Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 150 (2014) 467–486

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Coarse decision making and overfitting ☆

Nabil I. Al-Najjar [a,*], Mallesh M. Pai [b,1]

[a] *Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA*

[b] *Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, USA*

Received 28 March 2011; final version received 25 April 2013; accepted 12 August 2013

Available online 24 December 2013

Abstract

We study decision makers who willingly forgo decision rules that vary finely with available information, even though these decision rules are technologically feasible. We model this behavior as a consequence of using classical, frequentist methods to draw robust inferences from data. Coarse decision making then arises to mitigate the problem of over-fitting the data. The resulting behavior tends to be biased towards simplicity: decision makers choose models that are statistically simple, in a sense we make precise. In contrast to existing approaches, the key determinant of the level of coarsening is the amount of data available to the decision maker. The decision maker may choose a coarser decision rule as the stakes increase.

JEL classification: D81

Keywords: Coarse decision making; Statistical learning; Overfitting; VC-dimension; Bounded rationality

1. Introduction

Despite the breadth of phenomena they explain, classical models of decision making struggle with a large class of observed behavior we shall refer to as *coarse decision making*. By this

☆ The authors thank an anonymous editor and referee for comments that greatly improved the manuscript. They also thank Xavier Gabaix, George Mailath, Andy Postlewaite and Andrea Prat for helpful comments and discussions.

* Corresponding author.
E-mail addresses: al-najjar@northwestern.edu (N.I. Al-Najjar), mallesh@econ.upenn.edu (M.M. Pai).
URLs: http://www.kellogg.northwestern.edu/faculty/alnajjar/htm/index.html (N.I. Al-Najjar), http://www.malleshmpai.com/ (M.M. Pai).

[1] Early versions of this paper were written while the author was a graduate student at the Department of Managerial Economics and Decision Sciences, Kellogg School of Management, Northwestern University.

0022-0531/$ – see front matter
http://dx.doi.org/10.1016/j.jet.2013.12.003

we mean the phenomenon of decision makers making coarse choices—their choice does not finely vary with the information they have, even though doing so would be informationally and technologically feasible. They opt instead for decision rules that are less sensitive to state by state variations—'coarse' rules in our terminology.

There is a large literature documenting and modeling manifestations of coarse decision making, we limit ourselves here to a few motivating examples.[2] The literature on bounded rationality studies outcomes when decision makers use *rules of thumb* (or similar coarsenings such as *decision heuristics*, *routines*, and *analogies*).[3] 'Style investing' investment strategies based on asset categories or 'styles' rather than the assets themselves. The literature on rational inattention in macroeconomics studies agents with limited or costly information processing capacity, and therefore rationally choose to ignore some available information.

The goal of this paper is to provide a simple alternate theory of coarse decision making. The central idea is to view decision makers as learning from data using classical/frequentist methods, similar to empirical work. We study a setting where a decision maker must choose a decision rule i.e. a mapping from observables to actions. He has relevant sample data to assist him in this choice. We study a two-stage decision procedure:

1. *Model-selection stage:* Select a 'model,' or 'decision frame' $\mathcal{F}$ consisting of a set of decision rules.
2. *Inference stage:* Select a rule f in $\mathcal{F}$ based on its fit with the sample data.

The decision maker balances two conflicting objectives: (1) A rich decision frame $\mathcal{F}$ improves ability to 'fit' observed samples, but, (2) an unrestricted $\mathcal{F}$ results in 'over-fitting' the sample. Coarseness of the decision frame $\mathcal{F}$ is the result of a compromise between these two concerns.[4]

The more common view of coarse decision making in the literature is that it is a consequence of cognitive and computational limitations suffered by the agents. Our explanation, i.e. that it results from difficulties of inference from limited data, is complementary to this view. In most situations of interest, it is likely that agents have both limited cognition and limited data. We focus here solely on the difficulties posed by learning from limited data because it generates novel insights into the problem.

A first implication of our framework is that behavior will be biased towards statistically simple rules, in a sense we make precise. In the case of categorization, the decision maker relies on a coarse partition of the observables to counter the risk of selecting a rule that tracks the sample data too closely (over-fitting). This leads to under-sensitivity to information: decision makers do not respond to observable changes in signals that are finer than the coarse categories they have selected.

A second implication of our framework is how coarseness of the decision frame varies with the stakes for making the right decision. Decision makers for whom cognitive and computational limitations are binding will likely invest more resources in relaxing these constraints as the stakes increase. On the other hand, heuristics such as coarse categories may continue to be important even in decisions with very large stakes. Our learning-based model implies that increasing the

[2] A broader review of related work is in Section 5.1.

[3] See Cremer et al. [6], Jehiel [16], Mohlin [18] and Samuelson [25], among others.

[4] Although this tension is well-recognized in classical statistics, it has received little attention in the theory literature. The closest paper is Al-Najjar [1] which studies the asymptotic properties of uniform learning, but does not discuss over-fitting or applications to cognitive phenomena.

stakes may lead to coarser categories (Proposition 2). A decision maker who categorizes for learning-based reasons may not view cognitive or computational limitations as binding even when the stakes are high. This has potentially observable implications in terms of the incentives to invest in relaxing such constraints.

Third, our model allows a better understanding of the distinction between categorization and incomplete information. Lack of information is an exogenous objective constraint, whereas categories in our model are 'self-imposed' by the decision maker. In the absence of incentive and strategic motives, more information means more flexibility of choice, which cannot hurt and usually helps. By contrast, in Proposition 3 we show that refining the chosen categorization scheme may make the decision maker worse off, since this could lead them to overfit their limited data.

A final implication of our model which we do not discuss in detail pertains to disagreements in interpreting information and their persistence. There is substantial empirical evidence that questions the standard assumption that individuals agree on the interpretation of information—see Kandel and Pearson [17], Cutler et al. [7], Hong and Stein [14] among many others. Coarse decision making provides a potentially useful perspective on the problem of non-informational sources of disagreement. Individuals with different categorization schemes will interpret the same information differently, despite an accumulation of data and their knowledge of each other's model.

2. The setting

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[5]:

$$u : Y \times A \to \mathbb{R}.$$

For expository simplicity, we assume that X, Y and A are finite unless indicated otherwise.

