

# Coarse Decision Making

Nabil I. Al-Najjar\*

and

Mallesh Pai†

First draft: December 2008

This version: March 2009

## Abstract

We study decision makers who willingly forgo acts that finely vary with states, even though these acts are informationally and technologically feasible. They opt instead for coarse rules that are less sensitive to state by state variations. We model this coarse decision making as a consequence of individuals using classical, frequentist methods to draw robust inferences from scarce data. Our central theme is that coarse decision making arises to mitigate the problem of over-fitting the data. The main implication of this framework is behavior that is biased towards simplicity: decision makers choose *models*, or *decision frames* that are statistically simple, in a sense we make precise. We are also able to give a unified interpretation of many seemingly anomalous cognitive and decision making procedures, such as categorization, reliance on linear orders, and satisficing.

---

\* Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston IL 60208.

**E-mail:** <[al-najjar@northwestern.edu](mailto:al-najjar@northwestern.edu)>.

**Research page :** <http://www.kellogg.northwestern.edu/faculty/alnajjar/htm/index.htm>

† Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston IL 60208.

**E-mail:** <[m-pai@northwestern.edu](mailto:m-pai@northwestern.edu)>.

**Research page :** <http://www.kellogg.northwestern.edu/faculty/pai/>

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>The Model</b>	<b>6</b>
<b>3</b>	<b>So Many Theories, So Few Facts</b>	<b>8</b>
3.1	Learning, Naïve Empiricism, and Over-fitting . . . . .	8
3.2	Coarse Decision Making as Response to Over-fitting . . . . .	11
3.3	The Bayesian Methodology cannot Explain Coarse Decision Making . . . . .	14
<b>4</b>	<b>Satisficing, Categorization, and Linear Orders</b>	<b>16</b>
4.1	A Formal Model of Categorization . . . . .	16
4.2	Linear Orders . . . . .	18
4.3	Satisficing . . . . .	21
4.4	Discussion . . . . .	23
<b>5</b>	<b>Categorization and Heterogeneity</b>	<b>26</b>
<b>6</b>	<b>Structure, Identification, and the i.i.d. Assumption</b>	<b>29</b>
6.1	Absence of Built-in Structure . . . . .	29
6.2	Identification Problems . . . . .	30
6.3	The i.i.d. Assumption . . . . .	31
<b>A</b>	<b>A Technical Primer to VC-theory</b>	<b>32</b>
A.1	Vapnik Chervonenkis Dimension . . . . .	32
A.2	Pollard’s Pseudo-Dimension . . . . .	33
<b>B</b>	<b>Proofs of Theorems in Section 4</b>	<b>34</b>
<b>C</b>	<b>Proofs of Propositions in Section 5</b>	<b>41</b>

“*Ab uno disce omnes.*”<sup>1</sup>

## 1 Introduction

A central organizing principle in economic modeling is to explain observed behavior as a rational response to constraints. This methodology covers an astonishingly wide range of behaviors and institutions. In traditional price theory, consumer and firm choices are explained as responses to wealth and technological constraints, while game theory and information economics use informational constraints to uncover the incentives that shape the interactions of economic agents.

Despite the breadth of phenomena they cover, standard models struggle to explain what we shall refer to as *coarse decision making*. By this we mean the broad array of situations whose defining characteristic is decision makers willingly forgoing acts that finely vary with states, even though these acts are informationally and technologically feasible. Decision makers opt instead for acts that are less sensitive to state by state variations—‘coarse’ acts in our terminology.

A striking manifestation of coarse decision making is categorization, where decisions are based on a coarse partition of the state space into ‘categories.’ Categorization is universal in all aspects of human cognition and decision making, as attested by the vast psychological literature on the subject.<sup>2</sup> Economic examples of categorization are too numerous to comprehensively list here, so we limit ourselves to a few recent examples. In finance, there has been much interest in ‘style investing,’ where investors categorize available investment opportunities into asset classes they refer to as ‘styles.’ When making portfolio allocation decisions, investors shift funds between styles based on relative performance, rather than directly select which assets to invest into.<sup>3</sup> Sharpe (1992) pioneered the application of ‘style-based’ analysis to portfolio management problems, and showed that 90% of the variation in

---

<sup>1</sup>“From one, learn all.” Virgil’s Aeneid.

<sup>2</sup>See for example the early studies by Rosch and Lloyd (1976), Chi, Feltovich, and Glaser (1981) and Reed (1972).

<sup>3</sup>See Bernstein (1995) and Dimson and Nagel (2002) for a historical overview.

the return on mutual funds can be explained by these funds' exposure to a few asset classes. Barberis and Shleifer (2003) explain stylized facts about movements of asset prices assuming that investors are style investors. Another example of categorization is Fryer and Jackson (2008)'s model where agents in social settings categorize past experiences into a finite number of categories. Decisions are made based on the category to which a case belongs, rather than on the case itself. They show that agents who categorize will discriminate against minorities despite having no a priori bias favoring such discrimination.

Other manifestations of coarse decision making are the prevalence of linear orders and satisficing. By reliance on linear orders we refer to a decision maker organizing his environment along a set of linearly ordered dimensions. For example, a firm may plan according to sales, stock price, cost per unit; stocks are classified according to their P/E ratio, capitalization; and so on. The decision maker may then restrict attention to decision rules that are appropriately monotone with respect to these dimensions.

The concept of satisficing, originally introduced by Simon (1955), refers to procedures under which decision rules are evaluated based on whether they achieve some discrete payoff thresholds. For example, a firm that can control its cost, output and prices continuously may nonetheless plan based on attaining discrete profit targets.

Coarse decision making is also related with *rules of thumb* and other closely related ideas, such as *decision heuristics*, *routines*, and *analogies*.<sup>4</sup> These cognitive devices may be understood as 'reasonable' responses of a decision maker attempting to do well in an environment that is complex or poorly understood. They are instances of coarse decision making in that choices are restricted to a subset of all feasible rules, effectively lumping many distinct instances together, and ignoring the fine details of available information.

The goal of this paper is to provide a simple, unified theory of coarse decision making. The central idea is to view decision makers as learning

---

<sup>4</sup>See Tversky and Kahneman (1974), Nelson and Winter (1982), and Samuelson (2001), among others.

from evidence using classical, frequentist methods, like most econometricians and statisticians do. Decision makers in our model balance two conflicting objectives: (1) on the one hand, they would like to select among as broad class of decision rules as possible in order to improve fit; but (2) they need to restrict the class of rules to avoid over-fitting the data. The coarse class of rules they end up with is a compromise between fitting and over-fitting.<sup>5</sup>

Somewhat more formally, we model a decision maker who follows a two-stage decision procedure:

1. *Model-selection stage*: The agent chooses a ‘model,’ or ‘decision frame’  $\mathcal{F}$  which consists of a set of rules, each a function from the space of observables  $X$  to the set of feasible actions.
2. *Inference stage*: A decision rule in  $\mathcal{F}$  is selected by applying frequentist inference criteria to sample information.

This is analogous to the procedure followed in linear regression models: one first selects the regression model (identify a set of regressors and hypothesize that they determine the regressand linearly), then uses the OLS procedure to compute the estimated values of the parameters.

Coarse decision making is formally *defined* as the selection of a decision frame that restricts the set of rules to a strict subset  $\mathcal{F}$ . The central questions in our paper are:

- What must be true about the environment and the agent’s decision procedure for coarse decision making to arise?
- How is the degree of ‘coarseness’ affected by factors such as the amount of data available and the set of feasible rules?

---

<sup>5</sup>Although the importance of the problem of over-fitting is well-recognized in empirical work, the theory literature largely overlooked it (and predictably so, since over-fitting has little meaning when a decision maker is modeled as a Bayesian). The closest paper is Al-Najjar (2009) which studies the asymptotic properties of uniform learning. That paper, which deals with more abstract setting than ours, does not discuss over-fitting or applications to cognitive phenomena. Another related paper is the recent work by Gilboa and Samuelson (2008).

- What are observable, economically significant, consequences of coarse decision making?
- What can a theory of coarse decision making tell us about how behavior changes in response to changes in the environment?

As an answer to the first two questions, we propose that coarse decision making arises when the agent behaves like a classical statistician, with no prior beliefs about the true distribution  $P$ , and data is scarce relative to the set of all feasible decision rules. The decision maker would have liked to have enough data to evaluate all feasible rules. However, as large as it may be in absolute terms, the available data is still too small to accurately learn the performance of each possible rule in a ‘rich’ enough environment. The decision maker solves this problem by narrowing the scope of his frame, i.e. discarding all rules outside some subset  $\mathcal{F}$ .

As an answer to the third question, “*What are observable consequences of coarse decision making?*,” our model implies behavior that is biased towards simplicity: agents choose decision frames that are *statistically simple* in a sense we make precise in Section 2. Having made this general point, we then turn to specific classes of decision patterns that have been central in psychology and economics: categorization, linear orders, and satisficing. In Section 4 we show that they can all be understood as attempts to mitigate the problem of over-fitting in environments where the set of feasible decision rules is rich relative to available evidence. In the case of categorization, for example, the decision maker adopts a decision frame  $\mathcal{F}$  consisting of rules that are measurable with respect to a coarse partition of the observables. Although finer partitions are informationally and technologically feasible, the decision maker prefers coarser partitions to counter the risk of selecting a rule that tracks the sample data too closely (over-fitting). Similar arguments apply in the case of satisficing and linear orders, where statistical simplicity can be identified with the number of dimensions the decision maker considers.

The tension between improving fit while avoiding over-fitting on which this paper builds is fundamental to classical statistics. Returning to the linear regression analogy, when data is limited, incorporating more regressors, or considering more general functional forms, although helpful in improving

the empirical fit, harm the validity of the regression estimates. The modeler discards some potential regressors not because of lack of information or prohibitive computational cost, but because incorporating them undermines the quality of the inference he draws from the data. As we discuss in Section 3.3, this trade-off is meaningless to a Bayesian decision maker who subjectively believes he understands the ‘true’ data generating process.

As an answer to the fourth question, “*What can a theory of coarse decision making tell us about how behavior changes in response to changes in the environment?*,” our model makes predictions about the comparative statics of coarse decision making. Essentially, what determines the coarseness of the decision rules considered is the amount of data available and the ‘richness’ of the set of feasible rules. Thus, coarse rules of thumb may continue to hold great appeal even in high-stakes decision problems, not because decision makers have not had ample opportunity to contemplate and introspect, but because of the limited data they have relative to the statistical complexity of the underlying problem. In our model, the monetary stakes involved, or traditional sources of ‘bounded rationality’ like computational complexity and memory limitations play no *direct* role in shaping coarse decision making. They can, however, play an indirect role, as we discuss in Section 4.4.

In Section 5 we consider two applications to the literature on corporate culture. A principal (to be thought of as a firm) sets its corporate culture, modeled as a preferred action for each possible eventuality. However, the only way he can communicate this is by having his agents learn from the firm’s past history using (possibly different) decision frames. A principal who suffers a coordination cost when his agents miscoordinate, might want to ‘water down’ his corporate culture to be jointly measurable with respect to the two partitions. Even if the principal can (costlessly) train his agents to refine their partitions, he may choose not to, since the agents may then overfit the firm’s past history and learn less as a result.

We close this Introduction with a few methodological points. Choosing from a simple, or coarse class of rules is often viewed as an ‘anomaly,’ an aberration afflicting ‘boundedly rational’ or behaviorally biased agents. We do not take a position on the issue of bounded rationality, nor do we deny

its potential relevance. What we propose is a more unified approach based on the inherent difficulties of statistical inference from limited data.

A natural question when encountering a phenomenon in contradiction with standard theory is: *Why go through the trouble of investing in a new theory to explain the anomalous behavior when one can do with context-specific fixes and patches?* Almost thirty years ago, commenting on the need for theory, Lucas Jr. (1980, p. 697) wrote: “To the journalist, each year brings unprecedented new phenomena, calling for unprecedented new theories (where ‘theory’ amounts to a description of the new phenomena together with the assertion that they are new).” Like Lucas, we believe that “it is in our interest to take exactly the opposite viewpoint.” No disciplined theory can fit observed phenomena as well as a collection of disparate explanations, each tailor-made to accommodate a specific ‘bias’ or anomalous observation. A unified model, on the other hand, provides better insights into what drives behavior and how it changes in response to changes in the environment.

