde{e^{V/\lambda}}$ is the Fourier transform of a probability distribution, the solution gives the conditional distribution of $y \mid x$ a density proportional to $e^{V/\lambda}$;

ii otherwise the support of the distribution of x does not include the whole of $\mathbb{R}^n$.

Proposition 2 gives us a simple approach to tracking problems where $e^{V/\lambda}$ is everywhere nonzero, as it is when V is quadratic: Form $\tilde{g}/\widetilde{e^{V/\lambda}}/\lambda$, take its inverse Fourier transform. If the result is a probability distribution (which could be discrete), we have the solution's distribution for x. The solution then makes y the sum of x and a random variable independent of x with pdf proportional to $e^{V/\lambda}$. If the result is not a probability distribution, the support of x in the solution omits some part of $\mathbb{R}$.

Using this result allows us easily to get analytic results in some examples that may help guide intuition.

V.1. *The support of x need not be discrete*

While there are many examples of versions of our decision problem in which the solution gives x a continuous distribution, we state two well known ones here explicitly.

With quadratic objective function and Gaussian g, the solution will make x normal, unless information costs are so high that it is optimal to make x fixed at $E[y]$.

Corollary 2.1. *With a one-dimensional quadratic tracking V, when $Y \sim N(0, \sigma^2)$, X is normal and continuously distributed unless $\lambda > 2\sigma^2$, in which case no information is collected and $X \equiv 0$.*

Another well-known result is that with multivariate quadratic-loss tracking and y normal, we get the "water-filling" result.[5]

5. See Cover and Thomas (2006), p.348-9

Corollary 2.2. *If $y \sim N(0, \Sigma)$ and $V(z) = -\|Z\|^2$, the solution makes x normal with covariance matrix $\Sigma - 2\lambda I$, so long as λ is small enough that this matrix is positive semi-definite. With larger values of λ, x is normal, but with a singular covariance matrix whose rank decreases as λ increases.*

When the covariance matrix of x is singular, it lies in a linear space of lower dimension than the support of y, but in this quadratic-Gaussian case, x is never discretely distributed unless λ is so large that it is optimal to keep x fixed at the unconditional $E[y]$.

V.2. *Boundedness of the support of y is neither necessary nor sufficient for discreteness of x*

Theorem 1 uses the boundedness of the support of y to argue that $C(x)$ can't be constant. It could easily be extended to cover cases where instead $g(y)$ declines in the tails fast enough relative to $e^{V/\lambda}$, while still having full support on $\mathbb{R}^n$.

But there also cases where discrete x emerges despite g declining much more *slowly* in the tails than $e^{V/\lambda}$. For example, suppose g takes the form of a convolution of a discrete distribution with fat tails, say with probabilities proportional to $1/(1+n^2)$, with a normal distribution with variance σ^2. Then if $V = (x-y)^2$ and $\lambda = 2\sigma^2$, Proposition 2 tells us that the discrete distribution we used in the construction of g is the distribution of x. It seems likely that the distribution of x remains discrete for higher values of λ, though we have not shown this analytically. Proposition 2 guarantees only that for higher values of λ x will not have full support on $\mathbb{R}$.

V.3. *Similar decision problems may lead to sharply different x distributions*

In fact, there are cases where the solution does not determine the distribution of x uniquely, and there are both discrete x and continuous x solutions. Here is a simple example where this happens. It may provide some intuition about how discreteness in the solution arises.

Suppose y is in $(0, 2\pi)$ and $d(x,y) = \min(|y-x|, 2\pi - |y-x|)$. We can think of this as a problem where y is on the unit circle and d is the shortest arc between x and y. Suppose the objective function is

$$U(x,y) = \begin{cases} 0 & d(x,y) < a \\ -\infty & d(x,y) > a \end{cases}.$$

So this is a tracking problem on the unit circle, with zero losses if x is within a radians of y and infinite losses otherwise. Suppose g is uniform on the unit circle. A solution is then easy to see without any computation: Make x uniformly distributed on the circle and make $q(y \mid x)$ uniform on the arc of length $2a$ centered at x This makes losses zero, and has as information cost the difference in entropy between a uniform distribution on $(0, 2\pi)$ and a uniform distribution on $(-a, a)$, which is $\log(\pi/a)$.

This is the solution for any λ, since any smaller information cost would leave some probability of $d(y,x) > 1$, and thus infinite losses.

Suppose $a = \pi/n$ for some integer $n > 2$. Then there is another family of solutions, with the same 0 expected loss and the same information cost, but with discretely distributed x. In these solutions, x is distributed discretely on n equally spaced points on the circle, and as before $q(y \mid x)$ is uniform on an interval of length $2\pi/n$ centered at x, for every x in the support of its distribution. Because q is the same as in the continuous-x solution, the information cost, which is the difference between the entropy of the uniform initial distribution of y and the entropy of q, is the same in the continuous-x solution and each of the possible discrete-x solutions.

From the point of view of the decision maker, these solutions are all quite similar: She will get a signal telling her that y is in an interval of length $2\pi/n$ and choose as x the center of that interval. It is not important whether she knows that the signal will specify one of n a priori fixed intervals, or instead an interval of the same length randomly positioned on the circle.

Contrast this with what happens if we cut the circle, so that $d(x,y)$ is now simply $|x-y|$ Then a solution with continuously distributed x is no longer optimal. For x values near 0 or 2π, to avoid infinite losses we would need $q(y \mid x)$ to have support close to π/n in length, rather than $2\pi/n$, and thus with greater information cost. But one solution with discretely distributed x is still available: put the support points at $\pi/n, 3\pi/n, \ldots, (2-1/n)\pi$, with q uniform on an interval of length $2\pi/n$ centered at each of these points. This delivers the same zero loss, at the same information cost, as the solution for the original circle problem, and this solution is unique.

Though this example, with its infinite losses and non-analytic V, does not fit in the framework of Theorem 1, the way the discrete solution arises is similar to the mechanism operating in that theorem, with its bounded support for g. If we start with a problem with g spread over all of $\mathbb{R}$ and a solution that makes x continuous (like quadratic-Gaussian tracking), but then truncate g, we find the continuous-x solution does not work at the boundaries. Analyticity of C then forces the solution to be discrete.

We can illustrate this by contrasting a solution of the one-dimensional quadratic tracking problem with a normal distribution for y with a numerical solution of the same problem with a truncated normal distribution for y. Formally the problem is the one-dimensional version of the tracking problem of Theorem 1, with $V(z) = -z^2$ and g either a $N(0,\sigma_y^2)$ density or that same density truncated at $y = \pm 3\sigma_y$. It is well known that without the truncation in this problem it is optimal to make the joint distribution of X, Y Gaussian[6]. Furthermore if ω^2 is the conditional variance of $Y \mid X$

6. See Sims (2003a) or Cover and Thomas (2006).

(which is of course constant for a joint normal distribution), $I(X,Y) = \frac{1}{2}\log(\sigma_y^2/\omega^2)$ (where the units are bits if the log is base 2 and nats if the base is e). It is also optimal to make $E[Y \mid X] = X$.

With no information cost and no truncation, 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, but still no truncation of Y's normal distribution, we can use the characteristics of the solution we have laid out above to see that the problem becomes

$$\max_{\omega<\sigma_y} \omega\left(-\tfrac{1}{2}\omega^2 - \lambda(\log\sigma_y^2 - \log\omega^2)\right). \tag{V.15}$$

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 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 without truncation 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 the problem with the prior distribution of Y 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 untruncated problem. And indeed this would give results very close to those of the optimal solution.

From Theorem 1 we know that whenever the support of Y is bounded in this problem with quadratic loss, the support of X is a finite set of points. 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 numerical calculation shows that 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[7], 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 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 always one of the four points of support in the optimal truncated solution with all cases of $|y| > 3$ assigned to the two extreme points among the four. This would mean that the two ± 1.0880 values of x are not in fact conditional expectations of Y given the signal, so the solutions could be improved upon even if restricted to four points of support. Nonetheless, the occurrence of $|y| > 3$ is so rare, and the losses from setting $x = \pm 1.0880$ in these rare cases so small, that overall losses from applying the discrete solution in this way to the untruncated problem are .5118 instead of .5. The information cost from this approximate discrete solution are lower than those for the optimal solution.

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

7. 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.

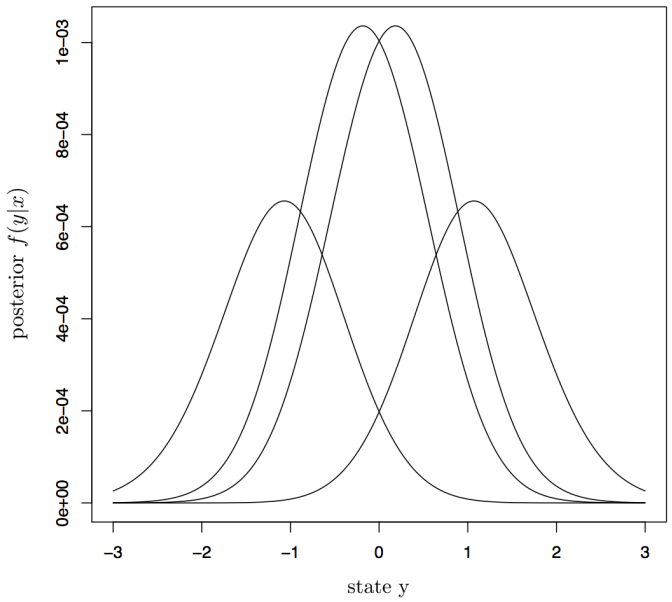


FIGURE 1

Weighted conditional pdf's for Y in tracking problem, $\lambda = .5$

Each pdf shown corresponds to one of the points in the discrete support of X. They are scaled by the probability on the support point, so that at any point on the x axis the sum of their heights is the truncated normal marginal pdf for y.

solutions differ sharply in the marginal distribution of the choice variable X. This kind of result reappears in other settings 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 can again make the optimal distribution for X discrete.