We distinguish x (the observables) from y (the outcome). 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

$$f : X \to A,$$

which determines an action $f(x)$ as a function of the observables x. Let $\mathbf{F}$ denote the set of all decision rules. To highlight issues connected with learning and statistical complexity, we assume that the decision maker is free to choose any decision rule in $\mathbf{F}$. This assumes away technological or informational factors that may limit the choice within $\mathbf{F}$.

An *environment* is a joint distribution P on $X \times Y$—it is unknown to the decision maker. We assume that, given P, a decision rule f is evaluated according to its expected payoff:

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

The decision maker has sample information to form an estimate of P and thus choose f. Specifically, there are t observations, each a pair (x, y) consisting of a vector of observables x and the corresponding outcome y. Past data about the relationship between observables and outcomes is represented by a sample:

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

[5] As currently written, X is not directly payoff relevant, it is only indirectly relevant via the inference that can be drawn about Y. Extending the model to include payoff relevant X comes at only a notational cost.

Let S^t denote the (finite) set of all such samples. We assume that samples are i.i.d. draws from the unknown P, i.e. S^t is a draw according to P^t.

Our decision maker's problem is to choose a procedure which selects a decision rule f based on the observed sample s^t. Examples of such procedures may be helpful in fixing ideas.

Example 1 *(Linear regression).* The space of observables X is identified with $\mathcal{R}^n$, while the space of outcomes Y and actions A are both identified with the real line $\mathcal{R}$. When the decision maker chooses action a, if the outcome y is realized, he receives a payoff of:

$$u(y,a) = -(y-a)^2.$$

A procedure used in practice for this sort of problem is linear regression. In our notation, a *regression model* is identified with a subset of regressors $I \subseteq \{1,2,\ldots,n\}$:

$$\mathcal{F}_I = \left\{ f \;\middle|\; f(x) = \sum_{i\in I} b^i x^i \text{ for some } b \in \mathcal{R}^I \right\},$$

where b^i and x^i represent the ith coordinate of the vectors b and x, respectively. Given a regression model $\mathcal{F}_I$ and sample $s^t = \{(x_1,y_1),\ldots,(x_t,y_t)\}$, the decision maker selects $\hat{f} \in \mathcal{F}_I$ according to the sum of least squares criterion:

$$\hat{f} \in \arg\max_{f\in\mathcal{F}_I} \sum_{j=1}^{t} -\big(y_j - f(x_j)\big)^2.$$

Example 2 *(Categorization).* The decision maker categorizes the space of observables, X, into K styles according to a categorization map

$$\kappa : X \to \{1,2,\ldots,K\}.$$

If an instance x is categorized as $\kappa(x)$, the decision maker takes an action that depends only on the category $\kappa(x)$, and not on x itself. A categorization map κ defines the model:

$$\mathcal{F}_\kappa = \big\{ f \;\big|\; f = g\circ\kappa, \text{ for some function } g : \{1,\ldots,K\} \to A \big\}.$$

The decision maker then selects the best decision rule $\hat{f}$ from $\mathcal{F}_\kappa$ based on the sample s^t:

$$\hat{f} \in \arg\max_{f\in\mathcal{F}_\kappa} \sum_{j=1}^{t} u\big(y_j, f(x_j)\big).$$

3. The decision procedure

In this paper we study a simple two-stage decision procedure that generalizes the examples we discussed above:

1. *Model-selection stage:* Select a 'model,' or 'decision frame' $\mathcal{F} \subseteq \mathbf{F}$ consisting of a set of decision rules.
2. *Inference stage:* A rule $f_{s^t}^{\mathcal{F}}$ in $\mathcal{F}$ is selected based on the observed sample s^t.

The main focus in this paper is the form of models selected in the model selection stage, i.e. what sort of $\mathcal{F} \subseteq \mathbf{F}$'s are chosen by a decision maker, given the inference stage that follows.

3.1. A frequentist inference stage

If the true distribution P is known, then the solution to our decision maker's problem is simple—he should select $f_P^\star$ where

$$f_P^\star \in \operatorname*{argmax}_{f\in\mathbf{F}} E_P f.$$

Our decision maker does not 'know' this true distribution P. The *empirical distribution* at a sample s^t assigns to each event $E \subset X \times Y$ its relative frequency in the sample

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

The empirical performance of a rule f is its average payoff over the sample

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

Instead of evaluating the expectations with respect to the unknown P, the decision maker uses the sample distribution $\nu(s^t)$. A *frequentist decision procedure* is a function:

$$\varphi : S^t \times 2^{\mathbf{F}} \to \mathbf{F}$$
$$\text{such that:}\quad \varphi\big(s^t; \mathcal{F}\big) \in \operatorname*{argmax}_{f\in\mathcal{F}} E_{\nu(s^t)} f.$$

That is, φ selects the best rule using the sample distribution $\nu(s^t)$ subject to the constraint that the selected rule must be in the selected model $\mathcal{F}$.

3.2. Model selection stage objectives

Our decision maker's objectives are to pick a frame $\mathcal{F}$ to minimize the difference in expected utility between $f_P^\star$ and $\varphi(s^t, \mathcal{F})$. Motivated by robustness, he takes a worst case *over all* probability distributions $P \in \Delta(X \times Y)$. Formally, given any frame $\mathcal{F}$, the decision maker is concerned with:

$$V(\mathcal{F}) = \sup_P \int_{s^t} \big(E_P f_P^\star - E_P \varphi\big(s^t, \mathcal{F}\big)\big)\, dP^t. \tag{1}$$

He would like to pick a frame $\mathcal{F}$ to make $V(\mathcal{F})$ as small as possible. To help interpret (1), it will be helpful to re-arrange the terms:

$$V(\mathcal{F}) = \sup_P \Bigg[\underbrace{\Big(E_P f_P^\star - \max_{f\in\mathcal{F}} E_P f\Big)}_{(1)} + \underbrace{\int_{s^t} \Big(\max_{f\in\mathcal{F}} E_P f - E_P \varphi\big(s^t, \mathcal{F}\big)\Big)\, dP^t}_{(2)} \Bigg]. \tag{2}$$

In words, picking $\mathcal{F}$ requires the decision maker to balance two conflicting criteria:

Term 1: We will refer to this term as the fit of a frame $\mathcal{F}$. Note that this improves as $\mathcal{F}$ becomes large (in the sense of inclusion). 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, and therefore this Term 1 equals zero.

