Prediction and Stochastic Choice

Abstract

In this paper, we study a non-parametric approach to prediction in stochastic choice models in economics. We show that VC complexity characterises the predictability of stochastic choice models. We establish prediction methods and provide corresponding rates of convergence.

Keywords: Statistical Learning; VC dimension; Stochastic Choice
2010 Mathematics Subject Classification: 62F10; 62G05

1 Introduction

The study of stochastic choice models in economics seeks to understand aggregate demand behaviour and preference heterogeneity (distribution over preferences) in markets (McFadden [1978],Berry et al. [1995]). A typical model of stochastic choice involves a set of alternatives from which an agent makes a choice and based on a set of underlying characteristics (x 2 X), the model prescribes the probabilities with which various items (a 2 A) would be chosen. This is defined in terms of a stochastic choice function _ : X ! _(A). The probabilities _(x) can be interpreted as the proportion of times a particular alternative is chosen compared to other feasible alternatives. Yet another interpretation involves the choice probabilities resulting from randomisation on part of the agent. The nature of this randomisation depends on the context and details of the decision making process (Manzini and Mariotti [2007],Fudenberg et al. [2015]). Online retail companies such as Amazon, are often interested in assessing consumer data $f(x_i; a_i)g_{i=1}^n$, to predict market shares and learn consumer preferences, by fitting the data through an economic model, in an effort to improve sales. In this paper, we study a novel machine learning approach to estimating choice probabilities, which correspond to a model of stochastic choice, based on ideas from statistical learning theory.

We study the problem of learning the map _ from finite data on choices. Hence, each data point consists of a pair $(x_i; a_i)$ where a_i is the alternative chosen when the characteristics were given by the vector x_i. We are interested in prediction methods defined on choice data which lead to accurate estimation of choice probabilities and the criterion we require is uniform consistency (Valiant [1984],Vapnik [1998],Kearns and Schapire [1994],Yamanishi [1992]). This requires that for any given precision parameters "; _ > 0, there exists a fixed data size N("; _), such that if the analyst were to apply the prediction method to a data set of size above N("; _), the probability would be at least 1 □ _ that the stochastic choice function conjectured by the prediction method would be close to the true stochastic choice function by at most ". Moreover, this holds true irrespective of the process that generates choice problems and the true function $_0$. Hence, working with amount of data at least N("; _) allows for robust estimates of the choice probabilities, when prediction methods are uniformly consistent. This is a desirable feature of the notion of consistency considered here. In many contexts, it is natural to assume that the analyst knows neither the distribution over characteristics nor the true stochastic choice function, and may wish to work with data sizes that gaurantee, in a robust manner, a certain degree of precision in estimation. The minimum number of samples N("; _), which provides such a gaurantee is called the sample complexity of the model of stochastic choice and is indeed a central object of study in the present paper.

The estimation strategy developed in this paper relies on the principle of empirical risk minimization (Vapnik [1998], Niyogi [2012], Niyogi [2006], Niyogi et al. [2011]). This involves setting up a risk function to evaluate the goodness of fit of the model. Ideally, these would be constructed in a way that the true choice probabilities would minimise expected risk. Then, since the data generating process is unknown, we approximate the expected risk with the empirical risk on the sample. For large enough samples, this would lead to good estimates for choice probabilities. We define risk in terms of incentive compatible scoring rules which are, in principle, reward schemes to incentivise truthful report of subjective probability judgements (Savage [1971],Gneiting and Raftery [2007]). Hence, when applied as a risk function, it ensures that true choice probabilities minimise expected risk. Furthermore, each scoring rule gives rise to a divergence function which allows us to quantify expected regret in terms of divergence in predictions between the true and estimated choice probabilities. The criterion for consistent learning, in turn, relies on this divergence converging uniformly to zero. Hence, scoring rules serve as a natural candidate for defining risk and a key result in the paper establishes that they admit uniform learning with respect to their associated divergence functions (Proposition 3.1). Finally, we obtain sample compexity bounds by applying complexity measures such as Rademacher complexity and Pollard dimension. The details of the approach can be found in Section 2.

There are several reasons as to why the above problem would be of relevance to economists. We discuss here a few points, highlighting how the present work contributes to the empirical and theoretical literature. The estimation of choice probabilities in discrete choice models has for a long

time been of interest to empiricists and econometricians interested in studying choice behaviour in markets (Manski [1975],McFadden [1978],Berry et al. [1995]). The novelty of the present approach lies in the introduction and assessment of sample complexity. This has two main advantages. Firstly, the sample size N(''; _) gaurantees robust estimates which are independent of the random process that generates the data. Hence, the analyst/econometrician can safely rely on such a sample size to achieve a predetermined level of accuracy in estimates ''; _ > 0. This is the key consequence of uniform consistency. Indeed, it also implies the usual consistency notion adopted in econometrics, where only convergence to the true parameter is required without a robustness gaurantee in terms of sample sizes. Hence, in the usual definition of consistency, depending on the data generating process, more or less number of samples may be needed for accurate estimation and a priori, the analyst may not know how much data would be needed, for which the present approach would provide gaurantees. Sample complexity yields uniform rates of convergence. Secondly, sample complexity also acts as a measure of complexity for stochastic choice models. Different models would have
See, for example, Shalev-Shwartz and Ben-David [2014].
Regret is defined simply as the difference between the expected risk of the estimated choice probabilities and the expected risk of the true choice probabilities.

different sample complexity and the models with higher sample complexity, would need more samples for good estimation. Hence, in a sense, we can compare different models based on such measures (say Logit v/s Probit) and this allows the analyst or econometrician to make a formal judgement as to which stochastic choice models are simple and which ones are complex.
We now discuss how the present work pertains to the theoretical literature on stochastic choice. Typically, in theoretical work, the stochastic choice function is treated as the primitive and is itself interpreted as the data (Luce [1959],Luce [1977]). However, real data sets only involve finitely many data points and typically, each data point represents a choice made by an individual. This suggests a gap between the assumption and the nature of real data sets and the logic that seems to justify the gap is that there is perhaps a ”law of large numbers” argument that one could rely on. However, from a formal standpoint, it is not clear whether this logic can always be implemented. For example, it may be that for some instances (such as observing a choice from a menu) there may be enough data points to evoke a law of large numbers argument but not for other instances. Hence, estimation of the entire map _ from finite data on choices seems far from obvious especially when keeping in mind the wide variety of decision procedures that the theoretical literature considers and which lead to very sophisticated stochastic choice models. Do all such models admit accurate estimation or predictability of choice probabilities?
This last question also brings us to a point of contrast between the theoretical and empirical literature. On the one hand, the theoretical models investigate varied contexts and decision procedures. If we assume these models as plausible in understanding choice behaviour, then perhaps a framework such as the present one could provide us a way to reason, in a simple manner, about the estimation and identification of all these models, allowing us to make good predictions about choice. However, on the other hand, empirical work still largely deals with more classical utility models and often for more complicated models, makes several distributional assumptions (on the data generating process) which are needed to obtain consistency results. In the present approach, we show that mild assumptions are needed for estimation. The main result of the paper is that in the context of stochastic preferences, if the preference class has finite VC dimension, then the corresponding stochastic choice model is predictable and allows us to recover the underlying distribution over preferences in the population, uniformly over the data generating process. Under some conditions, with a modified definition of VC dimension with strict preferences, we also obtain a converse result which says that a stochastic preference model is non-predictable if the dimension of the underlying preference class is infinite. This allows us to conclude that for any prediction method and any sample size, the worst case expected distance between true and estimated choice probabilities (in _(A)) is at least half times the maximum distance i.e 1=
p
2. In situations involving choice between exactly
two alternatives, this lower bound expected error is above the error level guaranteed by a rule which always makes the prediction that each alternative will be chosen with equal probability.
We provide several applications of our results for specific preference models commonly encountered in Industrial organisation, Decisions under Risk/Uncertainty and Spatial voting. We derive general bounds on rates of convergence based on Rademacher complexity. In particular, when the VC dimension is d (finite) and there are jAj many alternatives to choose from, then the Rademacher complexity associated with the stochastic choice model is at most O(
p
n□1jAj log(jAj)d log(d)),

which gives us an upper bound on the uniform rate of convergence of the estimator, which is defined via empirical risk minimisation with the quadratic scoring rule. We also show that for homothetic preferences specifically, convergence takes place in a stronger sense, the consistent estimates of the stochastic choice function converge in the sup norm, with a proportional rate of convergence. Lastly, we consider non-uniform learning and also address recoverability of cognitive heterogeneity in the setting of the bounded rationality model of choice with consideration sets.

Related Literature We discuss here, in some more detail, the relationship with prior work in the related literature. Typically, estimation of discrete choice models involves parametric or semiparametric assumptions (McFadden [1984],Manski [1975],Manski [1985]). The most widely used method for estimation is maximum likelihood. In our setting, maximum likelihood arises as a special case when the scoring rule applied is the log rule. We also show that a variant of Manski's maximum score estimation method (Manski [1975] corresponds to use of an incentive compatible scoring rule. Non-parametric approaches have also been studied in the economics literature (Matzkin [1992],Briesch et al. [2010]). The present paper has a non-parametric approach as well. However, the estimation/learning problem considered here is different in that we consider estimation of choice probabilities and our approach involves techniques and ideas from statistical learning theory at the core. In addition, as noted earlier, the consideration of sample complexity is one aspect that is novel in the present analysis. It is also interesting to compare the results of this paper with McFadden and Train [2000]. The paper establishes that choice probabilities of a random utility model can be approximated by choice probabilities corresponding to a mixed logit specification. In contrast, in the present setting, mixed logit being a subset of linear preferences, would be uniformly predictable (with linear complexity), while all continuous preferences would be non-predictable (with lower bounded predictive error). Hence, the dimensions of preferences (V C and V C_+ in this paper) play a critical role and its connection to predictability exhibits a perhaps interesting difference between the mixed logit model and stochastic choice from more general continuous preferences. Relatedly, De Blasi et al. [2010] consider a Bayesian non-parametric approach to the problem of recovering preference heterogeneity in the mixed logit framework via estimation of choice probabilities.

Scoring rules have been applied in prior work to study certain binary classification problems (Buja et al. [2005],Parry et al. [2016]) in other settings with different objectives. In contrast, the problem considered here is of consistent uniform learning for stochastic functions in the sense of PAC (probably approximately correct) learning (Valiant [1984]). This paper is hence, most closely related to prior work in probabilistic concept learning (Kearns and Schapire [1994],Abe et al. [1991]) and learning stochastic rules (Yamanishi [1992],Abe et al. [2001]). While the nature of the learning problem is similar, the present paper is different in that we apply scoring rules to study the learning problem. For instance, one interesting feature is that the related papers mention predictability with respect to specific divergences, which has a very natural connection to scoring rules. We show in this paper how the present approach exploits this connection and this is indeed a key aspect of our estimation strategy. Our approach is also more general and subsumes some of the learning problems studied in previous work. We should also emphasise that we are primarily interested here in stochastic choice models in economics and are motivated by problems in demand estimation.

The application of PAC learning in economics has been studied in different choice contexts in prior work (Kalai [2003],Beigman and Vohra [2006],Zadimoghaddam and Roth [2012],Basu and Echenique [2020]),Chambers et al. [2021], Chase and Prasad [2018]). However, the literature has only considered deterministic choice models. The distinguishing feature of the present work in relation to this strand of the literature is the consideration of stochastic choice models. We should also say that the present paper contributes to the growing literature on machine learning methods in economics (Athey and Imbens [2019]).

The outline of the paper is as follows. In section 2, we present the model and the learning problem. In section 3, we discuss consistent prediction methods, which we construct on the basis of the principle of empirical risk minimisation. Finally, in section 4, we apply our approach to a variety of stochastic choice models in economics.

