Theoretical Economics 15 (2020), 1279–1305 1555-7561/20201279

On the falsifiability and learnability of decision theories

Pathikrit Basu
Division of the Humanities and Social Sciences, Caltech

Federico Echenique
Division of the Humanities and Social Sciences, Caltech

We study the degree of falsifiability of theories of choice. A theory is easy to falsify if relatively small data sets are enough to guarantee that the theory can be falsified: the Vapnik–Chervonenkis (VC) dimension of a theory is the largest sample size for which the theory is "never falsifiable." VC dimension is motivated strategically. We consider a model with a strategic proponent of a theory and a skeptical consumer, or user, of theories. The former presents experimental evidence in favor of the theory; the latter may doubt whether the experiment could ever have falsified the theory.

We focus on decision-making under uncertainty, considering the central models of expected utility, Choquet expected utility, and max–min expected utility models. We show that expected utility has VC dimension that grows linearly with the number of states, while that of Choquet expected utility grows exponentially. The max–min expected utility model has infinite VC dimension when there are at least three states of the world. In consequence, expected utility is easily falsified, while the more flexible Choquet and max–min expected utility are hard to falsify. Finally, as VC dimension and statistical estimation are related, we study the implications of our results for machine learning approaches to preference recovery.

Keywords. Revealed preference theory, decision theory, machine learning.

JEL classification. C1, D1.

1. Introduction

We consider the smallest sample size needed for theories of choice to be falsifiable, even when the experimenter engages in selective experimental design. With small sample sizes it is harder to falsify a theory than with a large sample size: so for any given theory of choice, it makes sense to consider the smallest sample size needed for the theory to be falsifiable. Imagine two agents: a proponent and a consumer (or user) of theories of choice. The proponent designs a choice experiment and may present the consumer with

Pathikrit Basu: pathikritbasu@gmail.com
Federico Echenique: fede@hss.caltech.edu
We thank Fabio Maccheroni, Burkhard Schipper, and Adam Wierman for useful comments on a previous draft. We are also very grateful to the three anonymous referees for their suggestions and feedback. Echenique acknowledges NSF funding through Grants SES-1558757 and CNS-1518941, as well as the Linde Institute of Economic and Management Science at Caltech.

 https://doi.org/10.3982/TE3438

evidence in favor of the theory. The consumer is suspicious about the proponent's experimental design. There may exist designs for which no possible outcome could falsify the theory in question. The consumer's suspicion can only be addressed through a large enough sample size, because with enough data, it becomes difficult for the proponent to successfully manipulate the experimental design through selective testing.

Our paper calculates the smallest sample size needed to rule out manipulations; we focus on a setting that accommodates a general theory of preference relations and data from binary choice experiments. This smallest sample size is termed the *Vapnik–Chervonenkis* (VC) *dimension* of a theory (Vapnik and Chervonenkis 1971, Blumer et al. 1989). Roughly speaking, the VC dimension of a model is the largest cardinality of a data set that the theory can always rationalize. An experiment of given size that validates a theory with small VC dimension is more convincing than one that validates a theory with large VC dimension, as the former was, in principle, easier to falsify with the data at hand than the latter.

While the setting considered in the paper applies more generally, we focus on choice under uncertainty, and study resulting implications of falsifiability in experimental outcomes and VC dimension estimates. We first proceed to give a brief overview of choice under uncertainty and then describe our main results. It is hard to overstate the importance of the theory of choice under uncertainty. Many important models of economic behavior, markets, and institutions deal with the existence of uncertainty, and assume that agents conform to some theory of choice under uncertainty.[1] The most common model is subjective expected utility: economic agents choose among uncertain prospects as if they assigned a probability distribution to the different possible, and uncertain, events. Given a probability distribution, which is subjective and not observable, agents seek to maximize the expected reward obtained under each prospect. Subjective expected utility was famously axiomatized by Savage (1972).

While ubiquitous, subjective expected utility has some notable problems. Agents' attitude toward uncertainty is not always well captured by a probability distribution over uncertain events. The best known problems are illustrated by the Ellsberg paradox (Ellsberg 1961), a thought experiment in which agents' choices cannot be accommodated by a probability distribution because they exhibit *ambiguity aversion*. The Ellsberg paradox illustrates that while an agent may place a premium on events that have an objectively known probability, such a premium turns out to be incompatible with a probability distribution over unknown events. In response, decision theorists have sought to generalize the theory of subjective expected utility to allow for ambiguity aversion. The two best known alternatives are the models of max–min expected utility and Choquet expected utility. Hence, the outcome of the Ellsberg experiment, while falsifying subjective expected utility, is rationalizable by the max–min and Choquet models.

The model of max–min expected utility postulates agents who possess multiple probability distributions over uncertain events, giving each uncertain prospect multiple expected values. In the max–min theory, agents seek to maximize the minimum

[1] We do not address models of choice under risk, meaning choices over prospects that have stochastic consequences, with known and objective probabilities. Our paper focuses on (Knightian) uncertainty.

expected value. Given an uncertain prospect, the agent evaluates it in adversarial, pessimistic fashion, using the worst-case probability distribution in her set of possible distributions. By using more than one probability measure, it is easy to explain the Ellsberg paradox through the max–min model. Max–min was first axiomatized by Gilboa and Schmeidler (1989); it was also proposed and used in the statistical decision literature; see Wald (1950) and Huber (1981). The max–min model is a staple of modern decision theory and is used extensively in economic applications where agents face uncertainty.

Choquet expected utility assumes that agents have nonadditive beliefs over uncertain events. Instead of additive probability measures, as in the model of subjective expected utility, agents' beliefs are represented by a possibly nonadditive *capacity*. In Choquet expected utility theory, agents evaluate uncertain prospects according to the Choquet expectation with respect to their capacity. The Choquet model can accommodate the types of aversions to ambiguity exhibited in the Ellsberg paradox because nonadditivity allows an agent to place a premium on events that are less ambiguous than others. The model was first axiomatized by Schmeidler (1989).

The three models we have described are arguably the most important models of decision-making under uncertainty. Our purpose in the present paper is to understand the VC dimension of the theories of subjective expected utility, max–min, and Choquet expected utility. In all three cases, we assume an agent who is risk-neutral. If we were to include the utility function as an additional parameter of the theory, then the VC dimension of all three models would increase. In an effort to isolate the role of the prior beliefs in the theory, we focus here on the risk-neutral version of the three models.

For subjective expected utility theory, we show that the VC dimension is linear in the number of states of the world. This means that the sample size that is needed for a consumer of the theory to be convinced of theory-confirming evidence, beyond the dangers of manipulation, is relatively small. For example, for an experiment with 10 states of the world, a data set of size 11 suffices, and this number grows linearly with the number of states of the world.

The max–min expected utility (MEU) theory does not fare as well in our analysis as expected utility theory. It turns out that as long as there are at least three possible states of the world, then the MEU has infinite VC dimension. In consequence, *no* sample size can guarantee a consumer of the theory that evidence in favor of MEU could not have been the result of an experimental design that never had a chance of falsifying the theory. Put differently, the consumer's suspicions cannot be assuaged by empirical evidence, no matter how large the sample.

Of course, there are ways to avoid our negative conclusion. For example, the analyst could restrict the possible parameters of the theory. If the model is constrained by imposing additional restrictions on its parameters, then it will require smaller sample sizes to test. So our results can be read as saying that such additional restrictions are indeed needed. Another possibility is that the analyst has substantive information as to the process that generates the data. She could have knowledge of, or control over, the way in which the agent is presented with alternatives to choose from. In that case, one could read our results as saying that such knowledge is essential and must be precise.

	Learnable	VC dimension
Expected utility	✓	Linear
Choquet expected utility	✓	Exponential
Max–min (states > 2)	X	$+\infty$
Max–min (two states)	✓	—

TABLE 1. Summary of results.

The Choquet expected utility model fairs better than MEU: it has finite VC dimension and is, therefore, testable beyond a proponent's incentives to manipulate through selective testing. Unfortunately, Choquet has VC dimension that grows exponentially in the number of states. Therefore, it requires that the agent makes a sample size that is exponential in the number of states. To use the example we mentioned above, if we consider 10 states of the world, then we would now need a sample size of more than 250, and if the states of the world were 20, we would need more than 180,000 observations. In consequence, it quickly becomes impractical to assuage the consumer's suspicions by way of sample size.

Table 1 has a summary of our results.

We should emphasize that our paper focuses on the combination of risk neutrality and the absence of parametric restrictions on beliefs. If one adds a parametric model for beliefs, then the VC dimension will naturally decrease. The assumption of risk neutrality can be avoided if one works in an Anscombe–Aumann setting. The paper is also based on choices from binary menus. A brief discussion of more general choice functions is in Section 4.1.

Learning interpretation A second interpretation of our results is in terms of overfitting. Given a theory and a data set, we may want to fit an instance of the theory to the data or "learn" an instance of the theory. We adopt the paradigm of "probably approximately correct" (PAC) learning (Valiant 1984, Blumer et al. 1989) and assume an agent who is choosing among pairs of uncertain prospects. The question is whether choices made according to the theory of choice under uncertainty allow an outside analyst to recover the model of choice with high probability and in the limit as the number of choices made by the agent grows.

We are motivated by the notion that some models in behavioral economics are generalizations, meaning that they were formulated by relaxing relatively stringent economic models; this can make them prone to *overfitting*. Overfitting as a concern seems to be new in decision theory and behavioral economics. Economists are used to the idea that lax models may lead to theories that have few testable implications, but there are other potential dangers when working with flexible behavioral models. Consider an economist who fits a model to choice data, perhaps observed from an agent making choices in a laboratory experiment. If the model is very general and flexible, meaning that it contains many special cases, and can accommodate many particular behaviors, then it is possible that the economist fits a model that is too closely adapted to the ob-

served data. As a result, the model could then perform badly out of sample. The theory of PAC learning that we use in this paper seeks precisely to capture the presence or absence of overfitting.

A model with finite VC dimension is learnable; moreover, the VC dimension of the model controls the sample size needed to fit an instance of the model that is guaranteed to perform well in terms of out-of-sample predictions. So by computing the VC dimension of the model, we can understand how prone it is to overfitting.

Using the language of PAC learning, then, our results mean that subjective expected utility theory is learnable with relatively small sample sizes and is not prone to overfitting. MEU is not learnable: no matter how many choices are made by our subject, the analyst will be unable to learn the model generating choices with high probability. Put differently, the family of possible max–min parameters (the sets of multiple probabilities) is large and flexible enough that even with very large data sets, the model can wrap itself close to the data, while predicting badly out of sample. The Choquet expected utility model is PAC learnable, but it is still susceptible to overfitting because an analyst will require extremely large data sets so as to obtain good out-of-sample predictions.

