Learnability and stochastic choice

Pathikrit Basu*

November 29, 2018

Abstract

In this paper, we study a non-parametric approach to prediction in stochastic choice models in economics. We apply techniques from statistical learning theory to study the problem of learning choice probabilities. A model of stochastic choice is said to be *learnable* if there exist *learning rules* defined on choice data that are *uniformly consistent*. We construct learning rules via the procedure of empirical risk minimization, where risk is defined in terms of incentive compatible scoring rules. This approach involves mild distributional assumptions on the model, with the main requirement being a constraint on the capacity of the admissible set of choice probabilities. Further, the approach allows us to obtain bounds on the *sample complexity* for various models of stochastic choice i.e. the minimum number of samples needed to have a precise estimate of the true choice probabilities. This allows for distribution-free, robust estimates of choice probabilities for several well-known economic models of stochastic choice. We provide several applications and derive sample complexity upper bounds in closed form, in terms of the description and parameters of the underlying stochastic choice model.

1 Introduction

The study of stochastic choice models in economics seeks to understand aggregate demand behaviour and preference heterogeneity in markets (McFadden [1978]). 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 ($\mathcal{X}$), the model prescribes the probabilities with which various items (A) would be chosen. This is defined in terms

*I would like to thank Kalyan Chatterjee, Federico Echenique, Bob Sherman, Matt Shum and Adam Wierman for comments and suggestions.

of a stochastic choice function $\sigma : \mathcal{X} \to \Delta(A)$. The probabilities $\sigma(x)$ 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]). For example, in random utility models, the choice probabilities depend on structural parameters of the agents' preferences and idiosyncratic factors, in the form of random utility shocks, that may cause preferences to alter. In this paper, we study a novel approach to estimating choice probabilities, which correspond to a model of stochastic choice, based on data available on agents' choices among alternatives.

We study the problem of learning the map σ 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 learning rules defined on choice data which lead to accurate estimation of choice probabilities and the criterion we require is *uniform consistency*. This requires that for any given precision parameters $\varepsilon, \delta > 0$, there exists a fixed data size $N(\varepsilon, \delta)$, such that if the analyst were to apply the learning rule to a data set of size above $N(\varepsilon, \delta)$, the probability would be at least $1 - \delta$ that the stochastic choice function conjectured by the learning rule would be close to the true stochastic choice probability function by at most ε. Moreover, this holds true irrespective of the process that generates choice problems and the true function σ_0. Hence, working with amount of data at least $N(\varepsilon, \delta)$ allows for robust estimates of the choice probabilities, when learning rules 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 does not know either the data generating process or the true stochastic choice function, and may wish to work with data sizes that gaurantee, in a robust mannter, a certain degree of precision in estimation. The minimum number of samples $N(\varepsilon, \delta)$, 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.

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. The novelty of the present approach lies in the introduction and assessment of sample complexity. This has two main advantages.

Firstly, the sample size $N(\varepsilon, \delta)$ 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 $\varepsilon, \delta > 0$. 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. Secondly, sample complexity also acts as a measure of complexity for stochastic choice models. Different models would have 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. 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 σ 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 learnability 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, al-

lowing 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 and also provide relatively simple arguments to establish identification of choice probabilities for a variety of stochastic choice models such as Logit, Additive Perturbed Utility and Choice with Consideration Sets.

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 learning rules, which we construct on the basis of the principle of empirical risk minimisation. In our setting, we define objective risk in terms of incentive compatible scoring rules. Finally, in section 4, we apply our approach to a variety of stochastic choice models in economics.

2 Model

Let $\mathcal{X} \subseteq \mathbb{R}^k$ be a compact set of *characteristics* and $A = \{a_1, a_2, ..., a_m\}$ denote a finite set of *alternatives*. In certain contexts, $x = (x_I, x_A) \in \mathbb{R}^k$ shall be a vector including both individual characteristics (x_I) and product characteristics ($x_A = (x_a)_{a \in A}$). In other contexts, involving choice from menus, $\mathcal{X} \subseteq 2^A = \{0,1\}^A$. Hence, at each $x \in \mathcal{X}$ which is a menu, a choice of an alternative is made, $a \in x$. We define $\mathcal{Z} = \mathcal{X} \times A$ and will denote a typical element of $\mathcal{Z}$ as z.

A *stochastic choice function* is a map $\sigma : \mathcal{X} \to \Delta(A)$. The interpretation of σ is that at each x, $\sigma(x) = \{\sigma_a(x)\}_{a \in A}$ is a probability vector where $\sigma_a(x)$ denotes the probability that alternative $a \in A$ will be chosen given that the underlying characteristic is 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 $(x_i, a_i) \in \mathcal{Z}$. Hence, a *data set* is a finite sequence

$$z^n = \{(x_1, a_1), (x_2, a_2), ..., (x_n, a_n)\}. \tag{1}$$

Hence, a data set is a element $z^n = (z_1, ..., z_n)$ of $\mathcal{Z}^n$.

We now describe the data generating process. There is a probability measure $\pi \in \Delta(\mathcal{X})$[1] which defines the distribution of characteristics in the population. Given π and a stochastic choice map σ, the data is generated as follows. Independent across i, x_i is drawn according to π and then a_i is drawn according to the choice probablities given by $\sigma(x_i)$. Note here that the analyst only observes the characteristic and the alternative chosen i.e. (x_i, a_i) through the data in 1. The analyst knows neither the distribution π nor the stochastic choice function σ, but assumes that it satisfies certain properties. We denote as $\pi \otimes \sigma$, the probability measure induced by π, σ together on the set $X \times A$. This represents the joint distribution from which the data is generated, by taking n i.i.d. samples from $\pi \otimes \sigma$. We shall denote as $\pi \otimes \sigma^n$, the n-fold product measure induced by $\pi \otimes \sigma$ on $\mathcal{Z}^n$. This is essentially the distribution of the data z^n.

2.2 Learning

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

$$\hat{\sigma} : \bigcup_{n \geq 1} (\mathcal{X} \times A)^n \to \Sigma. \tag{2}$$

Suppose now that the true choice probabilities are given by σ_0 and suppose π_0 governs the distribution over characteristics. We shall require that a learning map $\hat{\sigma}$ be so that with enough data, $z^n = \{(x_i, a_i)\}_{i=1}^n$, the estimate of the learning rule $\hat{\sigma}(z^n)$ would be close to the true choice probabilities σ_0. Here, our notion of closeness between two choice probability maps σ, σ' will be given by

$$d_{\pi_0}(\sigma, \sigma') = \int_{\mathcal{X}} d(\sigma(x), \sigma'(x)) d\pi_0(x) < \varepsilon, \tag{3}$$

where $d : \Delta(A) \times \Delta(A) \to \mathbb{R}$ denotes a divergence function or metric on the space of all choice probability vectors on A i.e $\Delta(A)$. For example, d could be Euclidean distance, KL divergence or the total variation distance. This leads us to the following definition of consistency for learning rules.