## 2 The Model

We consider a decision problem characterized by a set of *observables* or *explanatory variables*  $X$ , a set of *outcomes*  $Y$ , a set of *actions*  $A$ ; and a payoff function:

$$u : X \times Y \times A \rightarrow \mathcal{R}.$$

For expository and technical simplicity, we assume that  $X, Y$  and  $A$  are finite.

We distinguish  $x$  (the observables) from  $y$  (the outcome) since the choice of actions can be conditioned on  $x$ , whereas  $y$  remains unobserved until after the action is chosen. A *decision rule*  $f$  is a contingent action plan

$$f : X \rightarrow A$$

which determines the action the decision maker takes,  $f(x)$ , as a function of the observables  $x$ . Let  $\mathbf{F}$  denote the set of all decision rules.

In line with our interest in learning and statistical complexity, we assume that the decision maker is free to choose any contingent plan  $f \in \mathbf{F}$ . This assumes away technological or informational factors that limit the decision maker’s freedom of choice within  $\mathbf{F}$ .

The decision maker's *environment* is a joint distribution  $P$  on  $X \times Y$ . We assume that, if  $P$  were known, he would evaluate a decision rule  $f$  according to its expected payoff:

$$E_P f \equiv E_P u(x, y, f(x)).$$

Our general setting corresponds to what is known as *treatment problems* in the econometrics literature (see for example Manski (2008)). We offer two examples to illustrate. First, think of  $x$  as the set of observable characteristics of a patient, and  $f(x)$  is the treatment (*e.g.*, a type of surgical operation, a diet, medical tests, and so on) chosen based on the observable characteristics. The set of outcomes  $Y$  consists of all payoff-relevant measures of the health of the patient which may be in part influenced by the treatment applied.

Another example is that of an investor who bases his investment in an asset on a vector of observables  $x$ . This vector may contain information about economic aggregates, decisions made by various companies, or regulatory changes. The outcome  $y$  is the actual return on the asset, which is unknown at the moment the investment decision is made. It is natural for the payoff function to take the form  $u(y, a)$ , so  $x$  is relevant only in so far as it influences the decision  $a$ .

An important special case is what we shall refer to as *forecasting problems* which include, among others, categorization problems and linear regression. See Sections 3.1.1 and 4. In forecasting problems the sets of actions and outcomes coincide, so  $A = Y$ , and the utility function takes the form  $u(y, a) = -\rho(y, a)$ , where  $\rho$  is a metric on  $Y$ .

The interpretation is that the decision maker formulates a rule  $f$  to forecast the value of  $y$  based on  $x$ . The metric  $\rho(y, a)$  measures the forecasting error, *i.e.*, the distance between the forecasted value  $a$  to the realized value  $y$ . Thus, the utility function  $u$  implicitly defines a similarity structure via  $\rho$  since none is present in  $X, Y$ , or  $A$ . For example, the utility function  $u(y, a) = -|proj(y) - proj(a)|$  first projects  $a$  and  $y$  to points in  $[0, 1]$  (via the function  $proj$ ). It then applies the usual metric on  $[0, 1]$  to judge the closeness of  $a$  and  $y$ . An alternate labeling  $proj'$  induces a different similarity relationship and thus a different utility function.

## 3 So Many Theories, So Few Facts

### 3.1 Learning, Naïve Empiricism, and Over-fitting

The decision maker uses sample information to form a frequentist estimate of  $P$ .<sup>6</sup> Specifically, the decision maker makes  $t$  observations. Each is a pair  $(x, y)$ , consisting of a vector of observables  $x$  and the corresponding outcome  $y$ .<sup>7</sup> This past experience of the relationship between observables and outcomes is represented by a sample:

$$s^t = \{(x_1, y_1), \dots, (x_t, y_t)\}.$$

We assume that the sampling is i.i.d. draws from the unknown  $P$ .

The decision maker cannot evaluate a rule  $f$  using its true performance  $E_P f$ , since the true distribution  $P$  is unknown to him. He relies instead on  $f$ 's performance under the *empirical distribution*:

$$\nu(s^t)(E) \equiv \frac{\#\{i : (x_i, y_i) \in E\}}{t}.$$

The empirical distribution assigns to each event  $E \subset X \times Y$  its relative frequency in the sample  $s^t$ . Then, the empirical performance of a rule  $f$  is:

$$E_{\nu(s^t)} f \equiv \frac{1}{t} \sum_{i=1}^t u(y_i, f(x_i)),$$

which is  $f$ 's average payoff over the sample.

The estimated performance  $E_{\nu(s^t)} f$  will typically differ from the true performance,  $E_P f$ , due to sampling error. Define the *empirical discrepancy*:

$$\Delta_t(f) \equiv \sup_P \int_{s^t} |E_{\nu(s^t)} f - E_P f| dP^t,$$

---

<sup>6</sup>We discuss the Bayesian approach, under which the decision maker has a prior belief on  $P$ , in Section 3.3.

<sup>7</sup>In the empirical literature on treatment problems, it is often assumed that  $y$  is not directly observed- the decision maker only sees the observables  $x$ , the action taken  $a$  and the realized payoff  $u(y, a)$ . This leads to an additional identification problem, which we avoid by assuming  $y$  is directly observed.

where  $P^t$  is the (product) distribution on i.i.d. samples of length  $t$ . This is the difference between  $E_P f$  and  $E_{\nu(s^t)} f$ , averaged over all samples under the worst-case distribution  $P$ .

A consequence of the law of large numbers is that for a large enough sample size  $t$ , the empirical estimate of the performance of  $f$  is close to its true performance with high probability for *any* probability distribution  $P$  and rule  $f$ . More precisely, for every  $\epsilon > 0$  there exists a sample size  $\bar{t}$  such that: <sup>8</sup>

$$\sup_{f \in \mathbf{F}} \Delta_t(f) < \epsilon, \quad \forall t \geq \bar{t}. \quad (1)$$

A *naïve frequentist* decision maker selects the rule that best fits the data:

$$f_{s^t}^* = \operatorname{argmax}_{\mathbf{F}} E_{\nu(s^t)} f,$$

intuitively hoping that  $E_P f_{s^t}^*$  is ‘close’ to  $E_P f_P^*$ .

A flawed argument in favor of this behavior is as follows: when (1) holds, the (average) empirical performance of each and every rule  $f$  is close to the true performance. As a result the empirical performance of  $f_{s^t}^*$  is close to its true performance, and therefore  $f_{s^t}^*$  performs nearly as well as the true optimal rule  $f_P^*$  for any true  $P$ .

The reason we call this decision maker naïve is that he ‘over-fits’ the data by picking rules tailored to the observed samples. Although  $f_{s^t}^*$  enjoys an unparalleled match with past data, its exceptional performance is an illusion. Intuitively, this is a result of closely tracking the particular sample  $s^t$  in a way that is uninformative of  $f$ ’s true performance  $E_P f$ .

As a formal illustration, fix  $Y = A = \{0, 1\}$  and  $u(y, a) = \chi\{y = a\}$ , where  $\chi$  is the indicator function. Then for any  $\frac{1}{2} > \epsilon, \delta > 0$ , there exists  $X$  with the associated set of functions  $\mathbf{F}$ , and a  $t$  such that:

$$\sup_{f \in \mathbf{F}} \sup_P \int_{s^t} \left| E_{\nu(s^t)} f - E_P f \right| dP^t = \epsilon, \quad (2)$$

yet:

$$\sup_P \int_{s^t} \left| \max_{\mathbf{F}} E_{\nu(s^t)} f - \max_{\mathbf{F}} E_P f \right| dP^t \geq \delta. \quad (3)$$

---

<sup>8</sup>The finiteness of  $X, Y$  and  $A$  are sufficient for this. In general, this would require suitable regularity conditions on  $u$ .

In other words, there exist natural settings in which the empirical discrepancy of any single decision rule is *small*, and yet the empirically best performing rule is not ‘close’ to the true best performing rule.

### 3.1.1 Illustration: The Case of Classical Regressions

An analogy with regressions in classical econometrics may help in illustrating these ideas. A decision maker picks a regression curve (not necessarily linear) to fit a set of observations. The fundamental problem is to reconcile the fact that data is limited with the need to make out-of-sample predictions.

If the decision maker is free to choose any regression curve, he is guaranteed to find one that fits the data perfectly. This is illustrated by the curved (green) line in the figure. However, this choice ‘over-fits’ the data. It picks a function that performs well on the available data, but with little or no predictive power ‘out of sample.’ This problem arises because the set  $\mathcal{F}_{\text{all}}$  of *all* regression curves is too ‘large’ to be resolved by the (limited) available data.

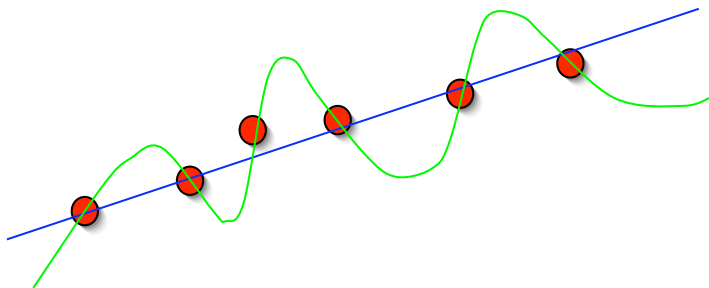


Figure 1: *Over-fitting*

In econometric research this problem is circumvented by restricting the space of regression curves—for example, to linear regressions  $\mathcal{F}_{\text{linear}}$  (with a small set of explanatory variables). Note the tension between fit and over-fit: although  $\mathcal{F}_{\text{linear}}$  will not fit the data as well as the unrestricted set  $\mathcal{F}_{\text{all}}$ , the sacrifice in fit is rewarded with robustness of the estimates. Regression

techniques ensure that, with high probability, the regression that fits the data best in  $\mathcal{F}_{\text{linear}}$  is close to the best regression *within*  $\mathcal{F}_{\text{linear}}$  for any data generating process.

This analogy, while suggestive, is imperfect because our framework does not impose any *a priori* order or any other structure on the sets  $X, Y$  and  $A$ . This is important since we may want to, among other things, *explain* why decision makers may prefer linear orders to handle problems with no obvious order structure.

## 3.2 Coarse Decision Making as Response to Over-fitting

### 3.2.1 Formal Criterion

In view of the risk of over-fitting (2)-(3), and in keeping with the regression intuition, for a given set of rules  $\mathcal{F} \subset \mathbf{F}$ , define:

$$\Delta_t(\mathcal{F}) \equiv \sup_P \int_{s^t} \sup_{f \in \mathcal{F}} |E_{\nu(s^t)} f - E_P f| dP^t. \quad (4)$$

Crucially, the ‘ $\sup_{f \in \mathcal{F}}$ ’ is inside the integral (rather than outside, as in  $\sup_{f \in \mathbf{F}} \Delta_t(f)$ ). This reflects the fact that we would like the empirical estimate of the performance of *each*  $f \in \mathcal{F}$  to be close to its true performance with high probability.

As we pointed out earlier, if  $\Delta_t(\mathcal{F})$  is ‘large,’ the empirical performance of  $f$  is a poor estimate of its true performance, and hence the true performance of the selected rule,  $E_P f_{s^t}^*$ , could be (much) lower than the performance of the true best rule,  $E_P f_P^*$ . This leads to the following definition:

**Definition 1** A model or decision frame is a pair  $(\mathcal{F}, \epsilon)$  where  $\mathcal{F} \subseteq \mathbf{F}$  and  $\epsilon > 0$  such that:

1.  $\Delta_t(\mathcal{F}) \leq \epsilon$ ;
2. Given the data  $s^t$ , he selects the empirically best performing rule in  $\mathcal{F}$ :

$$f_{s^t}^{\mathcal{F}} \in \operatorname{argmax}_{f \in \mathcal{F}} E_{\nu(s^t)} f.$$

When  $\epsilon$  is understood from the context, we refer to  $\mathcal{F}$  alone as the decision maker’s model or frame. The interpretation is that  $\mathcal{F}$  reflects the way the decision maker ‘frames’ the problem by determining how past evidence is used to make inferences and deciding which rules to consider and which to discard. The parameter  $\epsilon$  represents the degree of *robustness* the decision maker wishes to have in  $\mathcal{F}$ . A small  $\epsilon$  means that he is guaranteed that the expected difference between the estimated and true performances is small regardless of what the true distribution  $P$  is.

