Learnability and Models of Decision Making under Uncertainty[*]

Pathikrit Basu[†]   Federico Echenique[‡]

September 11, 2018

Abstract

We study whether some of the most important models of decision-making under uncertainty are uniformly learnable. Imagine an analyst who seeks to learn, or estimate, an agent's preference using data on the agent's choices. A model is learnable if the analyst can construct a learning rule to learn the agent's preference, when preferences conforms to the model, with enough data, and uniformly over processes that generate choice problems. We consider the Expected Utility, Choquet Expected Utility and Max-min Expected Utility models: arguably the most important models of decision-making under uncertainty. We show that Expected Utility and Choquet Expected Utility are learnable. Morever, the sample complexity of the former is linear, and of the latter exponential, in the number of states. This means that accurate estimation of Choquet Expected Utility may require very large sample sizes, while Expected Utility requires modest sample sizes. The Max-min Expected Utility model is learnable when there are two states, but not when there are three states or more. Our results exhibit a close relation between learnability and the axioms that characterise the model.

1   Introduction

We investigate whether some of the most important theories of choice under uncertainty are uniformly learnable. We adopt the notion 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 grow. Our results are mixed: some models in the theory are learnable while others are not. Table 1 provides a summary.

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 observed 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 preciseley to capture the presence or absence of overfitting. A key concept in the theory is the *VC dimension* of a model (Valiant [1984], Blumer et al. [1989]). Very roughly speaking, the VC dimension of a model is the largest cardinality of a dataset that the model can always rationalize. A model with infinite VC dimension is not learnable, and therefore prone to overfitting. A model with finite VC dimension is learnable; moreover, the VC dimension of the model controls the sample size needed in order 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 gain a rich understanding of its empirical performance. More specifically, if a model has a large VC dimension, then a large dataset is needed for accurate out of sample prediction, which indicates that the model is prone to overfitting. If a model has a small VC dimension, then smaller datasets suffice for accurate out of sample predictions, and we can say that the model is not subject to overfitting. The contribution of our paper essentially boils down to computing or bounding the VC dimension of the most popular models of decision under uncertainty.

We now 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 a model

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 towards 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 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.

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 expected value. Given an uncertain prospect, the agent evaluates it in adversarial, pessimistic, fashion, according to 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 used extensively in economic applications where agents face uncertainty.

Choquet expected utility assumes that agents have non-additive beliefs over uncertain events. Instead of additive probability measures, as in the model of subjective expected utility, agents' beliefs are represented by a possibly non-additive *capacity*. In the 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 non-additivity allows an agent to place a premium on events that are less ambiguous than others. The model was first axiomatized by Schmeidler [1989].

---

[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.

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 whether the models of subjective expected utility, max-min, and Choquet expected utility are PAC learnable. In all three cases, we assume an agent who is risk-neutral. If we were to include the utility function as an additional parameter to be learned, then the VC dimension of all three models would increase, and they would be harder to learn. In an effort to isolate the role of the priors in the theory, we focus here on the risk neutral version of the three models.[2]

We should emphasize that the literature of choice under uncertainty has proposed generalization of subjective expected utility in an effort to account for behavior that lies outside of the theory. As we mentioned in the beginning of the introduction, the inherent danger in generalizing a theory is twofold. The first is that the theory may have few testable implications. The second is that the generalizations may be prone to overfitting. The existing literature is axiomatic, and therefore captures testable implications well. In fact one can argue that the main focus of the theory of choice under uncertainty has been to describe the testable implications of its models: the typical paper in decision theory contains an axiomatic chatacterization. But overfitting has received no attention: we believe that ours is the first paper that seeks to deal with the possibility that, in generalizing subjective expected utility, the theory may be prone to overfitting.