For one more example of coexistence of solutions with very different supports for x, 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 3. *In a rotationally symmetric multivariate tracking problem with quadratic objective function, 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.*

Of course this does not prove that solutions with finite discrete support for x exist, but we have calculated numerical solutions for this problem in which discrete (and hence also continuous on circles) support for x emerges.

VI. CONSIDERATION SETS

Caplin et al. (2018) and Matějka and McKay (2015) find solutions concentrated on subsets of their discrete action spaces. Caplin et all show how to use the full set of first order conditions to solve for this support set. They interpret the concentration of probability on a subset of the action space as explaining how the concept of a "consideration set" can arise endogenously. Our results showing that solutions to problems with open subsets of $\mathbb{R}^n$ as action spaces often concentrate probability on finite sets of actions can be given a similar interpretation.

Our framework may provide some fresh insight into how consideration sets behave. The examples in section V.3 suggest that consideration sets may be sensitive to small variations in the specification of a decision problem. The truncated-normal example shows that with a continuum of actions available, eliminating a subset of the state space with prior probability less than one per cent can convert the consideration set from the entire real line to a finite, and small, set of points.

The examples of rotationally symmetric two-dimensional choice show that optimal behavior that concentrates probability on a finite set of actions may coexist with a pattern of behavior that is also optimal in which actions lie on a continuum of points. In those examples, the decision-maker can be thought of as possibly choosing a random orientation for a finite consideration set, then making a choice from that finite set. But an observer of the behavior would see only the continuously distributed (albeit on a low-dimensional set) behavior.

The coding example in Appendix C.1 shows that when coding is required to approach, without quite attaining, the Shannon bound on capacity, the sharp discreteness of the fully optimal solution may not arise. Instead we might see a kind of "fuzzy discreteness" in choices, with choices frequently very close to a finite set of points, but not concentrating entirely on the finite set in the full solution.

The Caplin et al paper observes that in its framework, it is possible to characterize the set of all priors consistent with a given consideration set. While we do not have an exactly equivalent result in our framework, there is a similar way to generate a large class of priors consistent with a given consideration set. Suppose we have a solution to the standard problem, with a density $p(x)$ over a support B (the consideration set), posteriors $q(y \mid x)$ for each x in B, and the $\theta(y)$ function that appears in the first order conditions (III.5)-(III.8). (Note that $h(y)$ in those equations is $e^{\theta(y)/\lambda}$.) The set of points in the support of the action distribution is determined by (III.10), in which p itself does not appear. Therefore, if we vary the density function $p(x)$, while keeping the support B of X, q, and θ fixed, all the first-order conditions for a solution remain satisfied, except for the adding-up constraint (III.8). But we can treat that equation as determining $g(y)$. As we let p vary over all possible densities on B, we generate a large family of priors g, all consistent with the same consideration set B.

We already have noted one instance of this construction. In the discussion following proposition 2 we noted that when $\theta(y)$ is constant in the one-dimensional tracking problem, the solution implies that $g(y)$ defines the distribution of X plus a random variable independent of X whose pdf is proportional to $e^{U/\lambda}$. In fact, in this case, B is the entire real line and the set of possible priors is that of random variables that are the sum of an arbitrary real-valued random variable and an independent "disturbance" with density proportional to $e^{U/\lambda}$.

VII. PORTFOLIO CHOICE

Microdata on household portfolio rebalancing reveal significant inaction. For example, only 8.6% of households adjust their portfolios every month and 71% every year (Bonaparte and Cooper, 2009). While this has traditionally been attributed to some form of adjustment cost, we show that rational inattention can also explain it. Alvarez et al. (2012) make a similar point, but they use a fixed observation cost of perfect information. Our results go beyond the infrequent adjustment. The optimally discrete distribution of portfolios that we find implies repeated returns to the same set of portfolios. Also our results show a pattern of "normal times" stock-picking, with occasional periods of "risk-on" or "risk-off" behavior in which little information about relative returns of individual assets is collected. This is similar to a finding by Kacperczyk et al. (2016).

The problem we consider is maximization of expected constant absolute risk aversion (CARA) utility by choosing portfolio weights on two risky, and one riskless, security. The returns on the risky

securities have two components: one, Z, that is inherently uncertain; another, Y, that is uncertain only because of limited information-processing capacity. There are no restrictions on short sales, and, because CARA utility is defined on the whole real line, no limits, other than risk aversion, on how leveraged the portfolio can be.

This problem differs from those discussed in section V in that it is not a tracking problem and has a multivariate choice vector.

Formally, the problem is

$$\max E[-\exp(-\alpha\theta'(Y+Z))] - \lambda I(\theta, Y) \tag{VII.16}$$

subject to

$$\theta'\mathbf{1} = 1\,, \tag{VII.17}$$

where θ is a vector of portfolio weights summing to one, and $Y + Z$ is the vector of random yields, of the same dimension as θ. Before information collection, the distribution of Y has pdf $g(y)$, which we assume has bounded support. $Z \sim N(0, \Sigma_z)$, with Y and Z independent, both before and after information collection. $I(\theta, Y)$ is the mutual information between Y and θ. Because there is no restriction on short sales or borrowing, 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, describe a professional high-frequency trader receiving market information over a costly data connection billed in bits per second.

We show in appendix Appendix A 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.

Our numerical solutions are for the case of three available securities, with the first security risk free and returning 1.03, with somewhat higher mean return 1.04 on each of the two risky assets. The prior distribution of the stochastic part of Y is $N(1.04 \cdot \mathbf{1}, .02^2 I)$, truncated to the square region within 3 standard deviations of the mean along each axis. We set the stochastic part of Z to have a $N(0, \sigma_z^2 I)$ distribution, independent of Y.

Results for several settings of λ (information cost), α (risk aversion parameter) and σ_z are displayed in Table 1. With these parameter values, the support of the distribution for the portfolio weights θ is just three or four portfolios.

The first three of these cases, A through C, are arranged in order of increasing leverage. Case A has a higher risk aversion parameter ($\alpha = 2$ instead of $\alpha = 1$) and generates no short selling of either

p	θ_1	θ_2
A) $\alpha = 2, \lambda = .05, \sigma_z = .0173$		
0.067	20.727	20.727
0.467	14.418	1.429
0.467	1.429	14.418
B) $\alpha = 1, \lambda = .1, \sigma_z = .0173$		
0.185	-23.006	-23.006
0.408	55.214	10.431
0.408	10.431	55.214
C) $\alpha = 1, \lambda = .05, \sigma_z = .0173$		
0.024	-77.975	-77.975
0.488	-0.091	45.748
0.488	45.748	-0.091
D) $\alpha = 1, \lambda = .05, \sigma_z = .03$		
0.361	19.328	19.328
0.248	19.890	-5.666
0.248	-5.666	19.890
0.142	-8.737	-8.737

TABLE 1

Portfolio weight distributions, CARA utility.
Note: The first column is the probability on each of the portfolios. The second and third are the portfolio weights on the two risky assets. Since the total value of the portolio is one, the amount invested in the riskless asset is minus the sum of the second and third columns, plus 1.

risky security. However, the sums of weights on the two risky securities are in all three portfolios far above one, implying that there is borrowing at a level 14 to 40 times the investor's net worth. The investor with over 90 per cent probability puts most of the portfolio in one or the other of the two risky assets, and with 7 per cent probability increases leverage even further to invest equally in the two risky assets.

With lower risk aversion, in cases B and C, the investor again most frequently leverages

investments mainly in just one of the two risky assets, but now the other portfolio shorts both risky assets, particularly strongly in case C, with both lower risk aversion and lower information cost. The conditional densities for returns in cases B and C are shown as contour plots in Figures 2 and 3. The lower information cost in case C leads to the investor choosing sharper boundaries between the conditional distributions, thereby limiting the risks of leveraging and leading to greater leveraging.

In Figures 4 and 5 we see that either higher risk aversion (case A) or higher irreducible risk from z (case D) makes leverage less attractive and thus reduces the incentive to collect information so that the conditional densities have little overlap. In case D we see the solution going to four, instead of three portfolios in the support of θ. The low information cost is used there to better separate the centers of the conditional distributions, with less emphasis on reducing overlap because leverage is going to be lower.

In all of these cases, if the investor is thought of as repeatedly applying the solution to this problem over time, he or she would be seen as often leaving their portfolio unchanged, and when changing most of the time switching between being long in one or the other of the two risky rates. Then relatively rarely (though not that rarely in case D) the investor would move to a "risk on" or "risk off" portfolio that weights the risky assets equally (either long or short). Episodes like this, in which investors occasionally buy or sell risky assets as a class, without paying as much attention as usual to differences among securities, are easily explained within this rational inattention framework.