We are not the first researchers to study PAC learning in economic models. Kalai (2003) considers a choice function, and connects learnability to substantive properties of choice. Beigman and Vohra (2006), Zadimoghaddam and Roth (2012), and Balcan et al. (2014) consider learning in the classical demand environment. Some of their results relate to learning linear utility functions, which is a point in common with our work, but none of these papers studies questions of choice under uncertainty. Our primitive model of choice is a preference relation, in contrast to demand behavior. As a result, our model of choice is in line with common practice in decision theory and experimental economics, where agents make choices over pairs of objects. The model of choice in the cited papers is more in line with the practice in the revealed preference theory of consumption. Last, in a different context, Salant (2007) considers PAC learning of the majority rule in a setting involving social choice.

We also point out prior work in the literature on estimating models of ambiguity aversion. The papers by Ahn et al. (2014), Chamberlain (2000), Mangelsdorff and Weber (1994), and Camerer and Weber (1992) all involve estimating the parameters of ambiguity aversion models based on real data from experiments and financial markets. The paper by Ahn et al. (2014) considers portfolio choice experiments, whereas Chamberlain (2000) undertakes an econometric analysis of the max–min model in the context of an autoregressive model with panel data. Mangelsdorff and Weber (1994) study the Choquet model and conduct experimental tests for various hypotheses regarding the capacities in the model. Preceding research in this vein was done by Camerer and Weber (1992), who also expand on various experimental and empirical applications of the ambiguity aversion models. Finally, large-scale experimental data sets on pairwise choice are available in, for example, Falk et al. (2018), Chapman et al. (2017), and Chapman et al. (2018).

2. Falsifiability and sample size

2.1 *Preliminary definitions*

Let X be a Euclidean space, endowed with its Borel σ-algebra $\mathcal{X}$. Denote the product space $X \times X$ by Z.

A *preference relation* on X is any binary relation $\succsim \subseteq Z$ such that $\succsim$ is measurable with respect to the product σ-algebra $\mathcal{Z}$ on Z. Denote by $\mathcal{P}^*$ the set of all preference relations on Z. A *theory* is a subset $\mathcal{P} \subseteq \mathcal{P}^*$. For example, the set of weak orders (complete and transitive preferences) is a theory. The set of preferences that have a linear, or a Cobb–Douglas, utility representation is another theory.

2.2 *A model*

Consider a conflict between two agents: Alice, a theorist, and Bob, an editor of a journal. Alice has a theory and wants to convince Bob that it is valid empirically. To that end, she runs an experiment and tries to use her theory to *rationalize* the outcome of the experiment. In our setting, Alice's theory will be about choice under uncertainty. The experiment will involve a subject making choices from a sequence of choice problems. Now it should be said that Alice has an agenda: she wants to prove her theory right. In pursuit of validating her theory, Alice cannot tamper with the subject's choices, but she can design the experiment so as to make any experimental result easier to rationalize by her theory.

Fix a finite set Ω of *states of the world*. Alice's theory of choice is about state-contingent monetary payoffs, that is, elements of $X = \mathbb{R}^{\Omega}$.[2] Her theory consists of a set $\mathcal{P}'$ of preference relations over X. For example, $\mathcal{P}'$ could consist of all the expected utility preferences over X or all the max–min expected utility preferences.

Now fix a number k, which we think of as a sample size. Alice sets up a choice experiment of size k, meaning a questionnaire

$$(x_1, y_1), \ldots, (x_k, y_k)$$

of size k. In other words, Alice chooses a "sample" of k pairwise choice situations $z_i = (x_i, y_i) \in Z$ that she presents to the subject in her experiment.

Given the questionnaire, the subject selects a choice from each problem. For example, x_1 from $z_1 = (x_1, y_1)$, y_2 from $z_2 = (x_2, y_2)$, and so on. We use an indicator function a_i to record the subject's choices from each problem. So $a_1 = 1$ because x_1 was chosen from z_1 and $a_2 = 0$ because x_2 was not chosen from z_2. Formally, a *size-k data set* is any finite sequence $D \in \bigcup_{n \geq 1} (Z \times \{0, 1\})^k$, so a data set takes the form

$$D = \big((z_1, a_1), (z_2, a_2), \ldots, (z_k, a_k)\big),$$

where $a_i \in \{0, 1\}$. The sequence D is interpreted as follows: for each i, if $z_i = (x_i, y_i)$, then the subject was asked to choose one of the alternatives in the set $\{x_i, y_i\}$, and $a_i = 1$ if and only if x_i is the alternative chosen.

[2] We use $\mathbb{R}^{\Omega}$ and $\mathbb{R}^{|\Omega|}$ interchangeably.

Finally, Alice, our theorist, tries to find an element $\succsim \in \mathcal{P}'$ of her theory to explain the data. We say that a size k data set $D = ((z_1, a_1), (z_2, a_2), \ldots, (z_k, a_k))$ is *rationalizable by* $\mathcal{P}'$ if there exists $\succsim \in \mathcal{P}'$ for which $a_i = 1$ if and only if $x_i \succsim y_i$, $1 \leq i \leq k$.

Imagine that Alice triumphantly offers Bob an experimental data set D_k with a rationalizing $\succsim \in \mathcal{P}'$. How should Bob react? Like any good editor, Bob is a skeptic. He worries that Alice has designed her experiment to make it easier for any subject to make choices that are consistent with her theory. Specifically, say that a questionnaire $\{z_1, \ldots, z_k\}$ is *always rationalizable by* $\mathcal{P}'$, or *shattered by* $\mathcal{P}'$, if for any vector $(a_1, a_2, \ldots, a_n) \in \{0, 1\}^k$, there exists a preference $\succsim \in \mathcal{P}'$ that rationalizes the data set $((z_1, a_1), (z_2, a_2), \ldots, (z_k, a_k))$. A questionnaire that is always rationalizable can never falsify the theory $\mathcal{P}$ because no matter how the subject chooses from the questionnaire, the choices can be explained by the theory. In other words, the questionnaire could never detect a violation of the theory.[3]

In sum, Alice has an agenda and, therefore, Bob is suspicious of the empirical evidence produced by Alice. How can the trust issues between Alice and Bob be resolved? The answer is simple and lies in the number k. When the sample size k is small, it is going to be easier for Alice to cook up a questionnaire that is always rationalizable by $\mathcal{P}'$. When k is large, there may exist subject choices that cannot be rationalized by $\mathcal{P}'$, thereby falsifying the theory. The crucial criterion is then the *smallest* k for which there exists a questionnaire that is always rationalizable by $\mathcal{P}'$, or (in other, more standard terminology) "shattered" by $\mathcal{P}'$. This number is called the Vapnik–Chervonenkis (VC) dimension of $\mathcal{P}'$.[4]

Formally, the *VC dimension* of a theory $\mathcal{P}'$, denoted as $\mathrm{VC}(\mathcal{P}')$, is defined as

$$\mathrm{VC}(\mathcal{P}') = \max\{k : \exists (z_i)_{i=1}^k \text{ which can be shattered by } \mathcal{P}'\}.$$

The VC dimension of a theory may be infinite. For example, suppose that $X = \mathbb{R}$ and let $\mathcal{P}_R$ be the set of rational preferences, i.e., all complete and transitive preference relations. This class of preferences has infinite VC dimension. To see this, let k be a given data size and select the z_is in $\mathbb{R}^2$ in such a way that for all $i \neq j$, it is the case that $x_i \neq y_i$, $x_i \neq x_j$, and $y_i \neq y_j$. Now, in a data set, no matter how the z_is are labelled by the a_is, we can always find a rational preference relation to rationalize the data.

Here is another example, this time of a theory with finite VC dimension. Again let $X = \mathbb{R}$, but now let $\mathcal{P}_{\mathrm{SP}}$ be the set of single-peaked preference relations. These are the preferences for which there exists a utility representation $u : X \to \mathbb{R}$ and a real number x^* (the "peak") such that u is strictly increasing in $(-\infty, x^*)$ and strictly decreasing in $(x^*, +\infty)$. Now consider (x_1, y_1) and (x_2, y_2), and suppose without loss of generality that $x_i < y_i$ for $i = 1, 2$. If $x_1 < y_2$, then it is easy to see that setting $a_1 = 1$ and $a_2 = 0$ cannot be rationalized by a single-peaked preference because $a_1 = 1$ means that the peak is to the

[3]See Chambers and Echenique (2016) for an overview of falsifiability and rationalizability. Kalai et al. (2002) provide a combinatorial approach to flexibility in rationalizing choice functions in the spirit of our paper. See also Rubinstein (1996).

[4]Our motivation for introducing VC dimension, by way of falsifiability and without resorting to its role in learning, is not new. See, for example, Vapnik (1998), Vapnik (2006), Corfield et al. (2005), and Corfield et al. (2009).

left of x_1, while $a_2 = 0$ implies that the peak is to the right of y_2. Alternatively, if $x_1 < y_2$ does not hold, then we must have $x_2 < y_2 \leq x_1 < y_1$, and the argument can be applied to $x_2 < y_1$.

In choosing and evaluating theories, and the experimental evidence in their favor, the VC dimension informs us of the degree of flexibility and falsifiability of the theory. Evidence confirming a theory that has a small VC dimension is more convincing than evidence in favor of a theory with large VC dimension. Moreover, a theory with infinite VC dimension or that has a VC dimension that grows very quickly with the description of the problem is clearly problematic, as it is very unlikely that a data set could ever disprove it.

Note that when offering evidence that supports her theory, Alice could argue that her experimental design could allow for the theory to be falsified. In a sense, our paper explores the theories and sample sizes for which such an argument is feasible.

2.3 *Theories of decisions under uncertainty*

We present a model of choice under uncertainty. Uncertainty is introduced through a *state space* Ω, a finite set. A subject chooses among uncertain prospects called acts. An *act* is a vector $x \in \mathbb{R}^{\Omega} =: X$. The interpretation is that the act x ensures a utility payoff $x(\omega)$ in state ω. A *preference relation* over acts is defined as a binary relation $\succsim \subseteq X \times X$. The preference relation encodes an agent's choices among pairs of acts. An exposition of the theory can be found in Kreps (1988) or Gilboa (2009).[5]

Throughout the paper, we restrict attention to preference relations that are *nontrivial*, meaning that there exists a pair $x, y \in X$ with $x \succsim y$ but not $y \succsim x$.

