

ROBUST EMPIRICAL RISK MINIMIZATION ON G-SETS

MARCIN PEŃSKI*

ABSTRACT. Vapnik and Chervonenkis' statistical learning theory finds simple necessary and sufficient conditions for consistent empirical risk minimization that is robust to any misspecification of the world (i.e. to the distribution of outcomes). In some learning problems, a statistician has an additional information about the structure of the world that is not accounted for by Vapnik-Chervonenkis theory. In those problems, the additional information restricts the set of worlds that are plausible and makes Vapnik-Chervonenkis conditions no longer necessary. This paper presents novel, simple, necessary and sufficient conditions for consistent empirical risk minimization that is robust to some worlds. As an application, I demonstrate that the various classes of predictors proposed to solve the Netflix recommendation problem, like clustering or factor models satisfy these conditions.

1. INTRODUCTION

A statistician wants to learn an unknown functional relationship between *instances* $\mathbf{x} \in \mathbf{X}$ (explanatory variables, inputs, decision problems) and *outcomes* $y \in Y$ (dependent variables, outputs, solutions). Empirical risk minimization (henceforth, ERM) is a simple induction principle to approach the learning problem. The statistician starts with a family of plausible models \mathcal{M} . He uses the sample data to choose a model $\theta \in \mathcal{M}$ that has the best fit in the sample data, i.e. minimizes the empirical risk. Statistical learning theory investigates when ERM is consistent, i.e. when the best fit in the sample implies, with a high probability, the good fit outside the sample.

Many classical estimation procedures can be represented as ERM. For example, ordinary least squares is an ERM applied to models that are linear in \mathbf{x} and the loss function is quadratic; nonlinear least squares enlarges the class of models but keeps the square loss function; maximum likelihood can be represented as ERM under the log-likelihood loss function; nonparametric regression or density estimations can be approached as

Key words and phrases. Statistical learning theory, statistical decision theory, empirical risk minimization.

VERY PRELIMINARY AND INCOMPLETE. *University of Chicago, Department of Economics. E-mail: mpeski@uchicago.edu. All remaining errors are my own.

ERM on families that are Lipschitz, have bounded derivatives, or satisfy some other assumptions.

Intuitively, the question of consistency of ERM is related to the size of family \mathcal{M} . If \mathcal{M} is too large, the statistician risks that the empirical minimizer will overfit. V. Vapnik and A. Y. Chervonenkis provide remarkably simple set of conditions that are necessary and sufficient for ERM to be consistent robustly to *all* worlds (i.e., distributions over instances and outcomes). The idea is to associate family \mathcal{M} with a certain natural number, called VC-dimension. This number depends only on combinatorial properties of family \mathcal{M} and, in particular, it extends the classical notion of dimension from linear models. It turns out that ERM is robustly consistent if and only if the VC-dimension of \mathcal{M} is finite. This allows to replace potentially difficult statistical problem, by arguably much easier, combinatorial one. Not surprisingly, families of models that are used in most applications have finite VC-dimension and the proof of this fact, in many cases, is almost trivial.¹

Robustly consistent ERM allows to perform the statistical analysis without any prior knowledge about the world. Nevertheless, in many situations, the statistician has additional information. Quite often, this information is a logical consequence of the way that the learning problem is posed and does not involve making any assumptions about the relationship between instances and outcomes. But, the additional knowledge may relax the learning problem and may cause Vapnik-Chervonenkis conditions to be no longer necessary.

The goal of this paper is to present novel, simple sufficient and necessary conditions for consistent ERM that is robust to *some* worlds. For this purpose, I introduce an extension of VC-dimension that is applicable in a wide class of learning problems. As an application, I check that some well-known families of models, like clustering or factor models, satisfy these conditions. This leads to a simple proof of the robust consistency of ERM in these cases.

In the rest of the Introduction, I sketch the statistical learning theory of Vapnik-Chervonenkis and illustrate the main results of the current paper on a particular application to recommendation problems. Section 2 presents the model and main assumptions. One of the contributions of this paper is to show that various learning problems can be analyzed by the same methods. Numerous examples of such learning problems are provided in Section 3. Section 4 contains the sufficient and necessary conditions for robustly consistent ERM when Y is binary. This is further extended to Y compact in

¹These includes linear or polynomials of bounded order when space \mathbf{X} is an euclidean space, neural networks, and others (see (Vapnik 1998), (Devroye, Györfi, and Lugosi 1996)).

Section 5. Section 6 contains some auxiliary results. Section 7 use the sufficient conditions of the main result to verify consistency of ERM in particular applications. Section ?? contains discussion of some additional issues. Section 8 and the Appendix contain proofs.

1.1. Statistical learning theory of Vapnik-Chervonenkis. To state the problem formally, let $X \subseteq \mathbf{X}$ be a finite subset of instances and let $\omega : X \rightarrow Y$ be the true relationship between instances and outcomes. A pair (X, ω) is called *world* and $|X| < \infty$ is the size of the world. A statistician attempts to describe the unknown relationship between instances and outcomes with a *model* $\theta : \mathbf{X} \rightarrow Y$. One measures the *risk* of model θ in world (X, ω) as

$$R_{(X, \omega)}(\theta) = \frac{1}{|X|} \sum_{\mathbf{x} \in X} l(\theta(\mathbf{x}), \omega(\mathbf{x})), \tag{1.1}$$

where $l(y, y') \in [0, 1]$ is the *loss* from predicting y when the correct outcome is equal to y' . Denote the universe of all possible worlds with Σ and let Σ_n be the universe of all worlds of size at least n .

The statistician has a family of plausible models $\mathcal{M} \subseteq Y^{\mathbf{X}}$. Let $X_\gamma \subseteq X$ be a sample of $\gamma|X|$ instances chosen randomly from X and let $R_{(X_\gamma, \omega)}$ be the *empirical risk* of the model θ defined on sample (X_γ, ω) . Empirical risk minimization selects the model with the lowest empirical risk:

$$\theta_{(X_\gamma, \omega)} \in \arg \min_{\theta \in \mathcal{M}} R_{(X_\gamma, \omega)}(\theta).$$

Say that ERM is *robustly consistent* if given sufficiently large worlds, with high probability, the true risk of model $\theta_{(X_\gamma, \omega)}$ is close to the true minimal risk: for any $\varepsilon, \gamma > 0$,

$$\limsup_n \sup_{(X, \omega) \in \Sigma_n} P_{X_\gamma} \left(R_{(X, \omega)}(\theta_{(X_\gamma, \omega)}) > \min_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) + \varepsilon \right) = 0. \tag{1.2}$$

where the probability P_{X_γ} is taken with respect to the sampling. Robust consistency is a strong but a natural requirement: If it holds, ERM can be applied without making any assumptions about the world.

I assume here that the world is unknown, but deterministic. This is without loss of generality. If the world is stochastically chosen from certain distribution, then (1.2) guarantees that ERM is consistent for *any realization* in this distribution.

It is easy to show that ERM is robust and consistent if family of models \mathcal{M} is finite. The interesting question is what happens when \mathcal{M} is infinite. The definite answer is given by the statistical learning theory of V. Vapnik and A. Y. Chervonenkis ((Vapnik and Chervonenkis 1971), (Vapnik 1998), (Bousquet, Boucheron, and Lugosi 2004),

(Boucheron, Bousquet, and Lugosi 2005)). Suppose for simplicity that $Y = \{0, 1\}$. For any finite $S \subseteq \mathbf{X}$, let $\mathcal{M}_S \subseteq \{0, 1\}^S$ be the set of model restrictions to set S :

$$\mathcal{M}_S = \{\theta|_S : \text{there is } \theta \in \mathcal{M}\}. \quad (1.3)$$

It is clear that $|\mathcal{M}_S| \leq 2^{|S|}$. Say that \mathcal{M} has a *VC-dimension* equal to k , write $\dim_{VC} \mathcal{M} = k$, if k is the smallest number, such that

$$\sup_{S \subseteq \mathbf{X}, |S| \leq k} |\mathcal{M}_S| < 2^k. \quad (1.4)$$

If such a k does not exist, say that family \mathcal{M} has an infinite VC-dimension.

(Vapnik and Chervonenkis 1971) shows that ERM is robustly consistent *if and only if* \mathcal{M} has a finite VC-dimension.² The necessary part of the result is almost trivial: if VC-dimension of family \mathcal{M} is infinite, then there is a world for which the empirical minimizer will certainly overfit the sample, i.e. it will have a perfect match to the sample observations. The real difficulty is in proving sufficiency.³

Next, I describe a Netflix recommendation problem. I show that a very natural family of models does not have a finite VC-dimension, hence does not satisfy the necessary part of the Vapnik-Chervonenkis theorem. Nevertheless, this family of models can be successfully used in a consistent ERM that is robust to some worlds.

1.2. Example: Netflix recommendation problem. Netflix is an Internet-based DVD rental company. In one of its services, it collects information on movie ratings

²Vapnik-Chervonenkis result is stronger than stated above. Precisely, if \mathcal{M} has finite VC-dimension, then, for any $\varepsilon > 0$,

$$\limsup_n \sup_{(X, \omega) \in \Sigma} P_{X_n} \left(R_{(X_n, \omega)}(\theta_{(X_n, \omega)}) > \min_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) + \varepsilon \right) = 0,$$

where X_n is the random sample of n instances. This implies that the bound on the sample size can be made absolute, not only relative to the size of the world. I state the sufficient part of the theorem in the weaker form (1.2) in order to facilitate the comparison with the main results of the current paper.

³Series of papers extend the original result from $Y = \{0, 1\}$ to other outcome spaces ((Haussler 1992), (Haussler and Long 1995), (Cesa-Bianchi and Haussler 1998), (Anstee and Furedi 1986), (Anstee 2006) (Kearns and Schapire 1994), (Alon, Ben-David, Cesa-Bianchi, and Haussler 1997)). This leads to the *uniform strong laws of large numbers*: general sufficient and necessary conditions for a class of functions to satisfy the thesis of Glivenko-Cantelli Theorem. VC dimension also gives rise to the *uniform central limit theorems* (see (Pollard 1985), (Dudley 1999), (Talagrand 2003), (Mendelson and Vershynin 2003)). The uniform theorems has been used in (Pakes and Pollard 1989) to prove asymptotics for optimization estimators. (Al-Najjar 2006) uses the VC dimension in a decision theoretic approach to complexity.

of its customers and uses the information to make personalized recommendations. Precise recommendations are highly profitable as they enhance the quality of experience of Netflix customers, which, in turn, increases the number of movie rentals.

The recommendation decision is formally equivalent to the treatment choice. (Manski 2004) discusses the treatment choice model as the decision problem under uncertainty and presents a formal approach in spirit of (Wald 1950) and (Blackwell and Girshick 1954): the statistician chooses treatment given the sample results and the choice optimizes certain criterion (as maximin or minimal regret). This can be naturally divided into two steps. First, the statistician uses the sample to estimate the treatment effects. In the terminology of the current paper, this corresponds to the choice of a model that minimizes the empirical risk. Next, the treatment choice is made given the estimates. Given the first step is properly addressed, the treatment choice is not difficult. In the current paper, I focus only the first step. Intuitively, in order to make good recommendations, Netflix must first predict the movie ratings.⁴

Let $Y = \{0, 1\}$ be the set of possible ratings and let

$$\mathbf{X} = C \times F \times \Xi,$$

be the set of instances, where C is the set of all customers, F is the set of all movies. For example, $\omega(c, f) = 0$ means that customer c does not like movie f .⁵

⁴Manski separately discusses one- and two-step decision problem. There are other differences between (Manski 2004) and this paper. Manski assumes that $|T| = 2$, whereas here T is infinite. He considers only specific models that depend only on observable covariates of the agents. The argument is that, at least in some cases, the decision maker may be constrained to offer treatments that depend only on prespecified characteristics (and do not depend on others, for example, race or age). In the Netflix problem, there is relative lack of observable covariates and the main challenge is to personalize recommendations to specific customers. This means that the statistician wants to make use of any heterogeneity, whether it is observed or latent. Finally, (Manski 2004) derives robust (i.e., distribution-free) sample estimates of the quality of the treatment rule chosen from a *particular* family of models. The Hoeffding inequality is used to show that robust estimates are consistent. The scope of the current paper is wider: I want to characterize *all* families of models, for which the sample estimates are robustly consistent.

⁵More generally, one could assume that

$$\mathbf{X} = C \times F \times \Xi,$$

where Ξ is the set of observable characteristics of movies and customers (for example, the average income of a customer, the release date of a movie or the distribution of the age of actors). Since the relative lack of observable characteristics makes them not so relevant for the Netflix problem, and their treatment is also somehow distracting from the main point of this example, I assume here that $\Xi = \emptyset$, i.e. there are no observable characteristics. The main body of the paper covers the general case.

Let \mathcal{M}^F be the family of all models for which prediction depends only on the movie:

$$\mathcal{M}^F = \{\theta : \theta(c, f) = \theta(c', f) \text{ for any } c, c' \in C, f \in F\}.$$

Clearly, for any sequence of different movies f_1, \dots, f_k , and any sequence of ratings $y_1, \dots, y_k \in \{0, 1\}$, there is a model $\theta \in \mathcal{M}^F$ that predicts that the rating of movie f_l is equal to y_l . But this means that \mathcal{M}^F does not have a finite VC-dimension:

$$|\mathcal{M}_S^F| = 2^{|S|}$$

for any subset of instances $S \subseteq \mathbf{X}$, such that for any different instances $(c, f), (c', f') \in S$, the corresponding movies are also different, $f \neq f'$ (Recall that the set of model restrictions is defined in (1.3).) The Vapnik-Chervonenkis theorem implies that ERM is not consistent robustly to *all* worlds.

However, observe that all worlds include, among others, those in which there is only one observation per movie. This is too strong. To make recommendations, Netflix should be able to predict ratings of *all* movies for *all* customers in their database. Therefore, Netflix is naturally interested only in worlds that have a rectangular shape. Precisely, I am going to consider only worlds (X, ω) where $X \subseteq \mathbf{X}$ is a product set:

$$\Sigma_n = \{(C' \times F', \omega) : C' \subseteq C, F' \subseteq F, \omega \in Y^{C' \times F'} \text{ and } n \leq |C'| < \infty, n \leq |F'| < \infty\}. \quad (1.5)$$

It is a simple exercise to show that for any $\varepsilon > 0$, any $\gamma > 0$, (1.2) holds. Hence, finite VC-dimension is not necessary for ERM that is robustly consistent to sufficiently large *rectangular* worlds.

It is important to emphasize that the restriction to rectangular worlds does not imply any prior knowledge about the functional relationship between instances and outcomes. In a sense, this is a restriction purely on the logical structure of the problem, not on its content.

Note that γ , the size of the sample relative to the size of the world, can be chosen any small. This makes the result potentially useful for Netflix: if sampling is costly, and good predictions are profitable, than the total cost of sampling can be made any small relative to the potential profits from good predictions.⁶

⁶In fact, Netflix has an access to sampling technology: It may (and it does) inquire a customer about preferences over a particular movie.

I assume here that the sampling is independent from the world, and, in particular, independent from the realization of preferences. Of course, customer will be able to rate a movie (most often) only if she has already watched it. Because the customer is also more likely to watch movies that she likes, the sample is not chosen independently from the ratings. To address the self-selection problem, one needs a model of sample choice.

The main result of this paper extends the notion of VC-dimension so to find the necessary and sufficient conditions for the consistent ERM that is robust to sufficiently large worlds with a particular shape. I discuss here an application of this extension to the Netflix example. Take any family of models \mathcal{M} . For any finite subsets of customers $A \subseteq C$ and movies $B \subseteq F$, let $\mathcal{M}_{A \times B} \subseteq \{0, 1\}^{|A| \times |B|}$ be the set of model restrictions defined in (1.3). Of course, for any $A \subseteq C, B \subseteq F$, $|\mathcal{M}_{A \times B}| \leq 2^{|A||B|}$. Say that family of models \mathcal{M} has a *finite matrix dimension* if there are k, l , such that

$$\sup_{A \subseteq C, B \subseteq F \text{ and } |A| \leq k, |B| \leq l} |\mathcal{M}_{A \times B}| < 2^{kl}.$$

For example, family \mathcal{M}^F has a finite matrix dimension: For any A and B , such that $|A| = 1$ and $|B| = 2$, $\mathcal{M}_{A \times B}^F = \{(0, 0), (1, 1)\}$ and $|\mathcal{M}_{A \times B}^F| = 2 < 2^2$.

I show that a family of models \mathcal{M} has a finite matrix dimension if and only if consistent ERM is robust to sufficiently large rectangular worlds, i.e. (??) holds with \mathcal{M}^F replaced by \mathcal{M} .

The matrix dimension is a strictly weaker concept than the VC-dimension. Note that VC-dimension constrains the size of the set of model restrictions to any subset of explanatory variables \mathbf{X} of certain size. The matrix dimension puts a constraint only on restrictions to product subsets of \mathbf{X} , i.e., subsets that generate rectangular worlds. Of course, there strictly more sets of size $|A| \times |B|$, then the rectangular sets $A \times B$.

2. MODEL AND ASSUMPTIONS

2.1. Model. Let \mathbf{X} be an infinite space of *instances* and Y be a space of *outcomes*. A *world* is a pair (X, ω) of finite subset of instances $X \subseteq \mathbf{X}$ and assignment of outcomes to instances $\omega : X \rightarrow Y$. The size of world (X, ω) is defined as $|X|$. Let Σ be the set of all worlds. For any world (X, ω) , define the *risk of model* $\theta : \mathbf{X} \rightarrow Y$ as the average loss from prediction as in (1.1). I write $R_{\Omega}^{(l)}$ when it is necessary to emphasize that the risk depends on a particular loss function $l : Y \times Y \rightarrow [0, 1]$. I consider separately two cases. In the binary case $Y = \{0, 1\}$, I assume that $l(y, y') = |y - y'|$. In the general case of Y compact and metrisable, I assume that l is a continuous function and $l(y, y) = 0$ for any $y \in Y$.

For any finite $X \subseteq \mathbf{X}$, any $\gamma > 0$, let X_{γ} denote the random sample of $\gamma|X|$ observations drawn uniformly from X without replacement.⁷ Let $R_{(X_{\gamma}, \omega)}(\theta)$ denote the *empirical risk* of model θ computed in the sample X_{γ} .

⁷Nothing changes in the subsequent analysis if observations are drawn uniformly with replacement and the sample X_{γ} is an ordered tuple of $\gamma|X|$ observations rather than a set.

To define the empirical risk minimizer, I consider separately both cases, binary and compact Y . Suppose first that Y is binary. Let $\mathcal{M} \subseteq Y^X$ be a family of models. Because the world and the sample are finite, $\min_{\theta \in \mathcal{M}} R_{(X_\gamma, \omega)}(\theta)$ is attained at certain model $\theta \in \mathcal{M}$. Denote the *empirical risk minimizer* as

$$\theta_{(X_\gamma, \omega)} \in \arg \min_{\theta \in \mathcal{M}} R_{(X_\gamma, \omega)}(\theta). \quad (2.1)$$

In general, there are many models that minimize sample risks. Model $\theta_{(X_\gamma, \omega)}$ is one of them. For the subsequent results, it does not matter how $\theta_{(X_\gamma, \omega)}$ is chosen, as long as the choice *depends only* on observations in sample X_γ .

Next, suppose that Y is compact and metrisable. In general, there is no guarantee that $\inf_{\theta \in \mathcal{M}} R_{\Omega_\gamma}(\theta)$ is attained. For any $\eta > 0$, define the set of empirical risk η -minimizers:

$$\Theta_{(X_\gamma, \omega)}^\eta = \left\{ \theta \in \mathcal{M} : R_{(X_\gamma, \omega)}(\theta) \leq \inf_{\theta' \in \mathcal{M}} R_{(X_\gamma, \omega)}(\theta') + \eta \right\}.$$

Let

$$\theta_{(X_\gamma, \omega)}^\eta \in \Theta_{(X_\gamma, \omega)}^\eta$$

be the *empirical risk η -minimizer* that is chosen in any way so that the choice *depends only* on observations in sample X_γ . If Y is binary, then $\theta_{(X_\gamma, \omega)}^0 = \theta_{(X_\gamma, \omega)}$.

The goal of this paper is to present necessary and sufficient conditions for robustly consistent ERM. For this purpose, let $\Sigma = \Sigma_0 \subseteq \Sigma_1 \subseteq \Sigma_2 \subseteq \dots$ be a decreasing sequence of universes, i.e. decreasing sequence of sets of worlds. Sequence (Σ_n) is interpreted as worlds that are of interest for the statistician and that have increasing size.

Definition 1. *Say that consistent ERM that is robust to sequence (Σ_n) is possible for family \mathcal{M} if and only if for any $\epsilon > 0$, any $\gamma > 0$ any $\eta > 0$*

$$\limsup_n \sup_{(X, \omega) \in \Sigma_n} P_{X_\gamma} \left(R_{(X, \omega)} \left(\theta_{(X_\gamma, \omega)}^\eta \right) > \inf_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) + \epsilon + \eta \right) = 0. \quad (2.2)$$

If Y is binary, I require the above to hold for $\eta = 0$.

This says that for any $\epsilon > 0$, there is a sufficiently large n , such that for all worlds $(X, \omega) \subseteq \Sigma_n$, with high probability, the true risk of the empirical risk minimizer is close to the true minimal risk across all models in \mathcal{M} .

