LEARNING THROUGH THEORIES

MARCIN PĘSKI*

ABSTRACT. This paper builds a decision-theoretic framework to examine the relationship between language and knowledge. A decision maker describes the world through theories. A theory consists of universal propositions called patterns, and it is formulated in some language. I look at two characteristics of a successful theory. A theory is *informative* if it allows agents to precisely predict outcomes of some process. A theory is *brief* if it consists of finitely many patterns. The main result of the paper identifies languages for which there is no trade-off between both characteristics: Any informative theory logically implies a theory that is informative as well as brief. I illustrate the main result on specific problems of reasoning under uncertainty: recommendation problems, binary preferences of a customer, or a stylized example of a chemical research.

1. INTRODUCTION

Theorizing, a process of thinking through general principles, is one of the most important tools of human cognition. There are at least two roles for theories. First, theories represent and compress information contained in a large number of individual facts. As such, they are communicated in journals, published in the form of books or taught in classrooms. Second, theories are used to predict novel facts. In this way, they influence the decision-making. A quality of a theory can be measured by how well it fulfils its roles. Among others,

- A good theory should be *correct*.
- It should be *informative*, i.e., it should lead to precise predictions. Theory "everything is possible" is tautologically correct, but it has no predictive value. I am interested here in a very strong notion of informativeness: one should be able to use the theory to deduce almost all relevant predictions.
- It should be *brief*, i.e. one should be able to express it using finitely many statements. To use the theory, the agent must be able to comprehend it. To

Preliminary and Incomplete. *University of Chicago, Department of Economics. E-mail: mpeski@uchicago.edu. I am grateful to Jeroen Swinkels, Larry Samuelson and Balazs Szentes and seminar participants in Caltech, UCLA and UT Austin for discussion. All remaining errors are my own.

communicate theory, it needs to be expressed in a finite medium. An infinite theory is clearly useless for any real-world applications.

Each theory must be stated in a language. The language defines the power and the form of expression, by characterizing which theoriies can be stated and how they are stated. There are various reasons why the relationship between language and theories is interesting. First, theorizing is rarely a goal itself. More often, it is a tool towards other goals, like storage of information, or prediction. As with any tool, it is natural to ask about its quality. Different features of a good theory lead to different trade-offs. One cannot understand this trade-offs without paying close attention to the language. In particular, it is impossible to define what it means that a theory is brief without being clear about how the theory is stated and what the alternative statements are. Second, the relation between language and knowledge is of foundational importance for the philosophy of science. Theories represent human knowledge about the world. In order to understand better the epistemic status of a theory, it is essential to know what are the constraints imposed on it by the language. Third, more or less explicit theorizing is used by all people on a daily basis to predict consequences of their actions. The language that the agent uses may affect how he processes information and how he makes decisions. This makes the language of interest for all behavioral sciences, including psychology, cognitive sciences and economics (see (Rubinstein 2000) and (Lipman 2003)). Fourth, people communicate knowledge in the language. The language will affect the quality of information transmission. Are there languages which are better for communication than other?

This paper uses a decision theoretic framework to examine the relationship between language and theories. In the first step, I characterize a class of theories that are brief and informative in the same time. Any such theory can be expressed using only one finitely stated proposition and such theory is sufficient to deduce almost all relevant predictions. Next, I characterize languages in which every informative theory belongs to the aforementioned class. In particular, any informative theory can be stated very briefly. Examples include a language used in the decision theory to discuss binary choices of a consumer, a language used to describe interactions between sets of objects and a language used to formulate knowledge in recommendation systems. A notable exception, which does not satisfy the assumptions of the main results, is a language used to express theories about events that develop in time.

In the rest of the introduction, I describe briefly the main ideas of the paper: the definition of a language, the decision theoretic model of learning, the formal results

and some more specific motivation. The introduction is concluded by brief overview of related literature.

1.1. **Language.** To focus the attention on the role of language, I assume that theories are deterministic. (Probabilistic theories involve additional complications, some of which are discussed in the last section of this paper.) The world is characterized by an assignment $\theta : X \to Y$ of properties $\theta(x) \in Y$ to objects $x \in X$, where X is infinite and Y is finite. A *language* is defined as a finite set $\mathcal{L} = \{R_i\}$ of k_i-nary relations $R_i \subseteq X^{k_i}$ on X. These relations, together with free variables $x_1, ..., x_k$, symbol θ, constants y, equality relation "=", $\{R_i\}$, and logical symbols, can be used to construct formulas (sentences) in the language. A theory consists of *patterns*. A pattern is any universal propositions

$$\forall_{x_1,...,x_k} F(x_1, ..., x_k, \theta), \tag{1.1}$$

where F is any formula with free variables $x_1, ..., x_k$ and function θ. For example, the theory "All objects have property y" is stated as a pattern

$$\forall_x \theta(x) = y. \tag{1.2}$$

Suppose that language $\mathcal{L}$ predefines binary relation R between objects. This allows to state the theory "If x has property y and xRx', then x' has property y":

$$\forall_{x,x'} \ (xRx' \wedge \theta(x) = y) \Longrightarrow \theta(x') = y. \tag{1.3}$$

Different sets of relations $\{R_i\}$ lead to different languages. In Section 2, I present an alternative, algebraic definition of a language as a group of all permutations on X that preserve relations $\{R_i\}$. That definition is more convenient to work with. In particular, one can analyze language through well-known algebraic tools. Both definitions are illustrated on 'examples. Section 2.1 discusses examples. The first language, *Time*, consists of one relation interpreted as "one period before". This allows to express patterns in discrete time series. The second example, *Product*, can be used to state patterns in the recommendation problems. The third example, *Ordered Graph,* is a language that allows to state various properties of binary preference relation, for instance, transitivity. The fourth example, *Graph*, is used to describe stylized problem of chemical research.

1.2. **Learning.** Theory (1.2) is trivially informative as it conveys very precise information about all objects. The informativeness of theory (1.3) is more complicated. This theory does not provide information about any particular object. However, if few past observations had been made, one can apply (1.3) to deduce future predictions. If the decision maker had observed an object x with property y, then (1.3) implies that all objects x' such that xRx', also have property y. Even if information about x cannot

be predicted correctly, the theory can still be useful if the cost of one mistake is small relative to the gain in information about other objects. In other words, such a theory might still be informative even if it does not provide precise information about every single object.

To capture formally such a notion of informativeness, I consider a model of sequential learning. The decision maker observes a process of objects $x_1, x_2, x_3, ...$ and sequentially predicts their properties. Her behavior is described by a learning rule, that is formally defined as a mapping from the history of past observations into current predictions. For any realization of properties, one can compute the long-run average number of mistakes committed while using the learning rule. A theory is *sufficient for deduction* if there is a learning rule that leads to almost no mistakes in the long-run and on average, for any realization of properties that satisfies the theory. Thus, an uncertainty-averse decision maker (as in (Gilboa and Schmeidler 1989)) who knows a theory that is sufficient for deduction expects to predict, in the long-run and on average, almost all properties.[1]

Whether a theory is informative or not, it depends on how many patterns the theory contains and how often these patterns can be applied to deduce outcomes of a sequential process. Clearly, there is a tension between brevity and informativeness. If a theory is informative and consists of infinitely many patterns, it might be easier to find an appropriate pattern for any given period. If a theory is brief and patterns can be used only rarely, it cannot be informative.

How often can a pattern (1.1) be applied in deduction? Define support of pattern (1.1) as the set of tuples $x_1, ..., x_k$ for which there exists θ, such that formula $F(x_1, ..., x_k, \theta)$ is not true. Patterns are informative only on their support because that's where they put restriction on θs. I identify a class of patterns *with generic support* that, informally, can be non-trivially applied to a "large" set of tuples $x_1, ..., x_k$. Patterns with generic support can be applied often.

1.3. **Results.** The key result is stated in Proposition 4: If $Y = \{0, 1\}$, then any pattern with generic support is alone sufficient for deduction. As a consequence, any theory that logically implies a pattern with generic support, is sufficient for deduction. (Proposition 4 is a combinatorial result that is closely related to the famous Sauer-Shelah's Lemma from the statistical learning theory. The connection is explained in Section 6.)

Are patterns with generic support exceptional or widespread? It depends on the language. In some, *tight*, languages, any informative theory logically implies at least one pattern with generic support. Section 5 shows that languages *Product, Graph* and

[1]This is a very strong notion of informativeness. An alternative, weaker notion is Popper's falsifiability, which asks for existence of at least one testable implication.

Ordered Graph are tight. Language *Time* does not contain any pattern with generic support, hence it is not tight.

The main results of the paper provide a complete characterization of theories that are and that are not sufficient for deduction. Consider first the binary case $Y = \{0, 1\}$. There, any theory that is sufficient for deduction logically implies a pattern with generic support; therefore, it logically implies a theory which is sufficient for deduction and which consists of only one pattern. In the general case of finite Y, any theory that is sufficient for deduction implies a theory that consists of finitely many patterns with generic support. In other words, any theory that is informative, but possibly consists of infinitely many patterns, implies a theory that is informative and brief.

The second result is concerned with theories that are not sufficient for deduction. Any tight language can be characterized by a constant $h > 0$ with the following property. Suppose that the decision maker's information is described by a theory that is not sufficient for deduction. Then, no matter what learning rule she uses, there is a realization of properties that is consistent with a theory and that leads to, on average, a mistake every $1/h$ predictions. Constant h depends only on the language but not on a theory.

1.4. **Applications.** The results are applicable to various specific questions.

Classification of theories: A decision maker (who uses a theory to make predictions), or a scientist (who describes his discovery), would both like to know how informative their theories are. The above results lead to a simple test of the sufficiency for deduction in tight languages. Namely, it is enough to check whether they imply patterns with generic supports. This might be quite helpful when patterns with generic support are easy to characterize. Examples of such situations are provided in Section 5).

Representation of information. If patterns allow the deduction of facts about the world, one can use them as a representation of information. This may lead to practical implications. First, patterns can be used to compress information. Second, knowledge of a pattern is valuable, hence patterns can be traded. Patterns with generic support are more informative than others; therefore these patterns should be more valuable.

Communication: Patterns can be used to communicate information. Consider a teacher who passes his knowledge about the world to a student. If the teacher's knowledge is specified in a tight language and it is informative, then the teacher will be able to pass his knowledge in a finite time. Loosely speaking, if there is anything interesting to be said, it can be said briefly.

Choice of language: In some situations, the agent may choose which language to use to express information. Suppose that the decision maker cares only about informative and brief theories (for example, because she cannot formulate theories that have infinitely

many patterns, and, she does not care for any theory that leads to a strictly positive amount of mistakes). She decides which of two languages, A or B, to use. If any theory in B that is informative and brief, can be also formulated in A, and A is tight, then she should prefer A. To see why, notice first that if theory is informative in B, it must be also informative in A (the deductive properties of a theory do not depend on the language in which it is stated). Since A is tight, this means that any informative theory in B can be restated as an informative and brief theory in A. The preferences for A will be strong if there are theories that are informative and brief in A, and they cannot be formulated briefly in B.

Bounded rationality: If there is a positive cost of memory, the decision maker may decide "what to know" (see (Rubinstein 1998), chapter 5). Consider two agents. The first one has limited memory and has capacity to understand only brief theories. The other one uses tight language to express her knowledge, but he has otherwise unbounded memory. The Theorem implies that the former agent is not necessarily at any disadvantage with respect to the latter. Knowing little, but knowing the right thing, is sufficient to compete with somebody who knows a lot.

1.5. **Related literature.** When computing payoffs from prediction, the decision maker uses the worst-case scenario criterion. This can be interpreted as an extreme version of the uncertainty aversion as in (Gilboa and Schmeidler 1989). Under different interpretation, this is an approach taken by two strands of the statistical literature: (a) the classic statistical decision theory of (Wald 1950), (Wald 1949) and (Blackwell and Girshick 1954) (see (Schwarz 1994)); and (b) the statistical learning theory of (Vapnik and Chervonenkis 1971) (see (Vapnik 1998) and (Bousquet, Boucheron, and Lugosi 2004)).

A. Rubinstein in (Rubinstein 1996) (see also Chapter 1 of (Rubinstein 2000)) asks what properties of binary relations make them appropriate for use in natural languages. (Binary relations correspond to assignments θ in my terminology; see example in Section 2.1.2.) Among others, he argues that binary relations should be easy to describe: One should be able to learn it from an universal proposition (theory, in my terminology) aided by few examples.[2] The ease of describibility is very closely connected to the sufficiency for deduction, as the latter means that one can predict many outcomes while making relatively few mistakes. Rubinstein looks for relations that are optimal with respect to

[2]In a similar spirit, (Blume 2004) postulates that a natural language, understood as an assignment of meanings to utterances, should be simple to learn. This implies that natural languages should have a grammar: utterances are built of words and each word is associated to a particle of meaning.

such a criterion. Here, the goal is to characterize the set of all theories that can be learned with few mistakes.

I. Gilboa and D. Schmeidler examine the role of similarity in induction (see, among others, (Gilboa and Schmeidler 1995), (Gilboa and Schmeidler 1996), (Gilboa and Schmeidler 2000), (Gilboa and Schmeidler 2003), (Billot, Gilboa, Samet, and Schmeidler 2005), and (Gilboa and Schmeidler 2001)). Although my model is very different, it is worthy to highlight a certain analogy. Suppose, for simplicity, that the language consists of a relation of equivalence R. One can distinguish between two notions of similarity. *Ex ante similarity* is predefined by the language: say that objects x and x' are ex ante similar, if xRx'. *Ex post similarity* depends on the outcomes of observations: say that x and x' are ex post similar if they share the same properties. The decision maker uses a theory to formulate connections between two types of similarity. For example, theory (1.3) predicts that objects that are ex ante similar are also ex post similar.

2. Language

The decision maker faces uncertainty about relationship between *instances* x (objects, problems) and *outcomes* y (properties, solutions). Let X be an infinite universe of instances and let Y be a finite space of outcomes. A mapping $\theta : X \to Y$ is called a *model of the world.* Thus, a model is an assignment of an outcome to each instance. Let $\mathcal{M} = Y^X$ be the space of all models. The decision maker describes her knowledge about the world in a language.

