Statistical Learnability and Rationality of Choice

Gil Kalai[*]
Institute of Mathematics and Center for Rationality
Hebrew University, Jerusalem, Israel

January 4, 2002

Abstract

In this paper we study the *learnability* of classes of choice functions that arise in theoretical economics using the basic concept of PAC-learnability from statistical learning theory. We prove that the class of rationalizable choice functions on N alternatives is statistically learnable from $O(N)$ examples. We prove that the class of rationalizable choice functions is optimal in terms of PAC-learnability among classes which are invariant under permutations of the elements and examine statistical learnability for more complex classes of choice functions.

1 Introduction

The purpose of this paper is to study the extent to which the concepts of choice used in economic theory imply "learnable" behavior.

In order to analyze learnability we will use a basic model of statistical learning theory introduced by Valiant called the model of PAC-learnability (PAC stands for "probably approximately correct", see Vidyasagar (1997)). Consider a family F of functions from a ground set U to another set Y. We assume that F is known and we want to learn a specific function $f \in F$ from examples. Let ϵ be a small positive real number and let ν be a probability distribution on U.

We say that F is learnable from t examples with probability of at least $1-\epsilon$ with respect to the probability distribution ν if the following assertion holds: For every $f \in F$ if $u_1, u_2, \ldots, u_t$ and u are chosen at random and

[*] I am thankful to Ron Holzman, Leonid Kontorovich, Michael Maschler, Piero La Mura, Bezalel Peleg, Ariel Rubinstein, Ran Spiegler, Naftali Tishby Andrew Thomason and the editor and referees of an earlier version for helpful suggestions.

independently according to the probability distribution ν and if $f' \in F$ satisfies $f'(u_i) = f(u_i), i = 1, 2, \ldots, t$, then $f'(u) = f(u)$ with probability of at least $1 - \epsilon$. We say that F is learnable from t examples with probability of at least $1 - \epsilon$ if this is the case with respect to *every* probability distribution ν on U.[1]

Statistical learnability of classes of functions that arise in an economic model is directly related to the ability to test the model empirically and to "calibrate" the model's parameters based on a small set of test cases. Learnability is related to the testable implications of choice models which have been studied primarily from the point of view of revealed preferences (see, e.g., Brown and Matzkin (1996) and Sprumont (2000)). The number of random examples needed for statistical learnability is closely related to the number of examples needed for testing a model from a small random sample of examples. We will explain this connection in Section 4.

Our analysis of statistical learnability is based on the fundamental combinatorial notion of the P-dimension which is defined in the next section.

Individual choice will be described using choice functions. Given a set X of N alternatives, a *choice function* c is a mapping which assigns to a nonempty subset S of X an element $c(S)$ of S. In decision theory we are primarily concerned with choice functions that are consistent with maximizing behavior. In other words, there is a linear ordering on the alternatives in X and $c(S)$ is the maximum among the elements of S with respect to this ordering. We will refer to such choice functions as *"rationalizable"*. Rationalizable choice functions are characterized by the Independence of Irrelevant Alternatives condition (IIA): the chosen element from a set is also chosen from every subset which contains it.

In Section 3 we prove the following theorem:

Theorem 1.1. *A rationalizable choice function can be statistically learned with high probability from a number of examples which is linear in the number of alternatives.*

In Section 5 we prove that the class of rationalizable choice functions has optimal learnability properties. A class of choice functions is *symmetric* if it is invariant under permutations of the alternatives.

Theorem 1.2. *Every symmetric class of choice functions requires at least a number of examples which is linear in the number of alternatives for learnability in the PAC-model.*

[1] Our definition of the probability ϵ that learning fails is less sensitive than the usual description in the literature (where the failure of learning is given in terms of two probabilities ϵ and δ).

Theorem 1.2 continues to hold even if we consider choice functions defined only on pairs of elements, i.e., asymmetric binary relations. Seeking optimal properties of the class of rationalizable choice functions and the class of order relations continues a line of research initiated by Rubinstein (1996, 2000) who presented similar results for a different notion of learnability. Rubinstein was seeking an explanation from theoretical economics for the more common appearance of order relations (compared to other relations) in natural languages. One of his proposed answers is: "There are forces (evolution or planner) which will favor optimal structures". Statistical learnability carries Rubinstein's argument one step further since it suggests ways to move from "optimality" to "commonly observed". We will discuss this matter in Section 5.

In Section 6 we study learnability of more complex classes of choice functions which arise in economics. Non-rationalizable choice functions may arise in models of bounded rationality, in various cases of strategic choice and when the choice reflects a complicated optimization procedure. They also arise in psychological models of individual choice and in social choice. Our hypothesis is that the choices of individuals as modeled in economic theory are statistically learnable from "a few" examples, namely a number of examples which is at most a polynomial in the number of alternatives. For such classes of choice functions learnability appears to reflect (in a concrete and quantitative way) the *structural* nature of individual choice as modeled in theoretical economics. We will use an example drawn from sports and gambling to demonstrate a general method for analyzing learnability in such a case and show how the argument applies in the analysis of outcomes of the Borda voting procedure.

2 The P-dimension

Our analysis of PAC-learnability will rely on the following fundamental combinatorial concept:[2] Let F be a family of functions from a ground set U to another set Y. The P-dimension of F, denoted by $\dim_P F$, is the maximal number of elements $u_1, u_2, \ldots, u_s$ of U and values $y_1, y_2, \ldots, y_s$ of Y such that for every subset B of $\{1, 2, \ldots, s\}$ there is a function $f_B \in F$ so that $f_B(u_i) = y_i$ if $i \in B$ and $f_B(u_i) \neq y_i$ if $i \notin B$.

Note that it follows from the definition that $u_1, \ldots, u_s$ must be distinct.

[2] We *only* consider the model of PAC-learnability in this paper. We will not use any probabilistic computations because our analysis is based on the connection with the P-dimension.