### 3.2.2 Robustness, Over-fitting and Coarse Decision Making

The essence of our account of coarse decision making is this: when data is scarce relative to the set of feasible rules, the decision maker corrects for the problem of over-fitting by restricting his choice to a subset  $\mathcal{F}$ , forgoing feasible rules outside  $\mathcal{F}$ . In our linear regression illustration, the econometrician rejects curves like the one in Figure 1 despite their superior (in fact, perfect) fit. He chooses instead to restrict himself to select the best fitting rule from the much smaller class  $\mathcal{F}_{\text{linear}}$ .

To further clarify the relationship between decision making, robustness, and over-fitting, decompose the discrepancy between the performance of the chosen rule  $f_{s^t}^{\mathcal{F}}$  and the true best rule  $f_P^*$  as

$$\begin{aligned} & \sup_P \int_{s^t} (E_P f_P^* - E_P f_{s^t}^{\mathcal{F}}) dP^t \\ = & \sup_P \left[ \underbrace{(E_P f_P^* - \max_{f \in \mathcal{F}} E_P f)}_{\text{1: measures fit}} + \underbrace{\int_{s^t} (\max_{f \in \mathcal{F}} E_P f - E_P f_{s^t}^{\mathcal{F}}) dP^t}_{\text{2: measures over-fit}} \right]. \quad (5) \end{aligned}$$

A decision frame  $\mathcal{F}$  must therefore balance two conflicting criteria:

1. Fit improves as  $\mathcal{F}$  becomes ‘large.’ In the extreme case where  $\mathcal{F} = \mathbf{F}$ , we trivially have  $E_P f_P^* = \max_{f \in \mathcal{F}} E_P f$  for each  $P$ . In this case, term 1 equals zero and we have perfect fit (as in case of set of all curves,  $\mathcal{F}_{\text{all}}$ , in our regression illustration).
2. The problem of over-fitting exacerbates as  $\mathcal{F}$  becomes ‘large’. If the set of rules is completely unrestricted, *i.e.*,  $\mathcal{F} = \mathbf{F}$ , the rule with the best

empirical fit,  $f_{s^t}^*$ , will track the data perfectly, giving little assurance that its performance is close to  $E_P f_P^*$ .

However, it can be shown that the over-fit term 2 satisfies:

$$\sup_P \int_{s^t} (\max_{f \in \mathcal{F}} E_P f - E_P f_{s^t}^{\mathcal{F}}) dP^t \leq 2\Delta_t(\mathcal{F}).$$

A decision maker can thus control the discrepancy between the performance of his chosen rule and the true best rule, by restricting himself to a class  $\mathcal{F}$  with ‘small’ empirical discrepancy  $\Delta_t(\mathcal{F})$ . Therefore, in this framework, coarse decision making arises when the decision maker must restrict the set of rules in order to control over-fitting. What this tells us about what must  $\mathcal{F}$  look like and how it depends on the environment is the subject of the next subsection.

However, we shall say little about how decision makers subjectively trade off fit for over-fit and select  $\mathcal{F}$ . One objective of future research is to provide a ‘decision theory’ of model selection.

### 3.2.3 Statistical Complexity and Vapnik-Chervonenkis Theory

So far we have talked loosely about  $\mathcal{F}$  being large or small, but we have not given a formal criterion for what this means. An elegant theory, originating with Vapnik and Chervonenkis (1971), characterizes the classes of  $\mathcal{F}$ ’s that can mitigate over-fitting. This is formally sketched in the appendix. Here we provide some intuition for the issues involved.

First we need to understand why  $\Delta_t(\mathcal{F})$  may be large, even though  $\Delta_t(f)$  is small for each  $f \in \mathcal{F}$ . The statement  $\sup_{f \in \mathcal{F}} \Delta_t(f)$  is small simply says that, fixing any  $f$ , the empirical estimate of the performance of  $f$  is a good indicator of its true performance on *most* samples. But to choose optimally within  $\mathcal{F}$ , a decision maker must use the sample he has to compare the performance of all rules in  $\mathcal{F}$ .

The fact that “the performance of  $f$ ,” has small estimation error for any  $f$  in  $\mathcal{F}$ , need not imply that “the performance of each  $f$  in  $\mathcal{F}$ ” has small estimation error. Roughly,  $\mathcal{F}$  is “statistically simple” if  $\Delta_t(\mathcal{F})$  must be small. Vapnik-Chervonenkis theory (herein VC theory) characterizes the structure of such  $\mathcal{F}$ ’s and thus relates overfitting to the amount of data available.

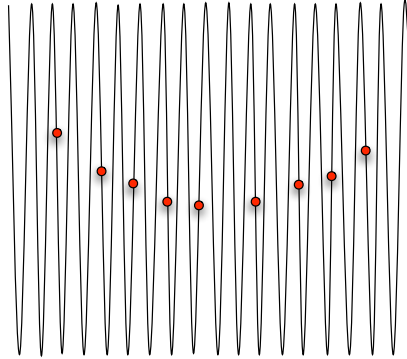


Figure 2: *Sine Curve*

We close by noting that statistical simplicity has at best a tenuous relationship to the cardinality of  $\mathcal{F}$ . The set of linear regressions  $\mathcal{F}_{\text{linear}}$  with a finite number of explanatory variables is statistically simple, even though its members are uncountable. On the other hand, consider the set:

$$\mathcal{F}_{\text{sin}} \equiv \{\sin(nx) : n = 1, 2, \dots\}.$$

This is a countable family of functions parameterized by the positive integer  $n$ . However, this can approximately fit any set of data in  $\mathfrak{R} \times [-1, 1]$ , as Figure 2 shows. The formal notion of statistical complexity is therefore more subtle.<sup>9</sup>

### 3.3 The Bayesian Methodology cannot Explain Coarse Decision Making

It may be useful at this point to compare the frequentist procedure  $(\mathcal{F}, \epsilon)$  with the standard Bayesian methodology. A Bayesian decision maker has a prior belief  $\pi$  over the set of data generating processes  $\mathcal{P}$ . He uses the sample  $s^t$  to update this prior to a posterior  $\pi(s^t)$  and selects the rule that

---

<sup>9</sup>For a brief, self-contained account, see Al-Najjar (2009). For textbook expositions, see Vapnik (1998) or Devroye, Györfi, and Lugosi (1996).

maximizes his expected payoff:

$$f \in \operatorname{argmax} E_{\pi(s^t)}[E_P f]$$

The inner expectation  $E_P f$  is the same as the one the frequentist uses. However, the Bayesian also takes expectations with respect to the posterior  $\pi(s^t)$ .

The frequentist, by contrast, takes a different approach to learning. Instead of assigning a prior  $\pi$ , he takes the approach of a classical statistician and considers a more restricted  $\mathcal{F}$ . Both the Bayesian and the frequentist make compromises: the Bayesian commits to a prior, while the frequentist commits to a frame  $\mathcal{F}$ . There is no ‘inductive’ free lunch! Since we do not aim to contribute to the decades-long doctrinal debate between different approaches to inference, the reader may take our modeling of decision makers as descriptive.

Crucially, the classical and Bayesian inductive principles can lead to strikingly different behaviors, with important economic consequences. A Bayesian decision maker would never choose to restrict himself to a strict subset of decision rules  $\mathcal{F}$ . Such a restriction cannot help—since expected utility maximization would pick the optimal rule. Further, this restriction will be sub-optimal if the maximizing rule with respect to  $\pi(s^t)$  happens to be outside  $\mathcal{F}$ . Intuitively, a Bayesian has no concern for over-fitting since he believes that his subjective belief  $\pi$  is ‘correct.’ In models with Bayesian decision makers, any restriction to a subset of rules  $\mathcal{F}$  must be due to exogenous constraints, such as wealth, technology, or information. For example,  $\mathcal{F}$  may be the set of rules measurable with respect to the agent’s information partition. This methodology cannot account for decision patterns such as rules of thumb or categorization, where the decision maker chooses from a restricted set of decision rules even though finer rules are informationally feasible.

## 4 Satisficing, Categorization, and Linear Orders

Our general model only points out the possible need to restrict the set of rules to some  $\mathcal{F} \subseteq \mathbf{F}$  to mitigate over-fitting, but does not point out the form such restrictions may take. Here we consider three phenomena of psychological and economic importance: categorization, linear orders, and satisficing. We show that these three phenomena can be understood as manifestations of coarse decision making.

We posit that the decision maker is faced with a treatment problem as defined in Section 2. As a normalization, we also require that  $\forall y \in Y, a \in A : u(y, a) \in [0, 1]$ .

### 4.1 A Formal Model of Categorization

Categorization is central in cognitive psychology and appears in numerous economic models. Categorization is also closely related to the ideas of similarity and analogy.

Intuitively, categories are groupings of objects or situations that are useful and informative in making judgements (Murphy and Medin (1985)). As seems evident from introspection and countless experiments, categorization is central in cognition: individuals tend to form judgements and make decisions based not on the identity of the object itself, but on the category it belongs to.<sup>10</sup>

We first formally define categorization within our framework:

**Definition 2** A categorization-based frame *consists of*:

- A categorization function:  $\kappa : X \rightarrow \{1, \dots, K\}$ ;
- A set of decision rules:

$$\mathcal{F}_\kappa = \left\{ f \mid f = g \circ \kappa, \text{ for some } g : \{1, \dots, K\} \rightarrow A \right\}$$

---

<sup>10</sup>The psychology literature on categorization is too vast to comprehensively cite. See for example the collections Vosniadou and Ortony (1989), and the papers by Reed (1972), Rosch and Lloyd (1976), Chi, Feltovich, and Glaser (1981), Rips (1989), Murphy and Medin (1985), Goldstone (1994), among many others.

Write  $X_k = \kappa^{-1}(k)$  to denote the  $k$ th category.

In a categorization-based frame, the decision maker first classifies the observables into one of  $K$  categories, then takes an action that depends only on the category. Note that any arbitrary decision rule  $f$  trivially defines a partition  $f^{-1}(y), y \in Y$ . What distinguishes categorization is the requirement that  $\mathcal{F}$  consists of all rules measurable with respect to one common categorization partition.

The next theorem shows that a decision maker who uses a categorization-based frame, and is concerned with robustness and overfitting, will necessarily rely on a *small* number of categories:

**Theorem 1** *For every  $t$  and  $\epsilon > 0$ , there exists an integer  $k^+$ , depending only on  $\epsilon$  and the amount of available data  $t$ , such that for every categorization function  $\kappa$ , with number of categories  $K$ ,*

$$\Delta_t(\mathcal{F}_\kappa) < \epsilon \implies K \leq k^+. \quad (6)$$

*Furthermore, for every integer  $k^- \leq \#X$  there exists  $T$  such that for every  $t \geq T$  and  $\epsilon > 0$ , there is a categorization rule  $\kappa$  with*

$$K = k^- \quad \text{and} \quad \Delta_t(\mathcal{F}_\kappa) \leq \epsilon. \quad (7)$$

If there are no constraints on the amount of available data, no coarse categorization arises. The decision maker can simply treat each singleton  $\{x\}$  as a separate category (thus, setting  $k^- = \#X$ ), in which case  $\mathcal{F}_\kappa$  coincides with the set of all rules  $\mathbf{F}$ , and still ensure that  $\Delta_t(\mathbf{F})$  is small.

The theorem has a bite when data is ‘scarce’. In particular, when  $k^+ \ll \#X$ , (6) says the decision maker *must* coarsely categorize. The intuition is that when data is scarce relative to the richness of the set of observables, an overly fine categorization will result in decision rules that overfit the data. Theorem 1 shows that such permissive classes of rules will have a large empirical discrepancy. A decision maker concerned with robustness will, as a result, choose a coarser categorization.

