NBER WORKING PAPER SERIES

A SPARSITY-BASED MODEL OF BOUNDED RATIONALITY

Xavier Gabaix

Working Paper 16911
http://www.nber.org/papers/w16911

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2011

I thank David Laibson for a great many enlightening conversations about bounded rationality over the years. For very good research assistance I am grateful to Alex Chinco, Tingting Ding and Farzad Saidi, and for helpful comments to Abhijit Banerjee, Daniel Benjamin, Douglas Bernheim, Vincent Crawford, Stefano Dellavigna, Alex Edmans, David Hirshleifer, Philippe Jéhiel, Botond Koszegi, Bentley MacLeod, Sendhil Mullainathan, Lasse Pedersen, Matthew Rabin, Antonio Rangel, Yuliy Sannikov, Andrei Shleifer, Jeremy Stein, Laura Veldkamp, and seminar participants at Berkeley, Cornell, Duke, Harvard, INSEAD, LSE, MIT, NBER, NYU, Princeton, the Stanford Institute for Theoretical Economics, and Yale. I also thank the NSF for support under grant DMS-0938185. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

A Sparsity-Based Model of Bounded Rationality
Xavier Gabaix
NBER Working Paper No. 16911
March 2011
JEL No. D03,D42,D8,D83,E31,G1

ABSTRACT

This paper proposes a model in which the decision maker builds an optimally simplified representation of the world which is "sparse," i.e., uses few parameters that are non-zero. Sparsity is formulated so as to lead to well-behaved, convex maximization problems. The agent's choice of a representation of the world features a quadratic proxy for the benefits of thinking and a linear formulation for the costs of thinking. The agent then picks the optimal action given his representation of the world. This model yields a tractable procedure, which embeds the traditional rational agent as a particular case, and can be used for analyzing classic economic questions under bounded rationality. For instance, the paper studies how boundedly rational agents select a consumption bundle while paying imperfect attention to prices, and how frictionless firms set prices optimally in response. This leads to a novel mechanism for price rigidity. The model is also used to examine boundedly rational intertemporal consumption problems and portfolio choice with imperfect understanding of returns.

Xavier Gabaix
New York University
Finance Department
Stern School of Business
44 West 4th Street, 9th floor
New York, NY 10012
and NBER
xgabaix@stern.nyu.edu

1 Introduction

This paper proposes a tractable model of some dimensions of bounded rationality (BR). It is designed to be easy to apply in concrete economic situations, and to inject a modicum of bounded rationality into existing models. It allows to study how and when bounded rationality makes an important difference for economic outcomes.

Its principles are the following. First, the decision maker in the model is not the traditional rational agent, but is best thought of as an economist building a simplified model of the world (a model-in-model). He builds a representation of the world that is simple enough, and thinks about the world through his partial model. Second, and most crucially, this representation is "sparse," i.e., uses few parameters that are non-zero or differ from the usual state of affairs.[1] I draw from the fairly recent literature on statistics and image processing to use a notion of "sparsity" that still leads to well-behaved, convex maximization problems. Third, the maximization can itself be imperfect, with a penalty that rises as the action taken becomes increasingly different from the default action, and it relies on the same sparsity criterion.

The decision maker simplifies his model of the world. For instance, he builds a model where some parameters are irrelevant (while they actually do matter to some degree), where some future cash flows do not occur, and where some variables are deterministic rather than random. He assumes convenient probability distributions rather than the complexity of reality: e.g., he might assume a distribution with two outcomes rather than a continuum of outcomes. These choices are controlled by an optimization of his representation of the world.

To motivate the model, I first consider a simple situation in which the decision maker wishes to make a decision that should be the weighted sum of many factors, such as his own income but also GDP growth in his country, the interest rate, recent progress in the construction of plastics, interest rates in Hungary, the state of the Amazonian forest, etc. Since it would be too burdensome to take all of these variables into account, he is going to discard most of them.[2] I study how to specify the cost of enriching the decision maker's

[1] The meaning of "sparse" is that of a sparse vector or matrix. For instance, a vector in $\theta \in \mathbb{R}^{100,000}$ with only a few non-zero elements is sparse.

[2] Ignoring variables altogether and assuming that they do not differ from their usual values are the same thing in the model. For instance, in most decisions we do not pay attention to the quantity of oxygen that is available to us because there is plenty of it. In the model, ignoring the oxygen factor is modeled as assuming that the quantity of oxygen available is the normal quantity. Indeed, the two are arguably the same thing.

representation of the world. Following antecedents in statistics and applied mathematics (Tibshirani 1996, Candès and Tao 2006, Donoho 2006), I show that one is particularly appealing: the ℓ_1 norm, i.e., the sum of absolute values of the non-zero updates in the variables. The reasons are as follows. First, a quadratic cost would not generate sparsity: small updates would have a miniscule penalty, hence under that model the decision maker would have non-sparse representations. Second, a fixed cost per variable would give sparsity but lose tractability; fixed costs lead to non-convex problems that make the solution very complicated in general. Instead, the ℓ_1 penalty both gives sparsity and maintains tractability. The model generates full or partial inattention to many variables.

The unweighted ℓ_1 criterion, used in the basic quadratic target problem, may not work in general: for instance, dimensions might not be comparable – e.g., the units could be different. I study how to generalize it. It turns out that, under some reasonable conditions, there is only one unique algorithm that (i) penalizes the sum of absolute values in the symmetric quadratic target problem, and (ii) is invariant to changes in units and various rotations of the problem. This is the algorithm I state as the "Sparse BR" algorithm. Hence, basic invariance considerations lead to an algorithm that is fairly tightly constrained. In addition, the algorithm involves just a simple optimization problem, so it is easy to apply.

I apply the model to a few of the main building blocks of economics, so that a modicum of bounded rationality can be injected into them and we can see when and how bounded rationality makes a difference for economic outcomes.

I first study intertemporal consumption. In this model, the agent may not think about all sources of income variables. Namely, he anticipates more about the usually important one, and less or nothing at all about the small ones. As a result, the marginal propensity to consume is different across income streams, whereas it would be the same in the traditional model. This is much like Thaler's (1985) "mental accounts." Also, this generates systematic deviations from Euler equations: they point towards inertia as agents will react in a dampened way to many future variables.

The next basic machinery of economics I apply BR to is a decision maker buying a vector of n goods. He is the traditional agent, except that he wishes to economize on thinking about all prices. The model generates a zone of insensitivity to prices: when prices are close to the average price, the decision maker does not pay attention to them. I then study how a firm will optimally price goods sold to such BR consumers. It is clear that the firm will not just choose any price strictly inside the zone of consumer inattention: it will rather select a price at its upper bound. Hence, a whole zone of prices will not be picked by firms. Even as the marginal cost of goods changes, there will be a zone of complete price rigidity. In addition, there is an asymmetry: there will sometimes be discrete downward jumps of the

price ("sales") but no corresponding upward jumps from the normal price (the asymmetry is due to the fact that the firm wants to keep a price as high as possible). Hence, we yield a tractable mechanism for price rigidity based on consumer bounded rationality rather than firms' menu costs.

Then, I also consider a few more psychological phenomena. One is that of cognitive overload: when the agent is confronted with too many decisions to make, the "cognitive budget constraint" becomes saturated and the quality of his decision making decreases. I also consider the endowment effect. In the model, the agent wishes to stay close to the default or status quo, which naturally generates an endowment effect. The value added by the model is that it yields a prediction of the size of the effect. As the Sparse-BR agent wishes to remain with the status quo when there is more model uncertainty, we obtain a higher endowment effect when the value of the good is more uncertain. This is different from prospect theory where the size of the effect depends only on the hedonic value of the good. Hence, the model explains why more experienced traders (List 2003) exhibit a much weaker endowment effect.

This paper tries to strike a balance between psychological realism and model tractability. The goal for the model is to be applicable without extensive complexity, and at the same time to capture some dimensions of bounded rationality. The central elements of this paper – the use of the ℓ_1 norm to model bounded rationality (rather than physical transaction costs), the accent on sparsity, and the Sparse BR algorithm – are, to the best of my knowledge, novel. I defer the discussion of the relationship between this paper and the rest of the literature to later in the paper when the reader is familiar with the key elements of the model.

The plan of the paper is as follows. Section 2 motivates the model in the context of a stylized model where the goal is to hit a target. Section 3 states the basic model. Section 4 applies the latter to a few basic economic problems. One is how a BR consumer selects a bundle of n goods while not completely processing the vector of prices. I also work out how a monopolist optimally sets prices given such a consumer: we will yield a novel source of real price rigidity, alongside occasional "sales" with large temporary changes in prices. Section 5 applies the idea of different representations to the simplification of random variables and categorization, using the language of "dictionaries" from the applied mathematics literature. Section 6 presents various enrichments of the model, for instance to discrete actions and models with constraints. It also discusses links with existing themes in behavioral economics. Section 7 discusses the limitations of this approach, and concludes. Many proofs are delegated to the appendix or the online appendix.

2 A Motivation: Sparsity and ℓ_1 Norm

We are developing a model where agents have sparse representations of the world, i.e., many parameters are set to "0," the default values. To fix ideas, consider the following decision problem.

Problem 1 *(Choice Problem with Quadratic Loss) The random variables x_i and weights μ_i are freely available to the decision maker, though perhaps hard to process. The problem is: pick a to maximize $V(a,x,\mu) = \frac{-1}{2}(a - \sum_{i=1}^{n}\mu_i x_i)^2$.*

If the x_i's are taken into account, the optimal action is

$$a(\mu) = \sum_{i=1}^{n} \mu_i x_i.$$

For instance, to choose consumption a (normalized from some baseline), the decision maker should consider not only his wealth, x_1, and the deviation of GDP from its trend, x_2, but also the interest rate, x_{10}, demographic trends in China, x_{100}, recent discoveries in the supply of copper, x_{200}, etc. There are $n > 10,000$ (say) factors $x_1, ..., x_n$ that should in principle be taken into account. However, most of them have a small impact on his decision, i.e., their impact μ_i is small in absolute value.

Hence, we want to model an agent that does not wish to bear the costs of analyzing all these dimensions. He will just analyze "the most important ones." Hence, he will calculate

$$a(m) = \sum_{i=1}^{n} m_i x_i$$

for some vector m that endogenously has lots of zeros, i.e., m is "sparse." For instance, if the agent only pays attention to his wage and the state of the economy, m_1 and m_2 will be non-zero, and the other m_i's will be zero.

Consider the expected loss from taking the imperfect (but parsimonious) policy $a(m)$ rather than the fully inclusive (but very expensive) policy $a(\mu)$: $L = \mathbb{E}[V(a(\mu), x, \mu) - V(a(m), x, \mu)]$. We have $L = \mathbb{E}[0 - \frac{-1}{2}(\sum_i m_i x_i - \sum_i \mu_i x_i)^2]$, and assuming for simplicity that the x_i's are i.i.d. with mean 0 and variance 1,

$$L = \frac{1}{2}\sum_i (m_i - \mu_i)^2.$$

We desire a systematic procedure to predict how an agent will pick the "important dimensions" that will receive non-zero weights m_i. We will formulate the choice of m as an

optimization problem:

$$\min_{m\in\mathbb{R}^n} \frac{1}{2}\sum_i (m_i-\mu_i)^2 + \kappa\sum_i |m_i|^\alpha \tag{1}$$

with $\kappa > 0$ and $\alpha \geq 0$. The first term is the utility loss from an imperfect representation of the world, m_i. The second term, $\kappa\sum_i |m_i|^\alpha$, represents a penalty for lack of sparsity: when the decision maker has a non-zero or large $|m_i|$, he pays a cost $\kappa |m_i|^\alpha$ where κ is a cost parameter.

Let us analyze what α would be appealing given that we want to capture that the decision maker has a sparse vector m. One natural choice would be $\alpha = 2$, which leads to a quadratic cost function. Then, we obtain $-(m_i-\mu_i) - 2\kappa m_i = 0$, i.e., $m_i = \mu_i/(1+2\kappa)$. This does not yield any sparsity: all features matter, regardless of whether μ_i is small or large. We just get some uniform dampening. Hence, we seek something else.

Another natural modeling choice would be $\alpha = 0$ (with the convention $|m|^\alpha = 1_{m\neq 0}$), which leads to a fixed cost function: the decision maker pays a cost κ for each non-zero element. Then, the solution is: $m_i = \mu_i$ if $|\mu_i| \geq \sqrt{2\kappa}$, and $m_i = 0$ otherwise. Now we do obtain sparsity. However, there is a large cost in terms of tractability. Problem 1 is no longer convex when $\alpha = 0$ (it is convex if and only if $\alpha \geq 1$). Its general formulation ($\min_{m\in\mathbb{R}^n} F(m) + \kappa\sum_i 1_{m_i} \neq 0$, for a convex F) is very hard to solve, and indeed generally untractable in a precise sense.[3]

Now consider the problem with $\alpha = 1$, i.e., a linear cost (with absolute values), as argued in the recent statistics and applied mathematics literature (Tibshirani 1996, Candès and Tao 2006, Donoho 2006). Then, problem (1) is convex. Let us solve it. Differentiating (1), we have:

$$-(m_i-\mu_i) - \kappa\cdot sign(m) = 0 \tag{2}$$

where $sign(m)$ is the sign of m ($sign(0)$ is the shorthand for some number between -1 and 1). Let us solve (2) when $\mu_i > 0$. When the solution is $m_i > 0$, we obtain $m_i = \mu_i - \kappa$, which requires $\mu_i > \kappa$. When $0 \leq \mu_i \leq \kappa$, $m_i = 0$. In general, we have:

$$m_i = \tau(\mu_i, \kappa) \tag{3}$$

for the function τ which is plotted in Figure 1 and defined as follows.

Definition 1 *The "anchoring and adjustment" function τ is*

$$\tau(\mu,\kappa) = (|\mu| - |\kappa|)_+ sign(\mu), \tag{4}$$

[3]It is "NP-complete" (Mallat 2009, chapter 12) in the terminology of complexity theory (if vector μ has 1,000 components, the brute-force solution would be to study the $2^{1000} \simeq 10^{300}$ subsets of fixed costs).

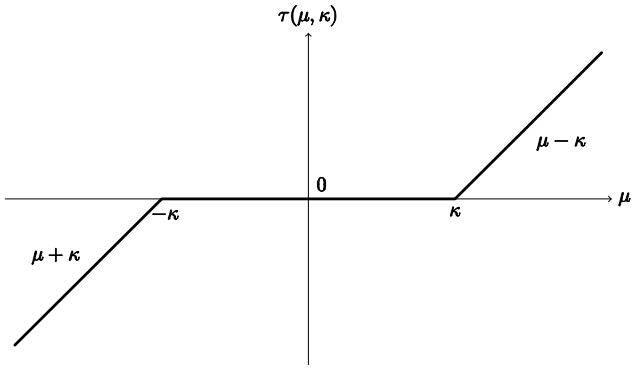

Figure 1: The anchoring and adjustment function τ

i.e., for $\kappa \geq 0$,

$$\tau(\mu, \kappa) = \begin{cases} \mu + \kappa & \text{if } \mu \leq -\kappa \\ 0 & \text{if } |\mu| < \kappa \\ \mu - \kappa & \text{if } \mu \geq \kappa \end{cases} . \tag{5}$$

The salient features of (3) are, first, that when $|\mu_i| < \kappa$, $m_i = 0$: all the small components are replaced by 0. This confers sparsity on the model. Second, for $\mu_i > \kappa$, $m_i = \mu_i - \kappa$. This corresponds to a partial adjustment towards the correct value μ_i. This motivates the term "anchoring and adjustment," a phenomenon demonstrated by Tversky and Kahneman (1974). In their experimental evidence there is anchoring on a default value and partial adjustment towards the truth (e.g., people pay only partial attention to the base rate when forming probability inferences).

Formulation (3) yields sparsity: all terms that have $|\mu_i| < \kappa$ are replaced by $m_i = 0$. For $\mu_i > \kappa$, we get $m_i = \mu_i - \kappa$, so there is a certain degree of dampening.[4]

The conclusion is that we can use the ℓ_1 norm, i.e., the one that corresponds to $\alpha = 1$ in (1), to generate sparsity and tractability at the same time. It is easy to check that sparsity is obtained if and only if $\alpha \in [0, 1]$, and tractability (a convex maximization problem) is obtained if and only if $\alpha \in [1, \infty)$. *Hence,* $\alpha = 1$ *(the* ℓ_1 *norm) is the only parametrization that yields both sparsity and tractability.*

We record the following lemma. Note that, in the notation m^d, d indicates a default value, not a power.

[4] Also, it is easy to see that m has at most $\min\left(\|\mu/\kappa\|_1, \|\mu/\kappa\|_2^2\right)$ non-zero components (because $m_i \neq 0$ implies $|\mu_i/\kappa| \geq 1$). Hence, even with infinite-dimensional μ and m, provided the norm of μ is bounded, m has a finite number of non-zero components, and is therefore sparse.

Lemma 1 *For $A > 0$, $K \geq 0$, and a real number m^d, the solution of*

$$\min_m \frac{A}{2}(m-\mu)^2 + K\left|m - m^d\right|$$

is

$$m = m^d + \tau\left(\mu - m^d, \frac{K}{A}\right)$$

where τ is the anchoring and adjustment function given in (4).

Proof. By shifting $m \to m - m^d$, $\mu \to \mu - m^d$, it is enough to consider the case $m^d = 0$. The f.o.c. is

$$A(m-\mu) + Ksign(m) = 0.$$

That is, $m = \tau(\mu, K/A)$. ■

Let me discuss the interpretation of the model. In this model, the decision maker is aware of dimension i, and even x_i: still, if its importance $|\mu_i|$ is less than κ, he discards that dimension. In that sense, he behaves like an economic modeler (or a physics modeler for that matter): an economic modeler is aware that there are many things outside his model, and he often knows how to model them; still, he wishes to discard those dimensions to keep the model simple. The decision maker does the same here.

Hence, the interpretation suggests some change of focus compared to the more conventional economic approach, which is that of "optimization under informational constraints" (see Veldkamp 2011 for an excellent survey of this literature). In the present model, the decision maker knows a lot, but prefers to discard a lot of minor information to keep his model sparse. The differences with existing approaches will be discussed in greater depth later.[5]

The τ function generates underreaction. It is worth seeing that this is a robust feature of models of noisy cognition. Take the canonical model where the agent receives a signal $s = \mu + \varepsilon$, with a non-degenerate noise ε uncorrelated with μ whose variance diminishes with cognitive effort. Then, to minimize a quadratic loss function $(m-\mu)^2$, it is well known that for (μ, ε) Gaussian with mean 0, the optimal signal extraction is $m(s) := \mathbb{E}[\mu \mid s] = \lambda s$ with $\lambda = var(\mu)/var(s) < 1$. This implies that $\mathbb{E}[m(s) \mid \mu] = \lambda\mu$, which generates dampening as $\lambda < 1$. This can increase our confidence that it is sensible for our model to generate dampening.