For example, the class F of linear functions of the form $f(x) = ax + b$ has P-dimension of two. Consider the two elements $u_1 = 0$ and $u_2 = 1$ and the two values $y_1 = 0$, $y_2 = 0$. The four functions 0, x, $1 - x$ and $1 + x$ show that the P-dimension is at least two. On the other hand, since the values of a linear function at two points determine its value at any other point, the P-dimension is not larger than two.

Remark: When Y contains precisely two elements, the P-dimension reduces to the more basic notion of the Vapnik-Chervonenkis (VC) dimension. In this case the P-dimension is simply the size of the largest subset Z of X with the property that every function from Z to Y is the restriction of a function $f \in F$ to Z.

A fundamental result from statistical learning theory states that PAC-learnability is asymptotically determined by the P-dimension.

Theorem 2.1. *For a fixed value of $\epsilon > 0$, the number of examples t needed to learn a class of functions with probability of at least $1 - \epsilon$ is bounded above and below by a linear function of the P-dimension.*

For further details and a description of the (mild) dependence of the required number of examples on ϵ, see Vidyasagar (1997) and Kearns and Vazirani (1994). In view of this theorem, we will henceforth consider the P-dimension to be our principal measure of learnability.

In order to relate the P-dimension to the number of functions in a class of choice functions we also require the following proposition:

Proposition 2.2. *Let F be a family of functions from U to Y. Then,*

$$\dim_{\mathrm{P}}(F) \leq \log_2 |F|.$$

Proof: The proof is obvious since in order for the P-dimension of F to be s we need at least 2^s distinct functions. □[3]

Remark: Let F be a family of functions from U to Y and let $U' \subset U$. Let F' be the family of all the functions from U' to Y which are the restrictions of functions $f \in F$ to U'. It follows immediately from the definitions that

$$\dim_{\mathrm{P}}(F') \leq \dim_{\mathrm{P}}(F).$$

Also, if F can be learned from t examples with probability of at least $1 - \epsilon$ then so can F' (since we can choose ν to be supported only on U').

[3] A reverse relation of a similar form which is less trivial is also known. $\dim_{\mathrm{P}}(F) \geq K \log |F|/n$ for some constant K.

3 Learnability of the class of rationalizable choice functions

We will consider now the class $\mathcal{C}$ of rationalizable choice functions defined on nonempty subsets of a set X of alternatives where $|X| = N$. Since the number of functions in $\mathcal{C}$ is $N!$ it follows from Proposition 2.2 that the P-dimension of this class is at most $\log_2(N!) < N \log_2 N$. Therefore, by Theorem 2.1 the number of examples required to learn a rationalizable choice function in the PAC-model is $O(N \log N)$. This bound can be improved by the following precise result concerning the P-dimension:

Theorem 3.1. *The P-dimension of the class of rationalizable choice functions is $N-1$.*

Proof: We will first show that the P-dimension of the class of rationalizable choice functions is at least $N-1$. To see this, consider the pairs $(a_1, a_2), (a_1, a_3), \ldots, (a_1, a_N)$, and consider the set R of conditions $c(a_1, a_k) = a_1, k = 2, \ldots, N$. It is clear that for every subset S of these conditions, there is a choice function which satisfies the conditions in S and violates the conditions in $R \backslash S$. To see this, simply order the elements so that a_1 is above those a_j's appearing in the pairs in S and below the others.

We will next show that the P-dimension is at most $N-1$. Suppose that the P-dimension of the class of rationalizable choice functions is N or more. Then there are N sets $A_1, A_2, \ldots, A_N$ and elements $a_1 \in A_1$, $a_2 \in A_2, \ldots$, $a_N \in A_N$ such that for every $S \subset \{1, 2, \ldots, N\}$ there exists a rationalizable choice function c such that $c(A_i) = a_i$ if $i \in S$ and $c(A_i) \neq a_i$ if $i \notin S$. In other words, there exists an order relation $<$ on X such that a_i is the maximal element in A_i with respect to the order if and only if $i \in S$.

In order to show that the P-dimension is smaller than N we need to show that for every N sets $A_1, \ldots, A_N$ and N elements $a_i \in A_i$, $i = 1, 2, \ldots, N$ there is a subset S of $\{1, 2, \ldots, N\}$ such that there is no linear order $<$ on the ground set X for which a_i is the maximal element of A_i if and only if $i \in S$. Let $X = \{x_1, x_2, \ldots, x_N\}$. Finding the set S requires the following argument from linear algebra (the remark following the proof may help the reader to visualize it): Let $s_k = |A_k|$. Clearly we can assume that $s_k > 1$ for every k. For every $j = 1, 2, \ldots, N$ consider the following vector:

$$v_j = (v_1^j, v_2^j, \ldots, v_N^j) \in \mathbb{R}^N.$$

The coefficients $v_1^j, \ldots, v_N^j$ are defined as follows:

- $v_k^j = 0$ if $x_k \notin A_j$,

- $v^j_k = s_j - 1$ if $x_k = a_j$ and
- $v^j_k = -1$ if $x_k \in A_j$ and $x_k \neq a_j$.

Note that all the vectors v_j belong to an $N-1$ dimensional space V of $\mathbb{R}^N$ of vectors whose sum of coordinates is 0. Therefore, the vectors $v_1, v_2, \ldots, v_N$ are linearly dependent.

Suppose now that

$$r_1 v_1 + r_2 v_2 + \cdots + r_N v_N = 0 \tag{3.1}$$

and that not all the r_j's equal zero. Let S be the set of j's such that r_j is positive. We will show that there is no rationalizable choice function c such that $c(A_k) = a_k$ when $k \in S$ and $c(A_k) \neq a_k$ when $k \notin S$. Assume to the contrary that there is such a rationalizable choice function c described by a linear ordering $<$ on the set of alternatives. Let $B = \cup\{A_j : r_j \neq 0\}$ and let m be the largest element in B with respect to the ordering $<$. Denote by y the mth coordinate in the linear combination $\sum_{j=1}^N r_j v_j$. We will show that y is positive.