2.1. **Examples.**

2.1.1. *Time.* A forecaster predicts weather in Chicago. Let $X = \mathbf{Z}$ be equal to the set of integers interpreted as periods. Let $Y = \{0, 1\}$, where 0 is interpreted as "rain" and 1 is interpreted as "sun". The forecaster may believe that

$$\text{It never rains three days in the row.} \tag{2.1}$$

2.1.2. *Binary choices.* An econometrician is interested in binary choices of a consumer.[3] Suppose that B is a countable set of products. Let X be a set of ordered pairs of distinct elements of B. Let $Y = \{0, 1\}$. Model $\theta : X \to Y$ assigns preferences to the pairs of products: if $\theta(a, b) = 0$, then, faced with choice between a and b, consumer prefers a;

[3]I am grateful for this example to Matias Iaryczower. Related learning problems are studied in (Kalai 2003) and (Salant 2007). The former paper addresses learnability of rational choice rules from sets. The latter deals with binary preferences and assumes that they are outcomes of majority voting.

$\theta(a,b) = 1$ means that he prefers b. The econometrician may know that consumer's preferences are transitive:

$$\begin{gathered}\text{For any products } a,b,c \in B, \\ \text{if } a \text{ is preferred to } b \text{ and } b \text{ is preferred to } c, \text{ then } a \text{ is preferred to } c.\end{gathered} \tag{2.2}$$

2.1.3. *Chemical research.* In a laboratory, chemists combine pairs of substances and observe what happens. Sometimes, but not always, salty water is produced in a stormy reaction. The chemists predict types of reactions using the acid-alkaline theory. According to this theory, there are two types of substances, acids and alkalines; the substances of different types react and the substances of the same types don't. Let B be a countable set of all tested substances. This theory is equivalent to two universal propositions:

$$\begin{gathered}\text{For any substances } a,b,c \in B, \\ \text{if } a \text{ and } b \text{ react, } a \text{ and } c \text{ react, then } b \text{ and } c \text{ don't react,}\end{gathered} \tag{2.3}$$

$$\begin{gathered}\text{for any substances } a,b,c \in B, \\ \text{if } a \text{ and } b \text{ react, } a \text{ and } c \text{ don't react, then } b \text{ and } c \text{ react.}\end{gathered} \tag{2.4}$$

2.1.4. *Netflix recommendation problem.* Netflix is an Internet-based DVD rental company. In one of its services, it makes personal recommendations to its customers. Formally, let C be the set of customers, M be the set of movies, $X = C \times M$ be the set of instances and let $Y = \{0,1\} = \{\text{doesn't like, like}\}$ be the set of preferences. Netflix's recommendation are based on its knowledge about the model of the world $\theta : X \to Y$, where $\theta(c,m) = 1$ is interpreted as "customer c likes movie m". Netflix may know that some universal propositions about preferences are true. For example,

$$\text{For any movie } m \text{ and any customers } c, c', \text{ if } c \text{ likes } m, \text{ then } c' \text{ likes } m, \tag{2.5}$$

$$\begin{gathered}\text{for any two customers } c \text{ and } c', \text{ for any two movies } m \text{ and } m', \\ \text{if } c \text{ likes } m \text{ but dislikes } m' \text{ and if } c' \text{ dislikes } m, \text{ then } c' \text{ dislikes } m'.\end{gathered} \tag{2.6}$$

describe universal propositions and are called patterns.

Next, I present two definitions of a language and patterns. Following the introduction, the first definition treats language as a finite set of relations on set X. The second definition describes language as a group of permutations on X. The latter definition is

more convenient and, under some conditions, is equivalent to the former. All definitions are illustrated with examples.

2.2. **Relational definition of language.** A *(relational) language* is defined as a finite set of relations $\mathcal{L} = \{R_i\}$ on X, where $R_i \subseteq X^{k_i}$ for some $k_i < \infty$ and each i. Language $\mathcal{L}$ can be used to formulate atomic formulas on free variables $x_1, x_2, ...$, function θ and constants $y \in \{0, 1\}$:

$$x_1 = x_2, R_i(x_1, .., x_{k_i-1}), \theta(x_1) = y, \text{ and not } A, \tag{2.7}$$

where A is one of the atomic formulas of the language. Let $F(x_1, ..., x_k, \theta)$ be a (not necessarily atomic) formula that is a disjunction of conjunctions of finitely many atomic formulas. A *(relational) pattern* is a universal proposition of the form (1.1).[4] A *(relational) theory* T is any subset of patterns.

Model θ *satisfies* pattern (1.1) if (1.1) is true.[5] Model θ satisfies theory T if it satisfies all patterns. For any theory $T \subseteq \mathcal{P}$, let $\mathcal{M}(T)$ denote the set of models that satisfy theory T. Set $\mathcal{M}(T)$ is interpreted as the range of uncertainty faced by the decision maker who knows that theory T holds and nothing else. Theory T is called *consistent* if it is satisfied by some models, i.e., $\mathcal{M}(T)$ is not empty. In what follows, I always assume that theory T is consistent. Theory T *logically implies* theory S, if $\mathcal{M}(T) \subseteq \mathcal{M}(S)$.

Any language induces an equivalence relation on tuples of instances. For any k, any $\bar{x}, \bar{x}' \in X^k$, say that tuples $\bar{x}$ and $\bar{x}'$ are *exchangeable*, if they are described by the same relations between its elements:

$$\forall_{l,l'}(x_l = x_{l'}) \Leftrightarrow (x'_l = x'_{l'}) \text{ and}$$
$$\forall_i \forall_{l_1,...,l_{k_i} \in \{1,...,k\}} R_i\left(x_{l_1}, ..., x_{l_{k_i}}\right) \Leftrightarrow R_i\left(x'_{l_1}, ..., x'_{l_{k_i}}\right).$$

Because the number of relations in the language is finite, the set of k-tuples can be partitioned into finitely many equivalence classes of exchangeability. Any relational language is completely characterized by these classes.

Although exchangeability fully characterizes relations between instances in the tuple, it does not capture relations between the instances in the tuple to the instances outside

[4]This definition of a language presumes that relations R_i are already equipped with a meaning. Mathematical logic defines language purely as a set of symbols (in particular, relational symbols). The symbols acquire meaning only when the language is interpreted in a model (see, for example, (Monk 2005)). I use a shortcut here to reach as quickly as possible a definition that corresponds to a daily use of the word "language".

[5]Formally, "model θ satisfies formula F" is defined starting from the atomic formulas and then, by induction, extending the definition to all formulas.

the tuple. A definition is useful. For any $\bar{x} \in X^k$ and $x \in X$, let $\bar{x}\hat{}x$ denote a $(k+1)$-tuple that is created by adding x at the end of tuple $\bar{x}$.

Definition 1. *Relational language $\mathcal{L}$ is* externally stable*, if, for any k, any exchangeable k-tuples $\bar{x}, \bar{x}' \in X^k$, any instance $x \in X$, there is an instance x', such that tuples $\bar{x}\hat{}x$ and $\bar{x}'\hat{}x'$ are exchangeable.*

External stability means that any two tuples with equivalent relations between its instances, can be also extended in an equivalent way.

2.2.1. *Examples.* **Time.** Let $X = \mathbf{Z}$ be the set of periods. Define relation $R \subseteq X \times X$ as

$$xRx' \text{ iff } x' = x + 1.$$

Then, statement (2.1) can be rewritten as a pattern

$$\forall x, x', x'' \left(xRx' \wedge x'Rx'' \wedge \theta(x) = 0 \wedge \theta(x') = 0 \right) \Longrightarrow \theta(x'') = 1. \tag{2.8}$$

Language $\mathcal{L}_{Time} = \{R\}$ is externally stable.

Binary choices. It is instructive to define a more general version of the binary preference example. Let B be a countable set of products and let

$$X = \{(b_1, ..., b_d) : b_i \neq j \text{ for any } i \neq j\}$$

be the set of ordered d-tuples of distinct products. For each $i, j \leq d$, define binary relations $R_{ij} \subseteq X \times X$: for any $x, x' \in X$, let

$$xR_{ij}x' \text{ iff } x_i = x'_j.$$

Thus, two instances x and x' are in relation R_{ij}, if the ith coordinate of x is equal to the jth coordinate of x'. Let $Y = \{0, ..., d-1\}$ be the space of outcomes. An interpretation of $\theta(b_1, ..., b_d) = d'$ is that the consumer, faced with choice among products $b_1, ..., b_{d'}$ chooses product $b_{d'+1}$. Denote this language as $\mathcal{L}_{Ordered\ Graph_d} = \{R_{ij}\}$ (terminology "*Ordered Graph*" is explained below). It is easy to check that this language is externally stable.

Consider a case $d = 2$. One can express the transitivity of θ as a pattern:

$$\forall x, x', x'' \left(xR_{21}x' \wedge xR_{11}x'' \wedge x'R_{22}x'' \wedge \theta(x) = 0 \wedge \theta(x') = 0 \right) \Longrightarrow \theta(x'') = 0.$$

This is because, Then, $xR_{21}x' \wedge xR_{11}x'' \wedge x'R_{22}x''$ implies that there are three different products $b_1, b_2, b_3 \in B$ such that $x = (b_1, b_2)$, $x' = (b_2, b_3)$ and $x'' = (b_1, b_3)$.

Chemical research. Let B be the set of substances More generally, let $X = \{x \subseteq B : |x| = d\}$ be the set of all d-element subsets of B. Let $Y = \{0, 1\}$ be the space of outcomes, where) is interpreted as "no reaction" and 1 means a "reaction".

Model θ describes outcomes of combining all combinations of d substances. For any $j \leq d$, define binary relations $R_j \subseteq X \times X$: for any $x, x' \in X$, xR_jx' iff $|x \cap x'| = j$. Thus, two instances x and x' are in relation R_j, if the intersection of x and x' has exactly j elements. Denote this language as $\mathcal{L}_{Graph_d} = \{R_j\}$ (terminology "*Graph*" is explained below). One checks that language $\mathcal{L}_{Graph_d}$ is externally stable.

Suppose that $d = 2$. Statements (2.3) and (2.4) of the acid-alkaline theory correspond to patterns, respectively,

$$\forall x, x', x'' \left(xR_1x' \wedge xR_1x'' \wedge x'R_1x'' \wedge \theta(x) = 1 \wedge \theta(x') = 1 \right) \Longrightarrow \theta(x'') = 0,$$
$$\forall x, x', x'' \left(xR_1x' \wedge xR_1x'' \wedge x'R_1x'' \wedge \theta(x) = 1 \wedge \theta(x') = 0 \right) \Longrightarrow \theta(x'') = 1.$$

Netflix recommendation problem. In the Netflix problem, there are two binary relations $R_C, R_M \subseteq X \times X$: for any (c, m),$(c', m') \in X$,

$$(c, m)\, R_C\, (c', m') \text{ iff } c = c',$$
$$(c, m)\, R_M\, (c', m') \text{ iff } m = m'.$$

One can use relational language $\mathcal{L} = \{R_C, R_M\}$ to formulate statements (2.5) and (2.6). For example, (2.5) is equivalent to theory (1.3) where R is replaced by R_M.

More generally, let $X = X^1 \times ... \times X^d$ be a product of d infinite sets. One can define binary relations $R_j \subseteq X \times X$: for any $x, x' \in X$, xR_jx' if and only if $x_j = x'_j$. Thus, two instances x and x' are in relation R_j, if the jth coordinate of x is equal to the jth coordinate of x'. Denote this language as $\mathcal{L}_{Product_d} = \{R_j\}$. This language is externally stable.

Suppose that $d = 2$. Consider statements (2.5) and (2.6). They can be restated as patterns:

$$\forall x, x' \left(xR_Mx' \wedge \theta(x) = 1 \right) \Longrightarrow \theta(x') = 0,$$
$$\forall x, x', x'', x''' \left(\begin{array}{c} xR_Mx' \wedge xR_Cx'' \wedge x'R_Cx''' \wedge x''R_Mx'' \\ \wedge \theta(x) = 1 \wedge \theta(x') = 0 \wedge \theta(x'') = 0 \end{array} \right) \Longrightarrow \theta(x''') = 0.$$

2.3. **Algebraic definition of language.** Next, I present an alternative definition of a language. The idea is to work with a class of transformations of X that preserve relations in the language. Such a class of transformations should form an algebraic group.[6] Formally, a *permutation* of X is any bijection from X to X. A set G of permutations is

[6]This idea is related to a model of learning from distinguishable histories in (Crawford and Haller 1990). More broadly, this is related to XIXth century Erlangen Program (see (Klein 1979)) that advocated the use of groups of symmetric transformations for a systematic approach to geomatric problems.

a *group* if (a) $\mathrm{id}_X \in G$, (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)). I refer to $G \longmapsto X$ as a *(algebraic) language*.

A *(algebraic) pattern* is a pair (S, τ) of finite set $S \subseteq X$ and function $\tau : S \to Y$. Set S is also called a *support* of pattern p. Let $\mathcal{P} = \bigcup_{S \subseteq X,\ S \text{ finite}} Y^S$ denote the set of all patterns.

Definition 2. *Model* θ omits pattern (S, τ) *if, for each permutation* $g \in G$*, there is* $x \in S$*, such that* $\theta(g \cdot x) \neq \tau(x)$*.*

Model θ omits pattern if, informally, it omits any of its permutations. For any model θ, any $g \in G$, define permutation of θ as $\theta_g(x) := \theta(g \cdot x)$. Thus, θ_g is a proper model of the world; if θ omits pattern (S, τ), then θ_g omits it as well.

A *(algebraic) theory* $T \subseteq \mathcal{P}$ is any subset of patterns. Set$\mathcal{M}(T)$ and the consistency of (algebraic) theory T are defined in an analogous way to the respective relational definitions. Two observations are useful: (a) $\mathcal{M}(T)$ is weakly decreasing in the set sense and (b) $\mathcal{M}(T)$ is invariant to permutations: For any model θ, any $g \in G$, if $\theta \in \mathcal{M}(T)$, then $\theta_g \in \mathcal{M}(T)$.

It turns out that relational and algebraic definitions of the language are equivalent when the space of instances is countable and the relational language is externally stable. This is made clear by two results. Say that language $\mathcal{L}$ *is generated by group action* $G \longmapsto X$ if for any k, any tuples $\bar{x}, \bar{x}' \in X^k$, tuples $\bar{x}$ and $\bar{x}'$ are exchangeable if and only if there is $g \in G$, such that

$$(g \cdot x_1, ..., g \cdot x_k) = (x'_1, ..., x'_k).$$

Proposition 1 shows any externally stable language is generated by a group of permutations. Proposition 2 shows that relational language and the corresponding algebraic language have the same powers of expression. These two results are proven in Appendix A.