We say that two acts $x, y \in X$ are *co-monotonic* if there do not exist states ω, ω' such that $x(\omega) > x(\omega')$ but $y(\omega) < y(\omega')$.

We focus our attention on preferences that satisfy a subset of the following axioms.

Axiom 1. Order. *For all $x, y \in X$, either $x \succsim y$ or $y \succsim x$ (completeness). Moreover, for all $x, y, z \in X$, if $x \succsim y$ and $y \succsim z$, then $x \succsim z$ (transitivity).*[6]

Axiom 2. Independence. *For all $x, y, z \in X$, and all $\lambda \in (0, 1)$,*

$$x \succsim y \quad \textit{if and only if} \quad \lambda x + (1-\lambda)z \succsim \lambda y + (1-\lambda)z.$$

Axiom 3. Continuity. *For all $x \in X$, the upper and lower contour sets*

$$U_x = \{y \in X \mid y \succsim x\} \quad \textit{and} \quad L_x = \{y \in X \mid x \succsim y\}$$

are both closed subsets of X.

Axiom 4. Monotonicity. *For all $x, y \in X$, if $x(\omega) \geq y(\omega)$ for all $\omega \in \Omega$, then*

$$x \succsim y.$$

[5]Our methods can also be applied to choice under risk, with objectively known probabilities.

[6]A preference relation that satisfies completeness and transitivity is called a *weak order*.

Axiom 5. Co-Monotonic Independence. *For all $x, y, z \in X$ that are pairwise co-monotonic and for all $\lambda \in (0,1)$,*

$$x \succsim y \quad \textit{if and only if} \quad \lambda x + (1-\lambda)z \succsim \lambda y + (1-\lambda)z.$$

Axiom 6. C-Independence. *For all $x, y \in X$, any constant vector $c \in X$, and for all $\lambda \in (0,1)$,*

$$x \succsim y \quad \textit{if and only if} \quad \lambda x + (1-\lambda)c \succsim \lambda y + (1-\lambda)c.$$

Axiom 7. Uncertainty Aversion. *For all $x, y \in X$, for all $\lambda \in (0,1)$, if $x \sim y$, then*

$$\lambda x + (1-\lambda)y \succsim x.$$

In this paper, we consider the following models of decision under uncertainty.

(a) *Expected Utility Model.* There exists a probability measure $p \in \Delta^{|\Omega|-1} \subseteq \mathbb{R}^{\Omega}$ such that $x \succsim y$ if and only if

$$p \cdot x \geq p \cdot y.$$

A preference relation $\succsim$ belongs to this model if and only if it satisfies Axioms 1–4.

(b) *Choquet Expected Utility Model.* A *capacity* is defined as a set function $\nu : 2^{\Omega} \to [0,1]$ such that $\nu(\varnothing) = 0$ and $\nu(\Omega) = 1$; $\nu(E) \geq \nu(F)$ whenever $F \subseteq E$. The Choquet expectation of an act x with respect to ν, denoted by $\mathbb{E}_{\nu}$, is defined as

$$\mathbb{E}_{\nu}(x) = \int_{-\infty}^{0} \big[\nu\big(\{\omega : x(\omega) \geq q\}\big) - \nu(\Omega)\big]\, dq + \int_{0}^{\infty} \nu\big(\{\omega : x(\omega) \geq q\}\big)\, dq.$$

In the Choquet expected utility model, an agent evaluates acts according to their Choquet expectation. Hence, $x \succsim y$ if and only if

$$\mathbb{E}_{\nu}(x) \geq \mathbb{E}_{\nu}(y).$$

A preference belongs to the Choquet expected utility model if and only if it satisfies Axioms 1 and 3–5.

We say that a capacity ν is *convex* if it satisfies $\nu(A \cup B) + \nu(A \cap B) \geq \nu(A) + \nu(B)$ for all events $A, B \subseteq \Omega$. When ν is convex, the Choquet integral takes a specific form. There exists a compact convex set of probability measures $\mathrm{Core}(\nu) \subseteq \Delta^{|\Omega|-1}$ such that

$$\mathbb{E}_{\nu}(x) = \min_{p \in \mathrm{Core}(\nu)} p \cdot x.$$

In fact, the set $\mathrm{Core}(\nu)$ is defined as $\mathrm{Core}(\nu) = \{p \in \Delta^{|\Omega|-1} : p(A) \geq \nu(A) \text{ for all } A\}$. This brings us to the next model of preferences we consider, the max–min model.

(c) *Max–min Expected Utility Model.* There exists a compact, convex set of probability measures $C \subseteq \Delta^{|\Omega|-1}$ such that $x \succsim y$ if and only if

$$\min_{p\in C} p \cdot x \geq \min_{p\in C} p \cdot y.$$

The max–min expected utility model is characterized by Axioms 1, 3, 4, 6, and 7.

We use the following notation for the models of decision-making under uncertainty.

- The term $\mathcal{P}_{\mathcal{I}}$ denotes the set of preferences satisfying Axioms 1 and 2.
- The term $\mathcal{P}_{\mathcal{EU}}$ denotes the set of preferences satisfying Axioms 1–4.
- The term $\mathcal{P}_{\mathcal{CEU}}$ denotes the set of preferences satisfying Axioms 1 and 3–5.
- The term $\mathcal{P}_{\mathcal{MEU}}$ denotes the set of preferences satisfying Axioms 1, 3, 4, 6, and 7.

Note that $\mathcal{P}_{\mathcal{EU}}$, $\mathcal{P}_{\mathcal{CEU}}$, and $\mathcal{P}_{\mathcal{MEU}}$ correspond to the expected utility, Choquet expected utility, and multiple priors models, respectively. However, $\mathcal{P}_{\mathcal{I}}$ satisfies only Axioms 1 and 2, and is (it turns out) strictly larger than the expected utility model $\mathcal{P}_{\mathcal{EU}}$. Interestingly, the model $\mathcal{P}_{\mathcal{I}}$ itself has some nice properties. For any preference $\succsim \in \mathcal{P}_{\mathcal{I}}$, there exist finitely many vectors $q_1, \ldots, q_K$, where $K \leq |\Omega|$ (see, for example, Blume et al. 1991) such that

$$x \succsim y \quad \text{if and only if} \quad (q_k \cdot x)_{k=1}^K \geq_L (q_k \cdot y)_{k=1}^K,$$

where $\geq_L$ denotes the lexicographic ordering on $\mathbb{R}^K$. For any two vectors $u, v \in \mathbb{R}^K$, we say that $u \geq_L v$ if either $u = v$ or $u_l > v_l$, where $l = \min\{i : u_i \neq v_i\}$. When $\succsim$ additionally satisfies monotonicity, then we have $q_1, \ldots, q_k \in \Delta^{|\Omega|-1}$ and the resulting model is called the *lexicographic expected utility* model ($\mathcal{P}_{\mathcal{LEU}}$). Further, if continuity is also satisfied, then we have $K = 1$, which would be the expected utility model. Hence, $\mathcal{P}_{\mathcal{EU}} \subseteq \mathcal{P}_{\mathcal{LEU}} \subseteq \mathcal{P}_{\mathcal{I}}$. As the VC dimension is monotonic with respect to set inclusion, our result on the upper bound on the VC dimension of $\mathcal{P}_{\mathcal{I}}$ has implications for these two models as well.

2.4 *Main result*

For a model of preferences $\mathcal{P}'$, let $\mathrm{VC}(\mathcal{P}')$ denote its VC dimension. Our main result is the following theorem.

THEOREM 1. *Let $\mathcal{P}_{\mathcal{I}}$, $\mathcal{P}_{\mathcal{MEU}}$, and $\mathcal{P}_{\mathcal{CEU}}$ be as defined at the end of the last section.*

(i) We have $\mathrm{VC}(\mathcal{P}_{\mathcal{EU}}) = |\Omega|$. *And,* $\mathrm{VC}(\mathcal{P}_{\mathcal{I}}) \leq |\Omega| + 1$.

(ii) We have $\binom{|\Omega|}{|\Omega|/2} \leq \mathrm{VC}(\mathcal{P}_{\mathcal{CEU}}) \leq (|\Omega|!)^2(2|\Omega| + 1)$.

(iii) If $|\Omega| = 2$, *then* $\mathrm{VC}(\mathcal{P}_{\mathcal{MEU}}) = 2$.

(iv) If $|\Omega| \geq 3$, *then* $\mathrm{VC}(\mathcal{P}_{\mathcal{MEU}}) = +\infty$.

This means that the conflict between Alice and Bob can be resolved for expected utility with relatively small sample sizes. Choquet expected utility fares much worse. There are sample sizes at which the skeptic Bob may be persuaded by Alice, but these grow very quickly with the complexity of the problem as expressed by the number of states of the world. Finally, for MEU, as long as there are at least three states, there is *no* sample size that would convince Bob of the confirmatory message in Alice's data.

Our results should be qualified by the formulation of our theories. By assuming risk neutrality, Alice has tied her hands with respect to the utility over money. More general classes of utilities would imply larger VC dimensions. Alternatively, the sets of priors in MEU are "nonparametric," and one can imagine a parametric specification of MEU that achieves a smaller VC dimension. For example, if one assumes that the set of priors has a finite number of extreme points (it is a polytope) and if we have an upper bound on the total number of extreme points, then the VC dimension will be finite. In fact, one way to read our results is that parametric assumptions are needed for the credibility of empirical findings on models that have otherwise large VC dimension.[7]

3. Learning

The ideas we have introduced so far have traditionally been used in machine learning to fit models to data in ways that optimize their out-of-sample predictions. We adopt the notion of probably approximately correct (PAC) learning (Valiant 1984, Blumer et al. 1989) and assume a subject who is choosing among pairs of uncertain prospects. The question is whether choices made according to the theory of choice under uncertainty allow an outside analyst to recover the model of choice with high probability and in the limit as the number of choices made by the agent grows. As a consequence of Theorem 1, we see that the results are mixed: some models in the theory are learnable, while others are not.

We imagine a subject making choices from finitely many ordered pairs (x_i, y_i), $i = 1, \dots, n$. The subject's choices are recorded in a collection of *labels* $a_i \in \{0, 1\}$. As before, a *data set* is any finite sequence $D \in \bigcup_{n\geq 1}(Z \times \{0, 1\})^n$. The set of all data sets is denoted by $\mathcal{D}$. The set of all data sets of size n is denoted by $\mathcal{D}_n$.