Our model of learning and data involves an agent who is choosing among pairs of uncertain prospects, and an analyst who is trying to learn her preferences. The agent could, for example, be a subject in a laboratory experiment. The subject in the lab could be presented with pairs of alternatives, each one carrying uncertain outcomes, and from each pair make a choice. A second example is that of an agent in the field, whose choices are observed and recorded. An analyst assumes that the agent behaves according to one of the theories of choice under uncertainty: subjective expected utility, max-min, or Choquet expected utility. The analyst seeks to recover the model from the agents' choices: that is, she seeks to learn the model behind the agent's choices, with high probability and in the limit as the number of choices made grows. Importantly, PAC learning requires the analyst to learn the model while being agnostic about the process that generates the alternatives that the agent has to choose from.

If the agent is an expected utility maximizer, then we show that it will be possible to infer the probability that she ascribes to the different states of the world after observing

---

[2]Alternatively, we could work in the Anscombe-Aumann framework, as is common in decision theory. With enough data on choice among constant acts, our results on subjective expected utility can be used to learn a utility function (the objective expected utility model is linear as well); presumably then one could turn to the problem of learning a risk-averse model, once payoffs are formulated in utils.

a number of choices that is *linear* in the number of states of the world. Here the agent is facing menus, or choice problems, consisting of pairs of alternatives. The analyst ignores the process by which the menus are conformed, and pairs are selected, but she postulates that the agent behaves according to subjective expected utility. She wants to learn the particular expected utility model that explains the agent's choices, despite ignoring how the agent was presented with the pairwise comparisons the was asked to make.

	Learnable	Sample Complexity
Expected utility	✓	Linear
Choquet expected utility	✓	Exponential
Max-min (states > 2)	X	$+\infty$
Max-min (2 states)	✓	-

Table 1: Caption: Summary of results

The max-min model 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 max-min model has infinite VC dimension, and is therefore not learnable. This means that, no matter how many choices are made by our subject, the analyst will be unable to learn the model generating choices with high probability. If there are only two possible states of the world, then the model is indeed learnable (this is reminiscent of the finding in the revealed preference literature, see Chambers et al. [2016], where the model with two states is significantly better behaved than the model with three or more). The max-min model not being learnable means that an analyst will be unable to estimate a particular instance of the model that will allow her to make good out-of-sample predictions, even if she has access to arbitrarily large datasets. Put differently, the family of possible max-min parameters (the sets of multiple probabilities) is large and flexible enough that, even with very large datasets, the model can wrap itself close to the data, while predicting badly out of 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 learn. So our results can be read as saying that such additional restrictions are needed. Another possibility is that the analyst has substantive information as to the process that generates the data. She could have knowledge, 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. We should highlight that our setting with the PAC criterion can be extended to handle this latter case, where the analysts assumes a particular class of distributions that could be governing the data.

The Choquet expected utility model fairs better than max-min: it has finite VC dimension and is therefore PAC learnable. Unfortunately, Choquet has VC dimension that grows exponentially in the number of states. Therefore it requires in the worst case that the agent makes a sample size that is exponential in the number of states, before the analyst can with high probability recover the non-additive probability that performs well out of sample. In other words, Choquet expected utility is not as flexible as to not be learnable, but it is still subject to overfitting because the analyst requires extremely large datasets in order to obtain good out-of-sample predictions.

We should emphasize that the model of PAC learning requires our analyst to learn the model with high probability *uniformly* over the laws that govern the pairs of choices that the analyst is presented with. We offer two defenses of this feature of the theory, as it applies to choice under uncertainty.

First, we view uniformity as a strong, but reasonable, requirement in many economic applications. When analyzing experimental data, it is hard for an analyst to know how the person carrying out the experiment selected the specific choices that were used in the experiment. Often, experimenters say little, or nothing, about how they selected the alternatives in the experiment. Even if they are explicit about how the alternatives were selected, it makes sense for the analyst to be skeptic, and to carry out the analysis while remaining agnostic about how the details of the experimental design were decided by the experimenter.

PAC learning also makes sense as a requirement when learning from field data. For field data, agents' choices are observed and recorded, but it is very hard to know how the agent encountered the choice problems in question. For example, consumption surveys record choices that were made, but not alternatives that were not chosen. It is impossible to know who created the menus of alternatives from which a consumer selects one choice, or how they proceeded to construct the menu. Hence, for data coming from the field, it seems reasonable to assume that the analyst wants to be agnostic about how the agents were presented with the different choice problems.