Definition 1. A learning rule $\hat{\sigma}$ is consistent (with respect to d and Σ) if for all $0 < \varepsilon, \delta < 1$,

[1] Throughout the paper, for any metric space Y, we will denote as $\Delta(Y)$, the set of all Borel probability measures on Y. For any $\nu \in \Delta(Y)$, we shall denote as ν^n, the n-fold product measure on Y^n defined by $\nu^n := \underbrace{\nu \times \nu \times .... \times \nu}_{n}$.

there exists $N(\varepsilon,\delta)$ such that for all $n \geq N(\varepsilon,\delta)$,

$$(\forall \pi_0 \in \Delta(\mathcal{X}))(\forall \sigma_0 \in \Sigma)\Big(\pi_0 \otimes \sigma_0^n\Big(\{z^n : d_{\pi_0}(\hat{\sigma}(z^n),\sigma_0) < \varepsilon\}\Big) > 1-\delta\Big). \tag{4}$$

We say that a model Σ is learnable with respect to d if there exists a learning rule $\hat{\sigma}$, which is consistent with respect to d and Σ. Finally, for a given $\varepsilon,\delta > 0$, we denote as $N(\varepsilon,\delta)$, the smallest n for which 4 holds. The function $N : (0,1)^2 \to \mathbb{N}$ is called the *sample complexity* of Σ (with respect to $\hat{\sigma}$)

In what follows, we shall discuss consistent learning rules for various stochastic choice models.

3 Consistent learning rules

3.1 Empirical Risk Minimization

We shall construct consistent learning rules based on the principle of *empirical risk minimization*. For a detailed treatment, see Vapnik [1998]. For the learning problem, we first define a loss function $V : \Sigma \times \mathcal{X} \times A \to \mathbb{R}$. Suppose the true distribution and choice probabilities are given by π_0, σ_0. Then, the *expected risk* corresponding to $\sigma \in \Sigma$ is defined as

$$\bar{V}(\sigma) = \int_{X\times A} V(\sigma,x,a) d\pi_0 \otimes \sigma_0(x,a). \tag{5}$$

A minimizer of expected risk σ^* is defined as

$$\sigma^* \in \arg\min_{\sigma\in\Sigma} \bar{V}(\sigma). \tag{6}$$

We shall consider loss functions V for which it will hold that $\sigma^* = \sigma_0$ i.e the true choice probabilities would minimise the expected risk (a property also known as Fisher consistency). The principle of empirical risk minimization involves estimating the expected risk by the empirical risk on the sample $z^n = \{(x_i,a_i)\}_{i=1}^n$. For each $\sigma \in \Sigma$, the empirical risk is given by

$$\hat{V}(\sigma) = \frac{1}{n}\sum_{i=1}^{n} V(\sigma,x_i,a_i) \tag{7}$$

The learning rule that corresponds to empirical risk minimization, denoted by $\hat{\sigma}_E$, is defined as

$$\hat{\sigma}_E(z^n) \in \arg\min_{\sigma\in\Sigma} \hat{V}(\sigma) \tag{8}$$

The minimum in 8 need not always exist. However, if it holds for some M that $V(\sigma, x, a) \geq M$ for all $\sigma \in \Sigma$ and $(x, a) \in \mathcal{Z}$, then the infimum exists and we can define an *almost-ERM* learning rule as follows. We have $\{\varepsilon_n\}_n$ such that $\varepsilon_n > 0$ and $\lim_{n\to\infty} \varepsilon_n = 0$. The almost-ERM rule selects $\hat{\sigma}_E(z^n)$ such that

$$\hat{V}(\hat{\sigma}_E(z^n)) \leq \inf_{\sigma\in\Sigma_{\mathcal{M}}} \hat{V}(\sigma) + \varepsilon_n \tag{9}$$

In this context, consistency relies on $\hat{\sigma}_E(z^n)$ being close to the true choice probability function σ^* as the function $\hat{V}$ approximates $\bar{V}$, for large enough n. Consistency here is defined as follows, in terms of V.

Definition 2. A learning rule $\hat{\sigma}$ is said to be consistent (with respect to V and Σ) if for all $0 < \epsilon, \delta < 1$, there exist an $N(\epsilon, \delta)$ such that for all $n \geq N(\epsilon, \delta)$, it holds that

$$(\forall \pi_0 \in \Delta(\mathcal{X}))(\forall \sigma_0 \in \Sigma)\Big(\pi_0 \otimes \sigma_0^n\Big(\{z^n : \bar{V}(\hat{\sigma}(z^n)) < \inf_{\sigma\in\Sigma} \bar{V}(\sigma) + \varepsilon\}\Big) > 1 - \delta\Big). \tag{10}$$

We say that a model Σ is learnable with respect to V if there exists a learning rule $\hat{\sigma}$, which is consistent with respect to V and Σ.

It turns out that a model $\mathcal{M}$ is learnable if the family of real-valued functions $V \circ \Sigma = \{V(\sigma, ., .) : \sigma \in \Sigma\}$ is a Glivenko-Cantelli class of functions (see, for example, Vapnik [1998]; Shalev-Shwartz and Ben-David [2014]). Each $f \in V \circ \Sigma$ is a function of the form $f : \mathcal{X} \times A \to \mathbb{R}$. We provide a definition below (see also Dudley (1996))

Definition 3. A class of real valued functions $\mathcal{F}$ on $\mathcal{Z}$ is said to be a Glivenko-Cantelli class of functions if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mu\in\Delta(\mathcal{Z})} \mu^n\Big(z^n : \sup_{f\in\mathcal{F}} |1/n \sum_{i=1}^{n} f(z_i) - \mathbb{E}_\mu(f)| \geq \varepsilon\Big) = 0, \tag{11}$$

where $\mathbb{E}_\mu(f)$ denotes the expectation of f under μ.

A necessary and sufficient condition for a class of real-valued functions to be a Glivenko-Cantelli class of functions is that it have finite V_γ-dimension for each $\gamma > 0$ (see, for

example, Alon et al. [1997]). Other combinatorial measures and notions of dimensions and capacity also guarantee the Glivenko-Cantelli property for a class of functions for. eg. Vapnik's V-dimension, Pollard's P-dimension, P_γ-dimension, Covering numbers and Metric Entropy. 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 $S : \Delta(A) \to \mathbb{R}^A$ (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 $p, q \in \Delta(A)$, we can define $S(p,q) := \sum_{a \in A} S_a(p) q_a$ to be the expected score when q is the true distribution over A. A scoring rule is said to be *incentive compatible* if

$$S(q,q) \geq S(p,q) \text{ for all } p, q \in \Delta(A) \tag{12}$$

i.e $p = q$ maximises the function $S(.,q)$. We say that S is *strongly incentive compatible*[2] if additionally, $p = q$ is the unique maximizer for each $q \in \Delta(A)$.

We now define a loss function based on S.

$$V^S(\sigma, x, a) := -S_a(\sigma(x)) \tag{13}$$

We can now prove the following lemma.