[5] In the Sims (2003) entropy framework with Gaussian x_i's, one may check that the DM's action (and signal) is $a = c\sum_i q_i x_i + \eta$ for a positive constant c and an independent Gaussian noise η; c and η are parametrized by the DM's information capacity. Hence, the DM pays attention to all x_i's. The decision is noisy but not sparse.

A tempting other formulation, ultimately not adopted here, is the following: use (1) to find which dimensions to eliminate, but for those that survive, use $m_i = \mu_i$, i.e., pay full attention (like in the fixed cost model). In other terms, use the "hard thresholding" function $\tau^H(\mu, \kappa) = \mu \cdot 1_{|\mu| \geq |\kappa|}$, rather than the "soft thresholding" function τ. Indeed, this τ^H function has been used in statistics (Belloni and Chernozhukov 2010). For some applications, this may be a useful model. However, it has several disadvantages. First, underreaction may actually be desirable, as argued above. Second, as it also seems to hold empirically, response functions are likely to be continuous, at least in the aggregate – and a goal of this paper is to find a tractable representation of a boundedly rational representative agent. Indeed, the soft thresholding function $\tau(\mu, \kappa)$ with ℓ_1 penalty and parameter κ can be seen as the representative agent aggregation of many heterogenous agents using the ℓ_0 penalty with different fixed costs k.[6] Third, and more technically, the fact that the hard thresholding function yields discontinuous response functions makes the model harder to handle. In contrast, the ℓ_1 formulation yields a convex decision problem, hence actions depend continuously on the environment.

Accordingly, I proceed with the ℓ_1 model and the soft thresholding, anchoring and adjustment function τ. I next generalize this idea to more general problems than the quadratic model.

3 The Basic Model

3.1 Model Statement

The decision maker has a value function $V(a, x, m)$, and wishes to select an action maximizing:

$$\max_a V(a, x, \mu). \tag{6}$$

The action $a \in \mathbb{R}^{n_a}$ is potentially multi-dimensional, i.e., maximization implies several actions: it could be the consumption of a good, the chosen allocation for a stock, etc.

The notation $m \in \mathbb{R}^{n_m}$ indicates the "representation of the world" (or "model-in-model") chosen by the agent, while the true (but potentially very complex) model is represented by a vector μ. In the previous examples, m_i is the importance on a dimension of the world: when $m_i = 0$, the agent does not think about dimension i, while when $m_i = \mu_i$, the agent fully pays attention to it. The value function is $V(a, x, m)$, and ideally the agent would like to maximize $V(a, x, \mu)$, i.e., the value function evaluated at the true model μ. However, his

[6] Indeed, if the distribution of k's is $f(k, \kappa) = 1_{k > \kappa} \kappa / k^2$, aggregation is exact: $\tau(\mu, \kappa) = \int_0^\infty \tau^H(\mu, k) f(k, \kappa) \, dk$.

concern for sparsity will make him choose a simpler (actually sparser) model m rather than the true model of the world μ. Vector $x \in \mathbb{R}^{n_x}$ is a series of quantitative features the decision maker might pay attention to (in a way modulated by m). When there is no such x (see Example 2 below), the value function is simply $V(a, m)$.

Finally, the decision maker has a representation $m^d \in \mathbb{R}^{n_m}$ and a default action $a^d \in \mathbb{R}^{n_a}$. A default representation could be $m_i^d = 0$, i.e., "do not think about dimension i." Typically, a good value for a^d is the best response given by the default model, $a^d = \arg\max_a V\left(a, x, m^d\right)$. That will often imply $a^d = 0$, "do nothing," or more precisely "do not deviate from the usual action." Arguably, such "do nothing" heuristics are among the most common decisions we make. It is directly at the core of the model as a default action.

It is useful to keep some examples in mind. The first one is the one we started with.

Example 1 *(Quadratic Target) We have* $V(a, x, m) = -\frac{1}{2}(a - \sum_i m_i x_i)^2$*, with true weights* μ *and sparser weights* m*.*

The next example shows how the model can capture "narrow framing."

Example 2 *(Narrow Framing): Let* a *be the optimal stock holding. Call* w *the baseline income wealth,* $\widetilde{r}$ *the excess stock return, and* $\widetilde{\varepsilon}$ *the labor income shock of the agent, so that time-1 consumption is* $c(m) = w + a\widetilde{r} + m\widetilde{\varepsilon}$*, with the true weight* $\mu = 1$*, and*

$$V(a, m) = \mathbb{E}u(w + a\widetilde{r} + m\widetilde{\varepsilon}).$$

In this example, $\mu = 1$ *means that when picking equities, the decision maker explicitly takes into account the other gambles in his life, such as labor income shocks. However, when* $m = 0$*, the decision maker uses "narrow framing" or "narrow bracketing" (e.g., Rabin and Weizsäcker 2009). The agent thinks about his optimal allocation in equities while forgetting about the other gambles in his life, such as future income shocks. Hence, the model can offer predictions about when the agent deviates from a narrow bracket.*

The third example demonstrates how the decision maker may not pay full attention to a variable of interest, such as the interest rate.

Example 3 *(Neglected Interest Rate) The decision maker starts with wealth* w*, consumes* a *at time* 1*, invests at a gross interest rate* R*, and consumes at time* 2*. His utility function is:*

$$V(a, R_t, m) = u(a) + v\left(R(m)(w - a)\right),\ R(m) = R^d + m\left(R_t - R^d\right).$$

When $m = \mu \equiv 1$*, the decision maker consciously uses the true interest rate* R_t*. However, when* $m = m^d \equiv 0$*, the decision maker does not pay attention to the interest rate; instead, he*

uses a default interest rate R^d. This interest rate might be the average historical real gross interest rate. Note also that the Euler equation fails.

To state the model, we assume a prior knowledge of the normal range of variation in the action, reflected by a variable η_a, and in the representation, indicated by η_m. I discuss them below. For X a random variable, I define: $\|X\|_\alpha = \mathbb{E}\left[|X|^\alpha\right]^{1/\alpha}$ for $\alpha \geq 0$. Unless specified otherwise, I take $\alpha = 2$.

I assume that the derivatives V_{aa} and V_{am} (i.e., the second derivatives with respect to a and m) are defined at $\left(a^d, x, m^d\right)$, and that V_{aa} is negative definite (which is the case if the function is locally strictly concave in a).

This paper proposes the following algorithm as a useful model of agents' behavior. It may be called the "Sparse Boundedly Rational" algorithm, or "Sparse BR" algorithm for short.

Algorithm 1 *(Sparse BR Algorithm) To solve the problem $\max_a V(a, x, \mu)$, the sparsity-seeking decision maker uses the following two steps:*

***Step 1. Choose an optimally sparse representation of the world**. Using the realism loss matrix Λ,*

$$\Lambda = -\mathbb{E}\left[V_{am} V_{aa}^{-1} V_{am}\right], \tag{7}$$

the agent chooses his optimally sparse representation of the world as the solution of:

$$\min_m \frac{1}{2}(m-\mu)' \Lambda (m-\mu) + \kappa[m]. \tag{8}$$

The first part is a measure of expected loss from an imperfect model m, while the second part is a psychic cost that is a penalty for the lack of sparsity in the model:

$$\kappa[m] = \kappa^m \sum_i \left|m_i - m_i^d\right| \left\|V_{m_i a} \eta_a\right\|. \tag{9}$$

***Step 2. Choose an optimal action**. The agent maximizes over the action a:*

$$\max_a V(a, x, m) - \kappa[a] \tag{10}$$

where the psychic cost of deviations from the default $\kappa[a]$ is

$$\kappa[a] = \kappa^a \sum_i \left|a_i - a_i^d\right| \left\|V_{a_i m} \eta_m\right\|. \tag{11}$$

In (7), (9), and (11), all derivatives of V are evaluated at the default $\left(a^d, m^d\right)$. The

unitless parameters κ^m and κ^a indicate the cost of deviations from the default. When $\kappa^a = \kappa^m = 0$, the decision maker is simply the traditional frictionless agent.

Let me comment on the parts of the model.

First-pass intuition for the model When $\kappa^m = 0$, the decision maker's model of the world is the correct one: $m = \mu$. When $\kappa^a = 0$, the maximization is perfect, conditional on the model-in-model. Hence, the model continuously includes the traditional model with no cognitive friction. When cognition costs κ^m are non-zero, the model exhibits inertia and conservatism: the model-in-model is equal to the default, and so is the action.

For many applications, it might be enough to just turn on either Step 1 or Step 2 of the model. In most of this paper, only Step 1 will be turned on, i.e., I will assume perfect maximization given the representation of the world ($\kappa^a = 0$).

When selecting m, the decision maker uses a quadratic approximation of the objective function The expression $L^{quad}(m) = \frac{1}{2}(m-\mu)' \Lambda (m-\mu)$ is the quadratic approximation of the expected loss from an imperfect model m. More specifically, consider a function V with no x, and $a(m) = \arg\max_a V(a, m)$, the best action under model m. The utility loss from using the approximate model m rather than the true model μ is $L(m) = V(a(\mu), \mu) - V(a(m), \mu)$. A Taylor expansion shows that for m close to μ, $L(m) = L^{quad}(m)$ to the leading order.[7] This motivates the use of the first term in (8): it is a representation of the utility loss from an imperfect representation.

One modeling decision in writing the Sparse BR algorithm is to use $L^{quad}(m)$ rather than the exact loss $L(m)$, which would be very complex to use for both the decision maker and the economist. The decision maker uses a simplified representation of the loss from inattention. This is one way to escape Simon's "infinite regress problem" – that optimizing the allocation of thinking cost can be even more complex than the original problem. I cut that Gordian knot by assuming a simpler representation of it, namely a quadratic loss around the default.

Finally, in evaluating (7), it is sometimes useful to take the expectation $\mathbb{E}$ over the distribution of x's (as in the quadratic model in Section 2), or to just take the realized values of x (then, $\mathbb{E}$ is simply conditional on x, i.e., it could be suppressed).

[7] As a solves $V_a(a, m) = 0$, the implicit function theorem gives $V_{aa}\delta a + V_{am}\delta m = 0$, i.e., $\delta a = -V_{aa}^{-1} V_{am} \delta m$ with $\delta m = m - \mu$. Hence, the loss is:

$$L = -V_a \delta_a - \frac{1}{2}\delta_a V_{aa} \delta_a = 0 + \frac{1}{2}(m-\mu)' \Lambda (m-\mu).$$

Defaults The model requires a default action a^d and a default representation m^d. In the applications below, the default action will be "do nothing" or "do as usual" while the default representation is "do not think about dimension i," $m_i^d = 0$. This said, richer defaults could be considered: the literatures on learning and in behavioral economics contain insightful theorizations of such defaults (Koszegi and Rabin 2006). Social and other processes might affect defaults in interesting ways.

Units and scaling Sparsity penalties κ^m and κ^a are unitless numbers. The model has the correct units: equations (8)-(11) all have the dimensions of V. Also, the equations are independent of the units in which the components of m and a are measured. For instance, if $|m_i|$ does depend on the units of m_i, $|m_i| \|V_{m_i a}\|$ does not. More generally, the model is invariant (for small changes) to reparametrizations of the action: for instance, if the agent picks consumption or log consumption, the representation chosen by the decision maker is the same. This adds some robustness and ease of use to the model.

The term $V_{m_i a}$ denotes by how much a change in the dimension m_i affects marginal utility V_a (i.e., $\frac{\partial V_a}{\partial m_i}$). Hence, it is a measure of how important dimension i is. However, the concept of marginal utility V_a is not unit-independent: it has the units of utils divided by actions. As we do need a unit-independent concept, (9) writes the penalty as $V_{m_i a}\eta_a$, where η_a represents the demeaned range of the action a. For instance, if $a \in [0, 100]$, then we could have η_a a random variable uniform on $[-50, 50]$. Written this way, the term $V_{m_i a}\eta_a$ becomes unit-independent. Variables η_a and η_m largely ensure that the model has the right units and scaling properties. They are typically not crucial in applications. To fully close the model, the following choices prove sensible: when they are one-dimensional, we can have $\eta_a = \sigma_a$, the standard deviation of a. One can typically say that η_m simply follows the distribution of μ, and a follows the distribution of $a^d(\mu, x)$, for instance.

However, the model is not invariant to the representations of the world m: some will be better for the agent than others. That is arguably a desirable feature of the model, and offers a simple way to model framing. For instance, suppose that w is real wage growth, π inflation, $w^{nom} = w + \pi$ is nominal wage growth, and that the agent has to guess real wage growth w. If the agent has access to w (say $x_1 = w$), he will use it, and his problem is simple. However, if the agent (which is more realistic in many contexts) has only direct access to nominal wage growth $x_1 = w^{nom}$ and inflation $x_2 = \pi$, with say $m_1^d = 1$ and $m_2^d = 0$, then his task will be harder, and will typically feature an incomplete adjustment for inflation.

Isn't the algorithm complex? The algorithm has been designed to be easy to use in economic applications. Also, it is not hard to use for the agent. For instance, Step 1

involves the maximization of a linear-quadratic problem (with an absolute value), and uses only the properties of the value function around the default. It is still not a completely trivial task, but it is simpler than the task of the traditional agent. In many cases, it is simply a collection of n_m independent maximization problems, for which the solution can be readily written down using the τ function (as we shall see below).

Step 2 is indeed rather complex, but not really more so than the traditional agent's problem. In some cases, it is simplified by the term $\kappa[a]$, which anchors many actions at their default and thus reduces the effective dimension of the action set to optimize on.

Why is the model set this way? The algorithm is written, first of all, to have some descriptive realism. That will be argued in the rest of the paper. Also, it is designed to have the following properties (for notational simplicity I drop the dependence on x in the remainder of this section):

(i) It generalizes the loss function of the quadratic problem in Section 2, as we shall soon see.

(ii) It gives the same answer irrespective of whether the decision maker maximizes $V(a,m)$ or $V(a,m)+B(m)$ for an arbitrary function B: it should do so because adding such a number $B(m)$ does not change the problem.

(iii) The model does not depend on third- and higher-order derivatives. This is to keep the model simple, and in some sense independent (at least locally) of various details like the third derivatives.

(iv) The model is invariant to the units of the components m and a.

The following proposition, proven in the appendix, says that there is a *unique* algorithm, namely the Sparse BR algorithm, that satisfies the above four criteria. In that sense, the model is tightly constrained, and equation (9) is rather necessary.

Proposition 1 *Normalize $m^d=0$. Suppose that the determination of m is*

$$\min_m \frac{1}{2}(m-\mu)'\Lambda(m-\mu)+K\left((m_i)_{i=1...n},\eta_a,V,V_m,(V_{am_i})_{i=1...n},V_{aa}\right) \tag{12}$$

for a penalty function K evaluated at V and its derivatives at point $\left(a^d,m^d\right)$. Suppose also that K satisfies:

(i) (Invariance with the units of m and a, and invariance by rotations of a) The value of K is unchanged under linear reparametrizations of m_i (for $i=1...n_m$) and of a: for all $\lambda_i\in\mathbb{R}$ and $A\in\mathbb{R}^{n_a\times n_a}$,

$$K(\lambda_i m_i,A'\eta_a,V,V_m,V_{am_i},V_{aa})=K(m_i,\eta_a,V,\lambda_i V_{m_i},\lambda_i AV_{am_i},AV_{aa}A') \tag{13}$$

(*ii*) *(Degree-1 scaling by affine transformations of V) Given a real $s > 0$ and a function $b(m)$ differentiable at m^d, a change $V(a,m) \to sV(a,m) + b(m)$ multiplies K by s.*

(*iii*) *(ℓ_1 norm in the basic quadratic problem) When the cost function K is evaluated for $V = -\frac{1}{2}(a_1 - m \cdot x)^2$ with $\|x_i\| = 1$ for all i and $\|\eta_{a_1}\| = 1$, we have*

$$K = \kappa^m \sum_i |m_i| . \tag{14}$$

Then, the penalty of m must be the one in Step 1 of Algorithm 1, i.e.,

$$K(m_i, m_a, V, V_{am_i}, V_{aa}) = \kappa^m \sum_i |m_i| \, \|V_{m_i a} \eta_a\| .$$

Proposition 1 justifies in some sense Step 1 of the algorithm. We match the basic quadratic targeting of the earlier section, and the model satisfies scale invariance. That leads to the formulation of $\kappa[m]$ in Step 1 of the algorithm.[8]

Step 2 is justified, heuristically, by using the idea that penalties for changing one's representation and penalties for changing one's action are treated symmetrically. This is why (11) is simply the rewriting of (9) by changing the roles of actions and representations.

The above might be a formal convenience, or perhaps it might reflect something slightly deeper in people's decision making: the "basic" algorithm would be given by the penalty (14), and then the mind would simply use the core algorithm after rescaling for the particular units of a situation. That leads the mind to the algorithm in (9).

Welfare In behavioral models, the welfare is often hard to assess, e.g., because of the existence of multiple selves in one agent (Bernheim and Rangel 2009). In the present model, this is relatively simpler as one might say that "fundamental" utility remains $V(a, \mu, x)$, not $V(a, m, x)$ under the chosen model. Put differently, if a benevolent advisor were to suggest perfect default models and actions, the DM would be better off, and would just follow the advisor.

Potential variants The online appendix discusses some variants that can be useful in some contexts but that I did not choose for the core model. For instance, rather than to have (9) satisfy unit-invariance, one could think of modeling the penalty $\kappa[m]$ as:

$$\kappa[m] = \sum_i \kappa_i \left| m_i - m_i^d \right| \tag{15}$$

[8]Note that the K function cannot depend on V_a as this value is generally 0 in the default policy.

where κ_i would be in utils over the units of m_i. This proposal may appear simpler than (9), but it turns out to be much more problematic to apply in practice: to use (15), at each stage one needs to take a stance on the value of κ_i for each i. Also, in a dynamic problem involving growth (say with a utility $\sum \rho^t c_t^{1-\gamma}/(1-\gamma)$), we should require κ_{it} to be proportional to $c_t^{1-\gamma}$ in order to make the model scale-invariant on the balanced growth path. This requires to set $\kappa_{it} \propto c_t^{1-\gamma}$ more or less manually. On the other hand, the adoption of $\kappa_i = \kappa^m \|V_{m_i a}\eta_a\|$ automatically provides the problem with a sensible scaling. Hence, it confers some parsimony on the model as there is no decision to make dimension-by-dimension (there is just one key parameter, κ^m, or, if one wishes, $\kappa^m \eta_a$, which is the same across dimensions m_i). At the same time, we shall see from the consequences of the model that it leads to sensible economic and psychological results.

This said, it is clear that some tasks (e.g., computing the 100th decimal of $\sqrt{2}$) are much harder than others; in some economic situations this is an important force, which could be formulated with a higher κ_i. However, dispensing with that additional degree of freedom does not significantly impact the model's economic realism.

Let us now apply the model to a concrete problem, so we can better see how it works.