Categorization can be readily interpreted as a similarity relationship between instances, under which the decision maker categorizes all instances in

a component  $\kappa^{-1}(k) \subseteq X$  as similar (regardless of the sample realization). Under this interpretation, our theorem suggests that similarity is irrelevant when data is abundant. Non-trivial similarity arises only when data is scarce relative to the (statistical) complexity of the problem, in which case the decision maker will lump instances together into categories.

Of course, in any realistic problem, there is a virtually infinite variety of possible similarity structures, a problem that is well-recognized in the cognitive psychology literature. Murphy and Medin (1985) vividly highlight the inadequacy of undisciplined use of similarity by noting: “Suppose that one is to list the attributes that *plums* and *lawnmowers* have in common in order to judge their similarity. It is easy to see that the list could be infinite: Both weigh less than 10,000 kg (and less than 10,001 kg, ...), both did not exist 10,000,000 years ago (and 10,000,001 years ago, ...), both cannot hear well, both can be dropped, both take up space, and so on. Likewise, the list of differences could be infinite.” Our model therefore provides a formal argument why unfettered reliance on similarity in decision making is unlikely to arise: rational decision makers concerned with drawing robust inferences will recognize that such reliance will lead to severe overfitting of the scant data available to them.

We note, finally, that the theorem is silent on which specific categorization rule  $\kappa$  the decision maker should select. This is akin to the Bayesian decision model which restricts the form of the decision procedure (namely, to maximize expected utility with respect to a subjective prior), but is silent on which prior decision makers adopt. The statistical learning approach does not provide a recipe for which specific rule should be selected, but an inductive *principle*, a framework for selecting decision rules, whose implications are to be contrasted with alternatives like the Bayesian model. We comment further on this in Section 4.4.

## 4.2 Linear Orders

Another fundamental cognitive phenomenon is the apparent reliance of decision makers on linear orders. By this we mean organizing the state space by embedding it in a multi-dimensional Euclidean space (with the usual par-

tial order). Numerous examples illustrate this phenomena. In management, several companies base their management decisions on what are known as ‘balanced scorecards,’ where various factors potentially affecting the decision are assigned a numerical score, and the decision is made based on a weighted average of the scores.<sup>11</sup> Similarly, several investment funds base their strategy on investing in companies with certain financial parameters, and advertise themselves as such. For example, ‘value investing’ funds invest in stocks that have high dividend yields, low price-to-earning multiples or low price-to-book ratios. A very different illustration can be found in statistical practice with the prevalence of linear regressions, despite the availability of several more ‘sophisticated’ estimators.

The importance of linear orders received attention in economics through the works of Rubinstein ((1996), (2000)) and Kalai (2003) among others. These works cite the psychological literature on the subject. In this section we use our framework to explain why linear structures emerge to aid decision making, even when the underlying space of observables and outcomes has no such structure intrinsically.

A linearization of the observables is a function  $v : X \rightarrow \mathcal{R}^n$  that embeds the abstract set of observables  $X$  into the Euclidean space  $\mathcal{R}^n$ . For example, if  $x$  represents a particular individual, the vector  $v(x)$  may denote  $n$  linearly ordered observable attributes of the individual, such as his height, age, number of years of schooling, and so on. We define linearization on outcomes similarly, for expositional simplicity, we require a one dimensional attribute  $w : Y \rightarrow \mathcal{R}$ .

The definitions above are vacuous if we allow a representation of high enough dimensionality ( $n$  to be arbitrarily large). Intuitively, when we think of decision makers as ‘organizing their environment along linear dimensions,’ we also mean that the dimensionality of the space is small. This is what we formalize below.

**Definition 3** *A linear attribute frame  $(v, w)$  is a pair of functions:*

$$v : X \rightarrow \mathcal{R}^n \quad \text{and} \quad w : A \rightarrow \mathcal{R},$$

---

<sup>11</sup>For more details on balanced scorecards and their prevalence in management, we refer the interested reader to Kaplan and Norton((1992), (1996)).

and the collection of all monotone decision rules:

$$\mathcal{F}_{v,w} = \left\{ f \mid v(x') \geq v(x) \implies w(f(x')) \geq w(f(x)) \right\}.$$

The next theorem shows that, assuming that the decision maker chooses a linear attribute frame, statistical learning implies that he must limit attention to a *small* set of dimensions to mitigate the problem of overfitting:

**Theorem 2** *For every  $t$  and  $\epsilon > 0$ , there is an integer  $n^+(\epsilon, t)$  such that for any linear attribute frame  $(v, w)$ ,  $\Delta_t(\mathcal{F}_{v,w}) < \epsilon$ , implies*

$$n \leq n^+(\epsilon, t).$$

The statement is analogous to that of Theorem 1 and the strategy of proof is similar. The interpretation, of course, is quite different. Categorization corresponds to partitions and thus resembles, at least superficially, models of incomplete information. With linear orders, the set of decision rules  $\mathcal{F}_{v,w}$  do not correspond to a partition on the space of observables  $X$ . Rather they reflect an order relationship which, in turn, has no counterpart for categorization problems.

In a rich environment, two instances  $x$  and  $x'$  can be compared on a virtually infinite number of dimensions. The theorem implies that the decision maker must ‘focus,’ via the function  $v$ , on a small number of dimensions. Notably, limiting attention to a few dimensions is not a result of bounded rationality, computational complexity, and the like, but a reflection of the decision maker’s concern for overfitting and desire for robust decision rules.

A model with a similar flavor is that of *rational inattention*, due to Sims (2003).<sup>12</sup> Using information theoretic concepts, Sims models the idea that individuals have limited information processing capability. As a result, they may rationally decide not to pay attention to all available information. This model has been applied to explain price and wage rigidities. The details of our model are quite different from Sims but, at least qualitatively, our conclusions bear some similarity. A decision maker who desires robustness in

---

<sup>12</sup>For examples of applications, see Moscarini (2004) and Mackowiak and Wiederholt (2009).

our sense will restrict the number of dimensions of the problem he considers, and thus will behave as if he is inattentive to other dimensions of information though they are available to him. Note that the source of inattentiveness in our model is the scarcity of data and the desire for robustness— unlike Sims’ framework, our decision makers do not suffer information processing constraints.

### 4.3 Satisficing

Simon (1955) proposed the idea of *satisficing* whereby a decision maker uses a plan which, while suboptimal, represents an attempt to do ‘reasonably well.’ Simon specifically proposed that a decision maker uses ‘coarse’ utility functions, in the sense that he treats as equivalent plans that meet a given threshold utility level (his satisfaction point).

Simon’s motivation is that real-world decision makers are unable to figure out the optimal plans due to computational complexity.

*We see that, by the introduction of a simple pay-off function ... the process of reaching a rational decision may be drastically simplified from a computational standpoint. (...) If, instead of requiring that the pay-off be maximized, we require only that the pay-off exceed some given amount, then we can find a program that satisfies this requirement by the usual methods of feasibility testing.*

He also refers to a reduced cost of information gathering in his setting from such coarsening.

*If the information-gathering process is not costless, then one element in the decision will be the determination of how far the mapping is to be refined. Now in the case of simple payoff functions (...), the information gathering process can be streamlined in an important respect.*

We do not question the role of computational considerations as possible sources of satisficing-type behavior. What we propose is that similar behavior can also arise as an instance of coarse decision making using the framework of this paper. As we discuss in Section 4.4, this complementary explanation generates interesting implications not possible when computational complexity is the sole force behind satisficing.

To make these ideas formal, we consider a forecasting problem with:

$$A = Y = \left\{ 0, \frac{1}{k}, \frac{2}{k}, \dots, 1 \right\},$$

and a payoff function given by the usual distance  $u(y, a) = -|y - a|$ . We consider a linear attribute frame, as in Section 4.2, with a one-dimensional linear order  $v$  on  $X$ , and the standard ordering on  $Y$ . Thus, the decision maker considers monotone decision rules:

$$\mathcal{F} = \left\{ f \mid v(x') \geq v(x) \implies f(x') \geq f(x) \right\}^{13}$$

Restricting attention to the one-attribute case abstracts from issues arising in the multi-dimensional linear orders discussed in the last section. This will help us focus on the role of satisficing instead.

When  $k$  is ‘large’ relative to the available data, the empirical discrepancy  $\Delta_t(\mathcal{F})$  of the class of rules  $\mathcal{F}$  is large, and the decision maker runs the risk of over-fitting the data. Satisficing then arises as a learning strategy to overcome this problem. Specifically, the decision maker limits his predictions to a subset  $A' \subseteq A$ ,  $|A'| = k' < k$ . This gives rise to the set of decision rules:

$$\mathcal{F}_{k'} = \left\{ f \mid f : X \rightarrow A'; v(x') \geq v(x) \implies f(x') \geq f(x) \right\}.$$

Limiting predictions to  $A'$  corresponds to a ‘coarsening’ of the utility function from  $u$  to  $u' : Y \times A' \rightarrow \Re$  ( $u'(y, a') = u(y, a')$ ). Fix  $y$ , and  $a \in A/A'$  s.t.  $u(y, a) > \max_{a' \in A'} u(y, a')$ . By only considering decision rules in  $\mathcal{F}_{k'}$ , the decision maker is effectively ignoring the possible additional payoff  $u(y, a) - \max_{a' \in A'} u(y, a')$ . This better fitting action,  $a$ , is available to him in  $\mathcal{F}$ , but he chooses not to consider this to avoid over-fitting. Thus, under satisficing, the decision maker is content with making a coarse approximation of  $y$ , rather than pinning its value precisely.

We can now state our main theorem:

**Theorem 3** *For every  $t$  and  $\epsilon > 0$ , there is an integer  $k^+(\epsilon, t)$  such that for any  $k$ , and any satisficing frame  $\mathcal{F}_{k'}$  ( $k' \leq k$ ):  $\Delta_t(\mathcal{F}_{k'}) < \epsilon$  only if:*

$$k' \leq k^+(\epsilon, t).$$

---

<sup>13</sup>The set  $\mathcal{F}$  depends on the index of the the grid of utilities  $k$  and the linear order  $v$ . We suppress reference to  $k$  and  $v$  since they will be fixed throughout.

Further, there exists  $\bar{\epsilon} > 0$  such that for every  $t$  and  $\epsilon \leq \bar{\epsilon}$ , for  $k$  large enough,  $\Delta_t(\mathcal{F}) > \epsilon$ .

To link this with Simon’s original motivation, we consider his example where  $x$  is a vector of inputs, say the effort and resources applied by a firm’s management, and  $y$  as the output resulting from applying  $x$ . Let  $\phi : X \rightarrow Y$  denote the true relationship between inputs and outputs. The firm does not know  $\phi$ , and would like as accurate an estimate of  $\phi$  as possible. That is, the firm chooses an  $f \in \mathcal{F}$  so the difference between  $f(x)$  and  $\phi(x)$ , averaged over all  $x$ ’s, is small. A firm can achieve this by limiting itself to ‘coarse’ predictions of the output.

For example, if  $Y = \{0, 1, 2, \dots, 1,000,000\}$ , the firm might want to satisfice and only predict to the nearest 1000, i.e.  $A' = \{0, 1000, \dots, 1,000,000\}$ . The coarsening limits the ability to overfit, and hence provides more robust predictions. The tradeoff is that this coarsening forces small prediction errors—in our example, even if the firm’s best prediction is 500, it must predict either 1000 or 0. Our result shows that the firm is potentially willing to accept these small errors in return for doing ‘reasonably well’. This is similar to the intuition of Simon—albeit in a learning rather than optimization setting.

## 4.4 Discussion

Instances of coarse decision making are ubiquitous in economics and cognitive psychology. Phenomena that involve categorization, similarity, rules of thumb, or satisficing are too numerous to comprehensively list here.