First note that if $r_j = 0$ or if $m \notin A_j$ then the contribution of $r_j \cdot v_j$ to y is zero. Assume now that $r_j \neq 0$ and $m \in A_j$. Note that m must be the largest element in A_j with respect to the ordering $<$ or, in other words, $c(A_j) = m$. There are two cases to consider:

- If $r_j > 0$ then by the definition of S, $j \in S$. By the requirement on c, $c(A_j) = a_j$. However, $c(A_j) = m$ and therefore $m = a_j$ and the contribution of the mth coordinate of $r_j v_j$ to y is $r_j \cdot (s_j - 1)$ which is positive.
- If $r_j < 0$ then by the definition of S, $j \notin S$. By the requirement on c we have $m = c(A_j) \neq a_j$ and the contribution of the mth coordinate of $r_j v_j$ to y is $r_j \cdot (-1)$ which is again positive.

Therefore, y, the m-th coordinate in $\sum_{j=1}^N r_j v_j$, is positive. This is a contradiction. □

Theorem 1.1 follows from Theorems 3.1 and Theorem 2.1.

Remark: A convenient way to visualize the proof is as follows: Regard the elements of X as people. The group of people A_j has a_j as their "leader". The vector v_j stands for the following transaction: each member of A_j pays one dollar to a_j. (Thus, $-2v_j$ represents a payment of two dollars from the leader to each of the others.) The relation $\sum r_j v_j = 0$ implies that after the money has changed hands no one has gained or lost. However, m must gain.

In the linear ordering $<$, $a_j = \max A_j$ if and only if $r_j > 0$. Therefore, for every set A_j that m belongs to, if $r_j > 0$ then m is the leader ($m = a_j$) and hence collects money from the others and if $r_j < 0$, m is not the leader and again he collects money.

There is a special case of Theorem 3.1 that deserves special attention. Consider the class of rationalizable choice functions restricted to subsets of X of cardinality two or, in other words, the class of order relations on X.

Corollary 3.2. *The P-dimension of the class of order relations on N alternatives is $N - 1$.*

Proof: The example which showed that the P-dimension of the class of rationalizable choice functions is at least $N - 1$ consists of pairs of elements and it therefore applies in our case as well. In this case, there is a much simpler proof to show that the P-dimension is at most $N - 1$. Consider N sets A_1, $A_2, \ldots, A_N$, where now $|A_i| = 2$ for every $i = 1, 2, \ldots, N$. Consider a graph G whose vertices are the elements of X and whose edges correspond to the sets $A_1, A_2, \ldots, A_N$. We must prove that we cannot realize by order relations all possible 2^N choice functions on $A_1, A_2, \ldots, A_N$. Every such choice function c corresponds to orienting the edges of G such that A_i is oriented towards $c(A_i)$. Since G has N vertices and N edges, G contains a cycle and therefore not all the orientations of G are realized by order relations. □

Theorem 1.1 can be described as follows: Assume that a consumer has a rationalizable choice function on subsets of a set X of possible alternatives and an observer can watch as the consumer makes choices from random subsets S of alternatives. After having observed the consumer's choices in $K \cdot N$ random examples the observer is able to predict with high probability the consumer's behavior for a large proportion of decision problems. The value of K depends on the probability and the proportion we wish to achieve.

For example, if the examples are drawn uniformly from among the pairs of elements of X then after having observed the order relations among $K \cdot N$ random pairs of elements the order relation is determined with high probability for a large proportion of all pairs. This particular case can be proved directly and it is also easy to prove that $o(N)$ examples will not suffice in this case. The strength of the PAC-learnability model (and the notion of the P-dimension) is that the conclusion reached holds for every probability distribution according to which the examples are drawn. Direct probabilistic arguments for arbitrary distributions appear to be difficult.

The ability to statistically learn rationalizable choice functions (and order relations) from a relatively small number of examples depends on an appropriate notion of statistical learnability. Rubinstein (1996, 2000) briefly considered a naive statistical concept of learnability: The expected number of examples needed to determine the function *precisely* when the examples are drawn uniformly and independently at random. When we adopt this concept, rationalizable choice functions and order relations are especially *difficult* to learn: The choice between the two last elements in order cannot be determined before an example consisting only of those two elements is observed. Therefore, the expected number of examples required to learn a rationalizable choice function in Rubinstein's sense is at least 2^N.

We will now discuss the question whether learning rationalizable choice functions can be achieved *efficiently*. The the fact that a function can be statistically learned from a few examples does not generally imply that there is a polynomial-time algorithm for learning (see Kearns and Vazirani (1994)).[4] However, statistically learning a rationalizable function from $O(N)$ examples can be done by a simple polynomial-time algorithm. The reason is that from an example "$c(A) = x$" we learn that x is larger in order than every other element of A. We can find the transitive closure of all the order relations we obtained from certain examples in polynomial-time and any relation which cannot be derived is still undetermined from the examples.

The bounds for the P-dimension of the class of rationalizable choice functions provide an answer to the following question: "Given that the choice is rationalizable how many examples are needed for it to be (statistically) learned?" The results quoted in the next section will show the relation to another natural question: "How many examples are needed to statistically learn that an observed choice behavior is rationalizable?"

4 Learnability and testability

Statistical learnability and the P-dimension are closely related to the question of finding testable implications of economic models and are directly related to the question of "how much data is needed to test a model". There

[4] Using the notation from the Introduction, we can explain what we mean by efficient learning as follows: We are given the random examples $u_1, u_2, \dots u_t$, the values of $f(u_1), \dots, f(u_t)$ and another random element u. We wish to determine the value of $f(u)$. The value $f(u)$ is uniquely determined with high probability. Therefore, an algorithm for finding this value is to examine the functions in F one by one until a function with the required values for $u_1, \dots, u_t$ is found. This algorithm can be very inefficient: The number of steps can be as large as the number of functions in F. We say that the family is efficiently learnable if there exists an algorithm which requires a number of steps which is polynomial in t.

are several papers (e.g. Brown and Matzkin (1996)) which deal with the testable implications of various concrete models of choice involving prices, utilities, etc. and it would be interesting to examine the statistical learnability of functions that arise in such models. The results presented in this paper are closer to the more abstract models of choice, such as those considered in Sprumont (2000).