The second justification is that PAC learning provides a simple and tractable framework to discuss overfitting, an important phenomenon that has been largely overlooked in the choice and revealed preference literature. Learnability translates into the study of VC dimension, which can be calculated by common arguments in choice theory.

We are not the first to study PAC learning in economic models. Kalai [2003] considers a choice functions, 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 neither of these papers study questions of choice under uncertainty. Our primitive model of choice is a preference relation, in contrast with 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.

2 Model

2.1 Preliminaries

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

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 as $\mathcal{P}$, the set of all preference relations on Z. A *model* is a subset $\mathcal{P}' \subseteq \mathcal{P}$. For example, the set of weak orders (complete and transitive preferences) is a model. The set of preferences that have a linear, or a Cobb-Douglas, utility representation is another model.

2.2 Learning

An agent makes choices among alternatives in X. An analyst observe these choices and seeks to infer the preference that may be guiding them.

We imagine an agent making choices from finitely many ordered pairs (x_i, y_i), $i = 1, \ldots, n$. The agent's choices are recorded in a collection of *labels* $a_i \in \{0, 1\}$. If the agent chooses x_i from the set $\{x_i, y_i\}$ we set $a_i = 1$. If she does not choose x_i, then we set $a_i = 0$.[3]

Formally, a *dataset* is any finite sequence $D \in \bigcup_{n \geq 1}(Z \times \{0, 1\})^n$; so a dataset takes the form:

$$D = ((z_1, a_1), (z_2, a_2), \ldots, (z_n, a_n)),$$

where $a_i \in \{0, 1\}$. A dataset is interpreted by the analyst as follows: for each i, if $z_i = (x_i, y_i)$ then the agent 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.

---

[3]One might instead want to keep track of when the agent strictly chooses x_i, y_i, or is indifferent. We believe that the same formalism can be used to capture this case, and that the results will not substantially change. The message about the learnability of the models under consideration does not change.

The set of all datasets is denoted by $\mathcal{D}$. The set of all datasets 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 agent is presented with choice problems. All the analyst knows is that choice problems are selected in an i.i.d. fashion from a probability distribution μ on Z, but μ is unknown. We shall assume that μ has full support.

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

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

The analyst would like σ to be such that, for a dataset 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 dataset $D \in \mathcal{D}_n$, 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 dataset 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

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

In words, the probability, according to μ, of drawing a sample $(x_1,y_1),\ldots,(x_n,y_n)$ such that, when labeled 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 agents preferences are in some model $\mathcal{P}'$, then she would choose a learning rule whose range lies in $\mathcal{P}'$.

We say that a model $\mathcal{P}'$ is *learnable* if the analyst can design a learning rule such that, whenever the agent's preference belongs to $\mathcal{P}'$, large samples of the agent'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 model 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 $s(\epsilon, \delta) \in \mathbb{N}$ such that for all $n \geq s(\epsilon, \delta)$,*

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

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 $(Z \times \{0,1\})^n$; σ_n is the prediction made by the learning rule σ on a dataset of size n.

It is important to note the role of $s(\epsilon, \delta)$ in the definition above. It represents a lower bound on the number of samples needed for condition in 1 to hold under the learning rule σ. We shall be interested in the *sample complexity* of a learning rule, which is a function $s_\sigma : (0,1)^2 \to \mathbb{N}$, such that for all ϵ, δ, the $s_\sigma(\epsilon, \delta)$ is the *minimum* number of samples, n, such that 1 holds. In what follows, we will characterize the sample complexity associated with the preference models considered in this paper (see section 3).

It is well-known that a model is learnable if and only if it has finite *Vapnik-Chervonenkis* (VC) dimension (see Blumer et al. [1989]). The VC dimension of a model is defined as the largest data size n, such that there exists a collection of n choice instances $(z_1, z_2, ... z_n)$ such that *any* dataset of size n with this collection of instances can be rationalized by some preference from the model the analyst has in mind.