VIII. COMPARISON TO OTHER MODELS OF COSTLY INFORMATION

VIII.1. *People as Shannon Channels*

A Shannon channel is defined by an input "alphabet" and a mapping from that alphabet to a distribution on its output alphabet. The alphabets can be finite lists of discrete characters, the set of all real numbers, functions of discrete or continuous time — anything that can be given a probability distribution can be the input or output alphabet. Shannon showed that once we know the input alphabet and the mapping from it to distributions on the output alphabet, we can calculate the distribution on the input alphabet that maximizes what he defined as the "mutual information" between the input and the output. This maximized mutual information is a real number that he called the "capacity" of the channel. Furthermore, whatever we may wish to send through the channel — music, stock prices, pictures — there is a way to "code" it, i.e. create a mapping from what we wish to send into the input alphabet of the channel, so that information is transmitted at arbitrarily close to the capacity rate.

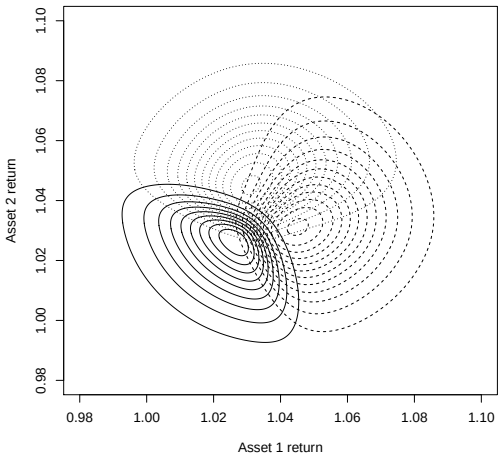


FIGURE 2

Conditional densities of y for case B of Table 1

Note: Each line type displays the contours of the joint pdf of the returns of the two risky assets for one of the three portfolios in the support of the optimal distribution of portfolios shown in the table. The continuous lines, for example, show the distribution of risky yields when the optimal portfolio heavily shorts both risky assets; most of the probability, in this case, is on the region where the risky yields are both below the 1.03 yield on the riskless asset.

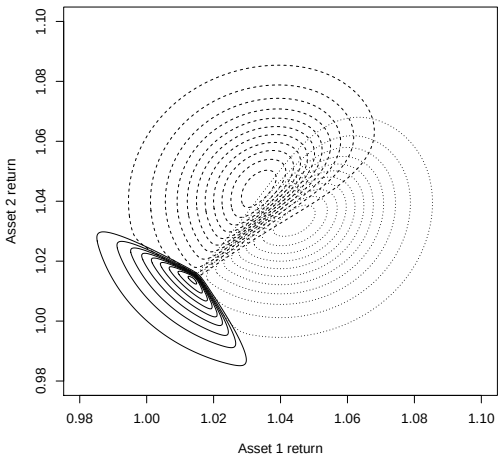


FIGURE 3

Conditional densities of y for case C of Table 1

Note: See note to Figure 2. Here, with lower information cost than case B, the investor still uses three portfolios, but has less overlap in conditional distributions, making leveraging less risky and thereby allowing more leverage.

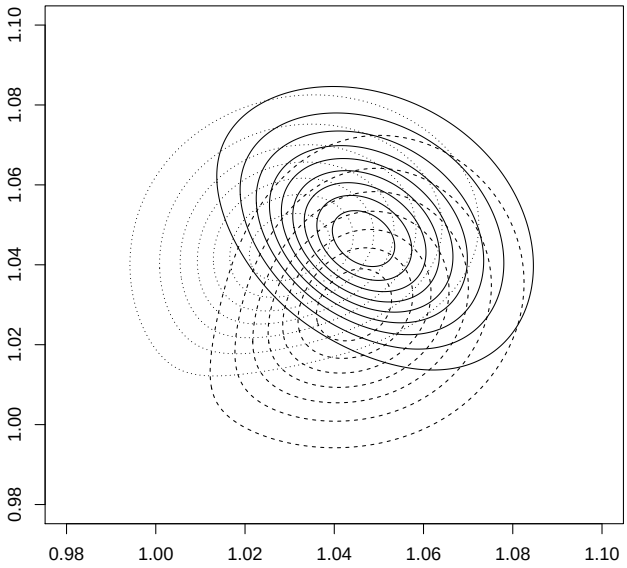


FIGURE 4

Conditional densities of y for case A of Table 1

Note: See note to Figure 2. With higher risk aversion than case B, the investor in this case will use less leverage, making the payoff from sharply separating the conditional distributions lower.

It is the underlying idea of a Shannon channel that gives Shannon mutual information its unique appeal. To the extent that this paper's results apply to observed human behavior, they depend on the idea that this static decision problem recurs over time, or can be grouped with multiple other such problems being solved simultaneously. We are assuming that people, as optimizing agents, manage to "code" the states they are observing optimally for the physical characteristics of their human information-processing apparatus, so that the limiting behavior characterized by Shannon's theorems is a good approximation to their behavior, regardless of the psychological or biological details of their information processing.

VIII.2. *Non-Shannon information processing costs*

One can think of the model in section III as a special case of a more general setup, in which the Shannon mutual information between X and Y is replaced by any reasonable measure of "information cost" and constraints on the joint distribution of X and Y beyond (III.3) are possibly

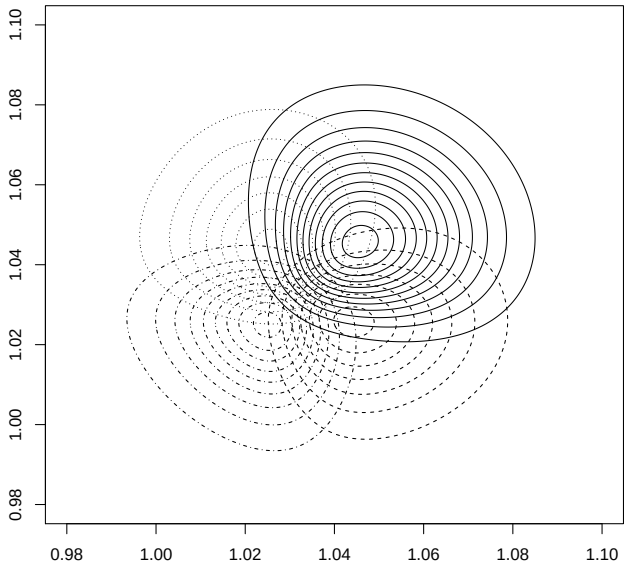


FIGURE 5

Conditional densities of y for case D of Table 1

Note: See note to Figure 2. Irreducible uncertainty from Z is higher than case C, though risk aversion and information cost are the same. Here the investor has four portfolios as points of support, with more overlap in the conditional distributions than in case C.

imposed.[8] For example, a manager might be contemplating investment in a new product and have the option of commissioning a survey to estimate demand. The cost of the survey might be linear in the number N of respondents, with the results (if the number of respondents is large) well approximated as Gaussian with error variance inversely proportional to N.[9] The cost of the survey is certainly a kind of cost of information, and this problem, if costs are linear in N, and expected profits quadratic in the error of the demand estimate, has a well defined solution for any given prior distribution the manager might have for the demand (which for the example's sake we treat as an unknown real number). The cost of information here is not Shannon mutual information, and the results of this paper do not apply. In this problem, unlike the same problem with Shannon mutual information as the information cost, with quadratic loss the manager's action is continuously distributed for continuously distributed priors, even if the prior has bounded support. This is true

8. Caplin and Dean (2015) consider an extended class of measures of information cost that might apply to a decision problem.

9. Pomatto et al. (2018) develops a notion of information cost that takes it as axiomatic that in repeated i.i.d. sampling information cost is linear in the size of the sample.

even if we make the cost proportional to $\log(N)$, which makes it coincide with Shannon mutual information in the special case where the prior is Gaussian (and hence has unbounded support).

While this example is a reasonable instance of decision-making under uncertainty, with an information cost and endogenous choice of posterior distribution, it cannot be interpreted as reflecting the information-processing constraints of the manager herself. The example is one where the costs of information are external to the decision maker, and there is no good argument for giving Shannon mutual information special emphasis in such a framework. Shannon mutual information is a useful measure of costs of information *processing*, turning information into action, when there is no other cost to obtaining the information.

Shannon mutual information depends only on the joint distribution of the unknown state and the action or observation; it does not depend on how the states are "labeled". For example, we might expect that people asked to identify white and black disks flashed on a screen at a certain rate could do so with negligible error, while if the disks are instead two very similar shades of gray they would make frequent errors. If the white and black, or light and dark gray, disks are presented as an i.i.d. sequence with .5 probability on each, the mapping from signal (white or black, light gray or dark gray) to action (the person's identification of what she has seen) is transmitting one bit of information per transmission if there are no errors, but less, maybe much less, if there are a lot of errors. Shannon mutual information between signal and action is greater when there is no error, which suggests that the "information cost" of tracking the light-gray/dark-gray sequence is higher than that of tracking the white-black sequence. This is sometimes taken as a criticism of the Shannon measure, since it implies that capacity, or the cost of information, could depend on the labeling. That is, there can be "hard to distinguish" and "easy to distinguish" states, and it may be claimed that a good measure of information costs must, unlike Shannon mutual information, recognize this distinction.