While no formal model is needed to assert that people engage in such behavior, why they do so is less obvious and is potentially the more important question. Do individuals categorize or satisfice because of computational complexity, limited memory, or lack of information? Understanding what drives coarse decision making may not be important for descriptive accounts of its various manifestations. But such understanding is essential if we want to even begin to answer questions like: Are these disparate phenomena, or are they related? How would these “biases” change as a function of the decision maker’s environment? And are these persistent aspects of decision making,

or transient phenomena likely to disappear as decision makers engage in learning, introspection, and competitive interaction?

*Bounded rationality as a source of coarse decision making:* Much of the existing literature views coarse decision making as an expression of “bounded rationality:” a decision maker suffers a cognitive and computational limitations that restricts him to decision rules that condition only on coarse features of the problem, even though finer rules are available. Several sources of such limitation have been proposed: memory limitations (imperfect and bounded recall, limited number of memory states), computational limitations (automata, turing machines, perceptrons) and contemplation costs, among many others.

A perennial problem with the bounded rationality approach is that we have a poor understanding of how to model decision makers’ cognitive limitations. The psychology literature offers a bewildering variety of (often incompatible) experimental findings of biases and heuristics. Likewise, there is a similar variety of computational models, none of which seems to reflect particularly well how human decision makers make choices. Since assumptions about the nature and details of these cognitive and computational limitations are critical for the conclusions of bounded rationality models, the result is all too often a patchwork of separate independent models, each tailored to fit a particular phenomenon.

A model based on statistical learning theory, by contrast, proposes a unified explanation of what drives coarse decision making, namely the desire to mitigate over-fitting when data is scarce. While a rich enough collection of ad hoc models will necessarily fit observed phenomena better than one based on statistical learning, the latter has the potential of offering an improved modeling discipline and insightful comparative statics.

*The comparative statics of coarse decision making:* Models of coarse decision making based on cognitive and computational limitations suggest that as the stakes in the decision problem increase, decision makers will devote greater efforts to overcome these limitations, ending up with progressively finer decision rules. This implies that things like categorization or rules of

thumb should diminish in importance as the stakes become large. This seems at odds with findings that coarseness persists even in ‘important’ decisions. For instance, phenomena we discussed earlier, such as style investing and the use of balanced scorecards by firms are clearly ones that affect ‘large stakes’ decisions.

An approach based on statistical learning offers different comparative statics. It implies that there is no *direct* connection between coarseness of decision rules and things like the monetary stakes agents have in the outcome, their depth of introspection or contemplation costs.<sup>14</sup> Thus, in problems with abundant data and simple outcome spaces, agents use fine decision rules regardless of how high or low the stakes are. On the other hand, coarse rules of thumb may continue to hold great appeal even in high-stakes decision problems, not because decision makers have not had ample opportunity to contemplate and introspect, but because of the limited data they have relative to the statistical complexity of the underlying problem.

The statistical learning model therefore suggests that one should not expect systematic relationship between the payoff-importance of decision problems and the refinement of decision rules used. Instead, it predicts a systematic connection between the statistical complexity of the set of feasible rules and the amount of data available.

*Coarse decision making vs. incomplete information:* A natural approach to study cognitive limitations is what one may refer to as *partitional models*, where a decision maker is assumed to be constrained by a partition that reflects his coarse or limited understanding of the environment. Many bounded rationality models, (involving analogies, similarities, memory limitations, or bounded recall) are of the partitional variety. The attractiveness of partitional models is obvious: they frame bounded rationality in the familiar language of informational problems.

Viewing bounded rationality as another instance of incomplete information, although comforting, can be misleading. Lack of information is an

---

<sup>14</sup>Although *indirect* connections are possible; for example, if we expand our model to include experimentation or data gathering then higher stakes may induce the agent to gather more data. Our point is that the only *direct* channel of causality here is data availability.

exogenous objective constraint, whereas bounded rationality has to do with information processing. The latter is inherently more nebulous, constantly changing with learning, introspection and competitive pressures. A methodology based on statistical learning offers a potentially fruitful alternative. In the categorization setting, for example, a decision maker who willingly limits his search to a subset of the set of feasible rules would display symptoms of bounded rationality, such as the coarse lumping of outcomes or the reliance on similarity and analogies. A Bayesian theorist, who must appeal to exogenous constraints like computational cost or cognitive limitations to explain this behavior, has few alternatives besides declaring such behavior boundedly rational. A decision maker concerned with drawing robust inferences from scarce data, on the other hand, will find it natural to limit his search to a strict subset of rules. This is not because he is not smart enough or the rules are too complicated, but because admitting such rules will overfit the data. There is no need to appeal to labored informational stories, ad hoc criteria of complexity and simplicity, or speculative accounts of the decision maker’s cognitive limitations.

## 5 Categorization and Heterogeneity

As an application of our framework, we consider a simple model with a principal and 2 agents who use models, or decision frames, in the sense of Definition 1, to learn from past data. Our results in this section can be thought of as an operationalization of Kreps’s desiderata on corporate culture Kreps (1990). To quote Kreps:

*“Consistency and simplicity being virtues, the culture/principle will reign even when it is not first best ... will be taken into areas where it serves no purpose except to communicate or reinforce itself ... that is to communicate the principle, we administer it repeatedly so that others learn it.”*

The setup is that of a principal who wants his two agents to coordinate on the same contingent action. As in the basic model, we assume a large finite space of business situations or problems, denoted  $X = \{x_1, \dots, x_N\}$ . For

simplicity, we assume that a problem is drawn at random from  $X$  according to a known probability distribution  $\pi$ .

At each  $x \in X$ , one of two actions  $\{0, 1\}$  can be taken.<sup>15</sup> For any given situation, the principal has a preferred action that he would like to see agents undertake. We denote the preferred action at situation  $x$  as  $A^*(x)$ . When a situation  $x$  realizes, the two agents must each take an action- if both agents take action  $A^*(x)$ , then the principal gets a payoff of 1, in any other event he gets a payoff of 0.

The trouble is that neither agent knows the preferred action mapping of the principal. Instead, they each get to view a history of  $t$  past situations, and the (correct) action that the firm took in each, i.e.  $(x_i, A^*(x_i)), i = 1 \dots t$ . In keeping with past notation, we denote this past history as  $s^t \in (X \times \{0, 1\})^t$ . From this they attempt to reconstruct the principal's preferred action mapping  $A^*$ .

For concreteness, we assume that both agents use categorization-based frames as in Definition 2. Agent  $i$  partitions the space of situations as  $\mathcal{C}_i$ . Critically, the partitions of the 2 agents may be different. Therefore agent  $i$ , on viewing past history  $s^t$ , for any category  $C_i^k \subseteq X$  will pick action 0 if

$$|\{i : s_i = (x, 0), x \in C_i^k\}| \geq |\{i : s_i = (x, 1), x \in C_i^k\}|,$$

and action 1 otherwise. In words, for each category (partition element), the agent will take the action that is more prevalent in the history corresponding to that category.

We will refer to the 4-tuple of  $\pi$ , the distribution over business situations, the preferred action mapping for the principal  $A^*$ , and the partitions for the two agents,  $\mathcal{C}_1$  and  $\mathcal{C}_2$ , as the environment.

We may interpret  $\mathcal{C}_i$  as the set of categories agent  $i$  uses to interpret past evidence, in the sense that two situations that belong to the same category are lumped together as similar. We study two comparative statics on this basic model.

First, consider a case where the principal has the flexibility to change his preferred action (at a cost). If the principal changes his preferred action to

---

<sup>15</sup>It will be clear that the assumptions that there are two actions and two agents do not play an important role in the qualitative point we make.

$A'$ , the agents observe a history of  $(x_i, A'(x_i)), i = 1 \dots t$ . In particular, he would rather change his preferred action than have the agents miscoordinate or end up on the wrong action. He can change his rule  $A^*$  to some other  $A'$ . However if the agents co-ordinate on  $A'(x) \neq A^*(x)$  for some  $x$ , he gets a payoff of  $\alpha < 1$ . We show that the principal may prefer to change his preferred actions to an  $A'(\cdot)$  that is measurable with respect to the meet of the two partitions.

**Proposition 4** *Suppose the principal can change his preferred action for any business situation  $x$ . If he does so he only gets a benefit  $\alpha < 1$  if agents co-ordinate on it. There exist environments  $(\pi, A^*, \mathcal{C}_1, \mathcal{C}_2)$ , such that for any  $\alpha \in [0, 1)$ , the principal would prefer to change his preferred actions to an  $A' : X \rightarrow \{0, 1\}$  where  $A'$  is measurable with respect to  $\mathcal{C}_1 \wedge \mathcal{C}_2$ .*

Second, suppose the principal maintains  $A^*$ , but has the ability to ‘train’ the agents to refine their partitions. For simplicity, we consider that he will re-train them to consider the finest possible partition, i.e. where each partition element has exactly one business situation. We show that he may prefer not to train the agents even if such training is costless. Roughly speaking, under the refined frame, agents have fewer training samples per partition element. As a result there may be more sampling error, and they may ‘learn’ the incorrect decision on more partitions. <sup>16</sup>

**Proposition 5** *Suppose the principal can costlessly refine the partitions employed by the two agents to  $\mathcal{C}^*$ , the finest possible partition. Then there exists an environment  $(\pi, A^*, \mathcal{C}_1, \mathcal{C}_2)$  and an integer  $n^-(t)$  such that if the number of business situations  $N \geq n^-$ , the principal would strictly prefer not to refine the partitions.*

We relate this results to a recent paper by Crémer, Garicano, and Prat (2007). Their paper studies properties of (coarse) ‘corporate codes,’ which emerge when agents are boundedly rational and can only remember at most  $K$  words. The implied welfare result in their paper is that it would be overall

---

<sup>16</sup>It will be clear from the proof that the result remains true for any suitably ‘fine’ refinement of the partitions. We use the finest possible partition for expository simplicity.

welfare improving if agents could learn more words, since they would be able to partition the state space more finely. However, they assume that agents know the mapping between the code and the implied events. Our result can be thought of modeling a situation where things like ‘codes’ in organizations predate the agents they hire, and these agents need to learn them. In this case, despite there being no restriction on the richness of codes an agent can learn, a principal designing the code may want to restrict the code used to ease the agents’ learning. Further, even if the agents use partitions that are too coarse for them to learn the first best code, the principal may not want to refine these partitions.

## 6 Structure, Identification, and the i.i.d. Assumption

### 6.1 Absence of Built-in Structure

We make a point of *not* requiring  $X$  to have any *a priori* presumed structure. For example,  $X$  need not have linear ordering, lattice or any other structure. The reader may find this at odds with typical economic or game theoretic problems where information and choice variables are rarely introduced as abstract ‘sets.’

The reason for our modeling choice is that our goal is to understand how structures like categorization or linear orders might arise from decision makers’ attempts to deal with the complexity of their environment. In the case of categorization, for instance, our model identifies forces that can lead decision makers to categorize (namely, the desire to mitigate over-fitting). This will, hopefully, help understand why various categorizations might emerge, and how they might vary with the fundamentals. A structure-free environment makes it possible to model this cleanly by ensuring that no exogenous structure outside the decision makers’ choices contaminates the analysis. It should be apparent that our analysis can be extended to settings where exogenous structures are assumed.

## 6.2 Identification Problems

We have assumed that observations of the outcome  $y$  are generated by  $P$  independently of the action taken. This makes sense in example like stock market investing where the price of the stock is independent of the investment strategy of a (small) investor. What this rules out is the important problem of *identification in treatment response* studied in the Econometrics literature (see, for instance, Manski (2008)).

As an example, suppose that  $x$  represents the characteristics of a worker,  $y$  his income level, and  $a$  his level of education. A ‘natural’ sample consists of observations of workers’ characteristics and their income levels *given* the actual level of education they had. What is not observed is the counterfactual: what would income be had these workers been given different education levels. For example, if a policy is implemented that ensures a common level of education  $a_1$  to all workers, then one cannot tell from a sample, no matter how large, what the distribution of income would be under an alternative policy with education level  $a_2$ .