Term 2: We will refer to this as the 'over-fit.' Roughly speaking, the inference stage will in general select a decision rule that is different (and worse performing) than the best decision rule in $\mathcal{F}$ if P was known. A 'large' (in a sense we make precise shortly) $\mathcal{F}$ relative to P exacerbates the problem of over-fitting. For instance, when $\mathcal{F} = \mathbf{F}$, the rule with the best empirical fit, $f^*_{s^t}$, will track the data perfectly and its performance need not be close to $E_P f^\star_P$.

For a given set of rules $\mathcal{F} \subset \mathbf{F}$, define:

$$\Delta_t(\mathcal{F}) \equiv \sup_P \int_{s^t} \sup_{f\in\mathcal{F}} |E_{\nu(s^t)} f - E_P f| \, dP^t. \tag{3}$$

The observation below shows that $\Delta_t(\mathcal{F})$ directly defines the overfit of a class $\mathcal{F}$ (we defer the proof to Appendix B).

Observation 1. If $\Delta_t(\mathcal{F}) \leqslant \epsilon$ for a model $\mathcal{F}$, then the overfit of the class:

$$\int_{s^t} \Big(\max_{f\in\mathcal{F}} E_P f - E_P \varphi\big(s^t, \mathcal{F}\big) \Big) \, dP^t \leqslant 2\epsilon.$$

Therefore if $\Delta_t(\mathcal{F})$ is large, the empirical performance of f is a poor estimate of its true performance and the selected rule from $\mathcal{F}$ may be much worse than the true best rule in $\mathcal{F}$.

Since both Term 1 and Term 2 are positive, picking an $\mathcal{F}$ with small overfit is a necessary condition to minimize $V(\mathcal{F})$. The analytical properties of Term 1 depend on the particular decision problem, while Term 2 can be analyzed more generally as we describe below. Therefore, we focus on properties of $\mathcal{F}$ such that their overfit is small.

Definition 1. A *model* or *decision frame* is a pair $(\mathcal{F}, \epsilon)$ where $\mathcal{F} \subseteq \mathbf{F}$ and $\epsilon > 0$ such that $\Delta_t(\mathcal{F}) \leqslant \epsilon$. Given a frame, an integer t, and data s^t, the decision maker selects the decision rule $\varphi(s^t, \mathcal{F})$.

The decision maker's frame $(\mathcal{F}, \epsilon)$ determines how inferences are drawn from past evidence. The set $\mathcal{F}$ represents the set of rules or patterns he considers, while the parameter ϵ represents the desire to avoid overfitting. We use the theory of uniform learning, originating with Vapnik and Chervonenkis [32], to characterize such models.[6]

Before we discuss the sorts of $\mathcal{F}$'s that perform well for a decision maker, it is useful to build some intuition for why some restriction is necessary. The following subsection suggests why a model of the set of all possible rules, i.e. $\mathcal{F} = \mathbf{F}$, can perform badly.

3.3. Naïve empiricism and over-fitting

The empirically estimated performance of a rule f will typically differ from its true performance due to sampling error. A consequence of the law of large numbers is that for a large

[6] For a brief, self-contained account, see Al-Najjar [1]. For textbook expositions, see Vapnik [31] or Devroye et al. [8]. See Harman and Kulkarni [12] (from which Fig. 1 was taken) for an informal introduction as well as connections to learning and induction.

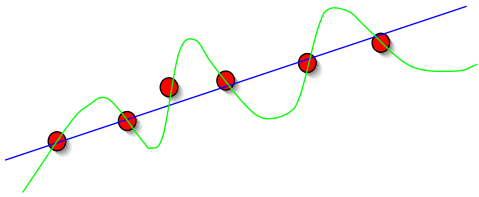

Fig. 1. Over-fitting.

enough sample size t, the empirical estimate of the performance of any single 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[7]:

$$\sup_{f \in \mathbf{F}} \Delta_t(f) < \epsilon, \quad \forall t \geqslant \bar{t}. \tag{4}$$

A "naïve" decision maker selects the rule that best fits the data without any constraints on $\mathbf{F}$. A flawed argument for this procedure is as follows: when (4) holds, the empirical performance of each rule f is close to its true performance for any P. In particular, the empirical performance of $\varphi(\cdot; \mathbf{F})$ is close to its true performance, and therefore $\varphi(s^t; \mathbf{F})$ performs nearly as well as the optimal rule $f_P^\star$ for any P. However this argument erroneously switches the order of limits, that is, it assumes that

$$\sup_{\mathbf{F}} \sup_{P} \int_{s^t} |E_{\nu(s^t)} f - E_P f| \, dP^t = \sup_{P} \sup_{\mathbf{F}} \int_{s^t} |E_{\nu(s^t)} f - E_P f| \, dP^t,$$

which is not true in general. Indeed, a bound on the right hand side is needed to conclude the (approximate) optimality of $\varphi(\cdot; \mathbf{F})$. For this, one needs theorems from the literature on uniform learning, also known as *uniform laws of large numbers*. An example may be helpful.

Example *(Regression revisited).* In the context of regression (Example 1) a decision maker who is free to choose any continuous regression curve is guaranteed to find one that fits perfectly—however, this choice 'over-fits' the data. Fig. 1 illustrates this.

This problem may be circumvented by both restricting attention to a 'small' set of explanatory variables, i.e. ignoring some components of X and a small class of functions e.g. linear in the chosen explanatory variables. The intuition we develop in this paper extends beyond linear regression because we do not impose any a priori order or any other structure on the sets X, Y and A, or make any assumptions regarding P.

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: in the case of linear regression, one may reject general continuous curves in favor of the smaller class $\mathcal{F}_{\text{linear}}$.

4. Categorization

We focus on categorization as an important illustration of coarse decision making. Many of the intuitions and results we develop extend to other, more general settings.

[7] The finiteness of X, Y and A are sufficient for this. In general, this would require suitable regularity conditions on u.

4.1. A formal model of categorization

Definition 2. A decision frame $(\mathcal{F}_\kappa, \epsilon)$ is a *categorization frame* if there is

$$\kappa : X \to \{1, \ldots, K\}$$
$$\text{such that} \quad \mathcal{F}_\kappa = \left\{ f \;\middle|\; f = g \circ \kappa, \text{ for some function } g : \{1, \ldots, K\} \to A \right\}.$$

In this case, refer to κ as the categorization function, and write $X_k = \kappa^{-1}(k)$ to denote the kth category.