Indeed, the P-dimension of a class of functions is related to the number of examples needed to determine with high accuracy how "far" an *arbitrary* function is from the class. Let F be a family of functions from a ground set U to another set Y. Let ϵ, δ be small positive real numbers and let ν be a probability distribution on U. Finally, let g be an arbitrary function from U to Y. Define the distance from g to F, $\mathrm{dist}(g, F)$, to be the minimum probability over all $f \in F$ that $f(x) \neq g(x)$, with respect to ν.

Given t random elements $u_1, u_2, \ldots, u_t$ (drawn independently according to ν), define the *empirical distance* of g from F, $\mathrm{dist}_{\mathbf{emp}}(g, F)$, as the minimum over all $f \in F$ of the quantity $|\{i : f(u_i) \neq g(u_i)\}|/t$.

Theorem 4.1. *There exists $K(\epsilon, \delta)$ such that for every probability distribution ν on U and every function $g : U \rightarrow Y$, the number of independent random examples t needed such that*

$$|\mathrm{dist}(g, F) - \mathrm{dist}_{\mathbf{emp}}(g, F)| < \delta,$$

with probability of at least $1 - \epsilon$, is at most

$$K(\epsilon, \delta) \cdot \dim \mathrm{P}(F).$$

Corollary 4.2. *For every probability distribution ν on U and every function $g : U \rightarrow Y$, if g agrees with a function in F on t independent random examples and*

$$t \geq K(\epsilon, \delta) \cdot \dim_P(F),$$

then

$$\mathrm{dist}(g, F) < \delta$$

with probability of at least $1 - \epsilon$.

Theorem 4.1 and Corollary 4.2 are fundamental results in statistical learning theory which demonstrate the relation between the P-dimension and the testability of a class of functions. A class F of functions from U to Y, has a "testable implication" in the sense used in the literature, if there are functions from U to Y that are not in that class. We are interested in cases where the domain U is very large (for choice functions on

N alternatives, $|U| = 2^N - 1$) and would like to know if testable implications arise from a small random sample. The size of the required sample is at most proportional to the P-dimension. For statistical testability (unlike statistical learnability), the P-dimension only provides an upper bound. It is possible for a class to be "statistically testable" from a few examples while having a large P-dimension.

Returning to the consumer and the observer from the previous section, if after having observed the consumer's choices in $K \cdot N$ random examples the observed choices are rationalizable (or even if only a large proportion of them are) then the observer is able to conclude with high probability that the consumer's behavior agrees with a rationalizable choice for a large proportion of all decision problems. (Again, the value of K depends on the probability and the proportion we wish to achieve.)

5 Optimality of rationality and order

The class of rationalizable choice functions is symmetric under relabeling of the alternatives. Mathematically speaking, every permutation π on X induces a symmetry among all choice functions given by $\pi(c)(S) = \pi^{-1}c(\pi(S))$. A class of choice functions will be called symmetric if it is closed under all permutations of the ground set of alternatives X. All the classes of choice functions considered in this paper are symmetric. We can expect that a model will lead to a symmetric class of choice functions if there is no a priori structure on the set of alternatives.

Our aim in this section is to derive a lower bound for the P-dimension of symmetric families of choice functions. From this point on we will consider choice functions defined on pairs of elements or in other words asymmetric preference relations. Every choice function describes an asymmetric preference relation by restricting it to pairs of elements. Therefore, lower bounds on the P-dimension (or on the number of examples needed for learning in the PAC-model) for symmetric classes of preference relations immediately extend to symmetric classes of choice functions (see the remark at the end of Section 2). Every choice function defined on pairs of elements of X describes a *tournament* whose vertices are the elements of X, such that for two elements x and y in X, if $c(\{x, y\}) = x$ then there is an edge oriented from x to y. (A tournament is a graph which has precisely one oriented edge between every two vertices.)

The following theorem shows that order relations are optimal among symmetric classes of preference relations in terms of the P-dimension:

Theorem 5.1. *(1) The P-dimension of every symmetric class $\mathcal{C}$ of preference relations (considered as choice functions on pairs) on N alternatives is at least $[N/2]$.*

(2) When $N \geq 8$ the P-dimension is at least $N-1$.

(3) When $N \geq 68$, if the P-dimension is precisely $N-1$, then the class is the class of order relations.

Proof: We will give a simple self-contained proof for part (1) which already implies Theorem 1.2 and will use recent results from graph theory to deduce parts (2) and (3).

(1) Let $X = \{x_1, x_2, \ldots, x_N\}$ and $m = [N/2]$. Let $A_1 = \{x_1, x_2\}$, $A_2 = \{x_3, x_4\}, \ldots, A_m = \{x_{2m-1}, x_{2m}\}$. Let c be a choice function in $\mathcal{C}$ and assume without loss of generality that $c(A_1) = x_1$, $c(A_2) = x_3$, $\ldots, c(A_m) = x_{2m-1}$. Let $R \subset \{1, 2, \ldots, m\}$. We wish to find a permutation π_R such that for the choice function $c' = \pi_R(c)$ we have $c(A_k) = x_{2k-1}$ if $k \in R$ and $c(A_k) \neq x_{2k-1}$ if $k \notin R$. Define π_R as follows: If $k \in R$ then $\pi_R(x_{2k-1}) = x_{2k-1}$ and $\pi_R(x_{2k}) = x_{2k}$ and if $k \notin R$ then $\pi_R(x_{2k-1}) = x_{2k}$ and $\pi_R(x_{2k}) = x_{2k-1}$ (if N is odd define $\pi_R(x_N) = x_N$). $\pi = \pi_R$ is the required permutation. To see that this is indeed the case first note that $\pi(A_k) = A_k$ for every k. If $k \in R$, then

$$\pi(c)(A_k) = \pi^{-1}(c(\pi(A_k))) = \pi^{-1}(c(A_k)) = \pi^{-1}(x_{2k-1}) = x_{2k-1} = c(A_k).$$