Formally, the VC dimension is defined as follows. We say that a sequence $(z_1, z_2, ... z_n)$ in Z is *shattered* by a model of preferences $\mathcal{P}'$ if, for any vector $(a_1, a_2, ... a_n) \in \{0,1\}^n$, there exists a preference $\succsim \in \mathcal{P}'$ such that for each $z_i = (x_i, y_i)$,

$$x_i \succsim y_i \text{ if and only if } a_i = 1,$$

i.e., the model $\mathcal{P}'$ *rationalizes* the dataset $D = (z_i, a_i)_{i=1}^n$. The *VC dimension* of a model $\mathcal{P}'$,

9

denoted as $VC(\mathcal{P}')$, is defined as

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

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

Theorem 1. *A model of preferences $\mathcal{P}'$ is learnable if and only if it has finite VC dimension.*[4]

The VC dimension of a model may be infinite. For example, suppose $X = \mathbb{R}$ and let $\mathcal{P}_R$[5] 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 n be a given data size and select the z_i's 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 dataset, no matter how the z_i's are labelled by the a_i's, we can always find a rational preference relation to rationalize the data.

Lastly, we will note that there is a strong connection between the VC dimension of a learnable model and its associated sample complexity. In the PAC setting (see Ehrenfeucht et al. [1989]), it turns out that the sample complexity of any learning rule, σ, such that 1 holds, is of the order of

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

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 model well, we will be able to characterise its sample complexity from 2. Indeed, our main result corresponds to achieving this by providing bounds for the VC dimension.

2.3 Decisions under uncertainty

We present a model of choice under uncertainty. Uncertainty is introduced through a *state-space* Ω, a finite set. An agent 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].

---

[4]The theorem also requires an additional measurability hypothesis on the model $\mathcal{P}'$. We discuss this issue in Section 4.1, and show that it is satisfied for the models we focus on.

[5]We impose that all preference relations in $\mathcal{P}_R$ be Borel measurable.

Throughout the paper, we restrict attention to preference relations that are *non-trivial*, 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 *comonotonic* 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.

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]

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

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

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

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

are both closed subsets of X.

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

$$x \succsim y.$$

5. (*Comonotic Independence*) For all $x, y, z \in X$ that are pairwise comonotonic and for all $\lambda \in (0, 1)$,

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

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

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

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 shall consider the following models of decision under uncertainty.

---

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

1. *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 satisfies the Order, Independence, Continuity and Monotonicity axioms.

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

$$\mathbb{E}_v(x) = \int_{-\infty}^{0} \left[v(\{\omega : x(\omega) \geq q\}) - v(\Omega) \right] dq + \int_0^\infty v(\{\omega : x(\omega) \geq q\} dq$$

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

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

   A preference belongs to the Choquet expected utility model if and only if it satisfies Order, Comonotonic Independence, Continuity and Monotonicity axioms.

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

$$\mathbb{E}_v(x) = \min_{p \in Core(v)} p \cdot x.$$

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

3. *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 the Order, C-Independence, Continuity, Monotonicity and Uncertainty Aversion axioms.

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

- $\mathcal{P}_{\mathcal{I}}$ denotes the set of preferences satisfying the Order and Independence axioms.

- $\mathcal{P}_{\mathcal{EU}}$ denotes the set of preferences satisfying the Order, Independence, Continuity and Monotonicity axioms.

- $\mathcal{P}_{\mathcal{CEU}}$ denotes the set of preferences satisfying Comonotonic Independence, Continuity and Monotonicity.

- $\mathcal{P}_{\mathcal{MEU}}$ denotes the set of preferences satisfying Order, Monotonicity, C-independence, Continuity and Uncertainty Aversion.