3.2 Application: Quadratic Target Problem

We detail the application of the model to the quadratic target problem, Example 1. The online appendix develops Examples 2 and 3. The problem is:

$$\max_a V(a, x, \mu), \qquad V(a, x, m) = \frac{-s}{2}(a - m \cdot x)^2$$

where $s > 0$ indicates the size of stakes and the x_i's are uncorrelated with mean 0 and variances σ_i^2. The agent has access to a vector of information x. Vector m represents the weights to put on x, whose true value is μ. Instead, the agent will use $V(a, x, m)$, with m possibly sparse: $m_i = 0$ corresponds to not thinking about dimension i. The decision maker's response is as follows (the proof is in the appendix).

Proposition 2 *(Quadratic Loss Problem) In the quadratic optimization problem, the representation is*

$$m_i^* = m_i^d + \tau\left(\mu_i - m_i^d, \kappa^m \frac{\sigma_a}{\sigma_i}\right), \tag{16}$$

and the action taken is

$$a = a^d + \tau\left(\sum_i m_i^* x_i - a^d, \kappa^a \sqrt{\sum_i \sigma_{m_i}^2 \sigma_i^2}\right). \tag{17}$$

When $\kappa^m = 0$, $m = \mu$, and when $\kappa^a = 0$,

$$a = \sum_i m_i^* x_i.$$

Equation (16) features anchoring on the default value m_i^d and partial adjustment towards the true value μ_i: $m_i \in \left[m_i^d, \mu_i\right]$. For most applications where dimension i is non-salient, $m_i^d = 0$ is probably the right benchmark.

The decision maker does not deviate from the default iff $\left|\mu_i - m_i^d\right| \sigma_i < \kappa^m \sigma_a$, i.e., when dimension i cannot explain more than a fraction $(\kappa^m)^2$ of the variance of action a. *It is the relative importance of attribute i in decision a that matters for whether or not the decision maker will pay attention to attribute i, not its absolute importance in terms of, say, a dollar payoff.*

The model is scale-invariant in V (e.g., equation 8 is homogenous of degree 1 in V). As a result, the total amount of attention concerning decision a is the same whatever the stakes s. People will pay attention to say 80% of attributes, whether it is for a small decision (e.g., buying at the supermarket) or a big decision (e.g., buying a car). I conjecture that this feature is a good benchmark which would be interesting to evaluate empirically (I do not claim it will work perfectly, but I conjecture that it will hold more likely than the polar opposite prediction that people would be 1 million times more precise for a good that costs 1 million times more). Some evidence consistent with that is presented by Samuelson and Zeckhauser (1988), who show similar percentage price dispersion between cheap and expensive goods, and by Tversky and Kahneman (1981): people consider a \$5 discount more worthy of an extra shopping trip if it is for a \$15 calculator than for a \$125 jacket. Finally, there is casual evidence that many people do not spend more than 1 hour on retirement planning. Still, in some cases the scale-independence feature of the model may not be appropriate, and Section 6.1.3 endogenizes κ and renders attention more important for more expensive goods.

Equation (17) indicates that *when there is more uncertainty about the environment, the action is more conservative and closer to the default*: when $\sqrt{\sum_i \sigma_{m_i}^2 \sigma_i^2}$ is higher, a is closer to a^d. In the model, for a given amount of information $(m \cdot x)$, the power of default is higher when there is more residual uncertainty in the environment. This implication might be testable in the rich literature on defaults (Madrian and Shea 2001). In general, σ_m is the amount of model uncertainty for the decision maker, in a way that will be more specific in examples that are described below.

Calibration We can venture a word about calibration. As a rough baseline, we can imagine that people will search for information that accounts for at least $\xi^2 = 10\%$ of the

variance of the decision, i.e., if $|\mu_i|\,\sigma_i < \xi\sigma_a$. Then, using (16), we find $\kappa^m \simeq \xi$. That leads to the baseline of $\kappa^m \simeq 0.3$. The reader may find that, rather than 10%, $\xi^2 = 1\%$ is better (though this may be very optimistic about people's attention), which corresponds to $\kappa^m = \sqrt{1\%} = 0.1$ – a number still in the same order of magnitude as $\kappa^m \simeq 0.3$. By the same heuristic reasoning, we can have as a baseline $\kappa^a \simeq 0.3$. As it turns out, in subsequent work (Gabaix 2011), the a-priori calibration $\kappa^m \simeq 0.3$ works quite well in predicting subject's behavior in experimental games.

To conclude, the model generates inattention and inertia that respond to the local (i.e., for the decision at hand) costs and benefits. We now explore the model's consequences in a few applications.

4 Some Applications of the Model

4.1 Myopia in an Intertemporal Consumption Choice Problem

The agent has initial wealth w and future income x, he can consume c_1 at time 1, and invest the savings at a gross interest rate R. Hence, the problem is as follows.

Example 4 *(2-Period Consumption Problem). Given initial wealth w, solve*

$$\max_{c_1} u(c_1) + v(x + R(w - c_1))$$

where income is $x = x_* + \sum_{i=1}^{I} x_i$: there are I sources of income x_i, and we normalize $\mathbb{E}[x_i] = 0$.

Let us study the solution of this problem with the Sparse BR algorithm. The decision maker observes the income sources sparsely: he uses the model $x(m) = x_* + \sum_{i=1}^{K} m_i x_i$ with m_i to be determined. The action is the date-1 consumption c_1. We assume $u(c) = -e^{-\gamma c}$ and $v(c) = -e^{-\rho} e^{-\gamma c}$ where γ is the coefficient of absolute risk aversion and ρ the rate of time preference. The value function is:

$$V(c_1, x, m) = u(c_1) + v\left(x_* + \sum_{i=1}^{K} m_i x_i + R(w - c_1)\right).$$

We apply the basic Sparse BR algorithm for the case $\kappa^a = 0$ (frictional understanding of the world, frictionless maximization given that understanding). Calculations in the appendix show the following proposition.

Proposition 3 *(2-Period Consumption Model) With full maximization of consumption, the time-1 consumption is:*

$$c_1 = \frac{1}{1+R}\left(D + x_* + \sum_{i=1}^{I} m_i x_i\right) \tag{18}$$
$$m_i = \tau\left(1, (1+R)\frac{\kappa^m \sigma_{c_1}}{\sigma_{x_i}}\right)$$

with the constant $D = Rw + \frac{\rho - \ln R}{\gamma}$*. The marginal propensity to consume (MPC) at time 1 out of income source* i*,* $\partial c_1/\partial x_i$*, is:*

$$\left(\frac{\partial c_1}{\partial x_i}\right)^{BR} = \left(\frac{\partial c_1}{\partial x_*}\right)^{ZC} \cdot m_i \tag{19}$$

where $\left(\frac{\partial c_1}{\partial x_i}\right)^{BR}$ *is the MPC under the BR model and* $\left(\frac{\partial c_1}{\partial x_i}\right)^{ZC}$ *is the MPC under the zero cognition cost model (i.e., the traditional model). Hence, in the BR model, unlike in the traditional model, the marginal propensity to consume is source-dependent.*

Different income sources have different marginal propensities to consume – this is reminiscent of Thaler's (1985) mental accounts. Equation (19) makes another prediction, namely that consumers pay more attention to sources of income that usually have large consequences, i.e., have a high σ_{x_i}. Slightly extending the model, it is plausible that a shock to the stock market does not affect the agent's disposable income much – hence, there will be little sensitivity to it.[9]

There is a similarity of this model with models of inattention based on a fixed cost of observing information (Duffie and Sun 1990), in particular with the optimal rules of the allocation of attention developed by Abel, Eberly, and Panageas (2010), Gabaix and Laibson (2002), and Reis (2006). Because of the fixed cost, in those models the rules are of the type "look up the information every D periods." Those models are relatively complex (they necessitate many periods and either many agents or complex non-linear boundaries for the multidimensional s, S rules) whereas the present model is simpler and can be applied with one or several periods. As a result, the present model, with an equation like (19), lends itself more directly to empirical testing. The presence of different models of boundedly rational behavior may be helpful for empirical research in that area.

The Euler equation will only hold with the "modified" parameters. Hence, we have $\mathbb{E}_m\left[Rv'\left(c_2\right)/u'\left(c_1\right)\right] = 1$, but only using the expectation under model m. Note that it

[9] In other cases, the default policy might be to consume what is in one's wallet, up to keeping some minimum amount. Then, the MPC of a dollar bill found on the sidewalk would be 1.

features underreaction to future news, especially small future news.

4.2 Choosing n Consumption Goods

We next study a basic static consumption problem with n goods.

Example 5 *Suppose that the vector of prices is $p \in \mathbb{R}^n_{++}$, and the budget is y. The frictionless decision problem is to choose the optimal consumption bundle $c \in \mathbb{R}^n$: $\max_{c \in \mathbb{R}^n} u(c)$ subject to the budget constraint $p \cdot c \leq y$.*

The price of good i is $p_i^d + \mu_i$, where p_i^d is the usual price and μ_i is some price change. The decision maker may pay only partial attention to the price change, and consider the price of good i to be $p_i^d + m_i$. If $m_i = 0$, he proceeds as if the true price were the default price p_i^d, rather than the actual price $p_i^d + \mu_i$. For instance, in the experimental setup of Chetty, Looney, and Kroft (2009), μ_i could be a tax added to the price.

In this subsection, we study the case where the utility function is quasi-linear in money: there is a good n ("money") with constant marginal utility λ and price $p_i^d = 1$. This assumption will be relaxed in Section 6.1.2. We apply the model of Section 3.1, with an action $a = c$ and value function $V(c, m) = u(c) - \lambda \sum_{i=1}^{n-1} \left(p_i^d + m_i\right) c_i$.

Proposition 4 *In the decision maker's model, the deviation from the normal price is:*

$$m_i = \tau\left(\mu_i, \kappa^m \frac{p_i \sigma_{\ln c_i}}{\psi_i}\right). \tag{20}$$

If demand depends only on prices,

$$m_i = \tau\left(\mu_i, \kappa^m \sigma_{p_i}\right). \tag{21}$$

Equation (20) says that controlling for the volatility of consumption, inattention is greater for less elastic goods. The intuition is that for such goods the price is a small component of the overall purchasing decision (whose range is measured by $\sigma_{\ln c_i}$). Equation (21) indicates that in order to be remarked, a given price change has to be large as a fraction of the normal price volatility. It would be insightful to test those predictions. Chetty, Looney, and Kroft (2009) present evidence for inattention, but do not investigate empirically a relation like (20) and (21).

4.3 Optimal Monopoly Pricing and BR-Induced Price Stickiness and Sales

I next study the behavior of a monopolist facing a BR consumer who has the utility function $u(Q,y) = y + Q^{1-1/\psi}/(1-1/\psi)$ when he consumes a quantity Q of the good and has a residual budget y. So, if the price is p, the demand is $D(p) = p^{-\psi}$ where $\psi > 1$ is the demand elasticity.[10] The consumer uses the Sparse BR algorithm; by the previous analysis, his demand is:

$$D^{BR}(p) = D\left(p^d + \tau\left(p - p^d, \kappa\right)\right) \tag{22}$$

where, by (20), $\kappa = \kappa^m p^d \sigma_{\ln Q}/\psi$. Hence, the consumer is insensitive to price changes when $p \in (p^d - \kappa, p^d + \kappa)$.[11] The default price p^d will be endogenized later to be the average price.

The monopolist picks p to maximize profits: $\max_p (p-c) D^{BR}(p)$ where c is the marginal cost (in this section, to conform to the notations of the optimal pricing literature, c denotes a marginal cost rather than consumption). The following proposition describes the optimal pricing policy.

Proposition 5 *With a BR consumer, the monopolist's optimal price is:*

$$p(c) = \begin{cases} \frac{\psi c + \kappa}{\psi - 1} & \text{if } c \leq c_1 \\ p^d + \kappa & \text{if } c_1 < c \leq c_2 \\ \frac{\psi c - \kappa}{\psi - 1} & \text{if } c > c_2 \end{cases} \tag{23}$$

where $c_1 = c^d - 2\sqrt{c^d \kappa/\psi} + O(\kappa)$ *solves equation (43), and* $c_2 = c^d + \kappa$ *with* $c^d := (1-1/\psi)p^d$ *is the marginal cost that would correspond to the price* p^d *in the model without cognitive frictions. The pricing function is discontinuous at* c_1 *and continuous elsewhere.*

Let us interpret Proposition 5. When $p \in \left(p^d - \kappa, p^d + \kappa\right)$, the demand $D^{BR}(p)$ is insensitive to price changes. Therefore, the monopolist will not charge a price $p \in \left(p^d - \kappa, p^d + \kappa\right)$: he will rather charge a price $p = p^d + \kappa$. We yield a whole interval of prices that are not used in equilibrium, and significant bunching at $p = p^d + \kappa$. There, the price is independent

[10] Previous work on rational firms and inattentive consumers includes Heidhues and Koszegi (2010) with loss-averse consumers, L'Huillier (2010) with differently-informed consumers, and Matejka (2010) with a Sims (2003)-type entropy penalty. Their models are quite different from the one presented here in specific assumptions and results. Still, there is a common spirit that behavioral consumers can lead to interesting behavioral by rational firms. A minimo, the present paper offers a particularly transparent and tractable version of this theme. Chevalier and Kashyap (2011) offer a theory of price stickiness and sales based on agents with heterogeneous search costs.

[11] This is a testable implication: the price elasticity of demand is the smaller the closer the price is to its default.

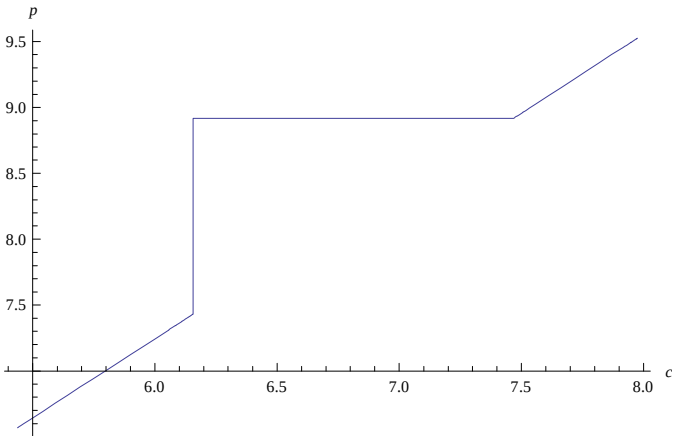

Figure 2: Optimal price p set by the monopolist facing boundedly rational consumers, as a function of the marginal cost c.

of the marginal cost. This is a real "stickiness."[12] This effect is illustrated in Figure 2.[13] We see that a whole zone of prices is not used in equilibrium: there is a gap distribution of price deviations from the norm.

For low enough marginal cost c, the price falls discretely, like a "sale." There is a discrete jump below the modal price but not above it: the asymmetry is due to the fact that in the inattention region $(p^d - \kappa, p^d + \kappa]$ the firm wishes to set a high price $p^d + \kappa$ rather than a low price. Hence, when we leave the inattention region, the price can rise a bit over $p^d + \kappa$, or otherwise has to jump discretely below $p^d - \kappa$.

The cutoff c_1 is much more below c^d than c_2 is above it. It deviates from the baseline c^d proportionally to $\sqrt{\kappa}$ whereas $c_2 = c^d + \kappa$.[14]

This simple model seems to account for a few key stylized facts. Prices are "sticky," with a wide range being insensitive to marginal cost. This paper predicts "sales:" a temporary large fall in the price after which the price reverts to exactly where it was (if c goes back to (c_1, c_2)). This type of behavior is documented empirically by Eichenbaum, Jaimovich, and Rebelo (forth.), Kehoe and Midrigan (2010), Klenow and Malin (forth.), and by Goldberg and Hellerstein (2010), which demonstrates the existence of local-currency price stickiness and sales in the domestic market of an exporter, consistent with this paper's view of cognitive frictions coming from the consumer side. In addition, the model says that the typical size of

[12] If the consumer's default is in nominal terms and mentally adjusting for inflation is costly, this model can easily yield nominal stickiness.

[13] The assumed values are $\psi = 6$, $p^d = 8.7$, and $\kappa = 0.025p^d$. They imply $\kappa = 0.22$, $c^d = 7.25$, $c_1 = 6.16$, $c_2 = 7.46$, $p(c_1) = 7.43$, and $p(c_2) = 8.92$.

[14] There is also a more minor effect. For very low marginal cost, consumers do not see that the price is actually too low: they replace p by $p + \kappa$. Hence, they react less to prices than usually (demand is less elastic), which leads the monopolist to raise prices. For high marginal cost, consumers replace the price by $p - \kappa$, so their demand is more elastic, and the price is less than the monopoly price.

a sales will be $p(c_2) - p(c_1)$, i.e., to the leading order

$$p(c_2) - p(c_1) = 2\sqrt{\frac{\kappa p^d}{\psi - 1}} \tag{24}$$

where $\kappa = \kappa^m p^d \sigma_{\ln Q}/\psi$. Hence, the model makes the testable prediction that the gap in the distribution of price changes, and the size of sales, is higher for goods with high consumption volatility and for goods that are less price elastic. The intuition is that for those goods price is a less important factor in the overall purchasing decision.

To close the model, one needs a theory of the default price. In a stationary environment, the simplest is to specify p^d to be the average empirical price

$$p^d = \mathbb{E}\left[p\left(\widetilde{c}, p^d\right)\right] \tag{25}$$

given the distribution over the marginal costs $\widetilde{c}$. By the implicit function theorem, for sufficiently small κ and a smooth non-degenerate distribution of costs, there is a fixed point p^d. In the small κ limit one can show that $p^d = \frac{\psi}{\psi-1}\overline{c} + \frac{2\overline{c}f(\overline{c})+2F(\overline{c})-1}{\psi-1}\kappa + o(\kappa)$ with $\overline{c} := \mathbb{E}[c]$ (the derivations are in the online appendix). Hence, the default price is higher than it would be in the absence of bounded rationality.

The model is robust to some form of consumer heterogeneity. The key is that the aggregate demand function $D(p)$ has kinks. Hence, if there are, for example, two types of agents – two $p_i^d + \kappa_i$ with $i \in \{1, 2\}$ – then we might also expect two reference prices.

This example illustrates that it is useful to have a tractable model, such as the Sparse BR algorithm, to think about the consequences of bounded rationality in market settings.[15] Also, the Sparse BR model is designed to generate inattention in the first place, not price stickiness and sales. Rather, it generates a potential new approach to price stickiness as an unexpected by-product.

4.4 Trading Inertia and Freezes

Step 2 of the Sparse BR algorithm indicates that the decision maker sticks with the default action when there is more model uncertainty (a higher $|\eta_m|$). Let us illustrate that effect in the context of trading freezes – the stylized fact that in moments of higher uncertainty many

[15] For instance, much of the analysis will carry over to a closely related setup where consumers are inattentive to the decimal digits of the price, i.e., $D^{BR}(n + x) = D^{BR}(n)$ for n a positive integer and $x \in [0, 1)$. There will be bunching at a price like \$2.99. Likewise, the online appendix solves the model with a fixed cost of thinking. It still yields price rigidity but loses the "sales" effect: there are two discontinuities in the optimal price function, rather than one.

agents simply withdraw from trading.