In the Introduction, I described the classical statistical learning theory of Vapnik and Chervonenkis. This theory requires robustness to *all* worlds. It is natural to define Σ_n as the set of all worlds that have size at least n . In Section 1.2, I argued that Netflix is interested in predicting outcomes only in rectangular worlds. There, Σ_n is defined in (1.5) and it consist of worlds (X, ω) , where X is a product of sets of size at least n .

This ends the presentation of the primitives of the model. To summarize, these are spaces \mathbf{X}, Y , loss function l and sequence of universes (Σ_n) that is of interest for the statistician. In the rest of this Section, I impose assumptions on space \mathbf{X} and the sequences of universes (Σ_n) . These assumptions culminate in the definition of *acceptable sequence* (Definition 4). Because of the statements and the proofs, it is convenient to state these assumptions in an abstract manner using certain tools from algebra. To facilitate the presentation, I discuss how these assumptions fit into the Netflix example discussed in Section 1.2.

2.2. Indexed worlds. Assume that set of instances \mathbf{X} is a product of two sets:

$$\mathbf{X} = I \times \Xi,$$

where I is an infinite set of indices and Ξ is a set of additional characteristics of an instance.

Let $J \subseteq I$ be a finite set of indices. Say that world (X, ω) is indexed by J , write $\Omega \in \Sigma(J)$, if there is an assignment of observable characteristics $\xi : J \rightarrow \Xi$, such that

$$X = \{(i, \xi(i)) : i \in J\}. \tag{2.3}$$

In other words, (X, ω) is indexed by J , if (a) only observations index with indices in J can be part of the world and (b) for each index $i \in J$, the world consists exactly one observation with this index.

In the *Netflix example*, $I = C \times F$ and each observation is indexed with a tuple (c, f) , where $c \in C$ is a customer and $f \in F$ is a movie. Ξ may include additional observable characteristics of customers, like age, credit rating, education, or additional observable characteristics of movies as year of release or age of actors.

2.3. Permutations. A *permutation* of I is any bijection from I to I . A set G of permutations is a *group* if (a) $\text{id}_I \in G$, (b) $g^{-1} \in G$ for any $g \in G$ and (c) $g \circ g' \in G$ for any $g, g' \in G$. (See (Lang 2002) for references.) Let G be a group of permutations of I , $G \mapsto I$. Group action $G \mapsto I$ induces a group action $G \mapsto 2^I$ on the set 2^I of all subsets of I : for any $g \in G$, any $S \subseteq I$, let

$$g \cdot S = \{g \cdot i : i \in S\} \subseteq I.$$

Abusing terminology, I say that set $g \cdot S$ is a permutation of S .

Say that group action $G \mapsto I$ is *transitive* if for any two $i, i' \in I$, there is a permutation g , such that $g \cdot i = i'$. For any subset of instances $U \subseteq I$ define a subgroup of permutations that keep set U invariant:

$$G_U = \{g \in G : g \cdot U = U\}.$$

Definition 2. Say that finite $U \subseteq I$ is local (under group action G), if for any $S \subseteq U$,

$$\{g \cdot S : g \cdot S \subseteq U, g \in G\} = \{g \cdot S : g \in G_U\}.$$

Take any finite $U \subseteq I$. Suppose that for some subset $S \subseteq U$, there is a permutation $g \in G$, such that $g \cdot S$ is contained in U . In general, there might be an index $i \in U$, such that $g \cdot i \notin U$. Set U is local, if for any $S \subseteq U$ and g , such that $g \cdot S \subseteq U$, there is a permutation g_U , such that $g \cdot S = g_U \cdot S$ and $g_U \cdot U = U$. In a sense, U is local if the action of the "local" group G_U behaves in the same way as the action of the original group G .

Definition 3. Say that action $G \mapsto I$ is locally generated if there is an increasing sequence of local sets $I_1 \subseteq I_2 \subseteq \dots, I_n$, such that for any finite $S \subseteq I$, there is a permutation $g \in G$ and n , such that $g \cdot S \subseteq I_n$.

Locally generated group action can be approximated by group actions on finite sets. Any increasing sequence of local sets with the property stated in the Definition is called a *generating sequence*.

Consider two examples of group actions.

Example 1 (Symmetric group). Let I be an infinite set and let Π_I be the group of all bijections from I to I . This group is known in the literature as the symmetric group of I .

Any finite $U \subseteq I$ is local under the symmetric group action $\Pi_I \mapsto I$. Also, any (strictly) increasing sequence of subsets of I is a generating sequence. This implies that $\Pi_I \mapsto I$ is locally generated.

Example 2 (*Product_k*). Let $I = I^1 \times \dots \times I^k$ be a product of $k \geq 2$ infinite sets I^j , $j = 1, \dots, k$. Let $G = \Pi_{I^1} \times \dots \times \Pi_{I^k}$ be a group product of k symmetric groups: for any $(g_1, \dots, g_k) \in G$, any $(i_1, \dots, i_k) \in I$, let

$$(g_1, \dots, g_k) \cdot (i_1, \dots, i_k) := (g_1 \cdot i_1, \dots, g_k \cdot i_k).$$

It is easy to check, that any product set $U = U^1 \times \dots \times U^k$ for finite $U^k \subseteq I^k$ is local under the group action $G \mapsto I$. Consider any (strictly) increasing sequences of finite sets $I_1^j \subset I_2^j \subset \dots \subset I^j$ for $j = 1, \dots, k$. Then, sets $I_n = I_n^1 \times \dots \times I_n^k$ form a generating sequence.

In the Netflix example, $k = 2$ and $I_1 = C$ is a set of customers and $I_2 = F$ is a set of movies. For any finite $C' \subseteq C, F' \subseteq F$, product set $C' \times F'$ is local. Take increasing sequences of finite sets $C_1 \subset C_2 \subset \dots$ and $F_1 \subset F_2 \subset \dots$. Then, sets $C_n \times F_n$ form a generating sequence.

Not all group actions are locally generated. The most important example is the shift action on time indices:

Example 3 (Shift). Let $G = I = \mathbf{Z}$, where \mathbf{Z} is the set of integers. For any $g \in G, i \in I$, let $g \cdot i := g + i$. Then, no finite subset of I is local and group action $G \mapsto I$ is not locally generated.

2.4. Acceptable sequences. The above definitions lead to an assumption on sequences of universes.

Definition 4. Let $\Sigma = \Sigma_0 \subseteq \Sigma_1 \subseteq \Sigma_2 \subseteq \dots$ be a decreasing sequence of universes. Say that sequence (Σ_n) is acceptable (under locally generated group action $G \mapsto I$), if there is a generating sequence $I_1 \subseteq I_2 \subseteq \dots$, such that

$$\Sigma_n \subseteq \bigcup_{\substack{U \supseteq I_n, \\ U \text{ is local}}} \bigcup_{g \in G} \Sigma(g \cdot U). \quad (2.4)$$

Sequence of universes is acceptable if the universes consists of worlds indexed with permutations of all local sets that contain elements of a generating sequence. An acceptable sequence consists of worlds that are indexed with sufficiently large local sets of indices. This restricts the sets of instances that may appear in the worlds from the sequence to those that have a particular shape. It is important to emphasize that this does not restrict in any way the relationship between instances and outcomes.

In particular, consider the Netflix example and a sequence of universes Σ_n of rectangular worlds that is defined in (1.5). This sequence is acceptable under group action $\Pi_C \times \Pi_F \mapsto C \times F$. To see it, notice that for any finite sets $C' \times F', C'' \times F'' \subseteq C \times F$, if $|C'| = |C''|$ and $|F'| = |F''|$, then there is a permutation $g \in \Pi_C \times \Pi_F$, such that

$$g \cdot (C' \times F') = C'' \times F''.$$

Take any generating sequence of indices $C_n \times F_n \in I$, such that $|C_n| = |F_n| = n$. Then, Σ_n satisfies Definition (2.4) and is acceptable.

3. EXAMPLES

In this Section, I show the various assumptions presented above are satisfied for a wide class of learning problems. I list here examples of such problems. In each case, I describe the primitives: spaces \mathbf{X}, Y and the sequence of universes that is of interest for the statistician. Then, I show that the sequence of universes is acceptable under certain locally generated and transitive group action.

3.1. Marriage market. Let M be a set of men and W be a set of women and let $I = M \times W$. Let Ξ contains observable characteristics of agents (for example, age, religion, education level, income). Let $(y_M, y_W) \in Y = [0, 1]^2$ be the utility of each of the partners in the match. Model $\theta : I \times \Xi \rightarrow Y$ describes the utility of men and women from each possible match.

The formal structure of this learning problem is exactly the same as the structure of the Netflix problem and it is not surprising that sequence of universes,

$$\Sigma_n = \bigcup_{\substack{M' \subseteq M, \\ n \leq |M'| < \infty}} \bigcup_{\substack{W' \subseteq W, \\ n \leq |W'| < \infty}} \Sigma(M' \times W').$$

is acceptable sequence under the product group action $\Pi_M \times \Pi_W \mapsto M \times W$.

(Hitsch, Hortacsu, and Ariely 2006) estimate preferences of partners in the match using revealed choices of the partners. Their specification allows for dependence of preferences on a rich set of observable characteristics Ξ . Alternative specifications could possibly address unobserved heterogeneity among agents.

3.2. Community detection. Let B be a community of agents and let

$$I = \{(b_1, b_2) : b_1, b_2 \in B, b_1 \neq b_2\} \quad (3.1)$$

be the set of ordered pairs of different community members. Let Y be the set of attitudes between agents. Model $\theta : I \rightarrow Y$ describes attitudes between all agents, where $\theta(b_1, b_2) = y$ means that " b_1 has attitude y towards b_2 ".

This example is an application of the so called community detection problem. In this problem, the statistician wants to know whether there the community can be divided into groups in such a way that individuals interact mostly inside the groups and only rarely between groups. The problem has a large literature in sociology (see (Scott 2000)), computer science ((Newman 2004), (Newman 2006); see also (Copic, Kirman, and Jackson 2006)) and graph theory (see (Alon and Shapira 2005)).⁸

For any finite B' , let $I(B') \subseteq I$ be the set of ordered pairs of different members of B' :

$$I(B') = \{(b_1, b_2) : b_1, b_2 \in B', b_1 \neq b_2\}. \quad (3.2)$$

Consider the following sequence of universes:

$$\Sigma_n = \bigcup_{B' \subseteq I \text{ and } n \leq |B'| < \infty} \Sigma(I(B')). \quad (3.3)$$

⁸For another example, suppose that B is a set of individuals who are eavesdropped by the National Security Agency. Let $Y = \{0, 1, 2\}$ where 0 is interpreted as an "no relationship", 1 is "accidental relationship" and 2 is "criminal relationship".

This is sequence of worlds that are indexed with all interactions between agents from n' -element sets of agents, $n' \geq n$.

The community detection problem is formally different than the learning problems discussed so far. In the Netflix problem (as well as in the marriage market), an index $i \in I$ is an ordered pair of elements of two disjoint sets of customers and movies and the set of indexes is a product of two sets. Here, an index $i \in I$ is an ordered pair of elements of *the same* set and I can be represented as the set of ordered edges on a graph with nodes B .

To show that (Σ_n) is an acceptable sequence, consider the following group action:

Example 4 (*Ordered Graph₂*). Let B and I be as above. Let $G = \Pi_B$ be the symmetric group on B . Define action $\Pi_B \mapsto I$: for any $g \in G$, any $(b_1, b_2) \in I$, let

$$g \cdot (b_1, b_2) = (g \cdot b_1, g \cdot b_2).$$

This group action acts on the ordered edges of the graph through permutations of nodes. It is easy to check that for any finite $B' \subseteq B$, $I(B')$ is local under this group action. For any increasing sequence of finite sets $B_1 \subset B_2 \subset \dots \subset B$, sets $I(B_n)$ form a generating sequence. A generating sequence induced by sets B_n , $|B_n| = n$, gives rise to sequence of acceptable universes (Σ_n) defined in (3.3).

3.3. Binary preferences over products. Let B be an infinite set of products and let I be the set of ordered pairs of different products as in (3.1).⁹ Let $Y = \{', >'\}$. Interpret any model $\theta : I \rightarrow Y$ as a set of statements about binary preferences of certain customer: for any $(b_1, b_2) \in I$, $\theta(b_1, b_2) = '>'$ if the customer prefers b_2 to b_1 .

For any finite $B' \subseteq B$, let $I(B')$ be the set of all binary comparisons between products in B' (this set is formally defined in (3.2)). Let Σ_n be the set of all worlds that are indexed with sets $I(B')$, where $n \leq |B'| < \infty$ (3.3). This is an acceptable sequence under the Ordered Graph₂ group action from Example 4.

Say that model $\theta \in \{0, 1\}^I$ respects transitivity if for any $a, b, c \in B$, if $\theta(a, b) = \theta(b, c) = '>'$, then $\theta(a, c) = '>'$. Consider family $\mathcal{M}^T \subseteq \{0, 1\}^I$ of all models that respect transitivity. In Section 6, I demonstrate that ERM that is consistent robustly to any acceptable sequence is possible for family \mathcal{M}^T . This result corresponds to an analogous finding in (Kalai 2003). There, it is shown that rational choice functions are learnable: the number of mistakes committed while predicting the results of choices from sets is small relative to the number of correct predictions if the choice function is known to be rational (but nothing else is known a priori).

⁹I am grateful for this example to Matias Iaryczower.

3.4. Utility of bundles of products. Let B be an infinite set of products. Let $I = \{A \subseteq B : |A| = k\}$ be the set of k -element bundles of products. Let $Y = [0, 1]$ be the space of utilities. Model $\theta : I \rightarrow Y$ is interpreted as the utility of a customer from each of the possible k -bundle of products.

Take any finite set $B' \subseteq B$ and define $I(B') \subseteq I$ as the set of all k -element subsets of B' :

$$I(B') = \{A \subseteq B' : |A| = k\}.$$

Consider the following sequence of universes:

$$\Sigma_n = \bigcup_{B' \subseteq B \text{ and } n \leq |B'| < \infty} \Sigma(I(B')). \quad (3.4)$$

This is the sequence of worlds that are indexed with sets of interactions of n' -element sets of products, $n' \geq n$.

To show that (Σ_n) is an acceptable sequence, consider the following group action:

Example 5 (*Graph_k*). Let B and I be as above. Define group action $\Pi_B \mapsto I$: for any $g \in \Pi_B$, for any $\{b_1, \dots, b_k\} \in I$, let

$$g \cdot \{b_1, \dots, b_k\} := \{g \cdot b_1, \dots, g \cdot b_k\} \in I.$$

It is easy to check that $I(B')$ is local for any finite $B' \subseteq B$. For any increasing sequence of finite sets $B_1 \subset B_2 \subset \dots \subset B$, sets $I(B_n)$ form a generating sequence. A generating sequence induced by sets B_n , $|B_n| = n$, gives rise to sequence of acceptable universes (Σ_n) defined in (3.4).

3.5. Team assignment. Let J be a set of jobs and let B be a set of workers. Let S_B^k be a set of k -element subsets of workers, i.e. *teams* of k workers. Let $I = J \times S_B^k$ and let $Y = [0, 1]$. Here, $\theta(j, \{b_1, \dots, b_k\}) \in Y$ is interpreted as the productivity of team $\{b_1, \dots, b_k\} \subseteq B$ assigned to job j .

For any finite $B' \subseteq B$, let $I(B')$ be the set of k -element subsets of B' defined as in the example above. Take any increasing sequence of natural numbers m_n , $\lim_{n \rightarrow \infty} m_n = \infty$ and consider the following sequence of universes:

$$\Sigma_n = \bigcup_{\substack{B' \subseteq B, \\ n \leq |B'| < \infty}} \bigcup_{\substack{J' \subseteq J, \\ m_n \leq |J'| < \infty}} \Sigma(J' \times I(B')). \quad (3.5)$$

This is the sequence of worlds that are indexed with sets of interactions of n' -element sets of agents, $n' \geq n$.

To show that (Σ_n) is an acceptable sequence, consider the following group action:

Example 6. Let J, B and I be as above. Define group action $\Pi_J \times \Pi_B \mapsto J \times I$: for any $(g_J, g_B) \in \Pi_J \times \Pi_B$, for any $(j, \{b_1, \dots, b_k\}) \in I$, let

$$(g_J, g_B) \cdot (j, \{b_1, \dots, b_k\}) := (g_J \cdot j, \{g_B \cdot b_1, \dots, g_B \cdot b_k\}) \in I.$$

One checks that $J' \times I(B')$ is local for finite $J' \subseteq J$ and $B' \subseteq B$. For any increasing sequences of finite sets $J_1 \subseteq J_2 \subseteq \dots \subseteq J$ and $B_1 \subseteq B_2 \subseteq \dots \subseteq B$, such that $\lim_{n \rightarrow \infty} \min(|J_n|, |B_n|) = \infty$, sets $J_n \times I(B_n)$ form a generating sequence. Any generating sequence of this form, and such that $|J_n| = m_n$ $|B_n| = n$, gives rise to the sequence of acceptable universes (Σ_n) defined in (3.5).

4. DIMENSION AND RESULTS WHEN $Y = \{0, 1\}$

In this Section, I discuss the binary case $Y = \{0, 1\}$. I present a definition of G -dimension that generalizes both the VC -dimension and the matrix dimension from Section 1.2. This definition is later used to present the sufficient and necessary conditions for robustly consistent ERM.

4.1. G -dimension when $\mathbf{X} = I$. Here, I consider the case when there are no observable characteristics of instances, or, in other words, when $\mathbf{X} = I$. The general case follows. Let $\mathcal{M} \subseteq \{0, 1\}^I$ be a family of models. For any finite $S \subseteq I$, define set of model restrictions \mathcal{M}_S as in (1.3). Then, $|\mathcal{M}_S| \leq 2^{|S|}$.

Definition 5. G -dimension of family $\mathcal{M} \subseteq \{0, 1\}^I$, write $\dim_G \mathcal{M}$, is defined as a collection of all finite $S \subseteq I$, such that

$$\sup_{g \in G} |\mathcal{M}_{g \cdot S}| < 2^{|S|}.$$

In other words, finite set of indices S belongs to G -dimension of \mathcal{M} , if, for *any* permutation of set S , model restrictions $\mathcal{M}_{g \cdot S}$ omit at least one configuration of outcomes.

Different groups of permutations lead to different definitions of the dimension. Before I discuss applications, a general comment is in order. Suppose that G' is a proper subgroup of group G ¹⁰. Then, $\dim_G \mathcal{M} \subseteq \dim_{G'} \mathcal{M}$. This is because $S \in \dim_G \mathcal{M}$ is the more restrictive, the larger group G is. In other words, for any finite $S \subseteq I$, there will be fewer families of models that satisfy $S \in \dim_G \mathcal{M}$, then those that satisfy $S \in \dim_{G'} \mathcal{M}$. In particular, $S \in \dim_G \mathcal{M}$ has the largest bite if G is equal to the symmetric group Π_I of all permutations on I .

The first example demonstrates that Definition 6 encompasses the VC -dimension. Recall the formal definition from the Introduction. Let Π_I be the symmetric group on

¹⁰ G' is a proper subgroup of G if $G' \subsetneq G$ and G' is a group (i.e., it satisfies the group axioms).

I . Then, for any two finite sets $S, S' \subseteq I$ of the same cardinality, $|S| = |S'|$, there is a permutation $g \in \Pi_i$, such that $S' = g \cdot S$. Hence,

$$\begin{aligned} \dim_{\Pi_I} \mathcal{M} &= \left\{ S : \sup_{g \in \Pi_I} |\mathcal{M}_{g \cdot S}| < 2^{|S|} \right\} \\ &= \{ S : |\mathcal{M}_S| < 2^{|S|} \} \\ &= \{ S : |S| \geq \dim_{VC} \mathcal{M} \}. \end{aligned}$$

In other words, Π_I -dimension of \mathcal{M} consists of all finite subsets of I that contain at least k elements, where k is equal to the VC dimension of \mathcal{M}_X .

Consider now the Netflix example and let $G = \Pi_C \times \Pi_F$ be the product of two symmetric groups on C and F , respectively. Suppose that the G -dimension of family \mathcal{M} contains a finite $S \subseteq C \times F$, or, in other words, for any permutation $g \in G$, $|\mathcal{M}_{g \cdot S}| < 2^{|S|}$. Because S is finite, there are finite sets $A \subseteq C, B \subseteq F$, such that $S \subseteq A \times B$. By the definition of permutation, $g \cdot S \subseteq g \cdot (A \times B)$. Of course, $|\mathcal{M}_{g \cdot S}| < 2^{|S|}$ implies that

$$|\mathcal{M}_{g \cdot (A \times B)}| < 2^{|A| \times |B|}.$$

Therefore, we get the equivalence: \mathcal{M} has nonempty G -dimension if and only if it has finite matrix dimension.

4.2. G -dimension. Next, I consider the general case with nonempty set of observable characteristics X . Let $\mathcal{M} \subseteq \{0, 1\}^{\mathbf{X}}$ be a family of models. For any finite $S \subseteq I$, for any assignment of observable characteristics $\xi : S \rightarrow \Xi$, define set of model restrictions

$$\mathcal{M}_{S, \xi} = \left\{ \tau \in \{0, 1\}^S : \text{there is } \theta \in \mathcal{M}, \text{ st. } \tau(i) = \theta(i, \xi(i)) \text{ for each } i \in S \right\}. \quad (4.1)$$

Then, $|\mathcal{M}_{S, \xi}| \leq 2^{|S|}$. This is equivalent to definition in equation (1.3) when $\mathbf{X} = I$.