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 the Order and Independence axioms, 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 $q_1, \dots, q_K$ where $K \leq |\Omega|$ (see for example, Blume et al. [1991]) such that

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

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 $u_k > v_k$ where $k = \min\{i : u_i \neq v_i\}$. When $\succsim$ additionally satisfies monotonicity, then we have $q_1, \dots q_k \in \Delta^{|\Omega|-1}$ and the resulting model is called the *Lexicographic Expected Utility* model ($\mathcal{P}_{\mathcal{LEU}}$). Further, if continuity is also satisfied, then we have $K = 1$. Hence, $\mathcal{P}_{\mathcal{EU}} \subseteq \mathcal{P}_{\mathcal{LEU}} \subseteq \mathcal{P}_{\mathcal{I}}$. As a consequence, our result on the upper bound on the VC dimension of $\mathcal{P}_{\mathcal{I}}$ will have implications for these two models as well.

3   Main result

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

Theorem 2. *Let $\mathcal{P}_{\mathcal{EU}}$, $\mathcal{P}_{\mathcal{MEU}}$ and $\mathcal{P}_{\mathcal{CEU}}$ be as defined at the end of the last section.*

1. *$VC(\mathcal{P}_{\mathcal{I}}) \leq |\Omega| + 1$.*

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

3. *If $|\Omega| = 2$, then $VC(\mathcal{P}_{\mathcal{MEU}}) \leq 20$, and $\mathcal{P}_{\mathcal{MEU}}$ is learnable.*

4. *If $|\Omega| \geq 3$, then $VC(\mathcal{P}_{\mathcal{MEU}}) = +\infty$, and $\mathcal{P}_{\mathcal{MEU}}$ is not learnable*

The proof of Theorem 2 is presented in Section 4. The proof of part (1) is based on a standard result about the VC dimension of the set of all half-spaces in a Euclidean space.

The model $\mathcal{P}_{\mathcal{I}}$ has VC dimension at most $|\Omega| + 1$. This implies that the VC dimension for the Expected Utility model is at most $|\Omega| + 1$. The same upper bound applies also to the Lexicographic Expected Utility model. One can also argue that the VC dimension of $\mathcal{P}_{\mathcal{EU}}$ is at least $|\Omega| - 1$. Consider the unit vectors $\{e_i : i \in \{1, 2, ..., |\Omega| - 1\}\}\}$ in $\mathbb{R}^{\Omega}$ and data points

$$\{(-e_i, \mathbf{0})\}_{i=1}^{|\Omega|-1}.$$

This set of points can be shattered. For any labelling $(a_i)_{i=1}^{|\Omega|-1}$, let p^a denote the uniform probability measure on the set $I = \{i : a_i = 0\}$. We then have $a_i = 1$ if and only if $-p^a.e_i \geq 0$. Since we had $\mathcal{P}_{\mathcal{EU}} \subseteq \mathcal{P}_{\mathcal{LEU}} \subseteq \mathcal{P}_{\mathcal{I}}$, this implies that

$$|\Omega| - 1 \leq VC(\mathcal{P}_{\mathcal{EU}}) \leq VC(\mathcal{P}_{\mathcal{LEU}}) \leq VC(\mathcal{P}_{\mathcal{I}}) \leq |\Omega| + 1.$$

Finally, Theorem 2 has the following corollary about sample complexity.

Corollary 3. *$\mathcal{P}_{\mathcal{I}}$, $\mathcal{P}_{\mathcal{CEU}}$ and, when $|\Omega| = 2$, $\mathcal{P}_{\mathcal{MEU}}$ are learnable. $\mathcal{P}_{\mathcal{EU}}$ 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 $|\Omega| \geq 3$.*

Finally, 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 models of preferences which satisfy comonotonic independence 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 $R^{2^{\Omega}}$. This representation allows us to tighten the VC dimension bound for $\mathcal{P}_{\mathcal{CEU}}$ to $2^{|\Omega|} + 1$ by applying part 1 of Theorem 2.[7]

4 Proofs

4.1 Measurability requirement on $\mathcal{P}'$

For the equivalence result of Theorem 1, an additional measurability requirement is needed on the model $\mathcal{P}'$. A model $\mathcal{P}'$ is said to be *image admissible Souslin* if it can be

---

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

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 1 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 4. *Let $\mathcal{P}'$ be a model of preferences. Suppose there exists an uncountable complete seperable 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 $\theta \in \Theta$,*

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

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

Proof of Lemma 4. : 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 follows :

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

Hence, we obtain

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

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 an analytic set. ∎

Now, consider the three models of decision under uncertainty. Each satisfies the hypothesis of Lemma 4. The corresponding set Θ and functions m, V are as follows.