To be definite, take an agent with logarithmic preferences, selecting his equity share a (i.e., action a) when the risk premium is π and stock volatility is σ:

$$V(a, \pi, m) = \pi(m)\, a - \frac{1}{2}\sigma^2 a^2.$$

The decision maker is uncertain about the value of $\pi(m)$. Assume that Step 1 is done with $\kappa^m = 0$ (no friction), but that the decision maker remembers the model uncertainty $\|\eta_\pi\| = \sigma_\pi > 0$. Then, Step 2 of the Sparse BR algorithm is:

$$\max_a \pi a - \frac{1}{2}\sigma^2 a^2 - \kappa^a \left|a - a^d\right| \sigma_\pi.$$

Differentiating in a, $\pi - \sigma^2 a - \kappa^a sign\left(a - a^d\right)\sigma_\pi = 0$, and

$$a = a^d + \frac{\tau\left(\pi - \pi^d, \kappa^a \sigma_\pi\right)}{\sigma^2}, \qquad \pi^d = a^d \sigma^2. \tag{26}$$

We see that, indeed, the portfolio is "frozen" at $a = a^d$ whenever $\left|\pi - \pi^d\right| < \kappa^a \sigma_\pi$. The "freeze" range is higher when there is more uncertainty σ_π about fundamentals.

Trading freezes are often attributed to asymmetric information (lemons style) or Knightian uncertainty (as in Caballero and Krishnamurthy 2008), but here trading freezes come from bounded rationality.

4.5 Endowment Effect

The model generates an endowment effect alongside some additional predictions. Call $a \in [0, 1]$ the quantity of mugs owned (the prototypical good used by Kahneman, Knetsch and Thaler 1990), $x \geq 0$ the (random) utility for having a costless mug, and p the mug price. Net utility is $V(a, x) = a(x - p)$, and the decision problem is $\max_{a\in[0,1]} V(a, x) = a(x - p)$. Using Step 2 of the Sparse BR algorithm (equation 11), the problem is:

$$\max_{a\in[0,1]} a(\mathbb{E}[x] - p) - \kappa^a \sigma_x \left|a - a^d\right|$$

where $\sigma_x = \|\eta_x\|$ is the uncertainty about x.

The solution is simple and yields the willingness to pay (WTP) as well as the willingness to accept (WTA) for the mug. If $a^d = 0$ (i.e., the agent does not already own the mug), the solution is to buy iff $p \leq WTP = \mathbb{E}[x] - \kappa^a \sigma_x$. If $a^d = 1$ (i.e., the agent already owns the mug), the solution is to sell iff $\mathbb{E}[x] \leq WTA = \mathbb{E}[x] + \kappa^a \sigma_x$. The discrepancy between the

two,

$$WTA - WTP = 2\kappa^a \sigma_x, \tag{27}$$

is the endowment effect. In contrast, with loss aversion, the discrepancy is

$$WTA - WTP = (\lambda - 1)\,\mathbb{E}[x] \tag{28}$$

where $\lambda \simeq 2$ is the coefficient of loss aversion. (With loss aversion λ, selling the good creates a loss of $\lambda \mathbb{E}[x]$ whereas getting it creates a gain of only $\mathbb{E}[x]$.)

Hence, this paper's approach predicts that the endowment effect is increasing in the uncertain subjective utility (σ_x) of a good.

There is some consistent evidence: for instance, there is no endowment effect for, say, dollar bills which have a known hedonic value. Liersch et al. (2011) find a large endowment effect when extra noise (corresponding to a higher σ_x in the model) is added. Cao et al. (2011) obtain a related prediction in a Gilboa and Schmeidler (1989) type of setting, and propose that it helps understand a variety of behavioral finance phenomena. Finally, professional traders (List 2003) do not exhibit an endowment effect – according to this theory, that is because the value of the good is known.

To conclude, we have seen that the same model can shed light on a variety of situations and propose comparative statics for them: the determinants of inattention to prices, price rigidity, trading inertia and freezes, and the endowment effect. We can now turn to two other rather different uses of the model.

5 Other Consequences of Sparsity-seeking Simplification

I now show two fairly different instances of the theme that the decision maker simplifies reality to make decisions.

5.1 Dictionaries and Stereotypical Thinking

One particular interpretation of m is potentially interesting. Following the image processing literature (Mallat 2009), we could have a "dictionary" of prototypes: $(x_i(m_i))_{i \in I}$. The resulting representation is:

$$X(m) = \sum_{i \in I} x_i(m_i).$$

Note that the dictionary might be "redundant," i.e., $x_i(m_i)$ need not form a basis. For

instance, take a geometrical example and the plane $\mathbb{R}^2$. We could have: $x_1(\alpha, \beta, R)$ a circle with center (α, β) and radius R (the index is in $\mathbb{R}^3$); $x_2(\alpha, \beta, \alpha', \beta')$ a square starting with two "top" edges (α, β) and (α', β') (the index is then in $\mathbb{R}^4$). The total figure is the sum of all those primitive figures. We describe a picture from the basic constituents.

In a more social setting, we could denote by x an n-dimensional vector of attributes such as profession, nationality, income, social background, ethnicity, gender, height, etc. Then, the primitive words in the dictionary could be x_{Eng} for a stereotypical engineer, x_{Asian} for an Asian person, etc.

The key is that it is simple (sparse) to think in terms of "ready-made" categories, but harder (less sparse) to think in terms of a mix of categories. For instance, suppose that attributes are $x = (y_1, y_2)$, where y_1 is how good the person is at mathematics and y_2 is how good she is at dancing. Say that there exists a "type" engineer with characteristics $x_{\text{Eng}} = (8, -3)$, i.e., engineers are quite good at math, but are rather bad dancers (on average). Take a person called Johanna. First, we are told she is an engineer, and the first representation is $x_J = x_{\text{Eng}}$. Next, we are told she is actually a good dancer, with level $+4$ in dancing. Her characteristics are $x_J = (8, 4)$. How will she be remembered? We could say $x(m) = x_{\text{Eng}} + m x_{\text{Dancer}}$ where $x_{\text{Dancer}} = (0, 1)$, but full updating to $m = \mu = 7$ is costly. Hence, the information "good dancer" may be discarded, and only x_{Eng} will be remembered. The "stereotype" of the engineer eliminates the information that she is a good dancer.

More precisely, suppose that one wishes to maximize $V = -(a_1 - x_1)^2 - \gamma(a_2 - x_2)^2$, i.e., have a good model of the person with a weight γ on the dancing ability. We start from $x^d = x_{\text{Eng}} = (8, -3)$, and plan to move to $\left(x_1^d, x_2\right)$ (it is clear that the first dimension need not change). Applying the algorithm, we have $\max_m -\frac{1}{2}\gamma(a_2 - m_2)^2 - \kappa\sigma_{a_2}\gamma\left|m_2 - x_2^d\right|$. Hence, using Lemma 1, $x_2 = x_2^d + \tau\left(x_2^\mu - x_2^d, \kappa\sigma_{a_2}\right)$, i.e.,

$$x_2 = -3 + \tau(7, \kappa\sigma_{a_2}).$$

Thus, we get partial adjustment, with x_2 between the stereotypical level of dancing (-3) and Johanna's true level (4).

Hence, a model of sparsity-seeking thinking with a dictionary would be the following. Given a situation $x \in \mathbb{R}^{n_x}$, find a sparse representation that approximates x well, e.g., find the solution to:

$$\min_m ||x(m) - x||^2 + \sum_i \kappa_i \left|m_i - m_i^d\right|.$$

Then, people will remember $x(m) = \sum_{i \in I} x_i(m_i)$ rather than the true x. This generates a simplification of the picture, using simple traits. The above may be a useful mathematical

model of categorization. For instance, we might arrive at a model of "first impressions matter." The first impression determines the initial category. Then, by the normal inertia in this model, opinions are adjusted only partially. I note that some of these effects can be obtained in other models of categorization (Mullainathan 2001, Fryer and Jackson 2008). An advantage here is that categorization comes naturally from a general model of sparsity.

It is also clear that it is useful to have a dictionary of such stereotypes: they make thinking or, at the very least, remembering sparser. One may also speculate that education and life events provide decision makers with new elements in their dictionaries, and that as some dictionaries are more helpful to face new situations than others, "habits of thoughts" and "cultural references" might be usefully modeled by dictionaries.

5.2 Simplification of Random Variables

5.2.1 Formalism

Consider a random variable Y with values in $\mathbb{R}^n$. In his model-in-model, the decision maker might replace it with a random variable X that is "simpler" in some sense.

(i) X might have a different, arguably simpler distribution: for instance, we could replace a continuous distribution with a one-point distribution (e.g., $X = \mathbb{E}[Y]$ with probability 1) or with a two-point distribution $X = \mathbb{E}[Y] \pm m$. We could even have X to be a certainty equivalent of Y.

(ii) X might have independent components. For example, we could have $X_i \overset{d}{=} Y_i$, but the components $(X_i)_{i=1...n}$ are independent while the components $(Y_i)_{i=1...n}$ are not.

To formalize (i), call F and G the CDFs of X and Y, respectively. Then, $U = G(Y)$ has a uniform $[0,1]$ distribution, and we can define $X = F^{-1}(U)$ with the same U, so that X and Y are maximally affiliated.[16]

The choice of the model is subject to the same cost-benefit principles as in the rest of the model, e.g., one can use the same criterion to pick a simplified $X(m)$:[17]

$$\min_m \frac{1}{2}\Lambda \mathbb{E}\left[(Y - X(m))^2\right] + \kappa^m \sum_i \left|m_i - m_i^d\right| \left\|V_{m_i a}\eta_a\right\|. \tag{29}$$

[16] To formalize (ii), it is useful to use the machinery of copulas. For an n-dimensional vector Y, let us write $Y = \left(G_1^{-1}(U_1), ..., G_n^{-1}(U_n)\right)$ with U_i having the copula $C(u_1, ..., u_n)$, so that $\mathbb{E}[\phi(Y)] = \int \phi\left(G_1^{-1}(u_1), ..., G_n^{-1}(u_n)\right) dC(u_1, ..., u_n)$. In the simplified distribution, the marginals G_i^{-1} and the copula could be changed. To express X, we could have $X = \left(G_1^{-1}(U_1'), ..., G_n^{-1}(U_n')\right)$, where the U_i''s have the copula of independent variables, $C^\theta(u_1, \ldots, u_n) = u_1 \cdots u_n$. If we wish to have X_i's marginals to be simpler than Y_i's, like in (i), we can set $X = \left(F_1^{-1}(U_1'), ..., F_n^{-1}(U_n')\right)$ for some F_i.

[17] This way, $\mathbb{E}\left[(Y - X(m))^2\right]$ is the (squared) Wasserstein distance between the distributions of Y and $X(m)$, which has many good properties.

Eyster and Weizsäcker (2010) present experimental evidence for correlation neglect, i.e., the use of simplification (ii). The next example illustrates the possible relevance of simplification (i).

5.2.2 Application: Acquiring-a-company Game

Samuelson and Bazerman (1985) propose the following ingenious problem.

Example 6 *(Acquiring-a-company) The company is worth Y (uniformly distributed on $[0, 100]$) to Ann, and worth $1.5Y$ to you (you are a better manager than Ann). You can make a take-it-or-leave-it offer a to Ann, who knows Y. What offer do you make?*

In addition, the experimental setup makes sure that "Ann" is a computer, so that its answer can be assumed to be rational. Before reading the next paragraph, interested readers are encouraged to solve (without paper and pencil) Example 6 for themselves.

Experimentally, subjects respond with a mode around 60 and a mean around 40 (Charness and Levin 2009). However, the rational solution is $a = 0$. This is an extreme case of asymmetric information (related to the winner's curse).

Let us generalize the problem and state the BR solution to it. Assume that the company is worth $Y \sim U\left[\underline{Y}, \overline{Y}\right]$, with $\underline{Y} < \overline{Y}$, and define the mean payoff $\mathbb{E}[Y] = \left(\underline{Y} + \overline{Y}\right)/2$. The company is worth $(\lambda - 1)Y > 0$ more to the decision maker than to Ann. Hence, in the original problem $\underline{Y} = 0, \overline{Y} = 100$, and $\lambda = 1.5$.

Let us see how to state the model-in-model. We will see how, if the agent uses a simpler representation of probabilities, we account for the non-zero experimental value. This is a different explanation from existing ones (Eyster and Rabin 2005, Crawford and Iriberri 2007) which emphasize the assumption that the other player is irrational whereas the decision maker is rational. However, there is no "other player" in this game, as it is just a computer, and then those models predict a bid of 0 (Charness and Levin 2009).

The agent uses a representation of the dispersion in values, $X(m)$, simpler than the true distribution, Y. For instance, the agent might form a model of the situation by simplifying the distribution and replacing it by a distribution with point mass $X(0) = \mathbb{E}[Y]$. Then, the best response is $a = \mathbb{E}[Y]$, which is 50 in the basic game. This is not too far from the empirical evidence.

In a richer model-in-model, let us replace the distribution $Y \sim U\left[\underline{Y}, \overline{Y}\right]$ by a distribution $X(m) = \mathbb{E}[Y] \pm m$ with equal probability, for some $m \in [0, \Delta]$ with $\Delta \equiv \mathbb{E}[Y] - \underline{Y} = \overline{Y} - \mathbb{E}[Y]$ (we leave it to be an empirical matter to see what m is – the same way it is an empirical matter to see what the local risk aversion is). Given this model, the agent

maximizes $V(a,m) = \mathbb{E}\left[(\lambda X(m) - a)\, 1_{X(m)\leq a}\right]$. The resulting action is stated here and derived in the appendix.

Proposition 6 *In the acquiring-a-company problem, the Sparse BR bid by the decision maker is:*

$$a^* = \begin{cases} \mathbb{E}[Y] + m & \text{if } m \in \left[0, \frac{\lambda-1}{3-\lambda}\mathbb{E}[Y]\right] \\ \mathbb{E}[Y] - m & \text{if } m \in \left(\frac{\lambda-1}{3-\lambda}\mathbb{E}[Y], \Delta\right] \end{cases}$$

as long as $\lambda < 3$, and $a^* = \mathbb{E}[Y] + m$ otherwise. In particular, in the basic problem with support in $[0, 100]$ and $\lambda = 1.5$,

$$a^* = \begin{cases} 50 + m & \text{if } m \in [0, 16.66\ldots] \\ 50 - m & \text{if } m \in (16.66\ldots, 50] \end{cases}.$$

On the other hand, the model does not explain parts of the results in the Charness and Levin (2009) experiments. In a design where the true distribution of X is 0 or 1 with equal probability, the rational choice is $a = 0$. However, subjects' choices exhibit two modes: one very close to $a = 0$ and another around $a = 1$. The model explains the first mode but not the second one. It could be enriched to account for that additional randomness, but that would take us too far afield. One useful model is the contingency-matching variant in the online appendix: with equal probability, the decision maker predicts that the outcome will be 0 or 1, and best-responds to each event by playing 0 and 1 with equal probability. Hence, reality seems to be reasonably well accounted for by a mixture of the basic model and its contingencies-matching actions. All in all, the model is useful to describe behavior in the basic acquiring-a-company game even though it does not account for all the patterns in the other variants.

6 Complements and Discussion

6.1 Some Extensions of the Model

This subsection presents extensions of the Sparse BR algorithm that may be useful in some situations.

6.1.1 Discrete Actions

The model is formulated with a Euclidean action space, which is the substrate in many economic problems and confers a nice structure (e.g., a metric) on them. It extends to a discrete action space, as I illustrate here; the online appendix provides further details.

Action $a \in \{1, ..., A\}$ generates utility $V(a, x, \mu)$ of which the agent may use an imperfect $V(a, x, m)$. To formulate the model, some notations are useful: for a function $f(a)$, define $\|\Delta_{\eta_a} f(a)\| := \left(\frac{1}{A}\sum_{a=1}^{A} \mathbb{E}\left[\left(f(a) - f\left(a^d\right)\right)^2\right]\right)^{1/2}$ to be the dispersion of f across actions. Then, define $\sigma_i^m = \left\|\Delta_{\eta_a} V_{m_i}\left(a, x, m^d\right)\right\|$, so that σ_i^m is analogous to $\|V_{m_i a}\eta_a\|$ in Algorithm 1: it is the typical size of the marginal enrichment m_i. Define $\sigma_V = \left\|\Delta_{\eta_a} V(a, x, \mu)\right\|$, a scale for the dispersion in values across actions, which is analogous to $\|\eta_a V_{aa}\eta_a\|$ in the main algorithm. A natural analogue of Step 1 is:

$$\text{Step 1' : } \max_m \sum_i \frac{1}{2}\frac{\left(\sigma_i^m\right)^2}{\sigma_V}\left(m_i - \mu_i\right)^2 + \kappa^m \sum_i \left|m_i - m_i^d\right| \sigma_i^m.$$

Applying Lemma 1, it yields:

$$m_i^* = m_i^d + \tau\left(\mu_i - m_i^d, \frac{\kappa^m \sigma_V}{\sigma_i^m}\right). \tag{30}$$

Step 2 is simply $\max_a V(a, m^*, x)$ in the baseline case with $\kappa^a = 0$, and we can have a soft maximum otherwise, e.g., the probability p_a of picking a could be $p_a = e^{\beta V(a,m^*,x)} / \sum_{a'} e^{\beta V(a,m^*,x)}$ with $\beta > 0$ (this is further discussed in Gabaix 2011).

To illustrate this formalism, consider the choice between A goods: good $a \in \{1...A\}$ has a value:

$$V(a, m, x) = \sum_{i=1}^{n} m_i x_{ia}$$

with the x_{ia}'s i.i.d. across goods a, normalized to have mean 0 and standard deviations σ_i. The dimensions $i \in \{1, ..., n\}$ are (normalized) hedonic dimensions, e.g., price, weight, usefulness, esthetical appeal of each good. The default is $m^d = 0$. Applying the above Step 1', we obtain $\sigma_i^m = \sigma_i$ and finally:

Proposition 7 *Suppose that the agent chooses among A goods where good $a \in \{1...A\}$ has value $V(a, \mu, x) = \sum_{i=1}^{n} \mu_i x_{ia}$. Then, the boundedly rational perception of a good a is*

$$V(a, m^*, x) = \sum_{i=1}^{n} \tau\left(\mu_i, \frac{\kappa^m \sigma_V}{\sigma_i}\right) x_{ia} \tag{31}$$

with $\sigma_V = \left(\sum_{i=1}^{n} \mu_i^2 \sigma_i^2\right)^{1/2}$.