We can map the identification problem to our setting by using  $Y^A$ , instead of  $Y$ , as outcome space. With  $K$  actions, say, an outcome now is a vector  $(y_{a_1}, \dots, y_{a_K})$  and an environment is a probability distribution  $P$  on  $X \times Y^A$ . Our analysis goes through unaltered if an observation is defined as  $(x, (y_{a_1}, \dots, y_{a_K}))$ . This, however, means that we observe the outcome not only at the action (or treatment) actually taken, but the counterfactuals of what would happen under different actions. The identification problem arises when all that is observed  $(x, y_{a_k})$ , where  $a_k$  is the action actually taken, so  $y_{a_l}$ ,  $l \neq k$  is not observed.

The identification problem above is orthogonal to the problem of statistical complexity that is our main focus. Under lack of identification the problem is not the amount of data available, but that observed data does not reveal the counterfactuals. In this case, learning may fail even with infinite amount of data.

### 6.3 The i.i.d. Assumption

An apparent limitation of our analysis is that observations are generated in an i.i.d. manner. One could argue that, although the i.i.d. case is obviously important, many problems of interest involve correlation and/or a stochastic structure that changes over time.

One response is obvious: the i.i.d. case is important and seems like a good starting point. A more subtle argument consists of viewing i.i.d.'ness as a completeness assumption. By saying that “when the distribution  $P$  is repeatedly sampled, the samples are i.i.d.,” we mean that if the same vector of observables  $x$  is seen again, the distribution of  $y$  remains the same and is independent of all other sample information. But this may be interpreted to mean that  $x$  is an exhaustive description of all the factors that can influence  $y$ . Thus, we can conceive of the i.i.d. assumption as saying that any correlation or time changes are already reflected in  $x$ .

Stated differently, fixing the space of observables  $X$ , the i.i.d. assumption is indeed quite restrictive. But if we think of a decision maker who wants to build a subjective frame to help him understand his environment, then the i.i.d. assumption is just saying that the decision maker’s frame is complete. This interpretation, if accepted by the reader, lends additional credibility to our arguments, since a space  $X$  that reflects an exhaustive description of the problem, and the corresponding set of feasible decision rules  $\mathbf{F}$ , will likely be very large. This makes the case for restricting to a subset of rules  $\mathcal{F}$  even more compelling.

## A A Technical Primer to VC-theory

This primer contains definitions and key theorems that we will need for the proofs of our propositions. For a fuller treatment, including proofs of the theorems and intuition, we refer the reader to Vapnik (1998) or Haussler (1995).

### A.1 Vapnik Chervonenkis Dimension

Consider a set  $X$  and a set of subsets of  $X$ ,  $\mathcal{C} \subseteq 2^X$ . We say that  $\mathcal{C}$  *shatters*  $(x_1, x_2, \dots, x_d) \in X^d$  if for each  $b = (b_1, \dots, b_d) \in \{0, 1\}^d$  there exists  $C_b \in \mathcal{C}$  such that:

$$x_i \in C_b \iff b_i = 1.$$

Therefore,  $\mathcal{C}$  shatters  $(x_1, x_2, \dots, x_d)$  if each subset can be contained in some member of  $\mathcal{C}$ .

**Definition 4** *The Vapnik-Chervonenkis dimension of  $\mathcal{C}$ ,  $VC(\mathcal{C}) = d$  if there exists  $(x_1, x_2, \dots, x_d) \in X^d$  such that  $\mathcal{C}$  shatters  $(x_1, x_2, \dots, x_d)$ , and there does not exist any  $(x_1, x_2, \dots, x_d, x_{d+1}) \in X^{d+1}$  such that  $\mathcal{C}$  shatters  $(x_1, x_2, \dots, x_d, x_{d+1})$ .*

*In other words, the VC dimension of  $\mathcal{C}$  is the length of the longest string it can shatter. If  $\mathcal{C}$  can shatter strings of arbitrary length, we say its VC-dimension is infinity.*

A central result in statistical learning theory is that a class of events  $\mathcal{C}$  is uniformly learnable if and only if it has finite VC-dimension.

**Theorem 6** *Consider a set  $X$ , and  $\mathcal{C} \subseteq 2^X$ . Suppose the VC dimension of  $\mathcal{C}$  is  $d$ . Then for any  $\epsilon > 0$ , and any integer  $t > 0$ :*

$$\sup_P P^t \left\{ s^t : \sup_{A \in \mathcal{C}} |\nu(s^t)(A) - P(A)| > \epsilon \right\} \leq K t^d e^{-t\epsilon^2/32}, \quad (8)$$

where  $K$  is a universal constant.<sup>17</sup>

---

<sup>17</sup>Tighter bounds are available, but the above version is sufficient for our purposes, see also Devroye, Györfi, and Lugosi (1996).

In order to see how this impacts our setting, consider the simplest possible version of our model-  $X$  is some finite set,  $Y = A = \{0, 1\}$  and

$$u(y, a) = \begin{cases} 1 & \text{if } y = a, \\ 0 & \text{otherwise.} \end{cases}$$

The set of all possible decision rules  $\mathbf{F} = \{f | f : X \rightarrow \{0, 1\}\}$ , and suppose the decision maker considers  $\mathcal{F} \subseteq \mathbf{F}$ . For any  $f \in \mathcal{F}$ , and any true probability distribution  $P$ :

$$E_P u(y, f(x)) = P((f^{-1}(0) \times \{0\}) \cup (f^{-1}(1) \times \{1\})).$$

Define  $\mathcal{X} = X \times \{0, 1\}$  and  $\mathcal{C} = \{A | A = (f^{-1}(0) \times \{0\}) \cup (f^{-1}(1) \times \{1\}), f \in \mathcal{F}\}$ . Therefore, it follows from Theorem 6 (see also Corollary 12.1 of Devroye, Györfi, and Lugosi (1996)):

$$\Delta_t(\mathcal{F}) \leq 16 \sqrt{\frac{V_{\mathcal{C}} \log t + 4}{2t}}.$$

## A.2 Pollard's Pseudo-Dimension

In the case of a more general  $Y, A, u$ , the Vapnik-Chervonenkis bounds do not directly apply. Various notions of dimension in the general setting are reviewed in Bendavid, Cesabianchi, Haussler, and Long (1995).

We will use Pollard's Pseudo dimension, sometimes referred to in the literature as the Pollard dimension (Pollard (1990)). Let  $\mathcal{F}$  be some set of functions from  $X$  to  $\mathfrak{R}$ . We say that  $\mathcal{F}$  *pseudo-shatters* a string  $(x_1, x_2, \dots, x_d)$  if there exists  $c = (c_1, \dots, c_d) \in \mathfrak{R}^d$  such that for each  $b = (b_1, \dots, b_d) \in \{0, 1\}^d$ , there exists  $f_b \in \mathcal{F}$  satisfying: <sup>18</sup>

$$\forall 1 \leq i \leq d : f_b(x_i) > c_i \text{ iff } b_i = 1.$$

**Definition 5** *The Pseudo-Dimension of  $\mathcal{F} = d$  if there exists  $(x_1, x_2, \dots, x_d) \in \mathcal{X}^d$  such that  $\mathcal{F}$  pseudo-shatters  $(x_1, x_2, \dots, x_d)$ , and there does not exist any  $(x_1, x_2, \dots, x_d, x_{d+1}) \in \mathcal{X}^{d+1}$  such that  $\mathcal{F}$  pseudo-shatters  $(x_1, x_2, \dots, x_d, x_{d+1})$ .*

<sup>18</sup>The literature uses the term shatter in this setting as well. We refer to the concept as pseudo-shattering to remove any ambiguity.

In other words, the Pseudo-dimension of  $\mathcal{F}$  is the length of the longest string it can pseudo-shatter. If  $\mathcal{F}$  can pseudo-shatter strings of arbitrary length, we say its Pseudo-dimension is infinity.

The following inequality is Corollary 2 of Haussler (1995) restated in our notation:

**Theorem 7** Consider a set of real-valued functions  $\mathcal{F}$  of bounded range  $[0, M]$ . Suppose the pseudo dimension of  $\mathcal{F}$  is  $d$ . Then for any  $\epsilon > 0$ , and any integer  $t > 0$ :

$$\sup_P P^t \left\{ s^t : \sup_{f \in \mathcal{F}} |E_{\nu(s^t)} f - E_P f| > \epsilon \right\} \leq 8 \left( \frac{32eM}{\epsilon} \ln \left( \frac{32eM}{\epsilon} \right) \right)^d e^{-\epsilon^2 t / 64M^2}. \quad (9)$$

## B Proofs of Theorems in Section 4

In preparations for the proofs, we will need a couple of preliminary lemmas.

**Lemma B.1** Suppose a non-negative random variable  $Z$  satisfies

$$\forall \epsilon > 0 : \mathbb{P}(Z > \epsilon) \leq c\epsilon^{-2d} e^{-k\epsilon^2}.$$

for some  $c, d, k \geq 1$ ,  $\ln ck > 1$ . Then:

$$\mathbb{E}(Z) \leq \sqrt{\frac{d \ln ck + 1}{k}}.$$

**Proof of Lemma B.1:** Since, for all  $\epsilon > 0$

$$\mathbb{P}(Z > \epsilon) \leq c\epsilon^{-2d} e^{-k\epsilon^2},$$

we have that:

$$\begin{aligned}
\mathbb{E}(Z^2) &= \int_0^\infty P(Z^2 > t) dt \\
&= \int_0^u P(Z^2 > t) dt + \int_u^\infty P(Z^2 > t) dt \quad \forall u > 0 \\
&\leq u + \int_u^\infty P(Z^2 > t) dt \\
&\leq u + \int_u^\infty ct^{-d} e^{-kt} dt \\
&\leq u + cu^{-d} \int_u^\infty e^{-kt} dt \\
&= u + u^{-d} \frac{c}{k} e^{-uk}
\end{aligned} \tag{10}$$

Plugging  $u = \frac{d \ln ck}{k}$  into inequality (10), we have:

$$\begin{aligned}
\mathbb{E}(Z^2) &\leq \frac{d \ln ck}{k} + \left( \frac{d \ln ck}{k} \right)^{-d} \frac{c}{k} \frac{1}{(ck)^d} \\
&= \frac{d \ln ck}{k} + \frac{1}{k} \frac{1}{(d \ln ck)^d c^{d-1}} \\
&\leq \frac{d \ln ck + 1}{k}. \quad (\ln ck \geq 1)
\end{aligned}$$

Finally note that

$$\mathbb{E}(Z) \leq \sqrt{\mathbb{E}(Z^2)}$$

by Jensen's inequality, giving us the desired result. ■

**Lemma B.2** *Suppose a set of real valued functions  $\mathcal{F}$  is such that for each  $f \in \mathcal{F}$ ,  $\text{range}(f) \subseteq [0, 1]$ . If the pseudo dimension of  $\mathcal{F}$  is less than  $d$ , then:*

$$\Delta_t(\mathcal{F}) \leq 8 \sqrt{\frac{d^2 \ln 32e + d \ln \frac{te}{8}}{t}}. \tag{11}$$

**Proof of Lemma B.2:** By Pollard's inequality, (9):

$$\begin{aligned}
\sup_P P^t \left\{ s : \sup_{f \in \mathcal{F}} |E_{\nu(s^t)} f - E_P f| > \epsilon \right\} &\leq 8 \left( \frac{32e}{\epsilon} \ln \left( \frac{32e}{\epsilon} \right) \right)^d e^{-\frac{\epsilon^2 t}{64}} \\
&\leq 8 \left( \frac{32e}{\epsilon} \right)^{2d} e^{-\frac{\epsilon^2 t}{64}} \\
&= (8(32e)^{2d}) \epsilon^{-2d} e^{-\frac{\epsilon^2 t}{64}}.
\end{aligned}$$

Then (11) follows from Lemma B.1. ■

Next, recall that the Pollard's pseudo-dimension applies to real-valued functions. Given the decision maker's utility function  $u$ , any decision rule  $f : X \rightarrow A$  induces a real valued function  $u_f : X \times Y \rightarrow \mathfrak{R}$ ,  $u_f(x, y) = u(y, f(x))$ ; and therefore  $\mathcal{F}$  induces a set of real valued functions  $\mathcal{U}_{\mathcal{F}}$ .