But rational inattention is based on the idea that rational agents, faced with a repeated problem of using information to guide decisions, will optimally *code* the states they are responding to, recognizing their own strengths and weaknesses as Shannon channels. The white-black-gray example describes a simple channel. A white input produces a white output without error; a black input produces a black output without error, a light gray or dark gray input produces either a light gray or dark gray output, but never white or black. If the person actually needs to react to a sequence of light and dark gray disks, with none white or black, she can simply map light gray to white and dark gray to black (or vice versa), and transmission will be without error. If instead black, white, light gray and dark gray transmissions occur in the input stream with equal probability, the person could for example map black to black, white to white, and light or dark gray to a gray signal that means the next transmission will be white if the input was light gray, black if the input was dark gray. Even if

the person has no ability at all to distinguish light and dark gray, so that seeing a gray disk of either shade leaves her with equal probabilities on light or dark gray for the signal, this coding scheme will transmit the white-black-light-gray-dark-gray sequence without error. Of course this is at the cost of taking, on average, 1.5 transmissions for each disk observation in the sequence of states. The Shannon capacity of this channel, assuming white and black inputs transmit without error and any gray input resolves no uncertainty about whether it is light or dark, is $\log_2(3) = 1.58$ bits per period. A channel in which all four states transmitted without error would have 2 bits per period capacity. The coding scheme we have described here, using a gray observation as a prefix, announcing that the state is gray and the next signal will be white or black and resolve the uncertainty about whether the gray is light or dark, achieves a transmission rate of $3/2$ bits per period. Shannon's coding theorem then tells us that we could somewhat improve on this simple coding scheme, but could not increase the transmission rate by more than $1.58 - 3/2 = .08$ bit per period. We discuss coding in more detail in the online appendix.

Of course there are decision problems in which options for coding signals are limited, or where the problem is isolated, not repeating, so that investing in improved coding is not necessarily optimal. In such problems Shannon capacity and mutual information may not be the best approach to modeling information processing costs, even those internal to the decision maker. But as also should be clear from this example, there will not be an obvious, unique alternative to the Shannon measures. Modeling information processing costs will depend on details of the specific problem in these cases.

There are a number of ways to frame sets of axioms that imply the Shannon measure, as laid out in Csiszár (2008). Suppose $H(Y)$ is a measure of uncertainty in the random variable Y and that it depends only on the distribution of Y, not on labels. In the case of a finite state space, this means it depends on the set $\{p_1, \ldots, p_n\}$ of probabilities on the points in the state space and is invariant to permutations of the p_i's in the set. Let $H(Y \mid X)$ be the expectation, over the distribution of X, of H applied to the conditional distribution of Y given X. If H has the property that $H(Y, X) = H(X) + H(Y \mid X)$, then, subject to some regularity conditions, it is Shannon entropy. This is the property, called Strong Additivity by Csiszar, that the information in observing X, plus the information in observing Y after updating our distribution for Y based on observing X, is the same as the information in seeing X and Y simultaneously. When modeling a flow of information over time, or across many simultaneous observations, this additivity property is essential to the measure's usefulness, as it is required to support Shannon's coding theorem.

Nonetheless, one can consider other plausible measures of information flow in a static decision problem. One might wonder whether any such measure, based on the joint distribution of x and

y and invariant to monotone transformations of x and y, might lead to results like our main ones in this paper. In section Appendix C.4 of the online appendix we consider what emerges if we measure information flow by the expected total variation distance between the prior density g and the posterior densities $q(y \mid x)$. We find that in this case, at least, neither our Lemma 1 nor our Theorem 1 result holds. We have carried out some numerical calculations with other measures of information flow based on the joint distribution, and have in some cases found apparent discrete support for x. But the result we find with Shannon mutual infomation, that bounded support for y and analytic U implies discrete support for x, regardless of λ, is clearly not generic with other static measures of quantity of information.

IX. 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. Lemma 2 and Proposition 1 provide general tools for assessing it, and we apply these tools in our examples. Proposition 1 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 example 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.

REFERENCES

Alvarez, F., L. Guiso, and F. Lippi (2012). Durable consumption and asset management with transaction and observation costs. *American Economic Review 102*(5), 2272–2300.

Alvarez, F. E., F. Lippi, and L. Paciello (2011). Optimal price setting with observation and menu costs. *The Quarterly Journal of Economics 126*(4), 1909–1960.

Berger, T. (1971). *Rate Distortion Theory: Mathematical Basis for Data Compression*. Prentice Hall.

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.

Caplin, A., M. Dean, and J. Leahy (2018). Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, rdy037. Accepted.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*. Hoboken, NJ: Wiley.

Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy 10*(2008), 261–73.

Eichenbaum, M., N. Jaimovich, and S. Rebelo (2011). Reference Prices, Costs, and Nominal Rigidities. *American Economic Review 101*(1), 234–62.

Fix, S. (1977, September). *Rate Distortion Functions for Continuous Alphabet Memoryless Sources*. Ph. D. thesis, University of Michigan.

Fix, S. (1978, Oct). Rate distortion functions for squared error distortion measures. *Proc. 16th Annu. Allerton Conf. Commun., Contr., Comput.*.

Hellwig, C. and L. Veldkamp (2009). Knowing what others know: Coordination motives in information acquisition. *The Review of Economic Studies 76*(1), 223–251.

Kacperczyk, M., S. Van Nieuwerburgh, and L. Veldkamp (2016). A rational theory of mutual funds' attention allocation. *Econometrica 84*(2), 571–626.

Krantz, S. G. (1992). *Function Theory of Several Complex Variables* (2nd ed.). AMS Chelsea Publishing.

Luo, Y. (2008). Consumption dynamics under information processing constraints. *Review of Economic Dynamics 11*(2).

Mackowiak, B. and M. Wiederholt (2009). Optimal sticky prices under rational inattention. *American Economic Review 99*(3), 769–803.

Maćkowiak, B. and M. Wiederholt (2015). Business cycle dynamics under rational inattention. *The Review of Economic Studies*, rdv027.

Mankiw, N. G. and R. Reis (2002). Sticky information versus sticky prices: a proposal to replace the new keynesian phillips curve. *The Quarterly Journal of Economics 117*(4), 1295–1328.

Matějka, F. (2015). Rigid pricing and rationally inattentive consumer. *Journal of Economic Theory 158*, 656–678.

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

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

Mondria, J. (2010). Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory 145*(5).

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

Myatt, D. P. and C. Wallace (2011). Endogenous information acquisition in coordination games. *The Review of Economic Studies 79*(1), 340–374.

Pomatto, L., P. Strack, and O. Tamuz (2018, December). The cost of information. Technical report, California Institute of Technology.

Rose, K. (1994, November). A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory 40*(6), 1939–1952.

Sims, C. A. (2003a). Implications of rational inattention. *Journal of Monetary Economics 50*(3).

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

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

Sims, C. A. (2010). Rational inattention and monetary economics. In *Handbook of Monetary Economics*. Elsevier.

Stevens, L. (2015, November). Coarse pricing policies. Research Department Staff Report 520, Federal Reserve Bank of Minneapolis.

Tutino, A. (2009). *The Rigidity of Choice: Lifetime Savings under Information-Processing Constraints*. Ph. D. thesis, Princeton University.

Tutino, A. (2013). Rationally inattentive consumption choices. *Review of Economic Dynamics 16*(3), 421–439.

Van Nieuwerburgh, S. and L. Veldkamp (2010). Information acquisition and under-diversification. *Review of Economic Studies 77*(2), 779–805.

Veldkamp, L. L. (2011). *Information choice in macroeconomics and finance*. Princeton University Press.

Woodford, M. (2009). Information-constrained state-dependent pricing. *Journal of Monetary Economics 56*(Supplement 1).

Yang, M. (2015). Coordination with flexible information acquisition. *Journal of Economic Theory 158*, 721–738.

APPENDIX A. PROOFS

Appendix A.1. *Proof of Proposition 1*

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 $v(x, y)$ 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 $\int v(x, y)\mu_y(dy)dy$ is analytic on S^*, and an analytic function on S^* is real analytic when restricted to S.

Appendix A.2. *Proof of Corollary 1.1*

Proof. In this case the role of $v(x, y)$ is played by $\gamma(y - x)$, which is clearly analytic in x over the same domain (the whole real line) for every y. A real analytic function is always extensible to a complex analytic function over some domain, and since here the real analytic function is the same except for a location shift for every y, the domain is the same except for a

location shift for all y. Since the radius of convergence of γ is bounded below by some $\delta > 0$, all the location-shifted versions of γ are extensible to the open and connected

$$\{z \in \mathbb{C} \mid \mathrm{Im}(z) < \delta\} \ ,$$

and the fact that p and γ are both integrable means their convolution is integrable, so the result follows from the proposition.

Appendix A.3. *Proof of Theorem 1*

Proof. g, as a probability density, must be integrable. $\int p(x)e^{V(x-y)/\lambda}\mu_x(dx)$ is positive and by Corollary 1.1 analytic in y and therefore bounded below away from zero on the bounded support of Y. From III.8, therefore, we can conclude that h is integrable. Therefore $C(x) = \left(e^{V/\lambda} * h\right)(x)$ is analytic, again by Corollary 1.1, because h, being integrable and of bounded support, is a scaled probability density. 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. Thus the support of the distribution of x contains no open sets, and if x and y are one-dimensional, the support consists of a finite number of points.