In a categorization frame, the decision maker first classifies the observables into one of K categories, then takes an action that depends only on the category. The remainder of this section discusses some implications of our model of categorization and contrasts them with the literature.

4.2. Why do we see coarse categories?

Every decision problem admits a trivial categorization, namely one where each x is its own category (that is, $K = |X|$ and $\kappa(x) = x$). The psychology literature cited earlier and the economic and finance uses of this concept (e.g., Barberis and Shleifer [2], and Hong et al. [15]) suggest that observed behavior exhibits reliance on few, coarse categories.

The following result shows that a decision maker who uses a categorization frame, and is concerned with robustness and overfitting, will necessarily rely on a 'small' number of categories, in a sense made precise in the theorem.

Proposition 1. *Suppose the decision maker has t samples, and restricts attention to models $\mathcal{F}$ with overfit at most $\epsilon > 0$, i.e. $\Delta_t(\mathcal{F}) \leqslant \epsilon$. Then, there are two functions $k^+(t, \epsilon)$ and $k^-(t, \epsilon)$ such that*:

1. *Categorization must be coarse*: *$\mathcal{F}$ can have at most $k^+(t, \epsilon)$ categories, i.e. for every categorization function κ with K categories*

$$\Delta_t(\mathcal{F}_\kappa) < \epsilon \quad \Rightarrow \quad K \leqslant k^+(t, \epsilon). \tag{5}$$

2. *Coarse categorization is possible*: *Any $\mathcal{F}$ with $k^-(t, \epsilon)$ categories has small overfit, i.e. for any categorization rule κ with $K = k^-(t, \epsilon)$, we have that*:

$$\Delta_t(\mathcal{F}_\kappa) \leqslant \epsilon. \tag{6}$$

Simple comparative statics follow easily from the proof.

4.2.1. How does categorization depend on data?

Corollary 1. *Fixing ϵ, $\kappa^+(t, \epsilon)$ and $\kappa^-(t, \epsilon)$ are increasing in t.*

We leave the details to Appendix B—see Section B.3 for details, and (11) for the formal inequality.

However, it is useful to build some intuition for how large the upper bound is to be sure that our statements are not generally vacuous. Fixing the particular decision problem (i.e. X, Y, A

and u) and the level of overfit ϵ the decision maker is willing to tolerate, one can show that κ^+ grows linearly in the number of data points available t, i.e. $\kappa^+(t,\epsilon) = O(t)$.

Our theorems therefore have bite when the cardinality of the space of observables is larger than the number available sample data points. In a lot of 'real-life' decision problems, this is easily the case—for instance a health study that encodes medically relevant characteristics of individuals (sex, age, height, weight, basic vitals and medical history) already has a state space cardinality that is orders of magnitude larger than the population of the earth! Some categorization is therefore necessary.

The question of identifying useful constraints on similarity has a long history in several fields (see Section 5.1)—a common theme is that the unconstrained use of similarity makes this concept all but useless.[8] The framework of this paper suggests a reason why.

4.2.2. How does categorization depend on overfit?

A related comparative static further clarifies the trade off between the amount of overfit ϵ the decision maker is willing to tolerate and the amount of sample data available.

Corollary 2. *Fixing the amount of data t, $\kappa^+(t,\epsilon)$ and $\kappa^-(t,\epsilon)$ are increasing in ϵ.*

In words, for the same amount of sample data available to the decision maker, he can consider models with more categories if he is willing to accept more overfit ϵ.

4.3. Categorization in high-stake decisions

Our account of coarse categorization in terms of learning and the desire to avoid over-fitting is but one possible explanation. The more popular explanation in the literature is that it is the result of cognitive and computational limitations: decision makers coarsely categorize because they lack the sophistication or resources to carry out unrestrictedly fine contingent planning.[9] Do these explanations lead to observable differences in behavior?

To answer this question, we consider the comparative statics of increasing the stakes in making the right decision while keeping other components of the decision problem fixed. Consider a forecasting problem with $X = Y = A$ where the decision maker observes an instance x and receives a payoff $u(y, y')$ that depends on his forecast $y' = a(x)$ and the realized value y. Define the payoff function:

$$u_\theta(y, y') = \begin{cases} 0, & y = y', \\ -\theta, & \text{otherwise.} \end{cases}$$

The parameter $\theta \geqslant 1$ reflects the stakes involved, with $\theta = 1$ serving as a useful benchmark. As θ increases, the decision maker is penalized more heavily for an incorrect forecast. How

[8] In their seminal paper on the subject, Murphy and Medin [19] note that "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."

[9] This argument has a long history in the study of bounded rationality. It appears in the classic works of Simon [27,28] where he appeals to computational limitations to explain the prevalence of rules of thumb, satisficing, and other behaviors that could be thought of as evidence of coarse decision making. Dye [10] and many others explicitly model computational costs to explain coarse contracts.

does the heightened incentive to get it right impact categorization? A decision maker driven by computational, cognitive, or memory constraints will devote greater efforts to overcome these limitations as the stake he has in taking the correct action increases. As the stakes increase, cognitive costs become increasingly trivial, leading to progressively finer decision rules. For example, see Dye [10], where a unit cost is paid for each computation step. In our model, the theorem shows that the decision maker uses coarse categories and rules of thumb even in important decisions with high stakes.[10]

Proposition 2. *For any ϵ (the maximum amount of empirical discrepancy the decision maker will tolerate) and t, let $\kappa_\theta^+(\epsilon, t)$ be the maximum number of categories the decision maker can have in classification problem with payoff u_θ if he has t samples and will accept an empirical discrepancy of ϵ. Then*

$$\kappa_\theta^+(\epsilon, t) = \kappa_1^+\left(\frac{\epsilon}{\theta}, t\right).$$

Therefore, $\kappa_\theta^+(\epsilon, t)$ is decreasing in θ.

In words, the maximum number of categories *decreases* as the penalty for making the wrong decision increases, ceteris paribus. The level of coarseness depends on $\frac{\epsilon}{\theta}$, i.e. the ratio of the empirical discrepancy the decision maker will tolerate and the 'stakes' of the problem.

There is little doubt that cognitive and computational limitations are real and play an important role in behavior. The main takeaway from this section is that a learning based model may provide opposite comparative statics to a cognitive limitations based model of coarse decision making. Where higher stakes in the latter give the agent incentives to spend more resources to 'get it right,' in the former class of models they may imply coarsening to avoid 'getting it wrong.' We think this difference is worth noting. As we highlighted earlier, there are several large stakes decisions where 'simple rules' are observed. Understanding whether limited data or limited cognition/computation is the relevant 'binding constraint' would help us understand how an agent should spend resources to improve the quality of his decision making.