Proposition 1. *If* X *is countable and relational language* $\mathcal{L}$ *is externally stable, then it is generated by a certain group of permutations* $G \longmapsto X$*.* G *can defined as the set of all permutations on* X *that preserve all relations on* X *: for any* i*, any* $\bar{x} \in X^{k_i}$*, any* $g \in G$

$$R_i(x_1, ..., x_{k_i}) \Leftrightarrow R_i(g \cdot x_1, ..., g \cdot x_{k_i}). \tag{2.9}$$

Proposition 2. *Suppose that* $\mathcal{L}$ *is generated by* $G \longmapsto X$*. For any algebraic pattern* (S, τ)*, there is a relational pattern (1.1) such that model* θ *omits the group pattern if and only if it satisfies the relational one.*

For any relational pattern (1.1), there is a finite set of algebraic patterns P *such that*

model θ satisfies the relational pattern if and only if it omits any group pattern from set P.

Although non-standard, the algebraic definition is more convenient for the purposes of this paper. Below, I always use the algebraic definition of language.

2.3.1. *Examples.* **Time.** Relational language $\mathcal{L}_{Time}$ is generated by the following group:

Example 1 (*Time*)**.** *Let $G = \mathbf{Z}$ be the set of integers and let $G \longmapsto X$ act as a shift: for any $g \in G$, any $x \in X$, let*

$$g \cdot x = g + x.$$

The relational pattern (2.8) corresponds to algebraic pattern (S, τ), where

$$S = \{t, t+1, t+2\} \text{ for some } t \text{ and}$$
$$\tau(t) = \tau(t+1) = \tau(t+2) = 0.$$

Binary choices. Relational language $\mathcal{L}_{Ordered\ Graph_d}$ is generated by the following group:

Example 2 (*Ordered Graph$_d$*)**.** *Let $G = \Pi_B$ be the group of all permutations on B. (Π_B is also known as the symmetric group on B.) Define action $\Pi_B \longmapsto I$: for any $g \in G$, any $(b_1, ..., b_d) \in I$, let*

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

When $d = 2$, one treat X as the set of all ordered edges in a graph with set of nodes B. Language *Ordered Graph$_2$* allows to state transitivity. Namely, it corresponds to omitted pattern (S, τ), where

$$S = \{(a,b), (a,c), (b,c)\} \text{ and} \tag{2.10}$$
$$\tau_0(a,b) = 0, \tau(b,c) = 0, \tau(a,c) = 1,$$

Chemical research. Relational language $\mathcal{L}_{Graph_d}$ is generated by the following group:

Example 3 (*Graph$_k$*)**.** *Let B be a countable set and let Let $G = \Pi_B$ be the group of all permutations on B. Define group action $G \longmapsto I$: for any $g \in G$ and for any $\{b_1, ..., b_d\} \in X$, let*

$$g \cdot \{b_1, ..., b_d\} := \{g \cdot b_1, ..., g \cdot b_d\} \in X.$$

When $d = 2$, one can think about X as the set of all (unordered) edges in a graph with set of nodes B. Language $Graph_2$ can be used to state the acid-alkaline theory. Specifically, (2.3) corresponds to pattern (S, τ), where

$$S = \{\{a, b\}, \{a, c\}, \{b, c\}\} \text{ and} \tag{2.11}$$
$$\tau(\{a, b\}) = 1, \tau(\{a, c\}) = 1, \tau(\{b, c\}) = 1,$$

and (2.4) corresponds to pattern (S, τ') with the same support and where

$$\tau'(\{a, b\}) = 1, \tau'(\{a, c\}) = 0, \tau'(\{b, c\}) = 0. \tag{2.12}$$

Netflix recommendation problem. Relational language $\mathcal{L}_{Product_d}$ is generated by the following group:

Example 4 ($Product_d$)**.** *For each j, let Π_{X^j} be the group of all permutations on set X^j.*[7] *Let $G = \Pi_{X^1} \times .... \times \Pi_{X^d}$, and define the group action $G \longmapsto X$: for any $(g_1, ..., g_d) \in G$, any $(x_1, ..., x_d) \in X$, let*

$$(g_1, ..., g_d) \cdot (x_1, ..., x_d) := (g_1 \cdot x_1, ..., g_d \cdot x_d).$$

Language $Product_2$ can be used to state patterns (2.5) and (2.6). Specifically, model θ satisfies (2.5) if it omits pattern (S, τ), where

$$S = \{(c, m), (c', m)\} \text{ and } \tau(c, m) = 1, \tau(c', m) = 0,$$

for some $c \in C, m, m' \in M$. Model θ satisfies (2.6) if it omits pattern (S', τ'), where

$$S' = \{(c, m), (c, m'), (c', m), (c', m')\} \text{ and} \tag{2.13}$$
$$\tau'(c, m) = 1, \tau'(c', m) = 0, \tau'(c', m) = 0, \tau'(c', m') = 1.$$

At this moment, the reader may wonder about the difference between languages $Product_d$, $Graph_d$, and *Ordered* $Graph_2$. Let me just say here that both languages are formally different because the groups of permutations are defined in a different way. In Section 5, I examine the properties of both languages more closely and I describe important, structural differences between them.

[7] This group of permutations has been used in a generalization of de Finetti's exchangeability in (Aldous 1981) (see also (Kallenberg 2005)). In (Blume 2004), it is used to study properties of natural languages.

2.4. **Group properties.** Here, I describe some useful properties of group languages. Group action $G \longmapsto X$ induces a group action $G \longmapsto 2^X$ on the set 2^X of all subsets of X: for any $g \in G$, any $S \subseteq X$, let

$$g \cdot S = \{g \cdot x : x \in S\} \subseteq X.$$

Abusing terminology, I say that set $g \cdot S$ is a permutation of S. Similarly, for each k, group action $G \longmapsto X$ induces a group action $G \longmapsto X^k$ on the set of k-tuples of elements of X: for any $g \in G$, any $\bar{x} = (x_1, ..., x_k) \in X^k$, let

$$g \cdot \bar{x} = (g \cdot x_1, ..., g \cdot x_k) \subseteq X^k.$$

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

$$G_U = \{g \in G, g \cdot U = U\}. \tag{2.14}$$

Group G_U describes a local behavior of group G on set U.

Definition 3. *Finite $U \subseteq X$ is* local (under group action G) *if for any k, any pair of k-tuples $\bar{x}, \bar{x}' \in U^k$ such that $g \cdot \bar{x} = \bar{x}'$ for some $g \in G$, there exists $g_U \in G_U$ such that $g_U \cdot \bar{x} = \bar{x}'$.*

Take any two tuples $\bar{x}, \bar{x}'$ of elements of U and suppose that there is a permutation g that takes one tuple into the other, $g \cdot \bar{x} = \bar{x}'$. If U is local, then permutation g can be chosen so that it keeps all elements of set U inside 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. Note for future reference, that if U is local and $g \in G$ is a permutation, then $g \cdot U$ is also local; in fact, $G_{g \cdot U} = g \cdot G_U \cdot g^{-1}$.

Definition 4. *Group action $G \longmapsto X$ is* locally generated *if there is a sequence of local sets $X_1 \subseteq X_2 \subseteq ... \subseteq X$, such that for any finite subset $S \subseteq X$, there is $n \geq 1$ and a permutation $g \in G$ such that $g \cdot S \subseteq X_n$.*

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 language $Product_d$ from Example 4. Of course, this language is transitive. It is easy to verify that any product set $U = U^1 \times ... \times U^d$ for finite $U^d \subseteq X^d$ is local under the group action $G \longmapsto X$. Take any (strictly) increasing sequences of finite sets

$X_1^j \subset X_2^j \subset ... \subset X^j$ for $j = 1, ..., d$. Given this, sets $X_n = X_n^1 \times ... \times X_n^d$ form a generating sequence.

Next, consider language *Ordered Graph*$_d$ from Example 2. Take any finite set $B' \subseteq B$ and define $X(B') \subseteq I$ as the set of all ordered d-tuples of distinct elements of B' :

$$X(B') = \{(b_1, ..., b_d) : b_j \in B', b_j \neq b_{j'} \text{ for } j \neq j'\}. \tag{2.15}$$

One easily checks that $X(B')$ is local for any finite $B' \subseteq B$. For any increasing sequence of finite sets $B_1 \subset B_2 \subset ... \subset B$, sets $X(B_n)$ form a generating sequence. Hence, *Ordered Graph*$_d$ is locally generated. Clearly, it is also transitive. By similar arguments, language *Graph*$_d$ from Example 3 is transitive and locally generated.

Not all group actions are locally generated. No finite subset of X is local under group action from Example 1. Hence, *Time* is not locally generated.

3. Learning

3.1. **Model.** An *instance process* is a sequence of instances $\bar{x}_\infty = (x_1, x_2, ...)$ such that no instances get ever repeated, $x_s \neq x_t$ for $s \neq t$. The decision-maker observes an instance process. In period t, she predicts the outcome of instance x_t. After the prediction is made, the decision maker is informed about the true outcome $\theta(x_t)$. To fix attention, assume that the decision maker receives a payoff of 1 if her prediction is correct and 0 if not. (More general payoffs do not change any of the results.) In period t, the decision maker has access to the database of past observations from periods $s < t$. A mapping $l : \bigcup_t (X \times Y)^{t-1} \times X \to \Delta Y$ is called a *learning rule*. In general, I allow for randomized predictions; for example,

$$l\left((x_s, y_s)_{s<t}, x_t\right)(y)$$

is the probability that learning rule assigns to the fact that outcome of x_t is equal to y, based on the database of past observations $(x_s, y_s)_{s<t}$.

Pure learning rule l makes deterministic predictions, i.e. after each history, it assigns probability 1 to certain outcome. Such learning rules have an important property. In each period t, the decision maker needs to remember only those past observations in which the prediction was incorrect; all other observations can be deduced from the assumption that the prediction was correct. (If $Y = \{0, 1\}$, then the decision maker needs to remember only periods in which mistakes were committed.) If the number of mistakes is small, this reduces the burden imposed on the decision maker's memory.

For any instance process $\bar{x}_\infty$, any learning rule l, any model θ, define

$$U_t\left(l,\bar{x}_\infty,\theta\right)=\frac{1}{t}\sum_{s=1}^{t}l\left(\left(x_u,\theta\left(x_u\right)\right)_{u<s},x_s\right)\left(\theta\left(x_s\right)\right)$$

This is the t-period average payoff from learning rule l given instance process $\bar{x}$ and model θ.

3.2. **Sufficiency for deduction.** Take any theory $T\subseteq\mathcal{P}$. Consider a decision maker who (a) cares only about the long-run average payoffs from prediction and (b) knows that theory T holds, but does not know which of the models that satisfy theory T is the true one. In particular, she considers all models which omit theory T as plausible. Following (Gilboa and Schmeidler 1989), I assume that the decision maker exhibits an extreme version of the uncertainty aversion and evaluates learning rule l according to the worst-case payoff.[8]. In particular,

$$\inf_{\theta\in M(T)}U_t\left(l,\bar{x}_\infty,\theta\right) \tag{3.1}$$

is the worst-case t-period payoff of a decision maker who knows that theory T holds, uses learning rule l, and observes instances according to the process $\bar{x}_\infty$. Alternatively, one can think about (3.1) as a payoff of a statistician in the zero-sum game against Nature ((Wald 1950), (Wald 1949), (Blackwell and Girshick 1954), (Schwarz 1994)).

By definition, expression (3.1) is never larger than 1. A learning rule that each period chooses an outcome from the same uniform distribution on Y guarantees payoff $\frac{1}{|Y|}$ for each model of the world, instance process and any number of periods.

One is especially interested in theories that are informative, i.e. theories that can be used to correctly predict almost all outcomes of an instance process.

Definition 5. *Theory T is* sufficient for deduction *if there is an instance process $\bar{x}$ and a learning rule l, such that*

$$\liminf_{t\to\infty}\inf_{\theta\in\mathrm{M}(T)}U_t\left(l,\bar{x}_\infty,\theta\right)=1. \tag{3.2}$$

Theory T is sufficient for deduction if there is an instance process $\bar{x}_\infty$ and a learning rule l such that for any $\varepsilon>0$ there is a period t' such that for any $t\geq t'$ the t-period average payoff from using learning rule l is guaranteed to be not smaller than $1-\varepsilon$ for

[8](Chen and Epstein 2002), (Epstein and Schneider 2003), (Epstein 2006) and (Epstein, Noor, and Sandroni 2006) develop a systematic study of axiomatic foundation of learning under ambiguity with discounting. Also, the robust control literature (see (Hansen and Sargent forthcoming)) uses a discounted version of formula (3.2) both in a theoretical and empirical framework.

any model of the world $\theta \in M(T)$.[9] Suppose that the decision maker knows theory T, and T is sufficient for deduction. Given this, she can choose an instance process and a learning rule so to be sure to deduce almost all outcomes of the instance process.

Definition 5 is quite weak. The decision maker can choose any learning rule that depends on all past observations and that may make randomized predictions. The decision maker can also choose an instance process. The latter can be particularly troubling. In some situations (for example, chemists in the laboratory), the decision maker may influence the order of observations. On the other hand, the definition seems to be too weak if the instance process is given as a parameter of the learning problem. In general, it would be to strong to require that (3.2) holds for *any* instance process. Instead, one can draw an analogy to statistics, where the consistency of an estimator can be shown only for datasets of sufficient size. Below, I describe instances processes that satisfy certain sufficient data condition. I show that if a theory is sufficient for deduction in the sense of Definition 5, then there is a learning rule l that guarantees almost perfect payoffs for *any* instance process that satisfies the sufficient data condition.