Definition 6. G -dimension of family $\mathcal{M} \subseteq \{0, 1\}^{\mathbf{X}}$ (write $\dim_G \mathcal{M}$) is defined as a collection of all finite $S \subseteq I$, such that

$$\sup_{g \in G} \sup_{\xi: g \cdot S \rightarrow \Xi} |\mathcal{M}_{g \cdot S, \xi}| < 2^{|S|}.$$

This definition generalizes the definition of G -dimension presented above. Here, finite set of indices S belongs to G -dimension of \mathcal{M} if, for *any* permutation of set S and *any* assignment of observable characteristics, model restrictions omit at least one configuration of outcomes.

To illustrate this definition, I show yet another connection between Π_I -dimension and VC -dimension. Take any family of models without indices, $\mathcal{M}_{\Xi} \subseteq \{0, 1\}^{\Xi}$ and let

$(\mathcal{M}_\Xi)^I \subseteq \{0, 1\}^{\Xi \times I}$ be the product of I copies of \mathcal{M}_Ξ . Then,

$$\begin{aligned} \dim_{\Pi_I} (\mathcal{M}_\Xi)^I &= \left\{ S : \sup_{g \in G} \sup_{\xi: g \cdot S \rightarrow \Xi} \left| \left((\mathcal{M}_\Xi)^I \right)_{g \cdot S, \xi} \right| < 2^{|S|} \right\} \\ &= \left\{ S : \sup_{\xi: S \rightarrow \Xi} \left| \left\{ \theta \circ \xi \in \{0, 1\}^S : \theta \in \mathcal{M}_\Xi \right\} \right| < 2^{|S|} \right\} \\ &= \{ S : |S| \geq \dim_{VC} \mathcal{M}_\Xi \}. \end{aligned}$$

In other words, Π_I -dimension of $(\mathcal{M}_\Xi)^I$ consists of all finite subsets of Ξ that contain at least $\dim_{VC} \mathcal{M}$ elements.

4.3. Sufficient conditions for learning. The fact that G -dimension of family \mathcal{M} contains a finite set of indices S means that not "everything goes" and model restrictions are constrained on permutations of S . Different sets S lead to different restrictions and not all of them are equally powerful. It is useful to distinguish here a special class of subsets.

Definition 7. *Say that finite set $S \subseteq I$ is generic if for any $\varepsilon > 0$, there is a local set U , such that for any subset $D \subseteq U$, $|D| \geq \varepsilon |U|$, there is $g \in G$, such that $g \cdot S \subseteq D$.*

Set S is generic, if for any $\varepsilon > 0$, there is a local set U , such that any subset of U with at least $\varepsilon |U|$ elements contains a permutation of S . In a sense, S is generic if its permutations can be fit almost everywhere. Note a simple fact that a subset of a generic set is also generic.

To illustrate the definition, consider the following examples (the claims made in these examples are proven in Section 6).

Example 7 (*Product_k*). *Consider Example 2. Any finite $S \subseteq I$ is generic.*

Example 8 (*Graph_k*). *Consider Example 5. Take any disjoint finite subsets $B^1, \dots, B^k \subseteq B$, $B^l \cap B^{l'} = \emptyset$ for any $l \neq l'$. Define set of indices*

$$I(B^1, \dots, B^k) = \{ i \in I : |i \cap B^l| = 1 \text{ for each } l \leq k \}.$$

This set is generic.

The next Theorem finds the sufficient conditions for consistent ERM that is robust to some worlds.

Theorem 1. *Fix a sequence of universes (Σ_n) that is acceptable under locally generated and transitive group action $G \mapsto I$. Then, for any family of models \mathcal{M} , if G -dimension of \mathcal{M} contains a generic set, then, consistent ERM that is robust to sequence (Σ_n) is*

possible.

Moreover, for any $\varepsilon > 0$, any $\gamma > 0$, any generic S , there exists a local set $U \subseteq I$, such that if $S \in \dim_G \mathcal{M}$ for some family \mathcal{M} , then

$$\sup_{\substack{V \supseteq U, \\ V \text{ is local}}} \sup_{g \in G} \sup_{(X, \omega) \in \Sigma(g \cdot V)} P_{X_\gamma} \left(R_{(X, \omega)}(\theta_{(X_\gamma, \omega)}) > \inf_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) + \varepsilon \right) \leq \exp \left(-\frac{1}{4} \gamma \varepsilon^2 |X| \right). \quad (4.2)$$

The Theorem says that if G -dimension of family of models contains a generic set, then ERM is consistent robustly to sufficiently large worlds indexed with local sets of indices. This extends the statistical learning theory of Vapnik and Chervonenkis. It has been shown above that, the Vapnik-Chervonenkis dimension coincides with G -dimension when G is equal to the symmetric group on the set of indices I . When $G \subsetneq \Pi_I$, the Theorem yields novel set of conditions.

As an application, consider the Netflix example from Section 1.2. Let G be the product of symmetric groups on the sets of customers and movies. I show above that family of models has finite matrix dimension if and only if it has nonempty G -dimension. By Example 7, G -dimension is nonempty if and only if it contains a generic set. Thus, finite matrix dimension is sufficient for consistent ERM that is robust to large rectangular worlds! This shows the sufficient part of the claim made in the end of Section 1.2.

Proof. The first statement follows from the second. Indeed, fix $\varepsilon > 0$, $\gamma > 0$ and generic S and find local set $U \subseteq I$ from the second statement. Take any acceptable sequence of universes (Σ_n) and let $I_1 \subseteq I_2 \subseteq \dots$ be a generating sequence that gives rise to (Σ_n) , i.e.

$$\Sigma_n \subseteq \bigcup_{\substack{V \supseteq I_n, \\ V \text{ is local}}} \bigcup_{g \in G} \Sigma(g \cdot V).$$

Since (I_n) is a generating sequence, there is n^* high enough and a permutation g^* , so that $g^* \cdot U \subseteq I_{n^*}$. Then, for any $n \geq n^*$,

$$\begin{aligned} & \sup_{(X, \omega) \in \Sigma_n} P_{X_\gamma} \left(R_{(X, \omega)}(\theta_{(X_\gamma, \omega)}) > \inf_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) + \varepsilon \right) \\ &= \sup_{\substack{V \supseteq I_n, \\ V \text{ is local}}} \sup_{g \in G} \sup_{(X, \omega) \in \Sigma(g \cdot V)} P_{X_\gamma} \left(R_{(X, \omega)}(\theta_{(X_\gamma, \omega)}) > \inf_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) + \varepsilon \right) \\ &\leq \sup_{\substack{V \supseteq I_n, \\ V \text{ is local}}} \sup_{g \in G} \sup_{(X, \omega) \in \Sigma(g \cdot V)} \exp \left(-\frac{1}{4} \gamma \varepsilon^2 |X| \right) \\ &\leq \exp \left(-\frac{1}{4} \gamma \varepsilon^2 |I_n| \right). \end{aligned}$$

(The first inequality is a consequence of the fact that for each local $V \supseteq I_n$, $U \subseteq (g^*)^{-1} \cdot V$ is also a local set.) Finally, the fact that $\lim_{n \rightarrow \infty} |I_n| = \infty$ completes the proof of the first statement.

The proof of the second statement has two parts. The first one is a combinatorial result that is proven in Section 8.2:

Lemma 1. *For any $\epsilon > 0$, any generic $S \subseteq I$, there is a local set U , such that for any family \mathcal{M} , if $S \in \dim_G \mathcal{M}$, then for any local $V \supseteq U$, any permutation $g \in G$, any assignment of observable characteristics $\xi : g \cdot V \rightarrow \Xi$,*

$$\log |\mathcal{M}_{g \cdot V, \xi}| \leq \epsilon |V|. \quad (4.3)$$

This means that the cardinality of model restrictions $\mathcal{M}_{g \cdot V}$ is bounded by $e^{\epsilon |V|}$.¹¹

The second part is an application of the Hoeffding inequality. Let $\epsilon = \frac{1}{4} \gamma \epsilon^2$. Find a local set $U \subseteq I$, such that (4.3) holds. Take any local set $V \supseteq U$, a permutation $g \in G$ and a world $(X, \omega) \in \Sigma(g \cdot V)$. Let $\xi : g \cdot V \rightarrow \Xi$ be an assignment of observable characteristics such that (2.3) is satisfied. Notice that

$$\begin{aligned} P_{X_\gamma} \left(R_{(X, \omega)}(\theta_{(X_\gamma, \omega)}) > \inf_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) + \epsilon \right) \\ \leq P_{X_\gamma} \left(\sup_{\theta \in \mathcal{M}} |R_{(X, \omega)}(\theta_{(X_\gamma, \omega)}) - R_{(X, \omega)}(\theta)| > \epsilon \right). \end{aligned}$$

For any model $\theta \in \mathcal{M}$, there is a model restriction $\tau = \theta|_X \in \mathcal{M}_{g \cdot V, \xi}$, such that for any subset $X' \subseteq X$,

$$R_{(X', \omega)}(\tau) := \frac{1}{|X'|} \sum_{i: (i, \xi(i)) \in X'} l(\tau(i), \omega(i, \xi(i))) = R_{(X', \omega)}(\theta).$$

¹¹The Lemma extends a corollary to Sauer-Shelah's lemma ((Sauer 1972), (Shelah 1972), (Vapnik and Chervonenkis 1971)). The Lemma says that if $\dim_{VC} \mathcal{M} = k$, then, for any finite $J \subseteq I$

$$|\mathcal{M}_J| \approx O(|J|^k).$$

As a corollary, we get that for any finite $U \subseteq I$,

$$\sup_{V \subseteq I, |V| \geq |U|} \frac{\log |\mathcal{M}_V|}{|V|} \leq \frac{\log |U|}{|U|}.$$

Hence,

$$\begin{aligned}
& P_{X_\gamma} \left(R_{(X,\omega)}(\theta_{(X_\gamma,\omega)}) > \inf_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon \right) \\
& \leq P_{X_\gamma} \left(\sup_{\tau \in \mathcal{M}_{g,V,\xi}} |R_{(X_\gamma,\omega)}(\tau) - R_{(X,\omega)}(\tau)| > \varepsilon \right) \\
& \leq \sum_{\tau \in \mathcal{M}_{g,V,\xi}} P_{X_\gamma} (|R_{(X_\gamma,\omega)}(\tau) - R_{(X,\omega)}(\tau)| > \varepsilon) \\
& \leq |\mathcal{M}_{g,V,\xi}| \max_{\tau} P_{X_\gamma} (|R_{(X_\gamma,\omega)}(\tau) - R_{(X,\omega)}(\tau)| > \varepsilon).
\end{aligned}$$

The probability in the last line can be bounded using the Hoeffding inequality¹²: For any $\tau \in \{0, 1\}^{g \cdot V}$,

$$P_{X_\gamma} (|R_{\Omega_\gamma}(\tau) - R_{(X,\omega)}(\tau)| > \varepsilon) \leq \exp \left(-\frac{1}{2} \gamma \varepsilon^2 |X| \right).$$

Hence,

$$\begin{aligned}
& P_{X_\gamma} \left(R_{(X,\omega)}(\theta_{(X_\gamma,\omega)}) > \inf_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon \right) \\
& \leq \exp \left(\left(\varepsilon - \frac{1}{2} \gamma \varepsilon^2 \right) |X| \right) \leq \exp \left(-\frac{1}{4} \gamma \varepsilon^2 |X| \right).
\end{aligned}$$

□

4.4. Necessary conditions. It can be shown that the sufficient conditions are also necessary for consistent ERM that is robust to *all* acceptable sequences of universes. Here, I am going to demonstrate a stronger statement: if the sufficient conditions are not satisfied, then there is no consistent ERM that is robust to *any* acceptable sequence of universes. I need a definition first.

Definition 8. *Say that a group action $G \mapsto I$ is tight if it is locally generated and transitive and for any local $U \subseteq I$, there is $S \subseteq U$ that is generic and $|S| \geq \delta |U|$.*

¹²Let $\mu = \frac{1}{m} \sum_{i=1}^m v(i)$ be a population mean of variables $v(i) \in [0, 1]$. Let $l_1^r, \dots, l_n^r \in \{1, \dots, m\}, n < m$ be random sample drawn uniformly with ($r = w$) or without ($r = o$) replacement. (Hoeffding 1963) shows that for both $r = w, o$,

$$P_r \left(\left| \frac{1}{n} \sum_{i=1}^n v(l_i^r) - \mu \right| > \varepsilon \right) \leq 2 \exp(-2n\varepsilon^2).$$

This says that the same inequality can be applied also to sampling with replacement and all the results go through.

Using examples presented above, it is easy to show that the symmetric group, $Product_k$ and $Graph_k$ are all tight group actions. These claims and also some other statements about tightness are shown formally in Section 6.

Theorem 2. *Suppose that $G \mapsto I$ is tight. If G -dimension of family of models \mathcal{M} does not contain any generic set, then, for any $\varepsilon > 0$, there are $c_\varepsilon > 0$ and $\gamma_\varepsilon > 0$, such that for any finite $J \subseteq I$, any $\gamma \leq \gamma_\varepsilon$, there are a permutation $g \in G$ and a world $(X, \omega) \in \Sigma(g \cdot J)$, such that*

$$P_{X_\gamma} \left(R_{(X,\omega)}(\theta_{(X_\gamma,\omega)}) < \min_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon \right) \leq \exp(-c_\varepsilon |J|).$$

The Theorem says that if G -dimension of a family of models does not contain generic set, then for *any sufficiently large* $J \subseteq I$, there are permutation $g \in G$ and a world $\Omega \in \Sigma(g \cdot J)$, such that, with high probability, the true risk of the empirical risk minimizer is ε -far to the true minimal risk. Hence, should not be used unless one is ready to make restrictive assumptions about the relationship between instances and outcomes. The proof of the Theorem can be found in Section 8.4.

The Theorem implies immediately that:

Corollary 1. *Suppose that G -dimension of family of models \mathcal{M} does not contain any generic set. Take an acceptable sequence (Σ_n) . Then, ERM that is consistent robustly to (Σ_n) is impossible for family \mathcal{M} .*

Proof. Let $I_1 \subseteq I_2 \subseteq \dots$ be a generating sequence that gives rise to (Σ_n) . Then,

$$\begin{aligned} & \sup_{(X,\omega) \in \Sigma_n} P_{X_\gamma} \left(R_{(X,\omega)}(\theta_{(X_\gamma,\omega)}) \geq \min_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon \right) \\ & \geq 1 - \inf_{(X,\omega) \in \Sigma_n} \exp(-c_\varepsilon |X|) \geq 1 - \exp(-c_\varepsilon |I_n|). \end{aligned}$$

When $n \rightarrow \infty$, the above expression converges to 1. □

Consider the Netflix example: Let G be the product of symmetric groups on the sets of customers and movies and suppose that G -dimension of family \mathcal{M} does not contain any generic set. By Example 7, this means that G -dimension of \mathcal{M} is empty. Recall that any two rectangles $C' \times F'$, such that $|C'| = k$, $|F'| = l$ are permutations for each other. The Theorem implies that for any sufficiently large k and l , there is $C' \times F'$, such that $|C'| = k$, $|F'| = l$, such that with high probability the true risk of the empirical risk minimizer is ε -far from the true minimal risk across all models. In other words, with high probability, ERM fails to be consistent.

To see some intuition for this result, notice that, if G -dimension of family \mathcal{M} is empty, then for any finite $J \subseteq I$, there are a permutation $g \in G$ and an assignment of observable characteristics $\xi : g \cdot J \rightarrow \Xi$, such that

$$|\mathcal{M}_{g \cdot J, \xi}| = 2^{|J|}.$$

Let $X = \{(i, \xi(i)) : i \in g \cdot J\}$. The above says that for any true functional relationship between instances and outcomes $\omega : \mathbf{X} \rightarrow Y$, for any subset $J' \subseteq g \cdot J$, there is a model $\theta \in \mathcal{M}$, such that

$$\begin{aligned} \theta(i, \xi(i)) &= \omega(i, \xi(i)) \text{ for any } i \in J' \text{ and} \\ \theta(i, \xi(i)) &\neq \omega(i, \xi(i)) \text{ for any } i \in g \cdot J \setminus J'. \end{aligned}$$

Hence, for any $X' \subseteq X$, $|X'| \leq \frac{1}{2}$, there are models $\theta_{\min}, \theta_{\max} \in \mathcal{M}$, such that

$$\begin{aligned} R_{(X', \omega)}(\theta_{\min}) &= R_{(X', \omega)}(\theta_{\max}) = 0 \text{ and} \\ R_{(X, \omega)}(\theta_{\min}) &= 0 \text{ and } R_{(X, \omega)}(\theta_{\max}) = 1 - \frac{|X'|}{|X|} \end{aligned}$$

Both models $\theta_{\min}, \theta_{\max}$ minimize the sample risk. In fact, in the sample they are indistinguishable. However, only θ_{\min} minimizes the true risk. Given that ERM chooses model independently from the realization of outcomes outside the sample, there is no guarantee that it will pick θ_{\min} rather than θ_{\max} .

This problem is otherwise known in the statistical literature as *overfitting*: if the family of models is chosen too large, then for any sample, one can easily find a model that has perfect fit on the sample, but absolutely no fit outside the sample.

5. DIMENSION AND RESULTS WHEN Y IS COMPACT

(Alon, Ben-David, Cesa-Bianchi, and Haussler 1997) propose an extension of VC-dimension that is applicable for $Y = [0, 1]$. Here, I use similar idea to extend Definition 6 to the case of Y compact and metrisable. I show that this extension leads to the sufficient and necessary conditions for learning.

5.1. (G, δ) -dimension. Let Y be a compact space with metric $d : Y \times Y \rightarrow [0, 1]$. Let $\mathcal{M} \subseteq Y^{\mathbf{X}}$ be a family of models. Say that finite subset of instances $S \subseteq I$ is δ -shattered by family \mathcal{M} if there is an outcome $y \in Y$ and a function $\xi : S \rightarrow \Xi$, such that for any subset $S' \subseteq S$, there is a model $\theta \in \mathcal{M}$ and such that

$$\begin{aligned} d((\theta(i, \xi(i)), y) &\leq \delta \text{ for any } i \in S' \text{ and} \\ d(\theta(i, \xi(i)), y) &\geq 2\delta \text{ for any } i \in S \setminus S'. \end{aligned} \tag{5.1}$$

Definition 9. (G, δ) -dimension of family $\mathcal{M} \subseteq Y^{\mathbf{X}}$ (write $\dim_{(G, \delta)} \mathcal{M}$) is a collection of all finite sets $S \subseteq I$, such that for any permutation $\in G$, $g \cdot S$ is not δ -shattered by \mathcal{M} .

This definition generalizes the definition of G -dimension (Definition 6). Indeed, note that if $Y = \{0, 1\}$ and $d(y, y') = |y - y'|$, then for any $\mathcal{M} \subseteq \{0, 1\}^{\mathbf{X}}$, for any $\delta < \frac{1}{2}$,

$$\dim_G \mathcal{M} = \dim_{(G, \delta)} \mathcal{M}.$$

5.2. Sufficient and necessary conditions for learning. Theorem 1 extends in the following way (the proof can be found in Section 8.3):

Theorem 3. Fix a sequence of universes (Σ_n) that is acceptable under locally generated and transitive group action $G \mapsto I$. Then, for any family of models \mathcal{M} , if for any $\delta > 0$, (G, δ) -dimension of \mathcal{M} contains a generic set, then, consistent ERM that is robust to sequence (Σ_n) is possible.

Similarly as in the Theorem 1, the rate of convergence can be bounded exponentially. Next, I show that the sufficient conditions are also necessary.

Theorem 4. Suppose that $G \mapsto I$ is tight. For any $\varepsilon > 0$, there are $c_\varepsilon > 0$ and $\gamma_\varepsilon > 0$, such that for any family of models \mathcal{M} , if, for some $\delta > 0$, (G, δ) -dimension of \mathcal{M} does not contain any generic set, then, there is a loss function l , such that for any finite $J \subseteq I$, there are a permutation $g \in G$ and a world $(X, \omega) \in \Sigma(g \cdot J)$, such that

$$P_{X_\gamma} \left(R_{(X, \omega)}^{(l)} \left(\theta_{(X_\gamma, \omega)}^\eta \right) < \inf_{\theta \in \mathcal{M}} R_{(X, \omega)}^{(l)}(\theta) + \varepsilon^* \right) \leq \exp(-c_\varepsilon |J|),$$

for any $\gamma \leq \gamma_\varepsilon$ and any $\eta > 0$.

The Theorem says that if G -dimension of a family of models does not contain generic set, then there is a loss function, such that for any sufficiently large $J \subseteq I$, there are permutation $g \in G$, world $\Omega \in \Sigma(J)$, such that the true risk of the empirical risk minimizer is *not* close to the true minimal risk with high probability.

The thesis of Theorem 3 holds for any loss function. On the other hand, Theorem 4 implies that if the sufficient conditions of Theorem 3 are not satisfied, then its thesis is also not satisfied for some loss function. Of course, Theorem 4 does not hold for all loss functions: For example, if the loss is identically equal to 0, $l \equiv 0$, then, quite trivially, ERM is robustly consistent for any family of models.