Hence, we obtain a dimension-by-dimension dampening, with small dimensions (small σ_i) dampened more or fully, very much in the spirit of the initial example we started from, but for discrete actions. Compared to process models of discrete choice with partial attention (e.g., Payne, Bettman and Johnson 1993, Gabaix, Laibson, Moloche and Weinberg 2006),

this model eschews sequential search (which typically does not lead to a closed form for the perceived value) and is thus much more tractable. Indeed, an equation such as (31) could be fairly directly estimated: when $\kappa^m = 0$, it is the rational actor model, while for $\kappa^m \to \infty$ the agent is fully inattentive. Empirical agents are likely to be in between.

6.1.2 Model with Constraints

We extend the model, so that it handles maximization under constraints. The decision maker wishes to solve:

$$\max_a V(a, x, \mu) \text{ subject to } B^k(a, x, \mu) \geq 0 \text{ for } k = 1...K. \tag{32}$$

For instance, B^1 could be a budget constraint, $B^1 = y - p(\mu) \cdot c$ where $p(\mu)$ is a vector of prices under the true model μ. As usual, we assume that V and $-B^k$ are concave in a.

We use Lagrange multipliers to formulate the extension of the model to constraints.

Algorithm 2 *(Sparse BR Algorithm with Constraints) To solve the problem in (32), the agent uses the following three steps.*

1. *Transformation into an unconstrained problem. Select the Lagrange multiplier $\lambda^* \in \mathbb{R}^K$ associated with the problem at the default model m^d:*

$$\max_a V\left(a, x, m^d\right) + \lambda^* \cdot B\left(a, x, m^d\right).$$

2. *BR-solve the new, unconstrained problem. Use the Sparse BR algorithm 1 for the value function V^* defined as:*

$$V^*(a, x, m) := V(a, x, m) + \lambda^* \cdot B(a, x, m)$$

without constraints. That returns a representation m and an action a.

3. *Adjustment to take the constraints fully into account. Call $b = (V^*_{aa})^{-1} B_a$ the $n_a \times K$ adjustment matrix and, for a vector of weights $\xi \in \mathbb{R}^K$, $a(\xi) = a + b\xi$. Pick a ξ that ensures that the K budget constraints are satisfied (typically, there is just one such ξ, but otherwise take the utility-maximizing one).*

Step 1 of Algorithm 2 picks a Lagrange multiplier λ^*, using the default representation m^d. This way, in Step 2 we can define a surrogate value function V^* that encodes the importance of the constraints by their Lagrange multipliers: V^* can be maximized without constraints, so that the basic Sparse BR algorithm can be applied.

The resulting recommended action may not respect the budget constraint. Hence, Step 3 adjusts the recommended action, so that all budget constraints are satisfied. The form $a(\xi)$ chosen is the linear form that returns the right answer to the benchmark case where $\kappa = 0$, as developed in Lemma 3 of the online appendix. Again, the model is in that sense quite constrained. The interpretation of action b is easiest to see in the case of just one budget constraint: suppose that there is a small income shock δy, so that the budget constraint becomes $B(a) + \delta y \geq 0$. Then, to the first order, the optimal action is $\delta a = b\xi$ for some ξ that ensures that the budget constraint is binding ($B_a \delta a + \delta y = 0$, so $\xi = -(B_a' b)^{-1} \delta y$): action b is proportional to $\partial a / \partial y$, the marginal response of the action to a change in income.

As an illustration, let us revisit the basic problem of maximizing a utility function subject to a budget set, which was developed in Section 4.2 by assuming a linear utility for residual money but which we can now solve with a budget constraint. Recall that the true vector of prices is $p = p^d + \mu$ and the problem is $\max_{c \in \mathbb{R}^n} u(c)$ subject to $y - p \cdot c \geq 0$. So, $V(c) = u(c)$, and there is one constraint, $B(c, m) = y - \left(p^d + m\right) \cdot c$.

In Step 1, we pick the Lagrange multiplier λ^* that corresponds to the problem: $\max_c u(c) + \lambda \left(y - p^d \cdot c\right)$ under the default price vector p^d. Then, we define:

$$V^*(c, m) = u(c) + \lambda^* \left(y - \sum_i \left(p_i^d + m_i\right) c_i \right). \tag{33}$$

This gives us a quasi-linear utility function, with linear utility for residual money.

Step 2 is as in Section 4.2, and yields a representation m (given by (20)) and as action the consumption vector $c\left(p^d + m\right)$. In Step 3 (applied with $c^d = 0$), the decision maker picks consumption $c = c\left(p^d + m\right) + \xi b$ with $b = \partial c\left(p^d, y\right) / \partial y$, and the scale factor $\xi \in \mathbb{R}$ ensures that the budget constraint holds: $\xi = (y - p \cdot c) / (p \cdot b)$. Psychologically, the decision maker thinks "I missed my budget by δy dollars, so I am going to make the regular adjustment $\delta c = b \delta y$ to that change in income in order to match by budget constraint." In that sense, the algorithm has a commonsensical psychological interpretation.

6.1.3 Cognitive Overload and Decisions under Stress

I present a way to model "cognitive overload" and the impact of decisions under stress.[18] This may be useful for analyzing bad decisions of people under stress, e.g., very poor individuals with difficult accidents in their lives or financiers in hectic markets (Hirshleifer, Lim, and Teoh 2009).

A slight and natural generalization of the Sparse BR model is required. Step 1 of the

[18] I thank Abhijit Banerjee for suggesting this application.

Sparse BR model becomes, using the notation $\kappa_i = \kappa^m \|V_{m_i a}\eta_a\|$:

$$\max_{\Theta\geq 0,\, m} \frac{-1}{2}(m-\mu)'\Lambda(m-\mu) + \kappa_0\Theta \tag{34}$$

$$\text{subject to } \Theta + \sum_i \frac{\kappa_i}{\kappa_0}\left|m_i - m_i^d\right| \leq \mathcal{C} \tag{35}$$

where $\Theta \geq 0$ is a measure of "cognitive leisure" (e.g., how much time is left to enjoy oneself rather than to think about decisions), κ_0 is the value of leisure in utils, and $\mathcal{C}$ is the decision maker's cognitive capacity.

In (34), the first term is the loss due to an imperfect model m while the second term $\kappa_0\Theta$ is the enjoyment of cognitive leisure.[19] The budget constraint (35) reflects that the cognitive capacity $\mathcal{C}$ is allocated between cognitive leisure Θ and the cost of processing m_i.

Let us solve the problem in the separable case, with Λ a diagonal matrix $\Lambda = diag\left(\Lambda_i\right)_{i=1...n}$:

$$\max_{m,\Theta} \frac{-1}{2}\sum_i \Lambda_i\left(m_i-\mu_i\right)^2 + \kappa_0\Theta - \lambda\left(\Theta + \sum_i \frac{\kappa_i}{\kappa_0}\left|m_i - m_i^d\right| - \mathcal{C}\right) + \pi\Theta$$

where λ and π are the Lagrange multipliers associated with (35) and $\Theta \geq 0$, respectively. Maximizing over Θ, we have $\kappa_0 - \lambda + \pi = 0$, i.e., if $\Theta > 0$, $\lambda = \kappa_0$, and $\lambda > \kappa_0$ otherwise. Next, maximizing over m_i and using Lemma 1, we have:

$$m_i = m_i^d + \tau\left(\mu_i - m_i^d, \frac{\lambda\kappa_i}{\Lambda_i\kappa_0}\right). \tag{36}$$

When the cost of cognition λ increases, the quality of decisions falls. To see this more analytically, consider the case where the decision maker has to make n decisions with the same $\kappa_i = \kappa$, $\Lambda_i = \Lambda$, $\mu_i = \mu > 0$ for all i, $m_i^d = 0$, and the decisions are important enough, so that $\Lambda\mu > \kappa$. Let us vary the number of decisions to be made (which is a way to model periods of stress) while keeping the cognitive capacity constant.

Proposition 8 *(Cognitive Overload) The attention paid to the problems $i = 1...n$ is:*

$$m_i = \min\left(\mu - \frac{\kappa}{\Lambda}, \frac{\mathcal{C}\kappa_0}{\kappa n}\right).$$

In particular, when $n \geq n^* = \mathcal{C}\kappa_0/\left(\kappa\left(\mu - \frac{\kappa}{\Lambda}\right)\right)$, the quality of decision making for each problem declines with the total number of problems n.

[19] The units of κ_0 and κ_i are in utils, so we might have $\kappa_0 = \|\eta_a W_{aa}\eta_a\|$ to get a definite value for κ_0.

Hence, in situations of extreme stress ($n > n^*$), the performance of all decisions declines because the decision maker hits his cognitive capacity.

Note that the formulation (34)-(35) may be useful in other domains. In particular, relaxing $\kappa_i = \kappa^m \|V_{m_i a}\eta_a\|$ may be useful when some dimensions are significantly harder to think about than others. For instance, a "salient" dimension could be modeled as having a lower κ_i. When the decision maker thinks about a dimension, the fact that λ and κ_i enter multiplicatively in (36) implies that the impact of salience is greater under a higher cognitive load. Finally, if the poor lead more stressful lives and therefore have a depleted amount of cognition $\mathcal{C}$, then the quality of their decision making is hampered and they are likely poorer as a result. Hence, we may have multiple equilibria, like in the poverty traps discussed in development economics. Currie (2009) reviews evidence that poor health leads to lower human capital (corresponding to a proxy for a higher κ in the model). Banerjee and Mullainathan (2010) document the hypothesis that the poor are more subject to behavioral biases, and derive some of its implications.

6.1.4 Diagonal Simplification for Λ

The following simplification is sometimes useful. Rather than Λ defined in (7), use a diagonal matrix with diagonal elements Λ_i instead:

$$\Lambda^{diag} = diag\left(\Lambda_1, \ldots, \Lambda_n\right), \qquad \Lambda_i = \max_k \frac{-V_{m_i a_k}^2}{V_{a_k a_k}}. \tag{37}$$

Then, Step 1 of the Sparse BR algorithm becomes:

$$\min_m \frac{1}{2}\sum_i \Lambda_i \left(m_i - \mu_i\right)^2 + \kappa^m \sum_i \left|m_i - m_i^d\right| \|V_{m_i a}\eta_a\| \tag{38}$$

whose solution is:

$$m_i = m_i^d + \tau\left(\mu_i - m_i^d, \kappa^m \frac{\|V_{m_i a}\eta_a\| V_{a_{k_i} a_{k_i}}}{V_{m_i a_{k_i}}^2}\right)$$

where k_i is the maximand in (37). When a is unidimensional,

$$m_i = m_i^d + \tau\left(\mu_i - m_i^d, \kappa^m \frac{\|\eta_a\| V_{aa}}{V_{m_i a}}\right). \tag{39}$$

The intuition is as follows. For each dimension m_i, select the "key action" that is related to it. That is the one with the maximum $\frac{-V_{m_i a_k}^2}{V_{a_k a_k}}$, in virtue of Footnote 7. The term Λ^{diag} is simple to calculate, and does not involve the matrix inversion of the general Λ in (7).

For instance, take the consumption example of Section 4.2. We have $\frac{-V_{m_i c_k}^2}{V_{c_k c_k}} = -\frac{\lambda^2}{V_{c_k c_k}}$

if $i = k$, and 0 otherwise. Hence, the key action corresponding to the price m_i is the consumption of the good c_i. Therefore, $\Lambda^{diag} = \lambda^2 diag\left(-1/u_{c_i c_i}\right)$, which is simple to use. Without the key action, the cross partials $u_{c_i c_j}$ matterℓ and things are more complex to derive for the paper-and-pencil economist, and also perhaps for the decision maker.

6.2 Links with Themes of the Literature

6.2.1 Links with Themes in Behavioral Economics

In this section, I discuss the ways in which the Sparse BR approach meshes with themes in behavioral economics: it draws from them, and is a framework to think about them.

Anchoring and adjustment The model exactly features anchoring and adjustment for expectations and decisions: the anchor is the default model-in-model m^d and action a^d, the adjustment is dictated by the circumstances. In this way, the model is a complement to other models; for instance, Gennaioli and Shleifer (2010) model what "comes to mind" to the decision maker, so that their work is a model of m^d, while the present model is about how the decision maker deviates from that simplified model m^d.

Power of defaults Closely related to anchoring and adjustment, it has now been well established that default actions are very often followed even in the field (Madrian and Shea 2001, Carroll et al. 2009). This model prominently features that stylized fact.

Rules of thumb Rules of thumb are rough guides to behavior, such as "invest 50/50 in stocks and bonds," "save 15% of your income," or "consume the dividend but not the principal" (Baker, Nagel, and Wurgler 2007). They are easily modeled as default actions a^d. The advantage is that the Sparse BR model generates deviations from the rule (the default action) when the circumstances call for it with enough force: for instance, if income is very low, the agent will see that the current marginal utility is very high, and he should save less.

Temptation vs BR The present model is about bounded rationality, rather than "emotions" such as hyperbolic discounting (Laibson 1997) or temptation. Following various authors (e.g., Fudenberg and Levine 2006, Brocas and Carillo 2008), we can imagine an interesting connection, though, operationalized via defaults. Suppose that System 1 (Kahneman 2003), the emotional and automatic system, wants to consume now. This could be modeled as saying that System 1 resets the default action to high consumption now (it will likely also shift the default representation). System 2, the cold analytical system, operates like the Sparse BR model. It partially overrides the default when cognition costs are low,

but will tend to follow it otherwise. While many papers have focused on modeling System 1, this paper attempts to model System 2.

Mental accounts Some of the above has a flavor of "mental accounts" (Thaler 1985). For instance, in Section 4.1, the marginal propensity to consume out of income is source-dependent.

Availability The theory is silent about the cost κ_i^m of each dimension, which is constant at κ^m in the benchmark model. It is, however, plausible that more "available" dimensions will have a lower κ_i^m. For instance, availability is greater when a variable is large, familiar, and frequently used.

Endowment effect The model generates an endowment effect (cf. Section 4.5), with the additional feature (compared to the common explanation based on prospect theory) that the better understood the good the lower the endowment effect.

$1/n$ **heuristics** This heuristic (Bernatzi and Thaler 2001, Huberman and Jian 2006) is to allocate an amount $1/n$ when choosing over n plans, irrespective of the plans' correlation: for instance, the agent allocates $1/3, 1/3, 1/3$, no matter whether the offering is one bond fund and two stock funds or one stock fund and two bond funds. The model can generate this by using the "simplification of variables" (cf. Section 5.2). Here, the simplification would be that the variables are treated as independent (or i.i.d.) rather than correlated.

6.2.2 Links with Other Approaches to Bounded Rationality and Inattention

This paper is another line of attack on the polymorphous problem of bounded rationality (see surveys in Conslik 1996 and Rubinstein 1998). The present paper is best viewed as a complement rather than a substitute for existing models. For instance, there is a vast literature on learning (Sargent 1993, Fudenberg and Levine 2009) that sometimes generates a host of stylized facts because agents may not set up their models optimally (Fuster, Laibson, and Mendel 2010). One could imagine joining those literatures in a model of "sparse learning" where the agent pays attention only to a subset of the world and thus perhaps learns only partially about the world.

This said, some of the most active themes are the following.[20]

[20] I omit many models here, in particular "process" models, e.g., Bolton and Faure-Grimaud (2009), Compte and Postlewaite (2011), Gabaix, Laibson, Moloche, and Weinberg (2006), and MacLeod (2002). They are instructive conceptually and descriptively, but yield somewhat complex mappings between situations and outcomes.

Limited understanding of strategic interactions. In several types of models, the BR comes from the interactions between the decision maker and other players, see Eyster and Rabin's (2005) cursed equilibrium, Jéhiel's (2005) analogy-based equilibrium, k-level of thinking models surveyed in Crawford, Costa-Gomes, and Iriberri (2010), and the related work of Camerer, Ho, and Chong (2004).[21] These models prove very useful for capturing naiveté about strategic interactions, e.g., the winner's curse in auctions or beauty contests. However, they can only be one part (albeit an important one) of the problem of BR: indeed, in a single-person context, they model the decision maker as fully rational. In contrast, in the present paper the decision maker is boundedly rational even in isolation.

Decisions within ε of the maximal utility. The near-rational approach of Akerlof and Yellen (1985) is based on the premise that agents will tolerate decisions that make them lose some ε utility, and still proves useful for empirical work (e.g., Chetty 2009). However, it implies that the decision maker's action will fall in a band (that gives him close to maximal utility), but it does not yield definite predictions about what actions the decision maker will take. This is in contrast to this paper.

Inattention and information acquisition. This paper is also related to the literature on modeling inattention (see Veldkamp 2011 for a comprehensive survey). There are several ways to model inattention. One strand of that literature uses fixed costs (Duffie and Sun 1990, Gabaix and Laibson 2002, Mankiw and Reis 2002, Reis 2006). I have argued that a key benefit of this paper's approach, with the ℓ_1 penalty, is the tractability it confers. Another influential proposal made by Sims (e.g., in 2003) is to use an entropy-based penalty. This has the advantage of a nice foundation; however, it leads to non-deterministic models (agents take stochastic decisions), and the modeling is very complex when it goes beyond the linear-Gaussian case. The Sparse BR model presents some important differences. One is that the model generates sparsity. Another is that the model is deterministic: in a given situation, ex-ante identical agents remain identical ex post. This makes the analysis much simpler.

Uncertainty aversion and concern for robustness. Hansen and Sargent (2007) have shown that many consequences (e.g., prudent allocation to stocks) stem from the assumption that the agents understand that they do not know the right model and have concerns for "robustness," which they model as optimization under the worse potential model. In contrast, the decision maker is biased towards simplicity, not pessimism, in the present model.

It may be interesting to note that while all those frameworks are inspired by psychology,

[21] See also models of naive hyperbolic discounting (O'Donoghue and Rabin 1999, DellaVigna and Malmendier 2004). Relatedly, an interesting literature studies BR in organizations (e.g., Radner and Van Zandt 2001), and aims at predictions on the level of large organizations rather than individual decision making. See also Madarász and Prat (2010) for a recent interesting advance in BR in a strategic context.

some are also inspired by modeling advances in applied mathematics. The Sims framework is based on Shannon's information theory of the 1940s. The Hansen-Sargent framework is influenced by the engineering literature of the 1970s. The present framework is inspired by the sparsity-based literature of the 1990s-2000s (Tibshirani 1996, Candès and Tao 2006, Donoho 2006, Mallat 2009), which shows that sparsity has many properties of tractability and near-optimality.[22] The present paper is the first paper in economic theory to use the recent sparsity-based literature from statistics.

7 Conclusion

This paper proposes a tractable model with some boundedly rational features. Its key contribution is to formulate a tractable version of the costs and benefits of thinking (captured as an enrichment of the agent's mental model). On the benefit side, the decision maker uses a quadratic approximation of his utility function, which circumvents Simon's infinite-regress problem. On the cost side, the decision maker uses an ℓ_1 norm to obtain sparsity and tractability (drawn from a recent literature in applied mathematics), with feature-specific weights that make the model largely invariant to many changes in scales, units, and reparametrization. This formulation leads to linear-quadratic problems (with a sparsity-inducing absolute value) which are easy to solve in many cases of interest. At the same time, it arguably features some psychological realism: we all simplify reality when thinking about it, and this model represents one way to do that – indeed, it is the simplest tractable way that I could devise.