If $k \notin R$, then

$$\pi(c)(A_k) = \pi^{-1}(c(\pi(A_k))) = \pi^{-1}(c(A_k)) = \pi^{-1}(x_{2k-1}) = x_{2k} \neq c(A_k).$$

(2) Havet and Thomassè (2000) proved a conjecture made by Rosenfeld in 1972 which asserts that, when $N \geq 8$, for every path P on N vertices with an arbitrary orientation of the edges, every tournament on N vertices contains a copy of P. This result implies part (2) of our theorem as follows: Let c be a choice function in the class and consider the tournament T described by c. Let $A_1 = \{x_1, x_2\}$, $A_2 = \{x_2, x_3\}$, $\ldots, A_{N-1} = \{x_{N-1}, x_N\}$. Every choice function c' on A_1, $A_2, \ldots, A_{N-1}$ describes a directed path P. Suppose that a copy of P can be found in our tournament and that the vertices of this copy (in the order they appear on the path) are $x_{i_1}, x_{i_2}, \ldots, x_{i_N}$. Define a permutation π by $\pi(x_j) = x_{i_j}$ The choice function $\pi(c)$ will agree with c' on $A_1, A_2, \ldots A_{N-1}$.

Remark: Thomason(1986) proved Rosenfeld's conjecture when $N > 10^{38}$ and gave a relatively simple argument that every directed path of order N is contained in every tournament of order $N+1$. (This implies that the P-dimension in our theorem is at least $N-2$.)

(3) This part follows from a stronger conjecture by Rosenfeld that every non-transitive tournament on $N > 8$ vertices contains every orientation of a cycle with N vertices. Havet (2000) proved this conjecture for $N > 68$. □

The optimality of the class of rationalizable choice functions and order relations in terms of the P-dimension is deeper than what simply follows from the fact that there are "few" order relations. Let c be an arbitrary choice function. Consider the class of choice functions which are obtained from c by permutations of the elements of X. The P-dimension of this class is at most $\log_2(N!) < N \log_2 N$. It turns out that for "most" choice functions the P-dimension indeed behaves like $N \log N$. This is the case even for preference relations and (as pointed out by Andrew Thomason) can be derived using the arguments above combined with the combinatorial results of Linial, Saks and Sos (1983).

At this point I would like to discuss the connections between the results of this section and those of Rubinstein (1996, 2000, Ch. 1). Rubinstein's intuitively appealing notion of "describability" is defined (in our setting) as follows: Let F be a family of functions from U to Y. F can be *described* by t examples if for every function f in F there are t values $u_1, u_2, \ldots, u_t$ such that if $f' \in F$ and $f(u_i) = f'(u_i), i = 1, \ldots, t$ then $f = f'$. In words, every function in the family is uniquely determined by t examples or less. The order relation $x_1 > x_2 > \cdots > x_N$ can be described by the $N-1$ relations $x_1 > x_2$, $x_2 > x_3$, $\ldots, x_{N-1} > x_N$, and it is easy to see that the class of order relations and the class of rationalizable choice functions need precisely $N-1$ examples to be described.[5]

Rubinstein conjectured that apart from a few small examples every symmetric class of preference relations requires at least $N-1$ examples to be described. He presented a proof that $N - O(n/\log N)$ examples suffice.

Rubinstein proposed the following explanation why order relations are more common in natural languages:

"There are forces (evolution or planner) which make it more likely that structures which are "optimal" with regard to the function of binary relations will be observed in natural languages."

[5] Rubinstein's notion of describability is more intuitive and simple but is less robust than the P-dimension which governs statistical learnability. For example, for the class of all binary relations that disagree with the relation $1 < 2 < 3 < \ldots < N$ in *at most* one place, $N(N-1)/2$ examples are needed for describability. But when you consider those relations which disagree in exactly one place you need just one example (and when you consider those that disagree in exactly five places you need five examples). In contrast, the value of the P-dimension can change by at most one when a single function is added to a family.

This is a far-reaching hypothesis. It provides an interpretation of Rubinstein's conjecture mentioned above concerning the optimality of order relations in terms of the number of examples needed for their description and a motivation for studying other similar results.

Using statistical learnability rather than a notion of describability[6] may hint at possible mechanisms for moving from optimality to "commonly observed" since the learning of relations from random examples seems closer to the way language is learned: When a child learn the relation "to be bigger" among N words (representing physical objects), it is not that the $N-1$ pairs of elements a and b where a is just barely bigger than b are described to the child and the entire order relation is deduced by transitivity. Learning the words themselves and the relations of the form "a is bigger than b" in some random manner seems more realistic.

It is worth noting that a very simple algorithm is sufficient to obtain good results for statistical learning with respect to the uniform distribution on pairs. The algorithm is as follows: In choosing between two alternatives a and b consider the last K examples in which either a or b was chosen and choose the one that was picked more often. (Take K to be odd.) The algorithm will produce mistakes; however, the probability of a mistake diminishes as K is increased. For $K = 1$ this algorithm is a naive form of mimicking and it already yields a substantial gain over random guesses. The ability to mimic appears to be relevant in the process of learning a language.

Statistical learnability (especially Corollary 4.2) suggests another "force" that may explain why optimal structures are more commonly observed. More commonly observed relations may represent not only an objective fact about the language but also their greater capacity for being recognized by an observer.

Finally, let us briefly return to choice functions in general. Theorem 5.1 implies the following Corollary:

Corollary 5.2. *The P-dimension of every symmetric class of choice functions on N alternatives, $N \geq 8$, is at least $N-1$.*

We can expect that there will be a simpler proof in this case (for every $N \geq 2$,) but I was unable to find it. When we consider statistical learnability as measured by the P-dimension, the class of choice functions is optimal compared to other classes. It is worth noting that this is not the case for

[6] The question whether natural languages are better learned by "rules" or by statistical methods is a central issue in machine learning.

describability in Rubinstein's sense. Consider the class of choice functions which represent choosing the *median* element according to some fixed order relation: In other words, let $c(S)$ be the median element among S according to some fixed ordering of the elements in X. (For simplicity assume that $|S|$ is odd.) This class of choice functions was considered in Kalai, Rubinstein and Spiegler (2001). Yuval Salant pointed out that the median choice requires at most $2n/3$ examples (but not less than $n/2$ examples) to be described. Moreover, there are symmetric classes of choice functions on N alternatives that can be described by $\log_2 N + 1$ examples, much less than the $N - 1$ examples required for the class of rationalizable choice functions.