1. *Expected Utility* : $\Theta = \Delta^{|\Omega|-1}$ and $m(\theta)$ is 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,

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

2. *Choquet Expected Utility* : Here, Θ is the set of all non-additive 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 non-additive 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.

3. *Max-min Expected Utility* For the max-min priors, the set Θ is the set of all non-empty 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 4 can be shown. We know for any $\succsim \in \mathcal{P}_{\mathcal{I}}$, there exist $q = (q_k)_{k=1}^K$ 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 analytic. Finally, we can identify the set of all q's with the unit interval $[0,1]$ as in the proof of the Lemma 4.

4.2   Technical Lemmas

The following lemmas are used in proving Theorem 2.

Lemma 5. *Suppose that $\succsim$ satisfies the Order and Independence axioms, then the following hold.*

1. *$x \succsim y$ if and only if $x - y \succsim 0$.*

2. *For each x, the upper and lower contour sets U_x and L_x defined as*

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

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

Proof. Consider part 1. Suppose $x \gtrsim y$. Then, by Independence, it follows that

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

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

Now, suppose $x - y \gtrsim 0$. Then, by Independence, it follows that

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

Again, applying Independence, we get $x \gtrsim y$.

Now, consider part 2. Let $y, z \in U_x$ and $\lambda \in [0,1]$. By Independence, since $y \gtrsim x$, we obtain

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

Since $z \gtrsim x$, by Independence, it also follows that

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

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 6. *Suppose $\gtrsim$ is a preference relation over acts satisfying Order, Comonotonic Independence, Continuity and Monotonicity. Then the following hold.*

1. *If $x \gtrsim y$ and $z \gtrsim w$ such that x, z are comonotonic and y, w are also comonotonic. Then, for all $\lambda \in [0,1]$,*

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

2. *If $x > y$ and $z > w$ such that x, z are comonotonic and y, w are also comonotonic. Then, for all $\lambda \in [0,1]$,*

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

Proof. We only prove the first part. The second part follows analogously.

The continuity and monotonicity of $\succsim$ implies 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 comonotonic with any act.

First note that $x \sim c_x$, and that x, c_x, and z are comonotonic. Then $\lambda x + (1 - \lambda)z \sim \lambda c_x + (1 - \lambda)z$, by comonotonic independence. 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 comonotonic. 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 comonotonic. 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 7. *Let K be a closed convex cone[8] in $\mathbb{R}^\Omega$ such that $\mathbb{R}_+^\Omega \subseteq K \subsetneq \mathbb{R}^\Omega$. Then, there exists a preference $\succsim$ which belongs to the max-min model, such that*

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

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 non-empty, 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 shall show that the maxmin preference $\succsim$ induced by the set of priors C indeed satisfies condition (3).

---

[8]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 $\alpha_1 x + \alpha_2 y \in K$.

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

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

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, $U_\mathbf{0} = K$. ∎

Lemma 8. *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 $x^1, x^2, ..., x^n$ on the plane $L = \{x : x_1 + x_2 + x_3 = 1\}$ such that for any set $I \subseteq \{1, 2, ..., n\}$, it holds that*

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

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 $r^i = \frac{1}{i+1}$ for each $i \in \{1, 2, .., n\}$. Now, define a set of points $x^1, x^2, \ldots, x^n$ in $\mathbb{R}^3$ by

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

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

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

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

$$x^j = \sum_{k=1}^m \alpha^k y^k \text{ and } \sum_{i=1}^m \alpha^k = 1. \tag{4}$$

By the equation above 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

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

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

4.3 Proof of Theorem 2

In this section, we provide the proof of Theorem 2. We shall make use of the technical lemmas established in the appendix above. Lemma 5 pertains to part (1) and Lemma 6 pertains to part (2). Lemmas 7 and 8 pertain to part (4).