The simplicity of the core model allows for the formulation of a BR version of a few important building blocks of economics. For instance, we can study BR-optimal choice of consumption bundles (the agent has an imperfect understanding of prices); and BR asset allocation (with inertia and trading freezes). The model leads to a theory of price rigidity based on stickiness in the consumer's mind, rather than stickiness in the price-setting technology of firms. In ongoing work, I formulate a way to do BR dynamic programming, where the agent builds on a simplified model with few state variables.

No doubt, the model could and should be greatly enriched. In the present work, there is simply a lone agent. In work in progress, I extend the model to include multi-agent models and the limited understanding of general equilibrium effects by agents. In addition, the model is silent about some difficult operations such as Bayesian (or non-Bayesian) updating

[22] For instance, somewhat miraculously, one can do a regression with fewer observations than regressors (like in genetics, or perhaps growth empirics) by assuming that the number of non-zero regressors is sparse and using an ℓ_1 penalty for sparsity (see Belloni and Chernozhukov 2010), as in $\min_\beta \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\beta' x_i\right)^2+\lambda\left\|\beta\right\|_1$.

and learning (see Gennaioli and Shleifer 2010 for recent progress in that direction), and memory management (Mullainathan 2002).

Indeed, the model is a complement rather than a substitute for other models: one could as well devise a model of BR learning or robustness with a sparsity constraint. Those extensions are left to future research (e.g., Gabaix 2011). However, despite these current limitations, given its tractability and fairly good generality, the Sparse BR model might be a useful tool for thinking about the impact of bounded rationality in economic situations.

8 Proof Appendix

Proof of Proposition 1. We need the following lemmas. Here, n and p are positive integers, and S is a set.

Lemma 2 *(a) Consider a function $f : \mathbb{R}^n \times \mathbb{R}^{n\times p} \to S$ such that for all $x \in \mathbb{R}^n$, $y \in \mathbb{R}^{n\times p}$, and $A \in \mathbb{R}^{n\times n}$, $f(Ax, y) = f(x, A'y)$. Then, there exists a function $g : \mathbb{R}^p \to S$ such that $f(x, y) = g(x'y)$. (b) Consider a function $f : \mathbb{R}^n \times \mathbb{R}^{n\times p} \times \mathbb{R}^{n\times n} \to S$ such that for all $x \in \mathbb{R}^n$, $y \in \mathbb{R}^{n\times p}, z \in \mathbb{R}^{n\times n}$, and $A \in \mathbb{R}^{n\times n}$, $f(Ax, y, z) = f(x, A'y, x, A'zA)$. Then, there exists a function $g : \mathbb{R}^p \times \mathbb{R} \to S$ such that $f(x, y, z) = g(x'y, x'zx)$.*

Proof . Let us prove (b), which is more general than (a). Define $e_1 = (1, 0_{n-1})'$ and, for a row vector $Y \in \mathbb{R}^p$ and a scalar $Z \in \mathbb{R}$, $g(Y, Z) := f(e_1, e_1 Y, e_1 Z e_1')$. We have:

$$\begin{aligned} f(x, y, z) &= f(xe_1'e_1, y, z) \text{ as } e_1'e_1 = 1 \\ &= f(e_1, e_1x'y, e_1x'zxe_1') \text{ using the assumption with } A = xe_1' \\ &= g(x'y, x'zx). \end{aligned}$$

□

Hypothesis (ii) implies that K is independent of the values V and V_m evaluated at the default. Hence, one can write $K((m_i)_{i=1...n}, \eta_a, (V_{am_i})_{i=1...n}, V_{aa})$ for some function K (by a minor abuse of notation).

We use the invariance to reparametrization λ_1 in hypothesis (i), and apply Lemma 2(a) to $K(m_1, V_{am_1}, Z_1)$ where Z_1 represents the other arguments. This implies that we can write $K(m_1, V_{am_1}, Z_1) = K(m_1 V_{am_1}, Z_1)$ for a new function K. Proceeding the same way for (m_i, V_{am_i}) for $i = 2...n$, we see that we can write $K = K(\eta_a, (m_i V_{am_i})_{i=1...n}, V_{aa})$. We next apply Lemma 2(b) to $x = \eta_a$ and $y = (m_i V_{am_i})_{i=1...n_m}, z = V_{aa}$. It implies that we can write:

$$K = k\left((\eta_a \cdot V_{am_i} m_i)_{i=1...n_m}, \eta_a' V_{aa} \eta_a\right) \tag{40}$$

for some function $k : \mathbb{R}^{n_m} \times \mathbb{R} \to \mathbb{R}$.

Let us next use assumption (iii). When $\|x_i\|$ and $\|\eta_{a_1}\|$ are non-zero, define $\widehat{x}_i = x_i / \|x_i\|$, $\widehat{a_1} = a_1 / \|\eta_{a_1}\|$, $\widehat{\eta}_{a_1} = \eta_{a_1} / \|\eta_{a_1}\|$, and $\widehat{m}_i = m_i \|x_i\| \|\eta_{a_1}\|$. Then, the problem associated with $(\widehat{a}_1, \eta_{\widehat{a}_1}, \widehat{x}_i, \widehat{m}_i)$ has $\|\widehat{x}_i\| = 1$ and $\|\eta_{\widehat{a}_1}\| = 1$. Hypothesis (iii) indicates that $k\left((\widehat{\eta}_{a_1} \widehat{x}_i \widehat{m}_i)_{i=1...n}\right) = \kappa^m \sum_i |\widehat{m}_i|$. Hence:

$$k\left((\eta_{a_1} x_i m_i)_{i=1...n}, 1\right) = k\left((\widehat{\eta}_{a_1} \widehat{x}_i \widehat{m}_i)_{i=1...n}\right) = \kappa^m \sum_i |\widehat{m}_i| = \kappa^m \sum_i \|\eta_a x_i m_i\|.$$

This implies that $k\left((y)_{i=1...n}, 1\right) = \kappa^m \sum_i \|y_i\|$. Using the homogeneity of degree 1 part of (ii), we have that

$$k\left((y)_{i=1...n}, z\right) = \kappa^m \sum_i \|y_i\|.$$

Using (40), we have:

$$K = \kappa^m \sum_i \|\eta_a x_i m_i\| = \sum_i \left\| \left(m_i - m_i^d\right) V_{m_i a} \cdot \eta_a \right\| = \sum_i |m_i| \left\| V_{m_i a} \cdot \eta_a \right\|. \blacksquare$$

Proof of Proposition 2 By homogeneity, it is enough to consider the case $s = 1$.

Step 1: Representation. We calculate

$$V_a = -a + m \cdot x, \qquad V_m = x\left(a - m \cdot x\right)$$
$$V_{aa} = -1, \qquad V_{am} = x, \qquad V_{am_i} = x_i$$

so when x is one-dimensional, $\Lambda = -\mathbb{E}\left[V_{am} V_{aa}^{-1} V_{am}\right] = \mathbb{E}\left[x^2\right]$. With n dimensions for m, drawn independently, we have by the same calculation that $\Lambda = Diag\left(\sigma_i^2\right)$.

Thus,

$$\kappa\left[m\right] = \kappa^m \sum_i \left|m_i - m_i^d\right| \left\|V_{m_i a} \eta_a\right\| = \kappa^m \sum_i \left|m_i - m_i^d\right| \|x_i\| \|\eta_a\| = \sum_i K_i \left|m_i - m_i^d\right|$$

with $K_i \equiv \kappa^m \sigma_a \sigma_i$.

So, the maximization (8) is

$$\max_m - \sum_i \frac{1}{2} \sigma_i^2 \left(m_i - \mu_i\right)^2 - \sum_i K_i \left|m_i - m_i^d\right|.$$

We use Lemma 1, which gives (16).

Step 2: Approximate maximization. We calculate $\kappa[a]$ from equation (11):

$$\kappa[a] = \kappa^a \left|a - a^d\right| \|V_{am}\eta_m\| = \kappa^a \left|a - a^d\right| \|x \cdot \eta_m\| = \kappa^a \left|a - a^d\right| \sqrt{\sum_i \sigma_{m_i}^2 \sigma_i^2} \equiv Q \left|a - a^d\right|.$$

Step 2 gives: $\max_a -\frac{1}{2}(a - m \cdot x)^2 - Q\left|a - a^d\right|$. This yields $a = a^d + \tau\left(m \cdot x - a^d, Q\right)$.

■

Proof of Proposition 3. We calculate:

$$V_{c_1} = u'(c_1) - v'(c_2) R, \qquad V_{c_1c_1} = u''(c_1) + v''(c_2) R^2, \qquad V_{c_1m_i} = -v''(c_2) R x_i.$$

Let us proceed with the part related to future income. The program (8) is:

$$\min_{m_i} \sum_i \frac{v''\left(c_2^d\right)^2 R^2}{-2V_{c_1c_1}^d} \sigma_{x_i}^2 (m_i - 1)^2 + \kappa^m \sum_i |m_i| \left|v''\left(c_2^d\right) R \sigma_{c_1}\right| \sigma_{x_i}$$

where $V_{c_1c_1}^d$ is a shorthand for $V_{c_1c_2}\left(c_1^d, c_2^d\right)$. Hence, the solution is:

$$m_i = \tau(1, \kappa_i) \qquad \kappa_i = \kappa^m \frac{V_{c_1c_1}^d}{v''\left(c_2^d\right) R} \frac{\sigma_{c_1}}{\sigma_{x_i}}. \tag{41}$$

We now use the functional form $u(c) = -e^{-\gamma c}$ and $v(c) = -e^{-\rho} e^{-\gamma c}$. Because under the default $u'\left(c_1^d\right) - v'\left(c_2^d\right) R = 0$, the exponential specification gives $u''\left(c_1^d\right) = v''\left(c_2^d\right) R$, and we have $V_{c_1c_1}^d = u''\left(c_1^d\right)(1 + R)$, so (41) gives:

$$\kappa_i = \frac{\kappa^m (1 + R) \sigma_{c_1}}{\sigma_{x_i}}.$$

In Step 2, the agent solves

$$\max_{c_1} -e^{-\gamma c_1} - e^{-\rho} e^{-\gamma(x(m) + R(w - c_1))}$$

which gives $e^{-\gamma c_1} - e^{-\rho} e^{-\gamma(x(m) + R(w - c_1))} R = 0$ and

$$\begin{aligned} c_1 &= x(m) + R(w - c_1) + \frac{\rho - \ln R}{\gamma} \\ &= x(m) - Rc_1 + D \text{ with } D = Rw + \frac{\rho - \ln R}{\gamma} \end{aligned}$$

as well as

$$\begin{aligned} c_1 &= \frac{1}{1+R}\left(D + x\left(m\right)\right) \\ &= \frac{1}{1+R}\left(D + x_* + \sum_i \tau\left(1, \kappa^m\left(1+R\right)\frac{\sigma_{c_1}}{\sigma_{x_i}}\right)x_i\right). \end{aligned}$$

Proof of Proposition 4 We calculate:

$$V_{c_i} = u_i - \lambda\left(p_i^d + m_i\right), \qquad V_{c_i c_j} = u_{ij}, \qquad V_{c_i m_j} = -\lambda 1_{i=j}.$$

Hence, the components of the loss matrix are $\Lambda_{ii} = \frac{\lambda^2}{-u_{ii}}$ in two cases: namely, if the utility function is separable in the goods ($u\left(c\right) = \sum_i u^i\left(c_i\right)$) or, for a non-separable utility function, if we apply the "key action" enrichment developed below in Section 6.1.4 (the key action corresponding to p_i is c_i).

Calling $\sigma_{c_i} = \|\eta_{c_i}\|$, the allocation of attention is:

$$\min_m \sum_i \left[\frac{\lambda^2\left(m_i - \mu_i\right)^2}{2\left|u_{ii}\right|} + \kappa^m \lambda\left|m_i\right|\sigma_{c_i}\right],$$

so we have

$$m_i = \tau\left(\mu_i, \frac{\kappa^m\left|u_{ii}\right|\sigma_{c_i}}{\lambda}\right) = \tau\left(\mu_i, \frac{\kappa^m u_i\left|\frac{c_i u_{ii}}{u_i}\right|\frac{\sigma_{c_i}}{c_i}}{\lambda}\right).$$

Using $u_i = \lambda p_i$ and calling $\psi_i = u_i/\left(-c_i u_{ii}\right)$ the price elasticity of demand of good i, we obtain (20).

To proceed further, we examine the case where preferences are separable, so the f.o.c. $u_i\left(c_i\right) = \lambda p_i$ implies that a change in price dp_i implies $u_{ii}dc_i = \lambda dp_i$, and thus $\left|u_{ii}\right|\sigma_{c_i} = \lambda\sigma_{p_i}$. Equation (21) follows.

Proof of Proposition 5. The monopolist solves

$$\max_p \pi\left(p\right), \qquad \pi\left(p\right) = \left(p - c\right)\left(p^d + \tau\left(p - p^d, \kappa\right)\right)^{-\psi}.$$

Consider first the interior solutions with $p \notin \left(p^d - \kappa, p^d + \kappa\right)$. Call $\varepsilon = sign\left(p - p_d\right)$. Then, $p^d + \tau\left(p - p^d, \kappa\right) = p - \varepsilon\kappa$ (equation 5). Therefore, $\partial_p \tau\left(p - p^d, \kappa\right) = 1$, and the f.o.c. is $p - \varepsilon\kappa - \psi\left(p - c\right) = 0$, i.e.,

$$p = p^{int} \equiv \frac{\psi c - \varepsilon\kappa}{\psi - 1}. \tag{42}$$

The profit is

$$\pi\left(p^{int}\right)=\left(\frac{\psi c-\varepsilon\kappa}{\psi-1}-c\right)\left(\frac{\psi c-\varepsilon\kappa}{\psi-1}-\varepsilon\kappa\right)^{-\psi}=\psi^{-\psi}\left(\frac{(c-\varepsilon\kappa)}{\psi-1}\right)^{1-\psi}.$$

Next, it is not optimal for the monopolist to have $p\in\left(p^d-\kappa,p^d+\kappa\right)$ as $p=p^d+\kappa$ yields the same demand and strictly higher profits. The profit is

$$\pi\left(p^d+\kappa\right)=\left(p^d+\kappa-c\right)\left(p^d\right)^{-\psi}.$$

It is optimal to choose p^{int} rather than $p^d+\kappa$ iff $R\geq 1$ where

$$\begin{aligned}R\left(c,c^d,\kappa\right)&=\frac{\pi\left(p^{int}\right)}{\pi\left(p^d+\kappa\right)}=\frac{\psi^{-\psi}\left(\frac{(c-\varepsilon\kappa)}{\psi-1}\right)^{1-\psi}}{\left(\frac{\psi}{\psi-1}c^d+\kappa-c\right)\left(\frac{\psi}{\psi-1}c^d\right)^{-\psi}}\\&=\frac{(c-\varepsilon\kappa)^{1-\psi}}{\left[\psi c^d+(\psi-1)(\kappa-c)\right](c^d)^{-\psi}}.\end{aligned}$$

The cutoffs c_1 and c_2 are the solution to $R\left(c_i,c^d,\kappa\right)=1$. The c_2 bound is easy to find because it is clear (as the profit function is increasing for $p<p^{int}$) that c_2 must be such that $p^{int}(c_2)=p^d+\kappa$, i.e., $\frac{\psi c_2-\kappa}{\psi-1}=\frac{\psi c^d}{\psi-1}+\kappa$, so $c_2=c^d+\kappa$. The more involved case is the one where $c<c^d$ as then there can be two local maxima (this is possible as the demand function is not log-concave). Hence, the cutoff c_1 satisfies, with $\varepsilon=-1$,

$$R\left(c_1,c^d,\kappa\right)=1 \tag{43}$$

and $c_1<c_d$. To obtain an approximate value of c_1, note that $R(c,c,0)=1$: when $\kappa=0$, the cutoff corresponds to $c=c^d$. Also, calculations show $R_1(c,c,0)=0$ and $R_{11}(c,c,0)\neq 0$. Hence, a small κ implies a change δc_1 such that, to the leading order, $\frac{1}{2}R_{11}\cdot(\delta c)^2+R_3\cdot\kappa=0$, i.e., $c_1=c^d-\sqrt{\frac{-2R_3\kappa}{R_{11}}}++O(\kappa)$. Calculations yield $c_1=c^d-2\sqrt{c^d\kappa/\psi}+O(\kappa)$. ■

Proof of Proposition 6 It is clear that the optimal solution a belongs to $\{\mathbb{E}[Y]-m,\mathbb{E}[Y]+m\}$. If the offer is $a=\mathbb{E}[Y]-m$, the offer is accepted only if $X=\mathbb{E}[Y]-m$ (in the model-in-model), so:

$$V^{M2}\left(\mathbb{E}[Y]-m\right)=\frac{1}{2}\left(\lambda\left(\mathbb{E}[Y]-m\right)-\left(\mathbb{E}[Y]-m\right)\right)=\frac{\lambda-1}{2}\left(\mathbb{E}[Y]-m\right).$$

If the offer is $a = \mathbb{E}[Y] + m$, the buyer gets the firm for sure, which has a value to him of $\lambda\mathbb{E}[Y]$ in expectation, so:

$$V^{M2}(\mathbb{E}[Y] + m) = \lambda\mathbb{E}[Y] - (\mathbb{E}[Y] + m) = (\lambda - 1)\mathbb{E}[Y] - m.$$

Note that $V^{M2}(\mathbb{E}[Y] + m) > V^{M2}(\mathbb{E}[Y] - m)$ if $\lambda \geq 3$. Once we have $\lambda < 3$, the two profits $V^{M2}(\mathbb{E}[Y] - m)$ and $V^{M2}(\mathbb{E}[Y] + m)$ are the same if and only if $m = \frac{\lambda-1}{3-\lambda}\mathbb{E}[Y]$. Thus, the optimal decision is as announced in the proposition. The maximum paid is $\mathbb{E}[Y] + \frac{\lambda-1}{3-\lambda}\mathbb{E}[Y] = \frac{2}{3-\lambda}\mathbb{E}[Y]$.

References

Abel, Andrew, Janice C. Eberly, and Stavros Panageas, "Optimal Inattention to the Stock Market with Information Costs and Transactions Costs," Working Paper, University of Pennsylvania, 2010.

Akerlof, George A., and Janet L. Yellen, "Can Small Deviations from Rationality Make Significant Differences in Economic Equilibria?" *American Economic Review*, 75 (1985), 708-720.

Aragones, Enriqueta, Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler, "Fact-Free Learning," *American Economic Review*, 95 (2005), 1355-68.

Baker, Malcolm, Stefan Nagel and Jeffrey Wurgler, "The Effect of Dividends on Consumption," *Brookings Papers on Economic Activity*, 38 (2007), 231-292.