We finally note that higher stakes may have *indirect* effects in a learning model. If we expand our model to include costly data gathering then a decision maker facing decisions where the stakes are high may choose to gather more data. The total effect on the number of categories is then ambiguous.

4.4. Merging categories

The categorization model presented in Section 4.1 has formal similarities to models of incomplete information: The categorization function κ defines an information partition, with $\mathcal{F}_\kappa$ being the set of all rules measurable with respect to this information. Many models of limited cognition use partitions to represent analogies, similarity, memory limitations or, more generally, decision makers' coarse understanding of the environment.

A natural question is whether partitional models of limited cognition reduce to standard Bayesian models with limited information. In this section, we illustrate how our categorization

[10] Phenomena such as style investing and the use of balanced scorecards by firms are clearly instances of coarse decision making that affect 'large stakes' decisions.

frames substantively differ in at least one important respect. Consider a standard incomplete information setting where two individuals are endowed with distinct information partitions $\mathcal{S}_1, \mathcal{S}_2$ of X. If these individuals pool their information, the resulting information structure is the common refinement $\mathcal{S}_1 \vee \mathcal{S}_2$. Absent commitment issues or other strategic motives, each individual weakly prefers the finer information partition $\mathcal{S}_1 \vee \mathcal{S}_2$. More information means more flexibility of choice, and flexibility cannot hurt—and usually helps.

Consider now two individuals who use categorization frames κ_1, κ_2, with corresponding partitions $\mathcal{Q}_1, \mathcal{Q}_2$. In contrast with information partitions, making decisions based on the pooled categorization $\mathcal{Q}_1 \vee \mathcal{Q}_2$ can be harmful. Specifically, there is a distribution P where the first decision maker's expected payoff under $\mathcal{Q}_1 \vee \mathcal{Q}_2$ is lower than his expected payoff under the coarser initial partition $\mathcal{Q}_1$.

To illustrate, consider *a forecasting problem* with $X = \{x_1, \ldots, x_N\}$, two outcomes $A = Y = \{0, 1\}$, and $u(x, a) = 1$ if $a = x$ and 0 otherwise. For a pair of categorization frames κ_1, κ_2, define $\kappa_1 \vee \kappa_2$ to be the frame that maps each instance x to the element of $\mathcal{Q}_1 \vee \mathcal{Q}_2$ that contains it.

Proposition 3. *There exists a function $n^-(t)$, such that for every forecasting problem with $|X| = N > n^-(t)$ there exists a probability distribution P and categorization frames κ_1, κ_2 such that*:

$$E_P\varphi\big(s^t, \kappa_1\big) > E_P\varphi\big(s^t, \kappa_1 \vee \kappa_2\big),$$

i.e. the decision maker's expected payoff from the best fitting act with the coarser partition κ_1 is larger than his payoff from the finer partition $\kappa_1 \vee \kappa_2$.

In words, the decision maker's expected utility at P is larger under the coarser set of rules. To be clear, the claim of Proposition 3 is stronger than saying that the decision maker is better off not refining under the decision criterion defined in this paper. That already follows from Theorem 1. This proposition asserts that even the decision maker's *expected payoff* is better when not refining. If partitions represented information sets, then refining the set of decision rules makes the decision maker at least weakly better off for *every* probability distribution. This also distinguishes our model from bounded rationality accounts of categorization where finer categories are unambiguously better.

The intuition for this result is that a decision maker with partition $\mathcal{Q}$ will pick action 0 for some x in partition element Q if the sample has more 0's than 1's in that partition element. If $\mathcal{Q}$ is 'coarse,' then with high probability, the decision maker will get several data points for each partition element, and pick the best average action for each. Suppose instead $\mathcal{Q}$ is fine, i.e. it has several partition elements. In this case, with high probability there will be many partition elements with few observations. This increases the likelihood of taking the wrong action due to sampling error. The proof of the proposition constructs probability distributions and categorization frames where the decision maker is worse off in expectation.

5. Discussion and conclusion

In this paper, we presented a simple alternate explanation of coarse decision making, i.e. that it arises from the decision maker faced with limited data restricting his model to avoid overfitting. The resulting implications and comparative statics are different from cognitive and computational limitation based explanations that are prevalent in the literature. We would like to close with two short notes.

First, for the interested reader, we include a brief literature review in Section 5.1. This goes over some of the prominent work on coarse decision making.

Second, we should note that our theorems provide no guidance on what frame the decision maker should actually pick, only that it should not be too fine (or too coarse) as made precise in Proposition 1. To actually 'apply' this model, additional criteria are required. This is similar to how, in applications, Bayesian decision makers are modeled with a 'common prior' or 'rational expectations.' We conclude in Section 5.2 with an exploratory discussion of the converse possibility—using coarse decision making to inform prior selection.

5.1. Literatures on coarse decision making

As we suggested in the introduction, there is a vast literature in psychology, economics and finance that documents and studies behavior that fits within our definition of coarse decision making. While a comprehensive review is outside the purview of this paper, we provide a sampling of relevant works from these literatures.

Categorization in psychology: Categorization is a decision procedure in which problems or situations are grouped into categories and decisions are made based on the categories, rather than the original problem or situation. Categorization is central in cognition psychology, for example see the collection of Vosniadou and Ortony [33], and the papers by Reed [22], Rosch and Lloyd [24], Chi et al. [5], Rips [23], Murphy and Medin [19], Goldstone [11], among many others.

Style investing in finance: In a pioneering work, Sharpe [26] showed that 90% of the variation in the return on mutual funds can be explained by investors basing their investment strategies on asset categories, or 'styles.' See Bernstein [4] and Dimson and Nagel [9] for a historical overview, and Barberis and Shleifer [2] for a model that uses style investing patterns to explain movements of asset prices.

Investing paradigms and model revision: In Hong et al. [15] asset returns are governed by a multi-variate process but the decision maker is restricted to using a univariate investing rule.

Rational inattention: Sims [29] proposed a model where limited information processing capability results in decision makers paying attention to only a subset of the available information. It has been applied to explain price and wage rigidities.