3.3. **Entropy characterization.** For any $A \subseteq X$, any theory consistent T, define

$$\mathcal{M}(A;T) := \left\{\tau \in Y^A : \tau = \theta|_A \text{ for some } \theta \in \mathcal{M}(T)\right\}, \tag{3.3}$$

$$\mathcal{E}(A;T) := \frac{1}{|A|}\log|\mathcal{M}(A;T)|. \tag{3.4}$$

Set $\mathcal{M}(A,T)$ consists of all restrictions of models $\theta \in \mathcal{M}(T)$ to instances in set A. Clearly, the more models in theory T, the larger is set $\mathcal{M}(A;T)$. $\mathcal{E}(A;T)$ is an entropy of theory T on set of instances A; $1-\mathcal{E}(A;T)$ measures the informational content of

[9]Alternatively, one can define the long-run payoffs as

$$\inf_{\theta\in\mathcal{M}(P)} \lim \inf_{T\to\infty} U_T(l, \bar{x}_\infty, \theta).$$

This puts less weight on the short-run payoffs than formula (3.2). To see it, consider a simple example. Suppose that set $\mathcal{M}$ consists of all functions $\theta : \mathbf{Z} \to \{0,1\}$ such that there is $z \in \mathbf{Z}$, so that $\theta(z') = 1$ for each $z' \geq z$. Consider an instance process $\bar{z}_\infty = 1, 2, 3, ...$ and a learning rule l that always predicts 1, $l \equiv 1$. Then,

$$\inf_{\theta\in\mathcal{M}} \lim \inf_{T\to\infty} U_T(l, \bar{x}_\infty, \theta) = 1.$$

On the other hand, for any learning rule l and any number of periods T, there is a model θ such that the T-period average payoff is not higher than $\frac{1}{2}$ (the best strategy of statistician in such a zero-sum game is to predict probability $\frac{1}{2}$ for each of the outcomes):

$$\sup_l \lim \inf_{T\to\infty} \inf_{\theta\in\mathcal{M}(P)} U_T(l, \bar{x}_\infty, \theta) = \frac{1}{2}.$$

Arguably, the criterion used in this paper is more appriopriate when the DM is patient but does not live infinitely long.

theory T on set A. Note that the entropy is invariant to permutations: for any $g \in G$, $\mathcal{E}(A;T) = \mathcal{E}(g \cdot A;T)$. Finally,

Definition 6. Entropy of theory T *is the infimum entropy across all finite sets* A :

$$\mathcal{E}(T) := \inf_{A \subseteq X, A \text{ is finite}} \mathcal{E}(A;T).$$

Here, $1 - \mathcal{E}(T)$ is a measure of the informational content of theory T. The next two results establish a connection between the entropy of a theory and the payoffs from prediction. The first Lemma shows that entropy leads to an upper bound on payoffs and the second Lemma finds a lower bound.

Lemma 1. *For any* $c \geq 1$, *define function* $h_c : [0,1] \to [0,1]$ *as the unique solution to the equation*

$$e := h_c(e)\left(c + 2 + \log\frac{1}{h_c(e)}\right).$$

Function h_c *is well-defined, continuous, and increasing. For any theory* T, *any instance process* $\bar{x}_\infty$, *any learning rule* l *and any* t,

$$\begin{aligned}\inf_{\theta \in \mathrm{M}(T)} U_t(l, \bar{x}_\infty, \theta) &\leq 1 + \frac{1}{t} - h_{\log|Y|}\left(\mathcal{E}\left(\{x_1, ..., x_t\}\, T\right)\right). \\ &\leq 1 + \frac{1}{t} - h_{\log|Y|}\left(\mathcal{E}(T)\right).\end{aligned}$$

The Lemma is relatively standard; its proof can be found in Appendix C.

Lemma 2. *For any theory* T, *any instance process* $\bar{x}_\infty$, *and any increasing sequence* $1 = t_0 < t_1 < ...$, *there is a pure learning rule* l, *such that for any* t

$$\inf_{\theta \in \mathrm{M}(T)} U_t(l, \bar{x}_\infty, \theta) \geq 1 - \sum_{k=1}^{k^*} \frac{t_k - t_{k-1}}{t} \mathcal{E}\left(\left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\}; T\right) - \frac{t - t_{k^*}}{t}, \tag{3.6}$$

where $k^* = \max\{k : t_k \leq t\}$.

The proof is contained in Appendix D. The idea is fairly simple. In each period $t_{k-1} < t \leq t_k$, let

$$U_t = \left\{\tau \in \mathcal{M}\left(\left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\}; T\right) : \tau(x_s) = y_s, t_{k-1} < s < t\right\}$$

be the set of model restrictions $\tau \in \mathcal{M}\left(\left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\}; T\right)$, which agree with observed outcomes $y_{t_{k-1}+1}, ..., y_{t-1}$. In particular, $U_{t_{k-1}}$ is equal to the whole set of model restrictions $\mathcal{M}\left(\left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\}; T\right)$. For any outcome $y \in Y$, let $n_t(y)$ be the number of model restrictions in U_t for which $\tau(x_t) = y$. Learning rule l predicts that outcome of instance x_t is equal to some $y_{\max} \in \arg\max_y n_t(y)$. Each period, either the prediction

of l is correct or not. If the former, then the decision maker does not make a mistake. If the latter, then the decision maker makes a mistake, but she also reduces U_t by at least $\frac{1}{2}$, $|U_{t+1}| \leq \frac{1}{2}U_t$. Because the number of such reductions is constrained by the entropy, learning rule l achieves the appropriate bound.

3.4. **Locally generated and transitive languages.** If language $G \longmapsto X$ is locally generated and transitive, a sharper characterization is possible.

Definition 7. *Instance process $\bar{x}_\infty$ satisfies* sufficient data condition*, if there is $\delta > 0$, a generating sequence of sets $X_1 \subseteq X_2 \subseteq ...$, an increasing sequence of periods $1 = t_0 < t_1 < ...$, a sequence of permutations $g_k \in G$, and a function $n : \mathbf{N} \to \mathbf{N}$ such that*

(1) *for any $k \geq 1$, $\left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\} \in g_k \cdot X_{n(k)}$ and $t_k - t_{k-1} \geq \delta \left|X_{n(k)}\right|$,*
(2) $\lim_{k\to\infty} n(k) = \infty$,
(3) $\lim_{k\to\infty} \frac{t_k}{t_{k-1}} = 0$.

The sufficient data condition requires that the data collected are related to each other in a particular way. Specifically, the instance process can be divided into intervals $\left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\}$ such that elements of kth interval belong to a local set $g_k \cdot X_{n(k)}$ (recall that any permutation of a local set is local). Parts (2) and (3) guarantee that the size of local sets increases but not too quickly. Definition 7 is not vacuous - there are instance processes that satisfy the sufficient data condition.

Proposition 3. *Suppose that language $G \longmapsto X$ is transitive and locally generated. Theory T is sufficient for deduction if only if $\mathcal{E}(T) = 0$. Moreover, if T is sufficient for deduction, then for any process $\bar{x}_\infty$ that satisfies sufficient data condition, there is a pure learning rule l such that (3.2) holds.*

The first part of the Proposition leads to a test on whether theory T is sufficient for deduction. It turns out that this is characterized completely by entropy of T, $\mathcal{E}(T)$. In particular, T is sufficient for deduction if for any $\epsilon > 0$, there exists a finite $A \subseteq X$ such that

$$\mathcal{E}(A; T) \leq \varepsilon. \tag{3.7}$$

Moreover, if theory is sufficient for deduction, then, for *any* instance process that satisfies Definition 7, there exists a learning rule that guarantees almost perfect predictions. This substantially strengthens the requirement in Definition 5.

The proof of the Proposition is based on the following result (to the best of my knowledge, the result is novel).

Lemma 3. *Suppose that $G \longmapsto X$ is transitive. For any local U, and any $A \subseteq U$, any theory T, $\mathcal{E}(U; T) \leq \mathcal{E}(A; T)$.*

This says that entropy on a local set is not larger than entropy on any of its subsets. The proof of the Lemma can be found in Appendix B.

Proof of Proposition 3. The "only if" direction of the first part of the Proposition follows from Lemma 1. To show the "if" part, suppose that $\mathcal{E}(T) = 0$. Then, there is a sequence of finite sets $A_m \subseteq X$ such that $\mathcal{E}(A_m; T) \leq \frac{1}{m}$. Let $\bar{x}_\infty$ be an instance process that satisfies the sufficient data condition with $\delta > 0$, generating sequence $X_1 \subseteq X_2 \subseteq ...$, increasing sequence $1 = t_0 < t_1 < ...$, a sequence of permutations $g_k \in G$, and a mapping $n : \mathbf{N} \to \mathbf{N}$. Because X_n is a generating sequence, there is a function $m : \mathbf{N} \to \mathbf{N}$ and permutations $g'_m \in G$ such that for any sufficiently high n, $g'_{m(n)} \cdot A_{m(n)} \subseteq X_n$ and $\lim_{n\to\infty} m(n) = \infty$. By Lemma 3 and by the permutation invariance,

$$\begin{aligned}
&\mathcal{E}\left(\left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\}; T\right) \\
&= \frac{1}{t_k - t_{k-1}} \log\left|\mathcal{M}\left(\left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\}; T\right)\right| \\
&\leq \frac{1}{t_k - t_{k-1}} \log\left|\mathcal{M}\left(X_{n(k)}; T\right)\right| \leq \frac{\left|X_{n(k)}\right|}{t_k - t_{k-1}} \mathcal{E}\left(X_n; T\right) \\
&\leq \frac{1}{\delta} \mathcal{E}\left(A_{m(n)}; T\right) \leq \frac{1}{\delta} \frac{1}{m\left(n\left(k\right)\right)}.
\end{aligned}$$

By Lemma 2, there is a pure learning rule l such that bound (3.6) holds for any t. Let $k_t = \sup\left(k : t_k \leq t\right)$. Then,

$$\inf_{\theta \in M(T)} U_t\left(l, \bar{x}_\infty, \theta\right) \geq 1 - \sum_{k=1}^{k_t} \frac{t_k - t_{k-1}}{t} \frac{2}{\delta} \frac{1}{m\left(n\left(k\right)\right)} - \frac{t - t_{k_t}}{t}.$$

Because $\lim_{k\to\infty} m\left(n\left(k\right)\right) = \infty$, convergence (3.2) follows. This also implies the second part of the Proposition. □

4. Main results

This Section contains the main results of the paper.

4.1. **Generic sets.**

Here and in the rest of the paper I always assume that language $G \longmapsto X$ is transitive and locally generated.

Definition 8. *Finite set $S \subseteq X$ is* generic *if for each $\varepsilon > 0$, 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 there is a local set U, such that any of its subsets of sufficient size contains a permutation of S. In a sense, generic sets can be found almost everywhere. Section 5 presents examples of generic sets in different languages.

Generic sets are especially interesting as supports of patterns. One might expect that theories that consist of patterns with generic support are more informative than theories without such patterns. This intuition is confirmed by the next result.

Proposition 4. *Suppose that* $Y = \{0,1\}$. *If* S *is generic, then, for any* $\varepsilon > 0$, *there is a local* $U \subseteq X$ *such that for any* $\tau : S \to Y$,

$$\mathcal{E}(U, \{(S,\tau)\}) \leq \varepsilon.$$

In particular, $\mathcal{E}(\{(S,\tau)\}) = 0$.

The Proposition is central to the analysis of this paper. To see exactly what the Proposition says, consider any one-pattern theory $T = \{(S,\tau)\}$. For any set $U \subseteq X$, theory T imposes a restriction on the cardinality of the set of model restrictions $\theta|_U$ that are consistent with theory T. In particular, if $U = S$, then the restriction is not trivial but not necessarily a very serious one:

$$\mathcal{E}(S, \{(S,\tau)\}) \leq \frac{1}{|S|} \log\left(2^{|S|} - 1\right) < 1$$

for large S. However, if S is generic, then there is a sufficiently large local set U, on which the cardinality of model restrictions of models that satisfy theory T is small. In other words, the restriction imposed by patterns with generic support becomes significant on sufficiently large local sets U.

Section 6 states a slightly stronger version of the Proposition and explains how it relates to results from the Vapnik-Chervonenkis statistical learning theory.[10]

How many subsets are generic? This depends on the language $G \longmapsto X$.

Definition 9. *Define* tightness *of transitive and locally generated language* $G \longmapsto X$ *as the largest number* $t(G) \in [0,1]$ *such that for any finite* $A \subseteq X$, *there is a generic* $S \subseteq A$ *such that* $|S| \geq t(G)|A|$. *Language* $G \longmapsto X$ *is* tight *if* $t(G) > 0$.

For any language $G \longmapsto X$, $t(G) \in [0,1]$. Section 5 characterizes the tightness of various languages.

[10]When $G = Graph_2$, Proposition 4 can be restated as the *omitted subgraph problem* known in the graph theory ((Promel and Steger 1991), (Promel and Steger 1993) and (Promel and Steger 1992); see also (Balogh, Bollobas, and Weinreich 2000), (J. Balogh and Weinreich 2001) and (Scheinerman and Zito 1994)). When $G = Product_2$, N. Alon suggested the Proposition can be obtained as a corollary to Lemma 6.3 from (Alon, Fischer, and Newman 2005). The advantage of the current result is that it applies widely to all transitive and locally generated languages, including $Product_d$, $Graph_d$, for $d > 2$. It also points to the genericity of the support of a pattern as the main force behind the result. (The definition of a generic set is, to the best of my knowledge, novel.)

4.2. **Case $Y = \{0,1\}$.** It is helpful to consider first only the binary case, $Y = \{0,1\}$.

Theorem 1. *Suppose that $Y = \{0,1\}$ and that language $G \longmapsto X$ is tight. Theory T is sufficient for deduction if and only if it logically implies a pattern with generic support. If theory T is not sufficient for deduction, then for any learning rule l, any instance process $\bar{x}_\infty$, and any period t*

$$\inf_{\theta \in \mathcal{M}(T)} U_t\left(l, \bar{x}_\infty, \theta\right) \leq 1 + \frac{1}{t} - h_{\log 2}\left(t\left(G\right)\right).$$

Proof. The "if" direction of the Theorem is a corollary to Propositions 3 and 4.