2 Model

Let X _ R_K be a compact set of characteristics and A = fa_1; a_2; :::; a_mg denote a finite set indexing finitely many alternatives. In certain contexts, x = (x_I ; x_A) 2 R_K shall be a vector including both individual characteristics (x_I) and product characteristics (x_A = (x_a)$_{a2A}$). In other contexts, involving choice from menus, X _ 2_A = f0; 1g$_A$. Hence, at each x 2 X which is a menu, a choice of an alternative is made, a 2 x. We define Z = X _ A and will denote a typical element of Z as z.

A stochastic choice function is a map _ : X ! _(A). The interpretation of _ is that at each x,

_(x) = f_a(x)g_{a2A} is a probability vector where _a(x) denotes the probability that alternative a 2 A will be chosen when the underlying characteristics are given by x.

2.1 Data Generating Process

The analyst has access to a finite data set of choices from a population. Each data point consists of an individual's characteristic and the alternative chosen i.e (x_i; a_i) 2 Z. Hence, a data set is a finite sequence

z_n = f(x_1; a_1); (x_2; a_2); :::::; (x_n; a_n)g: (2.1)

Hence, a data set is a element z_n = (z_1; :::; z_n) of Z_n.

We now describe the data generating process. There is a probability measure _ 2 _(X) which defines the distribution of characteristics in the population. Given _ and a stochastic choice map _, the data is generated as follows. Independent across i, x_i is drawn according to _ and then a_i is drawn according to the choice probablities given by _(x_i). Note here that the analyst only observes the characteristic and the alternative chosen i.e. (x_i; a_i) through the data in 2.1. The analyst knows neither the distribution _ nor the stochastic choice function _, but assumes that it satisfies certain properties. We denote as _ _, the probability measure induced by _; _ together on the set X _ A. This represents the joint distribution from which the data is generated, by taking n i.i.d. samples from _ _. We shall denote as _ $_n$, the n-fold product measure induced by _ _ on Z_n. This is essentially the distribution of the data z_n.

2.2 Prediction

A model is any family of stochastic choice functions _. The objective of the analyst is to learn the true choice probabilities based on the data and a model _ represents the analyst's hypothesis. Formally, a prediction method is a map

^_ :

[

$_{n\_1}$

(X _ A)$_n$! _: (2.2)

Suppose now that the true choice probabilities are given by $_0$ and suppose $_0$ governs the distribution over characteristics. We shall require that a learning map ^_ be so that with enough data, z_n = f(x_i; a_i)g$_n$

$_{i=1}$, the estimate of the prediction method ^_(z_n) would be close to the true choice probabilities $_0$. Here, our notion of closeness between two choice probability maps _; $_0$ will be given by

d$_0$ (_; $_0$) =

Z

X

d(_(x); $_0$(x))d$_0$(x); (2.3)

Throughout the paper, for any metric space Y , we will denote as _(Y), the set of all Borel probability measures on Y . For any _ 2 _(Y), we shall denote as $_n$, the n-fold product measure on Y $_n$ defined by $_n$:= _| _ _ _{z:::: _ _}

n

.

where d : _(A) _ _(A) ! R denotes a divergence function or metric on the space of all choice probability vectors on A i.e _(A). For example, d could be squared Euclidean distance, KL divergence or the total variation distance. This leads us to the following definition of consistency for prediction methods.

Definition 2.1. A prediction method ^_ is consistent (with respect to d and _) if for all 0 < "; _ < 1, there exists N("; _) such that for all n _ N("; _),

(8$_0$ 2 _(X))(8$_0$ 2 _)

_

$_0$ $_n$

0

□

fz_n : d$_0$ (^_(z_n); $_0$) < "g

_

_ 1 □ _

_

: (2.4)

We say that a model _ is predictable with respect to d if there exists a prediction method ^_, which is consistent with respect to d and _. Finally, for a given "; _ > 0, we denote as N("; _), the smallest n for which 2.4 holds. The function N : (0; 1)$_2$! N is called the sample complexity of _ (with respect to ^_).

The above definition of predictability is based on PAC predictability (see, for example, Valiant

[1984],Yamanishi [1992]). In what follows, we shall discuss consistent prediction methods for various stochastic choice models.

3 Consistent prediction methods

3.1 Empirical Risk Minimization

We shall construct consistent prediction methods based on the principle of empirical risk minimization. For a detailed treatment, see Vapnik [1998]. Much of this section contains standard ideas from machine learning. However, one should acknowledge Proposition 3.1, which provides a novel way to solve the problem of PAC learning stochastic functions w.r.t divergences via the use of scoring rules (see Yamanishi [1992],Kearns and Schapire [1994],Abe et al. [1991] Abe et al. [2001]).

For the learning problem, we first define a loss function V : _ _ X _ A ! R. Suppose the true distribution and choice probabilities are given by _0; _0. Then, the expected risk corresponding to _ 2 _ is defined as

_ V (_) =

Z

X_A

V (_; x; a)d_0 _0(x; a): (3.1)

A minimizer of expected risk _ is defined as

_ 2 arg min

_2_

_ V (_): (3.2)

We shall consider loss functions V for which it will hold that _ = _0 i.e the true choice probabilities would minimise the expected risk (a property also known as Fisher consistency). In the next section, we will introduce risk in a specific manner via the application of incentive compatible scoring rules.

The principle of empirical risk minimization involves estimating the expected risk by the empirical risk on the sample zn = f(xi; ai)gn

i=1. For each _ 2 _, the empirical risk is given by

^ V (_) =

1

n

Xn

i=1

V (_; xi; ai) (3.3)

The prediction method that corresponds to empirical risk minimization, denoted by ^_E, is defined as

^_E(zn) 2 argmin

_2_

^ V (_) (3.4)

The minimum in 3.4 need not always exist. However, if it holds for some M that V (_; x; a) _ M for all _ 2 _ and (x; a) 2 Z, then the infimum exists and we can define an almost-ERM prediction method as follows. We have f"ngn such that "n > 0 and limn!1 "n = 0 . The almost-ERM rule selects ^_E(zn) such that

^ V (^_E(zn)) _ inf

_2_M

^ V (_) + "n (3.5)

In this context, consistency relies on ^_E(zn) being close to the true choice probability function _ as the function ^ V approximates _ V , for large enough n. Consistency here is defined as follows, in terms of V .

Definition 3.1. A prediction method ^_ is said to be consistent (with respect to V and _) if for all 0 < _; _ < 1, there exist an N(_; _) such that for all n _ N(_; _), it holds that

(8_0 2 _(X))(8_0 2 _)

_

_0 _n

0

□

fzn : _ V (^_(zn)) < inf

_2_

_ V (_) + "g

_

> 1 □ _

_

: (3.6)

We say that a model _ is predictable with respect to V if there exists a prediction method ^_, which is consistent with respect to V and _.

It turns out that a model _ is predictable if the family of real-valued functions V _ _ = fV (_; :; :) : _ 2 _g is a uniform Glivenko-Cantelli class of functions (see, for example, Vapnik [1998]; Shalev-Shwartz and Ben-David [2014]). Each f 2 V __ is a function of the form f : X _A ! R. We provide a definition below (see also Dudley [2014]).

Definition 3.2. A class of real valued functions F on Z is said to be a uniform Glivenko-Cantelli class of functions if for all "; _ > 0, there exists N("; _) such that for all n _ N("; _), for all _ 2 _(Z),

$$\underline{\_}^{n}$$

$$\underline{\_}$$

$$Z_n : \sup_{f2F} \left| \frac{1}{n} \sum_{i=1}^{n} f(z_i) - E_{\_}(f) \right| < "$$

$$\underline{\_}$$

$$\_ \; 1 - \_; \quad (3.7)$$

where $E_{\_}(f)$ denotes the expectation of f under _.

A necessary and sufficient condition for a class of real-valued functions to be a uniform Glivenko-Cantelli class of functions is that it have finite V-dimension for each > 0 (see, for example, Alon et al. [1997]). Other combinatorial measures and notions of dimensions and capacity also guarantee the uniform Glivenko-Cantelli property for a class of functions for. eg. VC-dimension, Pollard's pesudodimension, P-dimension, Covering numbers and Rademacher Complexity. We shall define and apply these notions and their study implications for sample complexity bounds in the subsequent sections.

3.1.1 Scoring Rules

In this section, we will define loss functions based on scoring rules. A scoring rule is a function $S : \_(A) \to R_A$ (see, for example, Savage [1971], Selten [1998]). The interpretation of S is that it is a mechanism for eliciting subjective probability judgements. If p is a probabalistic prediction about the alternative to be chosen in A and suppose a is the alternative chosen, then $S_a(p)$ is the reward obtained. For each p; q 2 _(A), we can define $S(p; q) := \sum_{a2A} S_a(p) q_a$ to be the expected score when q is the true distribution over A. A scoring rule is said to be incentive compatible if

$$S(q; q) \geq S(p; q) \text{ for all } p; q \in \_(A) \quad (3.8)$$

i.e p = q maximises the function S(:; q). We say that S is strongly incentive compatible if additionally, p = q is the unique maximizer for each q 2 _(A).

We use the terminology of Selten [1998]. Incentive compatible (strongly) scoring rules are also referred to as proper (strictly) scoring rules. See, for example, Gneiting and Raftery [2007]

We now define a loss function based on S.

$$V_S(\_; x; a) := -S_a(\_(x)) \quad (3.9)$$

We can now prove the following lemma.

Lemma 3.1. Let $\__0 \in \_(X)$ be the true distribution over characteristics and let $\__0 \in \_$ be the true choice probability function. Suppose S is incentive compatible. Then, $\__0$ minimises $\_ \; V_S(\_)$. Furthermore, if S is strongly incentive compatible and $\_\_ \in \_$ minimizes $\_ \; V_S(:)$, then $\_\_(x) = \__0(x)$ with probability one according to $\__0$.

Proof. For any $\__0 \in \_$, we have

$$\_ \; V_S(\__0) = \int_{X\_A} V_S(\__0; x; a) d\__0 \; \__0(x; a)$$

$$= -\int_X \left[\sum_{a2A} S_a(\__0(x)) \__{0;a}(x) \right] d\__0(x)$$

$$= -\int_X S(\__0(x); \__0(x)) d\__0(x)$$

_ □
Z
X
S(_0(x); _0(x))d_0(x) (3.10)
= _ V S(_0);
where the inequality 3.10 follows from the fact that S is an incentive compatible scoring rule. Now, suppose __ minimizes _ V S(:) and let E = fx : __(x) 6= _0(x)g . Since S is strongly incentive compatible, note that S(_0(x); _0(x)) > S(__(x); _0(x)) for all x 2 E. We already have that S(_0(x); _0(x)) _ S(__(x); _0(x)) for all x 2 X. Hence, if _0(E) > 0, then we will have that _ V S(_0) < _ V S(__). This contradicts that __ minimizes _ V S(:).
The above result shows that _0 is the minimiser of expected risk. Each incentive compatible scoring rule leads to a divergence function
dS(p; q) = S(q; q) □ S(p; q) for all p; q 2 _(A): (3.11)
Hence, from Lemma 1, it follows that
_ VS(_) □ inf
_2_M
_ VS(_) = _ VS(_) □ _ VS(_0)
=
Z
X
dS(_(x); _0(x))d_0(x):
Hence, if we can construct consistent prediction methods with respect to _ VS, then we can construct consistent prediction methods with respect to the divergence function dS.
The prediction method based on empirical risk minimisation corresponds to maximising the empirical score based on the data D = f(xi; ai)gn
i=1. Hence, ^_E maximises
□^ V S(_) =
1
n
Xn
i=1
Sai (_(xi)); (3.12)
which can be thought of as the empirical score. For a scoring rule S and stochastic choice function _, we define the function S _ _(x; a) := Sa(_(x)). Consistency of the prediction method ^_E with respect to dS relies on the nature of the real-valued function class
S _ _ = fS _ _j_ 2 _g:
The following is a key result.