The analyst assumes that the population of choice instances z is distributed according to an *unknown* probability distribution $\mu \in \Delta(Z)$. In other words, the analyst ignores the nature of the process by which the subject is presented with choice problems. All the analyst knows is that choice problems are selected in an independent and identically distributed (i.i.d.) fashion from a probability distribution μ on Z, but μ is unknown. We assume that μ has full support.

When the analyst observes a data set D, she makes a conjecture about the subject's preference $\succsim$. The objective of the analyst is to precisely learn the preference of the subject. A *learning rule* is a map $\sigma : \mathcal{D} \to \mathcal{P}$. For a data set D, $\sigma(D)$ is the preference relation that the analyst believes is guiding the subject's choices and is what the analyst will use to make out-of-sample predictions.

[7] The problem cannot, however, be reduced to parameter counting. There are parametric models with a single parameter and infinite VC dimension.

We denote by σ_n the restriction of σ to $\mathcal{D}_n$.

The analyst would like σ to be such that, for a data set of size n, if n is large, then *the out-of-sample predictions of the learning rule σ_n, should with high probability be accurate* (that is, close to the choices made by the underlying preference $\succsim$).

Specifically, we consider the distance between the choices made by a conjectured relation $\succsim'$ and $\succsim$ defined by $d_\mu(\succsim, \succsim') = \mu(\succsim \triangle \succsim')$, a pseudometric, where

$$\succsim \triangle \succsim' = \{(x, y) \in Z : x \succsim y \text{ and } x \not\succsim' y\} \cup \{(x, y) \in Z : x \not\succsim y \text{ and } x \succsim' y\},$$

and $\triangle$ denotes the symmetric difference between the preference relation between the preference relations $\succsim$ and $\succsim'$. Note that $\mu(\succsim \triangle \succsim')$ is essentially the probability that the choices made according to each of the preferences will differ.

Now, given a data set $D \in \mathcal{D}_n$, we want to control the size of the out-of-sample prediction error $d_\mu(\sigma_n(D), \succsim) = \mu(\sigma_n(D) \triangle \succsim)$. Note that, given D and σ_n, the error is deterministic. The data set D is, however, drawn at random according to n i.i.d. draws from μ. So the probability of an error of size larger than ϵ is

$$\mu^n\big(\{(x_1, y_1), \ldots, (x_n, y_n) \in Z^n : d_\mu\big(\sigma_n\big(((x_1, y_1), 1_{x_1 \succsim y_1}), \ldots, ((x_n, y_n), 1_{x_n \succsim y_n})\big), \succsim\big) > \epsilon\big)\}\big).$$

In words, the probability, according to μ, of drawing a sample $(x_1, y_1), \ldots, (x_n, y_n)$ such that, when labelled according to $\succsim$, σ predicts a preference that differs from $\succsim$ by more than ϵ. Below, we write this expression succinctly as $\mu^n(d_\mu(\sigma_n, \succsim) > \epsilon)$.

Learnability If the analyst believes that the subject's preferences are in some theory $\mathcal{P}'$, then she would choose a learning rule whose range lies in $\mathcal{P}'$.

We say that a theory $\mathcal{P}'$ is *learnable* if the analyst can design a learning rule such that whenever the subject's preference belongs to $\mathcal{P}'$, large samples of the subject's choices would allow him to have a precise estimate of the preference with high probability. A precise estimate means, in accordance with our previous discussion, that the out-of-sample predictions made according to the learning rule are accurate with high probability. Furthermore, this should be the case despite the analyst not knowing the distribution μ. We next formally define the notion of learnability we consider here.

Definition 1. A theory of preferences $\mathcal{P}' \subseteq \mathcal{P}$ is *learnable* if there exists a learning rule σ such that for all $(\epsilon, \delta) \in (0, 1)^2$, there exists an $N(\epsilon, \delta) \in \mathbb{N}$ such that for all $n \geq N(\epsilon, \delta)$,

$$\big(\forall \succsim \in \mathcal{P}'\big)\big(\forall \mu \in \Delta^f(Z)\big)\big(\mu^n\big(d_\mu(\sigma_n, \succsim) > \epsilon\big) < \delta\big), \tag{1}$$

where $\Delta^f(Z)$ is the set of all full support probability measures on Z, μ^n represents the product measure induced by $\succsim$ and μ on $(Z \times \{0, 1\})^n$, and σ_n is the prediction made by the learning rule σ on a data set of size n.

It is important to note the role of $N(\epsilon, \delta)$ in the definition above. It represents a lower bound on the number of samples needed for the condition in (1) to hold under the learning rule σ. We are interested in the *sample complexity*[8] of a learning rule, which is

[8]See, for example, Shalev-Shwartz and Ben-David (2014).

a function $N_\sigma : (0,1)^2 \to \mathbb{N}$, such that for all ϵ, δ, the $N_\sigma(\epsilon,\delta)$ is the *minimum* number of samples, n, such that (1) holds. In what follows, we characterize the sample complexity associated with the preference theories considered in this paper (see Section 2.4).

The following theorem is due to Blumer et al. (1989), adapted to the current setting involving preferences.

THEOREM 2. *A theory of preferences $\mathcal{P}'$ is learnable if and only if it has finite VC dimension.*[9]

Last, we note that there is a strong connection between the VC dimension of a learnable theory and its associated sample complexity. In the PAC setting (see Ehrenfeucht et al. 1989 and, more recently, the work by Hanneke 2016), it turns out that the sample complexity of any learning rule, σ, such that (1) holds, is of the order of

$$\Theta\left(\frac{\mathrm{VC}(\mathcal{P}') + \ln(1/\delta)}{\epsilon}\right). \tag{2}$$

Hence, the sample complexity is linear in the VC dimension and independent of the particular learning rule σ. This implies that if we can estimate the VC dimension of a theory well, we will also be able to characterize its sample complexity from (2). Indeed, our main result corresponds to achieving this by providing bounds for the VC dimension.

Theorem 1 has the following implication for sample complexity.

COROLLARY 3. *Preferences $\mathcal{P}_{\mathcal{EU}}$, $\mathcal{P}_{\mathcal{CEU}}$, and, when $|\Omega| = 2$, $\mathcal{P}_{\mathcal{MEU}}$ are learnable. Preference $\mathcal{P}_{\mathcal{EU}}$ requires a minimum sample size that grows linearly with $|\Omega|$, while $\mathcal{P}_{\mathcal{CEU}}$ requires a minimum sample size that grows exponentially with $|\Omega|$. Finally, $\mathcal{P}_{\mathcal{MEU}}$ is not learnable when $|\Omega| \geq 3$.*

It is important to note the role played by the independence axiom in the above results. For example, our upper bound for the VC dimension of $\mathcal{P}_{\mathcal{CEU}}$ applies to all theories of preferences that satisfy co-monotonic independence (Axiom 5) and guarantee the existence of a certainty equivalent for every act. Hence, the bound applies more generally. Indeed, from Gilboa and Schmeidler (1994), we know that the Choquet integral can be represented as a linear functional defined on vectors in R^{2^Ω}. This representation allows us to tighten the VC dimension bound for $\mathcal{P}_{\mathcal{CEU}}$ to $2^{|\Omega|} + 1$ by applying part (i) of Theorem 1.[10]

4. DISCUSSION

4.1 *Choice functions*

Let X be a Borel space of *alternatives*, endowed with σ-algebra $\mathcal{X}$. A *menu* or *budget* of alternatives is any finite set $A \subseteq X$. A choice function on a *domain* $\mathcal{B} \subseteq 2^X$ is a map

[9]The theorem also requires an additional measurability hypothesis on the theory $\mathcal{P}'$. We discuss this issue in Section 5.1 and show that it is satisfied for the theories we focus on.

[10]We thank Burkhard Schipper for pointing this out.

$c : \mathcal{B} \to 2^X \setminus \{\varnothing\}$. The interpretation is that for a menu A, the set $c(A)$ consists of the alternatives chosen by the subject from A.

A choice function represents data that are available on the subject's choices. The choices themselves may be guided by an underlying preference relation held by the subject. Hence, through several instances of observed choice from menus, an outside analyst seeks to recover the underlying preference to the extent that he can make precise predictions about the subject's optimal choice. For each preference relation $\succsim$, we denote as $c_{\succsim}$ its associated choice function defined on the domain of menus of size at most $K \geq 2$, i.e., $\mathcal{B} = \{A \subseteq X \mid 2 \leq |A| \leq K\}$. In other words, we restrict attention to binary choice.

The choice function associated with a preference relation $\succsim$ is defined as

$$c_{\succsim}(A) = \{x \in A \mid x \succsim y \text{ for all } y \in A\}$$

for each $A \in \mathcal{B}$. For any theory of preferences $\mathcal{P}$, we can define the class of all choice functions that are generated by the preferences in the theory as

$$\mathcal{C}_{\mathcal{P}} = \{c_{\succsim} \mid \succsim \in \mathcal{P}\}.$$

We require a notion of how rich or flexible a family of choice functions is. We borrow the notion of P dimension from Kalai (2003): essentially an application of VC dimension to the graph of a function. Let X and Y be two sets, and let $F \subseteq Y^X$ be a collection of functions $f : X \to Y$. The P dimension of F is the largest number n with the property that there is $x_1, \ldots, x_n$ in X and $y_1, \ldots, y_n$ in Y such that for any $I \subseteq \{1, \ldots, n\}$ there is $f \in F$ with $y_i = f(x_i)$ if and only if $i \in I$. The P dimension of a class of functions F is denoted by $P(F)$.

Our next result connects the VC dimension of a theory with the P dimension of the class of choice functions associated with that theory. It implies that the negative conclusions we have obtained about theories with large VC dimension carry over to the choice functions defined by the theory, but that the theories that we have shown are learnable (and, therefore, have finite VC dimension) also have finite P dimension. Essentially, a theory $\mathcal{P}$ is learnable if and only if its associated choice $\mathcal{C}_{\mathcal{P}}$ functions are learnable.

Proposition 4. *We have* $\mathrm{VC}(\mathcal{P}) \leq P(\mathcal{C}_{\mathcal{P}})$. *Further, if* $\mathrm{VC}(\mathcal{P}) < +\infty$, *then* $P(\mathcal{C}_{\mathcal{P}}) < +\infty$.

In a followup paper, Basu (2019) shows that in the context of stochastic choice, finite VC dimension also suffices to learn preference heterogeneity in the population. If the preference theory has finite VC dimension, then we can recover the distribution over preferences that generates the choice data via a decision process involving stochastic preferences. This requires further complexity notions from statistical learning theory and empirical processes, namely Rademacher complexity.