Appendix A.4. *The portfolio problem of section VII*

The problem is

$$\begin{aligned} \max_{q,p,\mu\theta}\Big(\int -e^{-\alpha\theta'(Y+Z)}q(y \mid \theta)\phi(z,\Sigma)dy\,\mu_\theta(d\theta)dz \\ + \int \log(g(y \mid \theta)q(y \mid \theta)p(\theta)\mu_\theta\theta(d\theta) - \int \log(g(y))g(y)dy\Big) \end{aligned} \tag{A18}$$

$$\text{subject to: } \forall(x)\int q(y \mid \theta)dy = 1) \tag{A19}$$

$$\forall(y)\int p(\theta)q(y \mid \theta)\mu_\theta(d\theta) = g(y) \tag{A20}$$

$$\forall(y,\theta)f(y,\theta) \geq 0\,, \qquad \theta'\mathbf{1} = 1\,. \tag{A21}$$

Here $\phi(z,\Sigma)$ is the density of a normal distribution with covariance matrix Σ and mean zero. θ is the vector giving proportions of total wealth 1 invested in each security. Total yields of the securities are $Y + Z$, with uncertainty about Y being reducible by paying an information cost, while Z is irreducible uncertainty.

Proposition 4. *Suppose*

i the prior density g of Y has mean vector μ_y;
ii $\mu_y\mu_y' + \Sigma$ is positive definite; and
iii g has bounded support.

Then the solution to the portfolio choice problem above gives X a distribution whose support contains no open sets.

Proof. To get this into the notation of our general decision problem (which had no z), we first integrate with respect to z, which makes the objective function

$$E\left[-e^{-\alpha\theta'Y+\alpha^2\frac{1}{2}\theta'\Sigma\theta}\right]. \tag{A22}$$

The problem is then in our standard form, so we can apply the FOC's (III.7) and (III.8). The FOC's imply that

$$h(y) = \frac{g(y)}{\int p(\theta)\exp\left(-e^{-(\alpha/\lambda)\theta'y+(\alpha^2/\lambda)\frac{1}{2}\theta'\Sigma\theta}\right)\mu_\theta(d\theta)}\,. \tag{A23}$$

The denominator is an integral whose integrand is everywhere positive, and thus positive for every y and continuous in y. Since y has bounded support, the denominator is bounded away from zero on the support of y, and the ratio is therefore integrable. $h(y)$, being integrable, is proportional to a probability density. Thus

$$C(\theta) = \int h(y) \exp\left(-e^{-(\alpha/\lambda)\theta' y + (\alpha^2/\lambda)\frac{1}{2}\theta'\Sigma\theta}\right) dy \tag{A24}$$

fits the assumptions of our Proposition 1. The integrand is integrable in y because of the bounded support. If Σ is positive definite, so there is no riskless asset, the integrand is integrable in θ for every fixed y because the quadratic term in θ dominates, making the integrand drop rapidly to zero as $\|\theta\|$ increases. It is for each y analytic in θ on the whole of $\mathbb{C}^n$ plane (where n is the dimension of θ), so the regularity conditions in the proposition about the domain of analyticity are automatically satisfied. Thus $C(\theta)$ is analytic in θ. It cannot be constant, because it is positive, yet as $\|\theta\| \to \infty$, $C(\theta)$ clearly goes to zero.

If one of the assets is riskless, we need to interpret y and z as determining the yields on only the risky assets, remove the $\theta'\mathbf{1} = 1$ constraint, and add the term $-\alpha(1 - \theta'\mathbf{1})\rho$ to the exponent in the objective function, where ρ is the return on the riskless asset. This makes the algebra a little messier, but the argument remains essentially the same.

APPENDIX B. 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 (III.2)-(III.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 (III.7) and (III.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 with the on-line appendix.

APPENDIX C. ONLINE APPENDIX: MORE EXAMPLES

Appendix C.1. *Example of coding*

It may help in understanding the nature and practical importance of Shannon's result to consider two simple, but very different channels with the same capacity. One is a "telegraph key": Its input alphabet is the set $\{0,1\}$, its output alphabet is the same, and its output is simply equal to its input. This channel has capacity one "bit" per time period or per transmission.

Another channel with the same capacity is one whose input alphabet is real numbers, with the restriction that the variance of the distribution from which the numbers are drawn must be less than or equal to one. (In electronic communication, such a restriction arises from limits on the power of the input sequence.) For this channel the output from an input value y is an output value $z = y + \varepsilon$, where ε is a normally distributed "noise" of variance $1/3$ that is independent of y. The telegraph key channel attains its capacity rate of transmission when its input makes the probability 1 or 0 both the same, i.e. .5 each. The normal-error channel attains its capacity rate of transmission when the input has a $N(0,1)$ distribution.

It might seem that if we have a message to send that consists of zeros and ones, e.g. a digital computer file or a Morse code message, the telegraph-key channel would be useful, as it could send the message without error. The normal-errors channel might seem far less useful, as our discretely distributed input could not match the channel's optimal distribution of inputs, and its output would always be contaminated by the channel's inherent noise. Similarly, if we wished to send a message drawn from an i.i.d. sequence of $N(0,1)$ random variables and were satisfied with a standard deviation of $1/\sqrt{(3)}$ on the output signal's error, the normal-error channel would work easily. The telegraph key, which might seem to have no way to trade off accuracy for speed of transmission, would apparently be much worse. Shannon's coding theorem, and its generalization to what is called rate-distortion theory, shows that these two channels can each approximate the other's behavior arbitrarily well.[10] More generally, given any sequence of random variables (not necessarily defined on the channel's alphabet) as inputs and any objective function or measure of accuracy that is a function of the joint distribution of input and output, the accuracy we can achieve is defined by the channel capacity. *Coding* maps the message to be sent into sequences or sets of values of the channel's input alphabet. Transmission at a rate arbitrarily close to the capacity rate is achieved by optimally coding the message to be sent. None of the details of the channel's alphabets, internal noise, or mapping from input alphabet to output alphabet matter to the accuracy that can be achieved.

This generality is appealing as a basis for modeling the effects of limits on information processing by human beings. People process all kinds of information, all the time. Understanding how the five senses, brain structure, and nervous system map external signals to effects on human behavior, i.e. the details of the human information channel, is prohibitively complicated. But if we postulate that individuals have finite Shannon capacity, and that they use this capacity optimally, we can derive some conclusions about the effects of their finite capacity without knowing the details of their channel structure.

There are many other notions of a "cost of information" that have a legitimate claim to that label. In the paper's section VIII and appendix section Appendix C.4 we compare Shannon's measure to some others and discuss the degree to which our discreteness result generalizes to other measures of information cost. Shannon's measure is unique, though, in its connection to the ideas of coding and channel capacity, and thereby to the separation of information processing costs from biological and psychological details. We think this justfies our focus on this measure of information cost.

In order to understand the coding theorem and its implications, we can look further at our telegraph key and noisy Gaussian channels, aiming to make concrete the notion of coding to achieve transmission at close to the capacity rate. Consider the problem of matching x to y, where y is distributed as $N(0,1)$ and our losses are $E[\frac{1}{2}(y-x)^2]$. We can transmit information about y only via a channel with capacity one bit per time period — in one case, the telegraph-key channel, in the other the Gaussian channel we described above. If we consider one channel and one period in isolation, the best we can do with the Gaussian channel is to transmit y, observing it as the channel output $y + \varepsilon$. One bit of information is transferred, and by setting our decision variable as $x = .75 \cdot (y + \varepsilon)$ we achieve an expected loss of .125. It can be shown that this level of expected loss is the best that can be achieved with any 1-bit channel.

But what if we have only the telegraph key? If we consider this channel and a single transmission in isolation, the best we can do is code $y > 0$ as 1 and $y < 0$ as 0. We would then set $x = E[y \mid y > 0]$ if we observe the output 1 and $x = E[y \mid y < 0]$

10. For more discussion of Shannon's theory, and of these two example channels, see Cover and Thomas (2006).

if we observe the output zero. This produces expected loss of $.5 \cdot (1 - 2/\pi) = .182$, almost 50% greater than the optimum. But Shannon's theory tells us that any 1-bit-per-period channel can get arbitrarily close to the ideal value of the loss function attained by the Gaussian channel. How can this be achieved?

The ideal transmission rate can be approximated with the discrete telegraph-key channel by grouping transmissions, either across time or, if multiple such problems are being solved at once, across multiple similar channels. The naive solution that considers the channel and period in isolation is the crudest version of *quantization*, that is, a mapping of the continuously distributed y into the discrete, finite set $\{0, 1\}$. If we had, say, six channels like this running simultaneously, mapping six values of y into six zeros and ones, we could just assign each element of the six-dimensional y vector to a channel and reproduce the single-channel, single-period loss of .182 for each channel. But we can do better than that. With six channels, there are 64 possible sequences of six zeros and ones. The best solution with six channels treated jointly will quantize the 6-dimensional y vectors, but it will do so more efficiently than the naive solution. It will partition $\mathbb{R}^6$ into 64 sets and map each one into a distinct sequence of six zeros and ones. It is possible to choose these sets so that they deliver smaller variance of $y - x$ than the naive solution. Indeed as the number of periods or channels considered jointly increases, the performance of the grouped telegraph-key channels comes arbitrarily close to that of the Gaussian channel. In order to do so, the joint distribution of y and x for each individual channel approaches the joint Gaussian distribution that characterizes the optimal, Gaussian channel.