We can also seek an interpretation of the optimality of the class of rationalizable choice functions. Are there forces which will lead to optimality when it comes to individual choice? The discussion in Börgers (1996) appears relevant. Again, it is important to remember that we are considering classes of choice functions which are symmetric, namely we assume that the alternatives are a priori indistinguishable.

Compare, for example, the following two types of behaviors: The choice of one agent is rationalizable while for another agent, $c(S)$ is the median element among S according to some fixed ordering of the elements in X. We can ask the following: Given that we know the choice pattern but not the specific ordering, is it the case that it will be easier to learn the choices of the first agent based on a small number of random examples? Will it be easier for an economist or a psychologist by observing a small number of random examples to support or refute that an agent's choice is rationalizable than to support or refute a "median" behavior? Suppose that the type of behavior is "wired in" but the ordering is unknown and the choices of a novice agent are based on mimicking the behavior of experienced agents. Does rationalizable choice have an advantage over "median" choice?

Our objective is not to offer definitive answers. Such answers would depend on specific models. For the last (and perhaps most interesting) question we must take into account how "successful" each type of behavior is to start with and not only the learning factor. The ability of novice agents to quickly learn the choice may occasionally give an advantage to a rival... Our purpose is rather to propose PAC-learnability and the P-dimension as appropriate mathematical concepts for studying such questions.

6 Analyzing the learnability of more complex choice

6.1 Examples and motivation

There are various cases in which individual choices are modeled by choice functions which are not rationalizable. See Kalai, Rubinstein and Spiegler (2001) for several examples and for a more general notion of rationalization.

Consider the following three examples:

1. The decision maker chooses the alternative from a set S of alternatives which is 'second best' according to her utility function. (This example is discussed in Sen (1993).)

2. In order to choose one person to hire from a group of candidates, an exam is given and the decision is made between the two candidates with the highest scores. (Formally, there are two order relations, $<_1$ and $<_2$, on the alternatives. $c(A)$ is the maximal element according to $<_1$ between the two largest elements according to $<_2$.)

3. A committee of three people chooses between pairs of alternatives by majority vote.

In these examples upper bounds on the P-dimension can easily be derived by elementary counting. A choice functions described by the first example is determined by an order relation on the alternatives. The class of choice functions is of cardinality $N!$ and therefore (by Proposition 2.2) the P-dimension is at most $N \log_2 N$ and by Theorem 2.1 the number of examples needed for learning in the PAC-model is $O(N \log N)$. Choice functions in the second and third examples are determined by two- and three- order relations, respectively. The cardinalities of the classes of choice functions that arise are at most $N!^2$ and $N!^3$ and again the number of examples needed for learning in the PAC-model is at most proportional to $N \log N$.

Note that learnability reflects the structural nature of these examples and may fail miserably for simple unstructured extensions. The first two examples might suggest that we consider the class $\mathcal{U}_2$ of all choice functions with the property that there exists a linear ordering for which $c(A)$ is either the first or the second element of A. It is easy to see from the freedom of choice for subsets of cardinality larger than one that the P-dimension of $\mathcal{U}_2$ is already $2^N - N - 1$ and therefore an exponential number of examples is needed to learn a choice function in $\mathcal{U}_2$ according to the PAC-model.

Remark: Sprumont (2000) considered abstract choice functions representing the choices of several interacting agents and, in particular, choices that can be rationalized by the notion of Nash equilibrium. A simple counting argument (combined, of course, with Theorem 2.1) implies statistical learnability from a "few" examples for the class of choice functions which are Nash-rationalizable in Sprumont's sense. The method described below allows us to analyze learnability of "mixed Nash-rationalizable" choices as also defined by Sprumont.

Equipped with the examples above we are ready to discuss in some detail the original motivation behind the results of this section. The notion of a rationalizable choice function consists of a substantial abstraction which allows various choice procedures and models to be grouped together. For many economic applications, once we know the choice is rationalizable the specific model for deriving the utility functions, as complex as it may be, becomes irrelevant. Can a similar level of abstraction be reached for classes of choice functions that are not rationalizable? or for choices of two interacting agents? A natural extension would be to find more general forms of rationalization or to consider weaker forms of the IIA axiom which allow the inclusion of additional choice functions.

The first approach was adopted in Kalai, Rubinstein and Spiegler (2001) where choice functions that can be rationalized by multiple rationales were considered. For example, there are various interesting classes of choice functions that can be rationalized by two order relations, i.e., there are two order relations $<_1$ and $<_2$ such that for every S, $c(S)$ is either the maximal element of S according to $<_1$ or according to $<_2$. Note that the class of choice functions that can be rationalized by two order relations is already highly non-learnable. This means that along with interesting examples, a vast number of additional choice functions is introduced. Finding axiomatic descriptions for classes of choice functions which are not rationalizable also appears to be difficult (for example, I am not aware of any satisfactory axiomatic description for any of the classes of choice functions considered in this section,) and simple weakening of the IIA condition often lead to highly non-learnable classes of choice functions.

We examine this issue from a different angle. We propose that the structural nature of individual choice as modeled in theoretical economics implies statistical learnability from "a few" examples, i.e., a number of examples which is at most a polynomial in the number of alternatives. The notions of PAC-learnability and the P-dimension can serve as concrete ways to define a class of functions as structural and to measure "how structural" the class is. In the rest of this section we describe a method to analyze statistical learnability when the choice is based on a complicated optimization procedure.