Proposition 3.1. Let _ be a model of stochastic choice and let S be an incentive compatible scoring rule. Suppose the class of functions S__ is bounded above i.e. there exists anM such that f(z) _ M for all f 2 S__ and z 2 Z. If S__ is a uniform Glivenko-Cantelli class, then the model _ is predictable with respect to the divergence function dS.
Proof. The proof is in the appendix. It follows from the fact that the ERM rule yields consistency (see Vapnik [1998]) in the sense of Definition 3.1 and from Lemma 3.1.
We now give some examples of incentive compatible scoring rules and their associated divergence functions.
1. (Log Rule) : S(p; a) = ln(pa). The log scoring rule is incentive compatible and its divergence function corresponds to KL divergence S(p; q) = dKL(pjjq) =
P
a2A
ln(pa)
ln(qa) pa. Note that
maximisation of the empirical score with respect to the log rule corresponds to choosing choice probability functions according to the conditional maximum likelihood procedure .
2. (Brier/Quadratic Scoring Rule) : S(p; a) = 2pa □
P
b p2
b . The scoring rule is called the
Brier or quadratic scoring rule and is strongly incentive compatible. The divergence function associated with S is the square of the euclidean distance between p and q i.e. dS(p; q) =
jjp □ qjj2
2.
Since all Lp metrics on Rd are equivalent (as all norms on a finite dimensional vector space

are equivalent), it follows that consistency of a prediction method with respect to d_S implies consistency with respect to any other L_p metric. Further, from Pinsker's inequality, we have that

jjp □ qjj$_{TV}$ _

r

1

2

d_{KL}(pjjq):

Recall the total variational norm is equal to half times the L_1 distance i.e. jjp □ qjj$_{TV}$ =

1

2 jjp □ qjj$_1$. Hence, the above inequality implies that empirical score maximisation using the log scoring rule, which corresponds to conditional maximum likelihood, leads to consistency with respect to all L_p metrics.

3. (Manski's score with tie breaking) : Let _ be a complete strict order on A and let p 2 _(A). Now, define another strict order $_p$ on A as follows : a $_p$ b if either $p_a > p_b$ or it is the case that $p_a = p_b$ and a _ b. Finally, we let W(1) _ W(2) _ :::: _ W(jAj □ 1).

The Manski scoring rule with tie breaking is defined as S(p; a) = W(jfcja $_p$ cgj). Note that a $_p$ b implies both p_a _ p_b and S(p; a) _ S(p; b). Now, by the rearrangement inequality, this means that that S(q; q) _ S(p; q) for all p; q 2 _(A) i.e. S is an incentive compatible scoring rule.

4. (Subgradient of a convex function) : Suppose we have a convex function f : _(A) ! R. Let rf(x) : _(A) ! R_A be a subgradient for the function f. We can now define a scoring rule based on f as follows.

S_a(p) = f(p) + rf(p)($_a$ □ p).

S is incentive compatible and its associated divergence function corresponds to the Bregman divergence of f. Moreover, any incentive compatible scoring rule corresponds to some convex function. For example, f(p) = jjpjj$_2$ leads to the Brier score.

3.2 Dimension and Sample Complexity

In this section, we discuss sufficient conditions for a class of real valued functions to be uniform Glivenko-Cantelli. These place bounds on the capacity or complexity of the function class. There exist several alternative measures to quantify such complexity and we will discuss the central notions here. In particular, we discuss two notions we will use extensively : Rademacher complexity and Pollard Dimension.

Rademacher Complexity. Consider a real-valued function class F defined on the set Z. For the sample z_n, the empirical Rademacher complexity is defined as

R_{z_n}(F) = E_

h

sup

f2F

1

n

X_n

i=1

_$_i$f(z_i)

i

:

Here, each $_i$ is an independent bernoulli random variable which takes value 1 with 0:5 probability and □1 with 0:5 probability. Rademacher complexity of the function class F (with respect to a distribution _ 2 _(Z)) is defined as

R_n(F) = E$_n$[R_{z_n}(F)]:

Now, it turns out that the Rademacher complexity of a function class is closely related to its Glivenko-Cantelli properties. Suppose F is uniformly bounded i.e. there exists a real number M > 0 such that jf(z)j _ M for all z 2 Z and f 2 F. Then, we have that (see, for example, Shalev-Shwartz and Ben-David [2014]), for each _ 2 _(Z)

_^{n}

z_n : sup

f2F

j1=n

X_n

i=1

f(zi) □ E_(f)j < 2Rzn(F) + 6M
r
log 4=_
2n
!
_ 1 □ _;
Hence, if the empirical Rademacher complexity Rzn(F) converges to zero (say irrespective of the sample zn), then we obtain the uniform Glivenko-Cantelli property for the function class F. For example, if one shows that Rzn(F) _ C=
p
n, where C is an absolute constant, then we get that the
above inequality is satisfied for sample size n above
1
"2
h
2C + 6M
r
log 4=_
2
i2
:
This has implications for the sample complexity of stochastic choice models. Indeed, the proof of Proposition 3.1 shows that for a stochastic choice model, the sample complexity N(_; _), is at most maxf3=";N0("=3; _)g, where N0("=3; _) is the threshold sample size that guarantees inequality 3.7 for values ("=3; _) in the Glivenko-Cantelli definition. We next discuss Pollard dimension, which is related to Rademacher complexity.

Pollard Dimension. We say that a real-valued function class F shatters a set of points (z1; :::; zn) if there exists a vector (t1; :::; tn) 2 Rn such that for all B _ f1; :::; ng, there exists fB 2 F such that
fB(zi) > ti for all i 2 B
fB(zi) _ ti for all i =2 B
The Pollard dimension of F, denoted as P(F), is defined as
P(F) = maxfn j 9(z1; :::; zn) which can be shattered by Fg:
This complexity measure was introduced by Pollard [1984]. It is a generalisation of the VC dimension (see Vapnik and Chervonenkis [1968],Vapnik [1998]), which is defined for sets. Indeed, for a class

of 0-1 functions, the Pollard and VC dimensions coincide. Hence, for a collection of sets P, we have that V C(P) = P(P). The more standard (equivalent) definition would involve the notion of shattering for sets. A class of sets P shatters a set of points (z1; ::::; zn) if for each B _ f1; :::; ng, there exists P 2 P such that zi 2 P if and only if i 2 B. Then, the VC dimension of P is defined as
V C(P) = maxfn j 9(zi)n
i=1 which can be shattered by Pg:
For the function class F with Pollard dimension d, the empirical Rademacher complexity is
Rzn(F) _ C
r
d
n
;
where C > 0 is an absolute constant (see Pollard [1984]). Hence, an immediate implication of this is that finite Pollard dimension implies the uniform Glivenko-Cantelli property. In view of Proposition 3.1, the sample complexity of a stochastic choice model _, to be learned with respect to dS, is upper bounded by a linear function of the Pollard dimension of S _ _.

4 Stochastic Choice Models

In this section, we will consider a variety of stochastic choice models. In particular, we are interested in the recoverability of heterogeneity. By heterogeneity, we mean the distribution over preferences or parameters of the model, that leads to the stochasticity in choice. The main result of the paper is that when we consider stochastic preferences, we can recover preference heterogeneity if the support of the distribution has finite VC dimension. We discuss several familiar examples for which the result applies. We also establish a converse result on the non-predictability of such models. Lastly, we also consider choice with consideration sets and the recoverablity of cognitive heterogeneity.

4.1 Stochastic Preferences

Let X _ R^k be a set of alternatives. Suppose P is a class of continuous preference relations (complete and transitive) on X. We are interested in defining stochastic choice functions where preferences are random in P. Formally, we have a probability measure _ on P i.e. _ 2 _(P). In this setting, the characteristic vector shall be a vector of alternatives. Hence, X _ X^A, as in earlier. We shall assume that for each x = $(x_a)_{a2A}$ 2 X we have x_a 6= x_b for all a 6= b. Further,

$$\_\Big(\Big[_{a2A} f\%2 P : x_a \_ x_b \text{ for all } b 6= a\Big) = 1: \quad (4.1)$$

These assumptions ensure that in the random choice process, there are no ties with probability one. This is natural and simplifies the analysis. We now define the stochastic choice function corresponding to _ as

$$\__a(x; \_) = \_(f\%2 P : x_a \_ x_b \text{ for all } b 6= a);$$

which is the probability of choosing a given characteristics x and the distribution _ over preferences. The stochastic choice model corresponding to a random preference model on the class of preferences P is defined as follows.

$\__P$ = f_(:; _) j _ 2 _(P) satisfies (4.1) for all x 2 Xg:

We show the following result.

The set _(P) denotes the set of all Borel probability measures on P, which has the subspace topology generated by the closed convergence topology on the closed subsets of X _ X. See also the appendix.

Proposition 4.1. Suppose P is a class of continuous preferences with finite VC dimension d. Then, $\__P$ is predictable with respect to euclidean distance. Moreover, for the Quadratic scoring rule S_{br}, the Rademacher complexity of S_{br} _ $\__P$ is at most O(

$$\sqrt{n^{-1}jAj \log(jAj)d \log(d)}).$$

Proof. Firstly, we shall define a class of deterministic choice rules based on P. Then, we will show that $\__P$ is derived by taking continuous convex combinations of the deterministic class. We will then bound the Rademacher complexity corresponding to S_{br} _ $\__P$.

For each %2 P and a 2 A, define the 0-1 valued function

$$\__a(x;\%) = \begin{cases} 1 \text{ if } x_a \_ x_b \text{ for all } b 6= a \\ 0 \text{ otherwise} \end{cases}$$

By assumption (4.1), we get that for all _, and a 2 A,

$$\__a(x; \_) = \int_P \__a(x;\%)d\_(\%):$$

We will now show that for each a, the class of functions S^P_a = f_(:;%) j%2 Pg also has finite VC dimension.

Consider a finite set of n points D = $fx_1; :::; x_ng$ _ X. Suppose that D can be shattered by S^P_a.

Then, all 2^n possible 0-1 valued labellings of D can be generated by P. Now, consider the following set of points in X _ X.

$$D_0 = f(x^i_a; x^i_b)g_{i2f1;:::;ng;b2Ana}:$$

Note that D_0 is a set of n(jAj − 1) points. Let d be the VC dimension of P. On the one hand, from Sauer's Lemma (see Kearns and Vazirani [1994]), we have that at most $(\frac{en(jAj-1)}{d})^d$ many labellings in D_0 can be generated by P. On the other hand, we have that since S^P_a shatters D, at least 2^n labellings are obtained by P in D_0. For large n, we obtain a contradiction as

$$\_\frac{en(jAj - 1)}{d}\Big)^d < 2^n: \quad (4.2)$$

Hence, we establish that S^P_a has finite VC dimension, say d_a. Now, from a well known result (see

Pollard [1984]), it follows that there is an absolute constant C_a such that the Rademacher complexity of S^P_a, say $R_{x_n}(S^P_a)$, is at most $C_a\sqrt{d_a/n}$ for all x_n. Now, consider the probability of a in the stochastic choice model i.e. the function class $\Lambda^P_a = \{\lambda_a(\cdot;\beta)\,|\,\beta \text{ satistfies (4:1)}\}$. Since, Λ^P_a is essentially derived by taking convex combinations (including continuous combinations) from S^P_a (see Lemma 6.3 in the appendix), we obtain that the Rademacher complexity of Λ^P_a is at most equal to that of S^P_a.

Hence, we have $R_{x_n}(\Lambda^P_a) \leq C_a\sqrt{d_a/n}$.