Optimal categorization: The recent paper of Mohlin [18] studies a model similar to ours, where agent picks a categorization of observable variables to make a prediction about an unobserved variables. He studies the design of rules that minimize expected error for a known probability distribution, trading off bias (from categorizing different observables together) with variance (limited sample size within category). Cremer et al. [6] study the design of an optimal language with a limited number of 'words' or 'codes.'

Coarse decision making is also related to *rules of thumb* and other similar ideas, such as *decision heuristics*, *routines*, and *analogies*. See Tversky and Kahneman [30], Nelson and Winter [20], and Samuelson [25], among others.

5.2. Coarse decision making as a prior selection criterion

A Bayesian decision maker in our setting is one with a prior belief π over $\mathcal{P}$ with choice rule given by expected utility maximization:

$$\beta(s^t, \pi) \in \operatorname*{argmax}_{f \in \mathbf{F}} \int_{\mathcal{P}} [E_P f]\, d\pi(P|s^t).$$

Here $\pi(\cdot|s^t)$ is the posterior over $\mathcal{P}$ given the sample s^t. The term $E_P f$ is the same as in the frequentist rule φ, but the Bayesian uses the posterior $\pi(\cdot|s^t)$ to weigh different P's.

Bayesian decision making is founded on consistency conditions on the decision maker's preference. These yield: (1) a utility u over consequences; (2) a belief π over parameters, and (3) the expected utility criterion to combine the two. The consistency conditions give no guidance for what the belief π 'should be,' only that there must be such a belief. In practice, economic and statistical models impose considerable structure on Bayesian beliefs based on considerations of tractability, simplicity, or other intuitive desiderata.

Concerns about overfitting can provide a systematic way to select priors. Given a decision frame $\mathcal{F}$, define $\mathcal{P}_{\mathcal{F}}$ to be the set of all probability distributions P such that $\beta(s^t, P) \in \mathcal{F}$. That is, $\mathcal{P}_{\mathcal{F}}$ is all the distributions that would lead a decision maker, Bayesian or not, to optimally select a rule $\mathcal{F}$. A Bayesian with prior π whose support is $\mathcal{P}_{\mathcal{F}}$ will also choose a rule in $\mathcal{F}$, so $\beta(s^t, \pi) \in \mathcal{F}$. Well-known results in Bayesian statistics guarantee, under general conditions, that one can find a prior π such that $E_P \beta(s^t, \pi) \geqslant E_P \varphi(s^t, \mathcal{F})$ for every $P \in \mathcal{P}_{\mathcal{F}}$.[11] That is, the Bayesian procedure with prior β does weakly better than the frequentist procedure φ at each distribution in $\mathcal{P}_{\mathcal{F}}$.

When we apply the above to a set of rules with $\Delta_t(\mathcal{F}) < \epsilon$, the behavior of a Bayesian with these beliefs will, by construction, display coarse decision making. A Bayesian's desire for simplicity or coarseness is now the result of his selection of a 'simple' prior. Our framework may therefore be viewed as providing a motivation for selecting priors with particular properties; namely those that put most of the mass on the sets of the form $P \in \mathcal{P}_{\mathcal{F}}$ where $\Delta_t(\mathcal{F}) < \epsilon$.

Two points should be emphasized. First, the improvement achieved by the Bayesian procedure over φ is marginal: by design, we have $E_P \varphi(s^t, \mathcal{F}) > E_P \beta(s^t, P) - \epsilon$ for all P, because the robust frequentist rule was selected to be (approximately) optimal for all distributions. For this reason, we shall continue using the simpler procedure φ, even though it can be (marginally) improved on by some Bayesian procedure. Second, mitigating overfitting as a criterion for prior selection is not inconsistent with the foundations of Bayesian theory (nor is it implied by these foundations). The motivation we provided for this criterion reflects classical statistics concerns (e.g., regression).

Appendix A. Statistical learning theory

Readers familiar with statistical learning theory can skip this section without any loss. For the reader's convenience, we collect here key definitions and theorems needed for our results. There are excellent textbook accounts for readers who wish to see a more thorough treatment; for example, Vapnik [31] or Devroye et al. [8].

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

$$x_i \in C_b \quad \Leftrightarrow \quad b_i = 1.$$

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

[11] This type of results is known as *complete class theorems*; see Berger [3].

Definition 3. The Vapnik–Chervonenkis dimension of $\mathcal{C}$, $VC(\mathcal{C}) = d$ if there exists $(x_1, x_2, \ldots, x_d) \in X^d$ such that $\mathcal{C}$ shatters $(x_1, x_2, \ldots, x_d)$, and there does not exist any $(x_1, x_2, \ldots, x_d, x_{d+1}) \in X^{d+1}$ such that $\mathcal{C}$ shatters $(x_1, x_2, \ldots, x_d, x_{d+1})$.

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 1. *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 $t > 0$:*

$$\sup_P P^t\left\{s^t\colon \sup_{A\in\mathcal{C}}\left|\nu\left(s^t\right)(A) - P(A)\right| > \epsilon\right\} \leqslant K\, t^d e^{-t\epsilon^2/32}, \tag{7}$$

where K is a universal constant.[12]

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 $\mathbf{F} = \{f \mid f : X \to \{0, 1\}\}$, and suppose the decision maker considers $\mathcal{F} \subseteq \mathbf{F}$. For any $f \in \mathcal{F}$, and any true probability distribution P:

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

Define $\mathcal{X} = X \times \{0, 1\}$ and $\mathcal{C} = \{A \mid A = (f^{-1}(0) \times \{0\}) \cup (f^{-1}(1) \times \{1\}),\ f \in \mathcal{F}\}$. Therefore, it follows from Theorem 1 (see also Corollary 12.1 of Devroye et al. [8]):

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

Next we turn to Pollard's pseudo-dimension. In the case of a more general Y, A, u, the Vapnik–Chervonenkis bounds do not directly apply.

We will use Pollard's pseudo dimension, sometimes referred to in the literature as the Pollard dimension (Pollard [21]). Let $\mathcal{F}$ be some set of functions from X to $\mathcal{R}$. We say that $\mathcal{F}$ *pseudo-shatters* a string $(x_1, x_2, \ldots, x_d)$ if there exists $c = (c_1, \ldots, c_d) \in \mathcal{R}^d$ such that for each $b = (b_1, \ldots, b_d) \in \{0, 1\}^d$, there exists $f_b \in \mathcal{F}$ satisfying[13]:

$$\forall 1 \leqslant i \leqslant d\colon \quad f_b(x_i) > c_i \quad \text{iff} \quad b_i = 1.$$

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

[12] Tighter bounds are available, but the above version is sufficient for our purposes, see also Devroye et al. [8].