Lemma 1. *Let $\pi_0 \in \Delta(\mathcal{X})$ be the true distribution over characteristics and let $\sigma_0 \in \Sigma$ be the true choice probability function. Suppose S is incentive compatible. Then, σ_0 minimises $\bar{V}^S(\sigma)$. Furthermore, if S is strongly incentive compatible and $\sigma^* \in \Sigma$ minimizes $\bar{V}^S(.)$, then $\sigma^*(x) = \sigma_0(x)$ with probability one according to π_0.*

[2] 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]

Proof. For any $\sigma' \in \Sigma$, we have

$$\begin{aligned}
\bar{V}^S(\sigma') &= \int_{\mathcal{X}\times A} V^S(\sigma', x, a) d\pi_0 \otimes \sigma_0(x, a) \\
&= -\int_{\mathcal{X}} \Big[\sum_{a\in A} S_a(\sigma'(x)) \sigma_{0,a}(x) \Big] d\pi_0(x) \\
&= -\int_{\mathcal{X}} S(\sigma'(x), \sigma_0(x)) d\pi_0(x) \\
&\geq -\int_{\mathcal{X}} S(\sigma_0(x), \sigma_0(x)) d\pi_0(x) \qquad (14) \\
&= \bar{V}^S(\sigma_0),
\end{aligned}$$

where the inequality 14 follows from the fact that S is an incentive compatible scoring rule. Now, suppose σ^* minimizes $\bar{V}^S(.)$ and let $E = \{x : \sigma^*(x) \neq \sigma_0(x)\}$. Since S is strongly incentive compatible, note that $S(\sigma_0(x), \sigma_0(x)) > S(\sigma^*(x), \sigma_0(x))$ for all $x \in E$. We already have that $S(\sigma_0(x), \sigma_0(x)) \geq S(\sigma^*(x), \sigma_0(x))$ for all $x \in X$. Hence, if $\pi_0(E) > 0$, then we will have that $\bar{V}^S(\sigma_0) < \bar{V}^S(\sigma^*)$. This contradicts that σ^* minimizes $\bar{V}^S(.)$. □

The above result shows that σ_0 is the minimiser of expected risk. Each strongly incentive compatible scoring rule leads to a divergence function

$$d_S(p, q) = S(q, q) - S(p, q) \text{ for all } p, q \in \Delta(A) \qquad (15)$$

Hence, from Lemma 1, it follows that

$$\begin{aligned}
\bar{V}_S(\sigma) - \inf_{\sigma\in\Sigma_{\mathcal{M}}} \bar{V}_S(\sigma) &= \bar{V}_S(\sigma) - \bar{V}_S(\sigma_0) \\
&= \int_{\mathcal{X}} d_S(\sigma(x), \sigma_0(x)) d\pi_0(x).
\end{aligned}$$

Hence, if we can construct consistent learning rules with respect to $\bar{V}_S$, then we can construct consistent learning rules with respect to the divergence function d_S.

The learning rule based on empirical risk minimisation corresponds to maximising the empirical score based on the data $D = \{(x_i, a_i)\}_{i=1}^n$. Hence, $\hat{\sigma}_E$ maximises

$$-\hat{V}^S(\sigma) = \frac{1}{n} \sum_{i=1}^{n} S_{a_i}(\sigma(x_i)), \qquad (16)$$

which can be thought of as the empirical score. For a scoring rule S and stochastic choice

function σ, we define the function $S \circ \sigma(x, a) := S_a(\sigma(x))$. Consistency of the learning rule $\hat{\sigma}_E$ with respect to d_S relies on the nature of the real-valued function class

$$S \circ \Sigma = \{S \circ \sigma | \sigma \in \Sigma\}.$$

The following is a key result.

Proposition 1. *Let Σ be a model of stochastic choice and let S be an incentive compatible scoring rule. Suppose the class of functions $S \circ \Sigma$ is bounded above i.e. there exists an M such that $f(z) \leq M$ for all $f \in S \circ \Sigma$ and $z \in \mathcal{Z}$. If $S \circ \Sigma$ is a Glivenko-Cantelli class, then the model Σ is learnable with respect to the divergence function d_S.*

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 2 and from Lemma 1. □

We now give some examples of incentive compatible scoring rules and their associated divergence functions.

1. (*Log Rule*) : $S_a(p) = \ln(p_a)$. The *log* scoring rule is incentive compatible and its divergence function corresponds to KL divergence $S(p, q) = d_{KL}(p\|q) = \sum_{a \in A} \frac{\ln(p_a)}{\ln(q_a)} p_a$. 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. (*Quadratic Scoring Rule*) : $S_a(p) = 2p_a - \sum_b p_b^2$. 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. $d_S(p, q) = \|p - q\|_2^2$.

   Since all L_p metrics on $\mathbb{R}^d$ are equivalent (as all norms on a finite dimensional vector space are equivalent), it follows that consistency of a learning rule with respect to d_S implies consistency with respect to any other L_p metric. Further, from Pinsker's inequality, we have that

   $$\|p - q\|_{TV} \leq \sqrt{\frac{1}{2} d_{KL}(p\|q)}.$$

   Recall the total variational norm is equal to half times the L_1 distance i.e. $\|p - q\|_{TV} = \frac{1}{2}\|p - q\|_1$. 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 $\succ$ be a complete strict order on A and let $p \in \Delta(A)$. Now, define another strict order $\succ_p$ on A as follows : $a \succ_p b$ if either $p_a > p_b$ or it is the case that $p_a = p_b$ and $a \succ b$. Finally, we let $W(1) \leq W(2) \leq .... \leq W(|A|-1)$.

   The Manski scoring rule with tie breaking is defined as $S_a(p) = W(|\{b|a \succ_p b\}|)$. Note that $a \succ_p b$ implies both $p_a \geq p_b$ and $S_a(p) \geq S_b(p)$. Now, by the rearrangement inequality, this means that that $S(q,q) \geq S(p,q)$ for all $p,q \in \Delta(A)$ i.e. S is an incentive compatible scoring rule.

3.2 Dimension and Sample Complexity

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

For $\gamma \geq 0$, we will say that a set a function class $\mathcal{F}$ V_γ-shatters a set of points $(z_1,...,z_n)$ if there exists a real number $\bar{r} \in \mathbb{R}$ such that for all $B \subseteq \{1,...,n\}$, there exists $f_B \in \{1,...,n\}$ such that

$$f_B(z_i) > \bar{r} + \gamma \text{ for all } i \in B$$
$$f_B(z_i) \leq \bar{r} - \gamma \text{ for all } i \notin B$$

We say a $\mathcal{F}$ V-shatters $(z_1,...,z_n)$ if it V_0-shatters $(z_1,...,z_n)$.

Similary, for $\gamma \geq 0$, we say that a function class $\mathcal{F}$ P_γ-shatters a set of points $(z_1,...,z_n)$ if there exists a vector $r = (r_1,...,r_n) \in \mathbb{R}^n$ such that for all $B \subseteq \{1,...,n\}$, there exists $f_B \in \{1,...,n\}$ such that

$$f_B(z_i) > t_i + \gamma \text{ for all } i \in B$$
$$f_B(z_i) \leq t_i - \gamma \text{ for all } i \notin B$$

We say $\mathcal{F}$ P-shatters $(z_1,...,z_n)$ if it P_0-shatters $(z_1,...,z_n)$.

The following provides definitions for dimension we consider here.

Definition 4. Fix a function class $\mathcal{F}$. The V_γ-dimension of $\mathcal{F}$, denotes the largest number of points n that can be V_γ-shattered by $\mathcal{F}$. The P_γ-dimension of $\mathcal{F}$, denotes the largest

number of points n that can be P_γ-shattered by $\mathcal{F}$. The V-dimension is the V_0-dimension of $\mathcal{F}$ and finally, the P-dimension is the P_0-dimension of $\mathcal{F}$ and is also known as the Pollard dimension.