Let $t\left(G\right) > 0$ be the tightness of language $G \longmapsto X$. Take any finite $A \subseteq X$ and find generic $S \subseteq A$, $|S| \geq t\left(G\right)|A|$. If $\mathcal{M}\left(T\right) \nsubseteq \mathcal{M}\left(\{(S,\tau)\}\right)$, then, for any $\tau : S \to Y$, there is a model $\theta_\tau \in \mathcal{M}\left(T\right)$ and a permutation g_τ such that for any $x \in S$, $\theta_\tau\left(g_\tau \cdot x\right) = \tau\left(x\right)$. Define model $\theta_\tau^*\left(x\right) := \theta_\tau\left(g_\tau \cdot x\right)$ for any $x \in X$. Then, $\theta_\tau^* \in \mathcal{M}\left(T\right)$ and $\theta_\tau^*|_S = \tau$. This implies that

$$\begin{aligned}
\mathcal{E}\left(A,T\right) &= \frac{1}{|A|}\log \mathcal{M}\left(A,T\right) \geq \frac{1}{|A|}\log \mathcal{M}\left(S,T\right) \\
&\geq \frac{1}{|A|}\log\left|\tau : \theta_\tau^* \in \mathcal{M}\left(T\right)\right| = \frac{1}{|A|}\log 2^{|S|} \\
&= \frac{1}{|A|}\left|S\right| \geq t\left(G\right) > 0.
\end{aligned}$$

Together with Lemma 1, this implies that T is not sufficient for deduction. The last part of the Theorem is a corollary to Lemma 1. □

Suppose that a decision maker expresses theory T in a tight language and that T is sufficient for deduction. The first part of the Theorem says that T implies a theory that consist of only one pattern. If she cares about the cost of storing information, then she can replace theory T with a shorter one.

Because one can always create one pattern out of finitely many, this result is trivial when T already consists of finitely many patterns. The real bite comes when T counts infinitely many patterns. In the latter case, the Theorem allows to replace infinitely many patterns by one, by definition, finitely stated, pattern. In fact, the proof of the Theorem suggests a bound on the size of the generic support of the pattern that replaces theory T.

Corollary 1. *Suppose that there is an instance process $\bar{x}$, a period t and a learning rule l such that*

$$\inf_{\theta \in \mathcal{M}(T)} U_t\left(l, \bar{x}_\infty, \theta\right) > 1 + \frac{1}{t} - h_{\log 2}\left(t\left(G\right)\right).$$

Then, theory T is sufficient for deduction, and it logically implies a pattern (S,τ) with generic support of cardinality at most $|S| \leq 1 + t \cdot t(G)$.

Note that, by definition, the hypothesis of the corollary is true for any theory sufficient for deduction.

Proof. Let $S' \subseteq \{x_1, ..., x_t\}$ be a generic set of size at least $t \cdot t(G)$ and let $S \subseteq S'$ be a subset such that $t \cdot t(G) \leq |S| \, 1 + t \cdot t(G)$. Set S is generic, because any subset of a generic set is generic. By Lemma 1,

$$\mathcal{M}(S;T) \leq \mathcal{M}(\{x_1, ..., x_t\};T) < 2^{t \cdot t(G)} \leq 2^{|S|}.$$

Hence, there exists $\tau : S' \to \{0,1\}$ such that all models that satisfy theory T omit (S',τ). □

The second part of the Theorem drives the uniform wedge between theories that are and that are not sufficient for deduction. In particular, for any tight language, there is $h > 0$ such that for each theory T there are two possibilities. If T is sufficient for deduction, then the decision maker is guaranteed almost perfect predictions. If it is not, then, in the worst-case, the decision maker will commit a mistake every $1/h$ periods, *no matter what learning rule she uses or what order she observes instances.* The value of h depends only on the language, not on the theory.

4.3. **Case Y finite.** Here, I extend Theorem 1 to the general case of Y finite. I show that any theory that is sufficient for deduction implies a theory that is sufficient for deduction and that consists of finitely many patterns.

Let $L = \lceil \log |Y| \rceil$ be equal to the smallest natural number not smaller than $\log |Y|$. Fix an injection $v : Y \to \{0,1\}^L$. Fix pattern (S,τ), where $\tau : S \to Y$. For any mapping $\hat{\tau} : S \to \{0,1\}^L$, any $l \leq L$, say that pattern (S,τ) is l-consistent with $\hat{\tau}$ if for any $x \in S$

$$\hat{\tau}_l(x) = 1 \text{ if and only if } \tau(x) \in \{y : v_l(y) = 1\}.$$

Define theory $T(S,\hat{\tau}) \subseteq \mathcal{P}$:

$$T(S,\hat{\tau}) = \bigcup_l \{(S,\tau) : (S,\tau) \text{ is } l\text{-consistent withy } \hat{\tau}\}.$$

There is at most $L|Y|^{|S|}$ patterns in theory $T(S,\hat{\tau})$.

Theorem 2. *Suppose that language $G \longmapsto X$ is tight. Theory T is sufficient for deduction if and only if there are $S \subseteq X$ and $\hat{\tau} : S \to \{0,1\}^L$ such that T logically implies $T(S,\hat{\tau})$.*

If theory T is not sufficient for deduction, then for any learning rule l, any instance process $\bar{x}_\infty$, and any period t

$$\inf_{\theta\in\mathcal{M}(T)} U_t\left(l,\bar{x}_\infty,\theta\right)\leq 1+\frac{1}{t}-h_L\left(t\left(G\right)\right).$$

Proof. Notice first that for any model θ for any instance x,

$$\theta\left(x\right)=v^{-1}\left(\left(v_1\circ\theta\right)\left(x\right),...,\left(v_L\circ\theta\right)\left(x\right)\right).$$

This means that any model θ is uniquely identifiable from a sequence of binary models $v_l\circ\theta$. For any $A\subseteq X$, any theory T, define

$$\begin{aligned}\mathcal{M}_l\left(T\right)&:=\left\{v_l\circ\theta:\theta\in\mathcal{M}\left(T\right)\right\},\\ \mathcal{M}_l\left(A;T\right)&:=\left\{\theta_l|_A:\theta\in\mathcal{M}_l\left(T\right)\right\},\\ \mathcal{E}_l\left(A;T\right)&:=\frac{1}{|A|}\log\left|\mathcal{M}_l\left(A;T\right)\right|.\end{aligned}$$

Then,

$$\begin{aligned}\mathcal{E}\left(A;T\right)&\leq\frac{1}{|A|}\log\left|\mathcal{M}\left(A;T\right)\right|\\ &\leq\frac{1}{|A|}\log\prod_{l=1}^{L}\left|\mathcal{M}_l\left(A;T\left(S,\hat{\tau}\right)\right)\right|\leq\sum_l\mathcal{E}_l\left(A;T\left(S,\hat{\tau}\right)\right).\end{aligned}$$

Second, because any model $\theta\in\mathcal{M}\left(T\left(S,\hat{\tau}\right)\right)$ omits any pattern that is l-consistent with $\hat{\tau}$, any binary model $\theta_l\in\mathcal{M}_l\left(T\left(S,\hat{\tau}\right)\right)$ omits binary pattern $\left(S,\hat{\tau}_l\right)$. By Proposition 4, for any $\varepsilon>0$, there is a local set U such that for any l,

$$\mathcal{E}_l\left(U;T\left(S,\hat{\tau}\right)\right)\leq\mathcal{E}\left(U;\left\{S,\hat{\tau}_l\right\}\right)\leq\varepsilon.$$

Hence, $\mathcal{E}\left(U;T\left(S,\hat{\tau}\right)\right)\leq L\varepsilon$ and $T\left(S,\hat{\tau}\right)$ is sufficient for deduction.

Suppose that there is no generic S such that $\mathcal{M}\left(T\right)\subseteq\mathcal{M}\left(T\left(S,\hat{\tau}\right)\right)$ for some $\hat{\tau}:S\to\left\{0,1\right\}^L$. Then, for any generic S, there exists l, such that

$$\left|\left\{v_l\circ\theta:\theta\in\mathcal{M}\left(T\right)\right\}\right|=2^{|S|},$$

and $\mathcal{E}\left(S,T\right)=1$. Take now any finite $A\subseteq X$. Then, there is a generic $S\subseteq A$ such that $|S|\geq t\left(G\right)|A|$. Moreover, $\mathcal{E}\left(A,T\right)\geq t\left(G\right)\mathcal{E}\left(S,T\right)=t\left(G\right)$. Proposition 3 implies that T is not sufficient for deduction. □

5. Tightness of various languages

In this Section, I discuss tightness of various languages. In particular, I show that locally generated and transitive languages from Section 2.1 are tight. Most of proofs are postponed till Appendix F.

5.1. **Products.** Suppose that $G_j \longmapsto X_j$ are group actions for $j = 1, ..., d$. Define a product group action $G_1 \times ... \times G_d \longmapsto X_1 \times ... \times X_d$. For any $(x_1, ..., x_d) \in X_1 \times ... \times X_d$, any $(g_1, ..., g_d) \in G_1 \times ... \times G_d$, let

$$(g_1, ..., g_d) \cdot (x_1, ..., x_d) := (g_1 \cdot x_1, ..., g_d \cdot x_d) .$$

In particular, language $Product_d$ from Example 4 is a product of symmetric groups Π_{X_j}, $j \leq d$.

The next result shows that tightness is preserved under products.

Lemma 4. *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, and*

$$t(G) \geq t(G_1) t(G_2) ... t(G_d) .$$

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

Corollary 2. *Consider language* Product_d *from Example 4. Any finite subset is generic and the language is tight.*

As in the Netflix recommendation problem, suppose that $Y = \{0, 1\}$ and that the decision maker uses language $Product_2$. The Corollary says that any nontrivial theory (i.e. any theory that contains at least one pattern) is sufficient for deduction. One can interpret this result in two ways. On one hand, this says that any positive amount of knowledge leads, in the long-run and on average, to perfect predictions. On the other hand, it also says that any positive knowledge is highly restrictive and, as one might expect, hard to come upon in practice.

Regardless of the interpretation, the result indicates very strong deductive properties of language $Product_d$.

5.2. **Ordered graph.** Next, I show that language $Ordered\ Graph_d$ from Example 2 is tight. Let B be a countable set, and let X be equal to the set of ordered d-element tuples of distinct elements from B. For any mutually disjoint finite sets $B^1, ...B^d \subseteq B$, define set of tuples with lth coordinate in set B^l for any $l \leq d$:

$$X\left(B^1, ..., B^d\right) = \left\{(b_1, ..., b_d) : b_l \in B^l\right\} .$$

Proposition 5. *Consider language* $G =$Ordered Graph_d *from Example 2. For any mutually disjoint finite sets* $B^1, ...B^d \subseteq B$, *set* $X\left(B^1, ..., B^d\right) \subseteq X$ *is generic. The language is tight, and*

$$t(G) \geq d^{-d}.$$

I use the Proposition to show that transitivity is sufficient for deduction. Recall that transitivity is equivalent to pattern (S,τ) defined in (2.10). The Proposition does not imply that $S=\{(a,b),(b,c),(a,c)\}$ is generic. (In fact, I argue below that it is not.) However, if model θ omits pattern (S,τ), then it also omits (S^*,τ^*), where

$$S^*=\{(a_1,b_1),(a_2,b_1),(a_1,b_2),(a_2,b_2)\} \text{ and}$$
$$\tau^*(a_1,b_1)=0,\tau^*(a_1,b_1)=0,\tau^*(a_1,b_1)=0,\tau^*(a_1,b_1)=1.$$

Indeed, the above pattern corresponds to a statement: "If a_1 is preferred to b_1, b_1 is preferred to a_2, a_2 is preferred to b_2, then a_1 is preferred to b_2," which is trivially implied by transitivity. By Proposition,

$$S^*=X(\{a_1,a_2\},\{b_1,b_2\})$$

is generic. and any theory that implies (S^*,τ^*) must be sufficient for deduction.

Note that, even in the binary case, Proposition 5 does not guarantee that any non-trivial theory is sufficient for deduction. Consider a theory $\{(S,\tau')\}$, where S is as in (2.10) and

$$\tau'(a,b)=1,\tau'(b,c)=1,\tau(a,c)=1.$$

Lemma 5. *$\{(S,\tau')\}$ is not sufficient for deduction. In particular, S is not generic.*

Proof. This theory is satisfied by all models θ for which there are disjoint sets $B^1\cup B^2=B$, $B^1\cap B^2=\varnothing$ such that for any $b,b'\in B^j$, $j=1,2$, $\theta(b,b')=0$. Indeed, if $\theta(a,b)=1$ and $\theta(b,c)=1$, it means that $a,c\in B^j$ for some $j=1,2$ and $\theta(a,c)=0$. It is important to notice that, because B^1 and B^2 are disjoint, this specification does not constrain outcomes of model θ for instances (b_1,b_2), where $b_j\in B^j, j=1,2$. Hence, for any finite $A\subseteq X$,

$$\left|\{\theta|_{A\cap B^1\times B^2}:\theta\in\mathcal{M}(T)\}\right|\geq 2^{\left|A\cap B^1\times B^2\right|}.$$

Take any finite $A\subseteq X$. W.l.o.g., assume that $A\subseteq A^1\times A^2$, where A^j are finite and even for any $j=1,2$. There exist sets $B^1\cup B^2=B$ such that $B^1\cap B^2=\varnothing$, $|B^j\cap A^j|=\frac{1}{2}|A^j|$ and

$$\left|A\cap B^1\times B^2\right|\geq\frac{|B^1\times B^2\cap A^1\times A^2|}{|A^1\times A^2|}|A|=\frac{1}{4}|A|.$$

(This can be proved directly, or via the first part of Lemma 8 from Appendix B.) Hence,

$$\begin{aligned}\mathcal{E}(A;T) &= \frac{1}{|A|}\log \mathcal{M}(A;T)\\ &\geq \frac{1}{|A|}\log |\{\theta|_A : \theta \in \mathcal{M}(T)\}|\\ &\geq \frac{1}{|A|}\log |\{\theta|_{A\cap B^1\times B^2} : \theta \in \mathcal{M}(T)\}|\\ &\geq \frac{1}{|A|}\log 2^{|A\cap B^1\times B^2|} \geq \frac{1}{4}.\end{aligned}$$

By the first part of Lemma 1, $\{(S,\tau')\}$ is not sufficient for deduction. □

This points to an important difference between languages *Product*$_d$ and *Ordered Graph*$_d$: in the former any single pattern leads to a theory that is sufficient for deduction, in the latter, this is not the case.

5.3. **Graph.** Here, I show the tightness of language *Graph*$_d$ from Example 3. This fact turns out to be a relatively straightforward corollary to Proposition 5. To explain it, I need a simple result. Let $G^* \longmapsto X^*$, $G \longmapsto X$ be group actions and $p_G : G^* \to G, p_X : X^* \to X$ be "onto" mappings such that for any $g \in G^*$ and any $x \in X$

$$p_X(g\cdot x) = p_G(g)\cdot p_X(x). \tag{5.1}$$

Mappings p_G and p_X are projections of group action $G^* \longmapsto X^*$ onto group action $G \longmapsto X$ that preserve group properties.