[13] 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 [13] restated in our notation:

Theorem 2. *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 $t > 0$:*

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

Appendix B. Proofs

B.1. Observation 1

Proof of Observation 1. To simplify notation define $f_P^{\mathcal{F}}$ as the best decision rule in $\mathcal{F}$ if P was known, i.e. $f_P^{\mathcal{F}} = \arg\max_{f\in\mathcal{F}} E_P f$. Note that:

$$\begin{aligned}
&\int_{s^t}\left(\max_{f\in\mathcal{F}} E_P f - E_P\varphi\left(s^t,\mathcal{F}\right)\right) dP^t\\
&\quad= \int_{s^t}\left(E_P f_P^{\mathcal{F}} - E_P\varphi\left(s^t,\mathcal{F}\right)\right) dP^t\\
&\quad= \int_{s^t}\Big(\underbrace{E_P f_P^{\mathcal{F}} - E_{\nu(s^t)} f_p^{\mathcal{F}}}_{(1)} + \underbrace{E_{\nu(s^t)} f_p^{\mathcal{F}} - E_{\nu(s^t)}\varphi\left(s^t,\mathcal{F}\right)}_{(2)}\\
&\qquad + \underbrace{E_{\nu(s^t)}\varphi\left(s^t,\mathcal{F}\right) - E_P\varphi\left(s^t,\mathcal{F}\right)}_{(3)}\Big) dP^t\\
&\quad\leqslant \epsilon + 0 + \epsilon = 2\epsilon.
\end{aligned}$$

However, terms (1) and (3) are weakly less than ϵ from the definition of $\Delta_t(\mathcal{F})$, while term (2) $\leqslant 0$ by the definition of φ. Summing together, we have the required conclusion. □

B.2. Preliminaries

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

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

$$\forall\epsilon > 0\colon\quad \mathbb{P}(Z > \epsilon) \leqslant c\epsilon^{-2d} e^{-k\epsilon^2},$$

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

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

Proof. Since, for all $\epsilon > 0$,

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

we have that:

$$\begin{aligned}
\mathbb{E}\big(Z^2\big) &= \int_0^\infty P\big(Z^2 > t\big)\, dt \\
&= \int_0^u P\big(Z^2 > t\big)\, dt + \int_u^\infty P\big(Z^2 > t\big)\, dt \quad \forall u > 0 \\
&\leqslant u + \int_u^\infty P\big(Z^2 > t\big)\, dt \\
&\leqslant u + \int_u^\infty ct^{-d} e^{-kt}\, dt \\
&\leqslant u + cu^{-d} \int_u^\infty e^{-kt}\, dt \\
&= u + u^{-d} \frac{c}{k} e^{-uk}. \qquad (9)
\end{aligned}$$

Plugging $u = \frac{d \ln ck}{k}$ into inequality (9), we have:

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

Finally note that $\mathbb{E}(Z) \leqslant \sqrt{\mathbb{E}(Z^2)}$ by Jensen's inequality, giving us the desired result. □

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

$$\Delta_t(\mathcal{F}) \leqslant 8\sqrt{\frac{d^2 \ln 32e + d \ln \frac{te}{8}}{t}}. \qquad (10)$$

Proof. By Pollard's inequality, (8):

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

$$= \left(8(32e)^{2d}\right)\epsilon^{-2d}e^{-\frac{\epsilon^2 t}{64}}.$$

Then (10) follows from Lemma 1. □

Next, recall that Pollard's pseudo-dimension applies to real-valued functions. Given the decision maker's utility function u, any decision rule $f : X \to A$ induces a real valued function $u_f : X \times Y \to \mathcal{R}$, $u_f(x, y) = u(y, f(x))$; 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 3. *Let $\kappa : X \to \{1, \ldots, K\}$ be a categorization rule, and $\mathcal{F}_\kappa$ be the associated categorization frame. For any utility function $u : Y \times A \to \mathcal{R}$, the pseudo dimension of $\mathcal{F}_\kappa$ is at most $K|Y|$.*

Proof. 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 $1 \leqslant k \leqslant K$, $\mathcal{F}_\kappa$ 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 $(x_i, y_i) \in (\kappa^{-1}(k) \times Y)$, $i = 1, \ldots, |Y| + 1$. Let the associated cutoffs be $c_i \in \mathcal{R}$, $i = 1, \ldots, |Y| + 1$, without loss of generality let $c_1 \leqslant c_2 \leqslant \cdots \leqslant c_{|Y|+1}$.

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 $u_f(x_i, y_i) \leqslant c_i$, and therefore $\mathcal{F}_\kappa$ cannot shatter it. □

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

B.3. *Proposition 1*

Proof of Proposition 1. We first prove the former part, i.e. (5). 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 $\forall P \in \mathcal{P}_{\kappa, y_1, y_2}$:

$$\forall k \in \{1, \ldots, K\}: \quad P\left(y_1 | \kappa^{-1}(k)\right) = 1 \vee P\left(y_2 | \kappa^{-1}(k)\right) = 1.$$

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 et al. [8]. In our notation, it states that[14]:

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

[14] Note that the VC-dimension of the class of categorization rules with K categories is K.

However,

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

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

$$K \leqslant \frac{2et^2}{(t-1)\delta}\epsilon + 1. \tag{11}$$

(5) follows by setting k^+ to the right hand side of the above inequality.

To see the latter part, i.e. (6), by Lemma 3, the pseudo-dimension of a categorization-based rule $\mathcal{F}_\kappa$ with K partitions is at most $K|Y|$. Therefore, applying Lemma 2,

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

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

B.4. *Proposition 2*

Proof of Proposition 2. From the proof of Proposition 1, we see that

$$k^+(\epsilon, t) = \frac{2et^2}{(t-1)\delta}\epsilon + 1.$$

From the definition of u_θ it follows that $\delta = \theta$. Therefore

$$k^+_\theta(\epsilon, t) = \frac{2et^2}{(t-1)}\frac{\epsilon}{\theta} + 1.$$

As a result $k^+_\theta(\epsilon, t)$ is decreasing in θ. Further, it follows from observation that $k^+_\theta(\epsilon, t) = k^+_1(\frac{\epsilon}{\theta}, t)$. □

B.5. *Proposition 3*

Proof of Proposition 3. The proof follows by constructing a particular categorization problem, and a distribution P such that even the *expected* payoff of a decision maker following our decision procedure will be worse with the finer partitional rule.