We now complete the proof by applying the Quadratic scoring rule and bounding the Rademacher complexity of the class $F = S_{br}\setminus\Lambda^P$. Firstly, consider the function class $F_a = \{f(\cdot;a)\,|\,f \in S_{br}\setminus\Lambda^P\}$. Note that

$$f(x;a) = 2\lambda_a(x;\beta) - \sum_{b\in A}\lambda_b(x;\beta)^2$$

for some β. Since Rademacher complexity is additive (applied to $f - \lambda^P_a g_a$) and has nice lipschitz composition properties (in this case y^2), we get that $R_{z_n}(F_a) \leq 2C_a\sqrt{d_a/n} + \sum_b 2C_b\sqrt{d_b/n}$ (see Bartlett and Mendelson [2002] and also Shalev-Shwartz and Ben-David [2014]). Define $C_0 = \max_{a\in A} 2C_a\sqrt{d_a} + \sum_b 2C_b\sqrt{d_b}$. Now, for a given $z_n = (x_i;a_i)_{i=1}^n$, define $x_{n_a} = (x_i)_{i\in\{i\,|\,a_i=a\}}$ and $n_a = |\{i\,|\,a_i = a\}|$. Using Lemma 6.4 in the appendix, we get

$$R_{z_n}(F) \leq \sum_{a\in A}\frac{n_a}{n}R_{x_{n_a}}(F_a) \leq \sum_{a\in A}\frac{n_a}{n}\left(C_0\sqrt{\frac{n}{n_a}}\right) \leq \sum_{a\in A} r\sqrt{\frac{n_a}{n}}$$

n
(C_0=
p
n)
_ (C_0=
p
n)
X
a2A
r
n_a
n
_ (C_0=
p
n)
p
jAj: (4.3)
The last inequality follows from the fact that p
y is concave. Hence, we establish that $\_^P$ is predictable
with respect to any L_p metric.
Finally to obtain the sample complexity bound, consider the inequality 4.2. Based on the inequality, we can in fact say that the following holds.
d_a _ 4
d
log 2
log
_ 2d
log 2
_
+ 2d
log e(jAj □ 1)
log 2
: (4.4)
The above follows from a fact of arithmetic that x _ 4a log(2a) + 2b implies x _ a log(x) + b (see, for example, Shalev-Shwartz and Ben-David [2014]) whenever a > 1 and b > 0. Hence, the derived inequality in 4.4, together with the conclusion in 4.3 implies that the Rademacher complexity of F is at most O(
p
n□1jAj log(jAj)d log(d)).
The above result also allows us to recover the distribution over preferences. For this, we will define the following notion of distance between distributions.
d_(_; $\__0$) :=
Z
x
d(_(x; _)); _(x; $\__0$)))d_{0}(x):
Then, the following result holds. (Recovering heterogeneity) Suppose P is a class of continuous preferences with finite VC dimension. Then, there exists an estimator ^_ (based on a prediction method ^_ derived from almost-ERM with the Quadratic scoring rule) of the true distribution $\__0$ over preferences such that
d_(^_; $\__0$) $!_p$ 0;
uniformly over $\__0$ and $\__0$. Moreover, the rate of convergence is given by the sample complexity bound derived from Theorem 4.1.
In this setting, the class of preferences P would depend on the context. Stochastic preferences supported on linear utility would, for example, lead to the random coefficients model. Several canonical utility based models such as Cobb-Douglas or CES preferences may be considered. If preferences are over risky prospects, then we may have stochastic preferences over expected utility models and models exhibiting ambiguity aversion. We would bound the VC dimension of the preference class (see also Basu and Echenique [2020]) and then apply the above result to obtain predictability of the corresponding stochastic choice model supported on the preferences.

Applications We discuss some applications of the above result below.

1. (Linear Preferences) Consider the model of preferences P_{LIN} defined as follows. For each

%2 P_{LIN} there exists a vector v 2 R^k such that
x % y if and only if v:x _ v:y:
Such preferences satisfy the well-studied independence axiom from decision theory. This means that for all x; y; z 2 X and _ 2 [0; 1],
x % y if and only if _x + (1 □ _)z % _y + (1 □ _)z:
From a result in Basu and Echenique [2020]), it follows that the VC dimension of the set of preferences that satisfy the independence axiom is at most k + 1. Hence, it follows that V C(P_{LIN}) _ k + 1. From Theorem 4.1, it follows that the corresponding stochastic choice model, in which the vector of coefficients is random, would be predictable.
Consider the following widely used econometric specification. The utility from an alternative a is
u(x; _; _) = _:x_a + $_a$;
where _ 2 R^k and _ = ($_a$)$_{a2A}$ 2 R^A are vectors that are randomly drawn according to the probability measure _. For example, for the Logit model, _ is deterministically fixed and each $_a$ independently follows a logistic distribution (similarly for Probit). For Mixed Logit, we would have that _ would follow an unknown distribution for eg. multivariate normal. Here we have v = (_; _) and one may rewrite alternatives as x'_a = (x_a; e_a), where e_a is the unit vector in R^A
where the value corresponding to a is one. This also ensures that x'_a 6= x'_b for all a 6= b. Note
that our formulation here essentially pertains to the case where x is independent of _.
2. (Random Utility) Suppose now that preferences are defined by an underlying utility function. Hence, suppose we have a class of utility functions U _ fu j u : X ! Rg. For an given utility function u 2 U, the corresponding preference relation $\succsim_u$ is defined as
x $\succsim_u$ y if and only if u(x) _ u(y):
The model of preferences generated by U is then defined as P_U := f$\succsim_u$j u 2 Ug. It turns out that if the class of utility functions has finite Pollard dimension, say d, then the VC dimension of P_U is also finite, and is of the order O(d). It can be shown that utility classes corresponding to canonical models such as Cobb-Douglas and CES preferences have finite Pollard dimension.
3. (Decisions under Risk/Uncertainty) Suppose that the set of alternatives X corresponds to a set of monetary acts over a finite state space (uncertainty) i.e X _ R or the lotteries over a finite set of consequences Z i.e. X _ _(Z) (risk). Such contexts have been studied extensively in decision theory (see, for example, Kreps [1988] or Gilboa [2009]). Several studies have investigated different models of preferences based on different notions of expected utility. As was shown in Basu and Echenique [2020], the standard expected utility and choquet expected utility models had finite VC dimension, where the latter has VC dimension exponential in jj.
In the context of risk, one may show that expected utility and Quadratic expected utility (see Chew et al. [1991]) preferences have finite VC dimension. Indeed, the study of Gul and Pesendorfer [2006] consider random preferences with expected utility. The related model of random choice with private information due to Lu [2016] can be studied within our framework of linear preferences as well.
Finally, one should note that the nature of axioms of the model and VC dimension bounds are closely related. As was noted earlier, in Basu and Echenique [2020], we show the preferences satisfying the independence axiom have finite VC dimension, linear in jj. In that paper, it

was also argued that if preferences satisfy comonotic independence and admit existence of certainty equivalents, then the VC dimension would be finite and upper bounded by a function exponential in jj. Hence, for Choquet Expected Utility, denoted by P_{CEU}, the bounds
□ jj
jj=2
_
_ V C(P_{CEU}) _ (jj!)^2(2jj + 1) follow.
4. (Euclidean Preferences) Consider the context of spatial voting (see Downs et al. [1957]) where X _ R^k is an ideological space and each candidate in an election is represented by a point in X (alternative). In this voting setting, preferences are defined through an ideal point x_* such that candidates closer to the ideal point are preferred by the voter. Hence, we have that
x $\succsim_{x_*}$ y if and only if jjx □ x_*jj _ jjy □ x_*jj:
Hence P_{EUC} = f$\succsim_{x_*}$j x_* 2 Xg. In this case, the probability measure _ essentially gives us a distribution over ideal points, which is also referred to as voter heterogeneity. Since euclidean preferences can be defined by finitely many arithmetic computations (see Proposition 6.2 in the appendix), it follows that the VC dimension is finite and of the order O(k). Our estimation

results provide us a way to recover voter heterogeneity in these settings.

4.2 Non-predictability

We now establish a result on the non-predictability of $\__P$. With some additional conditions, we may prove that a result converse to Theorem 4.1 follows. This requires defining a notion of shattering based on strict preferences. We say that a model of preferences P strictly shatters a set of pairs of alternatives $((x_i; y_i))_{i=1}^n$ if for any labelling $(a_i)_{i=1}^n 2 f0; 1g^n$ of the pairs, there exists a preference relation %2 P such that

$x_i \_ y_i$ whenever $a_i = 1$ and ;

$y_i \_ x_i$ whenever $a_i = 0$

We now consider the following modified definition of VC dimension, which we will call $V C^+(P)$. It is defined as follows.

$V C^+(P) = \max\{n \mid 9((x_i; y_i))_{i=1}^n$ which can be strictly shattered by Pg:

We will assume here for simplicity that $X = R^k$ and that $X = fx 2 X^A j x_a 6= x_b$ for all a 6= bg . A preference relation % is said to be monotonic if $x >> y$ implies that $x \_ y$. We now state and prove the result.

Proposition 4.2. Let P be a class of monotone continuous preferences such that $V C^+(P) = +1$. Furthermore, suppose that there exists a probability measure satisfying 4.1 that has full support in P. Then, the following hold.

1. The stochastic choice model $\__P$ is not predictable with respect to any L^p metric.
2. Suppose we fix the euclidean distance on $\_(A)$. Then, for any prediction method $\hat{\_}$ and sample size n, the worst case expected distance (error) between true and estimated choice probabilities is at least $1=\sqrt{2}$, which is half of the maximum possible distance. For $jAj = 2$, this lower bound for error is above what can be achieved by predicting that each alternative will be chosen with probability 1=2.

Proof. The proof applies the probabilistic method where we randomly choose a stochastic choice function from $\__P$ and show that expected error is bounded away from zero. Then, we argue that there exists some function in $\__P$ that generates error above a threshold level.

This assumption is not crucial for the result.