Together, Theorems 3 and 4 provide a simple way to check whether consistent ERM that is robust to acceptable sequences. I discuss some applications of these results in Section 7.

6. TIGHTNESS OF VARIOUS DATA STRUCTURES

In this Section, I state some results about tightness of various group actions. In particular, I am going to show that the statements about generic sets made in Examples 7 and 8 are correct. Also, I show that the sufficient and necessary conditions derived in Sections 4 and 5 apply to all examples from Section 3. The proofs of the results can be found in Appendix C.1.

Suppose that $G_j \mapsto I_j$ are group actions for $j = 1, \dots, d$. One can define a product group action $G_1 \times \dots \times G_d \mapsto I_1 \times \dots \times I_d$ in a natural way: for any $(i_1, \dots, i_d) \in I_1 \times \dots \times I_d$, any $(g_1, \dots, g_d) \in G_1 \times \dots \times G_d$, define

$$(g_1, \dots, g_d) \cdot (i_1, \dots, i_d) := (g_1 \cdot i_1, \dots, g_d \cdot i_d).$$

The next useful result shows that tightness is preserved under products (the proof can be found in Appendix C.1).

Theorem 5. *The product of local sets is local under the product of group action. The product of generic sets is generic under the product group action. The product of tight group actions is tight.*

The Theorem, the fact that any finite subset is generic under the symmetric group, and the fact that any subset of generic set is also generic, lead together to a simple corollary.

Corollary 2. *Consider Example 2. The product group action of symmetric groups is tight and any finite subset of I is generic.*

Recall that the sequence of universes in the Netflix problem and the marriage market are acceptable under product action from Example 2. The Corollary shows that this group action is tight, and that the necessary and sufficient conditions for robustly consistent ERM apply in these learning problems.

Proposition 1. *Consider Example 4. Group action $G \mapsto I$ is tight. For any mutually disjoint finite sets $B^1, B^2 \subseteq B$, set*

$$I(B^1, B^2) = \{(b_1, b_2) : b_j \in B_j \text{ for } j = 1, 2\}$$

is generic.

This shows that the group action used in the community detection problem is tight and that the sufficient and necessary conditions apply for this learning problem. The same applies to the problem of inferring binary preferences over products.

Proposition 2. *Consider Example 8. Group action $G \mapsto I$ is tight. For any mutually disjoint finite sets $B^1, \dots, B^k \subseteq B$, set $I(B^1, \dots, B^k) \subseteq I$ is generic.*

This shows that the group action used in the utility over bundles is tight. Together with Theorem 5, this also demonstrates the tightness of the group action used in the team assignment problem.

7. APPLICATIONS: *Product*₂

In October 2006, Netflix announced a public contest for improving their recommendation algorithm. To allow the contestants to train their algorithms on real data, they released dataset of "100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18 thousand movie titles."¹³ The Grand Prize is going to be awarded to an algorithm that improves the standard deviation of prediction from the true observations by 10% relative to the current result of Netflix own algorithm.

The literature discusses, among others, two possible algorithms: clustering and factor models. Both of these algorithms can be defined as the ERM on a particular family of models. The goal of this Section is to describe the families of models that correspond to each of these algorithms and show that they satisfy the sufficient conditions for learning.

Assume here that $I = I_1 \times I_2$, where I_1 is the set of customers and I refer to the first coordinate of index $i = (i_1, i_2)$ as a customer; similarly, I_2 is the set of movies and I refer to the second coordinate of i as a movie. Let $G = \Pi_{I_1} \times \Pi_{I_2}$ be the product of two symmetric groups (see Example 2).

7.1. Clustering. Assume that each customer has one of finitely many types $a_1 \in A_1$; similarly, each of the movies has one of finitely many types $a_2 \in A_2$. A set of customers or movies with the same type is called a *cluster*. There is a prediction function $\rho : A_1 \times A_2 \rightarrow Y$ with the following interpretation: if a customer has type a_1 and a movie has type a_2 , then the model predicts that the outcome is equal to $\rho(a_1, a_2)$. Each model is generated by the assignment of types to customers and movies and by the choice of prediction function.

One can generalize this to allow for infinite but compact sets of types and to make prediction dependent on observable characteristics $x \in \Xi$. Suppose that Ξ, A_1, A_2 are compact and metrisable spaces with metrics, respectively, d_Ξ, d_{A_1}, d_{A_2} . Ξ is the space of observable characteristics of customer-movie pairs. For each $j = 1, 2$, space A_j contains unobservable characteristics of j th coordinate of an index i . Let $\alpha_j : I_j \rightarrow A_j$ be an assignment of unobservable characteristics to j th coordinates.

¹³Source: www.netflixprize.com.

Let $\rho : \Xi \times A_1 \times A_2 \rightarrow Y$ be a prediction function. The interpretation is that each instance with observable and unobservable characteristics equal to (x, a_1, a_2) is predicted to have an outcome equal to $\rho(x, a_1, a_2)$. Say that prediction function is *Lipschitz with constant C* if for any $(x, a_1, a_2), (x', a'_1, a'_2) \in \Xi \times A_1 \times A_2$,

$$d_Y(\rho(x, a_1, a_2), \rho(x', a'_1, a'_2)) \leq C [d_\Xi(x, x') + d_{A_1}(a_1, a'_1) + d_{A_2}(a_2, a'_2)].$$

Say that model $\theta : \mathbf{X} \rightarrow Y$ is generated by assignments α_1, α_2 and prediction function ρ if

$$\theta(i, x) = \rho(x, \alpha_1(i_1), \alpha_2(i_2)) \text{ for any } i \in I, x \in \Xi.$$

For any $C > 0$, consider the family of models $\mathcal{M}^C \subseteq Y^{\mathbf{X}}$ generated by *all* assignments of unobservable characteristics and *all* prediction functions that are Lipschitz with constant C . I show that this family satisfies the sufficient conditions for learning.

Proposition 3. *For any $C > 0$, any $\delta > 0$, (G, δ) -dimension of \mathcal{M}^C contains a generic set.*

Proof. Let $\mathcal{V}_\Xi, \mathcal{V}_1, \mathcal{V}_2$ be finite coverings of sets, respectively, Ξ, A_1, A_2 with open balls of radius $\frac{\delta}{2C}$ (in their respective metrics). Such coverings exist, because sets Ξ, A_1 and A_2 are compact and metrisable.

Let $k_\Xi = |\mathcal{V}_\Xi|$ and $k = |\mathcal{V}_1| \times |\mathcal{V}_2|$. There is n , such that for any $U_1 \subseteq I_1, U_2 \subseteq I_2$ $|U_j| \geq n$, for any function $\xi : I \rightarrow \Xi$, there are $V \in \mathcal{V}_\Xi$ and subsets $S_j \subseteq U_j$ for both $j = 1, 2$ and $|S_j| = k + 1$, such that for any $i \in S_1 \times S_2$, $\xi(i) \in V$. Indeed, any set $U = U_1 \times U_2$ is local and, by Corollary 2, any subset $S = S_1 \times S_2 \subseteq U$ is generic. Hence, for sufficiently high n , for any subset $D \subseteq U$, $|D| \geq \frac{1}{k_\Xi} |U|$, there is a permutation $g \in G$, such that $g \cdot S \subseteq D$. Define $D_V = \{i \in U : \xi(i) \in V\}$ - there is at least one $V \in \mathcal{V}_\Xi$, such that $|D_V| \geq \frac{1}{k_\Xi} |U|$.

Let $S \subseteq U$ have the required property. Take subset $S' \subseteq S$, $|S'| = k + 1$ and such that for any $(i_1, i_2), (i'_1, i'_2) \in S'$, for each $j = 1, 2$, if $i_j \neq i'_j$, then $i_{-j} \neq i'_{-j}$. In other words, S' is a "diagonal" subset of S . I show that there is no outcome $y \in Y$ and model $\theta \in \mathcal{M}^C$, such that (5.1) holds. Take any $y \in Y$ and $\theta \in \mathcal{M}^C$ that is generated by assignments α_1 and α_2 and prediction function ρ . Because $|S'| = k + 1$, for any pairs of assignments α_1 and α_2 , there are open balls $V_1 \in \mathcal{V}_1, V_2 \in \mathcal{V}_2$ and two different indices $i, i' \in S'$, $i_j \neq i'_j$ and

$$\alpha_j(i_j), \alpha_j(i'_j) \in V_j$$

for both $j = 1, 2$. Consider an instance $\bar{i} = (i_1, i'_2) \in S \setminus S'$. Then,

$$\begin{aligned} & d_Y(\rho(\xi(i), \alpha_1(i_1), \alpha_2(i_2)), \rho(\xi(\bar{i}), \alpha_1(i_1), \alpha_2(i'_2)))) \\ & \leq C [d_{\Xi}(\xi(i), \xi(\bar{i})) + d_{A_1}(\alpha_1(i_1), \alpha_1(i_1)) + d_{A_2}(\alpha_2(i_2), \alpha_2(i'_2))] \\ & \leq C \left[\frac{\delta}{2C} + 0 + \frac{\delta}{2C} \right] \leq \delta. \end{aligned}$$

This demonstrates that any set $U = U_1 \times U_2$, $|U_j| \geq n$, is not δ -shattered by \mathcal{M}^C . This also ends the proof of the Proposition, because, by Corollary 2, U is generic. \square

7.2. Factor models. In order to describe the factor models, I assume for simplicity that $Y = [0, 1]$ and that $\mathbf{X} = I$, i.e. there is no observable characteristics of movies and customers. On the other hand, each customer and each movie has unobservable characteristics called *factors*. As it is standard in the literature, I assume that factors are elements of euclidean space. Fix the number of factors k . Let $\psi_1 : I_1 \rightarrow R^k$ and $\psi_2 : I_2 \rightarrow R^k$ be factor assignments for customers and movies, respectively. For any pair of factor assignments ψ_1, ψ_2 , define model θ_{ψ_1, ψ_2} : for any $(i_1, i_2) \in I_1 \times I_2$,

$$\theta_{\psi_1, \psi_2}(i_1, i_2) = \sigma((\psi_1(i) | \psi_2(i))),$$

where $(u|v)$ denotes the scalar product of two vectors $u, v \in R^k$ and $\sigma : R \rightarrow [0, 1]$ is continuous and strictly increasing function. Let $\mathcal{M}^{\sigma, k}$ be a family of all models of this form:

$$\mathcal{M}^{\sigma, k} = \{ \theta_{\psi_1, \psi_2} : \psi_i : I_i \rightarrow R^k \text{ for both } i = 1, 2 \}.$$

It is instructive to notice the differences between the clustering and the factor models. On one hand, the relationship between factors and the outcome is assumed here to be linear. This is restrictive given that in the clustering case, any prediction function is allowed. On the other hand, factors here are members of R^n and not just some compact subset of it. I do not make also an assumption that the factor models must be uniformly Lipschitz. Although both of these assumptions are standard in the literature, they are nevertheless restrictive. The ability to remove them is one of the advantages of the methods based on VC- or G -dimensions.

I show that this family satisfies the sufficient conditions for learning.

Proposition 4. *For any continuous and strictly increasing $\sigma : R \rightarrow [0, 1]$, for any k , there is a generic $S \subseteq I$, such that $S \in \dim_{(G, \delta)} \mathcal{M}^{\sigma, k}$ for any $\delta > 0$.*

The proof of the Proposition is based on a simple geometrical result. Take any $y_d < y_u$, $y_d, y_u \in R \cup \{-\infty, \infty\}$ and consider a mapping $\tau^k : R^k \rightarrow \{0, 1\}^k$ defined as

$$(\tau^k(a_1, \dots, a_k))_l = 1 \text{ iff } a_l \in [y_d, y_u] \text{ for any } (a_1, \dots, a_k) \in R^k.$$

Lemma 2. *Let $A \subseteq R^{k+1}$ be a k -dimensional linear subspace. Then,*

$$|\{\tau^{k+1}(a) : a \in A\}| < 2^{k+1}.$$

Proof. Suppose not. Then, there is at least $k+1$ vectors $a^1, \dots, a^{k+1} \in A$, such that

$$(\tau^{k+1}(a^l))_{l'} = 1 \text{ if and only if } l = l' \text{ for any } l, l' \leq k+1.$$

These vectors must be linearly independent. Indeed, let $\alpha = \sum_{l' \neq l} \beta_{l'} a^{l'}$ be a linear combination of vectors $\alpha_{l'}, l' \neq l$. Then,

$$\alpha = \sum_{l' \neq l} \beta_{l'} a^{l'} \in [y_d, y_u],$$

because $a^{l'} \in [y_d, y_u]$ for any $l' \neq l$. Hence, $\alpha \neq a_l$.

This establishes a contradiction with the fact that A is a k -dimensional linear subspace. \square

Proof of Proposition 4. Take any subsets $S_1 \subseteq I_1, S_2 \subseteq I_2$, such that $|S_1| = k+1$ and $|S_2| = 2^{k+1}$. I show that set $S_1 \times S_2$ is not δ -shattered by family $\mathcal{M}^{\sigma, k}$.

Take any $y \in [0, 1]$ and $\delta > 0$. Define

$$y_d := \sigma^{-1}(y - \delta), y_u := \sigma^{-1}(y + \delta),$$

with the convention that $y_d = -\infty$ or $y_u = \infty$ if the respective values do not belong to the image of function σ . For any factor assignment $\psi_1 : I_1 \rightarrow R^k$, define $(k+1)$ -dimensional vectors: for any $l \leq k$, $a^l \in R^{S_1}$ and

$$a^l(i_1) = (\psi(i_1))_l \text{ for any } i_1 \in S_1.$$

Let A^{ψ_1} be a k -dimensional linear subspace of R^{k+1} spanned by vectors a^l . Define

$$\mathcal{M}_S = \left\{ \tau \in \{0, 1\}^S : \begin{array}{l} \text{there are } \psi_i : I_i \rightarrow R^k, i = 1, 2, \text{ such that for any } (i_1, i_2) \in S \\ \tau(i_1, i_2) = 1 \text{ iff } (\psi_1(i_1) | \psi_2(i_2)) \in [y_d, y_u]. \end{array} \right\}$$

Then, for any $\tau \in \mathcal{M}_S$ there is a factor assignment $\psi_1 : I_1 \rightarrow R^k$, such that, for any $i_2 \in S_2$,

$$\tau(\cdot, i_2) \in \{\tau^{k+1}(a) : a \in A^{\psi_1}\}.$$

Consider any $\tau^* \in \{0, 1\}^S$, such that for any $i_2, i'_2 \in S_2$, $\tau(\cdot, i_2) \neq \tau(\cdot, i'_2)$. By the Lemma above, $\tau^* \notin \mathcal{M}_S$. But this means that for $S' = (\tau^*)^{-1}(1)$, there is no model $\theta \in \mathcal{M}^{\sigma, k}$, such that (5.1) holds.

This demonstrates that $S_1 \times S_2$ is not δ -shattered by family $\mathcal{M}^{\sigma, k}$ and ends the proof of the Proposition. \square

8. PROOFS

8.1. Combinatorial result. The proof of Theorem 1 is based on Lemma 1. In order to prove Theorem 3, a slightly more general version is needed. Let I be the space of indices and let the space of outcomes consists of 0, 1, and an additional outcome $*$. Let $\mathcal{M} \subseteq \{0, 1, *\}^I$ be a family of models. Then, for any finite $J \subseteq I$, $|\mathcal{M}_J| \leq 3^{|J|}$ and $|\mathcal{M}_J \cap \{0, 1\}^J| \leq 2^{|J|}$.

Say that finite $S \subseteq I$ belongs to G -dimension* of \mathcal{M} , write $S \in \dim_G^* \mathcal{M}$, if

$$\sup_{g \in G} |\mathcal{M}_{g \cdot S} \cap \{0, 1\}^{g \cdot S}| < 2^{|S|}.$$

In other words, S belongs to G -dimension* of \mathcal{M} , if for any permutation of set S , model restrictions $\mathcal{M}_{g \cdot S}$ omit at least one configuration $\tau \in \{0, 1\}^{g \cdot S}$.

Say that set of model restrictions $\mathcal{M}_J^* \subseteq \{0, 1\}^J$ approximates \mathcal{M} on set J up to outcome $*$ if for any $\theta \in \mathcal{M}$, there is $\tau \in \mathcal{M}_J^*$, such that for any $i \in J$, either $\theta(i) = \tau(i)$ or $\theta(i) = *$.

Proposition 5. For any generic S , for any $\epsilon > 0$, there is a local U , such that for any local $V \supseteq U$, any permutation g , any family of models $\mathcal{M} \subseteq \{0, 1, *\}^I$, if $S \in \dim_G^* \mathcal{M}$, then there is $\mathcal{M}_{g \cdot V}^* \in \{0, 1\}^{g \cdot V}$ that approximates \mathcal{M} on $g \cdot V$ up to $*$ and such that

$$\log |\mathcal{M}_{g \cdot V}^*| \leq \epsilon |V|.$$

The Proposition says that if G^* -dimension of $\mathcal{M} \subseteq \{0, 1, *\}^I$ contains a generic set, then, for sufficiently large local sets V , there is a set of model restrictions $\mathcal{M}_{g \cdot V}^* \in \{0, 1\}^{g \cdot V}$ that approximates \mathcal{M} on $g \cdot V$ and that has small cardinality (relatively to the cardinality of all possible model restrictions on $g \cdot V$). The Proposition is proven in Appendix B.

8.2. Proof of Theorem 1. Only Lemma 1 remains to be proven. Fix $\epsilon > 0$ and generic S . Find local U from the Proposition 5. Take any assignment of observable characteristics $\xi : I \rightarrow \Xi$ and any family \mathcal{M} of models, such that $S \in \dim_G \mathcal{M}$. Define $\mathcal{M}^{(\xi)} \subseteq \{0, 1\}^I$ as

$$\mathcal{M}^{(\xi)} = \left\{ \theta^{(\xi)} \in \{0, 1\}^I : \text{there is } \theta \in \mathcal{M}, \text{ st. } \theta^{(\xi)}(i) = \theta(i, \xi(i)) \text{ for all } i \in I \right\}.$$

Then, for any finite $J \subseteq I$, $\mathcal{M}_J^{(\xi)} = \mathcal{M}_{J, \xi|_J}$. By the definition of G -dimension, for any permutation g ,

$$|\mathcal{M}_{g \cdot S}^{(\xi)}| = |\mathcal{M}_{g \cdot S, \xi}| < 2^{|S|}.$$

i.e., $S \in \dim_G \mathcal{M}^{(\xi)}$, where G -dimension is applicable here to family $\mathcal{M}^{(\xi)} \subseteq \{0, 1\}^I$ (see Section 4.1). By Proposition 5, for any local $V \supseteq U$, any permutation g ,

$$\log |\mathcal{M}_{g \cdot V, \xi}| = \log \left| \mathcal{M}_{g \cdot V}^{(\xi)} \right| \leq \epsilon |V|.$$

8.3. Proof of Theorem 3. I begin with a lemma.

Lemma 3. *For any $\delta > 0$, any $\epsilon' > 0$, any generic S , there exists a local U , such that for any family of models $\mathcal{M} \subseteq Y^{\mathbf{X}}$ such that $S \in \dim_{(G, \delta)} \mathcal{M}$, any local $V \supseteq U$, any permutation g , for any assignment of observable characteristics $\xi : g \cdot V \rightarrow \Xi$, there is a set $\mathcal{T} \subseteq Y^{g \cdot V}$, such that*

$$\log |\mathcal{T}| \leq \epsilon' |V|,$$

and, for any model $\theta \in \mathcal{M}$, there is $\tau \in \mathcal{T}$, such that for any $i \in g \cdot V$,

$$d(\theta(i, \xi(i)), \tau(i)) \leq 2\delta.$$

Proof. For any $y \in Y$, any $\delta \geq 0$, let $B(y_k, \delta)$ be the open ball with center at y and radius δ . Because Y is compact, there is a finite cover of Y with K_δ balls of radius δ , i.e., there is a sequence $y_1, \dots, y_{K_\delta} \in Y$, such that

$$Y = \bigcup_{k=1}^{K_\delta} B(y_k, \delta).$$

For any $k = 1, \dots, K_\delta$, define function $v^k : Y \rightarrow \{0, 1, *\}$ as

$$v^k(y) = \begin{cases} 0, & y \in B(y_k, \delta), \\ 1, & y \notin B(y_k, 2\delta), \\ *, & \text{otherwise.} \end{cases}$$

Take any $\delta > 0$, any $\epsilon' > 0$ and any generic S . Use Proposition 5 to find a local U that satisfies the thesis of the Proposition for $\epsilon = \frac{\epsilon'}{K_\delta}$ and S .