We should emphasize that our discussion of choice function is restricted to choice from binary menus. Our conclusion might change if we allow for choices from arbitrarily large menus, because each choice $x \in c(A)$ would reveal a large number of binary

comparisons.[11] If, instead, the size of menus in $\mathcal{B}$ remains bounded as $|X|$ grows, then the message of our results remains the same.

4.2 *On the PAC model of learning*

We make here a few remarks about the PAC framework. In particular, we discuss two issues: (a) learning and (b) overfitting. When a preference theory is learnable, then it is the case that the learning rule that always finds a preference consistent with the data (when the model is correctly specified) leads to precise estimation of the preference in the PAC sense. One interpretation of the result is that the standard rationalizability exercise, which is to find a preference relation consistent with choice data, leads to consistent estimates. More so, this happens in a uniform sense as is the PAC criterion. This conclusion is quite nontrivial and the present framework aims to explain the key elements behind it. Sample complexity gives us rates of convergence for preference recovery and, hence, it becomes useful to study the VC dimension of different theories of preferences.

Now consider the problem of overfitting. Formally speaking, overfitting represents a situation where there are two preference relations $\succsim_1$ and $\succsim_2$, both of which rationalize the data set D, but one of which (say $\succsim_2$) has lower error in out-of-sample predictions, i.e., for the true preference $\succsim$, we have $d_\mu(\succsim_1, \succsim) > d_\mu(\succsim_2, \succsim)$. In this case, we say that $\succsim_1$ overfits the data. Hence, when the VC dimension is infinite, it means that for arbitrarily large sample sizes, the preference conjectured by the learning rule, $\sigma(D)$, has distance from the true preference $\succsim$, $d_\mu(\sigma(D), \succsim)$, that remains bounded away from zero. Hence, the VC dimension of a preference theory then becomes a useful tool to study overfitting in choice data.

5. Proofs

5.1 *Measurability requirement on $\mathcal{P}'$*

For the equivalence result of Theorem 2, an additional measurability requirement is needed on the model $\mathcal{P}'$. A class of sets $\mathcal{P}'$ is said to be *image admissible Souslin* if it can be parametrized by the unit interval, i.e., $\mathcal{P}' = \{P_t : t \in [0,1]\}$, in such a way that the set $Q = \{(z,t) : z \in P_t\}$ is an analytic set (see Dudley 2014, Pestov 2011). Whenever $\mathcal{P}'$ satisfies this condition, Theorem 2 holds. The following lemma provides a sufficient condition (satisfied by the models we consider in this paper) on $\mathcal{P}'$ for it to be image admissible Souslin.

Lemma 5. *Let $\mathcal{P}'$ be a model of preferences. Suppose there exists an uncountable complete separable metric space Θ, a bijection $m : \Theta \to \mathcal{P}'$, and a continuous function $V : \mathbb{R}^\Omega \times \Theta \to \mathbb{R}$ such that for each $\theta \in \Theta$,*

$$x \, m(\theta) \, y \quad \text{if and only if} \quad V(x,\theta) \geq V(y,\theta).$$

Then the model $\mathcal{P}'$ is image admissible Souslin.

[11] The difficulty in large menus is that if, for example x is chosen from $\{x, y, z\}$, then we infer a revealed preference relation among x and y, and x and z, but we learn nothing about how y and z are compared.

PROOF. Since Θ is an uncountable complete separable metric space, by the Borel isomorphism theorem (see Theorem 3.3.13 in Srivastava 2008), there exists a Borel measurable bijection $\sigma : [0,1] \to \Theta$. Now define the class $\{P_t\}_{t\in[0,1]}$ as

$$P_t = m\big(\sigma(t)\big).$$

Hence, we obtain

$$\begin{aligned} Q &= \big\{(z,t) : z \in P_t\big\} \\ &= \big\{(x,y,t) : V\big(x,\sigma(t)\big) \geq V\big(y,\sigma(t)\big)\big\}, \end{aligned}$$

where the latter set is Borel measurable since V is continuous and σ is Borel measurable. This implies that Q is a Borel set and, hence, is an analytic set. □

Now consider the three models of decision under uncertainty. Each satisfies the hypothesis of Lemma 5. The corresponding set Θ and functions m, V are as follows.

(a) *Expected Utility.* Here $\Theta = \Delta^{|\Omega|-1}$ and $m(\theta)$ is a unique preference relation on acts defined by the probability vector θ. The function V is defined as expected utility of the act x according to probabilities in θ:

$$V(x,\theta) = \theta \cdot x.$$

(b) *Choquet Expected Utility.* Here Θ is the set of all nonadditive measures on Ω, which is a complete and separable metric space when viewed as a subspace of $\mathbb{R}^{2^\Omega}$. Now $m(\theta)$ is the preference induced by the nonadditive measure θ. The function V is defined as

$$V(x,\theta) = \mathbb{E}_\theta(x).$$

Hence, $V(x,\theta)$ is the Choquet expectation of the x under θ.

(c) *Max–min Expected Utility.* For the max–min priors, the set Θ is the set of all nonempty compact convex subsets of $\Delta^{|\Omega|-1}$. Now Θ is complete and separable under the Hausdorff metric. For each $\theta \in \Theta$, $m(\theta)$ is the mutiple priors preference corresponding to the set of priors θ. Finally, the function V is defined as

$$V(x,\theta) = \arg\min_{p\in\theta} p \cdot x.$$

It is also possible to show that the models $\mathcal{P}_{\mathcal{LEU}}$ and $\mathcal{P}_{\mathcal{I}}$ satisfy the condition of being image admissible Souslin. A counterpart of Lemma 5 can be shown. We know that for any $\succsim \in \mathcal{P}_{\mathcal{I}}$, there exists $q = (q_k)_{k=1}^K$ such that $x \succsim y$ if and only if

$$\bigvee_{k=1}^{K} \left(\bigwedge_{l=1}^{k-1} q_l.x = q_l.y \right) \wedge (q_k.x \geq q_k.y).$$

The set of all (x,y,q) that satisfy the above condition is a Borel set and, hence, is analytic. Finally, we can identify the set of all qs with the unit interval $[0,1]$ as in the proof of Lemma 5.

5.2 *Technical lemmas*

The following lemmas are used in proving Theorem 1.

LEMMA 6. *Suppose that $\succsim$ satisfies Axioms 1 and 2. Then the following statements hold.*

(i) We have $x \succsim y$ if and only if $x - y \succsim \mathbf{0}$.

(ii) For each x, the upper and lower contour sets U_x and L_x defined as

$$U_x = \{y \in X : y \succsim x\} \quad \text{and} \quad L_x = \{y \in X : x \succsim y\}$$

are both convex. Moreover, the sets $X \backslash U_x$ and $X \backslash L_x$ are also convex.

PROOF. Consider part (i). Suppose $x \succsim y$. Then, by Axiom 2, it follows that

$$(1/2)(x - y) = (1/2)x + (1/2)(-y) \succsim (1/2)y + (1/2)(-y) = \mathbf{0}.$$

This means that $(1/2)(x - y) = (1/2)\mathbf{0} + (1/2)(x - y) \succsim (1/2)\mathbf{0} + (1/2)\mathbf{0} = \mathbf{0}$. Hence, by Axiom 2 again, $x - y \succsim \mathbf{0}$.

Now suppose $x - y \succsim \mathbf{0}$. Then, by Axiom 2, it follows that

$$(1/2)x = (1/2)(x - y) + (1/2)y \succsim (1/2)\mathbf{0} + (1/2)y = (1/2)y.$$

Again, applying Axiom 2, we get $x \succsim y$.

Now consider part (ii). Let $y, z \in U_x$ and $\lambda \in [0, 1]$. By Axiom 2, since $y \succsim x$, we obtain

$$\lambda y + (1 - \lambda)z \succsim \lambda x + (1 - \lambda)z.$$

Since $z \succsim x$, by Axiom 2, it also follows that

$$\lambda x + (1 - \lambda)z \succsim \lambda x + (1 - \lambda)x = x.$$

Hence, $\lambda y + (1 - \lambda)z \in U_x$.

The proofs for the convexity of L_x, $X \backslash U_x$ and $X \backslash L_x$ follow along similar lines. □

LEMMA 7. *Suppose $\succsim$ is a preference relation over acts satisfying Axioms 1–5. Then the following statements hold.*

(i) If $x \succsim y$ and $z \succsim w$ such that x, z are co-monotonic and y, w are also co-monotonic, then, for all $\lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda)z \succsim \lambda y + (1 - \lambda)w.$$

(ii) If $x \succ y$ and $z \succ w$ such that x, z are co-monotonic and y, w are also co-monotonic, then, for all $\lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda)z \succ \lambda y + (1 - \lambda)w.$$

PROOF. We prove only the first part; the second part follows analogously.

The continuity and monotonicity of $\succsim$ imply that for each x, there is a unique scalar c_x such that $x \sim c_x$, where c_x is viewed as a constant act. The proof relies on the observation that every constant act is co-monotonic with any act.

First note that $x \sim c_x$, and that x, c_x, and z are co-monotonic. Then $\lambda x + (1-\lambda)z \sim \lambda c_x + (1-\lambda)z$, by Axiom 5. Similarly, $z \sim c_z$ and we obtain that $\lambda c_x + (1-\lambda)z \sim \lambda c_x + (1-\lambda)c_z$. So $\lambda x + (1-\lambda)z \sim \lambda c_x + (1-\lambda)c_z$.

Now $c_z \succeq w$, and c_z, w, and y are co-monotonic. Thus, $(1-\lambda)c_z + \lambda y \succeq (1-\lambda)w + \lambda y$. Finally, $c_x \succeq y$, and c_x, c_z, and y are co-monotonic. Then $\lambda c_x + (1-\lambda)c_z \succeq \lambda y + (1-\lambda)c_z$.

Thus, we obtain that

$$\lambda x + (1-\lambda)z \sim \lambda c_x + (1-\lambda)c_z \succeq \lambda y + (1-\lambda)c_z \succeq \lambda y + (1-\lambda)w.$$

The proof follows from transitivity. □

LEMMA 8. *Let K be a closed convex cone*[12] *in $\mathbb{R}^\Omega$ such that $\mathbb{R}^\Omega_+ \subseteq K \subsetneq \mathbb{R}^\Omega$. Then there exists a preference $\succsim$ that belongs to the max–min model, such that*

$$U_{\mathbf{0}} = \left\{x \in \mathbb{R}^\Omega : x \succsim \mathbf{0}\right\} = K, \tag{3}$$

where $U_{\mathbf{0}}$ represents the upper contour set of the constant act of zeroes $\mathbf{0}$ for the preference $\succsim$.