Fix any prediction method $\hat{\_}$ and suppose that $0 < " < \frac{1}{\sqrt{2}} \square 1$ and $0 < \_ < 1 \square 1=\sqrt{2}$. Also, let n 2 N be a sample size. Now, further fix k 2 N such that $k \_ 2$. Since $V C^+(P) = +1$, there exists a strictly shattered set $((x_i; y_i))_{i=1}^k$. Finally, let t 2 (0; 1) such that $t^n \_ 0:5$. We will show then, that there exists a set of points $(x_i)_{i=1}^k \_ X$ such that for all $B \_ f1; 2; :::; kg$, there exists $\_^B 2 \__P$ such that

$\_^B_a(x_i) \_ t$ for all i 2 B and ;

$\_^B_a(x_i) \_ 1 \square t$ for all i =2 B; (4.5)

where a 2 A is fixed beforehand. We now show how to construct the set of points $(x_i)_{i=1}^k \_ X$. Let b 2 Anfag. We first define $x^a_i := x_i$ and $x^b_i := y_i$. Further, for each i we let $fx^c_i gc2Anfa;bg \_ R^k$ be such that all the vectors in $fx^c_i gc2A$ are distinct and also that $x^a_i >> x^c_i$ and $x^b_i >> x^c_i$ for all

c 2 Anfa; bg. Due to monotonicity, this ensures that from the set of alternatives in x_i either x_i^a or x_i^b is maximal.
Since P strictly shatters $((x_i^a; x_i^b))_{i=1}^n$, for each B _ f1; 2; :::; kng, there exists a preference relation % such that
$x_i^a \_ x_i^b$ whenever i 2 B and ;
$x_i^b \_ x_i^a$ whenever i =2 B:
Now, consider the set of preferences Q = f%j $x_i^a \_ x_i^b$ for all i 2 B and $x_i^b \_ x_i^a$ for all i =2 Bg. It is open in P under the closed convergence topology. Hence, it is measurable in the corresponding Borel sigma algebra. Now, consider a probability measure _ 2 _(P) that satisfies the regularity condition in 4.1 and has full support in P. Then, since Q is open, we must have that _(Q) > 0.
We will now construct another probability measure $\__B$ satisfying 4.1 such that its corresponding stochastic choice function $\__B$ satisfies 4.5. Firstly, let _ > 0 be such that $\_\_(Q^c) \_ 1 - t$. Then, let $\_ := \frac{1-\_\_(Q^c)}{\_(Q)}$. This implies that $\_\_(Q) + \_\_(Q^c) = 1$ and $\_\_(Q) \_ t$. Now, define the probability measure $\__B(E) := \_\_(Q \setminus E) + \_\_(Q^c \setminus E)$. Note that $\__B(E) = 0$ if and only if _(E) = 0, which implies that $\__B$ also satisfies 4.1. By definition, it now follows that the corresponding stochastic choice function $\__B$ satisfies 4.5.
We now proceed with the proof of non-predictability. Since all L_p metrics are equivalent on _(A), it suffices to show the claim for euclidean distance. Let _ be the uniform distribution over the set of points $\{x_i\}_{i=1}^n$. Also, let _ be the uniform distribution over all stochastic choice functions in $\{\__B\}_{B \subseteq \{1;2;:::;kn\}}$. We will lower bound the expected error after n samples.

$$E\_E\_{\_n}E\_\left[jj\hat{\_}(z_n)(x) - \_(x)jj \right]$$

:
Now, from Fubini's theorem, we will change the order of expectations. The above expectation then becomes

$$E\_{\_n}E\_E\_\left[k\hat{\_}(z_n)(x) - \_(x)jj \right]$$

$$= E\_{\_n+1}\left[E\_ jj\hat{\_}(z_n)(x) - \_(x)jj \; j \; x \; 2 \; x_n \right] P\_{\_n+1}(x \; 2 \; x_n) + E\_{\_n+1}\left[E\_ jj\hat{\_}(z_n)(x) - \_(x)jj \; j \; x =2 \; x_n \right] P\_{\_n+1}(x =2 \; x_n):$$

Since _ is uniform, it follows that $P\_{\_n+1}(x =2 \; x_n) \_ \frac{k-1}{k}$. We will now simply lower bound the conditional expectation

$$E\_{\_n+1}\left[E\_ jj\hat{\_}(z_n)(x) - \_(x)jj \; j \; x =2 \; x_n \right]$$

. Consider some fixed x =2 x_n. Now,
let us write the expectation E_
h
jj^_(z_n)(x) □ _(x)jj
i
more explicitly as
E__E$_{c_n_(x_n)}$
h
jj^_(x_n; c_n)(x) □ _(x)jj
i
:

where _(x_n) is the joint distribution induced on fa; bg_n by the product of the choice probabilities in
(_(y))$_{y2x_n}$. Consider some fixed _$_B$ for some B _ f1; 2; :::2ng. Now, let d(B; x_n) 2 fa; bg_n be such
that for each j 2 f1; :::::; ng we define d_j(B; x_n) = a if and only if x_j 2 (x_1; :::::; x_n) = x_n is such that
x_j = x_i for some i 2 B. Now, recall that for i 2 B, we have _$_a$(x_i) _ t and for i 2 f1; 2; :::2ngnB,
we have _$_B$
$_a$ (x_i) _ 1 □ t. Hence, by definition, it follows that _$_B$(x_n) draws the vector d(B; x_n) with
probability at least t_n. This gives us
E$_{c_n_B(x_n)}$
h
jj^_(x_n; c_n)(x) □ _$_B$(x)jj
i
_ t_n
h
jj^_(x_n; d_n(B; x_n))(x) □ _$_B$(x)jj
i
;
which implies
E__E$_{c_n_(x_n)}$
h
jj^_(x_n; c_n)(x) □ _(x)jj
i
_ t_nE$_{_B_}$
h
jj^_(x_n; d_n(B; x_n))(x) □ _$_B$(x)jj
i
:
Now, note in the latter expectation, since _ is uniform and x =2 x_n, it follows that the random variable
^_(x_n; d_n(B; x_n))(x) is independent of the event "_$_B$
$_a$ (x) _ t", with the latter event having probability
0:5. This is because d_n(B; x_n) only depends on the set of indexes B(x_n) = fi j x_i 2 x_ng. Hence,
conditional on ^_(x_n; d_n(B; x_n))(x), the expectation
E$_{_B_}$
h
jj^_(x_n; d_n(B; x_n))(x) □ _$_B$(x)jj j ^_(x_n; d_n(B; x_n))(x)
i
is such that with probability 0:5 we have jj^_(x_n; d_n(B; x_n))(x) □ _$_B$(x)jj with _$_B$
$_a$ (x) _ t and with
probability 0:5 we have jj^_(x_n; d_n(B; x_n))(x) □ _$_B$(x)jj such that _$_B$
$_a$ (x) _ 1 □ t . This implies that
value of the above conditional expectation is at least 0:5jj(t; 1 □ t) □ (1 □ t; t)jj = 0:5
p
2j1 □ 2tj .
Combining these, we get the lower bound
E$_{_n+1}$
h
E_ jj^_(z_n)(x) □ _(x)jj j x =2 x_n
i
_ 0:5t_n
p

2j1 □ 2tj:
Finally, since, $P_{\_n+1}(x = 2\ x_n) \_ k \square 1$
k , we get that
$E\_E\_{\_n}E\_$
h
$jj\hat{}\_(z_n)(x) \square \_(x)jj$
i

_
k □ 1
k
$0{:}5t_n$
p
2j1 □ 2tj:
This means that there exists $\_ 2 f\_{B}g_{B\_f1;2;:::;2ng}$ such that
$E\_{\_n}E\_$
h
$jj\hat{}\_(z_n)(x) \square \_(x)jj$
i

_
k □ 1
k
$0{:}5t_n$
p
2j1 □ 2tj: (4.6)
Taking k ! +1 and t ! 1, we get that the RHS converges to 0:5
p
2 = 1=
p
2. Since $p1$
2
$jj\hat{}\_(z_n)(x) \square$
_(x)jj 2 [0; 1], by an application of Markov's inequality, it follows that for any 0 < " <
p
2(1□1=
p
p 2) =
2 □ 1 and 0 < _ < 1 □ 1=
p
2, we can choose k ! +1 and t ! 1 so that
$P\_{\_n}$

E_
h
$jj\hat{}\_(z_n)(x) \square \_(x)jj$
i
> "
!
_ _:
Hence, the result obtains. From the lower bound derived in the RHS of 4.6, we can say that
inf
^_
sup
$\_{0}2\_(X);\_{0}2\_P$
$E_{\_0\_n}$
$_0 E_{\_0}$
h
$jj\hat{}\_(z_n)(x) \square \_{0}(x)jj$
i
_ 1=
p
2;
where the infimum is taken over all possible prediction methods ^_. Finally, note that when jAj = 2, a rule which always predicts probabilities $\_{a}(x) = 1=2$ on each alternative a, achieves expected error

at most 1=
p
2.
From Markov's inequality, one can show that for a random variable X 2 [0; 1] and " 2 (0; 1), we have that if E[X] _ 1 □ "2, then P[X > 1 □ "] _ 1 □ ".

One can argue that the class of all monotone, continuous and convex preferences P has V C+(P) = +1. In the context of uncertainty from a result in Basu and Echenique [2018], it follows that Max-min preferences (MEU) have infinite VC dimension whenever jj _ 3. However, the arguments can be modified to show that V C+(PMEU) = +1 as well. Of course, non-monotonic preferences can be non-predictable as well. Consider again the following econometric specification : an alternative has utility u(xa; _) + _a, where each _a independently follows a normal distribution with (_a; _2
a) as mean
and variance. Suppose that for the utility class U = fu(:; _)g_2_, the corresponding preference class has V C+(PU) = +1. We define now preferences of the form x0
a = (xa; ea) %_;_ (yb; eb) = y0
b
if and only if u(xa; _) + _a _ u(xb; _) + _b. Then , it follows that V C+(f%_;_g(_;_)2RA_) = +1.
Preference heterogeneity could be specified by a product probability measure __ __ 2 _(_ _ RA), where __ follows a multivariate normal distribution as a product of independent normal distributions with parameters f(_a; _2
a)ga2A. Then, without any restrictions on __ and not restricting f(_a; _2
a)ga2A,
it obtains that the implied stochastic choice model would be non-predictable. This follows from the same arguments as in the proof of Theorem 4.2.
For these models, the above result establishes non-predictability or irrecoverability of heterogeneity in a certain sense (uniform). The dimension being infinite has the consequence that the preference class is complex enough to significantly affect the predictability of the model. The variability in choice behaviour permitted by the model is large enough to deter good worst case guarantees on out-ofsample estimation error.

4.3 Cognitive Heterogeneity

In this section, we will study stochastic choice in the setting where there is data on choice from menus of alternatives. As before, A is a finite set of alternatives. Hence, we will have X = 2Anf;g. For a choice probability function _ : X ! _(A), for each menu of alternatives x = A 2 X, the probability vector _(A) has full support in A. For each a 2 A, the quantity _a(A) denotes the probability of alternative a being chosen from the menu A. We will assume the following holds for _.
Definition 4.1. (Positivity) A stochastic choice function _ satisfies positivity if for each menu A and a 2 A, we have _a(A) > 0.
In this context, the choice probabilities from the Logit or the Luce model are as follows. There exists a function w : A ! R++, which assigns weights to the various alternatives. For each menu A, the alternative a 2 A is chosen with probability
_a(A) =
w(a) X
b2A
w(b)
:
It is well-known that under positivity, a stochastic choice map _ has the Luce representation if and only if _ satisfies the Independence of Irrelevant Alternatives axiom (IIA). The IIA axiom says that if A and B are two menus and we have two alternatives a; b 2 A \ B. Then,
_a(A)
_b(A)
=
_a(B)
_b(B)
:
Another example is the stochastic choice model involving consideration sets (Manzini and Mariotti [2007]). The model is defined as follows. There is a function : A ! (0; 1) and a strict order _ on A. The choice probabilities are given as follows.
When jj = 2, the Max-min model has VC dimension equal to 2.

_;_

$_a(A) := (1 \square (a))$
$\prod_{b2A:b\_a}$
(b).
The interpretation is that $1 \square (a)$ is the probability with which alternative a will be considered in choosing from a given menu A. Once the considered alternatives are determined, the agent chooses according to the strict order _. The following result obtains in this setting.
Proposition 4.3. Suppose _ is either the Logit/Luce model or the model with consideration sets. Then, under the log rule S_{log}, the Pollard dimension of $S_{log} \_ \_$ is at most $O(jAj^2)$.
Proof. The proof is in the appendix and applies a result on the VC dimension of neural networks (see Proposition 6.2). The bound is derived by computing the total number of arithmetic operations needed to express the choice probabilities in the respective models. Given that the expressions can be written via addition, multiplication and division of real numbers, the number of operations, is at most linear in jAj. Together with the dimensionality of the parameter space, which is also linear in jAj (every strict order has a one-to-one numeric representation), we get the quadratic bound.
The techniques from the previous section can be extended to allow recoverability of cognitive heterogeneity. This type of heterogeneity corresponds to a distribution _ over (;_). By arguments similar to the proof of Theorem 4.1, one shows that finite pollard dimension of $S_{log} \_ \_$ allows for consistent estimation of the true distribution $\__0$, if we place a restriction that the choice probabilities are bounded below. Since we are applying the log scoring rule, the estimator is essentially based on the conditional maximum likelihood procedure.