From now on, fix a local V , a permutation $g \in G$ and a function $\xi : g \cdot V \rightarrow \Xi$. Notice that, for any model θ , $\theta(i, \xi(i)) \in Y$, hence $v^k(\theta(i, \xi(i))) = 0$ for at least one k . Notice also that for any $i \in g \cdot V$,

$$\text{if } v^k(\theta(i, \xi(i))) \in \{0, *\}, \text{ then } d(\theta(i, \xi(i)), y_k) \leq 2\delta. \quad (8.1)$$

Take any family $\mathcal{M} \subseteq Y^{\mathbf{X}}$ such that $S \in \dim_{(G, \delta)} \mathcal{M}$. Construct a family of models $\mathcal{M}^k \subseteq \{0, 1, *\}^I$:

$$\mathcal{M}^k = \{\theta^k \in \{0, 1, *\}^I : \text{there is } \theta \in \mathcal{M}, \text{ st. } \theta^k(i) = v^k(\theta(i, \xi(i))) \text{ for each } i \in I\}.$$

Because (G, δ) -dimension of \mathcal{M} contains generic set S , it must be that G -dimension* of \mathcal{M}^k contains S for any k . By Proposition 5, for any k , there is a set $\mathcal{T}^k \in \{0, 1\}^{g \cdot V}$, such that

$$\log |\mathcal{T}^k| \leq \epsilon |V| \quad (8.2)$$

and, for any $\theta \in \mathcal{M}$, there is $\tau_\theta^k \in \mathcal{T}^k$, such that

$$\text{for any } i \in g \cdot V, \text{ either } (v^k \circ \theta)(i) = \tau_\theta^k(i) \text{ or } (v^k \circ \theta)(i) = *. \quad (8.3)$$

Hence, whenever $\tau_\theta^k(i) = 0$, then, by (8.1), $d(\theta(i, \xi(i)), y_k) \leq 2\delta$.

For any vector $\eta \in \{0, 1\}^{K_\delta}$, define

$$k(\eta) = \min \{k : \eta_k \in \{0\}\}.$$

By the above, there is always k , such that $\tau_\theta^k(i) = 0$ and $k(\tau_\theta^1(i), \dots, \tau_\theta^{K_\delta}(i))$ is well-defined. For any sequence of $\tau_k \in \mathcal{T}^k$, $k \leq K_\delta$, define $\theta(\tau_1, \dots, \tau_{K_\delta}) \in Y^{g \cdot V}$ as

$$\theta(\tau_1, \dots, \tau_{K_\delta})(i) = y_{k(\tau_1(i), \dots, \tau_{K_\delta}(i))} \text{ for any } i \in g \cdot V.$$

Then, for any $i \in g \cdot V$

$$d\left(\theta(i, \xi(i)), \theta\left(\tau_\theta^1, \dots, \tau_\theta^{K_\delta}\right)(i)\right) = d\left(\theta(i, \xi(i)), y_{k(\tau_\theta^1(i), \dots, \tau_\theta^{K_\delta}(i))}\right) \leq 2\delta.$$

The second inequality is a consequence of (8.1) and (8.3).

Finally, take \mathcal{T} as the set of all $\theta(\tau_1, \dots, \tau_{K_\delta}) \in Y^{g \cdot V}$ of the above form:

$$\mathcal{T} = \{\theta(\tau_1, \dots, \tau_{K_\delta}) : \tau_k \in \mathcal{T}^k, k \leq K_\delta\}.$$

Then,

$$|\mathcal{T}| \leq |\mathcal{T}^1| \cdot \dots \cdot |\mathcal{T}^{K_\delta}|.$$

Together with (8.2) and the fact that $\epsilon' = \frac{\epsilon}{K_\delta}$, this implies the thesis of the Lemma. \square

The rest of the proof of the Theorem follows the same lines as the proof of Theorem 1.

Fix $\epsilon > 0$, $\eta > 0$ and $\gamma > 0$. Let $\delta = \frac{\epsilon}{6}$ and $\epsilon' = \frac{1}{36}\gamma\epsilon^2$. Suppose that $S \in \dim_{(G, \delta)} \mathcal{M}$ is generic and find local U that satisfies the thesis of the Lemma for S , ϵ' and $\delta > 0$. Fix a local $V \supseteq U$, a permutation g and a world $(X, \omega) \in \Sigma(g \cdot V)$. Let $\xi : g \cdot V \rightarrow \Xi$ be an assignment of observable characteristics defined implicitly by

$$(i, \xi(i)) \in \Omega \text{ for each } i \in g \cdot V.$$

By the Lemma, there is a set of model restrictions $\mathcal{T} \subseteq Y^{g \cdot V}$, such that

$$\log |\mathcal{T}| \leq \epsilon' |V|,$$

and for any $\theta \in \mathcal{M}$, there is $\tau_\theta \in T$, st.

$$d(\theta(i, \xi(i)), \tau_\theta(i)) \leq \frac{\varepsilon}{3}. \quad (8.4)$$

Then,

$$\begin{aligned} P_{X_\gamma} \left(R_{(X,\omega)} \left(\theta_{(X_\gamma,\omega)}^\eta \right) > \inf_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon + \eta \right) \\ \leq P_{X_\gamma} \left(\sup_{\theta \in \mathcal{M}} |R_{(X_\gamma,\omega)}(\theta) - R_{(X,\omega)}(\theta)| > \varepsilon \right) \\ \leq P_{X_\gamma} \left(\sup_{\theta \in \mathcal{M}} [|R_{(X_\gamma,\omega)}(\theta) - R_{(X_\gamma,\omega)}(\tau_\theta)| + |R_{(X_\gamma,\omega)}(\tau_\theta) - R_{(X,\omega)}(\tau_\theta)| + |R_{(X,\omega)}(\tau_\theta) - R_{(X,\omega)}(\theta)|] > \varepsilon \right) \\ \leq P_{X_\gamma} \left(\sup_{\theta \in \mathcal{M}} |R_{(X_\gamma,\omega)}(\tau_\theta) - R_{(X,\omega)}(\tau_\theta)| > \frac{\varepsilon}{3} \right), \end{aligned}$$

where the last inequality is a consequence of (8.4). Hence,

$$\begin{aligned} P_{X_\gamma} \left(R_{(X,\omega)} \left(\theta_{(X_\gamma,\omega)}^\eta \right) > \inf_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon + \eta \right) \\ \leq P_{X_\gamma} \left(\sup_{\theta \in \mathcal{M}} |R_{(X_\gamma,\omega)}(\tau_\theta) - R_{(X,\omega)}(\tau_\theta)| > \frac{\varepsilon}{3} \right) \\ \leq \sum_{\tau \in T} P_{X_\gamma} \left(|R_{(X_\gamma,\omega)}(\tau) - R_{(X,\omega)}(\tau)| > \frac{\varepsilon}{3} \right) \\ \leq |T| \max_{\tau} P_{X_\gamma} \left(|R_{(X_\gamma,\omega)}(\tau) - R_{(X,\omega)}(\tau)| > \frac{\varepsilon}{3} \right). \end{aligned}$$

The probability in the last line can be bounded using the Hoeffding inequality: for any $\tau \in Y^{g \cdot V}$,

$$P_{X_\gamma} \left(|R_{(X_\gamma,\omega)}(\tau) - R_{(X,\omega)}(\tau)| > \frac{\varepsilon}{3} \right) \leq \exp \left(-\frac{1}{18} \gamma |V| \varepsilon^2 \right).$$

Hence, for any local $V \supseteq U$, permutation $g \in G$ and world $\Omega \in \Sigma(g \cdot V)$,

$$\begin{aligned} P_{X_\gamma} \left(R_{(X,\omega)} \left(\theta_{(X_\gamma,\omega)}^\eta \right) > \inf_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon + \eta \right) \\ \leq \exp \left(|V| \left(\varepsilon' - \frac{1}{18} \gamma \varepsilon^2 \right) \right) \leq \exp \left(-\frac{1}{36} \gamma |V| \varepsilon^2 \right). \end{aligned}$$

Finally, one can finish the proof of the Theorem. Indeed, fix $\varepsilon > 0$, $\gamma > 0$, $\eta > 0$. Suppose that $S \in \dim_{(G, \frac{\varepsilon}{6})} \mathcal{M}$ is generic and find local U that satisfies the thesis of the Lemma for S , $\varepsilon' = \frac{1}{36} \gamma \varepsilon^2$ and $\delta = \frac{\varepsilon}{6}$. Take any acceptable sequence of universes (Σ_n) and let $I_1 \subseteq I_2 \subseteq \dots$ be a generating sequence that gives rise to (Σ_n) , i.e.

$$\Sigma_n \subseteq \bigcup_{\substack{V \supseteq I_n, \\ V \text{ is local}}} \bigcup_{g \in G} \Sigma(g \cdot V).$$

Since (I_n) is a generating sequence, there is n^* high enough and a permutation g^* , so that $g^* \cdot U \subseteq I_{n^*}$. Then, for any $n \geq n^*$,

$$\begin{aligned} & \sup_{(X,\omega) \in \Sigma_n} P_{X_\gamma} \left(R_{(X,\omega)} \left(\theta_{(X_\gamma,\omega)}^\eta \right) > \inf_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon + \eta \right) \\ &= \sup_{\substack{V \supseteq I_n, \\ V \text{ is local}}} \sup_{g \in G} \sup_{(X,\omega) \in \Sigma(g \cdot V)} P_{X_\gamma} \left(R_{(X,\omega)} \left(\theta_{(X_\gamma,\omega)}^\eta \right) > \inf_{\theta \in \mathcal{M}} R_{(X,\omega)}(\theta) + \varepsilon \right) \\ &\leq \sup_{\substack{V \supseteq I_n, \\ V \text{ is local}}} \sup_{g \in G} \sup_{(X,\omega) \in \Sigma(g \cdot V)} \exp \left(-\frac{1}{4} \gamma \varepsilon^2 |V| \right) \leq \exp \left(-\frac{1}{4} \gamma \varepsilon^2 |I_n| \right). \end{aligned}$$

(The first inequality is a consequence of the fact that for each local $V \supseteq I_n$, $U \subseteq (g^*)^{-1} \cdot V$ is also a local set.) Finally, the fact that $\lim_{n \rightarrow \infty} |I_n| = \infty$ completes the proof of the Theorem.

8.4. Proof of Theorem 2. Take any finite $J \subseteq I$ and assignment of observable characteristics $\xi : J \rightarrow \Xi$ and let $X = \{i, \xi(i) : i \in J\}$. For any two worlds $(X, \omega), (X, \omega') \in \Sigma(J)$, if $\omega(i, \xi(i)) = \omega'(i, \xi(i))$ for each $i \in J$, then, for any subset $X' \subseteq X$, any model θ

$$R_{(X,\omega)}(\theta) = R_{(X,\omega')}(\theta).$$

Take any $\omega^J \in \{0, 1\}^J$ and define

$$R_{(X,\omega^J)} := R_{(X,\omega)},$$

for some $\omega \in Y^{\mathbf{X}}$, such that $\omega^J(i) = \omega(i, \xi(i))$ for any $i \in J$. By the above, this is well-defined and $R_{(X,\omega^J)}$ does not depend on the choice of ω .

For any finite sets of indices $H \subseteq J \subseteq I$, for any world $(X, \omega) \in \Sigma(J)$, let $X_H = \{(i, \xi) \in X : i \in H\}$ be the sample of instances indexed with H . Let θ_H denote the empirical risk minimizer on (X_H, ω) :

$$\theta_H \in \arg \min_{\theta \in \mathcal{M}} R_{(X_H,\omega)}(\theta).$$

As in (2.1), there might be many empirical risk minimizers. Any choice of the minimizer is good as long as it does not depend on the realization of outcomes outside sample Ω_H .

Lemma 4. For any finite $J \subseteq I$, any $A, B \subseteq J$, $A \cap B = \emptyset$, any world $(X, \omega) \in \Sigma(J)$, any $\varepsilon > 0$

$$2^{-|J|} \sum_{\omega^J \in \{0,1\}^J} \mathbf{1} \left\{ \left| R_{(X_A,\omega^J)}(\theta_B) - \frac{1}{2} \right| > \varepsilon \right\} \leq \exp \left(-\frac{1}{2} |A| \varepsilon^2 \right)$$

Proof. Let μ denote an uniform distribution on set $\{0, 1\}^J$, i.e. a distribution that assigns equal probability to any $\omega^J : J \rightarrow \{0, 1\}$. The thesis of the Lemma bounds from above the probability that the sample risk on (X_A, ω^J) of the empirical minimizer θ_B is ε -far to $\frac{1}{2}$ when ω^J is chosen according to the distribution μ ,

$$\mu \left(\left| R_{(X_A, \omega^J)}(\theta_B) - \frac{1}{2} \right| > \varepsilon \right).$$

Note that distribution μ chooses outcome of each $i \in J$ as equal to 0 independently and with probability $\frac{1}{2}$. Because the choice of θ_B depends only on the realization of outcomes over indices B , and these outcomes are independent from the realization of outcomes over A , one has

$$\begin{aligned} & \mu \left(\left| R_{(X_A, \omega^J)}(\theta_B) - \frac{1}{2} \right| > \varepsilon \right) \\ &= \mu \left(\left| R_{(X_A, \omega^J)}(\theta) - \frac{1}{2} \right| > \varepsilon \mid \theta = \theta_B \right) \\ &= \mu \left(\left| R_{(X_A, \omega^J)}(\theta) - \frac{1}{2} \right| > \varepsilon \right) \text{ for any model } \theta \in \mathcal{M}. \end{aligned}$$

The thesis of the Lemma follows from the Hoeffding inequality. \square

Lemma 5. *For any $\delta > 0$, there is $c_\delta > 0$, such that the following holds. Suppose that $S \subseteq J \subseteq I$ and $\xi : J \rightarrow \Xi$ are such that $|S| \geq \delta |I|$ and $|\mathcal{M}_{S, \xi|_S}| = 2^{|S|}$. Then, for any $H \subseteq J$, $|H| \leq \frac{\delta}{2} |J|$, any world $(X, \omega) \in \Sigma(J)$,*

$$2^{-|J|} \sum_{\omega^J \in \{0, 1\}^J} \mathbf{1} \left\{ \left| R_{(X, \omega^J)}(\theta_H) - \min_{\theta \in \mathcal{M}} R_{(X, \omega^J)}(\theta) \right| < \frac{\delta}{12} \right\} \leq \exp(-|J| c_\delta).$$

Proof. As in the previous Lemma, let μ denote an uniform distribution on $\{0, 1\}^J$. The thesis of the Lemma ask to bound from above the probability that the true risk of the empirical minimizer θ_H is $\frac{\delta}{12}$ -close to the true minimal risk when ω^J is chosen from distribution μ ,

$$\mu \left(\left| R_{(X, \omega^J)}(\theta_H) - \min_{\theta \in \mathcal{M}} R_{(X, \omega^J)}(\theta) \right| < \frac{\delta}{12} \right).$$

Observe that

$$\begin{aligned} & \mu \left(\left| R_{(X, \omega^J)}(\theta_H) - \min_{\theta \in \mathcal{M}} R_{(X, \omega^J)}(\theta) \right| \geq \frac{\delta}{12} \right) \\ & \geq \mu \left(R_{(X, \omega^J)}(\theta_H) \geq \frac{1}{2} \left(1 - \frac{\delta}{2} \right) - \frac{\delta}{12} \right) \\ & + \mu \left(\min_{\theta \in \mathcal{M}} R_{(X, \omega^J)}(\theta) \leq \frac{1}{2} (1 - \delta) + \frac{\delta}{12} \right). \end{aligned}$$

I bound both probabilities separately. First, notice that

$$R_{(X,\omega^J)}(\theta_H) \geq \left(1 - \frac{\delta}{2}\right) R_{(X_{J/H},\omega^J)}(\theta_H)$$

and, by Lemma 4,

$$\begin{aligned} \mu \left(R_{(X,\omega^J)}(\theta_H) \geq \frac{1}{2} \left(1 - \frac{\delta}{2}\right) - \frac{\delta}{12} \right) \\ \geq \mu \left(R_{(X_{J/H},\omega^J)}(\theta_H) \geq \frac{1}{2} - \frac{\delta}{6(2-\delta)} \right) \\ \geq 1 - \exp(-|J|c_\delta^1), \end{aligned}$$

where

$$c_\delta^1 = \frac{1}{2} \left(1 - \frac{\delta}{2}\right) \left(\frac{\delta}{6(2-\delta)}\right)^2.$$

Second, observe that $R_{\Omega_S}(\theta_S) = 0$ due to the fact that $|\mathcal{M}_{S,\xi|_S}| = 2^{|S|}$. Hence,

$$\begin{aligned} \min_{\theta \in \mathcal{M}} R_{(X,\omega^J)}(\theta) &\leq \frac{|S|}{|J|} R_{(X_S,\omega^J)}(\theta) + \left(1 - \frac{|S|}{|J|}\right) R_{(X_{J \setminus S},\omega^J)}(\theta_S) \\ &\leq (1 - \delta) R_{\Omega_{J \setminus S}}(\theta_S). \end{aligned}$$

By Lemma 4,

$$\begin{aligned} \mu \left(\min_{\theta \in \mathcal{M}} R_{(X,\omega^J)}(\theta) \leq \frac{1}{2} (1 - \delta) + \frac{\delta}{12} \right) \\ \geq \mu \left(R_{(X_{J \setminus S},\omega^J)}(\theta_S) \leq \frac{1}{2} + \frac{\delta}{12(1-\delta)} \right) \\ \geq 1 - \exp(-|J|c_\delta^2), \end{aligned}$$

where

$$c_\delta^2 = \frac{1}{2} (1 - \delta) \left(\frac{\delta}{12(1-\delta)}\right)^2.$$

Finally,

$$\begin{aligned} \mu \left(\left| R_{(X,\omega^J)}(\theta_H) - \min_{\theta \in \mathcal{M}} R_{(X,\omega^J)}(\theta) \right| < \frac{\delta}{12} \right) \\ \leq 1 - (1 - \exp(-|J|c_\delta^1)) - (1 - \exp(-|J|c_\delta^2)) \\ \leq \exp(-|J|c_\delta^1) + \exp(-|J|c_\delta^2) - 1 \\ \leq \exp(-|J|c_\delta), \end{aligned}$$

where $c_\delta = \max(c_\delta^1, c_\delta^2)$. □

Because $G \mapsto X$ is tight, there is $\delta^* > 0$, such that any finite $J \subseteq I$ contains a subset $S \subseteq J$ that is generic and $|S| \geq \delta^* |J|$. Take

$$\delta = \min(12\varepsilon, \delta^*) > 0 \text{ and } \gamma_\varepsilon = \frac{\delta}{2}.$$

Because $S \notin \dim_G \mathcal{M}$, there is a permutation $g \in G$, and $\xi : g \cdot J \rightarrow \Xi$, such that

$$|\mathcal{M}_{g \cdot S, \xi|_{g \cdot S}}| = 2^{|S|}.$$

By Lemma 5, there is $c_\delta > 0$, such that for any $\gamma \leq \gamma_\varepsilon$,

$$\begin{aligned} & \binom{|J|}{\lceil \gamma |J| \rceil}^{-1} 2^{-|J|} \sum_{H \subseteq J: |H| = \lceil \gamma |J| \rceil} \sum_{\omega^J \in \{0,1\}^J} \mathbf{1} \left\{ \left| R_{(X, \omega^J)}(\theta_H) - \min_{\theta \in \mathcal{M}} R_{(X, \omega^J)}(\theta) \right| < \frac{\delta}{12} \right\} \\ & \leq \exp(-|J| c_\delta), \end{aligned}$$

where $\binom{|J|}{\lceil \gamma |J| \rceil}$ is the number of all subsets $H \subseteq J$ of size $\lceil \gamma |J| \rceil$. Therefore, there is $\omega^J \in \{0,1\}^J$, such that

$$\binom{|J|}{\lceil \gamma |J| \rceil}^{-1} \sum_{H \subseteq J: |H| = \lceil \gamma |J| \rceil} \mathbf{1} \left\{ \left| R_{(X, \omega^J)}(\theta_H) - \min_{\theta \in \mathcal{M}} R_{(X, \omega^J)}(\theta) \right| < \frac{\delta}{12} \right\} \leq \exp(-|J| c_\delta).$$

This ends the proof of the Theorem.

8.5. Proof of Theorem 4. Because $G \mapsto I$ is tight, there is $\delta^* > 0$, such that for any finite $A \subseteq I$, there is a generic $S \subseteq A$, $|S| \geq \delta^* |A|$. Define

$$\gamma^* = \frac{\min(12\varepsilon^*, \delta^*)}{2} \text{ and } c^* = \frac{1}{4} \gamma^* (\varepsilon^*)^2.$$

Lemma 6. *Suppose that for some $\delta > 0$, (G, δ) -dimension of \mathcal{M} does not contain any generic set. Then, for any local set U , there is an outcome $y^U \in Y$, such that for any $J \subseteq U$, there is a permutation $g \in G$, a subset $S \subseteq g \cdot J$, $|S| \geq \delta^* |J|$, an assignment $\xi : S \rightarrow \Xi$, such that for any subset $S' \subseteq S$, there is a model $\theta \in \mathcal{M}$, so that*

$$\begin{aligned} d((\theta(i, \xi(i)), y^*), y^*) &\leq \delta \text{ for any } i \in S' \text{ and} \\ d(\theta(i, \xi(i)), y^*) &\geq 2\delta \text{ for any } i \in S \setminus S'. \end{aligned} \tag{8.5}$$

Proof. Let $S_U \subseteq U$ be a generic set, such that $|S_U| \geq \delta^* |U|$. Because $S_U \notin \dim_{(G, \delta)} \mathcal{M}$, there is an outcome $y^U \in Y$, a permutation $g_U \in G$ and an assignment $\xi : g_U \cdot S_U \rightarrow \Xi$, such that for any $S' \subseteq S_U$, there is a model $\theta \in \mathcal{M}$, so that (8.5) holds.