PROOF. Consider $K^* = \{p \in \mathbb{R}^\Omega : p \cdot x \geq 0 \text{ for all } x \in K\}$, the dual cone of K. Since $\mathbb{R}^\Omega_+ \subseteq K$ and since the dual cone of $\mathbb{R}^\Omega_+$ is itself, it follows that $K^* \subseteq \mathbb{R}^\Omega_+$. Further, K^* is nonempty, which follows from our assumption that $K \subsetneq \mathbb{R}^\Omega$. Let $x \in \mathbb{R}^\Omega \backslash K$. Since K is closed and convex, there exists a hyperplane $p \neq 0$ such that $p \cdot x \leq p \cdot y$ for all $y \in K$. Note that it cannot be the case that $p \cdot y < 0$ for some $y \in K$. Otherwise, given that K is a cone, one could choose a large enough $\alpha > 0$ so that $\alpha y \in K$ and $p \cdot (\alpha y) < p \cdot x$. Hence, $p \cdot y \geq 0$ for all $y \in K$. This implies that $p \in K^*$.

Now define the following set of probability measures on Ω:

$$C := \Delta^{|\Omega|-1} \cap K^*.$$

We show that the max–min preference $\succsim$ induced by the set of priors C indeed satisfies condition (3).

The upper contour set at $\mathbf{0}$ for the preference $\succsim$ is

$$U_{\mathbf{0}} = \{x : p \cdot x \geq 0 \text{ for all } p \in C\}.$$

Now, by definition of C, $p \cdot x \geq 0$ for all $p \in C$ if and only if $p \cdot x \geq 0$ for all $p \in K^*$. The reason is that K^* is a cone. It follows that

$$U_{\mathbf{0}} = \left\{x : p \cdot x \geq 0 \text{ for all } p \in K^*\right\},$$

the dual cone of K^*. The set K is a closed and convex cone. So the dual cone of K^* is, in fact, K. Hence, $U_{\mathbf{0}} = K$. □

[12] A subset $K \subseteq R^\Omega$ is a *convex cone* if for any $x, y \in K$ and $\alpha_1, \alpha_2 \in \mathbb{R}_+$, it holds that $\alpha_1 x + \alpha_2 y \in K$.

LEMMA 9. *Let e^i denote the unit vector in $\mathbb{R}^3$ for coordinate $i \in \{1,2,3\}$. For every n, there exist n points $x^1, x^2, \ldots, x^n$ on the plane $L = \{x : x_1 + x_2 + x_3 = 1\}$ such that for any set $I \subseteq \{1,2,\ldots,n\}$, it holds that*

$$x^j \notin \operatorname{conv}\left(\left\{x^i\right\}_{i\in I} \cup \left\{e^1, e^2, e^3\right\}\right)$$

for all $j \notin I$.

PROOF. Let $f : \mathbb{R} \to \mathbb{R}$ be a strictly concave function such that $f(0) = 1$ and $f(1) = 0$, and define the real numbers $r^i = 1/(1+i)$ for each $i \in \{1,2,\ldots,n\}$. Now define a set of points $x^1, x^2, \ldots, x^n$ in $\mathbb{R}^3$ by

$$x^i = \left(r^i, f\left(r^i\right), 1 - r^i - f\left(r^i\right)\right), \quad i = 1,\ldots,n.$$

Clearly $\{x^i\}_i \subseteq L$. Note also that $(0, f(0), 0) = e^2$ and $(1, f(1), 0) = e^1$. Now let $I \subseteq \{1,2,\ldots,n\}$ and suppose, toward a contradiction, that there exists $j \notin I$ such that

$$x^j \in \operatorname{conv}\left(\left\{x^i\right\}_{i\in I} \cup \left\{e^1, e^2, e^3\right\}\right).$$

This means that there exist vectors $\{y^k\}_{i=k}^m \subseteq \{x^i : i \in I\} \cup \{e^1, e^2, e^3\}$,and positive weights $\{\alpha^k\}_k$ such that

$$x^j = \sum_{k=1}^m \alpha^k y^k \quad \text{and} \quad \sum_{i=1}^m \alpha^k = 1.$$

By the preceding equation and the definition of x^j,

$$\sum_{k=1}^m \alpha^k y_2^k = x_2^j = f\left(\sum_{i=k}^m \alpha^k y_1^k\right).$$

Note that if $y^k \neq e^3$, then $y_2^k = f(y_1^k)$, and that if $y^k = e^3$, then $y_2^k = 0 < 1 = f(y_1^k)$. Thus, either way, $\sum_{k=1}^m \alpha^k y_2^k \leq \sum_{k=1}^m \alpha^k f(y_1^k)$.

Finally, observe that $\alpha^i < 1$ for all i, as $x^j \notin \{x^i : i \in I\} \cup \{e^1, e^2, e^3\}$. Then $\alpha^i < 1$ and the strict concavity of f implies that

$$\begin{aligned} f\left(x_1^j\right) &= f\left(\sum_{k=1}^m \alpha^k y_1^k\right) \\ &> \sum_{k=1}^m \alpha^k f\left(y_1^k\right) \\ &\geq \sum_{k=1}^m \alpha^k y_2^k \\ &= x_2^j, \end{aligned}$$

which contradicts the fact that $x_2^j = f(x_1^j)$. □

5.3 *Proof of Theorem 1*

In this section, we provide the proof of Theorem 1. We make use of the technical lemmas established above. Lemma 6 pertains to part (i) and Lemma 7 pertains to part (ii). Lemmas 8 and 9 pertain to part (iv).

5.3.1 *Proof of part (i)* We will first show that $\mathrm{VC}(\mathcal{P}_{\mathcal{I}}) \leq |\Omega| + 1$. Let $n > |\Omega| + 1$ and let $(z_1, z_2, \ldots, z_n)$ be a set of points in X^2. Now, for each $z_i = (x_i, y_i)$, define the act

$$f_i := x_i - y_i.$$

Now consider the collection $\{f_i\}_{i=1}^n$ of acts.

Suppose it is the case that not all f_is are distinct. That is, there exist $j \neq k$ such that $f_j = f_k$. Now this means any data set $(z_i, a_i)_i$ where $a_j = 1$ and $a_k = 0$ cannot be rationalized by the model. This is because, from part (i) of Lemma 6, $a_j = 1$ requires $f_j \succsim \mathbf{0}$, but $a_k = 0$ requires $\mathbf{0} \succ f_k = f_j$.

Suppose now that all f_is are distinct. Since $n \geq |\Omega| + 2$, from Radon's theorem,[13] there exists a partition (I, J) of $\{1, \ldots, n\}$ such that $\mathrm{conv}(\{f_i\}_{i\in I}) \cap \mathrm{conv}(\{f_i\}_{i\in J}) \neq \varnothing$. Now, let $(a_i)_i$ be such that $a_i = 1$ for all $i \in I$ and $a_i = 0$ for all $i \in J$. We argue that the data set $(z_i, a_i)_i$ cannot be rationalized by the model. Now suppose, for contradiction, there is a preference relation $\succsim$ that satisfies Axioms 1 and 2, and rationalizes the data set. Now let $\bar{f} \in \mathrm{conv}(\{f_i\}_{i\in I}) \cap \mathrm{conv}(\{f_i\}_{i\in J})$. On the one hand, applying part (ii) of Lemma 6, we have $\bar{f} \succsim \mathbf{0}$, because $f_i \succsim \mathbf{0}$ for all $i \in I$. On the other hand, applying part (ii) again, we have $\mathbf{0} \succ \bar{f}$, because $\mathbf{0} \succ f_i$ for all $i \in J$. This gives us a contradiction.

A corollary of the above is that $\mathrm{VC}(\mathcal{P}_{\mathcal{EU}}) \leq |\Omega| + 1$. We can, however, argue that, in fact, $\mathrm{VC}(\mathcal{P}_{\mathcal{EU}}) = |\Omega|$. We first show the lower bound. If $n \leq |\Omega|$, then we can choose the vectors $x_i = e^i$, where e^i is the unit vector; $y^i = (1/n, 1/n, \ldots, 1/n)$ is the constant vector of $1/n$s. Then, no matter how we partition the set $\{1, \ldots, n\}$ into two sets I and J, all we need is $p \in \Delta(\Omega) = \Delta^{n-1}$ such that we have probability $p_i \geq 1/n$ if and only if $i \in I$. But this construction is always possible. Hence, $\mathrm{VC}(\mathcal{P}_{\mathcal{EU}}) \geq |\Omega|$. Next, so as to argue that $\mathrm{VC}(\mathcal{P}_{\mathcal{EU}}) \leq |\Omega|$, we make the simple observation that the VC dimension of $\mathcal{P}_{\mathcal{EU}}$ is the same as the VC dimension of the set of linear classifiers $\{\{f \in \mathbb{R}^\Omega \mid p.f \geq 0\}\}_{p\in\Delta(\Omega)}$. But, from a standard result, this is upper bounded by the dimension of the set of linear functions $\{p.x\}_{p\in\Delta(\Omega)}$, which is at most $|\Omega|$ (see, for example, Mohri et al. 2018). Hence, this shows that $\mathrm{VC}(\mathcal{P}_{\mathcal{EU}}) = |\Omega|$.

5.3.2 *Proof of part (ii)* We first show that the VC dimension is at most $(|\Omega|!)^2(2|\Omega| + 1)$.

We enumerate the set of states as $\Omega = \{\omega_1, \ldots, \omega_s\}$. We say that $\omega_i > \omega_j$ if $i > j$. For each permutation $\sigma : \Omega \to \Omega$, define the set X_σ to be the set of all acts that are nondecreasing with respect to the permutation σ (when the states are arranged according to σ). That is, $X_\sigma = \{x \in \mathbb{R}^\Omega : \sigma(\omega) < \sigma(\omega') \Rightarrow x(\omega) \leq x(\omega')\}$. Clearly, each X_σ contains all the constant vectors. Also, any two acts in X_σ are co-monotonic. Note that

$$X^2 = \bigcup_{\sigma,\sigma'} X_\sigma \times X_{\sigma'}. \tag{4}$$

[13] Radon's theorem states that any set of $|\Omega| + 2$ points in $\mathbb{R}^\Omega$ can be partitioned into disjoint subsets whose convex hulls have a nonempty intersection

Now, let $n > (|\Omega|!)^2(2|\Omega|+1)$. This of course implies $n \geq (|\Omega|!)^2(2|\Omega|+1)+1$. By the pigeonhole principle, if $\{z_1, \ldots, z_n\}$ are distinct points in X^2, then (4) implies that there exist permutations σ and σ' such that $|\{z_i\}_{i=1}^n \cap X_\sigma \times X_{\sigma'}| \geq 2|\Omega|+2$. By Radon's theorem, there is a partition (I, J) of the set $\{i : z_i \in X_\sigma \times X_{\sigma'}\}$, where I and J are nonempty, and such that the convex hulls of $(z_i)_{i\in I}$ and $(z_i)_{i\in J}$ intersect. Define a labelling $(a_i)_{i=1}^n \in \{0,1\}^n$ by $a_i = 1$ if and only if $i \in I$. Consider the data set $D = (z_i, a_i)_{i=1}^n$; we claim that D cannot be rationalized.