4.4 Non-uniform Prediction

The notion of consistency considered above requires that convergence should take place uniformly over both the distribution over characteristics $\__0$ and the true stochastic choice function $\__0$. The main advantage of such a requirement is that it leads to estimates that are robust with respect to the data generating process. Suppose, we instead consider non-uniform consistency which only requires that the divergence between true and estimated choice probabilities converges to zero. We provide a definition.
Definition 4.2. A prediction method $\hat{\_}$ is pointwise consistent (with respect to d and _) if for all $0 < "; \_ < 1$, and for all $\__0 2 \_(X)$, $\__0 2 \_$, there exists $N("; \_; \__0; \__0)$ such that for all $n \_ N("; \_; \__0; \__0)$,
$\__0^{\ \_^n}$
$_0$
$\square$
$fz_n : d_{\__0}(\hat{\_}(z_n); \__0) < "g$
$\_$
$\_ 1 \square \_$:
We say that a model _ is pointwise predictable with respect to d if there exists a prediction method $\hat{\_}$, which is pointwise consistent with respect to d and _.
The above is also a kind of PAC criterion but only requires that the random variable $d_{\__0}(\hat{\_}(z_n); \__0)$ converge in probability to zero. Note that we had a stronger notion of consistency before. Hence, if a stochastic choice model is predictable according to Definition 2.1, then it is also pointwise predictable. It turns out that one can study pointwise predictability via a weaker definition of Glivenko-Cantelli classes called universal Glivenko-Cantelli classes (see, for example, Dudley et al. [1991],van Handel [2013]). This requires that for a real-valued function class F (defined on a set Z) we have that $sup_{f2F} j1=n$
P_n
$_{i=1} f(z_i) \square E\_[f]j$ converges to zero for each probability measure $\_ 2 \_(Z)$. In contrast to uniform Glivenko-Cantelli classes, the convergence need not be uniform over _. Note, however, we still have uniformity over F as we have convergence of the supremum. This would also be closer to the uniformity criterion typically required in extremum estimation (see Amemiya [1985]). Similar to the proof of Proposition 3.1, one can argue that for a stochastic choice model _ and scoring rule S, if $S \_ \_$ is universal Glivenko-Cantelli, then _ is pointwise predictable w.r.t d_S.

A useful characterisation of universal Glivenko-Cantelli classes is given by van Handel [2013] in terms of Boolean _-independence. A sequence of functions $ff_jg_{j2N} \_ F$ is Boolean _-independent at levels $\_ > \_ > 0$ if for every subset $F \_ N$,
$\setminus$
$_{j2F}$
$fz \; j \; f_j(z) < \_g \setminus$
$\setminus$
$_{j \not=2F}$
$fz \; j \; f_j(z) > \_g \; 6= ;:$

It then holds that a function class is universal Glivenko-Cantelli if and only if there exists no Boolean _-independent sequence in F for any _; _. From this, it follows that a parametrised function class F = ff(:; _) j _ 2 _g, where _ is a compact metric space and f(z; :) is continuous in _ for each z, is a universal Glivenko-Cantelli class. We can then show that under some conditions, a stochastic preference model supported on continuous preferences would be pointwise predictable. As before, it will be convenient to work with the closed convergence topology on continuous preferences (see appendix). We state the result below.

Proposition 4.4. Let P_{CONT} be the set of all continuous preferences over the set of alternatives X, equipped with the closed convergence topology. Further, let M _ _(P_{CONT}) be any set of probability measures satisfying the condition 4.1, that is closed in the topology of weak convergence. Then, $\__{P_{CONT}}(M)$:= f_(:; _ j _ 2 Mg is pointwise predictable with respect to any L_p metric.

Proof. Can be found in the appendix.

It is interesting to compare the above result with the results in Theorem 4.1 and Theorem 4.2. Firstly, there may be continuous preferences with infinite VC dimension (for eg. all continuous convex preferences) and there may be discontinuous preferences with finite VC dimension (for eg. lexicographic preferences). Secondly, conditional on considering only continuous preferences, which allows pointwise predictability, robust estimation is characterised by the dimension of the preference classes (V C and V C_+). If the dimension is finite, we get uniform consistency but for infinite dimension, the estimates are non-robust, to an extent that the same estimation procedure that would guarantee pointwise learning, would lead to erroneous out-of-sample predictions in the worst case. Lastly, the techniques for obtaining rates of convergence for pointwise predictability, would involve deriving bounds on the bracketing or covering numbers of the function class S _ _ (see Dudley et al. [1991]). This would also be similar to PAC learning problems in the deterministic case with a non-uniform predictability criterion (see Blumer et al. [1987],Benedek and Itai [1994],Shalev-Shwartz and Ben-David [2014]) and also structural risk minimisation (see Vapnik and Chervonenkis [1974] and Vapnik [1998]).

5 Conclusion

This paper presents new learning rules for estimation of choice probabilities in various stochastic choice models. The main feature of preferences that is taken in account is its model complexity, for example, the VC dimension of the model. Perhaps interestingly, given that VC dimension is a characteristic of the collection of sets that defines the model, it also characterises uniform learning involving distributions over the collection, leading to random discrete choice. We also present a wide class of models and learning rules, allowing for applicability to discrete choice models used in statistics.

References

Naoki Abe, Manfred K Warmuth, and Jun-ichi Takeuchi. Polynomial learnability of probabilistic concepts with respect to the kullback-leibler divergence. In Proceedings of the fourth annual

workshop on Computational learning theory, pages 277–289. Morgan Kaufmann Publishers Inc., 1991.

Naoki Abe, Jun-ichi Takeuchi, and Manfred K Warmuth. Polynomial learnability of stochastic rules with respect to the kl-divergence and quadratic distance. IEICE Transactions on Information and Systems, 84(3):299–316, 2001.

Charalambos D Aliprantis and Kim C Border. Infinite Dimensional Analysis: A Hitchhiker's Guide. Springer Science & Business Media, 2006.

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. Journal of the ACM (JACM), 44(4):615–631, 1997.

Takeshi Amemiya. Advanced econometrics. Harvard university press, 1985.

Martin Anthony and Peter L Bartlett. Neural network learning: Theoretical foundations. cambridge university press, 2009.

Susan Athey and Guido W Imbens. Machine learning methods that economists should know about. Annual Review of Economics, 11, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research, 3(Nov):463–482, 2002.

Pathikrit Basu and Federico Echenique. Learnability and models of decision making under uncertainty. In Proceedings of the 2018 ACM Conference on Economics and Computation, pages 53–53. ACM, 2018.

Pathikrit Basu and Federico Echenique. On the falsifiability and learnability of decision theories. Theoretical Economics, 15(4):1279–1305, 2020.

Eyal Beigman and Rakesh Vohra. Learning from revealed preference. In Proceedings of the 7th ACM

Conference on Electronic Commerce, pages 36–42. ACM, 2006.
Gyora M Benedek and Alon Itai. Nonuniform learnability. Journal of Computer and System Sciences, 48(2):311–323, 1994.
Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. Econometrica: Journal of the Econometric Society, pages 841–890, 1995.
Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. Information processing letters, 24(6):377–380, 1987.
Kim C Border and Uzi Segal. Dynamic consistency implies approximately expected utility preferences. Journal of Economic Theory, 63(2):170–188, 1994.
Richard A Briesch, Pradeep K Chintagunta, and Rosa L Matzkin. Nonparametric discrete choice models with unobserved heterogeneity. Journal of Business & Economic Statistics, 28(2):291–307, 2010.
Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Working draft, November, 3, 2005.
Christopher P Chambers, Federico Echenique, and Nicolas S Lambert. Recovering preferences from finite data. Econometrica, 89(4):1633–1664, 2021.

Zachary Chase and Siddharth Prasad. Learning Time Dependent Choice. In Avrim Blum, editor, 10th Innovations in Theoretical Computer Science Conference (ITCS 2019), volume 124 of Leibniz International Proceedings in Informatics (LIPIcs), pages 62:1–62:19. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018. ISBN 978-3-95977-095-8.
Soo Hong Chew, Larry G Epstein, Uzi Segal, et al. Mixture symmetry and quadratic utility. Econometrica, 59(1):139–63, 1991.
Pierpaolo De Blasi, Lancelot F James, John W Lau, et al. Bayesian nonparametric estimation and consistency of mixed multinomial logit choice models. Bernoulli, 16(3):679–704, 2010.
Anthony Downs et al. An economic theory of democracy. 1957.
Richard M Dudley. Uniform central limit theorems, volume 142. Cambridge university press, 2014.
Richard M Dudley, Evarist Gin´ e, and Joel Zinn. Uniform and universal glivenko-cantelli classes. Journal of Theoretical Probability, 4(3):485–510, 1991.
Drew Fudenberg, Ryota Iijima, and Tomasz Strzalecki. Stochastic choice and revealed perturbed utility. Econometrica, 83(6):2371–2409, 2015.
Itzhak Gilboa. Theory of decision under uncertainty, volume 45. Cambridge university press, 2009.
Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.
Faruk Gul and Wolfgang Pesendorfer. Random expected utility. Econometrica, 74(1):121–146, 2006.
Werner Hildenbrand. On economies with many agents. Journal of economic theory, 2(2):161–188, 1970.
Werner Hildenbrand. Core and Equilibria of a Large Economy.(PSME-5). Princeton university press, 2015.
Gil Kalai. Learnability and rationality of choice. Journal of Economic theory, 113(1):104–117, 2003.
Yakar Kannai. Continuity properties of the core of a market. Econometrica (pre-1986), 38(6):791, 1970.
Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. Journal of Computer and System Sciences, 48(3):464–497, 1994.
Michael J Kearns and Umesh Vazirani. An introduction to computational learning theory. MIT press, 1994.
Alexander Kechris. Classical descriptive set theory, volume 156. Springer Science & Business Media, 2012.
David Kreps. Notes on the Theory of Choice. Routledge, 1988.
Jay Lu. Random choice and private information. Econometrica, 84(6):1983–2027, 2016.
R Duncan Luce. Individual choice behavior: A theoretical analysis. John Wiley, 1959.
R Duncan Luce. The choice axiom after twenty years. Journal of mathematical psychology, 15(3): 215–233, 1977.

Charles F Manski. Maximum score estimation of the stochastic utility model of choice. Journal of econometrics, 3(3):205–228, 1975.
Charles F Manski. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. Journal of econometrics, 27(3):313–333, 1985.
Paola Manzini and Marco Mariotti. Sequentially rationalizable choice. American Economic Review, 97(5):1824–1839, 2007.