By Lemma 8 in Appendix A, for any finite $J \subseteq U$, there is a permutation $g_J \in G$, such that

$$\frac{|g_J \cdot S_U \cap J|}{|U|} \geq \frac{|S_U|}{|U|} \frac{|J|}{|U|}.$$

Denote $S_J = g \cdot S_U \cap J$. Then, $|S_J| \geq \delta^* |J|$. The thesis of the Lemma is satisfied for permutation $g = g_U \circ (g_J)^{-1}$, subset $S = g \cdot S_J$ and assignment $\xi|_S$. \square

Lemma 7. *Suppose that for some $\delta > 0$, (G, δ) -dimension of \mathcal{M} does not contain any generic set. Then, there is outcome $y^* \in Y$, such that for any finite $J \subseteq I$, there is a permutation $g \in G$, a subset $S \subseteq g \cdot J$, $|S| \geq \delta^* |J|$, an assignment $\xi : S \rightarrow \Xi$, such that for any subset $S' \subseteq S$, there is a model $\theta \in \mathcal{M}$, so that*

$$\begin{aligned} d((\theta(i, \xi(i)), y^*), y^*) &\leq \frac{4}{3}\delta \text{ for any } i \in S' \text{ and} \\ d(\theta(i, \xi(i)), y^*) &\geq \frac{5}{3}\delta \text{ for any } i \in S \setminus S'. \end{aligned} \quad (8.6)$$

Proof. Y is compact, hence it is possible to find finite set $Y^* \subseteq Y$, such that for any $y \in Y$, there is $y' \in Y^*$, so that $d(y, y') \leq \frac{\delta}{3}$. By Lemma 6, for any local set, there is $y_U^* \in Y^*$, such that for any $J \subseteq U$, there is a permutation $g \in G$, a subset $S \subseteq g \cdot J$, $|S| \geq \delta^* |J|$, an assignment $\xi : S \rightarrow \Xi$, such that for any subset $S' \subseteq S$, there is a model $\theta \in \mathcal{M}$, so that (8.6) holds.

Take any generating sequence $I_1 \subseteq I_2 \subseteq \dots$ and find outcomes with the above properties. Because $Y^* = Y$ is finite, there is $y^* \in Y^*$, such that $y^* = y_{I_n}^*$ for infinitely many n . Let $N^* = \{n : y^* = y_{I_n}^*\}$.

Take any finite $J \subseteq I$. Then, there is $n \in N^*$, such that $J \subseteq g \cdot I_n$ for some $g \in G$. This ends the proof of the result. \square

Suppose that for some $\delta > 0$, (G, δ) -dimension of \mathcal{M} does not contain any generic set and let y^* be as in the Lemma. Take a loss function l , such that

$$\begin{aligned} l(y, y') &= 1 \text{ if } d(y, y^*) \leq \frac{4}{3}\delta \text{ and } d(y', y^*) \leq \frac{4}{3}\delta, \\ l(y, y') &= 0 \text{ if } d(y, y^*) \leq \frac{4}{3}\delta \text{ and } d(y', y^*) \geq \frac{5}{3}\delta, \\ l(y, y') &= 0 \text{ if } d(y, y^*) \geq \frac{5}{3}\delta \text{ and } d(y', y^*) \leq \frac{4}{3}\delta, \\ l(y, y') &= 1 \text{ if } d(y, y^*) \geq \frac{5}{3}\delta \text{ and } d(y', y^*) \geq \frac{5}{3}\delta. \end{aligned}$$

Let $R_{(X, \omega)}^{(l)}$ denotes the true risk of model θ with respect to loss function l .

Take any finite $J \subseteq I$ and find a permutation $g \in J$, a subset $S \subseteq g \cdot J$, $|S| \geq \delta^* |J|$, an assignment $\xi : S \rightarrow \Xi$, such that for any subset $S' \subseteq S$, there is a model $\theta \in \mathcal{M}$, so

that (8.6) holds. Pick any $y^{**} \in Y$, $d(y^{**}, y^*) \geq \frac{5}{3}\delta$. For any $S' \subseteq S$, pick $\omega^{S'} \in \{0, 1\}^J$ such that,

$$\begin{aligned}\omega^{S'}(i) &= y^* \text{ for each } i \in S' \text{ and} \\ \omega^{S'}(i) &= y^{**} \text{ for each } i \notin S \setminus S' .\end{aligned}$$

Let

$$\Omega^* = \left\{ \omega^{S'} : S' \subseteq S \right\} \subseteq \{0, 1\}^J .$$

Then, $|\Omega^*| = 2^{|S|}$.

CHECK IT! The rest of the proof follows the same lines as the proof of Theorem 2. By Lemma 5, there is $c_{\delta^*} > 0$, such that for any $\gamma \leq \gamma_\varepsilon$,

$$\begin{aligned}& \left(\frac{|J|}{\lceil \gamma |J| \rceil} \right)^{-1} \frac{1}{|\Omega^*|} \sum_{H \subseteq J: |H| = \lceil \gamma |J| \rceil} \sum_{\omega \in \Omega^*} \mathbf{1} \left\{ \left| R_{(X, \omega)}(\theta_H) - \min_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) \right| < \frac{\delta}{12} \right\} \\ & \leq \exp(-|J| c_\delta) .\end{aligned}$$

Therefore, there is $\omega \in \Omega^*$, such that

$$\left(\frac{|J|}{\lceil \gamma |J| \rceil} \right)^{-1} \sum_{H \subseteq J: |H| = \lceil \gamma |J| \rceil} \mathbf{1} \left\{ \left| R_{(X, \omega)}(\theta_H) - \min_{\theta \in \mathcal{M}} R_{(X, \omega)}(\theta) \right| < \frac{\delta}{12} \right\} \leq \exp(-|J| c_\delta) .$$

This ends the proof of the Theorem.

REFERENCES

- AL-NAJJAR, N. (2006): “Decision Makers as Statisticians: Diversity, Ambiguity and Belief Formation,” Kellogg School of Management, Northwestern University. 4
- ALON, N., S. BEN-DAVID, N. CESA-BIANCHI, AND D. HAUSSLER (1997): “Scale-Sensitive Dimensions, Uniform Convergence, and Learnability,” *Journal of the ACM*, 44, 615–631. 4, 22
- ALON, N., AND A. SHAPIRA (2005): “Every Monotone Graph Property is Testable,” *Proc. of the 37 ACM STOC*, pp. 128–137. 12
- ANSTEE, R. (2006): “A Survey of Forbidden Configuration Results,” preprint. 4
- ANSTEE, R., AND Z. FUREDI (1986): “Forbidden Submatrices,” *Discrete Mathematics*, 62, 225–243. 4
- BLACKWELL, D., AND A. GIRSHICK (1954): *Theory of Games and Statistical Decisions*. Wiley, New York. 5
- BOUCHERON, S., O. BOUSQUET, AND G. LUGOSI (2005): “Theory of Classification: A Survey of Recent Advances,” *ESAIM: Probability and Statistics*, p. forthcoming. 4

- BOUSQUET, O., S. BOUCHERON, AND G. LUGOSI (2004): “Introduction to Statistical Learning Theory,” in *Advanced Lectures in Machine Learning*, ed. by O. Bousquet, U. Luxburg, and G. Rätsch, pp. 169–207. Springer. 4
- CESA-BIANCHI, N., AND D. HAUSSLER (1998): “A Graph-Theoretic Generalization of the Sauer-Shelah Lemma,” *Discrete Applied Mathematics*, 86, 27 – 35. 4
- COPIC, J., A. KIRMAN, AND M. JACKSON (2006): “Identifying Community Structures from Network Data,” <http://www.stanford.edu/~jacksonm/netcommunity.pdf>. 12
- DEVROYE, L., L. GYÖRFI, AND G. LUGOSI (1996): *A Probabilistic Theory of Pattern Recognition*. Springer, New York. 2
- DUDLEY, R. M. (1999): *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics. Cambridge University Press, New York. 4
- HAUSSLER, D. (1992): “Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications,” *Information and Computation*, 100, 78 – 150. 4
- HAUSSLER, D., AND P. M. LONG (1995): “A Generalization of Sauer’s Lemma,” *Journal of Combinatorial Theory*, 71(2), 219–240. 4
- HITSCH, G. J., A. HORTACSU, AND D. ARIELY (2006): “What Makes You Click? Mate Preferences and Matching Outcomes in Online Dating,” MIT Sloan Research Paper No. 4603-06. 12
- HOEFFDING, W. (1963): “Probability Inequalities for Sums of Bounded Random Variables,” *Journal of the American Statistical Association*, 58, 13–30. 20
- KALAI, G. (2003): “Learnability and Rationality of Choice,” *Journal of Economic Theory*, 113, 104–117. 13
- KEARNS, M. J., AND R. E. SCHAPIRE (1994): “Efficient Distribution-Free Learning of Probabilistic Concepts,” *Journal of Computer and System Sciences*, 48, 464 – 497. 4
- LANG, S. (2002): *Algebra*, Graduate Texts in Mathematics. Springer, New York. 9
- MANSKI, C. F. (2004): “Statistical Treatment Rules for Heterogeneous Populations,” *Econometrica*, 72, 1221–1246. 5
- MENDELSON, S., AND R. VERSHYNIN (2003): “Entropy and the Combinatorial Dimension,” *Inventiones Mathematicae*, 152, 37–55. 4
- NEWMAN, M. E. J. (2004): “Detecting Community Structure in Networks,” *The European Physical Journal B - Condensed Matter and Complex Systems*, 38. 12
- (2006): “Modularity and Community Structure in Networks,” *Proceedings of National Academy of Science USA*, 103, 8577–8582. 12
- PAKES, A., AND D. POLLARD (1989): “Simulation and the Asymptotics of Optimization Estimators,” *Econometrica*, 57, 1027–1057. 4
- POLLARD, D. (1985): “New Ways to Prove Central Limit Theorems,” *Econometric Theory*, 1, 295–314. 4
- SAUER, N. (1972): “On the Density of Families of Sets,” *Journal of Combinatorial Theory (A)*, 13, 145–147. 19
- SCOTT, J. (2000): *Social Network Analysis: A Handbook*. Sage, London. 12
- SHELAH, S. (1972): “A Combinatorial Problem: Stability and Order for Models and Theories in Infinitary Languages,” *Pacific Journal of Mathematics*, 41, 247–261. 19

- TALAGRAND, M. (2003): “Vapnik-Chervonenkis Type Conditions and Uniform Donsker Classes of Functions,” *Annals of Probability*, 31, 1565–1582. [4](#)
- VAPNIK, V. N. (1998): *Statistical Learning Theory*. Wiley-Interscience. [2](#), [3](#)
- VAPNIK, V. N., AND A. Y. CHERVONENKIS (1971): “On the Uniform Convergence of Relative Frequencies of Events to their Probabilities,” *Theory of Probability and Applications*, 16, 264–280. [3](#), [4](#), [19](#)
- WALD, A. (1950): *Statistical Decision Functions*. John Wiley and Sons, London. [5](#)

APPENDIX A. TRANSITIVE GROUP ACTION

This part of the Appendix reminds some mixing properties of a transitive group action. Recall that group action $G \curvearrowright A$ is *transitive* if for any $a, a' \in A$, there is $g \in G$, such that $g \cdot a = a'$. Assume here that A is finite and $G \curvearrowright A$ is a transitive group action. W.l.o.g. assume that $G \subseteq \Pi_A$, hence G is also finite. For any function $f : A \rightarrow \mathbb{R}$, define

$$Ef := \frac{1}{|A|} \sum_{a \in A} f(a). \quad (\text{A.1})$$

Lemma 8. *For any two functions $f, h : A \rightarrow \mathbb{R}$,*

$$\frac{1}{|G|} \sum_g Ef(\cdot) h(g \cdot \cdot) = EfEh.$$

Proof. Compute

$$\begin{aligned} & \frac{1}{|G|} \sum_g Ef(\cdot) h(g \cdot \cdot) \\ &= \frac{1}{|A|} \sum_{a \in A} \frac{1}{|G|} \sum_g f(a) h(g \cdot a) \\ &= \frac{1}{|A|} \sum_{a \in A} \sum_{b \in A} \left[\frac{1}{|G|} \sum_{g \in G} \mathbf{1}\{g \cdot a = b\} \right] f(a) h(b). \end{aligned}$$

Since G is transitive, for any two $a, a', b \in A$,

$$\sum_{g \in G} \mathbf{1}\{g \cdot a = b\} = \sum_{g \in G} \mathbf{1}\{g \cdot a' = b\}.$$

Moreover, by the definition of permutation, if $g \cdot a = b$, then $g \cdot a' \neq b$ and for any $g \in G$, there is $a \in A$, such that $g \cdot a = b$. This implies that for any $a, b \in A$,

$$\sum_{g \in G} \mathbf{1}\{g \cdot a = b\} = \frac{|G|}{|A|}.$$

Hence,

$$\frac{1}{|G|} \sum_g E f(\cdot) h(g \cdot) = \frac{1}{|A|} \sum_{a \in A} f(a) \frac{1}{|A|} \sum_{b \in S} h(b) = E f E h.$$

□

Lemma 9. *For any two subsets $B, C \subseteq A$, there is a permutation $g \in G$, such that*

$$\frac{|B \cap g \cdot C|}{|B|} \leq \frac{|C|}{|A|}. \quad (\text{A.2})$$

For any two subsets $B, C \subseteq A$, for any function $f : A \rightarrow R$, there is $g \in G$, such that

$$E \mathbf{1}_C f(g \cdot) \geq \frac{1}{2} \frac{|C|}{|A|} E f \text{ and } E \mathbf{1}_B f(g \cdot) \leq 3 \frac{|B|}{|C|} E \mathbf{1}_C f(g \cdot). \quad (\text{A.3})$$

Proof. The above Lemma says that

$$\frac{|B|}{|A|} \frac{|C|}{|A|} = \frac{1}{|G|} \sum_{g \in G} \frac{|B \cap g \cdot C|}{|A|}.$$

Hence, there must be at least on $g \in G$, such that (A.2) holds.

he first part is standard. For the second part, for any set $S \subseteq A$, define $m_S : G \rightarrow [0, 1]$ as

$$m_S(g) = E \mathbf{1}_S f(g \cdot) \text{ and } m_S = \frac{1}{|G|} \sum_{g \in G} m_S(g).$$

Define sets

$$G_C = \left\{ g : m_C(g) \geq \frac{1}{2} m_C \right\} \text{ and } G_B = \left\{ g : m_B(g) \geq 3 \frac{m_B}{m_C} m_C(g) \right\}.$$

If the thesis of the lemma is not true, then $G_C \subseteq G_B$. But this leads to a contradiction:

$$\begin{aligned} m_B &\geq \frac{1}{|G|} \sum_{g \in G_B} m_B(g) \geq 3 \frac{1}{|G|} \sum_{g \in G_B} \frac{m_B}{m_C} m_C(g) \geq 3 \frac{m_B}{m_C} \frac{1}{|G|} \sum_{g \in G_C} m_C(g) \\ &= 3 \frac{m_B}{m_C} \left[m_C - \frac{1}{|G|} \sum_{g \notin G_C} m_C(g) \right] \geq \frac{3}{2} m_B. \end{aligned}$$

□

APPENDIX B. PROOF OF PROPOSITION 5

The proof of the Proposition is divided into two parts. In the next part of this Appendix, properties of generic sets are discussed. For any local U and generic S , I define a constant $a(U, S)$. I show that for any generic S , any $\epsilon > 0$, there is a local U , such that for any local $V \supseteq U$, any permutation $g \in G$,

$$a(g \cdot V, S) \leq \epsilon. \quad (\text{B.1})$$

In the remaining parts of this Appendix, I show a Lemma:

Lemma 10. *For any local U , any generic $S \subseteq U$, any $\mathcal{M} \subseteq \{0, 1, *\}^I$, such that $S \in \dim_G^* \mathcal{M}$, there is $\mathcal{M}_U^* \in \{0, 1\}^U$ that approximates \mathcal{M} on U up to $*$ and*

$$\log |\mathcal{M}_U^*| \leq |U| a(U, S).$$

B.1. Generic sets. This part of the Appendix describes some useful properties of generic sets. Let $G \mapsto I$ be a locally generated and transitive group action.

Let I^k be a set of k -tuples of indices i . A typical element of I^k is denoted as $\bar{i} = (i_1, \dots, i_k) \in I^k$. $G \mapsto I$ induces a group action on I^k : for any $g \in G$, any $\bar{i} \in I^k$, let

$$g \cdot \bar{i} = (g \cdot i_1, \dots, g \cdot i_k).$$

For any finite $U \subseteq I$, any tuple $\bar{i} \in I^k$, define set of those permutations of tuple \bar{i} that are contained in set U :

$$A_U(\bar{i}) = \{g \cdot \bar{i} : g \in G \text{ and } \{i_1, \dots, i_k\} \subseteq U\}. \quad (\text{B.2})$$

For any ε , any finite $U \subseteq I$, define

$$\delta(U, \bar{i}, \varepsilon) := \inf_{D \subseteq U : |D| \geq \varepsilon |U|} \frac{|A_D(\bar{i})|}{|A_U(\bar{i})|}. \quad (\text{B.3})$$

Lemma 11. *For any local sets $U \subseteq V \subseteq I$, any $\bar{i} \in I^k$, any $\varepsilon > 0$,*

$$\delta(V, \bar{i}, \sqrt{\varepsilon}) \geq \frac{1}{2} \sqrt{\varepsilon} \delta(U, \bar{i}, \varepsilon).$$

Proof. Let

$$G_V = \{g \in G : g \cdot V = V\}.$$

Because V is local and $G \mapsto I$ is transitive, $G_V \mapsto V$ is a transitive group action. By Lemma 8, for any subsets $D \subseteq V$,

$$\frac{|D|}{|V|} = \frac{1}{|G_V|} \sum_{g \in G_V} \frac{|D \cap g \cdot U|}{|U|}.$$

For any $D \subseteq V$, define

$$\alpha(D) = \frac{|\{g \in G_V : |D \cap g \cdot U| \geq \varepsilon |U|\}|}{|G_V|}.$$

Suppose that $|D| \geq \sqrt{\varepsilon} |V|$. Then, by (B.4),

$$\sqrt{\varepsilon} \leq \frac{|D|}{|V|} = \frac{1}{|G_V|} \sum_{g \in G_V} \frac{|D \cap g \cdot U|}{|U|} \leq \alpha(D) + \varepsilon(1 - \alpha(D));$$

hence,

$$\alpha(D) \geq \frac{\sqrt{\varepsilon} - \varepsilon}{1 - \varepsilon} = \sqrt{\varepsilon} \frac{1}{1 + \sqrt{\varepsilon}} \geq \frac{1}{2} \sqrt{\varepsilon}$$

Because V is local, group action $G_V \mapsto V$ induces transitive group action $G_V \mapsto A_V(\bar{i})$. Take any $D \subseteq V$, such that $|D| \geq \sqrt{\varepsilon}|V|$. By Lemma 8,

$$\frac{|A_D(\bar{i})| |A_U(\bar{i})|}{|A_V(\bar{i})| |A_V(\bar{i})|} = \frac{1}{|G|} \sum_{g \in G_V} \frac{|A_{D \cap g \cdot U}(\bar{i})|}{|A_V(\bar{i})|}.$$

By definition, for any set $D \subseteq I$, such that $|D \cap g \cdot U| \geq \varepsilon|U|$,

$$\frac{|A_{D \cap g \cdot U}(\bar{i})|}{|A_U(\bar{i})|} \geq \delta(U, \bar{i}, \varepsilon). \quad (\text{B.4})$$

Hence,

$$\frac{|A_D(\bar{i})|}{|A_V(\bar{i})|} = \frac{1}{|G|} \sum_{g \in G_V} \frac{|A_{D \cap g \cdot U}(\bar{i})|}{|A_U(\bar{i})|} \geq \alpha(D) \delta(U, \bar{i}, \varepsilon) \geq \frac{1}{2} \sqrt{\varepsilon} \delta(U, \bar{i}, \varepsilon).$$

□

Take any local set $U \subseteq I$ and a tuple $\bar{i} \in I^k$. Let $\gamma^l(U, \bar{i}, \varepsilon) \in [0, 1]$, $0 \leq l \leq k$ be a finite sequence of positive constants, such that

$$\gamma^0(U, \bar{i}, \varepsilon) = \delta(U, \bar{i}, \varepsilon)$$

and for each $1 \leq l \leq k$

$$\begin{aligned} \gamma^{l+1}(U, \bar{i}, \varepsilon) &= (\gamma^l(U, \bar{i}, \varepsilon))^2, \\ \gamma^*(U, \bar{i}, \varepsilon) &= \frac{1}{4k} (1 - \varepsilon) \gamma^k(U, \bar{i}, \varepsilon). \end{aligned} \quad (\text{B.5})$$

For any finite $S \subseteq I$, $|S| = k$, an *enumeration* of S is a tuple $\bar{i} = (i_1, \dots, i_k)$, such that $\{i_1, \dots, i_k\} = S$. There is $k!$ different enumerations of S . Notice that for any set S , if \bar{i}, \bar{i}' are two different enumerations of S , then $\delta(U, \bar{i}, \varepsilon) = \delta(U, \bar{i}', \varepsilon)$.