Lemma 6. *Suppose that $G \longmapsto X$ is transitive. If $U \subseteq X$ is local under $G^* \longmapsto X^*$, then $p_X(U)$ is local under $G \longmapsto X$. If S is generic under $G^* \longmapsto X^*$, then $p_X(S)$ is generic under $G \longmapsto X$. If $G^* \longmapsto X^*$ is tight, then $G \longmapsto X$ is tight and $t(G) \geq t(G^*)$.*

For any mutually disjoint finite sets $B^1, ...B^d \subseteq B$, define set of all d-element subsets of B, each with exactly one element in $B^l, l \leq d$:

$$X\left(B^1,...,B^d\right) = \left\{\{b_1,...,b_d\} : b_l \in B^l\right\}.$$

Proposition 6. *Consider language* Graph$_d$ *from Example 3. For any mutually disjoint finite sets $B^1, ...B^d \subseteq B$, set $X\left(B^1,...,B^d\right) \subseteq X$ is generic. The language is tight, and*

$$t(G) \geq d^{-d}.$$

Proof. Let $G^* \longmapsto X^*$ be as in language *Ordered Graph*$_d$. Then $G = G^*$ and let $p_G = \mathrm{id}$, $p_X(b_1,...,b_d) := \{b_1,...,b_d\}$ for any $(b_1,...,b_d) \in X^*$. One checks that (5.1) is satisfied. The Proposition follows from Lemma 6 and Proposition 5. □

I show that the acid-alkaline theory is sufficient for deduction. Both patterns (S,τ) and (S,τ') *from* equations (2.11) and (2.12), respectively, imply a pattern (S^*,τ^*) :

$$S^* = \{\{a_1,b_1\},\{a_2,b_1\},\{a_1,b_2\},\{a_2,b_2\}\} \text{ and}$$
$$\tau^*\{a_1,b_1\} = 1, \tau^*\{a_1,b_1\} = 1, \tau^*\{a_1,b_1\} = 1, \tau^*\{a_1,b_1\} = 0.$$

Indeed, this pattern corresponds to a statement: "If a_1 and b_1 react, a_2 and b_1 react, a_2 and b_2 react, then a_1 and b_2 react," which is a consequence of the acid-alkaline theory. Because $S^* = X(\{a_1,a_2\},\{b_1,b_2\})$ is generic, this pattern leads to a theory that is sufficient for deduction.

6. Relation to statistical learning theory

This Section explains the relationship between Proposition 4 and the statistical learning theory. For this purpose, I state a somehow stronger version of Proposition 4. Suppose that $\mathcal{M} \subseteq \{0,1\}^X$ is a family of models $\theta : X \to \{0,1\}$. For any subset $S \subseteq X$, let $\mathcal{M}_S = \{\theta|_S : \theta \in \mathcal{M}\}$ denote the set of model restrictions to set S. Say that G-dimension of family $\mathcal{M}$ includes S if for any permutation $g \in G$,

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

In particular, G-dimension of family $\mathcal{M}(\{S,\tau\})$ includes S. In Appendix E, I show the following result.

Proposition 7. *Suppose that S is generic. Then, for any $\varepsilon > 0$, there is a local $V \subseteq X$ such that for any permutation g, any local U such that $g \cdot V \subseteq U$, for any family $\mathcal{M}$ which G-dimension includes S,*

$$\frac{1}{|V|}\log|\mathcal{M}_V| \leq \varepsilon. \tag{6.1}$$

Here, phrase "there exists local V such that for any permutation g, any local U so that $g \cdot V \subseteq U$, ... " should be read as "for sufficiently large local U, ..." The Proposition bounds the size of model restrictions to sufficiently large local sets of family of models which G-dimension includes a generic set. Proposition 4 is a corollary to Proposition 7.

The definition of G-dimension is related to the famous VC-dimension of a family of models. Define the VC-dimension of family $\mathcal{M}$ as

$$\dim_{VC}\mathcal{M} = \inf\left\{k : |\mathcal{M}_S| < 2^k \text{ for any } S \subseteq X, |S| \leq k\right\}$$

if this set is not empty and ∞ if for each k, there exists $S \subseteq X, |S| \leq k$ such that $|\mathcal{M}_S| = 2^k$.

Proposition 8 (Sauer-Shelah's Lemma, (Sauer 1972), (Shelah 1972), (Vapnik and Chervonenkis 1971))**.** *For any finite* $V \subseteq X$,

$$|\mathcal{M}_V| \leq |V|^{\dim_{VC} \mathcal{M}}.$$

The Proposition provides foundation for the Vapnik-Chervonenkis statistical learning theory (see (Vapnik 1998)). It is used to prove various distribution-free asymptotic results. Quite often, a weaker thesis is needed: If the VC dimension of family $\mathcal{M}$ is finite, then, for any $\varepsilon > 0$, there is n, such that for any finite $V \subseteq X$, $|V| \geq n$, inequality (6.1) holds.

This weaker form of Sauer-Shelah's Lemma is a special case of Proposition 7 for $G = \Pi_X$, i.e., when the group of permutations consists of all permutations on X. To see this recall that any finite subset of X is generic and local under group Π_X. Then, $\dim_{VC} \mathcal{M} < \infty$ is equivalent to the existence of a generic $S \subseteq X$ such that Π_X-dimension of $\mathcal{M}$ includes S. Let U be a local set from the thesis of Proposition 7. Take $n = |U|$ and recall that for any finite $U, V \subseteq X$, there exists a permutation $g \in \Pi_X$ such that $g \cdot U \subseteq V$ if and only if $|U| \leq |V|$.

7. Comments and conclusion

The goal of this paper is to examine the relationship between knowledge and language. In the binary case $Y = \{0, 1\}$, a theory is sufficient for deduction if it implies a pattern with generic support. If the language is tight, then the relation goes also the other way: Theory is sufficient for deduction only if it implies a pattern with generic support. In the general case of finite Y, any theory that is sufficient for deduction implies a theory that is also sufficient for deduction and consists with finitely many patterns with the same generic support.

This is a good place to discuss some limitations of this paper. I consider here only deterministic patterns. This is a natural assumption if one is interested in deduction rather than probabilistic reasoning. On the other hand, the decision maker should be able to formulate universal statements about probabilistic properties of the world. For example, Netflix may understand that "For any customers c_1, c_2, any movie m, if customer c_1 likes movie m, then customer c_2 likes movie m with probability 0.7." The analysis of the probabilistic patterns leads different conceptual issues: How correctly define consequences of a probabilistic pattern? For example, are realizations of outcomes drawn independently for each pattern? What is the worst-case scenario? I leave these questions for future research. However, it is conjecture that the tightness of a language plays an important role in characterizing the informativeness of probabilistic patterns.

If the conjecture is true, then this paper should be understood as the first step towards a more general probabilistic approach.

I assume that the decision maker computes payoffs with respect to the long-run criterion. This is certainly a simplification; in the real world, people discount. One of the consequences of the current approach is that sufficiency for deduction is a zero–one notion and it divides theories into two groups: those that predict well with few mistakes and those that do not. A discounted model could allow for more refined divisions into theories that allow for correct predictions early and those that lead initially to many mistakes.

Some technical, but interesting, questions are left open. In the end of Section 2, I presented an example of a language that is transitive but not locally generated. In the rest of the paper, I assumed that the language is transitive, locally generated and tight. At this moment, it remains unclear whether there are any languages that are locally generated, transitive, and *not* tight. Second, the bound on the entropy obtained in Proposition 7 is very crude. In some special cases, much tighter bounds can be found (for example, Sauer-Shelah's Lemma; see also footnote ???). The question is how to tighten the bound in Proposition 7.

References

Aldous, D. (1981): "Representations for Partially Exchangeable Arrays of Random Variables," *Journal of Multivariate Analysis*, 11, 581–598.

Alon, N., E. Fischer, and I. Newman (2005): "Testing of Bipartite Graph Properties," http://www.math.tau.ac.il/ nogaa/PDFS/afn2.pdf.

Balogh, J., B. Bollobas, and D. Weinreich (2000): "The Speed of Hereditary Properties of Graphs," *Journal of Combinatorial Theory Ser. B*, 79, 131–156.

Billot, A., I. Gilboa, D. Samet, and D. Schmeidler (2005): "Probabilities as similarity-weighted frequencies," *Econometrica*, 73(4), 1125–1136.

Blackwell, D., and A. Girshick (1954): *Theory of Games and Statistical Decisions*. Wiley, New York.

Blume, A. (2004): "A Learning-Efficiency Explanation of Structure in Language," *Theory and Decision*, 57, 265 – 285.

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.

Chen, Z., and L. Epstein (2002): "Ambiguity, Risk and Asset Returns in Continuous Time," *Econometrica*, pp. 1403–1443.

Crawford, V. P., and H. Haller (1990): "Learning How to Cooperate: Optimal Play in Repeated Coordination Games," *Econometrica*, 58, 571–595.

Epstein, L. (2006): "An Axiomatic Model of Non-Bayesian Updating," *Review of Economic Studies*, 73, 413–436.

Epstein, L., J. Noor, and A. Sandroni (2006): "Non-Bayesian Updating: A Theoretical Framework," University of Rochester, mimeo.

Epstein, L., and M. Schneider (2003): "Recursive Multiple-Priors," *Journal of Economic Theory*, 113, 1–31.

Gilboa, I., and D. Schmeidler (1989): "Maxmin Expected Utility with Non-Unique Priors," *Journal of Mathematical Economics*, 18, 141–153.

Gilboa, I., and D. Schmeidler (1995): "Case-Based Decision Theory," *Quarterly Journal of Economics*, 110(3), 605–639.

——— (1996): "Case-based optimization," *Games and Economic Behavior*, 15(1), 1–26.

——— (2000): "Case-based knowledge and induction," *Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans*, 30(2), 85–95.

Gilboa, I., and D. Schmeidler (2001): *A Theory of Case-Based Decisions*. University Press, Cambridge, UK.

Gilboa, I., and D. Schmeidler (2003): "Inductive inference: An axiomatic approach," *Econometrica*, 71(1), 1–26.

Hansen, L. P., and T. Sargent (forthcoming): *Robustness*. Princeton University Press, Princeton.

J. Balogh, B. B., and D. Weinreich (2001): "The Penultimate Rate of Growth for Graph Properties," *European J. of Combinatorics*, 22(3), 277–289.

Kalai, G. (2003): "Learnability and Rationality of Choice," *Journal of Economic Theory*, 113, 104–117.

Kallenberg, O. (2005): *Probabilistic Symmetries and Invariance Principles*, Probability and Its Applications. Springer, New York.

Klein, F. (1979): *Development of Mathematics in the 19th Century*. Math Sci Press, Brookline, Mass.

Lang, S. (2002): *Algebra*, Graduate Texts in Mathematics. Springer, New York.

Lipman, B. (2003): "Language and Economics," in *Cognitive Processes and Rationality in Economics*, ed. by N. Dimitri, M. Basili, and I. Gilboa. Routledge, London.

Monk, J. (2005): *Mathematical Logic (Graduate Texts in Mathematics)*. Springer, New York.

Poizat, B. (2000): *A Course in Model Theory: An Introduction to Contemporary Mathematical Logic*, Universitext. Springer, New York.

Promel, H., and A. Steger (1991): "Excluding Induced Subgraphs: Quadrilaterals," *Random Strutures and Algorithms*, 2, 55–71.

——— (1992): "Excluding Induced Subgraphs III: A General Asymptotic," *Random Strutures and Algorithms*, 3, 19–31.

——— (1993): "Excluding Induced Subgraphs II: Extremal Graphs," *Discrete Applied Mathematics*, 44, 283–294.

Rubinstein, A. (1996): "Why Are Certain Properties of Binary Relations Relatively More Common in Natural Language?," *Econometrica*, 64, 343–355.

——— (1998): *Modeling Bounded Rationality*. MIT Press, Cambridge.

——— (2000): *Economics and Language*. Cambridge University Press, Cambridge, UK.

SALANT, Y. (2007): "On the Learnability of Majority Rule," *Journal of Economic Theory*, p. forthcoming.

SAUER, N. (1972): "On the Density of Families of Sets," *Journal of Combinatorial Theory (A)*, 13, 145–147.

SCHEINERMAN, E. R., AND J. ZITO (1994): "On the Size of Hereditary Classes of Graphs," *Journal of Combinatorial Theory Ser. B*, 61, 16–39.

SCHWARZ, G. (1994): *Game Theory and Statistics*vol. 2, chap. 21, pp. 769–779. Elsevier.

SHELAH, S. (1972): "A Combinatorial Problem: Stability and Order for Models and Theories in Infinitary Languages," *Pacific Journal of Mathematics*, 41, 247–261.

VAPNIK, V. N. (1998): *Statistical Learning Theory*. Wiley-Interscience.

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.

WALD, A. (1949): "Statistical Decision Functions," *Annals of Mathematical Statistics*, 20, 165–205.

——— (1950): *Statistical Decision Functions*. John Wiley and Sons, London.

APPENDIX A. RELATIONAL AND GROUP LANGUAGES

A.1. **Proof of Proposition 1.** Let G be the set of all permutations $g : X \to X$, such that for any i, any $\bar{x} \in X^{k_i}$, (2.9) holds. G contains identity (hence, it is not empty) and it is a group of permutations.

It is clear that for any tuples $\bar{x}, \bar{x}' \in X^k$, if $\bar{x}$ and $\bar{x}'$ are exchangeable, then $g \cdot \bar{x}$ and $g \cdot \bar{x}'$ are exchangeable. I show that for any two exchangeable tuples $\bar{x}, \bar{x}' \in X^k$, there is a permutation $g \in G$ such that $\bar{x}' = g \cdot \bar{x}$. Indeed, both tuples $\bar{x} = (x_1, ..., x_k)$ and $\bar{x}' = (x'_1, ..., x'_k)$ can be extended to sequences $(x_1, x_2, ...)$ and $(x'_1, x'_2, ...)$ such that (a) for each $d' \geq k$, tuples $(x_1, ..., x_{k'})$ and $(x'_1, ..., x'_{k'})$ are exchangeable and (b) $X = \{x_1, x_2, ...\} = \{x'_1, x'_2, ...\}$. (This is a simple exercise in the back-and-forth method; see, for example, (Poizat 2000).) Take now $g : X \to X$ such that $g(x_l) = x_l$ for any l. This is well-defined, because for any l, l', $x_l = x_{l'} \iff x'_l = x'_{l'}$.