The V-dimension is due to Vapnik (1989). It is a generalisation of the VC dimension, which is defined for sets i.e. binary valued functions. Indeed, the VC dimension of a class of subsets of $\mathcal{Z}$ equals the V-dimension of the indicator functions corresponding to the sets in the class. The P-dimension is due to Pollard and is commonly referred to as the Pollard dimension. We shall refer to the V-dimension, V_γ-dimension, P-dimension and P_γ-dimension of the function class $\mathcal{F}$ as $\dim_V(\mathcal{F})$, $\dim_{V_\gamma}(\mathcal{F})$, $\dim_P(\mathcal{F})$ and $\dim_{P_\gamma}(\mathcal{F})$ respectively. Note that $\dim_{V_\gamma}(\mathcal{F}) \leq \dim_V(\mathcal{F})$ and $\dim_{P_\gamma}(\mathcal{F}) \leq \dim_P(\mathcal{F})$ for all $\gamma > 0$. The following results are due to Alon et al. [1997] and Bartlett and Long [1995].

Proposition 2. *(Alon et al. [1997] and Bartlett and Long [1995]) Let $\mathcal{F}$ be a uniformly bounded class of real-valued functions defined on $\mathcal{Z}$ i.e., there exists M such that $|f(z)| \leq M$ for all $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. For all $\gamma > 0$, $\dim_{V_\gamma}(\mathcal{F}) \leq \dim_{P_\gamma}(\mathcal{F}) \leq (2\lceil \frac{1}{2\gamma} \rceil - 1)\dim_{V_{\gamma/2}}(\mathcal{F})$. The function class $\mathcal{F}$ is a Glivenko-Cantelli class if and only if $\dim_{V_\gamma}(\mathcal{F})$ is finite for all $\gamma > 0$. Hence, furthermore, a class $\mathcal{F}$ is a Glivenko-Cantelli class if and only if $\dim_{P_\gamma}(\mathcal{F})$ is finite for all $\gamma > 0$.*

Let $(\varepsilon, \delta) \in (0,1)^2$ and suppose for some $0 < \kappa < 1/4$, we have $\dim_{P_{(1/4-\kappa)\varepsilon}}(\mathcal{F}) = d$, which is finite. Suppose now that $n \in \mathbb{N}$ is to the order of

$$O\Big(\frac{1}{\varepsilon^2}\Big(d \ln^2 \frac{1}{\varepsilon} + \ln \frac{1}{\delta}\Big)\Big).$$

Then, for n,

$$\sup_{\mu \in \Delta(\mathcal{Z})} \mu^n\Big(z^n : \sup_{f \in \mathcal{F}} |1/n \sum_{i=1}^{n} f(z_i) - \mathbb{E}_\mu(f)| \geq \varepsilon\Big) < \delta.$$

The above proposition provides us with necessary and sufficient conditions for a function class to be Glivenko-Cantelli. Combining Proposition 1 and Proposition 2, we obtain the following corollary.

Corollary 1. *Let Σ be a model of stochastic choice and let S be a incentive compatible scoring rule. If $S \circ \Sigma$ has finite V_γ-dimension for all $\gamma > 0$, then Σ is learnable with respect to d_S. Furthermore, there exists an almost-ERM rule $\hat{\sigma}_E$ which is consistent with respect to d and Σ*

which has sample complexity at most to the order of

$$O\Big(\frac{1}{\varepsilon^2}\Big(d\ln^2\frac{3}{\varepsilon}+\ln\frac{1}{\delta}\Big)\Big),$$

where $d=\dim_{P_{(1/8)\varepsilon}}(\mathcal{F})$.

4 Applications

4.1 Holder Classes

We establish an upper bound for the V_γ-dimension of a class of Holder continuous functions. A choice probability function $\sigma:\mathcal{X}\to\Delta(A)$ is said to be *Holder continuous* with constants $K,\alpha>0$ if for all $x,y\in\mathcal{X}$,

$$\|\sigma(x)-\sigma(y)\|^\alpha\leq K\|x-y\|.$$

We will obtain a result on the learnability of classes of Holder continuous choice probability maps. Before establishing the result, we will need the following notions of covering and packing numbers. Consider a compact subset A of a metric space (X,d) and a real number $r>0$. We call as $N^c(A,r)$, *r-covering number* of A, to be the smallest number n, of points $x_1,...,x_n\in X$ such that $A\subseteq\cup_i B(x_i,r)$. The *r-packing number* of A, denoted as $N^p(A,r)$, is defined as the largest number n, points $x_1,...,x_n\in A$ such that $\|x_i-x_j\|\geq r$ for $i\neq j$. The covering and packing numbers satisfy the inequality $N^p(A,r)\leq N^c(A,r)$. Note that for compact sets, by definition, there exists a finite r-covering number and hence, from the inequality, also has a finite packing number. We state and prove the learnability result below.

Proposition 3. *Let* $\bar{K},\bar{\alpha}>0$ *and consider the following model*

$$\Sigma_{\bar{K},\bar{\alpha}}:=\{\sigma:\sigma\text{ is Holder continuous with constants }K,\alpha\text{ s.t }K\leq\bar{K},\alpha\leq\bar{\alpha}\}.$$

Further, let S^{br} *denote the Brier scoring rule. For* $0<\gamma<1/2$, *the* V_γ*-dimension of the function class* $S^{br}\circ\Sigma_{\bar{K},\bar{\alpha}}$ *is at most*

$$|A|N^p\Big(\mathcal{X},\frac{(2\gamma)^{\bar{\alpha}}}{8\bar{K}}\Big).$$

Hence, the model $\Sigma_{\bar{K},\bar{\alpha}}$ *is learnable with respect to any* L_p *metric on* $\Delta(A)$.

Proof. We will first argue that for any $f \in S^{br} \circ \Sigma_{\bar{K},\bar{\alpha}}$ and $a \in A$, the function $f(.,a) : \mathcal{X} \to \mathbb{R}$ is Holder continuous with constant K', α' such that $K' \leq 8\bar{K}$ and $\alpha' \leq \bar{\alpha}$. Now, since $f \in S^{br} \circ \Sigma_{\bar{K},\bar{\alpha}}$, there exists $\sigma \in \Sigma$ such that $f(.,a) = S^{br}(\sigma(.),a)$. Now, it is known that the composition $g \circ h$ of two Holder continuous functions g and h with constants K, α and L, β is Holder continuous with constants $K^{\beta}L$ and $\alpha\beta$. In our case, $g = S^{br}(.,a)$ and $h = \sigma(.)$. Hence, it suffices and we will indeed show that for the Brier score function, $S^{br}(p,a)$ is Lipschitz continuous in p with Lipschitz constant 8.

The Brier score $S^{br}(p,a) = 2p_a - \|p\|^2$, is concave as a sum of two concave functions (the square of the Euclidean norm is convex). It then follows that any $L \geq \sup_{p \in \Delta(A)} \|\nabla S(p,a)\|$ is a Lipschitz constant for $S(.,a)$. Now, the gradient is equal to $\nabla S(p,a) = (\nabla_b S(p,a))_{b \in A} = (2 - 2p_a, (-2p_b)_{b \neq a})$. Hence, for any $p \in \Delta(A)$, we have that $\|\nabla S(p,a)\|^2 \leq 4(1-p_a)^2 + \sum_{b \neq a} 4p_b^2 \leq 8$.