Corollary 3. *Suppose that S is generic. Then, for any $\varepsilon > 0$, there is a local set U , such that for any enumeration \bar{i} of S ,*

$$\inf_{\substack{V \supseteq U, \\ V \text{ is local}}} \inf_{g \in G} \delta(g \cdot V, \bar{i}, \varepsilon) > 0 \text{ and } \inf_{\substack{V \supseteq U, \\ V \text{ is local}}} \inf_{g \in G} \gamma^*(g \cdot V, \bar{i}, \varepsilon) > 0.$$

Proof. Fix $\varepsilon > 0$ and find local U , such that $\delta(U, \bar{i}, \varepsilon^2) > 0$ for any enumeration of S . Such local set exists, because S is generic. The result follows from Lemma 11. □

Lemma 12. *Take any local U , finite $S \subseteq I$ and enumeration \bar{v} of S , such that $\delta(U, \bar{v}, \varepsilon) > 0$ for some $\varepsilon \in (0, \frac{1}{2})$. Then, there are $i^* \in S$ and subsets $W, T \subseteq U$, such that*

$$|W| \leq \varepsilon |U|, \quad |T| \geq \frac{1}{2|S|}$$

and for any $i \in T$, there is a permutation $g \in G$, such that $g \cdot i^* = i$ and $g \cdot (S \setminus \{i^*\}) \subseteq W$.

Proof. Let $W \subseteq U$ be a maximal set among those that do not contain any permutation of S :

- (a) for any permutation $g \in G$, $g \cdot S \subseteq W$ and
- (b) for any $i \in U \setminus W$, there is a $g \in G$, such that $g \cdot S \subseteq W \cup \{i\}$.

There is at least one such a set and $|W| \leq \varepsilon |U|$ because $\delta(U, \bar{v}, \varepsilon) > 0$. For any $i^* \in S$, define sets $T_{i^*} \subseteq U$, such that for any $i \in T_{i^*}$, there is a permutation $g \in G$, such that $g \cdot i^* = i$ and $g \cdot (S \setminus \{i^*\}) \subseteq W$. Then, $\bigcup_{i^*} T_{i^*} = U \setminus W$ and there is $i^* \in S$, such that

$$|T_{i^*}| \geq \frac{1}{|S|} \frac{\varepsilon}{1 - \varepsilon} |U| \geq \frac{1}{2|S|}.$$

□

Take any finite $W \subseteq U \subseteq I$ and tuple $\bar{v} \in I^k$. For any $l \leq k$, consider set $T^l(U, W, \bar{v}) \subseteq U$ that consists of instances $g \cdot i_l$, where permutation g is such that $g \cdot \{i_1, \dots, i_{l-1}\} \subseteq W$ and $g \cdot \{i_l, \dots, i_k\} \in U$:

$$T^l(U, W, \bar{v}) := \{g \cdot i_l : g \in G \text{ and } g \cdot \{i_1, \dots, i_{l-1}\} \subseteq W \text{ and } g \cdot \{i_l, \dots, i_k\} \in U\}.$$

Define also

$$\omega^l(U, \bar{v}) = \inf_{W \subseteq U: |T^l(U, W, \bar{v})| \geq \frac{1}{2|S|}|U|} \frac{|W|}{|U|}. \quad (\text{B.6})$$

Then, Lemma 12 implies that if $\delta(U, \bar{v}, \varepsilon) > 0$ for some $\varepsilon \in (0, \frac{1}{2})$, then there is an enumeration \bar{v}^* , such that $\omega^k(U, \bar{v}) \leq \varepsilon$. Indeed, enumeration $\bar{v}^* = (i_1^*, \dots, i_k^*)$, should be chosen such that $i_k^* = i^*$. The next Lemma strengthens this observation.

Lemma 13. *Take any local $U \subseteq I$ and finite $S \subseteq I$, $|S| = k$, such that $\delta(U, \bar{v}, \varepsilon) > 0$ for some $\varepsilon \in (0, \frac{1}{2})$ and enumeration \bar{v} of S . There is an enumeration \bar{v}^* of S , such that for any $l \leq k$,*

$$\omega^l(U, \bar{v}^*) \leq \varepsilon.$$

Proof. This is a corollary to Lemma 12 and the fact that if $\delta(U, \bar{v}, \varepsilon) > 0$, then $\delta(U, \bar{v}', \varepsilon) > 0$ for any l -subtuple $\bar{v}' = (i_1, \dots, i_l)$ of $\bar{v} = (i_1, \dots, i_k)$, $l \leq k$. Indeed, let \bar{v}^k be any enumeration of S . By Lemma 12, there is an enumeration $\bar{v}^{k*} = (i_1^{k*}, \dots, i_{k-1}^{k*}, i_k^*)$, such

that

$$\omega^k (U, \bar{i}^{k*}) \leq \varepsilon.$$

Next, consider set $S^{k-1} = S \setminus \{i_k^*\}$. For any enumeration \bar{i}^{k-1} of S^{k-1} , $\delta(U, \bar{i}^{k-1}, \varepsilon) > 0$.

By Lemma 12, there is an enumeration $\bar{i}^{(k-1)*} = (i_1^{(k-1)*}, \dots, i_{k-2}^{(k-1)*}, i_{k-1}^*)$, such that

$$\omega^{k-1} \left(U, \left(i_1^{(k-1)*}, \dots, i_{k-2}^{(k-1)*}, i_{k-1}^*, i_k^* \right) \right) = \omega^{k-1} (U, \bar{i}^{(k-1)*}) \leq \varepsilon.$$

A repetition of this argument for $k-2, k-3, \dots, 1$, yields the Lemma. \square

Define entropy function: for any $t \in (0, 1)$

$$H(t) = t \log \frac{1}{t} + (1-t) \log \frac{1}{1-t}. \quad (\text{B.7})$$

Take any local set $U \subseteq I$ and generic $S \subseteq I$, $|S| = k$. Define

$$a(U, S) :=$$

$$\inf_{\varepsilon > 0} \inf_{\bar{i}^* \text{ is enumeration of } S} \left(H(2\varepsilon) - \frac{\log \gamma^*(U, \bar{i}^*, \varepsilon)}{|U|} + \frac{\log k}{|U| \gamma^*(U, \bar{i}^*, \varepsilon)} + \frac{\max_{l \leq k} [\omega^l(U, \bar{i}^*) + H(\omega^l(U, \bar{i}^*))]}{\gamma^*(U, \bar{i}^*, \varepsilon)} \right).$$

The final result of this part of the Appendix shows ()

Lemma 14. *For any generic S , any $\varepsilon > 0$, there is a local U , such that for any local $V \supseteq U$, any permutation $g \in G$,*

$$a(g \cdot V, S) \leq \varepsilon.$$

Proof. Fix generic S and $\varepsilon > 0$ and let $k = |S|$. Find $\varepsilon > 0$ small enough, so that

$$H(2\varepsilon) \leq \frac{\varepsilon}{3}.$$

Such ε exists, because $\lim_{t \rightarrow 0} H(t) = 0$.

First, find local set U_0 , such that

$$\gamma^* := \inf_{\substack{V \supseteq U_0, \\ V \text{ is local}}} \inf_{g \in G} \gamma^*(g \cdot V, \bar{i}, \varepsilon) > 0.$$

for some enumeration \bar{i} of generic set S . Such local set exists by the second part of Corollary 3.

Second, find local set $U_1 \supseteq U_0$ large enough, such that

$$-\frac{\log \gamma^*}{|U_1|} + \frac{\log k}{|U_1| \gamma^*} \leq \frac{\varepsilon}{3}.$$

Such a local set exists, because $G \mapsto I$ is locally generated.

Third, take $\varepsilon' \in (0, \frac{1}{2})$, such that

$$\frac{\varepsilon' + H(\varepsilon')}{\gamma^*} \leq \frac{\varepsilon}{3}.$$

Such ε' exists, because $\lim_{t \rightarrow o} H(t) = 0$. Find local $U \supseteq U_1$, such that

$$\inf_{\substack{V \supseteq U, \\ V \text{ is local}}} \inf_{g \in G} \delta(g \cdot V, \bar{v}, \varepsilon') > 0.$$

Such local set exists by the first part of Corollary 3 and the fact for any two sets finite $U_1, U' \subseteq I$, there is a local U that contains $U_1 \cup U'$. (This is because $G \mapsto I$ is locally generated.)

Then, by Lemma 13, for any local $V \supseteq U$, any permutation g ,

$$a(g \cdot V, S) \leq H(2\varepsilon) - \frac{\log \gamma^*}{|V|} + \frac{\log k}{|V| \gamma^*} + \frac{\varepsilon' + H(\varepsilon')}{\gamma^*} \leq \varepsilon.$$

□

B.2. Notation. The rest of Appendix B deals with the proof of Lemma 10. The idea is to construct a set of functions ρ with small cardinality and such that if the dimension of a family of models \mathcal{M} contains a generic set, then it can be spanned, up to minor modifications, by finitely many functions ρ . I begin with some notation. Then, I construct a set of indicators that help me to parametrize (and count) the aforementioned set of functions ρ . In the inductive step, I construct the set of functions ρ parametrized with indicators and use them to span models $\theta \in \mathcal{M}$. The last part finishes the proof of the Proposition.

From now on, fix local $U \subseteq I$ and generic $S \subseteq I$, $|S| = k$. Let $\{i_1^*, \dots, i_k^*\}$ be the enumeration of S from Lemma 13. Let $G_U \in \Pi_U$ denote the subgroup of permutations on U :

$$G_U = \{g|_U : g \in G \text{ and } g \cdot U = U\}.$$

Define set of k -tuples:

$$A = \{(g \cdot i_1^*, \dots, g \cdot i_k^*) : g \in G_U\}.$$

Set A consists of k -tuples of elements of U that are obtained as permutations of tuple $\bar{v}^* = (i_1^*, \dots, i_k^*)$. Thus, $A = A_U(\bar{v}^*)$, where set $A_U(\bar{v}^*)$ is defined in (B.2). Group action $G_U \mapsto U$ induces group action $G_U \mapsto A$. Because U is local, $G_U \mapsto A$ is transitive.

For any function $f : A \rightarrow R$, define Ef as in (A.1). For any subset $B \subseteq (U)^k$, let $\mathbf{1}_B : (U)^k \rightarrow R$ be the indicator of set B : $\mathbf{1}_B(\bar{v}) = 1$ iff $\bar{v} \in B$. For example, if $D \subseteq U$,

$\mathbf{1}_{D^k}$ is an indicator of tuples all elements of which belong to D . By definition (B.3), for any $\varepsilon > 0$,

$$\inf_{D \subseteq U: |D| \geq \varepsilon |U|} E \mathbf{1}_{D^k} = \delta(U, \bar{v}^*, \varepsilon). \quad (\text{B.8})$$

For any $l \leq k$, $i \in U$ define set of k -tuples:

$$A^l(i) = \{\bar{v} \in A : i_l = i\}.$$

This is a set of tuples from A , for which the l th element is equal to i . Observe that $\sum_{i \in U} \mathbf{1}_{A^l(i)} = \mathbf{1}$ and

$$Ef = Ef \left(\sum_{i \in U} \mathbf{1}_{A^l(i)} \right) = \sum_{i \in U} Ef \mathbf{1}_{A^l(i)}.$$

Let $\mathcal{M} \subseteq \{0, 1, *\}^I$ be a family of models, such that $S \in \dim_G^* \mathcal{M}$. Then, for each $\bar{v} = (i_1, \dots, i_k) \in A$, there is a function $\tau_{\bar{v}} : \{i_1, \dots, i_k\} \rightarrow \{0, 1\}$, such that for any $\theta \in \mathcal{M}$, $\theta|_{\{i_1, \dots, i_k\}} \neq \tau_{\bar{v}}$. For any $l \leq k + 1$, any $\theta : I \rightarrow \{0, 1, *\}$, define set of k -tuples:

$$A^l(\theta) = \{(i_1, \dots, i_k) \in A : \theta|_{\{i_1, \dots, i_{l-1}\}} = \tau_{\bar{v}}|_{\{i_1, \dots, i_{l-1}\}}\}. \quad (\text{B.9})$$

Here, $A^l(\theta)$ is a set of tuples \bar{v} , such that the values of function θ at the first $l - 1$ elements are equal to the corresponding values of $\tau_{\bar{v}}$. Note that,

$$A^{l+1}(\theta) = A^l(\theta) \cap \{\bar{v} : \theta(i) = \tau_{\bar{v}}(i_l)\}, \quad (\text{B.10})$$

and that

$$A^{k+1}(\theta) = \emptyset \text{ for any } \theta \in \mathcal{M}. \quad (\text{B.11})$$

B.3. Indicators. An *indicator* is any function $f : A \rightarrow R$. For any $\theta : I \rightarrow \{0, 1, *\}$, any $l \leq k$ and any permutation $g \in G_U$, define an operator $B_{g, \theta}^l$ on indicators: for any indicator $f : A \rightarrow R$, any $\bar{v} \in A$, let

$$B_{g, \theta}^l f(\bar{v}) := \mathbf{1}_{A^l(\theta)}(\bar{v}) f(g \cdot \bar{v}).$$

Take any indicator $f : A \rightarrow R$. For any $l \leq k$, define set of indicators $\mathcal{F}^l(f) :$

$$\mathcal{F}^l(f) := \{B_{g, \theta}^l f : g \in G \text{ and } \theta : I \rightarrow \{0, 1, *\}\}.$$

Lemma 15. *For any $l \leq k$, there exists indicator $f^l : A \rightarrow R$, such that*

- (a) $E f^l > \frac{1}{2k}$.
- (b) For each $i \in U$, $E f^l \mathbf{1}_{A^l(i)} \leq \frac{1}{|U|}$.
- (c) $\log |\mathcal{F}^l(f^l)| \leq |U| (\omega^l(U, \bar{v}^*) + H(\omega^l(U, \bar{v}^*)))$.

Recall that $\omega^l(U, \bar{v}^*)$ and $H(\cdot)$ are defined in equations (B.6) and (B.7).

Proof. Fix $l \leq k$. By definition (B.6), there is a set $W \subseteq U$, such that

$$|T^l(U, W, \bar{v})| \geq \frac{1}{2|S|} |U| \quad \text{and} \quad |W| \leq \omega^l(U, \bar{v}^*).$$

Define set of tuples $\bar{v} \in A$, for which the first $l - 1$ elements belong to set W ,

$$T^l := \{\bar{v} \in A : \{i_1, \dots, i_{l-1}\} \subseteq W\}.$$

Then,

$$\{i \in U : A^l(i) \cap T^l \neq \emptyset\} = T^l(U, W, \bar{v}). \quad (\text{B.12})$$

Define indicator $f^l : A \rightarrow R$ as

$$f^l(\bar{v}) = \begin{cases} \frac{|A|}{|U|} \frac{1}{|A(i) \cap T^l|} & \text{if } \bar{v} \in T^l, \\ 0 & \text{if } \bar{v} \notin T^l. \end{cases}$$

For any $i \in U$, such that $A^l(i) \cap T^l$ is empty, it must be that $E f^l \mathbf{1}_{A^l(i)} = 0$. For any $i \in U$, such that $A^l(i) \cap T^l$ is not empty, it must be that

$$E f^l \mathbf{1}_{A^l(i)} = \frac{1}{|A|} \sum_{\bar{v} \in A^l(i) \cap T^l} \frac{|A|}{|U|} \frac{1}{|A^l(i) \cap T^l|} = \frac{1}{|U|}.$$

Therefore, by (B.12) and Lemma ??,

$$\begin{aligned} E f^l &= \sum_{i \in U} E f^l \mathbf{1}_{A^l(i)} = \sum_{i \in U : A^l(i) \cap T^l \neq \emptyset} E f^l \mathbf{1}_{A^l(i)} \\ &= \frac{|\{i \in U : A^l(i) \cap T^l \neq \emptyset\}|}{|U|} = \frac{|T^l(U, W, \bar{v})|}{|U|} \geq \frac{1}{2k}. \end{aligned}$$

This demonstrates parts (a) and (b) of the Lemma.

I proceed not to derive a bound on $|\mathcal{F}^l(f^l)|$. Notice that for any $g, g' \in G$ and $\theta, \theta' : I \rightarrow Y$, such that

$$g \cdot W = g' \cdot W \quad \text{and} \quad \theta|_{g \cdot W} = \theta'|_{g \cdot W},$$

it must be that,

$$B_{g, \theta}^l f = B_{g', \theta'}^l f.$$

In other words, functions $f_{g, l, \theta}^l$ depend only on $g \cdot W$ and $\theta|_{g \cdot W}$. There are

- at most $\binom{|U|}{|W|}$ ways of choosing set $g \cdot W \subseteq U$ and
- at most $2^{|W|}$ ways of choosing different restrictions $\theta|_{g \cdot W}$.

The thesis of the Lemma follows from an application of the Stirling formula (??):

$$\frac{1}{|U|} \log |\mathcal{F}^l(f^l)| \leq \frac{1}{|U|} \log \left[\binom{|U|}{|W|} 2^{|W|} \right] \leq H(\omega^l(U, \bar{v}^*)).$$

□

From now on, I write \mathcal{F}^l instead of $\mathcal{F}^l(f^l)$.

Lemma 16. *For any $\theta : I \rightarrow \{0, 1, *\}$, any $l \leq k$, any $D \subseteq U$, there is an indicator $f \in \mathcal{F}^l$, such that*

$$\begin{aligned} f &= \mathbf{1}_{A^l(\theta)} f, \\ Ef \mathbf{1}_{A^l(i)} &\leq \frac{1}{|U|} \text{ for any } i \in U, \\ E \mathbf{1}_{D^k} f &\geq \frac{1}{4k} \frac{|A^l(\theta) \cap D^k|}{|A|}, \\ E \mathbf{1}_{A^{l+1}(\theta)} \mathbf{1}_{D^k} f &\leq 3 \frac{|A^{l+1}(\theta) \cap D^k|}{|A^l(\theta) \cap D^k|} E \mathbf{1}_{D^k} f. \end{aligned}$$

Proof. For any permutation $g \in G_U$, define $f_g := B_{g,\theta}^l f^l \in \mathcal{F}^l$. Then, for each $i \in U$,

$$\begin{aligned} Ef_g \mathbf{1}_{A^l(i)} &= E \mathbf{1}_{A^l(i)} \mathbf{1}_{A^l(\theta)} f^l(g \cdot \cdot) \leq E \mathbf{1}_{A^l(i)} f^l(g \cdot \cdot) \\ &= E \mathbf{1}_{A^l(g^{-1} \cdot i)} f^l \leq \frac{1}{|U|}, \end{aligned}$$

where the last equality follows from the fact that $G_U \mapsto A$ is transitive. Moreover, for each $g \in G_U$,

$$\begin{aligned} E \mathbf{1}_{D^k} f_g &= E \mathbf{1}_{A^l(\theta) \cap D^k} f^l(g \cdot \cdot), \\ E \mathbf{1}_{A^{l+1}(\theta)} \mathbf{1}_{D^k} f_g &= E \mathbf{1}_{D^k} \mathbf{1}_{A^{l+1}(\theta)} \mathbf{1}_{A^l(\theta)} f^l(g \cdot \cdot) = E \mathbf{1}_{A^{l+1}(\theta) \cap D^k} f^l(g \cdot \cdot). \end{aligned}$$

The last equality comes from the fact that $A^{l+1}(\theta) \subseteq A^l(\theta)$ (see (B.10)). By Lemma 15, $E f^l \geq \frac{1}{2k}$. By the second part of Lemma 9, there is $g \in G$, such that

$$\begin{aligned} E \mathbf{1}_{A^l(\theta)} \mathbf{1}_{D^k} B_{g,\theta}^l f^l &\geq \frac{1}{4k} \frac{|A^l(\theta) \cap D^k|}{|A|}, \\ E \mathbf{1}_{A^{l+1}(\theta)} \mathbf{1}_{D^k} B_{g,\theta}^l f^l &\leq 3 \frac{|A^{l+1}(\theta) \cap D^k|}{|A^l(\theta) \cap D^k|} E \mathbf{1}_{A^l(\theta)} \mathbf{1}_{D^k} B_{g,\theta}^l f^l. \end{aligned}$$

□

B.4. Inductive step. This part of the Appendix contains two lemmas that form the inductive step of the main argument.