A.2. **Proof of Proposition 2.** Let $\mathcal{L}$ be generated by $G \longmapsto X$. Take any group pattern (S, τ). Let $k = |S|$ and let $\bar{x}^* \in X^k$ be an enumeration of S, i.e. tuple such that $\{x^*_1, ..., x^*_k\} = S$. For any i, any $l_1, ..., l_{k_i} \in \{1, ..., k\}$, let $\rho^*_{i,(l_1,...,l_{k_i})} \in \{0, 1\}$ be defined as

$$\rho^*_{i,(l_1,...,l_{k_i})} = 0 \text{ iff } R_i\left(x^*_{l_1}, ..., x^*_{l_{k_i}}\right).$$

In the statement below, $(-)^1$ denotes a negation and $(-)^0$ makes the "minus" sign disappear. Model θ omits group pattern (S, τ) if and only if, for any tuple of instances $\bar{x}$ that is exchangeable with $\bar{x}^*$, there exists $l \leq k$ such that $\theta(x_l) \neq \tau(x^*_l)$. This further

holds if and only if θ satisfies a relational pattern:

$$\forall_{x_1,...,x_k}$$
$$\left(\bigwedge_i \bigwedge_{l_1,...,l_{k_i}\in\{1,...,k\}} (-)^{\rho^*_{i,(l_1,...,l_{k_i})}} R_i\left(x_{l_1},...,x_{l_{k_i}}\right)\right) \wedge \bigwedge_{l\leq k-1} \theta(x_l) = \tau(x^*_l)$$
$$\Longrightarrow \theta(x_k) \neq \tau(x^*_k).$$

This shows the first part of the Proposition.

Let $F(x_1,...,x_k,\theta)$ be a formula that is a disjunction of conjunctions of atomic formulas (2.7).

Lemma 7. *Suppose that F is satisfied for some tuple $x_1,...,x_k$ and model θ. Then, it is also satisfied for any tuple $x'_1,...,x'_k$ such that $\bar{x}$ is exchangeable with $\bar{x}'$ and any model θ' such that $\theta'(x'_l) = \theta(x_l)$ for any $l \leq k$.*

For any tuple $\bar{x} \in X^k$, define $\Theta(\bar{x}) \subseteq Y^{\{1,...,k\}}$

$$\Theta(\bar{x}) = \left\{\rho : \rho(l) = \theta(x_l) \text{ for any } l \leq k \text{ and some } \theta \in Y^X, \text{ st. } F(x_1,...,x_k,\theta) \text{ holds}\right\}.$$

By the Lemma, if $\bar{x}$ and $\bar{x}'$ are exchangeable, then $\Theta(\bar{x}) = \Theta(\bar{x}')$.

There is finitely many equivalence classes of exchangeability on X^k. Let M be the number of equivalence classes. Choose a representative $\bar{x}^{(m)}$ for any equivalence class $m = 1,...,M$. For any m, any $\rho \in Y^{\{1,...,k\}}$, construct group patterns (S_m, τ_ρ) :

$$S_m = \left\{x^{(m)}_1,...,x^{(m)}_k\right\},$$
$$\tau_\rho\left(x^{(m)}_l\right) = \rho(l) \text{ for any } l \leq k.$$

Then, model θ satisfies relational pattern $\bigvee_{x_1,...,x_k} F(x_1,...,x_k,\theta)$ if and only if it omits all group patterns (S_m,τ_ρ), where $m = 1,...,M$ and $\rho \notin \Theta\left(\bar{x}^{(m)}\right)$.

Appendix B. Transitive group action

This Appendix contains various results about finite transitive group actions. There are two parts. The first part is concerned with mixing and the second part describes entropy properties of finite transitive group actions.

B.1. **Mixing.** Assume that A is finite set and $G \longmapsto A$ is a transitive group action. For any function $f : A \to R$, define

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

Lemma 8. *Suppose that $G \longmapsto A$ is transitive. For any two subsets $B, C \subseteq A$, there is $g \in G$ such that*

$$\frac{|B \cap g \cdot C|}{|B|} \leq \frac{|C|}{|A|}. \tag{B.2}$$

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

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

Proof. Let $G_x = \{g : g \cdot x = x\}$ be a subgroup of G, which is called a *stabilizer* of $x \in A$. Since $G \longmapsto A$ is transitive, its size $|G_x|$ does not depend on the choice of $x \in A$ and $\frac{|G|}{|G_x|} = |A|$. For any set $S \subseteq A$, define $m_S : G \to [0,1]$ as

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

We compute

$$m_S = \frac{1}{|G|}\sum_{g \in G} m_S(g) = \frac{1}{|A|}\frac{1}{|G|}\sum_{a \in A}\sum_{b \in S}\left[\sum_{g \in G} \mathbf{1}\{g \cdot a = b\}\right] f(a).$$

Since G is transitive, for any two $a, b \in A$, $\sum_{g \in G} \mathbf{1}\{g \cdot x = y\} = |G_x|$. Hence,

$$m_S = \frac{1}{|A|}\sum_{a \in A} f(a) \sum_{b \in S}\frac{|G_x|}{|G|} = \frac{|S|}{|A|} Ef.$$

For the first part, assume that $f = \mathbf{1}_C$ is an indicator function of set C. The above says that

$$\frac{|B|}{|A|}\frac{|C|}{|A|} = m_B = \frac{1}{|G|}\sum_{g \in G} E\mathbf{1}_B f(g \cdot .) = \frac{1}{|G|}\sum_{g \in G}\frac{|B \cap g \cdot C|}{|A|}.$$

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

For the second part, 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 $G_C \subseteq G_B$. However, this leads to a contradiction, as the following calculations show:

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

□

Suppose that $G \longmapsto B$ is another group action of the same group G. Let $p : A \to B$ be an "onto" mapping. Say that p *preserves group action if* for each $g \in G$ and $a \in A$,

$$p(g \cdot a) = g \cdot p(a).$$

Lemma 9. *Suppose that $p : A \to B$ preserves group action. Then, for any set $C \subseteq B$,*

$$\frac{|C|}{|B|} = \frac{|p^{-1}(C)|}{|A|}.$$

Proof. This is a consequence of the fact that for any $b, b' \in B$, $|p^{-1}(b)| = |p^{-1}(b')|$. □

B.2. **Entropy under transitive actions.** For any subset $U \subseteq A$, let Y^U be a set of functions from U to Y. For any $\tau_U \in Y^U$, let

$$[\tau_U] = \left\{\tau \in Y^A : \tau|_U = \tau_U\right\}$$

be a set of functions $\tau \in Y^A$, whose restrictions to set U are equal to τ_U. Let $\mu \in \Delta F(A)$ be a probability distribution over functions. For any subset $U \subseteq A$, let

$$h_\mu(U) := -\sum_{\tau_U \in Y^U} \mu([\tau_U]) \log \mu([\tau_U]).$$

be the entropy of restrictions of functions $\tau \in Y^A$ to set U. I omit the reference to measure μ whenever it does not lead to ambiguity.

Lemma 10. *For any $U \subseteq A$, $h(U) \leq \left|\left\{\tau_U \in Y^U : \mu([\tau_U]) \neq 0\right\}\right|$. For any subsets $U, V \subseteq A$,*

$$h(U \cup V) - h(U) \leq h(V) - h(U \cap V).$$

Proof. The first part is standard property of the entropy. The second part is equally standard and I present its proof for the sake of completeness. Observe that

$$h(U \cup V) - h(U) = -\sum_{\tau \in Y^A} \mu(\tau) \log \frac{\mu([\tau|_{U\cup V}])}{\mu([\tau_U])} \text{ and}$$

$$h(V) - h(U \cap V) = -\sum_{\tau \in Y^A} \mu(\tau) \log \frac{\mu([\tau|_V])}{\mu([\tau|_{U\cap V}])}.$$

By Jensen's inequality,

$$\begin{aligned} &h(U \cup V) - h(U) - [h(V) - h(U \cap V)] \\ &= \sum_{\tau \in Y^A} \mu(\tau) \log \frac{\mu([\tau|_V])\,\mu([\tau_U])}{\mu([\tau|_{U\cap V}])\,\mu([\tau|_{U\cup V}])} \leq \log\left(\sum_{\tau \in Y^A} \mu(\tau) \frac{\mu([\tau|_V])\,\mu([\tau_U])}{\mu([\tau|_{U\cap V}])\,\mu([\tau|_{U\cup V}])}\right). \end{aligned}$$

Observe that

$$\sum_{\tau\in Y^A}\mu(\tau)\frac{\mu([\tau|_V])\,\mu([\tau_U])}{\mu([\tau|_{U\cap V}])\,\mu([\tau|_{U\cup V}])}$$
$$=\sum_{\tau_{U\cap V}\in Y^{U\cap V}}\mu([\tau_{U\cap V}])\sum_{\tau_{U\cup V}\in Y^{U\cup V},\text{ st. }\tau_{U\cup V}|_{U\cap V}=\tau_{U\cap V}}\frac{\mu([\tau_{U\cup V}])}{\mu([\tau_{U\cap V}])}\frac{\mu([\tau_{U\cup V}|_V])\,\mu([\tau_{U\cup V}|_U])}{\mu([\tau_{U\cup V}])\,\mu([\tau_{U\cap V}])}.$$

Since

$$\sum_{\tau_{U\cup V}\in Y^{U\cup V},\text{ st. }\tau_{U\cup V}|_{U\cap V}=\tau_{U\cap V}}\frac{\mu([\tau_{U\cup V}])}{\mu([\tau_{U\cap V}])}\frac{\mu([\tau_{U\cup V}|_V])\,\mu([\tau_{U\cup V}|_U])}{\mu([\tau_{U\cup V}])\,\mu([\tau_{U\cap V}])}$$
$$=\sum_{\tau^V\in Y^V,\tau^U\in Y^U,\text{ st. }\tau^V|_{U\cap V}=\tau^U|_{U\cap V}=\tau_{U\cap V}}\frac{\mu\left(\left[\tau^V\right]\right)}{\mu([\tau_{U\cap V}])}\frac{\mu\left(\left[\tau^U\right]\right)}{\mu([\tau_{U\cap V}])}$$
$$=\left(\sum_{\tau^V\in Y^V\text{ st.}\tau^V|_{U\cap V}=\tau_{U\cap V}}\frac{\mu\left(\left[\tau^V\right]\right)}{\mu([\tau_{U\cap V}])}\right)\left(\sum_{\tau^U\in Y^U\text{ st.}\tau^U|_{U\cap V}=\tau_{U\cap V}}\frac{\mu\left(\left[\tau^U\right]\right)}{\mu([\tau_{U\cap V}])}\right)$$
$$=1,$$

it must be that

$$\sum_{\tau\in Y^A}\mu(\tau)\frac{\mu([\tau|_V]\,|\,[\tau|_{U\cap V}])}{\mu([\tau|_{U\cup V}]\,|\,[\tau|_U])}=\sum_{\tau_{U\cap V}\in Y^{U\cap V}}\mu([\tau_{U\cap V}])=1.$$

This finishes the proof of the Lemma. □

Let $\mathcal{M}\subseteq Y^A$ be a set of functions from A to set Y. For any subset $U\subseteq A$, let

$$\mathcal{M}(U)=\left\{\tau_U\in Y^U:\tau_U=\tau|_U\text{ for some }\tau\in\mathcal{M}\right\}.$$

Say that set $\mathcal{M}$ is *invariant with respect to group action* if $\tau(g\cdot.)\in\mathcal{M}$ for each $\tau\in\mathcal{M}$ and $g\in G$.

Lemma 11. *Suppose that $\mathcal{M}$ is invariant with respect to group action. For any finite set $U\subseteq X_n$,*

$$\frac{\log|\mathcal{M}(U)|}{|U|}\geq\frac{\log|\mathcal{M}(A)|}{|A|}.$$

Proof. Let μ be a uniform distribution on $\mathcal{M}$. Then, $h(A)=\log|A|$ and $h(U)\leq\log|\mathcal{M}(U)|$ for each $U\subseteq A$ by the first part of Lemma 10. For any subset $U\subseteq A$ and any permutation g, by the invariance of set $\mathcal{M}$ it must be that $h(U)=h(g\cdot U)$.

We may find a subset $V\subseteq A$ such that (a) for each $U\subseteq A$,

$$\frac{h(V)}{|V|}\leq\frac{h(U)}{|U|}$$

and (b) for each $V \subseteq U$, $U \neq V$,

$$\frac{h(V)}{|V|} < \frac{h(U)}{|U|}.$$

Since A is finite, such a set clearly exists. If $V = A$, then, by property (a), for each set $U \subseteq A$

$$\frac{\log|\mathcal{M}(A)|}{|A|} = \frac{h(A)}{|A|} \leq \frac{h(U)}{|U|} \leq \frac{\log|\mathcal{M}(U)|}{|U|}$$

and the Lemma holds. Suppose otherwise that $V \neq A$. Since G acts transitively, there is $g \in G$, so that $g \cdot V \neq V$. By the second part of Lemma 10,

$$h(V \cup g \cdot V) \leq h(V) + h(g \cdot V) - h(V \cap g \cdot V).$$

Hence,

$$\begin{aligned}\frac{h(V \cup g \cdot V)}{|V \cup g \cdot V|} &\leq \frac{h(V) + h(g \cdot V) - h(V \cap g \cdot V)}{|V| + |g \cdot V| - |V \cap g \cdot V|} \\ &= \frac{h(V)}{|V|} + \frac{\left(\frac{h(V)}{|V|} - \frac{h(V \cap g \cdot V)}{|V \cap g \cdot V|}\right) \cdot |V \cap g \cdot V|}{2|V| - |V \cap g \cdot V|}.\end{aligned}$$

By property (a) of set V

$$\frac{h(V)}{|V|} \leq \frac{h(V \cap g \cdot V)}{|V \cap g \cdot V|}.$$

However, this implies that

$$\frac{h(V \cup g \cdot V)}{|V \cup g \cdot V|} \leq \frac{h(V)}{|V|},$$

which leads to a contradiction with property (b). □

Proof of Lemma 3. It is a corollary to Lemma 11. □

Appendix C. Proof of Lemma 1

C.1. **Function h_c.** Consider function $f_c(h) = h\left(c + 2 + \log\frac{1}{h}\right)$. Then, $f_c(h) = 0$, f_c is strictly increasing for any $h \leq e^{c+1}$, and $f_c(e^c) \geq 1$. Thus, there is an increasing inverse of f_c with domain $e \in [0, 1]$.