Banerjee, Abhijit, and Sendhil Mullainathan, "The Shape of Temptation: Implications for the Economic Lives of the Poor," Working Paper, MIT, 2010.

Belloni, Alexandre and Victor Chernozhukov, "High Dimensional Sparse Econometric Models: An Introduction", Working Paper, MIT, 2010.

Bernatzi, Shlomo, and Richard Thaler, "Naive Diversification Strategies in Defined Contribution Saving Plans," *American Economic Review,* 91 (2001), 79-98.

Bernheim, B. Douglas, and Antonio Rangel, "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *Quarterly Journal of Economics,* 124 (2009), 51-104.

Bolton, Patrick and Antoine Faure-Grimaud, "Thinking Ahead: The Decision Problem" *Review of Economic Studies* (2009), 76(4): 1205-1238

Brocas, Isabelle, and Juan Carillo, "The Brain as a Hierarchical Organization," *American Economic Review*, 98 (2008), 1312–1346.

Caballero, Ricardo, and Arvind Krishnamurthy, "Collective Risk Management in a Flight

to Quality Episode," *Journal of Finance,* 63 (2008), 2195–2230.

Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong, "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, 119 (2004), 861-898.

Candès, Emmanuel, and Terence Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Transactions on Information Theory*, 52 (2006), 5406-5425.

Cao, H. Henry, Bing Han, David Hirshleifer and Harold H. Zhang, "Fear of the Unknown: Familiarity and Economic Decisions," *Review of Finance* 15 (2011), 173-206.

Carroll, Gabriel D., James Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick, "Optimal Defaults and Active Decisions," *Quarterly Journal of Economics*, 124 (2009), 1639-1674.

Charness, Gary, and Dan Levin, "The Origin of the Winner's Curse: A Laboratory Study," *American Economic Journal: Microeconomics*, 1 (2009), 207–236

Chetty, Raj, "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply, "NBER Working Paper # 15616, 2009.

Chetty, Raj, Adam Looney, and Kory Kroft, "Salience and Taxation: Theory and Evidence," *American Economic Review,* 99 (2009), 1145-1177.

Chevalier, Judith and Anil Kashyap, "Best Prices," Yale Working Paper, 2011.

Compte, Olivier and Andrew Postlewaite, "Mental Processes and Decision Making," PSE Working Paper, 2011.

Conlisk, John, "Why Bounded Rationality?" *Journal of Economic Literature*, 34 (1996), 669-700.

Crawford, Vincent, Miguel Costa-Gomes and Nagore Iriberri, "Strategic Thinking," Working Paper, Oxford, 2010.

Crawford, Vincent, and Nagore Iriberri, "Level-k Auctions: Can Boundedly Rational Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica,* 75 (2007), 1721-1770.

Currie, Janet, "Healthy, Wealthy, and Wise? Socioeconomic Status, Poor Health in Childhood, and Human Capital Development," *Journal of Economic Literature*, 47 (2009), 87-122.

DellaVigna, Stefano, and Ulrike Malmendier, "Contract Design and Self-Control: Theory and Evidence," *Quarterly Journal of Economics*, 119 (2004), 353-402.

Donoho, David, "Compressed Sensing," *IEEE Transactions on Information Theory*, 52 (2006), 1289-1306.

Duffie, Darell, and Tong-sheng Sun, "Transactions costs and portfolio choice in a discrete continuous time setting," *Journal of Economic Dynamics & Control* 14 (1990), 35-51.

Eichenbaum, Martin, Nir Jaimovich, and Sergio Rebelo, "Reference Prices and Nominal Rigidities," *American Economic Review,* forthcoming.

Esponda, Ignacio, "Behavioral Equilibrium in Economies with Adverse Selection," *American Economic Review*, 98 (2008), 1269-91.

Eyster, Erik, and Matthew Rabin, "Cursed Equilibrium," *Econometrica*, 73 (2005), 1623-1672.

Eyster, Erik, and Georg Weizsäcker, "Correlation Neglect in Financial Decision-Making," Working Paper, London School of Economics, 2010.

Fryer, Roland, and Matthew O. Jackson, "A Categorical Model of Cognition and Biased Decision Making," *The BE Journal of Theoretical Economics,* 8 (2008), Article 6.

Fudenberg, Drew, and David Levine, "Learning and Equilibrium," *Annual Review of Economics,* 1 (2009), 385-420.

Fudenberg, Drew, and David Levine, "A Dual-Self Model of Impulse Control," *American Economic Review*, 96 (2006), 1449-1476.

Fuster, Andreas, David Laibson, and Brock Mendel, "Natural Expectations and Macroeconomic Fluctuations," *Journal of Economic Perspectives*, 24(2010), 67-84.

Gabaix, Xavier, "Game Theory with Sparsity-Based Bounded Rationality," Working Paper, NYU, 2011.

Gabaix, Xavier, and David Laibson, "The $6D$ bias and the Equity Premium Puzzle," *NBER Macroeconomics Annual*, 16 (2002), 257-312.

Gabaix, Xavier, David Laibson, Guillermo Moloche, and Stephen Weinberg, "Costly Information Acquisition: Experimental Analysis of a Boundedly Rational Model," *American Economic Review,* 96 (2006), 1043-1068.

Gennaioli, Nicola, and Andrei Shleifer, "What Comes to Mind," *Quarterly Journal of Economics*, 125 (2010), 1399-1433.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny, "Financial Innovation and Financial Fragility," Working Paper, Harvard University, 2010.

Gilboa, Itzhak and David Schmeidler, "Maxmin expected utility theory with non-unique prior," *Journal of Mathematical Economics* 18 (1989), 141–153

Goldberg, Pinelopi and Rebecca Hellerstein, "A Structural Approach to Identifying the Sources of Local-Currency Price Stability," Working Paper, Yale, 2010.

Hansen, Lars, and Thomas Sargent, *Robustness* (Princeton: Princeton University Press, 2007).

Heidhues, Paul and Botond Koszegi, "Regular Prices and Sales," Working Paper, Berkeley, 2010.

Hirshleifer, David, Sonya Lim and Siew Teoh, "Driven to Distraction: Extraneous Events

and Underreaction to Earnings News," *Journal of Finance*, 64 (2009), 2289-2325

Huberman, Gur, and Wei Jiang, "Offering versus Choice in 401(k) Plans: Equity Exposure and Number of Funds," *Journal of Finance*, 61 (2006), 763–801.

Jéhiel, Philippe, "Analogy-Based Expectation Equilibrium," *Journal of Economic Theory,* 123 (2005), 81–10.

Kahneman, Daniel, "Maps of Bounded Rationality: Psychology for Behavioral Economics," *American Economic Review*, 93 (2003), 1449-1475.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler, "Experimental Tests of the Endowment Effect and the Coase Theorem," *Journal of Political Economy*, 97 (1990), 1325-1348.

Kehoe, Patrick, and Virgiliu Madrigan, "Prices Are Sticky After All," National Bureau of Economic Research Working Paper, 2010.

Klenow, Peter and Benjamin Malin, "Microeconomic Evidence on Price-Setting," forth. *Handbook of Monetary Economics.*

Koszegi, Botond, and Matthew Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, 121 (2006), 1133-1166.

Laibson, David, "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics* 62 (1997), 443-77.

L'Huillier, Jean-Paul, "Consumers' Imperfect Information and Nominal Rigidities," Working Paper, Massachusetts Institute of Technology, 2010.

Liersch, Michael, Yuval Rottenstreich, Howard Kunreuther, and Min Gong, "Reference-Dependent versus Connection-Based Accounts of the Endowment Effect" Working Paper, NYU, 2011.

List, John A., "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics*, 118 (2003), 41-71.

MacLeod, Bentley. "Complexity, Bounded Rationality and Heuristic Search," *Contributions to Economic Analysis & Policy* Vol 1, No. 1, Article 1, 2002.

Madrian, Brigitte C., and Dennis Shea, "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior," *Quarterly Journal of Economics*, 116 (2001), 1149-1187.

Mankiw, N. Gregory, and Ricardo Reis, "Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve," *Quarterly Journal of Economics*, 117 (2002), 1295-1328.

Madarász, Kristóf, and Andrea Prat, "Screening with an Approximate Type Space," Working Paper, London School of Economics, 2010.

Mallat, Stéphane, *A Wavelet Tour of Signal Processing: The Sparse Way,* third edition,

(New York: Academic Press, 2009).

Matejka, Filip, "Rigid Pricing and Rationally Inattentive Consumer," Working Paper, Princeton University, 2010.

Mullainathan, Sendhil, "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 117 (2002), 735-774.

Mullainathan, Sendhil, "Thinking through categories," Working Paper, MIT, 2001.

O'Donoghue, Ted, and Matthew Rabin, "Doing It Now or Later," *American Economic Review*, 89(1999), 103-124.

Rabin, Matthew and Georg Weizsäcker, "Narrow Bracketing and Dominated Choices," *American Economic Review*, 99 (2009), 1508-1543

Radner, Roy, and Timothy Van Zandt, "Real-Time Decentralized Information Processing and Returns to Scale," *Economic Theory,* 17 (2001), 545-575.

Reis, Ricardo, "Inattentive Consumers," *Journal of Monetary Economics*, 53 (2006), 1761-1800.

Rosenthal, Robert W., "Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox," *Journal of Economic Theory*, 25 (1981), 92-100.

Rubinstein, Ariel, *Modeling Bounded Rationality* (Cambridge: MIT Press, 1998).

Samuelson, William F., and Max H. Bazerman, "The Winner's Curse in Bilateral Negotiations," in *Research in Experimental Economics: A Research Annual vol 3*, Vernon Smith, ed. (Greenwich: JAI Press, 1985).

Samuelson, William F., and Richard Zeckhauser, "Status Quo Bias in Decision Making," *Journal of Risk and Uncertainty*, 1 (1988), 7-59.

Sargent, Thomas, *Bounded Rationality in Macroeconomics* (Oxford: Oxford University Press, 1993).

Shubik, Martin, "The Dollar Auction Game: A Paradox in Noncooperative Behavior and Escalation," *The Journal of Conflict Resolution*, 15 (1971), 109-111.

Sims, Christopher, "Implications of Rational Inattention," *Journal of Monetary Economics*, 50 (2003), 665–690.

Thaler, Richard H., "Mental Accounting and Consumer Choice," *Marketing Science*, 4 (1985), 199-214.

Tibshirani, Robert, "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society B*, 58 (1996), 267–288.

Tversky, Amos, and Daniel Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, 185 (1974), 1124–1130.

Tversky, Amos, and Daniel Kahneman. "The Framing of Decisions and the Psychology of Choice," *Science*, 211 (1981), 453-458.

Veldkamp, Laura, *Information Choice in Macroeconomics and Finance* (Princeton: Princeton University Press, 2011).

Online Appendix for "A Sparsity-Based Model of Bounded Rationality," Not for Publication

March 27, 2011

This appendix presents additional derivations and some reasonable variants of the model.

9 Additional Derivations

9.1 Derivation of the Fixed Point p^d in the Monopoly Pricing Model of Section 4.3

In the small κ limit, I solve for the default price p^d, which is the fixed point $p^d = \mathbb{E}\left[p\left(\widetilde{c}, p^d\right)\right]$. The $p(\cdot)$ function is given by Proposition 5. Call F and $f = F'$ the CDF and PDF of c, and $\overline{c} = \mathbb{E}[c]$. Also, define $A = 2\sqrt{c^d/\psi}$, so that $c_1 = c^d - A\sqrt{\kappa} + O(\kappa)$. When $\kappa = 0$, $c^d = \overline{c}$, so for small κ we look for a solution c^d close to $\overline{c}$. We have:

$$\begin{aligned}
G &\equiv (\psi - 1)\,\mathbb{E}\left[p\left(c, p^d\right)\right] \\
&= \mathbb{E}[\psi c] + \kappa\mathbb{E}\left[1_{c<c_1}\right] - \kappa\mathbb{E}\left[1_{c>c_2}\right] + \mathbb{E}\left[\left((\psi-1)\left(p^d+\kappa\right) - \psi c\right)1_{c\in[c_1,c_2]}\right] \\
&= \psi\overline{c} + \kappa\left(F(c_1) - (1 - F(c_2))\right) + \mathbb{E}\left[\left(\psi c^d + (\psi-1)\kappa - \psi c\right)1_{c\in[c_1,c_2]}\right] \\
&= \psi\overline{c} + \kappa\left(2F(\overline{c}) - 1\right) + o(\kappa) + \psi\mathbb{E}\left[(c_2 - c)\,1_{c\in[c_1,c_2]}\right] - \kappa\mathbb{E}\left[1_{c\in[c_1,c_2]}\right].
\end{aligned}$$

We calculate:

$$\begin{aligned}
\mathbb{E}\left[(c_2 - c)\,1_{c\in[c_1,c_2]}\right] &= \int_{c_1}^{c_2} f(c)(c_2 - c)\,dc = \frac{1}{2}\left(f(\overline{c}) + O\left(\sqrt{\kappa}\right)\right)(c_2 - c_1)^2 \\
&= \frac{1}{2}\left(f(\overline{c}) + O\left(\sqrt{\kappa}\right)\right)\left(\kappa + A\sqrt{\kappa} + O(\kappa)\right)^2 = \frac{1}{2}A^2\kappa f(\overline{c}) + o(\kappa) \\
&= \frac{1}{2}\frac{4c^d}{\psi}f(\overline{c})\,\kappa + o(\kappa) = \frac{2\overline{c}f(\overline{c})}{\psi}\kappa + o(\kappa).
\end{aligned}$$

Given $\kappa\mathbb{E}\left[1_{c\in[c_1,c_2]}\right] = O\left(\kappa^{3/2}\right)$, we have:

$$G = \psi\overline{c} + \left(2\overline{c}f(\overline{c}) + 2F(\overline{c}) - 1\right)\kappa + o(\kappa).$$

Finally,

$$p^d = \frac{G}{\psi - 1} = \frac{\psi}{\psi - 1}\overline{c} + \frac{1}{\psi - 1}\left(2\overline{c}f(\overline{c}) + 2F(\overline{c}) - 1\right)\kappa + o(\kappa). \tag{44}$$

Monopoly pricing model of Section 4.3 with a fixed cost It may be interesting to compare the paper's model to a variant with a fixed cost of cognition. We will see that we maintain the stickiness, but we lose the "sales" effect: the pricing function stops exhibiting the "cliff" at c_1. Instead, it exhibits two symmetrical jumps at c_1 and c_2.

The agent sees the price p. If he pays a fixed cost K, he uses the price p in his decision. If he does not, he simply uses the default price p^d. With a fixed cost, the perceived price is

$$p^{BR}(p) = \begin{cases} p^d & \text{if } \left|p - p^d\right| \leq \kappa \\ p & \text{if } \left|p - p^d\right| > \kappa \end{cases}$$

for a constant κ related to K.[23] The monopolist's problem is $\max_p (p-c) D\left(p^{BR}(p)\right)$, with $D(p) = p^{-\psi}$. Its solution is as follows.

Proposition 9 *When the BR consumer has a fixed cost of cognition, the monopolist's optimal price is*

$$p(c) = \begin{cases} p^d + \kappa & \text{if } c_1 \leq c \leq c_2 \\ \frac{\psi c}{\psi - 1} & \text{if } c \notin (c_1, c_2) \end{cases} \tag{45}$$

where $c_1 < c_2$ solve equation (46), and are equal to $c_i = c^d \pm \sqrt{\frac{2c^d\kappa}{\psi}} + O(\kappa)$ for small κ. The pricing function is discontinuous at c_1 and c_2, and continuous elsewhere.

Proof. (Sketch) The proof is as above. When the consumer is inattentive, $p = p^d + \kappa$, and the profit is $\pi\left(p^d + \kappa\right) = \left(p^d + \kappa - c\right)\left(p^d\right)^{-\psi}$. When the consumer is attentive (and pays the fixed cost), $\pi = \psi^{-\psi}\left(\frac{c}{\psi - 1}\right)^{1-\psi}$ as above. So, the c_i's solve:

$$f(c_i, \kappa) = 0 \tag{46}$$

$$f(c, \kappa) := \frac{(\psi - 1)^{\psi - 1}}{\psi^{\psi}} c^{1-\psi} - \left(p^d + \kappa - c\right)\left(p^d\right)^{-\psi}.$$

For the Taylor expansion, observe that $f\left(c^d, 0\right) = 0$ for $c^d = p^d(\psi - 1)/\psi$ and, by the envelope theorem, $f_1\left(c^d, 0\right) = 0$. So a small κ implies a change δc_1 such that, to the leading order, $\frac{1}{2} f_{11} \cdot (\delta c)^2 + f_3 \cdot \kappa = 0$, i.e., $c_i = c^d \pm \delta c$ with $\delta c = \sqrt{\frac{-2f_3\kappa}{f_{11}}} + O(\kappa)$. Calculations then yield $\delta c = \sqrt{\frac{2c^d\kappa}{\psi}} + O(\kappa)$. ■

[23] The derivation is: the DM picks $\min_{p^{BR}} \frac{\left(p^{BR} - p\right)^2}{-2u_{QQ}} + K 1_{p^{BR} \neq p}$, so we obtain the expression for p^{BR}, with $\kappa = \sqrt{2Kp_d^{1+\psi}/\psi}$. Chetty, Looney, and Kroft (2009) have similar analytics for the DM's decision, but do derive the monopolist's response.

9.2 Derivations of Additional Examples

This section presents the solutions to some of the early examples in the paper.

Example 2. Consider the case $u(c) = -e^{-\gamma c}$, $r \sim N(\pi, \sigma^2)$, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$, and r, ε jointly Gaussian with covariance $\sigma_{r\varepsilon}$. Then, we have:

$$\begin{aligned} V(a,m) &= \mathbb{E}u(w + a\widetilde{r} + m\widetilde{\varepsilon}) \\ &= u\left(w + a\pi - \gamma\left(\frac{a^2\sigma^2 + m^2\sigma_\varepsilon^2}{2} + am\sigma_{r\varepsilon}\right)\right) \\ V_a(a,m) &= -\gamma u\left(w + a\pi - \gamma\left(\frac{a^2\sigma^2 + m^2\sigma_\varepsilon^2}{2} + am\sigma_{r\varepsilon}\right)\right)\left(\pi - \gamma m\sigma_{r\varepsilon} - \gamma a\sigma^2\right). \end{aligned} \tag{47}$$

The default model is $m^d = 0$ (the decision maker does not take into account the background risk). Hence, the default action, which is the optimal action under the default model, satisfies $V_a\left(a^d, 0\right) = 0$, i.e.,

$$a^d = \frac{\pi}{\gamma\sigma^2}. \tag{48}$$

Simple calculations yield:

$$\begin{aligned} V_{aa}\left(a^d, 0\right) &= V\left(a^d, 0\right)\gamma^2\sigma^2 \\ V_{am}\left(a^d, 0\right) &= V\left(a^d, 0\right)\gamma^2\sigma_{r\varepsilon}. \end{aligned}$$

Step 1 gives:

$$m^* = \tau\left(\mu, \kappa^m \frac{V_{aa}\left(a^d, 0\right)}{V_{am}\left(a^d, 0\right)}\sigma_a\right) = \tau\left(1, \kappa^m \frac{\sigma^2}{\sigma_{r\varepsilon}}\sigma_a\right)$$

where σ_a is the normal variation in allocation, e.g., coming from an underlying dynamic problem (for instance, it might depend on variations in the estimated equity premium π). Finally, using $\kappa^a = 0$, Step 2 gives that the optimal allocation a^* satisfies $V_a(a^*, m^*) = 0$, i.e.,

$$\begin{aligned} a^* &= \frac{\pi - \gamma m^* \sigma_{r\varepsilon}}{\gamma\sigma^2} \\ a^* &= \frac{\pi}{\gamma\sigma^2} - \tau\left(\frac{\sigma_{r\varepsilon}}{\sigma^2}, \kappa^m\sigma_a\right). \end{aligned} \tag{49}$$

As expected, if background risk covaries positively with stocks ($\sigma_{r\varepsilon}$ is higher), then the allocation in stocks (a^*) weakly falls. However, this effect is truncated: if $\left|\frac{\sigma_{r\varepsilon}}{\sigma^2}\right| \leq \kappa^m\sigma_a$ then $\tau\left(\frac{\sigma_{r\varepsilon}}{\sigma^2}, \kappa^m\sigma_a\right) = 0$, and there is no effect. Hence, the agent reacts only to large enough background risk.