Lemma 17. *For any $\varepsilon > 0$, any $D \subseteq U$, such that $|D| \geq \varepsilon |U|$, any $\theta \in \mathcal{M}$, there is $l \leq k$, so that*

$$E\mathbf{1}_{A^l(\theta)}\mathbf{1}_{D^k} \geq \gamma^l(U, \bar{v}^*, \varepsilon) \text{ and } E\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k} \leq \gamma^{l+1}(U, \bar{v}^*, \varepsilon). \quad (\text{B.13})$$

Proof. By definition (B.9), $A^1(\theta) = A$. Together with (B.8), this implies that

$$E\mathbf{1}_{A^1(\theta)}\mathbf{1}_{D^k} = E\mathbf{1}_{D^k} = \delta(U, \bar{v}^*, \varepsilon) = \gamma^0(U, \bar{v}^*, \varepsilon).$$

On the other hand, by (B.11)

$$E\mathbf{1}_{A^{k+1}(\theta)}\mathbf{1}_{D^k} = 0 \leq \gamma^{k+1}(U, \bar{v}^*, \varepsilon).$$

The result follows from the fact that constants $\gamma^l(U, \bar{v}^*, \varepsilon)$ are decreasing in l . \square

Define function $\mathcal{T}^l : A \rightarrow \{0, 1\}$ as

$$\mathcal{T}^l(\bar{i}) = 1 - \tau_{\bar{i}}(i_l).$$

For any $f \in \mathcal{F}^l$, any set $D \subseteq U$, define set

$$S^l(f; D) = \left\{ i \in U : Ef\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k} \geq \frac{1}{4k} \varepsilon \gamma_n^l(\varepsilon) \frac{1}{|U|} \right\}. \quad (\text{B.14})$$

Define a prediction function $p^l(f, D) : U \rightarrow [0, 1]$ as

$$p^l(f, D)(i) := \frac{E\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k} f \mathcal{T}^l}{E\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k} f}. \quad (\text{B.15})$$

Lemma 18. *For any $\varepsilon > 0$, any $\theta : I \rightarrow \{0, 1, *\}$, any $D \subseteq U$, if (B.13) holds for some $l \leq k$, then, there is an indicator $f \in \mathcal{F}^l$, such that*

$$|S^l(f; D)| \geq \gamma^*(U, \bar{v}^*, \varepsilon) |U|, \quad (\text{B.16})$$

$$\sum_{i \in S^l(f; D), \theta(i) \neq *} |\theta(i) - p^l(f, D)(i)| \leq \varepsilon |S^l(f; D)|. \quad (\text{B.17})$$

Proof. By Lemma 16 and (B.13), there is an indicator $f \in \mathcal{F}^l$, such that

$$f = \mathbf{1}_{A^l(\theta)} f, \quad (\text{B.18})$$

$$Ef\mathbf{1}_{A^l(i)} \leq \frac{1}{|U|} \text{ for each } i \in U \quad (\text{B.19})$$

$$E\mathbf{1}_{D^k} f \geq \frac{1}{4k} \gamma^l(U, \bar{v}^*, \varepsilon) \quad (\text{B.20})$$

$$E\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k} f \leq 3 \frac{\gamma^{l+1}(U, \bar{v}^*, \varepsilon)}{\gamma^l(U, \bar{v}^*, \varepsilon)} E\mathbf{1}_{D^k} f. \quad (\text{B.21})$$

By (B.19) and the definition of set $S^l(f; D)$

$$\begin{aligned} Ef\mathbf{1}_{D^k} &= \sum_{i \in S^l(f; D)} Ef\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k} + \sum_{i \notin S^l(f; D)} Ef\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k} \\ &\leq |S^l(f; D)| \frac{1}{|U|} + \frac{1}{4k} \varepsilon \gamma^l(U, \bar{v}^*, \varepsilon). \end{aligned}$$

Therefore, by (B.20),

$$\begin{aligned} \frac{1}{|U|} |S^l(f; D)| &\geq Ef\mathbf{1}_{D^k} - \frac{1}{4k} \varepsilon \gamma_n^l(\varepsilon) \geq (1 - \varepsilon) Ef\mathbf{1}_{D^k} \\ &\geq \frac{1}{4k} (1 - \varepsilon) \gamma^l(U, \bar{v}^*, \varepsilon) \geq \gamma^*(U, \bar{v}^*, \varepsilon). \end{aligned} \quad (\text{B.22})$$

This shows (B.16).

Observe that for any $\bar{v} \in A^l(\theta)$, if $\theta(i_l) \neq *$, then

$$|\theta(i_l) - \mathcal{T}^l(\bar{v})| = \begin{cases} 1, & \text{if } \bar{v} \in A^{l+1}(\theta) \\ 0, & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} &\sum_{i \in S^l(f; D), \theta(i) \neq *} |\theta(i) - p^l(f, D)(i)| \\ &= \sum_{i \in S^l(f; D), \theta(i) \neq *} \frac{|\theta(i) E\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k}f - E\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k}f\mathcal{T}^l|}{E\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k}f} \\ &= \sum_{i \in S^l(f; D), \theta(i) \neq *} \frac{E\mathbf{1}_{A^l(i)}\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}f}{E\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k}f}. \end{aligned}$$

Suppose that (B.17) does not hold. Then, there is subset $V \subseteq S^l(f; D)$, such that $|V| \geq \frac{\varepsilon}{2} |S^l(f; D)|$ and for each $i \in V$, $\theta(i_l) \neq *$ and

$$E\mathbf{1}_{A^l(i)}\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}f \geq \frac{\varepsilon}{2} E_n\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k}f.$$

But then,

$$\begin{aligned} E\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}f &= \sum_{i \in U} E\mathbf{1}_{A^l(i)}\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}f \\ &\geq \sum_{i \in V} E\mathbf{1}_{A^l(i)}\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}f \\ &\geq \sum_{i \in V} \varepsilon E_n\mathbf{1}_{A^l(i)}\mathbf{1}_{D^k}f \\ &\geq \frac{\varepsilon}{2} |V| \frac{1}{4k} \varepsilon \gamma^l(U, \bar{v}^*, \varepsilon) \frac{1}{|U|}. \end{aligned}$$

The last inequality is a consequence of the definition of set $|S^l(f; D)|$. Because $|V| \geq \frac{\varepsilon}{2} |S^l(f; D)|$,

$$E \mathbf{1}_{A^{l+1}(\theta)} \mathbf{1}_{D^k} f \geq \frac{1}{16k} \varepsilon^3 \gamma^l(U, \bar{v}^*, \varepsilon) \frac{|S^l(f; D)|}{|U|}.$$

By inequalities (B.21) and (B.22),

$$\gamma^{l+1}(U, \bar{v}^*, \varepsilon) \geq \frac{1}{12} \varepsilon^3 (\gamma^l(U, \bar{v}^*, \varepsilon))^2 (1 - \varepsilon \alpha).$$

But this yields a contradiction with the choice of constants (B.5).

CHECK THIS LAST INEQUALITY - SOME MINOR CHANGES □

B.5. Proof of Lemma 10. Fix $\varepsilon > 0$. Take finite sequences $l_1, \dots, l_m \leq k$ and $f_1, \dots, f_m, f_{m'} \in \mathcal{F}^{l_{m'}}$ for each $m' \leq m$ and define inductively sets $B_{m'}, D_{m'} \subseteq U$

$$D_{m'} := U \setminus \bigcup_{m'' < m'} B_{m''} \text{ and } B_{m'} := S_n^{l_{m'}}(f_{m'}, D_{m'}).$$

If $|D_{m+1}| \geq \varepsilon |U|$, then, by Lemmas 17 and 18, for any $\theta \in \mathcal{M}$, there are l_{m+1} and f_{m+1} , such that for $m' = m + 1$,

$$|B_{m'}| \geq \gamma^*(U, \bar{v}^*, \varepsilon) \text{ and } \sum_{i \in B_{m'}, \theta(i) \neq * } |\theta(i) - p^{l_{m'}}(f_{m'}, D_{m'})| \leq \varepsilon |B_{m'}|. \quad (\text{B.23})$$

holds.

Let $m^l(\varepsilon) = \left\lceil \frac{1}{\gamma^*(U, \bar{v}^*, \varepsilon)} \right\rceil$. Let Ψ be the space of all pairs sequences $(l_{m'} f_{m'})_{m' \leq m}$, where $m \leq m^l(\varepsilon)$, such that $l_{m'} \leq k$ and $f_{m'} \in \mathcal{F}^{l_{m'}}$ for each $m' \leq m$. Denote a typical element of Ψ as

$$\psi = \{m, (l_{m'} f_{m'})_{m' \leq m}\}.$$

For any $\psi \in \Psi$, let $\tau_\psi : U \rightarrow \{0, 1\}$, be defined as

$$\tau_\psi(i) = \begin{cases} 0 & \text{if } i \in B_{m'} \text{ and } p^{l_{m'}}(f_{m'}, D_{m'}) < \frac{1}{2}, \\ 1 & \text{if } i \in B_{m'} \text{ and } p^{l_{m'}}(f_{m'}, D_{m'}) \geq \frac{1}{2}, \\ 0 & \text{if } i \in D_{m+1}. \end{cases}$$

Then, (B.23) implies that for any $\theta \in \mathcal{M}$, there is $\psi \in \Psi$, such that

$$\begin{aligned} & \sum_{i \in U, \theta(i) \neq *} |\theta(i) - \tau_\psi(i)| \\ & \leq |D_{m+1}| + \sum_{m' \leq m} \sum_{i \in B_{m'}: p^{l_{m'}}(f_{m'}, D_{m'}) < \frac{1}{2}} \theta(i) + \sum_{m' \leq m} \sum_{i \in B_{m'}: p^{l_{m'}}(f_{m'}, D_{m'}) \geq \frac{1}{2}} 1 - \theta(i) \\ & \leq \varepsilon |U| + 2 \sum_{m' \leq m} \sum_{i \in B_{m'}} |\theta(i) - p^{l_{m'}}(f_{m'}, D_{m'})| \leq \varepsilon |U| + 2\varepsilon \sum_{m' \leq m} |B_{m'}| \leq 2\varepsilon |U|. \end{aligned}$$

For any $\psi \in \Psi$, any subset $W \subseteq U$, define

$$\tau_{\psi,W}(i) = \begin{cases} \tau_{\psi,W}(i), & \text{if } i \notin W, \\ 1 - \tau_{\psi,W}(i), & \text{if } i \in W. \end{cases}$$

Let

$$\mathcal{M}_U^* = \{\tau_{\psi,W} : \psi \in \Psi \text{ and } W \subseteq U, |W| \leq 2\varepsilon |U|\}.$$

Then, family \mathcal{M}_U^* approximates \mathcal{M} on U .

The only thing left is to bound the size of family \mathcal{M}_U^* . Notice that

$$\begin{aligned} |\mathcal{M}_U^*| &\leq |\Psi| |\{W \subseteq U : |W| \leq 2\varepsilon |U|\}|. \\ &\leq m^l(\varepsilon) \left(k \max_{l \leq k} |\mathcal{F}^l| \right)^{m^l(\varepsilon)} \binom{|U|}{2\varepsilon |U|}. \end{aligned}$$

Next, observe that by Lemma 15 and the definition of $m^l(\varepsilon)$

$$\begin{aligned} \frac{1}{|U|} \log |\Psi| &\leq \frac{1}{|U|} \log \left(m^l(\varepsilon) \left(k \max_{l \leq k} |\mathcal{F}^l| \right)^{m^l(\varepsilon)} \right) \\ &\leq \frac{-\log \gamma^*(U, \bar{v}^*, \varepsilon)}{|U|} + \frac{\log k}{|U| \gamma^*(U, \bar{v}^*, \varepsilon)} + \frac{\max_{l \leq k} \log |\mathcal{F}^l|}{|U| \gamma^*(U, \bar{v}^*, \varepsilon)} \\ &\leq \frac{-\log \gamma^*(U, \bar{v}^*, \varepsilon)}{|U|} + \frac{\log k}{|U| \gamma^*(U, \bar{v}^*, \varepsilon)} + \frac{\max_{l \leq k} [\omega^l(U, \bar{v}^*) + H(\omega^l(U, \bar{v}^*))]}{|U| \gamma^*(U, \bar{v}^*, \varepsilon)}. \end{aligned}$$

Also, by Stirling's formula,

$$\frac{1}{|U|} \log |\{W \subseteq U : |W| \leq 2\varepsilon |U|\}| \leq \frac{1}{|U|} \log \binom{|U|}{2\varepsilon |U|} \leq H(2\varepsilon).$$

The result follows.

APPENDIX C. PROOFS OF SECTION 6

C.1. Proof of Theorem 5. The proof is divided into three parts.

C.1.1. Product of local sets is local. It is sufficient to consider only the product of two groups, $d = 2$. Let U_j be local sets under group actions $G_j \mapsto I_j$ for both $j = 1, 2$. Observe that

$$\begin{aligned} G_{U_1 \times U_2} &= \{(g_1, g_2) : (g_1, g_2) \cdot U_1 \times U_2 = U_1 \times U_2\} \\ &= \prod_{j=1,2} \{g_j : g_j \cdot U_j = U_j\} = G_{U_1} \times G_{U_2}. \end{aligned}$$

Take any subset $S \subseteq U$ and suppose that there is $(g_1, g_2) \in G$, such that $(g_1, g_2) \cdot S \subseteq U$. Let $S_j \subseteq U_j$ be the projection of S on its j th coordinate:

$$S_j = \{i_j : (i_j, i_{-j}) \in S\}.$$

Because U_j is local, there is a permutation $g'_j \in G_{U_j}$, such that $g'_j \cdot S_j = g_j \cdot S_j$. This implies that $(g'_1, g'_2) \cdot S = (g_1, g_2) \cdot S$. Hence, $U_1 \times U_2$ is local under the product group action.

C.1.2. *Product of generic sets is generic.* It is sufficient to prove the result for $d = 2$. Fix any $\varepsilon > 0$. Denote $k_j = |S_j|$ and let $\bar{v}^{*j} = (i_1^{*j}, \dots, i_{k_j}^{*j})$ be an enumeration of S_j , i.e. $\{i_1^{*j}, \dots, i_{k_j}^{*j}\} = S_j$. For any local sets $U_j \subseteq I_j$, any $\varepsilon > 0$, define $\delta_j(U_j, \bar{v}^{*j}, \varepsilon)$ as in equation (B.3). Define also sets $A_{U_j}(\bar{v}^{*j})$ as in (B.2). Choose local set $U_1 \subseteq I_1$,

$$\delta_1\left(U_1, \bar{v}^{*1}, \frac{\varepsilon}{2}\right) > 0$$

and local set $U_2 \subseteq I_2$, such that

$$\delta_2\left(U_2, \bar{v}^{*2}, \frac{\varepsilon}{4}\delta_1\left(U_1, \bar{v}^{*1}, \frac{\varepsilon}{2}\right)\right) > 0.$$

Such local sets exist by Lemma ?? and because S_j are generic.

By the above, $U_1 \times U_2$ is local under the product group action. For any $D \subseteq U_1 \times U_2$, define sets

$$D_1 \subseteq A_{U_1}(\bar{v}^{*1}) \times U_2, \quad D_{12} \subseteq A_{U_1}(\bar{v}^{*1}) \times A_{U_2}(\bar{v}^{*2})$$

as follows:

$$\begin{aligned} D_1 &= \{(\bar{v}^1, i^2) : (i_l^1, i^2) \in D \text{ for any } l \leq k_1\} \text{ and} \\ D_{12} &= \{(i_1^1, \dots, i_{k_1}^1, i_1^2, \dots, i_{k_2}^2) : \text{for any } l \leq k_2, (i_1^1, \dots, i_{k_1}^1, i_l^2) \in D_1\} \\ &= \{(i_1^1, \dots, i_{k_1}^1, i_1^2, \dots, i_{k_2}^2) : \text{for any } l_1 \leq k_1, l_2 \leq k_2, (i_{l_1}^1, i_{l_2}^2) \in D\}. \end{aligned}$$

I will show that for any $|D| \geq \varepsilon |U_1 \times U_2|$, set D_{12} is not empty. This implies that then there is a permutation $(g_1, g_2) \in G_1 \times G_2$, such that $(g_1, g_2) \cdot (S_1 \times S_2) \subseteq D$. Thus, S is generic.

For any $i_2 \in U_2$, define

$$\alpha_2(i_2) = \frac{|\{i_1 : (i_1, i_2) \in D\}|}{|U_1|}.$$

Then, for any $\gamma_1 > 0$

$$\begin{aligned} |D| &\leq \{i_2 : \alpha_2(i_2) \geq \gamma_1\} |U_1| + \gamma_1 |U_1| |U_2|, \text{ and} \\ \frac{|D|}{|U_1| |U_2|} - \gamma_1 &\leq \frac{\{i_2 : \alpha_2(i_2) \geq \gamma_1\}}{|U_2|}. \end{aligned}$$

For any $\bar{i}_1 \in A_{U_1}(\bar{i}^{*1})$, define also

$$\alpha_1(\bar{i}_1) = \frac{|\{i_2 : (\bar{i}^1, i^2) \in D_1\}|}{|U_2|}.$$

Then, for any $\gamma_2 > 0$,

$$\begin{aligned} |D_1| &\leq \{\bar{i}_1 : \alpha_1(i_1) \geq \gamma_2\} |U_2| + \gamma_2 |A_{U_1}(\bar{i}^{*1})| |U_2|, \text{ and} \\ \frac{|D_1|}{|U_2| |A_{U_1}(\bar{i}^{*1})|} - \gamma_2 &\leq \frac{|\{\bar{i}_1 : \alpha_1(\bar{i}_1) \geq \gamma_2\}|}{|A_{U_1}(\bar{i}^{*1})|}. \end{aligned}$$

Observe that,

$$\begin{aligned} |D_1| &= \sum_{i_2 \in U_2} |\{(\bar{i}^1, i^2) : \text{for any } l \leq k_1, (\bar{i}_l^1, i^2) \in D\}| \\ &\geq \sum_{i_2 \in U_2} \delta\left(U_1, \bar{i}^{*1}, \frac{|\{i_1 : (i_1, i_2) \in D\}|}{|U_1|}\right) |A_{U_1}(\bar{i}^{*1})| \\ &= \sum_{i_2 \in U_2} \delta(U_1, \bar{i}^{*1}, \alpha_2(i_2)) |A_{U_1}(\bar{i}^{*1})| \\ &\geq \left(\frac{|D|}{|U_1| |U_2|} - \gamma_1\right) \delta(U_1, \bar{i}^{*1}, \gamma_1) |U_2| |A_{U_1}(\bar{i}^{*1})|. \end{aligned}$$

Similarly,

$$\begin{aligned} |D_{12}| &= \sum_{\bar{i}^1 \in A_{U_1}(\bar{i}^{*1})} |\{(\bar{i}^1, \bar{i}^2) : \text{for any } l \leq k_2, (\bar{i}_l^1, i_l^2) \in D_1\}| \\ &\geq \sum_{\bar{i}^1 \in A_{U_1}(\bar{i}^{*1})} \delta\left(U_2, \bar{i}^{*2}, \frac{|\{i_2 : (\bar{i}^1, i^2) \in D_1\}|}{|U_2|}\right) |U_2| \\ &= \sum_{\bar{i}^1 \in A_{U_1}(\bar{i}^{*1})} \delta(U_2, \bar{i}^{*2}, \alpha_1(\bar{i}_1)) |U_2| \\ &\geq \left(\frac{|D_1|}{|U_2| |A_{U_1}(\bar{i}^{*1})|} - \gamma_2\right) \delta(U_2, \bar{i}^{*2}, \gamma_2) |A_{U_1}(\bar{i}^{*1})| |U_2|. \end{aligned}$$

By the above,

$$\frac{|D_{12}|}{|A_{U_1}(\bar{i}^{*1})| |U_2|} \geq \left(\left(\frac{|D|}{|U_1| |U_2|} - \gamma_1\right) \delta(U_1, \bar{i}^{*1}, \gamma_1) - \gamma_2\right) \delta(U_2, \bar{i}^{*2}, \gamma_2).$$

Choose now

$$\gamma_1 = \frac{\varepsilon}{2} \text{ and } \gamma_2 = \frac{\varepsilon}{4} \delta_1(U_1, \bar{i}^{*1}, \gamma_1).$$

By the choice of local sets,

$$\frac{|D_{12}|}{|A_{U_1}(\bar{i}^{*1})| |U_2|} \geq \frac{\varepsilon}{4} \delta(U_1, \bar{i}^{*1}, \gamma_1) \delta\left(U_2, \bar{i}^{*2}, \frac{\varepsilon}{4} \delta_1(U_1, \bar{i}^{*1}, \gamma_1)\right) > 0.$$

C.1.3. *Product of tight group actions is tight.* It is sufficient to prove the result for $d = 2$. The fact that the product of transitive group actions is transitive is obvious. Suppose that for any $j = 1, 2$, there are constants $\delta_j > 0$, such that for any finite $A_j \subseteq I_j$, there is generic $S_j \subseteq A_j$, $|S_j| \geq \delta_j |A_j|$. Take any finite $A \subseteq I_1 \times I_2$. I show that it contains a generic subset with at least $\delta_1 \delta_2 |A|$ elements.

Indeed, there are local sets $U_j \in I_j$, such that $A \subseteq U_1 \times U_2$. By the above, $U_1 \times U_2$ is local, and group action $G_{U_1 \times U_2} \mapsto U_1 \times U_2$ is transitive. Let $S_j \subseteq U_j$ be generic sets such that $|S_j| \geq \delta_j |U_j|$ for both $j = 1, 2$. By Lemma 8, there is a permutation $g \in G_{U_1 \times U_2}$, such that

$$\frac{|(g \cdot (S_1 \times S_2)) \cap A|}{|U|} \geq \frac{|S_1 \times S_2| |A|}{|U| |U|} \geq \delta_1 \delta_2 \frac{|A|}{|U|}.$$

Let $S = (g \cdot (S_1 \times S_2)) \cap A$. The above implies that $|S| \geq \delta_1 \delta_2 |A|$. Since $S_1 \times S_2$ is generic as a product of generic sets, S is generic as a subset of a generic set.

This ends the proof of the Theorem.

C.2. **Proof of Proposition 1.** Tba

C.3. **Proof of Proposition 2.** TBA

C.4. UNIVERSITY OF CHICAGO, DEPARTMENT OF ECONOMICS

E-mail address: mpeski@uchicago.edu