The main result of this paper[11] implies that if instead of a $N(0, 1)$ distribution for y we had a truncated normal distribution, say truncated at ± 3, the optimal solution would not be exactly attained with the Gaussian channel and would indeed make the marginal distribution of the optimal x concentrate on a finite set of points.

Of course the telegraph-key channel, even if grouped into 6 joint transmissions, would be transmitting only 64 distinct possible messages each period, and thus the x_i values for each individual channel i would have finite support on 64 points. There would be finite suppport no matter how many channels or periods we group, though the number of points of support would grow exponentially with the number of channels or periods grouped. Because the solution would always be quantized (i.e. involve a deterministic mapping from y to sequences of zeros and ones, and hence to x values), it would be expected to distribute x values over 2^n possible points of support, where n is the number of channels or periods grouped. Our result about discreteness of the x distribution in the optimal solution implies that, as n increases the distribution of x_i values for a given individual channel i collapses on a finite set of points, with that limiting set of support points not increasing with n. It does not imply that at any given value of n the limiting finite support set is the full support of x in the approximate solution.

Figure 6 shows the distribution of one element of the six-dimensional x vector when a set of y values drawn from a $N(0, 1)$ distribution truncated at ± 2 is quantized with 6 bit discrete code (i.e. with a partition of $\mathbb{R}^6$ into 64 sets). The tendency toward a discrete distribution with two points of support is apparent, but the distribution is not completely concentrated on those two points. This coding succeeds in getting the loss function value to a value almost as low as what would be achieved with a continuous channel by naively using the linear mapping from the noisy observed data to x that would be optimal with untruncated data (0.118 vs 0.115). The naive variable-by-variable quantization would produce expected loss 0.131. This quantization was calculated by applying the k-means algorithm (using R's `cclust` function) to 3.3 million draws from a truncated $N(0, 1)$ distribution, arranged as a 550,000 by 6 array. Clearer discreteness would show up with grouping over more dimensions.

Note that when we group n channels or periods what is quantized is the mapping from 2^n subsets of $\mathbb{R}^n$ to 2^n sequences of n zeros and ones. Therefore the conditional distribution of the n-dimensional y vector given the n-dimensional x vector is concentrated entirely on one set in the partition of $\mathbb{R}^n$, with zero conditional density outside that set. But the conditional

11. For this univariate, quadratic objective function case, the result was known in the information theory literature, see Fix (1977) andRose (1994).

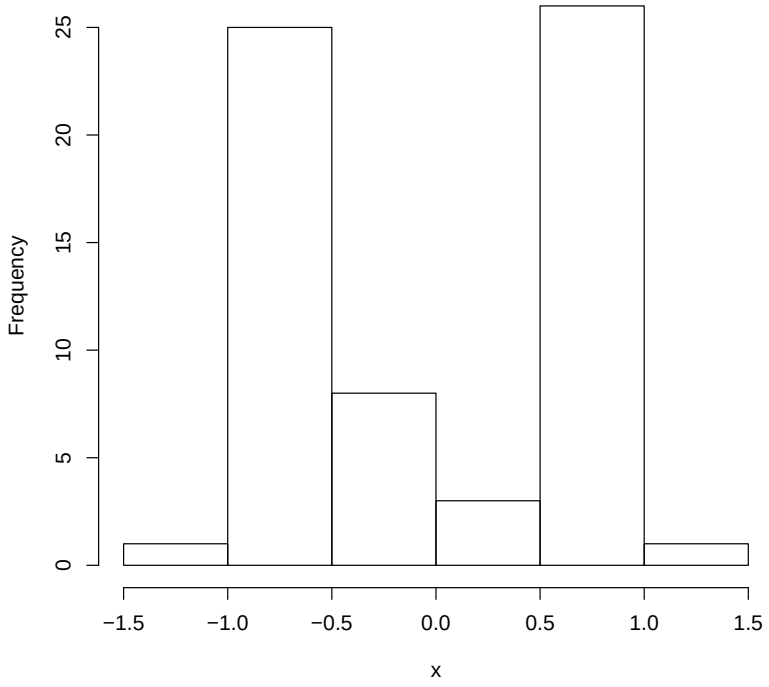


FIGURE 6

Histogram of one element of x, 6-bit code, N(0,1) truncated at 2

density of an individual element y_i given the corresponding element x_i of x is not quantized. That is, the support of y_i conditional on a given value of x_i generally overlaps the support of y_i given other possible values of x_i. In the ideal limiting joint distribution, so long as the cost of information is positive and losses are finite, the support of the $y_i \mid x_i$ distribution is the same as the support of the marginal distribution of y_i itself, for all values of x_i.[12]

Appendix C.2. *A risk-averse monopolist*

This example has interesting economic content and is outside the range of cases considered in the engineering rate-distortion literature.

12. See Lemma 1.

We could also have considered as an example of coding the reverse problem: Desiring to send a sequence of zeros and ones without error, as is possible with the telegraph-key channel, how do we do so, at the same rate, with the Gaussian channel? Shannon's theory tells us that by grouping across time or channels, the Gaussian channel can transmit at the same rate as the telegraph key, and with arbitrarily low error. Coding schemes that can accomplish this are more subtle than the n-dimensional quantization of our simple example. Our example is meant only to give initial insight into how the Shannon theory allows abstraction from the details of channel specification. Readers who want to pursue a deeper understanding of coding should start with Cover and Thomas (2006).

Suppose a risk-averse monopolist faces a demand curve $Q = q(X)$ and a constant returns production technology with unit cost W, where W 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 w to a choice of X, the problem becomes

$$\max E\Big[\log\big((X-W)q(X)\big)\Big] - \lambda I(X,W)\,, \tag{C25}$$

where the maximization is over the joint distribution of X and W. We assume W 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 III. To proceed further we need to assume an explicit form for the demand curve $q()$. Consider $q(x) = x^{-\theta}$ for $x > 0$.[13]

In section 3.2.1 below 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 W 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 W, but distributions for X whose support contains intervals of the form $(0,T)$ can emerge only with a restricted class of cost 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 W 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 W 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 $\lambda \to \infty$ eventually the solution does not have full support on any interval of the form $(0,T)$, for any $0 < T \leq \infty$.

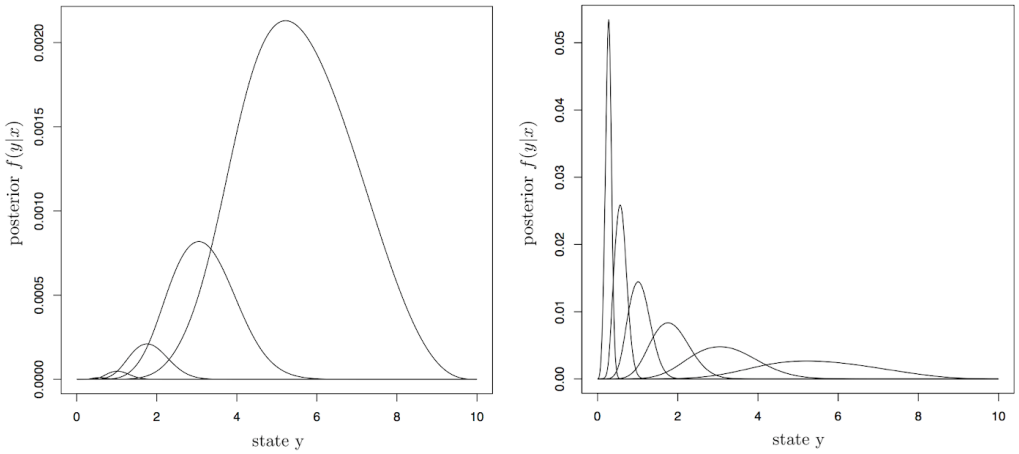
iii When there is an a priori known upper bound $\bar{w}$ on cost W, any solution in which X is continuously distributed over some interval of the form $(0,\bar{x})$ must have $\bar{x} = \bar{w}$. This might seem counterintuitive, since with $\lambda = 0$ we know that the solution is $x \equiv \theta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

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

We have not been able to show that when X is not continuously distributed, it must necessarily have countable discrete support, but in Proposition 6 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 $(w^*, \theta w^*/(\theta-1))$. With such a tight bound, the support of the X distribution lies outside the support of the W distribution, and it turns out this implies discreteness.

We solve this problem numerically with the distribution of W a Beta(4,4) distribution scaled to cover the interval (0,10). The prior distribution for costs is thus symmetric around $w = 5$. It is in the family of distributions shown in section 3.2.1 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.

13. Matějka (2016) 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 (2006).



[h]

FIGURE 7

Conditional pdf's for Y in the risk-averse monopolist problem.

Note: $\lambda = .05$, $\theta = 1.5$. On the left the pdf's are weighted by probabilities of the corresponding actions (so they sum to the prior); on the right, unweighted, so each integrates to one.

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 W given these 6 values for the decision variables are shown in Figure 7. 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 w is small and in each case certainty that W 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 W, including an upper bound on it, is therefore extremely valuable when W is in fact very low. The information-constrained monopolist simply sets a high price, enough to ensure a profit even if W is at its maximum possible level, 80% of the time, with his beliefs about W in this case spread broadly over the interval 3 to 10. But on the rare occasions when W 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 $\theta W/(\theta - 1)$, 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.