Suppose, toward a contradiction, that D is rationalized by a preference relation $\succsim$ that satisfies the axioms. Then $x_i \succsim y_i$ for all $i \in I$ and $y_i \succ x_i$ for all $i \in J$. Let $\bar{z} = (\bar{x}, \bar{y})$ be a point in the intersection of the convex hulls of $(z_i)_{i\in I}$ and $(z_i)_{i\in J}$, and let $(\lambda_i)_{i\in I}$ and $(\lambda'_j)_{i\in J}$ be probability vectors such that

$$\left(\sum_{i\in I}\lambda_i x_i, \sum_{i\in I}\lambda_i y_i\right) = (\bar{x}, \bar{y}) = \left(\sum_{i\in J}\lambda'_i x_i, \sum_{i\in J}\lambda'_i y_i\right).$$

On the one hand, from Lemma 7 part (i), we have $\bar{x} \succsim \bar{y}$, since $x_i \succsim y_i$ for all $i \in I$. On the other hand, applying Lemma 7 part (ii), we have $\bar{y} \succ \bar{x}$ since $y_i \succ x_i$ for all $i \in J$. Thus, we arrive at a contradiction.

We next show that the VC dimension of the Choquet expected utility model is at least $\binom{|\Omega|}{\lceil |\Omega|/2 \rceil}$.

Let $\mathcal{E}_{/2} \subseteq 2^\Omega$ be the set of all events with cardinality equal to $\lceil |\Omega|/2 \rceil$.

Let $n = \binom{|\Omega|}{\lceil |\Omega|/2 \rceil}$ and let E_i, $i = 1, \ldots, n$, enumerate the members of $\mathcal{E}_{/2}$. Now let

$$z_i = (\mathbf{1}_{E_i} - 1/2, \mathbf{0}),$$

where $\mathbf{1}_E$ denotes the indicator vector for the event E (i.e., $\mathbf{1}_E(\omega) = 1$ if and only if $\omega \in E$).

Let $\{a_i\}_{i=1}^n \in \{0,1\}^n$ be arbitrary and consider the data set $D = (z_i, a_i)_{i=1}^n$. We prove that the Choquet model can rationalize D.

Let $\mathcal{I} = \{i \in [n] : a_i = 1\}$. Let ν be a monotone nonadditive probability measure such that

$$\nu(E_i) \geq 1/2 \quad \text{for all } i \in \mathcal{I} \quad \text{and} \quad \nu(E_i) < 1/2 \quad \text{for all } i \notin \mathcal{I}.$$

Such a nonadditive measure can be constructed explicitly. For example, let $\nu(E) = 0$ for all E of cardinality strictly smaller than $\lceil |\Omega|/2 \rceil$, and let $\nu(E) = 1$ for all E of cardinality strictly greater than $\lceil |\Omega|/2 \rceil$. For the E that have cardinality $\lceil |\Omega|/2 \rceil$, and using our enumeration, we can set $\nu(E_i) = 1/2$ if $i \in \mathcal{I}$ and $\nu(E_i) = 1/3$ if $i \notin \mathcal{I}$.

Choquet expectations can now be calculated and turn out to be

$$\mathbb{E}_\nu[\mathbf{1}_{E_i} - 1/2] \geq 0 \quad \text{if } a_i = 1 \quad \text{and} \quad \mathbb{E}_\nu[\mathbf{1}_{E_i} - 1/2] < 0 \quad \text{if } a_i = 0.$$

Hence, the Choquet expected utility preference $\succsim$ that corresponds to the nonadditive measure ν rationalizes the data set D.

5.3.3 *Proofs of parts (iii) and (iv)* When there are only two states of nature i.e., $|\Omega| = 2$, a max–min preference can be represented as min of expected utilities corresponding to exactly two priors. There are two probabilities $(\underline{p}, \bar{p}) \in [0, 1]$ such that $\underline{p} \leq \bar{p}$ and the utility of a particular act (x_1, x_2) is

$$U(x_1, x_2) = \min\{\underline{p}x_1 + (1 - \underline{p})x_2, \bar{p}x_1 + (1 - \bar{p})x_2\}.$$

We show that for $|\Omega| = 2$, we have $\mathrm{VC}(\mathcal{P}_{\mathcal{MEU}}) = 2$. We first show that $\mathrm{VC}(\mathcal{P}_{\mathcal{MEU}}) < 3$ and then demonstrate a set of two data points that can be shattered.

Suppose, for contradiction, that there exist three data points $(x_i, y_i)_{i=1}^3$ that can be shattered by $\mathcal{P}_{\mathcal{MEU}}$. Now the point (x_i, y_i) is labelled $a_i = 1$ if and only if $U(x_i) \geq U(y_i)$. One can easily argue that this latter condition is satisfied for $(\underline{p}, \bar{p})$ if and only if $\alpha_i \underline{p} + \beta_i \bar{p} \geq \gamma_i$, where α_i, β_i, and γ_i depend solely on x_i and y_i. For example, if $x_{i1} \geq x_{i2}$ and $y_{i1} < y_{i2}$, then $U(x_i) = \underline{p}x_{i1} + (1 - \underline{p})x_{i2}$ and $U(y_i) = \bar{p}y_{i1} + (1 - \bar{p})y_{i2}$. Hence, we get that for $\alpha_i = x_{i1} - x_{i2}$, $\beta_i = y_{i2} - y_{i1}$, and $\gamma_i = y_{i2} - x_{i2}$, the condition $U(x_i) \geq U(x_i)$ is satisfied if and only if $\alpha_i \underline{p} + \beta_i \bar{p} \geq \gamma_i$. Finally, note that each such triple $(\alpha_i, \beta_i, \gamma_i)$ defines a hyperplane $\alpha_i \underline{p} + \beta_i \bar{p} = \gamma_i$. Consider the triangular region $T = \{(\underline{p}, \bar{p}) \in [0, 1]^2 \mid \underline{p} \leq \bar{p}\}$. For each i, the hyperplane defined by $(\alpha_i, \beta_i, \gamma_i)$ splits T into two regions: one in which $\alpha_i \underline{p} + \beta_i \bar{p} \geq \gamma_i$ and one in which $\alpha_i \underline{p} + \beta_i \bar{p} \leq \gamma_i$. If we label $a_i = 1$, we need to find $(\underline{p}, \bar{p})$ in the first region, and if we label $a_i = 0$, then we need to find $(\underline{p}, \bar{p})$ in the second region. Thus for each labelling of the data, there must exist $(\underline{p}, \bar{p}) \in T$ in the relevant region for each of the $(\alpha_i, \beta_i, \gamma_i)$. However, the three hyperplanes can generate at most seven regions of the triangle T.[14] This is a contradiction, since shattering the three points would have needed eight regions, i.e., the total number of labellings $\{0, 1\}^3$.

Now we show that a set of two points can be shattered. Consider the set of points $\{((-1, 2), (0, 0)), ((2, -1), (0, 0))\}$. This set admits all four labellings. For example, consider the labelling $(0, 1)$. This can be generated for $1/3 < \underline{p} < \bar{p} < 2/3$.

We next prove that for $|\Omega| \geq 3$, the max–min expected utility model is not learnable.

We prove the result for the case when $|\Omega| = 3$. If $|\Omega| > 3$, our construction can be embedded into a max–min preference in $\mathbb{R}^\Omega$ by simply ignoring all but three states when comparing acts. The axioms for max–min preferences will be satisfied by our construction. Hence, it is sufficient to prove the result for the case when $|\Omega| = 3$.

We prove that the VC dimension of the model is infinite. Let $n \in \mathbb{N}$ be any data size. Let $x^1, x^2, \ldots, x^n$ be the collection of points in $\mathbb{R}^3$ obtained from Lemma 9. Consider the data points

$$\{z_i\}_{i=1}^n = \{(x^i, \mathbf{0})\}_{i=1}^n.$$

Let $\{a_i\}_{i=1}^n \in \{0, 1\}^n$ be an arbitrary labeling of $\{z_i\}$, and consider the data set $D = \{(z_i, a_i) : i \in [n]\}$. We construct a max–min preference that rationalizes D.

Define $I = \{i \in [n] : a_i = 1\}$. Consider the set

$$K = \mathrm{cone}\left(\{x^i\}_{i \in I} \cup \{e^1, e^2, e^3\}\right) = \left\{\sum_{i \in I} \alpha^i x^i + \gamma^1 e^1 + \gamma^2 e^2 + \gamma^3 e^3 : \alpha^i \geq 0 \text{ and } \gamma^j \geq 0\right\},$$

[14] See, for example, Wetzel (1978) and the lazy caterer's sequence.

the cone generated by the vectors $\{x^i\}_{i\in I} \cup \{e^1, e^2, e^3\}$. Note that $\mathbb{R}^\Omega_+ \subseteq K$, as e^1, e^2, and e^3 are part of the generating vectors. By Lemma 8, there exists a max–min preference $\succsim$ such that

$$\{x \in \mathbb{R}^\Omega_+ : x \succsim 0\} = U_{\mathbf{0}} = K.$$

Observe that, by definition of K, $x^i \succsim 0$ for all $i \in I$. If we prove that $x^j \notin K$ for all $j \notin I$, then we are done. Suppose then, toward a contradiction, that $x^j \in K$ for some $j \notin I$. This implies that there exist vectors $y^1, y^2, \ldots, y^m$ in $\{x^i\}_{i\in I} \cup \{e^1, e^2, e^3\}$ and nonnegative weights $\eta^1, \eta^2, \ldots, \eta^m$ such that

$$x^j = \sum_{i=1}^m \eta^i y^i.$$

By definition of x^i and Lemma 9, each y^i satisfies that $y^i_1 + y^i_2 + y^i_3 = 1$. Further, we have

$$\sum_{k=1}^3 x^j_k = \sum_{k=1}^3 \sum_{i=1}^m \eta^i y^i_k = \sum_{i=1}^m \eta^i.$$

Now, since $x^j_1 + x^j_2 + x^j_3 = 1$, it follows that $\sum_{i=1}^m \eta^i = 1$. But this implies that

$$x^j \in \operatorname{conv}\big(\{x^i\}_{i\in I} \cup \{e^1, e^2, e^3\}\big),$$

contradicting Lemma 9.