4.3.1 Proof of part 1

Let $n > |\Omega| + 1$ and let $(z_1, z_2, ..., z_n)$ 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_i's are distinct. That is, there exist $j \neq k$ such that $f_j = f_k$. Now, this means any dataset $(z_i, a_i)_i$ where $a_j = 1$ and $a_k = 0$ cannot be rationalised

by the model. This is because, from part 1 of Lemma 5, $a_j = 1$ requires $f_j \gtrsim \mathbf{0}$ but $a_k = 0$ requires $\mathbf{0} > f_k = f_j$.

Suppose now that all f_i's are distinct. Since $n \geq |\Omega| + 2$, from Radon's Theorem[9], there exists a partition (I, J) of $\{1, ..., n\}$ such that $\mathrm{conv}(\{f_i\}_{i \in I}) \cap \mathrm{conv}(\{f_i\}_{i \in J}) \neq \emptyset$. 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 the dataset $(z_i, a_i)_i$ cannot be rationalised by the model. Suppose not. Hence, there is a preference relation $\gtrsim$ that satisfies Order and Independence axioms and rationalises the dataset. 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 2 of Lemma 5, we have $\bar{f} \gtrsim \mathbf{0}$ because $f_i \gtrsim \mathbf{0}$ for all $i \in I$. On the other hand, applying part 2 again, we have $\mathbf{0} > \bar{f}$ because $\mathbf{0} > f_i$ for all $i \in J$. This gives us a contradiction.

4.3.2 Proof of part 2

We first show that the VC dimension at most $(|\Omega|!)^2(2|\Omega| + 1)$.

We enumerate the set of states as $\Omega = \{\omega_1, ..., \omega_s\}$. 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 non-decreasing 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 comonotonic. Note that

$$X^2 = \bigcup_{\sigma, \sigma'} X_\sigma \times X_{\sigma'}. \tag{5}$$

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 $\{z_1, ..., z_n\}$ are distinct points in X^2, then (5) 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 collection $(a_i)_{i=1}^n \in \{0, 1\}^n$ by $a_i = 1$ if and only if $i \in I$. Consider the dataset $D = (z_i, a_i)_{i=1}^n$; we claim that D cannot be rationalized.

Suppose, towards a contradiction, that D is rationalized by a preference relation $\gtrsim$ that satisfies the axioms. Then, $x_i \gtrsim y_i$ for all $i \in I$ and $y_i > x_i$ for all $i \in J$. Let $\bar{z} = (\bar{x}, \bar{y})$ be a

---

[9]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 non-empty intersection

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 6 part 1, we have $\bar{x} \gtrsim \bar{y}$, since $x_i \gtrsim y_i$ for all $i \in I$. On the other hand, applying Lemma 6 part 2, we have $\bar{y} > \bar{x}$ since $y_i > x_i$ for all $i \in J$. Thus, we have arrived at a contradiction.

We next show that the VC dimension of the Choquet Expected Utility model is atleast $\binom{|\Omega|}{\lceil |\Omega|/2 \rceil}$.

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, $i = 1, \dots, n$ 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 dataset $D = (z_i, a_i)_{i=1}^n$. We shall prove that the Choquet model can rationalize D.
   Let $\mathcal{I} = \{i \in [n] : a_i = 1\}$. Let ν be a monotone non-additive probability measure such that

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

Such a non-additive measure can be constructed explicitly. For example, let $\nu(E) = 0$ for all E of cardinality strictly smaller than $\lceil |\Omega|/2 \rceil$, and $\nu(E) = 1$ 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

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

Hence, the Choquet Expected Utility preference $\gtrsim$ that corresponds to the non-additive measure ν, rationalises the dataset D.