3.2.1. Analytic results. Adapting the first order conditions for the general decision problem discussed in section above to this case gives us

$$p(c \mid x) = ((x-c)q(x))^{1/\lambda}\, h(c) \tag{C26}$$

$$\therefore \int_0^x ((x-c)q(x))^{1/\lambda}\, h(c)\, dc = 1 \quad \text{for any } x \text{ such that } p(x) > 0 \tag{C27}$$

$$\int p(x)\, ((x-c)q(x))^{1/\lambda}\, h(c)\, d\mu_x(x) = g(c) \quad \text{all } c\,, \tag{C28}$$

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 (C27) for all positive prices x. If we guess $h(c) = Kc^\gamma$, we find

$$\int_0^x \left((x-c)x^{-\theta}\right)^{1/\lambda} Kc^\gamma\, dc = Kx^{\frac{1-\theta}{\lambda}+\gamma+1} B\left(\gamma+1, \frac{1}{\lambda}+1\right)\,, \tag{C29}$$

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 5 that this $h(c)$ function is unique; no other $h(c)$ can satisfy (C29) over any interval of the form $(0, \bar{c})$ with $0 < \bar{c} \leq \infty$.

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

$$g(c) = \frac{\int_c^\infty \pi(x)\left((x-c)x^{-\theta}c^{\theta-1-\lambda}\right)^{1/\lambda}\, \mu_x(dx)}{B\left(\frac{\theta-1}{\lambda}, \frac{1}{\lambda}+1\right)}, \tag{C30}$$

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 5. *In the risk-averse monopolist problem of section* Appendix C.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* $0 < S \leq \infty$*, the* $h()$ *function that solves* (C27) *for all* x *in* $(0,S)$ *is* $h(c) = Kc^\gamma$*, with* $\gamma = (\theta-1)/\lambda - 1$ *and* $K = 1/\text{Beta}(\gamma+1, 1+\lambda^{-1})$*. 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 (C27) to derive the conclusion

$$x^{-\theta/\lambda} \int_0^x (x-c)^{1/\lambda} \tilde{h}(c) \mu_c(dc) = 0 \quad \text{all } x \in (0,S). \tag{C31}$$

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

$$\begin{aligned} 0 = \int_0^y (y-x)^\alpha \int_0^x (x-c)^{1/\lambda} \tilde{h}(c) \mu_c(dc) dx \\ &= \int_0^y \int_c^y (y-x)^\alpha (x-c)^{1/\lambda} dx \tilde{h}(c) \mu_c(dc) \\ &\int_0^y \int_0^{y-c} (y-c-x)^\alpha x^{1/\lambda} dx \tilde{h}(c) \mu_c(dc) \\ &= \int_0^y (y-c)^m \tilde{h}(c) \mu_c(dc) \int_0^1 (1-x)^\alpha x^{1/\lambda} dx \end{aligned}$$

$$= \text{Beta}(\alpha+1, 1/\lambda+1) \int_0^y (y-c)^m \tilde{h}(c)\mu_c(dc)\,, \quad \text{(C32)}$$

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

$$0 = \int_0^y \tilde{h}(c)\mu_c(dc) \quad \text{(C33)}$$

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

However, (C30) 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 (C30) 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 (C30) 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 (C30) 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 (C30) 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 6. *In the risk-averse monopolist problem of section* Appendix C.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 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

$$h(c) = \frac{g(c)}{\int p(x) x^{-\theta/\lambda}(x-c)^{1/\lambda}\mu_x(dx)}\,. \quad \text{(C34)}$$

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

$$\int_0^x x^{-\theta/\lambda} \cdot (x-c)^{1/\lambda} h(c)dc = \int_{c_1}^{c_2} x^{-\theta/\lambda} \cdot (x-c)^{1/\lambda} h(c)dc \quad \text{(C35)}$$

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.

Appendix C.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.

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

$$-\tfrac{1}{2}E[\|X-Y\|^2]-\lambda I(X,Y)\,, \tag{C36}$$

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, 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

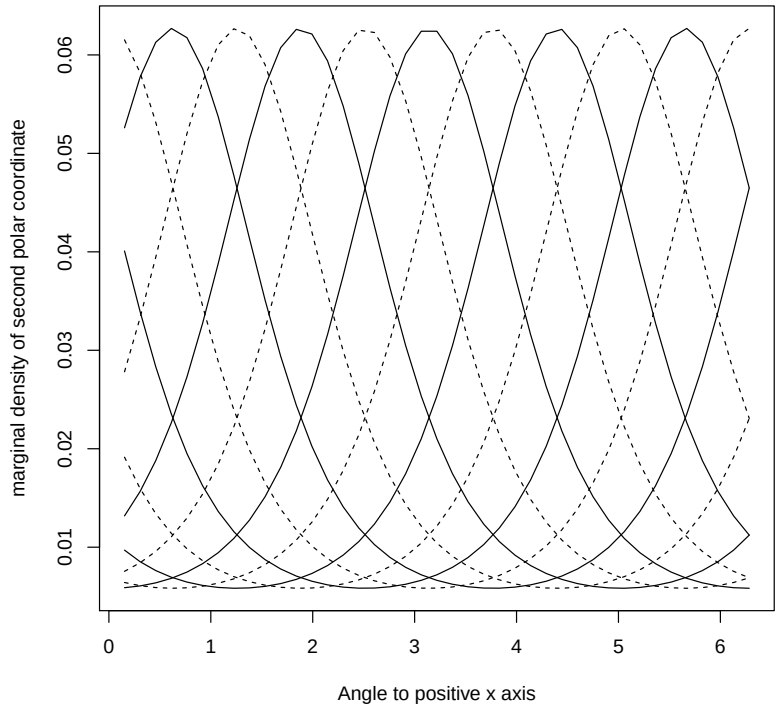
$$-\tfrac{1}{2}E\left[\sum_{i=1}^{n}\omega_i^2-\lambda(\log(\sigma_i^2/\omega_i^2)\right]. \tag{C37}$$

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 $\|Y\| \leq 3$, the support of X is likely to be one or more circles in $\mathbb{R}^2$. We show in proposition 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



[h]

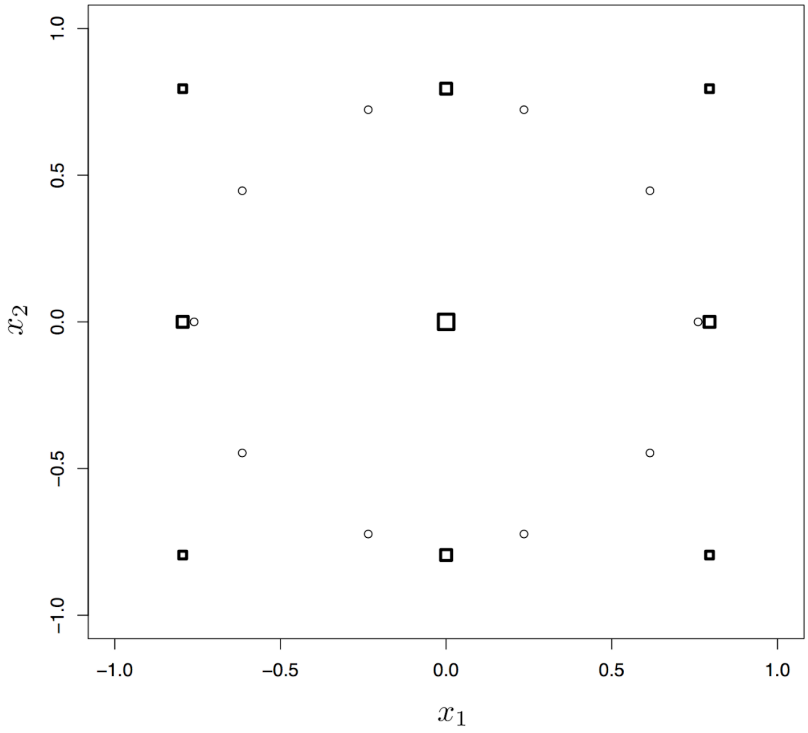
FIGURE 8

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

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 8 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.[14] 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 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 9 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

14. Details about the numerical methods are in appendix Appendix B.



[h]

FIGURE 9

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 $N(0, I)$.

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, 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.

Appendix C.4. L_1 *information costs*

This information cost formulation, as was noted in the text of the paper, sets the cost as

$$\int p(x) \, |(q(y \mid x) - g(y)| \, dy \, dx \, . \tag{C38}$$

We do not have an economic example where this has a clear grounding in psychology or optimality, but this is a metric on probability distributions that is widely applied in mathematics. It, like the Shannon measure, has the property that monotone transformations of x and y leave it unchanged — an important property when we are modeling optimizing agents. Like essentially every information cost measure other than the Shannon measure, it does not have what Csiszár (2008) calls the

strong additivity property — the information in observing a random variable X_1, plus the information in observing X_2 after updating the prior based on the observation of X_1, is not the same as the information in observing the pair X_1, X_2.

On the other hand, it delivers results that are quite different in some respects from what emerges from the Shannon measure. For the L_1 measure there is no result like Lemma 1; optimal $q(y \mid x)$ can, and generally does, have $q(y \mid x) = 0$ for intervals of y values for a given x. By the L_1 measure, the information in observing a continuously distributed random variable without error is finite, in fact always equal to 2.0, whereas with the Shannon measure the information in such an observation is infinite. Thus while with the Shannon measure any non-zero level of information costs will imply non-degenerate $q(y \mid x)$, with the L_1 measure a low enough information cost will make it optimal to observe y without error. Even when information costs are higher, solutions with the L_1 cost often imply a mixed continuous-discrete distribution for $y \mid x$. And finally, our main result, finite support for optimal x distributions when U is analytic on R and y has bounded support, does not hold with L_1 information cost.