In the sequel, given a utility function  $u$ , we shall abuse notation by referring the pseudo dimension etc. of  $\mathcal{F}$  directly, instead of the induced set of real valued functions  $\mathcal{U}_{\mathcal{F}}$ .

**Lemma B.3** *Let  $\kappa : X \rightarrow \{1, \dots, K\}$  be a categorization rule, and  $\mathcal{F}_{\kappa}$  be the associated categorization-based frame. For any utility function  $u : Y \times A \rightarrow \mathcal{R}$ , the pseudo dimension of  $\mathcal{F}_{\kappa}$  is at most  $K|Y|$ .*

**Proof of Lemma B.3:** We need to show that there is no string in  $(X \times Y)^{K|Y|+1}$  that  $\mathcal{F}_{\kappa}$  can pseudo-shatter. We show that for any  $1 \leq k \leq K$ ,  $\mathcal{F}_{\kappa}$  can pseudo-shatter at most  $|Y|$  elements in  $(\kappa^{-1}(k) \times Y)$  (the desired lemma clearly follows).

So suppose not. Fix  $k$ , and consider any  $|Y| + 1$  elements  $(x_i, y_i) \in (\kappa^{-1}(k) \times Y)$ ,  $i = 1, \dots, |Y| + 1$ . Let the associated cutoffs be  $c_i \in \mathfrak{R}$ ,  $i = 1, \dots, |Y| + 1$ , without loss of generality let  $c_1 \leq c_2 \leq \dots \leq c_{|Y|+1}$ .

By the Pigeon Hole Principle, there must be two elements  $(x_i, y_i), (x_j, y_j)$ ,  $i < j$  such that  $y_i = y_j$ . However, since  $x_i, x_j \in \kappa^{-1}(k)$ ,  $f(x_i) = f(x_j)$  for all  $f \in \mathcal{F}_{\kappa}$ . Hence,  $u_f(x_i, y_i) = u_f(x_j, y_j)$  for all  $f \in \mathcal{F}_{\kappa}$ . Clearly, there cannot exist  $f \in \mathcal{F}_{\kappa}$  such that  $u_f(x_j, y_j) > c_j$  and  $u_f(x_i, y_i) \leq c_i$ , and therefore  $\mathcal{F}_{\kappa}$  cannot shatter it. ■

We can now proceed to the proofs of the theorems in the paper.

**Proof of Theorem 1:** We first prove the former part of the theorem, i.e. (6). Note that since there are more than 2 actions and  $u$  is a real valued function, VC-theory does not directly apply. Our first step is to effectively reduce the number of outcomes to 2.

Let  $\delta = \min_{y \neq y'} |u(y, y) - u(y, y')|$  and let  $y_1, y_2 = \arg \min_{y \neq y'} \delta$ . Consider the subset of probability distributions  $\mathcal{P}_{\kappa, y_1, y_2} \subseteq \Delta(X \times Y)$ , such that  $\forall P \in \mathcal{P}_{\kappa, y_1, y_2}$ :

$$\forall k \in \{1, \dots, K\} : P(y_1 | \kappa^{-1}(k)) = 1 \bigvee P(y_2 | \kappa^{-1}(k)) = 1.$$

In words, the set  $\mathcal{P}_{\kappa, y_1, y_2}$  is the set of distributions such that the outcome  $y$  can be only one of  $y_1$  and  $y_2$ . Further,  $y$  depends only on the category  $x$  falls under (and not on  $x$  itself), and is deterministic conditional on the category of  $x$ .

Clearly for every  $P \in \mathcal{P}_{\kappa, y_1, y_2}$ , there exists a rule in  $\mathcal{F}_\kappa$  that is the best rule in  $\mathbf{F}$  for  $P$ . We now use Theorem 14.1 of Devroye, Györfi, and Lugosi (1996). In our notation, it states that:<sup>19</sup>

$$\sup_{P \in \mathcal{P}_{\kappa, y_1, y_2}} \int_{s^t} \sup_{f \in \mathcal{F}_\kappa} |E_{\nu(s^t)} f - E_P f| dP^t \geq \frac{(K-1)\delta}{2et} \left(1 - \frac{1}{t}\right).$$

However,

$$\begin{aligned} \Delta_t(\mathcal{F}_\kappa) &= \sup_P \int_{s^t} \sup_{f \in \mathcal{F}_\kappa} |E_{\nu(s^t)} f - E_P f| dP^t \\ &\geq \sup_{P \in \mathcal{P}_{\kappa, y_1, y_2}} \int_{s^t} \sup_{f \in \mathcal{F}_\kappa} |E_{\nu(s^t)} f - E_P f| dP^t \\ &\geq \frac{(K-1)\delta}{2et} \left(1 - \frac{1}{t}\right). \end{aligned}$$

Therefore for  $\Delta_t(\mathcal{F}_\kappa) \leq \epsilon$ , it must be that

$$K \leq \frac{2et^2}{(t-1)\delta} \epsilon + 1.$$

---

<sup>19</sup>Note that the VC-dimension of the class of categorization rules with  $K$  categories is  $K$ .

To see the latter part, i.e. (7), by Lemma B.3, the pseudo-dimension of a categorization-based rule  $\mathcal{F}_\kappa$  with  $K$  partitions is at most  $K|Y|$ . Therefore, applying Lemma B.2,

$$\Delta_t(\mathcal{F}_\kappa) \leq 8\sqrt{\frac{(K|Y|)^2 \ln 32e + K|Y| \ln \frac{te}{8}}{t}}.$$

Therefore for any  $\epsilon$  and any  $k^-$ , there exists  $t$  large enough such that  $\Delta_t(\mathcal{F}_\kappa) < \epsilon$  when  $\kappa$  has at most  $k^-$  partitions. ■

**Proof of Theorem 2:** The proof of this theorem proceeds analogously to the first part of the proof of Theorem 1, noting that in the 2 outcome case, linear-order frames with  $n$  dimensions have a VC-dimension of  $n$ . ■

**Proof of Theorem 3:** Suppose that the decision maker has  $t$  data points, and uses his original  $\mathcal{F}$ . Recall that  $Y = A = \{0, \frac{1}{k}, \frac{2}{k}, \dots, 1\}$ .

Our proof proceeds in two steps. We prove an upper bound on  $\Delta_t(\mathcal{F}_{k'})$ . We then prove a similar lower bound on  $\Delta_t(\mathcal{F})$  such that  $\Delta_t(\mathcal{F}) > \Delta_t(\mathcal{F}_{k'})$  for  $k'$  sufficiently less than  $k$ .

First, suppose the decision maker uses the ‘satisficing’ class  $\mathcal{F}_{k'}$ ,  $k' \leq k$ , where  $A$  has been reduced to  $\{y'_1, y'_2, \dots, y'_{k'}\}$ .

**Lemma B.4** *Let  $Y$  and  $A$  be as described before, and  $u(y, a) = -|y - a|$ . The Pseudo-dimension of  $\mathcal{F}_{k'}$  is  $k'$ .*

For simplicity, without loss of generality, we shall consider  $X = \{1, 2, \dots, |X|\}$  with the standard order, and assume that  $|X| > k'$ .

We claim that there is a string of size  $k'$  that can be pseudo-shattered. In particular, consider the string  $(i, y'_i), i = 1 \dots k'$ . Let  $c = (0, \dots, 0)$ . For any string  $b \in \{0, 1\}^{k'}$ , select a function  $f_b$  such that :

$$f_b(i) = \begin{cases} y'_i & \text{if } b_i = 0 \\ y'_{i+1} & \text{otherwise} \end{cases}$$

Note that such a  $f_b$  is in  $\mathcal{F}$  by definition.

Next we need to show that no string of size  $k' + 1$  can be pseudo-shattered. Suppose not, i.e. there exist  $(x_i, y_i) \in (X \times Y)$ ,  $i = 1 \dots k' + 1$  and  $c \in \mathfrak{R}^{k'+1}$  such that for each  $b \in \{0, 1\}^{k'+1}$ , there is a function  $f_b \in \mathcal{F}_{k'}$  such that:

$$\forall 1 \leq i \leq k' + 1 : f_b(x_i) > c_i \text{ iff } b_i = 1.$$

Without loss of generality, suppose  $x_1 < x_2 < \dots < x_{k'+1}$ .

Clearly for  $\mathcal{F}_{k'}$  to be able to pseudo-shatter, for each  $i$  there must exist  $y_i^1, y_i^2 \in Y'$  such that

$$|y_i - y_i^1| \leq c_i < |y_i - y_i^2|.$$

Since all functions in  $\mathcal{F}_{k'}$  must be monotonic, it must further be the case that for all  $i$ :

$$y_i^j \leq y_{i+1}^{j'}, j, j' \in \{1, 2\}.$$

Further, the collection  $(y_i^1, y_i^2, y_{i+1}^1, y_{i+1}^2)$  must have at least 3 distinct values: if there are only 2 distinct values  $y' < y''$ , there are exactly 3 possible combinations:

1.  $f(x_i) = f(x_{i+1}) = y'$ .
2.  $f(x_i) = y', f(x_{i+1}) = y''$ .
3.  $f(x_i) = f(x_{i+1}) = y''$ .

Clearly these 3 possible combinations cannot satisfy the 4 inequalities required to pseudo-shatter the points  $(x_i, y_i), (x_{i+1}, y_{i+1})$ .

Therefore,  $y_1^1, y_1^2, \dots, y_{k'+1}^1, y_{k'+1}^2$  must have at least  $k' + 1$  distinct values, which is clearly impossible. ■

We now have an upper bound  $\Delta_t(\mathcal{F}_{k'})$  by Lemma B.2,

$$\Delta_t(\mathcal{F}_{k'}) \leq 8\sqrt{\frac{k'^2 \ln 32e + k' \ln \frac{te}{8}}{t}}. \quad (12)$$

This concludes the proof of the first part of the theorem. We now need to show a lower bound on  $\Delta_t(\mathcal{F})$ . By Lemma B.4, we know that the pseudo dimension of  $\mathcal{F}$  is  $k$ .

By Theorems 11 and 12 of Li, Long, and Srinivasan (2001), (for all  $0 \leq \nu, \alpha \leq \frac{1}{100}; \frac{1}{5} \geq \delta > 0$ ), if  $t \leq \lfloor \max(\frac{1}{30\alpha^2\nu} \ln \frac{1}{6\delta\nu}, \frac{k'}{30\alpha^2\nu} \ln \frac{1}{3\nu}) \rfloor$ ,

$$P^t \left\{ \sup_{f \in \mathcal{F}} \frac{|E_{\nu(s^t)}f - E_P f|}{E_{\nu(s^t)}f + E_P f + \nu} > \alpha \right\} > \delta.$$

Therefore if  $t = \lfloor \frac{1}{2}(\frac{1}{30\alpha^2\nu} \ln \frac{1}{6\delta\nu} + \frac{k'}{30\alpha^2\nu} \ln \frac{1}{3\nu}) \rfloor$ .

$$P^t \left\{ \sup_{f \in \mathcal{F}} \frac{|E_{\nu(s^t)}f - E_P f|}{E_{\nu(s^t)}f + E_P f + \nu} > \alpha \right\} > \delta.$$

This implies that

$$P^t \left\{ \sup_{f \in \mathcal{F}} \frac{|E_{\nu(s^t)}f - E_P f|}{E_{\nu(s^t)}f + E_P f + \nu} > \alpha \right\} > \min \left( c' \nu^{-k} e^{-\frac{t\alpha^2\nu}{c}}, \frac{1}{5} \right),$$

for appropriate constants  $c, c' > 0$ .