Andreu Mas-Colell. On the continuous representation of preorders. International Economic Review, pages 509–513, 1977.
Rosa L Matzkin. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. Econometrica: Journal of the Econometric Society, pages 239–270, 1992.
Daniel McFadden. Modeling the choice of residential location. Transportation Research Record (673), 1978.
Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. Journal of applied Econometrics, 15(5):447–470, 2000.
Daniel L McFadden. Econometric analysis of qualitative response models. Handbook of econometrics, 2:1395–1457, 1984.
Ilya Molchanov. Theory of random sets, volume 19. Springer, 2005.
Partha Niyogi. The computational nature of language learning and evolution, volume 43. MIT press Cambridge, MA, 2006.
Partha Niyogi. The informational complexity of learning: perspectives on neural networks and generative grammar. Springer Science & Business Media, 2012.
Partha Niyogi, Stephen Smale, and Shmuel Weinberger. A topological view of unsupervised learning from noisy data. SIAM Journal on Computing, 40(3):646–663, 2011.
Matthew Parry et al. Linear scoring rules for probabilistic binary classification. Electronic Journal of Statistics, 10(1):1596–1607, 2016.
David Pollard. Convergence of stochastic processes. Springer Science & Business Media, 1984.
Leonard J Savage. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 66(336):783–801, 1971.
Reinhard Selten. Axiomatic characterization of the quadratic scoring rule. Experimental Economics, 1(1):43–61, 1998.
Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.
Michel Talagrand et al. The glivenko-cantelli problem. The Annals of Probability, 15(3):837–870, 1987.
Leslie G Valiant. A theory of the learnable. Communications of the ACM, 27(11):1134–1142, 1984.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In Weak convergence and empirical processes, pages 16–28. Springer, 1996.
Ramon van Handel. The universal glivenko–cantelli property. Probability Theory and Related Fields, 155(3-4):911–934, 2013.
Vladimir Vapnik. Statistical learning theory. 1998, volume 3. Wiley, New York, 1998.
Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.
Vladimir N Vapnik and Aleksei Yakovlevich Chervonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. In Doklady Akademii Nauk, volume 181, pages 781–783. Russian Academy of Sciences, 1968.
Kenji Yamanishi. A learning criterion for stochastic rules. Machine Learning, 9(2-3):165–203, 1992.
Morteza Zadimoghaddam and Aaron Roth. Efficiently learning from revealed preference. In International Workshop on Internet and Network Economics, pages 114–127. Springer, 2012.

6 Appendix

6.1 Technical

When dealing with preferences in economic contexts, a popular topology on the space of preference relations is that given by the topology of closed convergence (Kannai [1970], Hildenbrand [1970]). We provide some general definitions in this section and note some important results from the literature that will be useful in our setting. Let $(Y; \__Y)$ be a topological space. Let C denote the set of all closed subsets of Y . The closed convergence topology on C is defined as follows (see also Aliprantis and Border [2006]). For any compact $K \_ Y$ and finitely many open sets $fU_i g_{i2I}$ in Y , we define the following subset of C.

$O(K; fU_i g_i) = fF 2 C j F \setminus K = ;$ and $F \setminus U_i 6= ;$ for each i 2 Ig:

All sets of the form $O(K; fU_i g_i)$ constitute a basis for the closed convergence topology on C. We will denote this topology as $\__C$. When Y is a locally compact and seperable metric space, then it turns out that $(C; \__C)$ is a compact metrizable space (see Hildenbrand [2015]). Further, it also holds that $(C; \__C)$ is second countable (see Molchanov [2005]).

We now note how the closed convergence topology provides us a way to topologize the space of continuous preferences. Let $(X; \__X)$ be a topological space of alternatives. A preference relation on X is a binary relation $\%\_ X \_ X$ which is complete and transitive. A continuous preference relation

% is a preference relation that is closed in X _ X (under the product topology). Hence, a class of continuous preferences P _ 2X_X can be endowed with the subspace topology, say _P, generated by the topology of closed convergence on closed subsets of 2X_X.
The following theorem, due to Mas-Colell [1977], provides conditions under which a utility representation exists that is jointly continuous in the alternatives and the preferences.
Proposition 6.1. (Mas-Colell [1977]) Let X be a locally compact and second countable space of alternatives. Further, let PCONT denote the set of all continuous preference relations on X, equipped with the subspace topology induced by the closed convergence topology. Then, there exists a utility function U : X _ PCONT ! R, which is jointly continuous and each %2 PCONT is represented by U(:;%).

An immediate corollary of the above result is the following. For any finite set of alternatives (xa)a2A, the set of preferences f%2 PCONT j xa % xb for all b 6= ag is closed and f%2 PCONT j xa _ xb for all b 6= ag is open with respect to the closed convergence topology.
Now, consider a subspace of preferences P _ PCONT . We denote as _(P), the set of all Borel probability measures defined on P. We will endow _(P) with the topology of weak convergence (see Aliprantis and Border [2006]). Hence, for a sequence f_ngn and probability measure _ in _(P) we have that _n ! _ in the weak convergence topology if and only if for every bounded continuous function f : P ! R, the convergence E_n[f(%)] ! E_[f(%)] holds. Now, if the space of alternatives X is a locally compact and seperable metric space, then it is also the case that X _ X is a locally compact and seperable metric space. Hence, the subspace PCONT _ 2X_X is a secound countable compact metric space. This means that for any subspace P _ PCONT , the set of probability measures _(P) with the weak convergence topology is a seperable metrizable space (see Aliprantis and Border [2006]). We now show some measurability and continuity properties of the stochastic choice probabilities _a(x; _).
We will first define the regularity condition that we assume on _ 2 _(P). Consider the set of preferences where xa is strictly preferred from a menu of alternatives x = (xb)b2A i.e. Q(x; a) = f%2 P j xa _ xb for all b 6= ag. The regularity assumption imposes on _ that we have _([a2AQ(x; a)) = 1. Also, by the definition of choice probabilities, _a(x; _) = _(Q(x; a)). The following lemma is useful.
Lemma 6.1. Let X be a locally compact, seperable metric space of alternatives and let X _ XA be such that xa 6= xb for all x = (xc)c2A 2 X. Suppose P _ PCONT and let _r(P) _ _(P) be the set of all probability measures on the preference class P satisfying the regularity condition. The stochastic choice function _a(x; _) is continuous in _ for each x 2 X; continuous in x for each _ 2 _r(P). Hence, _a(x; _) is jointly measurable under the Borel sigma-algebra induced by the product topology on X _ _r(P).
Proof. Consider a fixed x 2 X and let f_ngn be a sequence and let _ be a probability measure in _r(P) such that _n ! _ in the weak convergence topology. It suffices to argue that Q(x; a) is a continuity set according to the measure _ i.e. _(@Q(x; a)) = 0, where the set @Q(x; a) is the boundary of the set Q(x; a). Suppose we have a convergent seqeuence f%ngn in Q(x; a) with limit %. As _ satisfies the regularity condition _([a2AQ(x; a)) = 1, it suffices to show that %2 f%02 PCONT j xa %0 xb for all b 6= ag. But this is true since the latter set is closed and contains the set Q(x; a) as noted above. It now follows that _a(x; :) is continuous.
Now, let _ 2 _r(P) and suppose fxng _ X is a convergent sequence with limit x 2 X. Note that since preferences are continuous, we have that Q(x; a) _ lim infn Q(xn; a). This implies that _(Q(x; a)) _ _(lim infn Q(xn; a)) _ lim infn _(Q(xn; a)). Hence, _a(:; _) is lower semi-continuous.
We now show upper-semicontinuity. Define the set Q(x;Anfag) := [b2AnfagQ(x; a). Note that since _ satisfies regularity, we have _(Q(y; a)) = 1 □ _(Q(y;Anfag)) for all y 2 X. Now, let xn ! x. Again, as preferences are continuous, we obtain that Q(x;Anfag) _ lim infn Q(xn;Anfag) which means _(Q(x;Anfag)) _ _(lim infn Q(xn;Anfag)) _ lim infn _(Q(xn;Anfag)). But this implies that _(Q(x; a)) _ lim supn _(Q(xn; a)). Hence, the function _a(:; _) is continuous as it is both lower and upper semi-continuous.
The two facts established above imply that _a(x; _) is a Caratheodory function (see, for example, Aliprantis and Border [2006], Kechris [2012]). Hence, _a(x; _) is jointly measurable in (x; _).
Now it follows that under the conditions satisfied by the above lemma, we obtain appropriate measurability of the function class F = Sbr _ _P for the Glivenko-Cantelli theorems to apply. From Van Der Vaart and Wellner [1996], a sufficient condition is a seperability condition called pointwise

measurability. It requires that there exist a countable subset F0 _ F such that for any f 2 F, there exists a sequence in F0 which converges to f pointwise. Since _(P) is a seperable metric space, it follows that _r(P) is a seperable metric space as well. Let _ _ _r(P) be a countable dense subset.

Since from Lemma 6.1 above, the choice probability $\rho_a(x; \lambda)$ is continuous in λ and from the definition of the Quadratic scoring rule S_{br}, it follows that the countable set $F_0 = \{S_{br} \circ (\cdot; \lambda) \mid \lambda \in \Lambda\}$ makes F pointwise measurable. Alternatively, Talagrand et al. [1987] requires that F belong to $L_1(\mu)$, for each $\lambda \in \Lambda(X \times A)$.

We will now discuss the existence of probability measures μ which satisfy the regularity condition. The following result applies when preferences are defined through a parametrised utility function $u : X \times \Lambda \to \mathbb{R}$. For any such function u and $\lambda \in \Lambda$, we will denote as $\succeq_\lambda$, the preference relation u induces on X. Hence, we define $x \succeq_\lambda y$ if and only if $u(x; \lambda) \geq u(y; \lambda)$. We will say that a probability measure $\mu \in \Delta(\Lambda)$ satisfies the regularity condition for u if for all $x = (x_b)_{b \in A} \in X^A$ with $x_a \neq x_b$ for all $a \neq b$, we have $\mu(\{\lambda \in \Lambda \mid u(x_a; \lambda) > \max_{b \neq a} u(x_b; \lambda)\}) = 1$.

The following result holds. We will assume that preferences are locally strict (see Border and Segal [1994]).

Lemma 6.2. Let X be a locally compact and seperable metric space of alternatives, $\mathcal{P}^{CONT}$ be the set of all continuous preference relations on X (endowed with the topology induced by the closed convergence topology) and Λ be a metric space of parameters. Let $u : X \times \Lambda \to \mathbb{R}$ be a jointly continuous utility function. Further, suppose that u is locally strict i.e. for all λ, if $u(x; \lambda) \geq u(y; \lambda)$, then every neighbourhood of $(x; y)$ contains a pair $(x_0; y_0)$ such that $u(x_0; \lambda) > u(y_0; \lambda)$. Then, the function $\lambda : \Lambda \to \mathcal{P}^{CONT}$ defined as

$$\lambda(\lambda) = \succeq_\lambda;$$

is a continuous function. Now, suppose $\mu \in \Delta(\Lambda)$ is a Borel probability measure on the space Λ which satisfies the regularity condition for u. Also suppose that the class of preferences $\mathcal{P}_u = \{\succeq_\lambda \mid \lambda \in \Lambda\}$ is Borel. Then, the induced probability measure μ_λ on the Borel subsets of $\mathcal{P}^{CONT}$, defined as

$$\mu_\lambda(B) = \mu(\lambda^{-1}(B));$$

satisfies the regularity condition on $\mathcal{P}_u$. If $\mathcal{P}_u$ is also closed in $\mathcal{P}^{CONT}$ and Λ has full support on Λ, then μ_λ has full support on $\mathcal{P}_u$.

Proof. To show that λ is continuous, it will suffice to show that the pre-image of any basic open set is open in Λ. Let $O(K; \{U_i\}_{i \in I})$ be a basic open set. Now, each U_i is open in $X \times X$. Hence, from local strictness of u it follows that for any λ, if $u(x; \lambda) \geq u(y; \lambda)$, where $(x; y) \in U_i$, there exists another pair $(x_0; y_0) \in U_i$ such that $u(x_0; \lambda) > u(y_0; \lambda)$. For convenience, define the function $\rho_u(x; y; \lambda) = u(x; \lambda) - u(y; \lambda)$. Then, we get that the pre-image of $O(K; \{U_i\}_{i \in I})$ can be written as

$$\lambda^{-1}(O(K; \{U_i\}_{i \in I})) = \{\lambda \in \Lambda \mid \max_{(x;y) \in K} \rho_u(x; y; \lambda) < 0 \text{ and } \sup_{(x;y) \in U_i} \rho_u(x; y; \lambda) > 0 \text{ for each } i \in I\}.$$

Now, as K is compact, from the theorem of the maximum, it follows that $\max_{(x;y) \in K} \rho_u(x; y; \lambda)$ is continuous in λ. Hence, the set $\Lambda_K = \{\lambda \in \Lambda \mid \max_{(x;y) \in K} \rho_u(x; y; \lambda) < 0\}$ is open in Λ. Now, note that $\sup_{(x;y) \in U_i} \rho_u(x; y; \lambda)$ is a supremum of lower semi-continuous functions and is hence, lower semi-continuous in λ. Hence, the set $\Lambda_i = \{\lambda \in \Lambda \mid \sup_{(x;y) \in U_i} \rho_u(x; y; \lambda) > 0\}$ is open, for each $i \in I$.