6.2 Who is the most likely winner?

The following example will be used to present a general method for analyzing learnability. Consider a situation with N tennis players such that for each two players i and j there is a probability p_{ij} that i beats j in a match between the two (and therefore, $p_{ji} = (1 - p_{ij})$). Among a set A of players, let $c(A)$ be the player who is most likely to win in a tournament involving the players in A. In this tournament, there will be a match between every pair of players and the player with the largest number of victories is the winner.[7] In the case of a tie no winner is declared.

Consider the class $\mathcal{W}$ of choice functions that arise in this model where the probabilities p_{ij} vary over all real numbers in the interval $[0, 1]$. The first thing to notice is that non-rationalizable choice functions can result in this situation. In fact, the choices for pairs of players can be prescribed in an arbitrary manner. In this example, the choice from a set A depends on a complicated computation based on $N(N-1)/2$ real parameters. Is this description sufficiently "structured" to imply learnability from a few examples? Does the model have any testable implications or perhaps the choice of every set can be prescribed in advance?

Theorem 6.1. *The class of choice functions $\mathcal{W}$ requires $O(N^3)$ examples for learning in the PAC-model.*

The proof relies on results from real algebraic geometry which were first applied in the context of learnability by Goldberg and Jerrum (1995). Note that to obtain crude upper bounds for learnability *all* that is needed is to assert that the restrictions on the class of choice functions in question are strong enough so that only a few choice functions are left and therefore it is easy to learn the implied class. Moving from the mathematical model to the conclusion that "a few choice functions are left" can be far from obvious.

6.3 Sign patterns of real polynomials

When the model for the choice functions is based on optimization involving certain real parameters, simple counting arguments may not apply. There are various methods that can be used for proving that the number of choice functions will nevertheless be small if the real functions involved in their definition are not overly complex. Perhaps the most general one is the

[7] The (smaller) class of choice functions that represent the special case in which all p_{ij} are either zero or one is considered in psychology, see Tversky and Shafir (1992).

application of results from real algebraic geometry. See Alon (1995) for a survey of this approach and references.

We will rely on the following theorem by Warren which is tailor-made for combinatorial applications. Consider m polynomials $Q_1(x_1, \ldots, x_r)$, $Q_2(x_1, \ldots, x_r)$, ..., $Q_m(x_1, \ldots, x_r)$, in r variables $x_1, x_2, \ldots, x_r$. For a point c in $\mathbb{R}^r$ the sign pattern $(s_1, s_2, \ldots, s_m)$ is a vector in $\{-1, 0, 1\}^m$ where $s_j = signQ_j(c)$, namely $s_j = 1$, if $Q_j(c) > 0$, $s_j = -1$ if $Q_j(c) < 0$ and $s_j = 0$ if $Q_j(c) = 0$.

Theorem 6.2. *If the degree of every Q_j is at most D and if $2m > r$ then the number of sign patterns given by the polynomials $Q_1, \ldots, Q_m$ is at most $(8eDm/r)^r$.*

We now show how Warren's theorem applies to the example considered above. The argument extends to various classes of choice functions (and other types of functions found in economics) which may be based on complicated optimization procedures.

First we show the precise computation for determining the most likely winner. Given a set A of s players, the probability that the k-th player will be the winner in a tournament between the players of A is described by a polynomial $Q(A, k)$ in the variables p_{ij} as follows:

First, we represent the outcome of all matches between the players of A using an s by s matrix $M = (m_{ij})$ such that $m_{ij} = 1$ if player i won the match against player j and $m_{ij} = 0$ otherwise. We put $m_{ii} = 0$ for every i. The number of such matrices is the number of possible outcomes which equals $2^{s(s-1)/2}$. The probability p_M that such a matrix M will represent the results of matches in a tournament is

$$p_M = \prod \{p_{ij} : i, j \in A, \quad m_{ij} = 1\}.$$

Define $Q(A, k)$ as the sum of all the expressions p_M for which player k is the winner. (Player k is the winner in A if when we restrict the matrix M to rows and columns that correspond to A, the row that corresponds to k has more 'ones' than any other row.) $Q(A, k)$ is the probability that player k will be the winner in a tournament involving the players in A. $c(A)$ is the player in A for which $Q(A, k)$ is maximal.[8]

[8] We can characterize choice functions in $\mathcal{W}$ by a sentence of the form: "There exist real probabilities p_{ij} such that $c(A)$ is the element k of A for which $Q(A, k)$ is maximal". In a sense, we would like to eliminate the quantifier "there exist p_{ij}". The issue and the mathematical tool we apply are to some extent close in spirit to those considered by Brown and Matzkin (1996).

Proof of Theorem 6.1: Given a set A of s players we described above the probability $Q(A,k)$ that the k-th player will be the winner in a tournament between the players of A. $Q(A,k)$ is a polynomial of degree s in the variables p_{ij}, $i,j \in A$.

Now consider all the polynomials of the form $Q(A,k,j) = Q(A,k) - Q(A,j)$ for all nonempty subsets A of players and all pairs of distinct players k and j in A. We have altogether less than $2^N \cdot N^2$ polynomials in $N(N-1)/2$ variables p_{ij}. (Note that $p_{ij} = 1 - p_{ji}$.) The degree of these polynomials is at most $N(N-1)/2$.

The crucial observation is that the choice function given by a vector of probabilities p_{ij} is *determined* by the sign pattern of all the polynomials $Q(A,k,j)$. Indeed, $c(A) = k$ (that is, k is the most likely winner in a tournament between the players in A) if and only if $Q(A,k,j)$ is positive for every $j \in A$, $j \neq k$.