Consider a partition X into k sub-blocks of $\frac{N}{k}$ elements each, $X_1, X_2, \ldots, X_k$. Partition $\mathcal{Q}_1$ is the k-element partition:

$$\mathcal{Q}_1 = \{X_1, X_2, X_3, \ldots, X_k\},$$

with κ_1 as the corresponding decision frame.

Define $\mathcal{Q}_2$ as the finest possible partition of X, i.e. each partition element is a single element in X, with κ_2 as the corresponding decision frame. Clearly:

$$\mathcal{Q}_1 \vee \mathcal{Q}_2 = \mathcal{Q}_2 = \big\{\{x_1\}, \{x_2\}, \ldots, \{x_N\}\big\}.$$

Finally to specify the distribution P, it is enough to define the marginal distribution on X and the conditional distribution on Y for each $x \in X$. We define P as follows: the marginal distribution on X is uniform, i.e. any x occurs with probability $\frac{1}{N}$. The conditional distribution on Y is specified thus: regardless of x, $P(y = 1) = p > \frac{1}{2}$.

As a result, the optimal rule if the decision maker knew P is to always take the action $a = 1$. This rule has an expected payoff of p.

Firstly, note that by Theorem 1, the decision maker's payoff when using the decision rule implied by the frame κ_1 is such that:

$$E_P\varphi\big(s^t, \kappa_1\big) \geqslant p - 16\sqrt{\frac{k\log t + 4}{2t}}.$$

However, if the decision maker switches to the finer frame $\kappa_1 \vee \kappa_2$, his expected payoff is

$$E_P\varphi\big(s^t, \kappa_1 \vee \kappa_2\big) \leqslant p\frac{t}{N} + \frac{1}{2}\frac{N-t}{N}.$$

To see why, note that the data contains at most t unique observables x. Since the partitions are single elements, for every situation they have not seen in the past, he can only guess the action to take, with expected payoff 0.5. Therefore the expected payoff is upper-bounded by the payoff when the agent takes the 'correct' action for ever x he sees in the data, and guesses otherwise. This results in the inequality above.

It follows that, fixing k and t, for N large enough, the expected payoff to an agent from using the finer frame $\kappa_1 \vee \kappa_2$ is strictly less than the expected payoff from using the coarser frame κ_1. □

References

[1] N.I. Al-Najjar, Decision makers as statisticians: Diversity, ambiguity and learning, Econometrica 77 (2009) 1339–1369.
[2] N. Barberis, A. Shleifer, Style investing, J. Finan. Econ. 68 (2) (2003) 161–199.
[3] J. Berger, Statistical Decision Theory and Bayesian Analysis, Springer, 1985.
[4] R. Bernstein, Style Investing: Unique Insight into Equity Management, John Wiley and Sons, 1995.
[5] M. Chi, P. Feltovich, R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. 5 (2) (1981) 121–152.
[6] J. Cremer, L. Garicano, A. Prat, Language and the theory of the firm, Quart. J. Econ. 122 (1) (2007) 373–407.
[7] D. Cutler, J. Poterba, L. Summers, What moves stock prices?, J. Portfol. Manage. 15 (3) (1998) 4–12.
[8] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, Berlin, 1996.
[9] E. Dimson, S. Nagel, Seeking out Investment Value in Style, London Business School, 2002.
[10] R.A. Dye, Costly contract contingencies, Int. Econ. Rev. 26 (1985) 233–250.
[11] R. Goldstone, The role of similarity in categorization: Providing a groundwork, Cognition 52 (2) (1994) 125–157.
[12] G. Harman, S. Kulkarni, Reliable Reasoning: Induction and Statistical Learning Theory, MIT Press, 2007.
[13] D. Haussler, 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, 1995.
[14] H. Hong, J. Stein, Disagreement and the stock market, J. Econ. Perspect. 21 (2) (2007) 109–128.
[15] H. Hong, J. Stein, J. Yu, Simple forecasts and paradigm shifts, J. Finance 62 (3) (2007) 1207–1242.
[16] P. Jehiel, Analogy-based expectation equilibrium, J. Econ. Theory 123 (2) (2005) 81–104.

[17] E. Kandel, N. Pearson, Differential interpretation of public signals and trade in speculative markets, J. Polit. Economy 103 (4) (1995) 831–872.
[18] E. Mohlin, Optimal categorization, Technical Report, Oxford University, 2013.
[19] G.L. Murphy, D.L. Medin, The role of theories in conceptual coherence, Psychol. Rev. 92 (3) (1985) 289–316.
[20] R.R. Nelson, S.G. Winter, An Evolutionary Theory of Economic Change, Harvard University Press, Cambridge, MA, 1982.
[21] D. Pollard, Empirical Processes: Theory and Applications, IMS, 1990.
[22] S. Reed, Pattern recognition and categorization, Cogn. Sci. 3 (3) (1972) 382–407.
[23] L. Rips, Similarity, typicality, and categorization, in: S. Vosniadou, A. Ortony (Eds.), Similarity and Analogical Reasoning, Cambridge University Press, 1989, pp. 21–59.
[24] E. Rosch, B. Lloyd, Cognition and Categorization, Lawrence Erlbaum Associates, Hillsdale, 1976.
[25] L. Samuelson, Analogies, adaptation, and anomalies, J. Econ. Theory 97 (2) (2001) 320–366.
[26] W. Sharpe, Asset allocation: Management style and performance measurement, J. Portfol. Manage. 18 (2) (1992) 7–19.
[27] H.A. Simon, A behavioral model of rational choice, Quart. J. Econ. 69 (1) (1955) 99–118.
[28] H.A. Simon, Theories of decision-making in economics and behavioral science, Amer. Econ. Rev. 49 (3) (1959) 253–283.
[29] C. Sims, Implications of rational inattention, J. Monet. Econ. 50 (3) (2003) 665–690.
[30] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, Science 185 (4157) (1974) 1124–1131.
[31] V.N. Vapnik, Statistical Learning Theory, John Wiley and Sons Inc., New York, 1998.
[32] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (1971) 264–280.
[33] S. Vosniadou, A. Ortony, Similarity and Analogical Reasoning, Cambridge University Press, 1989.