We now show the result. Let $0 < \gamma < 1/2$. Let $m > |A| N^p\left(\mathcal{X}, \frac{(2\gamma)^{\bar{\alpha}}}{8\bar{K}}\right)$. Let $\{(x_i, a_i)\}_{i=1}^m$ be a set of data points. Now, by the pigeon-hole principle, there exists $a \in A$ such that $n := |\{i | a_i = a\}| > N^p\left(\mathcal{X}, \frac{(2\gamma)^{\bar{\alpha}}}{8\bar{K}}\right)$.

Consider now the set of data points $D = \{(x_i, a_i)\}_{i \in \{i | a_i = a\}}$. Suppose the set D is V_γ-shattered. Hence, there exists $\bar{r}$ such that for any $i \neq j$, there exists $f \in S^{br} \circ \Sigma_{\bar{K},\bar{\alpha}}$ such that $f(x_i, a_i) > \bar{r} + \gamma$ and $f(x_j, a_j) \leq \bar{r} - \gamma$. Hence, this means $8\bar{K}\|x_i - x_j\| \geq |f(x_i, a_i) - f(x_j, a_j)|^{\alpha'} = |f(x_i, a) - f(x_j, a)|^{\alpha'} \geq (2\gamma)^{\alpha'} \geq (2\gamma)^{\bar{\alpha}}$. But this, in turn, implies that $\|x_i - x_j\| \geq \frac{(2\gamma)^{\bar{\alpha}}}{8\bar{K}}$ for all $i \neq j$. But this contradicts the fact $N^p\left(\mathcal{X}, \frac{(2\gamma)^{\bar{\alpha}}}{8\bar{K}}\right)$ is the $\frac{(2\gamma)^{\bar{\alpha}}}{8\bar{K}}$-packing number for A that now since $|D| = n > N^p\left(\mathcal{X}, \frac{(2\gamma)^{\bar{\alpha}}}{8\bar{K}}\right)$. □

Note that the above result allows us to obtain the minimum number of samples needed to have precise estimate of choice probabilities satisfying a Holder restriction. This is a fairly general class and the sample complexity bound derived from the above result serves as an upper bound on the sample complexity of sub-models which would be of interest in applications. For example, the class of stochastics maps corresponding to several random utility models with independent noise (see, for example, Fosgerau et al. [2017]).

Proposition 3 holds for the Brier score, which gaurantees learnability with respect to any L_p metric. A key aspect of the result is the fact that the Brier score yields a uniformly bounded class of functions $S^{br} \circ \Sigma_{\bar{K},\bar{\alpha}}$. Hence, if instead we use the Log rule $S^{\log}$, the corresponding class of functions, $S^{log} \circ \Sigma_{\bar{K},\bar{\alpha}}$, would not be uniformly bounded. Hence,

we would be unable to derive a counterpart of the above proposition for Log rule as Proposition 2 would not apply. However, if there exists a lower bound for all the choice proabilities in Σ, then we may apply the Log scoring rule to yield consistent estimates with respect to KL divergence. This essentially corresponds to consistency of the conditional maximum likelihood procedure in the present context.

One may also apply the result to construct stastical tests to test the assumptions of a model. For example, consider a sub-model of stochastic choice $\Sigma \subseteq \Sigma_{\bar{K},\bar{\alpha}}$. Given a consistent rule σ for the model $\Sigma_{\bar{K},\bar{\alpha}}$. Then, with enough data z^n, we may accept the model if $d(\sigma(z^n), \Sigma) := \inf_{\sigma' \in \Sigma} d(\sigma(z^n), \sigma')$ is below a threshold value $\bar{\epsilon}$, rejecting the model otherwise. Such tests would yield accurate results for large n, as uniform consistency of the learning rule gaurantees that estimates $\sigma(z^n)$ eventually converge to the true stochastic function σ_0.

4.2 Additive Perturbed Utility Models

The Additive Perturbed Utility model is due to Fudenberg et al. [2015] and is defined as follows. There is a utility function $v : \mathcal{X} \to \mathbb{R}^A$ and a cost function $c : \Delta(A) \to \mathbb{R}$. The stochastic choice function is defined as

$$\sigma_{v,c}(x) \in \arg\max_{p \in \Delta(A)} p.u(x) - c(p).$$

Here, a model Σ corresponds to a class of utility and cost functions v and c. Proposition 3 implies the following learnability result.

Proposition 4. *Let Σ be an Additive Perturbed Utility Model such that for each (v,c), the utility function v and cost function c are both Lipschitz continuous upto constant $\bar{K}$. Then, Σ is learnable.*

Proof. Can be found in the appendix. Follows from a theorem of maximum for Lipschitz functions and an application of Proposition 3. □

4.3 Random Utility Models

We focus on the Random Utility Model of Block et al. [1959] and McFadden [1978], allowing for non-seperability in the utility function. Let $\mathcal{X} \subseteq \mathbb{R}^k$ be a compact set of *characteristics* and a finite set of *alternatives* or *products* $A = \{a_1, a_2, ..., a_m\}$. The utility that an individual derives from a particular alternative depends on the characteristic and also

depends on an *idiosyncratic* variable, which acts as a utility shock. The set of all possible idiosyncratic variables is $\mathcal{E} \subseteq \mathbb{R}^l$. An individual's *utility function* is given by

$$u : \mathcal{X} \times \mathcal{E} \to \mathbb{R}^A. \tag{17}$$

Hence, if an individual's characteristic is $x \in \mathcal{X}$ and the idiosyncratic variable is $\varepsilon \in \mathcal{E}$, then the utility derived from the various alternatives is given by the vector $u(x, \varepsilon) = (u_a(x, \varepsilon))_{a \in A} \in \mathbb{R}^A$. This means that, given (x, ε), the alternative a will be chosen only if

$$u_a(x, \varepsilon) \geq u_b(x, \varepsilon) \text{ for all } b \in A. \tag{18}$$

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 by him i.e (x_i, a_i). Hence, a *data set* is a finite sequence

$$D = \{(x_1, a_1), (x_2, a_2), ..., (x_n, a_n)\}. \tag{19}$$

We now describe the data generating process. There is a joint distribution π over $\mathcal{X} \times \mathcal{E}$. This defines the distribution of (x, ε)'s in the population. The (x_i, ε_i)'s are drawn i.i.d accross i, according to π and the alternative a_i is chosen so as to maximise $u_a(x_i, \varepsilon_i)$, where the utilities are given by an underlying utility function u as in 17. For each i, the analyst only observes (x_i, a_i) but not ε_i. Further, the analyst does not know the distribution π and the utility function u but assumes that it satisfies certain properties.