5.4 *Proof of Proposition 4*

Let $n \leq \mathrm{VC}(\mathcal{P})$. Let $\{(x_i, y_i)\}_{i=1}^n$ be points such that for each $B \subseteq \{1, \ldots, n\}$, there exists a preference $\succsim \in \mathcal{P}$ such that $x_i \succsim y_i$ if and only if $i \in B$.

We now show that $n \leq P(\mathcal{C}_\mathcal{P})$. Consider the set of pairs

$$\big\{(\{x_i, y_i\}, \{y_i\})\big\}_{i=1}^n.$$

Suppose $B \subseteq \{1, \ldots, n\}$ and let $B' = B^c$. Now let $\succsim$ be such that $x_i \succsim y_i$ if and only if $i \in B'$. Consider the choice function $c_{\succsim}$ associated with $\succsim$. Note that $c_{\succsim}(\{x_i, y_i\}) \neq \{y_i\}$ for all $i \in B'$. Further, since $\succsim$ is complete, $y_i \succ x_i$ for all $i \in B$. Hence, $c_{\succsim}(\{x_i, y_i\}) = \{y_i\}$ for all $i \in B$.

The above result implies that with the PAC criterion, a model of preference $\mathcal{P}$ would be learnable if its corresponding class of choice functions is learnable. The next result establishes the converse.

Let $d = \mathrm{VC}(\mathcal{P})$ and suppose, for contradiction, that $P(\mathcal{C}_\mathcal{P}) = +\infty$. Let n be such that

$$\left(\frac{2enK(K-1)}{d}\right)^d < 2^n.$$

Now consider any collection of pairs

$$\{(A_i, c_i)\}_{i=1}^n,$$

where $A_i \in \mathcal{B}$ and $c_i \subseteq X$. We must have that $c_i \subseteq A_i$. Now consider the set of points

$$\mathcal{I} = \bigcup_{i=1}^{n} \bigcup_{x,y \in A_i; x \neq y}^{n} \{(x,y),(y,x)\}.$$

This is a set of at least $2n$ and at most $nK(K-1)$ distinct points. Define also the set $\Pi_{\mathcal{P}}(\mathcal{I}) = \{\succsim \cap \mathcal{I} \mid \succsim \in \mathcal{P}\}$. Now, since $P(\mathcal{C}_{\mathcal{P}}) = +\infty$, it follows that for any $B \subseteq \{1,\dots,n\}$, there exists $c \in \mathcal{C}_{\mathcal{P}}$ such that $c(A_i) = c_i$ if and only if $i \in B$. Hence, this means that $|\Pi_{\mathcal{P}}(\mathcal{I})| \geq 2^n$. Alternatively, since $d = \mathrm{VC}(\mathcal{P})$, from Sauer's lemma (see, for example, Chapter 3 in Kearns et al. 1994) we have that

$$\left|\Pi_{\mathcal{P}}(\mathcal{I})\right| \leq \left(\frac{2enK(K-1)}{d}\right)^d,$$

which yields a contradiction.[15]

References

Ahn, David, Syngjoo Choi, Douglas Gale, and Shachar Kariv (2014), "Estimating ambiguity aversion in a portfolio choice experiment." *Quantitative Economics*, 5, 195–223. [1283]

Balcan, Maria-Florina, Amit Daniely, Ruta Mehta, Ruth Urner, and Vijay V. Vazirani (2014), "Learning economic parameters from revealed preferences." In *Web and Internet Economics* (Tie-Yan Liu, Qi Qi, and Yinyu Ye, eds.), 338–353, Springer International Publishing, Switzerland. [1283]

Basu, Pathikrit (2019), "Learnability and stochastic choice." Available at SSRN 3338991. [1292]

Beigman, Eyal and Rakesh Vohra (2006), "Learning from revealed preference." In *Proceedings of the 7th ACM Conference on Electronic Commerce*, 36–42. [1283]

Blume, Lawrence, Adam Brandenburger, and Eddie Dekel (1991), "Lexicographic probabilities and choice under uncertainty." *Econometrica*, 59, 61–79. [1288]

Blumer, Anselm, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth (1989), "Learnability and the Vapnik–Chervonenkis dimension." *Journal of the ACM*, 36, 929–965. [1280, 1282, 1289, 1291]

Camerer, Colin and Martin Weber (1992), "Recent developments in modeling preferences: Uncertainty and ambiguity." *Journal of Risk and Uncertainty*, 5, 325–370. [1283]

Chamberlain, Gary (2000), "Econometric applications of maxmin expected utility." *Journal of Applied Econometrics*, 15, 625–644. [1283]

[15]The magnitude $|\Pi_{\mathcal{P}}(\mathcal{I})|$ is called the *growth function* of $\mathcal{P}$, and Sauer's lemma provides a (tight) bound on the growth function, which in turn provides the current bound by applying standard techniques. In fact, the bound we use is in Kearns et al. (1994).

Chambers, Christopher P. and Federico Echenique (2016), *Revealed Preference Theory*, volume 56 of Econometric Society Monographs. Cambridge University Press. [1285]

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer (2017), *Willingness to Pay and Willingness to Accept Are Probably Less Correlated Than You Think*. Working paper, NBER Working Paper No. 23954. [1283]

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer (2018), *Econographics*. Working paper, NBER Working Paper No. 24931. [1283]

Corfield, David, Bernhard Schölkopf, and Vladimir N. Vapnik (2005), "Popper, falsification and the VC-dimension." Working paper, Max Planck Institute for Biological Cybernetics. [1285]

Corfield, David, Bernhard Schölkopf, and Vladimir N. Vapnik (2009), "Falsificationism and statistical learning theory: Comparing the Popper and Vapnik–Chervonenkis dimensions." *Journal for General Philosophy of Science*, 40, 51–58. [1285]

Dudley, Richard M. (2014), *Uniform Central Limit Theorems*, volume 142. Cambridge University Press. [1293]

Ehrenfeucht, Andrzej, David Haussler, Michael Kearns, and Leslie Valiant (1989), "A general lower bound on the number of examples needed for learning." *Information and Computation*, 82, 247–261. [1291]

Ellsberg, Daniel (1961), "Risk, ambiguity, and the Savage axioms." *Quarterly Journal of Economics*, 75, 643–669. [1280]

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde (2018), "Global evidence on economic preferences." *Quarterly Journal of Economics*, 133, 1645–1692. [1283]

Gilboa, Itzhak (2009), *Theory of Decision Under Uncertainty*, volume 45 of Econometric Society Monographs. Cambridge University Press, New York, New York. [1286]

Gilboa, Itzhak and David Schmeidler (1989), "Maxmin expected utility with non-unique prior." *Journal of Mathematical Economics*, 18, 141–153. [1281]

Gilboa, Itzhak and David Schmeidler (1994), "Additive representations of non-additive measures and the Choquet integral." *Annals of Operations Research*, 52, 43–65. [1291]

Hanneke, Steve (2016), "The optimal sample complexity of PAC learning." *Journal of Machine Learning Research*, 17, 1319–1333. [1291]

Huber, Peter J. (1981), *Robust Statistics*. John Wiley and Sons, New York. [1281]

Kalai, Gil (2003), "Learnability and rationality of choice." *Journal of Economic Theory*, 113, 104–117. [1283, 1292]

Kalai, Gil, Ariel Rubinstein, and Ran Spiegler (2002), "Rationalizing choice functions by multiple rationales." *Econometrica*, 70, 2481–2488. [1285]

Kearns, Michael J., Umesh Virkumar Vazirani, and Umesh Vazirani (1994), *An Introduction to Computational Learning Theory*. MIT Press. [1302]

Kreps, David M. (1988), *Notes on the Theory of Choice*. Westview Press, Boulder. [1286]

Mangelsdorff, Lukas and Martin Weber (1994), "Testing Choquet expected utility." *Journal of Economic Behavior & Organization*, 25, 437–457. [1283]

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018), *Foundations of Machine Learning*. MIT Press. [1298]

Pestov, Vladimir (2011), "PAC learnability versus VC dimension: A footnote to a basic result of statistical learning." In *Neural Networks (IJCNN), the 2011 International Joint Conference on*, 1141–1145. [1293]

Rubinstein, Ariel (1996), "Why are certain properties of binary relations relatively more common in natural language?" *Econometrica*, 343–355. [1285]

Salant, Yuval (2007), "On the learnability of majority rule." *Journal of Economic Theory*, 135, 196–213. [1283]

Savage, Leonard (1972), *The Foundations of Statistics*, Second Revised edition. Dover Publications, New York, New York. [1280]

Schmeidler, David (1989), "Subjective probability and expected utility without additivity." *Econometrica*, 57, 571–587. [1281]

Shalev-Shwartz, Shai and Shai Ben-David (2014), *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, New York. [1290]

Srivastava, Sashi Mohan (2008), *A Course on Borel Sets*, volume 180. Springer Science & Business Media. [1294]

Valiant, Leslie G. (1984), "A theory of the learnable." *Communications of the ACM*, 27, 1134–1142. [1282, 1289]

Vapnik, Vladimir (1998), *Statistical Learning Theory*, 624. Wiley, New York. [1285]

Vapnik, Vladimir N. (2006), *Estimation of Dependences Based on Empirical Data*. Springer Science & Business Media. [1285]

Vapnik, Vladimir N. and Alexei Y. Chervonenkis (1971), "On the uniform convergence of relative frequencies of events to their probabilities." *Theory of Probability and Its Applications*, 16, 264–280. [1280]

Wald, Abraham (1950), *Statistical Decision Functions*. Wiley, Oxford, England. [1281]

Wetzel, John E. (1978), "On the division of the plane by lines." *American Mathematical Monthly*, 85, 647–656. [1300]

Zadimoghaddam, Morteza and Aaron Roth (2012), "Efficiently learning from revealed preference." In *Internet and Network Economics* (Paul W. Goldberg, ed.), 114–127, Springer Berlin Heidelberg, Berlin, Heidelberg. [1283]

Co-editor Ran Spiegler handled this manuscript.

Manuscript received 11 September, 2018; final version accepted 22 February, 2020; available online 27 February, 2020.