4.3.3 Proofs of parts 3 and 4

When there are only two states of nature i.e $|\Omega| = 2$, then the max-min model becomes a special case of the Choquet model. The reason is that any convex and compact set of priors in Δ^1 is can be identified with a subinterval $[\underline{p}, \bar{p}]$ of $[0,1]$. Here, a point $p \in [0,1]$ represents the probability of one of the two states, say $\omega_1 \in \Omega = \{\omega_1.\omega_2\}$. Now, define the non-additive measure ν as follows : $\nu(\{\omega_1\}) = \underline{p}$, $\nu(\{\omega_2\}) = 1 - \bar{p}$. One can verify that ν is convex and indeed $Core(\nu)$ corresponds to the interval of priors $[\underline{p}, \bar{p}]$.

Hence, from part 2, it follows as a corollary that when $|\Omega| = 2$, the max-min model is learnable. The upper bound derived in 2 applies here and hence, when $|\Omega| = 2$, the VC dimension of the max-min model equals at most 20.

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 shall 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 8. 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 dataset $D = \{(z_i, a_i) : i \in [n]\}$. We construct a max-min preference that rationalize D.

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

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

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 7, there exists a max-min preference $\succsim$ such that

$$\{x \in \mathbb{R}_+^{\Omega} : x \succsim 0\} = U_0 = K.$$

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

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

By definition of x^i and Lemma 8, each y^i satisfies that $y_1^i + y_2^i + y_3^i = 1$. Further, we have

$$\sum_{k=1}^{3} x_k^j = \sum_{k=1}^{3} \sum_{i=1}^{m} \eta^i y_k^i = \sum_{i=1}^{m} \eta^i.$$

Now, since $x_1^j + x_2^j + x_3^j = 1$, it follows that $\sum_{i=1}^{m} \alpha^i = 1$. But this implies that

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

contradicting Lemma 8.

References

Maria-Florina Balcan, Amit Daniely, Ruta Mehta, Ruth Urner, and Vijay V. Vazirani. Learning economic parameters from revealed preferences. In Tie-Yan Liu, Qi Qi, and Yinyu Ye, editors, *Web and Internet Economics*, pages 338–353, Cham, 2014. Springer International Publishing. ISBN 978-3-319-13129-0.

Eyal Beigman and Rakesh Vohra. Learning from revealed preference. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 36–42. ACM, 2006.

Lawrence Blume, Adam Brandenburger, and Eddie Dekel. Lexicographic probabilities and choice under uncertainty. *Econometrica: Journal of the Econometric Society*, pages 61–79, 1991.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Christopher P Chambers, Federico Echenique, and Kota Saito. Testing theories of financial decision making. *Proceedings of the National Academy of Sciences*, 113(15):4003–4008, 2016.

Richard M Dudley. *Uniform central limit theorems*, volume 142. Cambridge university press, 2014.

Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

Daniel Ellsberg. Risk, ambiguity, and the savage axioms. *The quarterly journal of economics*, pages 643–669, 1961.

Itzhak Gilboa. *Theory of decision under uncertainty*, volume 1. Cambridge university press Cambridge, 2009.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of mathematical economics*, 18(2):141–153, 1989.

Itzhak Gilboa and David Schmeidler. Additive representations of non-additive measures and the choquet integral. *Annals of Operations Research*, 52(1):43–65, 1994.

Peter J Huber. *Robust statistics*. John Wiley and sons, 1981.

Gil Kalai. Learnability and rationality of choice. *Journal of Economic theory*, 113(1):104–117, 2003.

David Kreps. *Notes on the Theory of Choice*. Westview press, 1988.

Vladimir Pestov. Pac learnability versus vc dimension: a footnote to a basic result of statistical learning. In *Neural Networks (IJCNN)*, *The 2011 International Joint Conference on*, pages 1141–1145. IEEE, 2011.

Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica: Journal of the Econometric Society*, pages 571–587, 1989.

Sashi Mohan Srivastava. *A course on Borel sets*, volume 180. Springer Science & Business Media, 2008.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Abraham Wald. *Statistical decision functions*. Wiley, 1950.

Morteza Zadimoghaddam and Aaron Roth. Efficiently learning from revealed preference. In Paul W. Goldberg, editor, *Internet and Network Economics*, pages 114–127, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-35311-6.