It shall be useful to work with an alternative but equivalent description of the data generating process. This is as follows. Fix a pair (π, u) and define the following stochastic map $\sigma_{\pi,u} : \mathcal{X} \to \Delta(A)$

$$\sigma_{\pi,u}(x)(a) = \Pr_{\varepsilon \sim \pi(.|x)}[u_a(x, \varepsilon) \geq u_b(x, \varepsilon) \text{ for all } b \in A]. \tag{20}$$

The map $\sigma_{\pi,u}$ tells us the choice probabilities as a function of the characteristic $x \in \mathcal{X}$. Hence, one may alternatively view the data generating process as follows : i) x_i is drawn according to $\pi_{\mathcal{X}}$, which is the marginal of π on $\mathcal{X}$. ii) The choice a_i is made according to probabilities given by $\sigma_{\pi,u}(x)$. Let $Q_{\pi,u}$ denote the resulting distribution over $X \times A$.

A *model* is any family $\mathcal{M}$ of pairs (π, u). We define the choice probabilities corresponding

to a given model $\mathcal{M}$ as

$$\Sigma_{\mathcal{M}} = \{\sigma_{\pi,u} | (\pi, u) \in \mathcal{M}\}.$$

Suppose the model is correctly specified and suppose the true distribution and utility function is (π_0, u_0). Then, the true choice probabilities are given by σ_{π_0,u_0}. For notational convenience, we shall sometimes denote σ_{π_0,u_0} as σ_0 and $\Sigma_{\mathcal{M}}$ as Σ. We say that Σ is learnable if it is learnable in the sense of Definition 1.

4.3.1 Multinomial Logit

In this section, we show that the Logit model is learnable with respect to KL divergence. We derive the sample complexity of the conditional maximum likelihood procedure, which corresponds to the ERM learning rule with the Log scoring rule as the loss functions. This is shown by establishing bounds for the Pollard dimension of the model and then, applying Proposition 2. We describe the model as follows.

The set of characteristics will have two components, namely, individual and product characteristics i.e. $x = (x_I, x_A)$. Let d_I denote the dimensionality of individual characteristic i.e. $x_I \in \mathbb{R}^{d_I}$. For each $a \in A$, d_a denotes the dimensionality of the characteristic of product a. Hence, $x_A = (x_a)_a$ where $x_a \in \mathbb{R}^{d_a}$ for each $a \in A$. Hence, the set of characteristics is a compact subset $\mathcal{X} \subseteq \mathbb{R}^{d_I} \times \prod_{a \in A} \mathbb{R}^{d_a}$. The idiosynctratics variables are $\mathcal{E} = \mathbb{R}^A$, a typical element being $\varepsilon = (\varepsilon_a)_a$. Lastly, there is a set of coefficients for the characteristics $\Theta \subseteq \mathbb{R}^{d_I} \times \prod_{a \in A} \mathbb{R}^{d_a}$. A typical element of Θ will be denoted as $\theta = (\theta_I, (\theta_a)_a) \in \Theta$. We shall assume that Θ is compact as well. Lastly, the utility function $u_a(x, \varepsilon; \theta)$ is parametrised by θ and is given below.

$$u_a(x, \varepsilon; \theta) = \theta_I . x_I + \theta_a . x_a + \varepsilon_a$$

The model is as follows. According to π, the characteristic x is independent of the indiosynctratic variable ε. Moreover, the ε_a's are independently and identically distribution across a, according to a logistic distribution. The resulting choice probabilities for $\theta \in \Theta$ are given as follows. For $a \in A$, the probability that alternative a will be chosen at x is

$$\sigma_a^{logit}(x; \theta) = \frac{e^{\theta_I . x_I + \theta_a . x_a}}{\sum_{b \in A} e^{\theta_I . x_I + \theta_b . x_b}}.$$

Now the model of stochastic choice probabilities is $\Sigma_{logit} = \{\sigma^{logit}(.;\theta)|\theta \in \Theta\}$. We now obtain the following result on the learnability of the Multinomial Logit model.

Proposition 5. *Suppose Σ_{logit} is the Logit model and S^{log} is the Log scoring rule. Let $\bar{d} := \sum_{a\in A} d_I + d_a$. Then, the class of real-valued functions $S^{log} \circ \Sigma_{logit}$ is uniformly bounded and has Pollard P-dimension at most equal to*

$$(2\bar{d} + |A| + 2)^2 \bar{d}(\bar{d} + 19\log_2(9\bar{d})).$$

Hence, the Logit model is learnable with respect to KL divergence, using the conditional maximum likelihood procedure. Moreover, the sample complexity of conditional maximum likelihood has an upper bound that is polynomial in the number of alternatives and the number of regressors.

Proof. Proof can be found in the appendix and follows from a result on the VC dimension of neural networks (see Proposition 8). □

The above model allows for heterogeniety in the coefficients of the product attributes. Moreover, the number of attributes can differ across products as well. Suppose we assume homogeneity and that each product has the same number of attributes (without individual characteristics). The probability of choosing alternative $a \in A$ becomes

$$\sigma_a^{logit}(x;\theta) = \frac{e^{\theta.x_a}}{\sum_{b\in A} e^{\theta.x_b}}.$$

We then obtain the following result.

Corollary 2. *Let Σ'_{logit} be the above Logit model with homogeneity in the coefficients and suppose S^{log} is the Log scoring rule. Then, the class of real-valued functions $S^{log} \circ \Sigma'_{logit}$ is uniformly bounded and has Pollard P-dimension at most equal to*

$$((2d + 1)|A| + 2)^2 d(d + 19\log_2(9d)).$$

Hence, the sample complexity of the Logit model is quadratic in the number of alternatives.

4.3.2 Non-linear utilities

The Logit model corresponds to the case where the utility function of the agent has an additive form and is linear in x. We can also consider non-linear utility forms and esblish

learnability of such models by providing upper bounds for Pollard dimension. Suppose we consider polynomial utilities which are of the form $u_a(x, \epsilon; \theta) = P_a(x; \theta) + \epsilon_a$, where each $P_a(x; \theta)$ is a polynomial in x and the parameter θ is a vector in a Euclidean space, denoting the coefficients of the polynomial. The choice probabilities would be given by

$$\sigma_a^{logit}(x; \theta) = \frac{e^{P_a(x;\theta)}}{\sum_{b \in A} e^{P_b(x;\theta)}}.$$

We obtain the following result establishing the learnability of such models.

Proposition 6. *Let Σ'_{logit} be the above Logit model with polynomial utilities, where each $P_a(x; \theta)$ is a polynomial of order at most m. Suppose S^{log} is the Log scoring rule. Then, the class of real-valued functions $S^{log} \circ \Sigma'_{logit}$ is uniformly bounded and has Pollard P-dimension at most equal to*

$$(m^{k+1} + |A| + 2)^2 m^k (m^k + 19 \log_2(9m^k)).$$

where k is the dimensionality of the characteristic x. Hence, the Logit model with polynomial utilities is learnable.

4.4 Choice from Menus

In this section, we will study stochastic choice in the setting where there is data on choice from a menus of alternatives. As before, A is a finite set of alternatives. Hence, we will have $\mathcal{X} = 2^A \backslash \{\emptyset\}$. For a choice probability function $\sigma : \mathcal{X} \to \Delta(A)$, for each menu of alternatives $x = A \in \mathcal{X}$, the probability vector $\sigma(A)$ which full support in A. For each $a \in A$, the quantity $\sigma_a(A)$ denotes the probability of alternative a being chosen from the menu A. We will assume the following holds for σ.

Definition 5. (*Positivity*) A stochastic choice function σ satisfies positivity if for each menu A and $a \in A$, we have $\sigma_a(A) > 0$.