To conclude, observe that $\lambda^{-1}(O(K; \{U_i\}_{i \in I})) = \Lambda_K \setminus (\bigcup_{i \in I} \Lambda_i)$ is open as an intersection of finitely many open sets. Hence, λ is continuous.

We now argue that μ_λ satisfies the regularity condition if μ satisfies the regularity condition for u. Let $B = \bigcup_{a \in A} Q(x; a)$. Then $\mu_\lambda(B) = \mu(\{\lambda \in \Lambda \mid u(x_a; \lambda) > \max_{b \neq a} u(x_b; \lambda)\}) = 1$. Now, further

assume that μ has full support on Λ and $\mathcal{P}_u$ is closed. Suppose for contradiction that there exists a closed strict subset $P \subset \mathcal{P}_u$ such that $\mu_\lambda(P) = 1$. But this means that $\mu(\lambda^{-1}(P)) = 1$ where the set $\lambda^{-1}(P)$ is closed as λ is continuous and P is closed in $\mathcal{P}^{CONT}$. But $\lambda^{-1}(P) \subset \lambda^{-1}(\mathcal{P}_u) = \Lambda$, where both sets are closed but $\mu(\lambda^{-1}(P)) = 1$ meaning that μ does not have full support, which is a contradiction.

6.2 Proofs from Section 3

The following is the proof for Proposition 3.1.

Proof. We show that μ is predictable with respect to d_S. We shall construct an almost-ERM learning $\hat{\lambda}_E$, which would be consistent with respect to d and μ. For each $n \in \mathbb{N}$, define $\varepsilon_n := 1/n$. Now, for $z_n = ((x_i; a_i))_{i=1}^n \in Z_n$, let $\hat{\lambda}_E(z_n)$ be such that

$$\hat{V}_S(\hat{\lambda}_E(z_n)) \geq \inf_{\lambda \in \Lambda} \hat{V}_S(\lambda) + \varepsilon_n;$$

where recall that for any $\lambda \in \Lambda$, $\hat{V}_n(\lambda) := 1/n \sum_{i=1}^n$

$\sum_{i=1} S_{a_i}(\_(x_i))$. Note that the infimum exists in the above definition since $S \_ \_$ is bounded above by M. Hence, the $\hat{}\_E$ is well-defined. We now show it is consistent.

Let $(\text{"}; \_) \in (0; 1)^2$. Since $S \_ \_$ is a Glivenko-Cantelli class of functions, for $(\text{"}=3; \_) \in (0; 1)^2$, there exists $N_0(\text{"}=3; \_)$ such that for all $n \_ N_0(\text{"}=3; \_)$, for all $\_ \in \_(Z)$

$$\_^n$$

$$\_$$

$$z_n : \sup_{f \in S\_\_} \left| \frac{1}{n} \sum_{i=1}^{n} f(z_i) - E\_(f) \right| < \text{"}=3$$

$$\_$$

$$\_ 1 - \_: \quad (6.1)$$

Now, let $N(\text{"}; \_) := \max\{3=\text{"}; N_0(\text{"}; \_)\}$. Now, suppose $n \_ N(\text{"}; \_)$. Let $\__0 \in \_(X)$ be a distribution over characteristics and suppose the true choice probabilities are given by $\__0 \in \_$. For convenience, denote $E_{z_n}(f) = 1=n \sum_{i=1}^{n} f(z_i)$.

Now suppose z_n is such that $\sup_{f \in S\_\_} |E_{z_n}(f) - E\_{\__0}(f)| < \text{"}=3$. Note that this happens with probability at least $1 - \_$ under $\__0^n$. Consider the almost-ERM rule $\hat{}\_E$. Define also $f_{E;z_n}(x; a) := S(\hat{}\_E(z_n)(x); a)$, $f_0(x; a) := S(\__0(x); a)$ and $\__0 = \__0 \ \__0$. It follows that

$$\left| E\_{\__0}(f_{E;z_n}) - E\_{\__0}(f_0) \right| = E\_{\__0}(f_0) - E\_{\__0}(f_{E;z_n}) \quad (6.2)$$

$$= E\_{\__0}(f_0) - E_{z_n}(f_{E;z_n}) + E_{z_n}(f_{E;z_n}) - E\_{\__0}(f_{E;z_n})$$

$$\_ E\_{\__0}(f_0) - E_{z_n}(f_{E;z_n}) + \text{"}=3 \quad (6.3)$$

$$\_ E\_{\__0}(f_0) - E_{z_n}(f_0) + 1=n + \text{"}=3 \quad (6.4)$$

$$\_ \text{"}=3 + \text{"}=3 + \text{"}=3 \quad (6.5)$$

$$= \text{"}:$$

Here, 6:2 follows since $\__0$ minimises expected risk as S is incentive compatible (Lemma 3.1). The inequality 6.3 follows by assumption that $\sup_{f \in S\_\_} |E_{z_n}(f) - E\_{\__0}(f)| < \text{"}=3$. This yields $E_{z_n}(f_{E;z_n}) - E\_{\__0}(f_{E;z_n}) < \text{"}=3$. Also, 6.4 follows since $\hat{}\_E$ is almost-ERM. Lastly, 6.5 follows since $n \_ 3=\text{"}$ and again from our assumption that $\sup_{f \in S\_\_} |E_{z_n}(f) - E\_{\__0}(f)| < \text{"}=3$, which yields $E_{z_n}(f_0) - E\_{\__0}(f_0) < \text{"}=3$.

Now, since we have

$$E\_{\__0}(f_0) - E\_{\__0}(f_{E;z_n}) = \int_X d_S(\hat{}\_E(z_n)(x); \__0(x)) \, d\_{\__0}(x);$$

it follows that $\hat{}\_E$ is consistent with respect to d_S and $\_$.

6.3 Proofs from Section 4

The following two lemmas will be useful.

Lemma 6.3. Suppose F is a uniformly bounded class of real-valued functions defined on the set Z and let$M_ \in _(F)$. For $\_ \in \_(F)$. Define the function $f\_$ as

$$f\_(z) = \int_F f(z) \, d\_(f):$$

Consider the function class $F_M = \{f\_ : \_ \in M\}$. Then,

$$R_{z_n}(F_M) \_ R_{z_n}(F):$$

for all $z_n \in Z_n$.

Proof. We first develop some notation. For z_n and a real-valued function f, define $f(z_n) = (f(z_1); \ldots; f(z_n)) \in R_n$. From the definition of Rademacher Complexity, it suffices to show that for any real vector $\_ \in R_n$,

$$\sup_{\_ \in M} \left| \int_F \_\_ f(z_n) \, d\_(f) \right| \_ \sup_{f \in F} \left| \_\_ f(z_n) \right|: \quad (6.6)$$

Suppose for contradiction that 6.6 does not hold. Then, $\sup_{\_ \in M}$

R
F __f(zn)d_(f) > supf2F __f(zn).
Then, there exists _ 2 M such that
R
F _ _ f(zn)d_(f) > supf2F _ _ f(zn). But this means that
there exists f_ such that _ _ f_(zn) > supf2F _ _ f(zn). This is a contradiction and hence, the result
follows.
The following is another useful lemma.
Lemma 6.4. Suppose, as in the present context, Z = X _ A, where jAj < 1. Further, let F be a
real-valued function class defined on Z. For each a 2 A, define the following function class on X:
Fa = ff(:; a) j f 2 Fg:
Then, for all zn = (xi; ai)n
i=1 2 Zn,
Rzn(F) _
X
a2A
na
n
Rxna (Fa);
where xna = (xi)i2fija i=ag and na = jfi j ai = agj.
Proof. Let zn 2 Zn. Define Na = fi j ai = ag and then na = jNaj. Then,
Rzn(F) = E_
h
sup
f2F
1
n
Xn
i=1
_if(xi; ai)
i
= E_
h
sup
f2F
1
n
X
a2A
X
i2Na
_if(xi; a)
i
_ E_
hX
a2A
sup
f2F
1
n
X
i2Na
_if(xi; a)
i
=
X
a2A
E_
h
sup
f2F
1
n
X
i2Na
_if(xi; a)

i
=
X
a2A
na
n
E_
h
sup
f2F
1
na
X
i2Na
_if(xi; a)
i
=
X
a2A
na
n
Rxna (Fa):
This is the set of all probability measures on F.

6.4 Proofs from Section 4.3

Proposition 4.3 follows straightforwardly from the following result on the VC dimension of neural networks.

Proposition 6.2. (Anthony and Bartlett [2009]) Let _ _ Rp. Suppose F = ff(w; _) : _ 2 _g is a parametrized class of 0-1 valued functions defined on a set W _ Rk. Further, suppose, for each f 2 F and each (w; _) 2 _, computing the value of f(w; _) takes no more than t many operations from the following

1. The arithmetic operations +;□;_ and n defined on real numbers.
2. Pairwise comparisons of two real numbers involving the relations _; >;_;< and =; 6=.
3. Output 0 or 1.

Then, the VC dimension of F is at most 4p(t + 2).

We now prove Proposition 4.4.

Proof. We will first show that if we have a parametrised function class F = ff(z; _) j _ 2 _g, where _ is a compact metric space and f(z; :) is continuous in _ for z, then F contains no Boolean _-independent sequence for any _; _. Suppose not. Then, there exists a sequence of parameters f_ng1
n=1 _ _ such that ff(:; _n)gn is a Boolean _-independent sequence in F for some 0 < _ < _. Since _ is compact, there exists a convergent subsequence f_nkg1
k=1. It then follows that ff(:; _nk)gk
is also a Boolean _-independent sequence at _; _. Now consider the sets KO = fk j k odd g and KE = fk j k even g. The following set is non-empty.

\
k2KO
fz j f(z; _nk) < _g \
\
k2KE
fz j f(z; _nk) > _g 6= ;:

Consider an element z in the above intersection. Further, let __ 2 _ be the limit of f_nkg1
k=1. Then,
limk!1 f(z; _nk) = f(z; __). However, from the above intersection, it follows that f(z; _nk) < _ for k odd and f(z; _nk) > _ for k even. Hence, we have two subsequences ff(z; _nk)gk2KO and ff(z; _nk)gk2KE with different limits. But this is a contradiction as all subsequences of a convergent sequence must have the same limit.

We now show the pointwise predictability of continuous preferences. Note that PCONT = f%_ X _X j% closed in X _Xg. With the the topology of closed convergence on PCONT , which is also the topology induced by the Hausdorff metric on PCONT (see, for example, Aliprantis and Border [2006]), it follows that PCONT is compact. Hence, _(PCONT) is compact in the weak convergence topology which can metrized by the Prokhorov metric. SinceMis now a closed subset of a compact metric space, it also compact. Now, we will define _ :=M.

Consider the Quadratic scoring rule S_{br} and the stochastic choice model $\rho_{PCONT}(M)$. Then, for any $f \in S_{br} \cap \rho_{PCONT}(M)$, we have

$$f(x; a; \lambda) = \sum_{b \in A} \lambda_a(x; \lambda)\, \lambda_b(x; \lambda)^2:$$

Now we need only show that if $\lambda_n \to \lambda$, then $f(x; a; \lambda_n) \to f(x; a; \lambda)$. This follows from Lemma 6.1.