The first order condition with respect to $q(y \mid x)$, in the L_1 cost version of the model is

$$p(x)U(x,y) = p(x)\lambda \operatorname{sign}\big(q(y \mid x) - g(y)\big) + p(x)\theta(y) + \phi(x)\,. \tag{C39}$$

This has to hold only where $q(y \mid x) > 0$. Generally, and certainly in the case of quadratic U, $U(x,y)$ as a function of y cannot match the variation in the single $\theta(y)$ function on the right hand side of this FOC for multiple x values. Thus q in the solution is likely to take the form of a function that is zero for some y's, equal to $g(y)$ for other y's, and neither zero nor $g(y)$ on a set of y values that is unique to the particular value of x.

The nature of the result is perhaps easiest to see if $g(y)$ has the unit circle as support and is uniform on it, and U is minus the squared distance between x and y along the circle. Then the L_1 solution gives $q(y \mid x)$ the form of a discrete lump of probability a on x, together with a uniform density on an interval $(x-b, x+b)$ centered on x. a and b must satisfy

$$\frac{b}{\pi} = 1 - a$$

in order that the height of the density on $(-b, b)$ match $1/(2\pi)$, the uniform density on the unit circle. As information costs decrease, b shrinks to zero. As they increase, b increases up to the point at which it is better to collect no information and set $q(y \mid x) = g(y)$ for all x. Of course if x is uniformly distributed on the unit circle, these conditional densities do satisfy the constraint that g is the weighted sum of the conditional densities.

With the Shannon information cost on this same unit-circle, quadratic tracking problem, the optimal form of $q(y \mid x)$ is proportional to $\exp(-(y-x)^2/\lambda)$, where "$y - x$" is interpreted as distance between y and x along the circle. It also leads to continuously distributed x arond the circle.

Now consider, though, what happens when we cut the circle, making g uniform on an interval $(0, 2\pi)$. Now there are end effects. With quadratic loss, we will never want to choose an x right at the boundary of the interval, unless we had exact knowledge that this was the value of y. With Shannon costs, it never pays to have such precise knowledge. As a result, with Shannon costs the distribution of x puts zero probability in the neighborhood of the ends of the intervals, and once we know this, the analyticity of $C(x)$ delivers us the result that the distribution of x is discrete.

With L_1 costs, though, it is possible that the $q(y \mid x)$ grows more concentrated around x at the ends of the interval, and even becomes entirely discrete at the boundary. Even if not, the presence of an interval with zero $p(x)$ density at the boundaries does not propagate into a requirement for discreteness with L_1 costs. We do not have an analytic proof of this, but numerical calculations seem to support this conclusion. When x and y are discretized, with 10 points in the x distribution and 100 in the y distribution, the solution maintains 10 distinct $q(y \mid x)$ shapes, all showing a region of y's where $q(y \mid x) = 0$, a region where $q(y \mid x) = g(y) = .01$, and an interval where $q(y \mid x)$ is larger. There are ten non-overlapping regions where $q(y \mid x)$ is larger than g, one such interval for each x. When λ is small enough that we do not end up with the no-information,

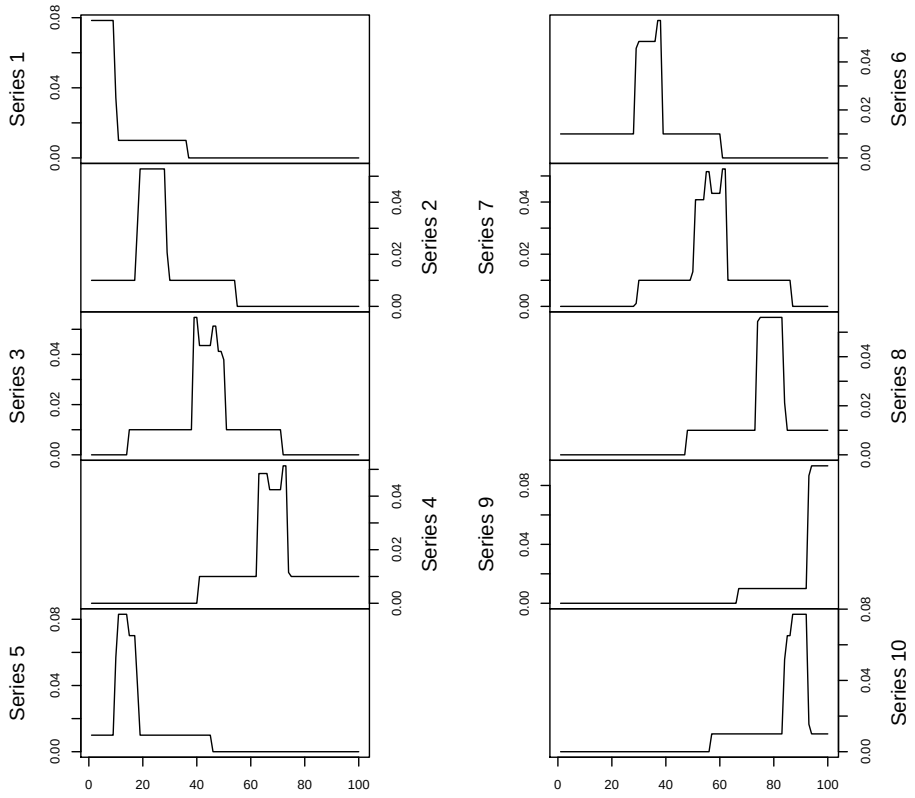


FIGURE 10

Conditional densities for y with 10-point x distribution, L_1 information cost, $\lambda = 4$. See text for interpretation of this and Figure 11.

$q \equiv g$ solution, increasing the number of points of support always seems to increase the objective function value. Figures 10 and 11 show numerical solutions for $\lambda = 4$ and $\lambda = 8$. The $\lambda = 8$ solution is close to the value at which the solution reverts to no-information, and in fact delivers almost exactly the same value of the objective function as the no-information solution, despite having non-trivial probability weight on each of the 10 points of support.

Note that in both figures the posterior densities are either .01 (the value of the uniform pdf under the prior), zero, or above .01. Each conditional density puts the posterior density above .01 on a unique interval, with these intervals not overlapping across the 10 x values. The lower information cost in Figure 10 is reflected in more x values ruled out, with posterior pdf zero. In these figures the computations were done with 10 points of support. Expanding the number of points of support to 15 still left all 15 distinct and with positive probabilities, unless the information cost was pushed above a critical level, in which case the solution collapsed to a single point of support.

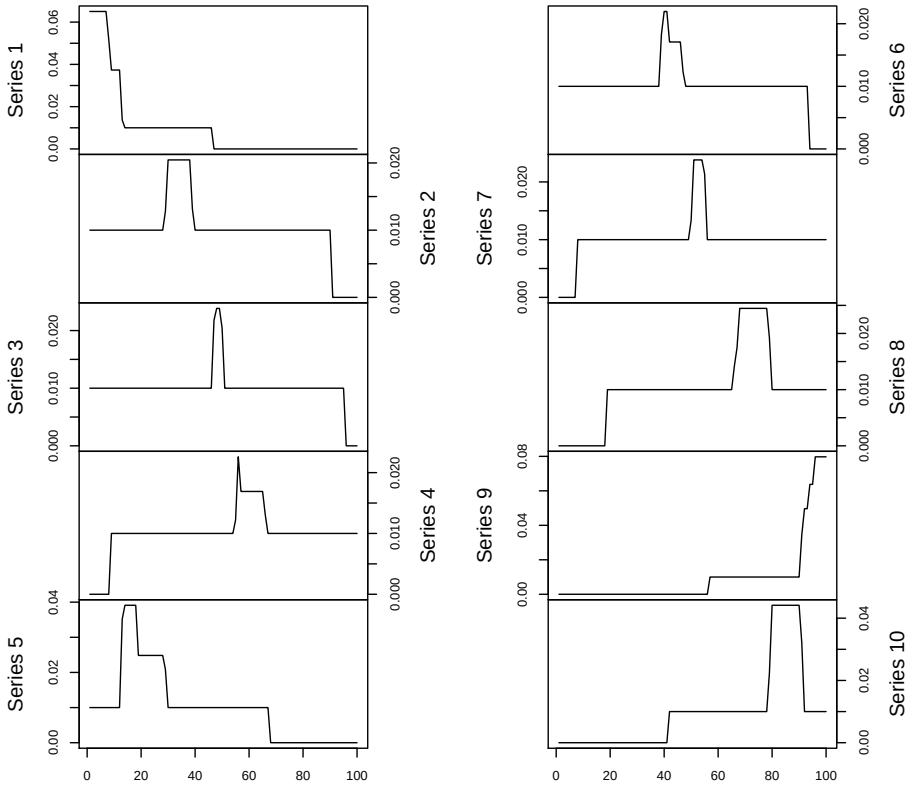


FIGURE 11

Conditional densities for y with 10-point x distrbiution, L_1 information cost, $\lambda = 8$.