In this context, the choice probabilities from the Logit or the Luce model are as follows. There exists a function $w : A \to \mathbb{R}_{++}$, which assigns weights to the various alternatives. For each menu A, the alternative $a \in A$ is chosen with probability

$$\sigma_a(A) = \frac{w(a)}{\sum_{b \in A} w(b)}.$$

It is well-known that under positivity, a stochastic choice map σ has the Luce representation if and only if σ 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 $a, b \in A \cap B$. Then,

$$\frac{\sigma_a(A)}{\sigma_b(A)} = \frac{\sigma_a(B)}{\sigma_b(B)}.$$

Another example is the stochastic choice model involving consideration sets (Manzini and Mariotti [2007]). The model is defined as follows. There is a function $\gamma : A \to (0,1)$ and a strict order $\succ$ on A. The choice probabilities are given as follows.

$$\sigma_a^{\gamma,\succ}(A) := (1 - \gamma(a)) \prod_{b:b\succ a} \gamma(b)$$

Proposition 7. *Suppose Σ is either the Logit/Luce model or the model with consideration sets. Then, under the log rule S^{log}, the Pollard dimension of $S^{log} \circ \Sigma$ is at most $O(|A|^2)$.*

Proof. The proof is similar to Proposition 5 and also applies a result from neural networks (see Proposition 8). □

References

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.

Peter L Bartlett and Philip M Long. More theorems about scale-sensitive dimensions and learning. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 392–401. ACM, 1995.

Henry David Block, Jacob Marschak, et al. Random orderings and stochastic theories of response. Technical report, Cowles Foundation for Research in Economics, Yale University, 1959.

Mogens Fosgerau, Emerson Melo, André de Palma, and Matthew Shum. Discrete choice and rational inattention: A general equivalence result. 2017.

Drew Fudenberg, Ryota Iijima, and Tomasz Strzalecki. Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409, 2015.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Paola Manzini and Marco Mariotti. Sequentially rationalizable choice. *American Economic Review*, 97(5):1824–1839, 2007.

Daniel McFadden. Modeling the choice of residential location. *Transportation Research Record*, (673), 1978.

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.

Vladimir Vapnik. *Statistical learning theory. 1998*, volume 3. Wiley, New York, 1998.

5 Appendix

5.1 Proofs from Section 3

The following is the proof for Proposition 1.

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

$$\hat{V}_S(\hat{\sigma}_E(z^n)) \leq \inf_{\sigma \in \Sigma} \hat{V}_S(\sigma) + \varepsilon_n,$$

where recall that for any $\sigma \in \Sigma$, $\hat{V}(\sigma) := 1/n \sum_{i=1}^n S_{a_i}(\sigma(x_i))$. Note that the infimum exists in the above definition since $S \circ \Sigma$ is bounded above by M. Hence, the learning rule $\hat{\sigma}_E$ is well-defined. We now show it is consistent.

Since $S \circ \Sigma$ is a Glivenko-Cantelli class of functions, for each $(\varepsilon, \delta) \in (0,1)^2$, there exists $N'(\varepsilon, \delta)$ such that for all $n \geq N'(\varepsilon, \delta)$, we have

$$\sup_{\mu \in \Delta(\mathcal{Z})} \mu^n \Big(z^n : \sup_{f \in S \circ \Sigma} |1/n \sum_{i=1}^n f(z_i) - \mathbb{E}_\mu(f)| \geq \varepsilon/3 \Big) < \delta. \tag{21}$$

Now, let $N(\varepsilon,\delta) := \max\{3/\varepsilon, N'(\varepsilon,\delta)\}$. Now, suppose $n \geq N(\varepsilon,\delta)$, let $\pi_0 \in \Delta(\mathcal{X})$ be a distribution over characteristics and suppose the true choice probabilities are given by $\sigma_0 \in \Sigma$. For convenience, denote $\mathbb{E}_{z^n}(f) = 1/n \sum_{i=1}^n f(z_i)$.

Now suppose z^n is such that $\sup_{f \in S \circ \Sigma} |\mathbb{E}_{z^n}(f) - \mathbb{E}_{\pi_0 \otimes \sigma_0}(f)| < \varepsilon/3$. Note that this happens with probability at least $1-\delta$ under $\pi_0 \otimes \sigma_0^n$. Consider the almost-ERM rule $\hat{\sigma}_E$. Define also $f_{E,z^n}(x,a) := S(\hat{\sigma}_E(z^n)(x),a)$, $f_0(x,a) := S(\sigma_0(x),a)$ and $\mu_0 = \pi_0 \otimes \sigma_0$. It follows that

$$\begin{aligned}
|\mathbb{E}_{\mu_0}(f_{E,z^n}) - \mathbb{E}_{\mu_0}(f_0)| &= \mathbb{E}_{\mu_0}(f_0) - \mathbb{E}_{\mu_0}(f_{E,z^n}) && (22)\\
&= \mathbb{E}_{\mu_0}(f_0) - \mathbb{E}_{z^n}(f_{E,z^n}) + \mathbb{E}_{z^n}(f_{E,z^n}) - \mathbb{E}_{\mu_0}(f_{E,z^n}) \\
&\leq \mathbb{E}_{\mu_0}(f_0) - \mathbb{E}_{z^n}(f_{E,z^n}) + \varepsilon/3 && (23)\\
&\leq \mathbb{E}_{\mu_0}(f_0) - \mathbb{E}_{z^n}(f_0) + 1/n + \varepsilon/3 && (24)\\
&\leq \varepsilon/3 + \varepsilon/3 + \varepsilon/3 && (25)\\
&= \varepsilon.
\end{aligned}$$

Here, 22 follows since σ_0 minimises expected risk as S is incentive compatible (Lemma 1). The inequality 23 follows by assumption that $\sup_{f \in S \circ \Sigma} |\mathbb{E}_{z^n}(f) - \mathbb{E}_{\mu_0}(f)| < \varepsilon/3$. This yields $\mathbb{E}_{z^n}(f_{E,z^n}) - \mathbb{E}_{\mu_0}(f_{E,z^n}) < \varepsilon/3$. Also, 24 follows since $\hat{\sigma}_E$ is almost-ERM. Lastly, 25 follows since $n \geq 3/\varepsilon$ and again from our assumption that $\sup_{f \in S \circ \Sigma} |\mathbb{E}_{z^n}(f) - \mathbb{E}_{\mu_0}(f)| < \varepsilon/3$, which yields $\mathbb{E}_{z^n}(f_0) - \mathbb{E}_{\mu_0}(f_0) < \varepsilon/3$.

Now, since we have

$$\mathbb{E}_{\mu_0}(f_0) - \mathbb{E}_{\mu_0}(f_{E,z^n}) = \int_{\mathcal{X}} d_S(\hat{\sigma}_E(z^n)(x), \sigma_0(x)) d\pi_0(x),$$

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

5.2 Proofs from Section 4

We now prove Proposition 5. The following result on the VC dimension of neural networks will be useful in our analysis. These can be found in Anthony and Bartlett (see Theorems 8.4 and 8.14).