Example 3. To reduce notational clutter, I solve this example with $u(c) = v(c) = c^{1-\gamma}/(1-\gamma)$, $R^d = 1$. The net interest rate is $r_t = R_t - 1$. We have:

$$V(a, R_t, m) = u(a) + u(R(m)(w-a)) = \frac{a^{1-\gamma}}{1-\gamma} + \frac{R(m)^{1-\gamma}(w-a)^{1-\gamma}}{1-\gamma}$$
$$V_a(a, R_t, m) = a^{-\gamma} - R(m)^{1-\gamma}(w-a)^{-\gamma}.$$

So at $m^d = 0$, $V_a\left(a^d, R_t, m^d\right) = 0$ gives $a^d = w/2$. Next,

$$\begin{aligned} V_{aa}\left(a^d, R_t, m^d\right) &= u''\left(a^d\right) + u''\left(a^d\right) R\left(m^d\right)^{1-\gamma} = -2u'\left(a^d\right)\frac{\gamma}{a^d} \\ V_{am}\left(a^d, R_t, m^d\right) &= \partial_m\left[-R(m)^{1-\gamma}\left(w-a^d\right)^{-\gamma}\right] \\ &= (\gamma-1)R'\left(m^d\right)u'\left(a^d\right) = (\gamma-1)r_t u'\left(a^d\right). \end{aligned} \tag{50}$$

Finally, Step 1 gives:

$$\begin{aligned} m^* &= \tau\left(1, \kappa^m \frac{V_{aa}\left(a^d, R_t, m^d\right)}{V_{am}\left(a^d, R_t, m^d\right)}\sigma_a\right) \\ &= \tau\left(1, \kappa^m \frac{2\frac{\gamma}{a^d}}{(\gamma-1)r_t}\sigma_a\right). \end{aligned}$$

Hence, the interest rate perceived by the decision maker is $R(m^*) = 1 + m^* r_t$, i.e.,

$$R(m^*) = 1 + \tau\left(r_t, \kappa^m \frac{2}{1-1/\gamma}\frac{\sigma_a}{a^d}\right).$$

The agent's attention to the interest rate is lower when the net interest rate r_t is small and when the agent's intertemporal elasticity of substitution, $1/\gamma$, is close to 1. Indeed, when $\left|\kappa^m \frac{2}{1-1/\gamma}\frac{\sigma_a}{a^d}\right| \geq |r_t|$ the decision maker does not pay attention to the interest rate at all.

The agent's optimal consumption at time 1 satisfies $u'(a) - u'(R(m)(w-a))R(m) = 0$, i.e.,

$$a^{-\gamma} = R(m^*)^{1-\gamma}(w-a)^{-\gamma},$$

hence

$$a^* = \frac{w}{1 + R(m^*)^{1/\gamma - 1}}. \tag{51}$$

9.3 Change in Lagrange Multiplier after a Shift

Consider the problem:

$$\max_a u(a,s) \text{ s.t. } B(a,s) \geq 0$$

where a is the action and s is a "shift" parameter (which is general and could represent a shift in income, price, taste, etc.), and the objection function u and the budget constraint $-B$ are concave in a. We will derive the change in action δa when there is an infinitesimal parameter shift δs. Define the Lagrangian:

$$L(a,s,\lambda) = u(a,s) + \lambda B(a,s). \tag{52}$$

We suppose that the constraint binds, $\lambda > 0$.

Lemma 3 *(Change in Action and Lagrange Multiplier after a Shift) After a change $\delta\lambda$, we have:*

$$\delta\lambda = \left(B_a' L_{aa}^{-1} B_a\right)^{-1} \left(B_s' \delta s - B_a' L_{aa}^{-1} L_{as} \delta s\right) \tag{53}$$

and

$$\delta a = -L_{aa}^{-1}\left(L_{as}\delta s + B_a \delta\lambda\right) \tag{54}$$

$$= -\left(L_{aa}^{-1} B_a\right)\left(B_a' L_{aa}^{-1} B_a\right)^{-1} B_s' \delta s - L_{aa}^{-1}\left[1 - B_a\left(B_a' L_{aa}^{-1} B_a\right)^{-1} B_a' L_{aa}^{-1}\right] L_{as}\delta s. \tag{55}$$

With other notations, $p = -B_a$, $\delta y = B_s' \delta s$ (the notations are inspired by the example $B(a,s) = y(s) - p \cdot a$), $b = L_{aa}^{-1} B_a$, we have:

$$\delta a = -L_{aa}^{-1} L_{as}\delta s - b\delta\lambda$$

where $\delta\lambda$ adjusts to satisfy the budget constraint:

$$-p'\delta a + \delta y = 0. \tag{56}$$

The interpretation is that b is the vector of the basis axis of adjustment when the dollar budget changes. Call $\delta c = -L_{aa}^{-1} L_{as}\delta s$ the "myopic" change without thinking about the budget constraint. Then,

$$\delta a = \delta c - b\delta\lambda$$

where $\delta\lambda$ solves (56):

$$\begin{aligned}\delta y &= p'\left(\delta c - b\delta\lambda\right)\\ &\Rightarrow \delta\lambda = \left(p'b\right)^{-1}\left(p'\delta c - \delta y\right),\end{aligned}$$

so

$$\begin{aligned}\delta a &= \delta c - b\left(p'b\right)^{-1}\left(p'\delta c - \delta y\right)\\ &= b\left(p'b\right)^{-1}\delta y + \delta c - b\left(p'b\right)^{-1}p'\delta c. \qquad (57)\end{aligned}$$

The interpretation of (54) is as follows. The first term, $-L_{aa}^{-1}L_{as}\delta s$, is the "myopic" change in action, using the same prices (Lagrange multiplier) as before the shift and forgetting about the budget constraint. The term is the change in action to satisfy the budget constraint. This interpretation motivates Step 3 in the Sparse BR algorithm with constraints (Algorithm 2).

In equation (55), the first term is a direct change of income, and the second is the change in the price, L_a, that is orthogonal to the price vector B_a.

Proof of Lemma : Differentiating $L_a\left(a, s, \lambda\right) = 0$,

$$0 = L_{aa}\delta a + L_{as}\delta s + L_{a\lambda}\delta\lambda,$$

so as $L_{a\lambda} = B_a$,

$$\delta a = -L_{aa}^{-1}\left(L_{as}\delta s + B_a\delta\lambda\right). \qquad (58)$$

The budget constraint $B\left(a, s\right)$ gives:

$$\begin{aligned}0 &= B_a'\delta a + B_s'\delta s\\ &= -B_a'L_{aa}^{-1}\left(L_{as}\delta s + B_a\delta\lambda\right) + B_s'\delta s,\end{aligned}$$

so

$$\delta\lambda = \left(B_a'L_{aa}^{-1}B_a\right)^{-1}\left(B_s'\delta s - B_a'L_{aa}^{-1}L_{as}\delta s\right).$$

Note that when there are K budget constraints, then $B_a'L_{aa}^{-1}B_a \in \mathbb{R}^{K\times K}$, and $B_s'\delta s$ and $B_a'L_{aa}^{-1}L_{as}\delta s \in \mathbb{R}^K$.

Finally, we have:

$$\begin{aligned}\delta a &= -L_{aa}^{-1}\left(L_{as}\delta s + B_a\delta\lambda\right)\\ &= -L_{aa}^{-1}\left(L_{as}\delta s + B_a\left(B_a'L_{aa}^{-1}B_a\right)^{-1}\left(B_s'\delta s - B_a'L_{aa}^{-1}L_{as}\delta s\right)\right)\\ &= -L_{aa}^{-1}B_a\left(B_a'L_{aa}^{-1}B_a\right)^{-1}B_s'\delta s - L_{aa}^{-1}\left(1 - B_a\left(B_a'L_{aa}^{-1}B_a\right)^{-1}B_a'L_{aa}^{-1}\right)L_{as}\delta s.\end{aligned}$$

10 Some Enrichments of the Model

10.1 Enrichments of the Basic Sparse BR Model

Operator language To express variants of the model, it is useful to use the following operators on a function $f(a,m)$:

$$\begin{array}{cc}\left(\Delta_{m_i}f\right)(m) = \left(m_i - m_i^d\right)\partial_{m_i}f(m), & \left(\Delta_{a_i}f\right)(m) = \left(a_i - a_i^d\right)\partial_{a_i}f(m),\\ \left(\Delta_{\eta_m}f\right)(a) = \eta_m\cdot\partial_m f(m), & \left(\Delta_{\eta_a}f\right)(a) = \eta_a\cdot\partial_a f(a).\end{array} \tag{59}$$

The notation $\partial_a f(a)$ is the differential of f at point a, and the dot $\cdot$ is the vector product; for instance, $\eta_a\cdot\partial_a f(a) = \sum_i \eta_{a_i}\frac{\partial f}{\partial a_i}(a)$. With that notation, in the Sparse BR model, the penalties (9) and (11) are equivalently expressed:

$$\kappa[m] = \kappa^m\sum_i\left\|\Delta_{m_i}\Delta_{\eta_a}V(a,x,m)\right\|, \qquad \kappa[a] = \kappa^a\sum_i\left\|\Delta_{a_i}\Delta_{\eta_m}V(a,x,m)\right\|.$$

However, the operator notation generalizes more easily.

Discrete sets, non-differential operators Sometimes (e.g., when the space underlying a is not continuous) it is useful to replace the differential operators used in Algorithm 1 by their non-differential counterparts (the superscript F is a shorthand for "finite"):

$$\begin{array}{cc}\left(\Delta_{m_i}^F f\right)(m) = f(m_i, m_{-i}) - f\left(m_i^d, m_{-i}\right), & \left(\Delta_{a_i}^F f\right)(m) = f(a_i, a_{-i}) - f\left(a_i^d, a_{-i}\right),\\ \left(\Delta_{\eta_m}^F f\right)(m) = f(m+\eta_m) - f(m), & \left(\Delta_{\eta_a}^F f\right)(a) = f(a+\eta_a) - f(a).\end{array}$$

How to define "$a+\eta_a$" when the action space A is finite? Assume that space A comes equipped with a distance $d(a,a')$: for instance, if $A = \{1,...,n\}$ ordered in $\mathbb{N}$, $d(a,a') = |a-a'|$, and if A is just a set of options with no clear metric (e.g., 4 options with no particular spatial ordering), we can have $d(a,a') = 1_{a\neq a'}$. Then, "$a+\eta_a$" stands for a random variable variable $\widetilde{a}$ with $\mathbb{P}(\widetilde{a} = a') = Ke^{-\beta d(a,a')}$ for some $\beta > 0$ and a constant K: it has a mode at a, and decreases away from a.

Likewise, sometimes (e.g., when dealing with functions with discrete support) it might be useful to have a non-differential version of the Λ matrix. A simple device is to consider values $a^*(m)$ and set:

$$\Lambda_{ii} = \frac{1}{(m_i - \mu_i)^2} \mathbb{E}\left[u\left(a^*(\mu), \mu\right) - u\left(a^*\left(m_i, \mu_{-i}\right), \mu\right)\right] \tag{60}$$

where $a^*(m)$ is the optimum under the model parametrized by m.

Averaging In the baseline model, Λ is evaluated at the default action and representation. We could extend that by averaging around the baseline. For instance, define $\Lambda(a, m, x) = -V_{am} V_{aa}^{-1} V_{am}$ and

$$\Lambda = \mathbb{E}\left[\Lambda\left(a^d + \eta_a, m^d + \eta_d, x\right)\right] \tag{61}$$

where the expectation is over η_a, η_d, and x. So, we add noise around a^d and m^d.

For instance, if we use the default action (no saving), there is no impact of the interest rate, the simple Λ is 0. But with averaging, the agent will see that for some other policies (non-zero saving) the interest rate does matter.

Enrichment via loss aversion One interesting enrichment is to use a loss-aversion-based penalty for negative outcomes but not positive ones. Denote $x^- = \max(-x, 0)$, i.e., $x^- = -x$ for $x < 0$ and 0 for $x \geq 0$. Call Δ_- the "loss aversion" operator, $(\Delta_- f)(x) = (f(x))^-$. Instead of the original formulation (11), $\kappa[a] = \kappa^a \sum_i \|\Delta_{a_i} \Delta_{\eta_m} V\|$, we could have for a complexity parameter $\kappa^{a,-}$:

$$\kappa[a] = \kappa^{a,-} \sum_i \|\Delta_- \Delta_{a_i} V\|.$$

This operator Δ_- may be useful, first, because loss aversion seems important in many parts of economic psychology. Also, it is serviceable in the (relatively rare) cases where a gamble is offered with no downside. To see this, take the problem where the agent can pick a quantity $a \in [0, 1]$ of a gamble g with non-negative support, i.e., the agent obtains utility $u(ag)$. It is clear that, whatever the complexity of g, by domination, picking $a = 1$ is the right thing to do. This is missed by the basic algorithm, but is detected with the loss aversion operator: normalizing $u(0) = 0$,

$$\kappa[a] = \kappa^{a,-} \|\Delta_- \Delta_a V\| = \kappa^{a,-} E\left[(u(ag) - u(0))^-\right] = 0$$

because $u(ag) - u(0) \geq 0$ almost surely. Then, it is clear that there is no penalty for complexity.

We can also mix and match, and replace (11) by

$$\kappa[a] = \kappa^a \sum_i \|\Delta_- \Delta_{a_i} \Delta_{\eta_m} V\| + \kappa^{a,-} \sum_i \|\Delta_- \Delta_{a_i} V\|.$$

This is adding a "loss aversion" operator to the previous operators. It seems that in many situations it is not worth bothering about the loss aversion operator Δ_- , which adds some algebraic complexity, but it is good to have it available when "domination" patterns are important.

Finally, the decision maker might restrict himself to a parametrization of the actions. For instance, if the underlying action is $A = (A_1, ..., A_T)$ where A_t is the savings rate at time t, we can have $A_t(a) = a_0 + a_1 t$, a savings rate that depends in an affine way on age, where (a_0, a_1) is a 2-dimensional parametrization of the agent's savings rate.

Contingencies-matching Suppose there is a random variable ε in the value function, $V(a, x, m, \varepsilon)$. Then, the following variant of Step 2 of Algorithm 1 may be useful.

Step 2': For each realization of the noise ε, pick the best action:

$$a(x, \varepsilon) \in \max_a V(a, x, m, \varepsilon) - \kappa[a] \tag{62}$$

and then play $a(\varepsilon)$ according to the probability of ε.

This variant accounts for "probability matching." In the paradigmatic game, a biased coin will be tossed and come out as heads with probability 0.7, say, and heads with probability 0.3. Subjects have to predict which side will be drawn. They tend to predict heads with probability 0.7. This is a deviation from rationality which implies betting on heads at all times. Step 2' above generates that behavior even when κ^a is set to 0: with probability 0.7 (resp. 0.3), the agent draws heads (resp. tails), and best-responds to it.

A Sparse BR model with fixed cost For some purposes, it may be useful to have a model with a fixed cost of thinking, rather than the ℓ_1 cost of thinking worked out in the paper. To this end, I propose the following model. It pays keen attention to the scaling of the various costs and benefits.

Algorithm 3 *(Sparse BR Algorithm with Fixed Costs) To solve the problem* $\max_a V(a, x, \mu)$*, the agent uses the following two steps:*

1. ***Optimize on the representation of the world.*** *Using the realism loss matrix* Λ *given in (7), determine* m *by solving for*

$$\min_{m} \frac{1}{2} (m - \mu)' \Lambda (m - \mu) + \kappa [m]. \tag{63}$$

The first part is a measure of expected loss from a poor simulation while the second part is the complexity cost of the representation with a fixed cost:

$$\kappa [m] = \kappa^m \sum_i \|\eta_{m_i} \eta_a V_{m_i a}\| \, 1_{m_i \neq m_i^d}. \tag{64}$$

2. ***Optimize on the action.*** *Maximize over the action* a*:*

$$\max_{a} V (a, x, m) - \kappa [a]$$

where the expectation is over the realizations of ε *and where the complexity cost of the action,* $\kappa [a]$*, is:*

$$\kappa [a] = \kappa^a \sum_i \|\eta_{a_i} \eta_m V_{m a_i}\| \, 1_{a_i \neq a_i^d}. \tag{65}$$

In the formulation of Algorithm 3, the costs $\kappa [m]$ and $\kappa [a]$ are fixed costs. The agent pays the cost only if $m_i \neq m_i^d$. The model includes some scaling of the fixed cost: that is to satisfy the invariance properties listed in Proposition 1. Here, η_m represents some variability of m.

Problem (63)-(64) is non-convex, so in general it is very difficult to solve. However, when Λ is a diagonal matrix, it allows a simple solution:

$$m_i^* = \begin{cases} m_i^d & \text{if } \left|m_i^d - \mu_i\right| \leq \kappa_i \\ \mu_i & \text{if } \left|m_i^d - \mu_i\right| > \kappa_i \end{cases} \tag{66}$$

where

$$\begin{aligned} \kappa_i &:= \sqrt{\frac{2\kappa^m \|\eta_{m_i} \eta_a V_{m_i a}\|}{\Lambda_{ii}}} \\ &= \sqrt{\frac{\kappa^m \|\eta_{m_i}\| \, \|V_{aa} \eta_a\|}{\|V_{m_i a}\|}} \text{ when } a \text{ is one-dimensional.} \end{aligned}$$

The idea is that agents pay the fixed cost only if the difference between the default and the optimal representation is large enough ($\left|m_i^d - \mu_i\right| > \kappa_i$), and if the action is important enough (high Λ_{ii}).