We can now invoke Warren's theorem with $r = D = \binom{N}{2}$ and $m \leq N^2 \cdot 2^N$. According to Warren's theorem the number of different sign patterns of the polynomials $Q(A,k,j)$ is at most $(e2^N N^2)^{N(N-1)/2}$. The logarithm of the number of choice functions described in this fashion is therefore smaller than N^3. □

6.4 The Borda rule

Let X be a set of politicians and consider a society in which each individual has an order relation over X. Consider a subset $A \subset X$ of candidates that are running for office. According to the plurality rule $c(A)$ is the candidate who placed first among the elements of A for the largest number of voters. According to the Borda rule $c(A)$ is chosen as follows: For each individual, rank the elements of A by the numbers $0, 1, \ldots, |A|-1$ (the candidate ranked '0' is the least favorable) and let $c(A)$ be the element for which the sum of the individual ranks is maximal.

Both rules lead to non-rationalizable classes of choice functions but the learnability of these classes differs sharply. Saari (1989) proved that the plurality rule for large societies gives rise to *all* choice functions. In contrast we present the following theorem:

Theorem 6.3. *The class $\mathcal{B}$ is learnable in the PAC-model from $O(N^3)$ examples.*

Proof of Theorem 6.3: For two alternatives $i, j \in X$ let b_{ij} be the number of voters that prefer i to j. For a subset A of X and for $i \in A$ define $Q(A,i) = \sum\{b_{ij} : j \in A\}$. $Q(A,i)$ is precisely the sum of rankings of candidate i in an election in which the members of A are the candidates. Therefore, $c(A)$ is the element i of A for which $Q(A, i)$ is maximal. The Borda choice for A is therefore expressed by the sign patterns of the (linear) polynomials $Q(A, i) - Q(A, j)$ in the variables b_{ij}. From this point the proof proceeds as in Theorem 6.1. (The number of sign patterns of n linear expressions in m variables is the number of regions determined by n hyperplanes in $\mathbb{R}^m$. Therefore, in this case, simpler tools than Warren's theorem are available.) □

Theorem 6.3 adds to a large literature concerning the advantage of the Borda rule (see Saari (1995) and references cited there). It is well-known (and perhaps was already known to Borda himself) that the choices for the Borda rule cannot be prescribed in an arbitrary manner: For example, if $x = c(A)$ then x cannot lose in a two-candidate election against every other element of A (see, Saari (1995)). Theorem 6.3 gives a quantitative upper bound to the amount of freedom in the outcomes of the Borda rule.

What accounts for the difference in learnability between the Borda rule and the plurality rule? According to the plurality rule the society's choice for a set A depends only on the individual choices for A. According to the Borda rule the society's choice for a set A also depends on the individual preferences for the elements of A. We say that a social choice function satisfies the "Irrelevance of Rejected Alternatives" (IRA) condition if the choice of the society depends only on the choices of the individuals.

In a subsequent paper we prove that under fairly general conditions and when the society is large, if (IRA) is assumed then the class of choice functions that arises includes all choice functions and hence requires 2^N examples for statistical learning with high probability.

7 Conclusion

This paper is centered around the mathematical concept of statistical learnability and especially around the P-dimension, a combinatorial concept used to analyze statistical learnability. We have chosen to study this concept in the context of abstract choice theory. A similar study for economic models of prices, demands and utilities would be an interesting direction for further research.

The main economic justification for this study is that statistical learnability is related to the ability to make predictions based on empirical data and to empirically test economic models.

Our main result determines the P-dimension of the class of rationalizable choice functions. A complementary result demonstrates the optimality of the class of rationalizable choice functions in terms of the P-dimension under the assumption of symmetry. This result is analogous to a conjecture of Rubinstein (1996, 2000) in a context which arguably is more suitable to Rubinstein's interpretation. Finally, we describe a mathematical method for analyzing the statistical learnability of complicated choice models and demonstrate this method for the outcomes of the Borda voting rule and for an example in which the choice is based on involved optimization.

References

[1] Alon, N. (1995) Tools from higher algebra, in *Handbook In Combinatorics* (Graham et als., eds.), Vol. II , 1749-1783, Elsevier, Amsterdam.

[2] Börgers T. (1966), On the Relevance of Evolution and Learning to Economic Theory, Economic Journal 106, 1274 - 1285.

[3] Brown D. and R. Matzkin (1996), Testable restrictions on the equilibrium manifold, Econometrica 64, 1249-1262.

[4] Goldberg P. W. and and M. R. Jerrum (1995), Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parameterized by Real Numbers, Machine Learning, 18, pages 131–148.

[5] Havet, F (2000), Oriented Hamiltonian cycles in tournaments. J. Combin. Theory Ser. B 80, 1–31.

[6] Havet, F and Thomassè, S. (2000), Oriented Hamiltonian paths in tournaments: a proof of Rosenfeld's conjecture. J. Combin. Theory Ser. B 78, 243–273.

[7] Kalai, G., A. Rubinstein and R. Spiegler (2001), Rationalizing choice functions by multiple rationales, preprint, 2001.

[8] Kearns M. and Vazirani U (1994), *An introduction to computational learning theory,* MIT Press, Cambridge, MA.

[9] Linial, N., M. Saks and V. Sos(1983), Largest digraphs contained in all n-tournaments, Combinatorica 3, 101-104.

[10] Rubinstein, A. (1996), Why are certain properties of binary relations relatively more common in natural language?", Econometrica, 64, 343-356

[11] Rubinstein, A. (2000), *Economics and Language*, Cambridge University Press, Cambridge.

[12] Saari, D., G. (1989), A dictionary of voting paradoxes, J. Economic Theory 48. 443-475.

[13] Saari, D. G. (1995), A chaotic exploration of aggregation paradoxes, SIAM Review 37, 37-52.

[14] Sen, A. (1993) Internal consistency of Choice, Econometrica, 61, 495-521.

[15] Sprumont, Y. (2000) On the testable implications of collective choice theories, J. Econom. Theory 93, 205–232.

[16] Thomason, A. (1986), Paths and cycles in tournaments, Trans. Amer. Math. Soc. 296, 167–180.

[17] Tversky, A., and E. Shafir (1992), Choice under conflict: The dynamics of deferred decision, Psychological Science, 3, 358-361.

[18] Vidyasagar M. (1997), A theory of learning and generalization. With applications to neural networks and control systems. Springer-Verlag London, Ltd., London.