Proposition 8. *Let $\Theta \subseteq \mathbb{R}^p$. Suppose $\mathcal{F} = \{f(w,\theta) : \theta \in \Theta\}$ is a parametrized class of 0-1 valued functions defined on a set $\mathcal{W} \subseteq \mathbb{R}^k$. Further, suppose, for each $f \in \mathcal{F}$ and each $(w,\theta) \in \Theta$, computing the value of $f(w,\theta)$ takes no more than t many operations from the following*

1. *The arithmetic operations $+, -, \times$ and $\setminus$ defined on real numbers.*

2. *Pairwise comparisons of two real numbers involving the relations* $\geq, >, \leq, <$ *and* $=, \neq$.

3. *Output* 0 *or* 1.

Then, the VC dimension of $\mathcal{F}$ *is at most* $4p(t+2)$.

Proposition 9. *Let* $\Theta \subseteq \mathbb{R}^p$. *Suppose* $\mathcal{F} = \{f(w,\theta) : \theta \in \Theta\}$ *is a parametrized class of 0-1 valued functions defined on a set* $\mathcal{W} \subseteq \mathbb{R}^k$. *Further, suppose, for each* $f \in \mathcal{F}$ *and each* $(w,\theta) \in \Theta$, *computing the value of* $f(w,\theta)$ *takes no more than* t *many operations from the following*

1. *The arithmetic operations* $+, -, \times$ *and* $\setminus$ *defined on real numbers.*

2. *Pairwise comparisons of two real numbers involving the relations* $\geq, >, \leq, <$ *and* $=, \neq$.

3. *Output* 0 *or* 1.

4. *Computing the exponential function* $x \to e^x$.

Then, the VC dimension of $\mathcal{F}$ *is at most* $t^2 d(d + 19\log_2(9d))$.

We now show Proposition 5.

Proof. Firstly, note that the Pollard dimension of $S^{log} \circ \Sigma_{logit}$ is the same as the Pollard dimension of the function class $\{\sigma(x)(a) : \sigma \in \Sigma_{logit}\,; (x,a) \in \mathcal{X} \times A\}$. This follows since $\ln(x)$ is strictly increasing in x. We will show the result by application of Proposition 9 and noting that choice probabilities in the Logit model is parametrised by θ, which is an element of $\Theta \subseteq \mathbb{R}^{\bar{d}}$. Hence, it suffices to compute the number of operations, t, needed to determine whether

$$\frac{e^{\theta_I.x_I + \theta_a.x_a}}{\sum_{b \in A} e^{\theta_I.x_I + \theta_b.x_b}} \geq r_i \tag{26}$$

Note that the computation of each $\theta_I.x_I + \theta_a.x_a = \theta_a^1.x_a^1 + \theta_a^2.x_a^2 + ... + \theta_a^{d_a}.x_a^{d_a} + \theta_a^1.x_a^1 + \theta_a^2.x_a^2 + ... + \theta_a^{d_a}.x_a^{d_a}$ takes $2(d_I + d_A) - 1$ steps to compute. Hence, it takes $2(d_I + d_A)$ steps to compute $e^{\theta_I.x_I + \theta_a.x_a}$. This implies a total number of $\sum_a 2(d_a + d_I) + |A|$ steps to compute $\frac{e^{\theta_I.x_I + \theta_a.x_a}}{\sum_{b \in A} e^{\theta_I.x_I + \theta_b.x_b}}$. This means that the total number of steps needed to determine whether 26 holds is equal to $\sum_a 2(d_a + d_I) + |A| + 2$. This means that $t = 2\bar{d} + |A| + 2$. Finally, applying Proposition 9 and observing that $\Theta \subseteq \mathbb{R}^{d_I + \sum_a d_a}$, the result obtains since $d_I + \sum_a d_a \leq \bar{d}$. □

Now, we prove Corollary 2.

Proof. In the case of the standard Logit model, we have $\Theta \subseteq \mathbb{R}^d$. The argument follows along the same lines as the proof of Proposition 5. □

We next establish the proof for Proposition 6 i.e. the learnability of non-linear models involving polynomial utilities.

Proof. The parameter θ gives us the coefficients of the polynomial. This means that if $n = (n_1, ..., n_k) \in \mathbb{N}^k$ such that $\sum_i n_i \leq m$, then $\theta(n)$ denotes the coefficient of the terms $\prod_{i=1}^k x_i^{n_i}$. Hence, $\theta = \{\theta(n)\}_{n \in \{n : \sum_i n_i \leq m\}}$ is vector in $\mathbb{R}^{m^k}$. We then indeed have that

$$P_a(x;\theta) = \sum_{n \in \{n : \sum_i n_i \leq m\}} \theta(n) \prod_{i=1}^{k} x_i^{n_i}.$$

The number of operations needed to compute $P_a(x;\theta) = \sum_{n \in \{n : \sum_i n_i \leq m\}} \theta(n) \prod_{i=1}^k x_i^{n_i}$ is at most m^k. Applying Proposition 9 again, we obtain the upper bound $(m^{k+1} + |A| + 2)^2 m^k (m^k + 19 \log_2(9m^k))$. □

The following technical lemma, about the Lipschitz continuity of optimal functions, shall be useful in our analysis.

Lemma 2. *Let $W : \mathcal{X} \times \Delta(A) \to \mathbb{R}$ be a function such that i) $W(., p)$ is Lipschitz continuous with constant $K_1 > 0$, for each $p \in \Delta(A)$; ii) $W(x, .)$ is Lipschitz continuous with constant $K_2 > 0$, for each $x \in \mathcal{X}$. Consider an optimal choice probability function $\sigma^* : \mathcal{X} \to \Delta(A)$ defined as*

$$\sigma^*(x) \in \arg\max_{p \in \Delta(A)} W(x, p).$$

Then, it follows that σ^* is lipschitz continuous with constant $K_1 K_2 > 0$.

Proof. Let $x, y \in \mathcal{X}$. Then,

$$\|\sigma^*(x) - \sigma^*(y)\| \leq K_2 (W(x, \sigma^*(x)) - W(x, \sigma^*(y))). \tag{27}$$

This follows from the Lipschitz continuity of $W(x, .)$ and from the fact that $\sigma^*(x)$ is optimal in $\Delta(A)$. Hence, $W(x, \sigma^*(x)) \geq W(x, \sigma^*(y))$. By the same argument, we also have for y that

$$\|\sigma^*(x) - \sigma^*(y)\| \leq K_2 (W(y, \sigma^*(y)) - W(y, \sigma^*(x))). \tag{28}$$

Combining 27 and 28 we get that

$$\begin{aligned}
\|\sigma^*(x) - \sigma^*(y)\| &\leq \frac{K_2}{2}(W(x,\sigma^*(x)) - W(x,\sigma^*(y)) + W(y,\sigma^*(y)) - W(y,\sigma^*(x))) \\
&\leq \frac{K_2}{2}|W(x,\sigma^*(x)) - W(y,\sigma^*(x))| + \frac{K_2}{2}|W(x,\sigma^*(y)) - W(y,\sigma^*(y))| \\
&\leq \frac{K_2}{2}K_1\|x-y\| + \frac{K_2}{2}K_1\|x-y\| \\
&= K_1K_2\|x-y\|.
\end{aligned}$$

□

Combining the conclusions of Lemma 2 and Proposition 3, we have that Proposition 4 follows as a corollary.