Next note that, for any  $\epsilon, \nu = \frac{1}{100}, \alpha = 100\epsilon$ ,

$$\frac{|E_{\nu(s^t)}f - E_P f|}{E_{\nu(s^t)}f + E_P f + \nu} > \alpha \Rightarrow |E_{\nu(s^t)}f - E_P f| > \epsilon.$$

Therefore, for any  $\epsilon \leq 10^{-4}$ :

$$P^t \left( \sup_{f \in \mathcal{F}} |E_{\nu(s^t)}f - E_P f| > \epsilon \right) > \min \left( c' \left( \frac{1}{100} \right)^{-k} e^{-\frac{t\epsilon^2}{2c}}, \frac{1}{5} \right).$$

Note that if for a non negative random variable  $x$ ,  $P(x \geq \epsilon) > \delta$  for some  $\epsilon, \delta$ , then  $E(x) \geq \epsilon\delta$ . Therefore,

$$\Delta_t(\mathcal{F}) \geq \min \left( \epsilon c' (100)^k e^{-\frac{t\epsilon^2}{2c}}, \frac{10^{-4}}{5} \right).$$

This concludes the proof of the second part of the theorem. ■

## C Proofs of Propositions in Section 5

The proofs of both propositions will be by example- we therefore first describe the partitions used by the two agents,  $A^*$  and elements of environment.

First, for economy of notation, we assume that  $\pi$ , the distribution over business situations, is uniform.

Consider a partition  $X$  into  $2k$  sub-blocks of  $\frac{N}{2k}$  elements each,  $X_1, X_2, \dots, X_{2k}$ . Each agent's partition has  $k$  elements, and is described below:

$$\begin{aligned} \mathcal{C}_1 &= \left\{ (X_1 \cup X_2), (X_3 \cup X_4), \dots, (X_{2k-1} \cup X_{2k}) \right\}, \\ \mathcal{C}_2 &= \left\{ (X_1 \cup X_4), (X_2 \cup X_3), \dots, (X_{2k-3} \cup X_{2k}), (X_{2k-2} \cup X_{2k-1}) \right\}. \end{aligned}$$

This implies that the meet and join of the two partitions are:

$$\begin{aligned} \mathcal{C}_1 \wedge \mathcal{C}_2 &= \left\{ \left( \bigcup_{i=1}^4 X_i \right) \dots \left( \bigcup_{i=2k-3}^{2k} X_i \right) \right\}, \\ \mathcal{C}_1 \vee \mathcal{C}_2 &= \{X_1, X_2, \dots, X_{2k}\}. \end{aligned}$$

Recall that the distribution on  $X$  is uniform, i.e. any of the business situations can occur with probability  $\frac{1}{n}$ - let us denote this distribution as  $\pi$ . The preferences,  $A^*$  of the principal are (partially) specified thus: in any sub-block  $X_i$  as

$$\pi(\{x | A^*(y) = 1, x \in X_i\}) = p > \frac{1}{2}.$$

**Proof of Proposition 4:** We will make a simplifying assumption in this proposition- agent  $i$ , on viewing past history  $s^t$ , for any category  $C_i^k \subseteq X$  will pick action 0 if

$$|\{i : s_i = (x, 0), x \in C_i^k\}| > |\{i : s_i = (x, 1), x \in C_i^k\}|,$$

and action 1 otherwise. In words, for each category (partition element), the agent will take the action that is more prevalent in the history corresponding to that category, breaking ties in favor of action 1.<sup>20</sup>

---

<sup>20</sup>Note that this bias is clearly helpful to the principal. The proposition remains true

First suppose the principal ‘waters down’ his preferred rule. Clearly, his preferred rule that is jointly measurable by both agents’ partitions is  $A'(x) = 1$  for all  $x$ . Further, if he undertakes this watering down, then both agents learn this rule with probability 1. Therefore the payoff to the principal in this event will be  $(1 - p)\alpha + p$ .

Now suppose he sticks to his originally preferred rule,  $A^*$ . Each agent  $i$  sees a history of  $t$  past decisions, and chooses the empirically best performing rule measurable with respect to his partition  $\mathcal{C}_i$ .

Firstly, note that the best possible rule each agent could pick measurable with respect to their partitions, if they knew the principal’s preference  $A^*$ , is 1 for all  $x$ . This would give the principal a payoff of  $p$ .

Therefore best performing rule,  $A_i(s^t)$ , that agent  $i$  could pick as a function of the data is such that:

$$E_{s^t} [\pi[x|A_i(s^t)(x) = A^*(x)]] < p.$$

As a result the payoff to the principal from the agents’ learning strategies is :

$$\begin{aligned} E_{s^t} \left[ \pi[x|(A_1(s^t)(x) = A^*(x)) \wedge (A_2(s^t)(x) = A^*(x))] \right] &< p \\ &\leq (1 - p)\alpha + p. \end{aligned}$$

Where the second inequality follows since  $0 \leq \alpha < 1$ .

**Proof of Proposition 5:** Firstly, note that by Theorem 6 if the principal does not refine agents’ partitions, each agent’s learned rule  $A_i(s^t)$  is such that:

$$E_{s^t} [\pi[x|A_i(s^t)(x) = A^*(x)]] \geq p - 16\sqrt{\frac{k \log t + 4}{2t}}.$$

---

even otherwise. To see this note that if the agents did not have this bias, they would have additional mis-coordination when both actions appear exactly the same number of times in the history corresponding to that category, since they would have to guess the action. Further the loss from this would be more in the principal’s preferred rule than in the ‘watered down’ version.

However,

$$\begin{aligned}
\pi \left[ x \mid \bigcap_{i=1,2} (A_i(s^t)(x) = A^*(x)) \right] &= \pi[x \mid (A_1(s^t)(x) = A^*(x))] + \pi[x \mid (A_2(s^t)(x) = A^*(x))] \\
&\quad - \pi[x \mid \bigcup_{i=1,2} (A_i(s^t)(x) = A^*(x))] \\
&\geq \pi[x \mid (A_1(s^t)(x) = A^*(x))] + \pi[x \mid (A_2(s^t)(x) = A^*(x))] - 1
\end{aligned}$$

Therefore

$$E_{s^t} \left[ \pi[x \mid \bigcap_{i=1,2} (A_i(s^t)(x) = A^*(x))] \right] \geq 2p - 32\sqrt{\frac{k \log t + 4}{2t}} - 1.$$

If the principal does refine agents' partitions (by assumption to the finest possible partition), each agents' learned rule  $A'_i(s^t)$  is such that:

$$\mathbb{E}_{s^t} [\mathbb{P}[x \mid A'_i(s^t)(x) = A^*(x)]] \leq \frac{t}{N} + \frac{0.5(N-t)}{N}.$$

This is because agents only view data for at most  $t$  unique situations. Since their partitions are single elements, for every situation they have not seen in the past, they can only 'guess' the correct response with probability 0.5. Therefore the expected payoff to the principal is

$$\mathbb{E}_{s^t} \left[ \mathbb{P}[x \mid \bigcap_{i=1,2} (A_i(s^t)(x) = A^*(x))] \right] \leq \frac{t}{N} + \frac{0.25(N-t)}{N}.$$

The result follows provided  $N$  is suitably larger than  $t$ . ■

## References

- AL-NAJJAR, N. I. (2009): “Decision Makers as Statisticians: Diversity, Ambiguity and Learning,” *Econometrica*, Forthcoming.
- BARBERIS, N., AND A. SHLEIFER (2003): “Style investing,” *Journal of Financial Economics*, 68(2), 161–199.
- BENDAVID, S., N. CESABIANCHI, D. HAUSSLER, AND P. LONG (1995): “Characterizations of Learnability for Classes of  $\{0, \dots, n\}$ -Valued Functions,” *Journal of Computer and System Sciences*, 50(1), 74–86.
- BERNSTEIN, R. (1995): *Style Investing: Unique Insight Into Equity Management*. John Wiley and Sons.
- CHI, M., P. FELTOVICH, AND R. GLASER (1981): “Categorization and representation of physics problems by experts and novices,” *Cognitive Science*, 5(2), 121–152.
- CRÉMER, J., L. GARICANO, AND A. PRAT (2007): “Language and the Theory of the Firm,” *Quarterly Journal of Economics*, pp. 373–407.
- DEVROYE, L., L. GYORFI, AND G. LUGOSI (1996): *A Probabilistic Theory of Pattern Recognition*. Springer, Berlin.
- DIMSON, E., AND S. NAGEL (2002): “Seeking out Investment Value in Style,” London Business School.
- FRYER, R., AND M. JACKSON (2008): “A Categorical Model of Cognition and Biased Decision Making,” *The BE Journal of Theoretical Economics*, 8(1), 6.
- GILBOA, I., AND L. SAMUELSON (2008): “Evolution of Simplicity,” Yale University.
- GOLDSTONE, R. (1994): “The role of similarity in categorization: Providing a groundwork,” *Cognition*, 52(2), 125–157.

- HAUSSLER, D. (1995): “Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications,” in *The Mathematics of Generalization: The Proceedings of the SFI/CNLS Workshop on Formal Approaches to Supervised Learning*. Perseus Books.
- KALAI, G. (2003): “Learnability and rationality of choice,” *Journal of Economic Theory*, 113(1), 104–117.
- KAPLAN, R., AND D. NORTON (1992): “The balanced scorecard—measures that drive performance.,” *Harv Bus Rev*, 70(1), 71–9.
- KAPLAN, R., AND D. NORTON (1996): “Using the balanced scorecard as a strategic management system,” *Harvard Business Review*, 74(1), 75–85.
- KREPS, D. (1990): “Corporate Culture and Economic Theory,” in *Perspectives on Positive Political Economy*, ed. by Alt, and Shepsle, pp. 90–143. Cambridge University Press.
- LI, Y., P. LONG, AND A. SRINIVASAN (2001): “Improved Bounds on the Sample Complexity of Learning,” *Journal of Computer and System Sciences*, 62(3), 516–527.
- LUCAS JR., R. (1980): “Methods and Problems in Business Cycle Theory,” *Journal of Money, Credit and Banking*, 12(4), 696–715.
- MACKOWIAK, B., AND M. WIEDERHOLT (2009): “Optimal sticky prices under rational inattention,” *American Economic Review* (forthcoming).
- MANSKI, C. (2008): *Identification for prediction and decision*. Harvard University Press.
- MOSCARINI, G. (2004): “Limited information capacity as a source of inertia,” *Journal of Economic Dynamics and Control*, 28(10), 2003–2035.
- MURPHY, G. L., AND D. L. MEDIN (1985): “The Role of Theories in Conceptual Coherence,” *Psychological Review*.
- NELSON, R. R., AND S. G. WINTER (1982): *An Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge, MA.

- POLLARD, D. (1990): “Empirical Processes: Theory and Applications,” *Ims*.
- REED, S. (1972): “Pattern recognition and categorization,” *Cognitive Psychology*, 3(3), 382–407.
- RIPS, L. (1989): “Similarity, typicality, and categorization,” *Similarity and analogical reasoning*, pp. 21–59.
- ROSCH, E., AND B. LLOYD (1976): *Cognition and categorization*. Hillsdale: Lawrence Erlbaum Associates.
- RUBINSTEIN, A. (1996): “Why Are Certain Properties of Binary Relations Relatively More Common in Natural Language?,” *Econometrica*, 64(2), 343–355.
- (2000): *Economics and Language*. Cambridge University Press.
- SAMUELSON, L. (2001): “Analogies, adaptation, and anomalies,” *J. Econom. Theory*, 97(2), 320–366, The evolution of preferences.
- SHARPE, W. (1992): “Asset Allocation: Management Style and Performance Measurement,” *Journal of Portfolio Management*, 18(2), 7–19.
- SIMON, H. A. (1955): “A Behavioral Model of Rational Choice,” *The Quarterly Journal of Economics*, 69(1), 99–118.
- SIMS, C. (2003): “Implications of rational inattention,” *Journal of Monetary Economics*, 50(3), 665–690.
- TVERSKY, A., AND D. KAHNEMAN (1974): “Judgment under Uncertainty: Heuristics and Biases,” *Science*, 185(4157), 1124–1131.
- VAPNIK, V. N. (1998): *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, A Wiley-Interscience Publication.
- VAPNIK, V. N., AND A. Y. CHERVONENKIS (1971): “On the Uniform Convergence of Relative Frequencies of Events to their Probabilities,” *Theory of Probability and its Applications*, XVI, 264–80.
- VOSNIADOU, S., AND A. ORTONY (1989): *Similarity and Analogical Reasoning*. Cambridge University Press.