C.2. **Combinatorial result.** Here, I prove a combinatorial result that is useful in the argument for Lemma 1. Define function $f : R \times N \to R$ as follows:

$$\begin{aligned}f(p, 1) &:= |Y| \max(p + 1, 0) \text{ and} \\ f(p, m + 1) &:= (|Y| - 1) f(p - 1, m) + f(p, m), \text{ for } m \geq 1.\end{aligned}$$

Lemma 12. *Function $f(p,m)$ is continuous, convex, and increasing in p and for any $p>0$, any $r \leq m, r \in N$,*

$$f(r,m) \leq (r+1)|Y|^r \begin{pmatrix} m \\ r \end{pmatrix}. \tag{C.1}$$

Proof. $f(p,1)$ is continuous, convex, and increasing in p by definition. When $m>1$, the continuity, convexity, and monotonicity of $f(p,m)$ follows by induction.

First, I show that for $0 \leq r \leq m$,

$$f(r,m) \leq |Y|^m (r+1).$$

Indeed, this is true for $m=1$ by the definition. Suppose that the above is true for m. Take $r \leq m+1$. Then,

$$f(r,m+1) \leq (|Y|-1) r |Y|^m + (r+1)|Y|^m \leq (r+1)|Y|^{m+1}.$$

In particular, this implies that

$$f(m,m) \leq (m+1)|Y|^m.$$

Notice also that, by the definition, for any m

$$f(0,m) = 1. \tag{C.2}$$

By induction on r and k, I show that for any $r \geq 1, k \geq 0$,

$$f(r,r+k) \leq (r+1)|Y|^r \begin{pmatrix} r+k \\ r \end{pmatrix}. \tag{C.3}$$

By the above, the statement is true for $k=0$ and any r. Suppose that the statement is true for k and any r. Then,

$$\begin{aligned} f(1,1+k) &= (|Y|-1) f(0,k) + f(1,k) \\ &\leq |Y| - 1 + 2|Y| k \leq 2|Y|(2+k) = 2|Y| \begin{pmatrix} k+2 \\ 1 \end{pmatrix}, \end{aligned}$$

where the inequality follows from the inductive step and (C.2). Suppose now that (C.3) is true for $k+1$ and $r \geq 1$. Then

$$\begin{aligned}
&f(r+1, r+1+k+1) \\
&\leq (|Y|-1) f(r, r+k+1) + f(r+1, r+k+1) \\
&\leq (|Y|-1)(r+1)|Y|^r \begin{pmatrix} r+k+1 \\ r \end{pmatrix} + (r+2)|Y|^{r+1} \begin{pmatrix} r+k+1 \\ r+1 \end{pmatrix} \\
&\leq (r+2)|Y|^{r+1} \left[\begin{pmatrix} r+k+1 \\ r \end{pmatrix} + \begin{pmatrix} r+k+1 \\ r+1 \end{pmatrix} \right] \\
&= (r+2)|Y|^{r+1} \begin{pmatrix} r+k+2 \\ r+1 \end{pmatrix},
\end{aligned}$$

where the first inequality comes from the inductive step. This shows the inductive step for $k+1$ and all r. This also finishes the proof of the Proposition. □

C.3. **Proof of Lemma 1.** Fix theory T, an instance process $\bar{x}_\infty$, a learning rule l, and period t. Define $A_t = \{x_1, ..., x_t\}$ and

$$\mathcal{M}(y_1, ..., y_{t-m}) := \{\tau \in \mathcal{M}(A_t; T) : \tau(x_1) = y_1, ..., \tau(x_{t-m}) = y_{t-m}\}, \tag{C.4}$$

$$u_m(y_1, ..., y_{t-m}) = \inf_{\tau \in \mathcal{M}(y_1,...,y_{t-m})} \sum_{s=t-m+1}^{t} l\left((x_u, \tau(x_u))_{u<s}, x_s\right)(\tau(x_s)) \leq m,$$

$$M_m(u) = \max_{y_1,...,y_{t-m}: u_t(y_1,...,y_{t-m}) \geq u} |\mathcal{M}(y_1, ..., y_{t-m})|,$$

with a convention that $M_m(u) = 0$ for any $u > m$. Here,

- $\mathcal{M}(y_1, ..., y_{t-m})$ is the set of model restrictions $\tau \in \mathcal{M}(A_t; T)$ for which the first $n-m$ outcomes are equal to $y_1, ..., y_{t-m}$. In other words, this is the set of model restrictions that are considered as plausible by the decision maker who observed that $\theta(x_1) = y_1, ..., \theta(y_{t-m}) = y_{t-m}$. Of course,
$$|\mathcal{M}(y_1, ..., y_{t-m})| = \sum_y |\mathcal{M}(y_1, ..., y_{t-m}, y)|;$$
- $u_m(y_1, ..., y_{n-m})$ is the worst-case payoff from correct predictions in periods $t = n-m+1, ..., n$, that are committed by learning rule l given that the model restriction belongs to set $\mathcal{M}(y_1, ..., y_{n-m})$. Then, u_m is characterized by a recursive formula:

$$u_m(y_1, ..., y_{t-m}) = \inf_y \left[l\left((x_u, y_u)_{u \leq t-m}, x_{t-m+1}\right)(y) + u_{m-1}(y_1, ..., y_{t-m}, y) \right]. \tag{C.5}$$

- $M_m(u)$ is the upper bound on the size of $\mathcal{M}(y_1, ..., y_{t-m})$ for these sequences of outcomes $y_1, ..., y_{t-m}$ that lead to continuation payoff at least u.

Lemma 13. *$M_m(u) \leq f(m-u, m)$ for any u and any $m = 1, ..., t$.*

Proof. $M_m(u)$ is (weakly) decreasing in u. Notice that for any $y_1, ..., y_{t-m}, y$,

$$\begin{aligned}|\mathcal{M}(y_1, ..., y_{t-m}, y)| &\leq M_{m-1}(u_{m-1}(y_1, ..., y_{t-m}, y)) \\ &= M_{m-1}(u_{m-1}(y_1, ..., y_{t-m}, y)) \\ &\leq M_{m-1}\left(u_m(y_1, ..., y_{t-m}, y) - l\left((x_u, y_u)_{u \leq t-m}, x_{t-m+1}\right)(y)\right),\end{aligned}$$

where the last inequality is a consequence of (C.5). This leads to a recursive bound,

$$M_m(u) \leq \max_{\delta \in \Delta Y} \sum_{y \in Y} M_{m-1}(u - \delta(y)). \tag{C.6}$$

The rest of the proof is by induction on m. The inductive hypothesis can be directly verified for $m = 1$. Indeed, by definition,

$$\begin{aligned}&M_1(u) \leq |Y| \leq |Y| \max(2-u, 0) \text{ for any } u \leq 1 \text{ and} \\ &M_1(u) = 0 \leq |Y| \max(2-u, 0) \text{ for any } u > 1.\end{aligned}$$

Suppose that the inductive step holds for m. Then, due to (C.6),

$$M_{m+1}(u) \leq \max_{\delta \in \Delta Y} \sum_{y \in Y} f(m - u + \delta(y), m).$$

Since function $f(x, m)$ is convex in x, the maximum above is obtained when $\delta(y) = 1$ for some y and $\delta(y') = 0$ for $y' \neq y$. Hence, by definition of function f,

$$M_{m+1}(u) \leq (|Y| - 1) \sum_{y \in Y} f(m-u, m) + f(m-u+1, m) = f(m+1-u, m+1),$$

which finishes the inductive step and the proof of the Lemma. □

Let

$$u^* := u_t(\varnothing) = \inf_{\theta \in \mathcal{M}(T)} tU_t(l, \bar{x}_\infty, \theta).$$

Then, $\mathcal{M}(A_t;T) = M_t(u^*)$. By Lemmas 12 and 13,

$$\begin{aligned}
&\mathcal{E}(A_t;T)\\
&= \frac{1}{t}\log \mathcal{M}(A_t;T) = \frac{1}{t}\log M_t(u^*)\\
&\leq \frac{1}{t}\log\left[(\lceil t-u^*\rceil+1)|Y|^{\lceil t-u^*\rceil}\begin{pmatrix} t \\ \lceil t-u^*\rceil\end{pmatrix}\right]\\
&\leq \frac{\lceil t-u^*\rceil}{t}\left(\log|Y| + 2 + \log\frac{t}{\lceil t-u^*\rceil}\right),
\end{aligned}$$

and the last inequality follows from Stirling's formula

$$\frac{1}{n}\log\begin{pmatrix} n \\ \lceil \varepsilon n\rceil\end{pmatrix} \approx \varepsilon\log\frac{1}{\varepsilon} + (1-\varepsilon)\log\frac{1}{1-\varepsilon}, \tag{C.7}$$

and the fact that that $(1-\varepsilon)\log\frac{1}{1-\varepsilon} \leq \varepsilon$ for any $\varepsilon \in (0,1)$. Because h_c is increasing in e, this shows that

$$\frac{\lceil t-u^*\rceil}{t} \leq h_{\log|Y|}(\mathcal{E}(A_t;T)) \leq h_{\log|Y|}(\mathcal{E}(T)).$$

This ends the proof of the first part of the Lemma.

Appendix D. Proof of Lemma 2

Fix a theory T, an instance process $\bar{x}_\infty$, and an increasing sequence of periods $1 = t_0 < t_1 ...$. For any k, let $A_k = \left\{x_{t_{k-1}+1}, ..., x_{t_k}\right\}$. I construct learning rule l such that for any k,

$$\inf_{\theta\in\mathcal{M}(T)}\sum_{t=t_{k-1}+1}^{t_k}\left[1 - l_t\left(\left(x_s,\theta(x_s)\right),x_t\right)(\theta(x_s))\right] \leq (t_k - t_{k-1})\,\mathcal{E}(A_k;T).$$

This will finish the proof of the second part of Lemma 1.

As in (C.4), define set

$$\mathcal{M}(y_1,...,y_s) := \left\{\tau\in\mathcal{M}\left(\left\{x_{t_{k-1}+1},...,x_{t_k}\right\},T\right) : \tau\left(x_{t_{k-1}+u}\right) = y_u \text{ for each } u\leq s\right\}.$$

$\mathcal{M}(y_1,...,y_s)$ consists of these model restrictions for which the initial s outcomes are equal to $y_1,...,y_s$, respectively. Then, $\mathcal{M}(\varnothing) = \mathcal{M}(A_k)$. Define learning rule l : for any $t = t_{k-1}+1,...,t_k$, let

$$l_t\left((x_s,y_s)_{s<t},x_t\right) := \arg\max_y \frac{\left|\mathcal{M}\left(y_{t_{k-1}+1},...,y_{t-1},y\right)\right|}{\left|\mathcal{M}\left(y_{t_{k-1}+1},...,y_{t-1}\right)\right|}.$$

Thus, the prediction of l is equal to the outcome $\tau(x_t)$ that is predicted by the largest subset of model restrictions from $\mathcal{M}\left(y_{t_{k-1}+1},...,y_{t-1}\right)$. At each period t,

$$\frac{\left|\mathcal{M}\left(y_{t_{k-1}+1},...,y_{t-1},y_t\right)\right|}{\left|\mathcal{M}\left(y_{t_{k-1}+1},...,y_{t-1}\right)\right|} \leq 1 \text{ if the prediction of } l \text{ is correct,}$$
$$\frac{\left|\mathcal{M}\left(y_{t_{k-1}+1},...,y_{t-1},y_t\right)\right|}{\left|\mathcal{M}\left(y_{t_{k-1}+1},...,y_{t-1}\right)\right|} \leq \frac{1}{2} \text{ if the prediction of } l \text{ is wrong.}$$

Therefore, the number of mistakes committed between periods $t_{k-1}+1$ and t_k cannot be higher than

$$\sum_{t=t_{k-1}+1}^{t_k}\left[1-l_t\left(\left(x_s,\theta\left(x_s\right)\right),x_t\right)\left(\theta\left(x_s\right)\right)\right] \leq \log\left|\mathcal{M}\left(\varnothing\right)\right|$$
$$= \left(t_k-t_{k-1}\right)\mathcal{E}\left(A_k;T\right).$$

Appendix E. Proof of Proposition 7

Fix locally generated group action $G \longmapsto X$.

A *predictor* is a tuple (C,p) of a finite subset $C \subseteq X$, called *support of the predictor*, and a function $p: C \to [0,1]$.

Definition 10. *Family of models $\mathcal{M}$ is* (ε',γ)-decomposable on local set U, *if, for any $D \subseteq U, |D| \geq \varepsilon'|U|$, there is a set of predictors $\mathcal{C}_D^*$ such that*

1. *for any $(C,p) \in \mathcal{C}_D^*$, $C \subseteq D$ and $|C| \geq \gamma|U|$;*
2. *$\log|\mathcal{C}_D^*| \leq \gamma\varepsilon'|U|$;*
3. *for any model $\theta \in \mathcal{M}$, there is $(C,p) \in \mathcal{C}_D^*$ so that*

$$\sum_{x\in C}\left|\theta\left(x\right)-p\left(x\right)\right| \leq \varepsilon'|C|. \tag{E.1}$$

Lemma 14. *Suppose that $\gamma \leq \varepsilon' \leq 1$ and family of models $\mathcal{M} \subseteq \{0,1\}^X$ is (ε',γ)-decomposable on local set U. Then, there exists a class of functions $\mathcal{T} \subseteq \{0,1\}^U$ such that $\log|\mathcal{T}| \leq \varepsilon'|U|$ and for each model $\theta \in \mathcal{M}$, there is $\tau \in \mathcal{T}$ such that*

$$\sum_{x\in U}\left|\theta\left(x\right)-\tau\left(x\right)\right| \leq 3\varepsilon'.$$

Proof. For any $D \subseteq U$, $|D| \geq \varepsilon'|U|$, let $\mathcal{C}_D^*$ be the family of subsets from the definition of decomposability. By decomposability, for any model $\theta \in \mathcal{M}$, there is an inductively constructed sequence of predictors (C_m,p_m), $m=0,...,n-1$, where $n \leq \frac{1}{\gamma}$, such that

- all sets C_m are mutually disjoint,

- $\left|U\setminus\bigcup_{m=0}^{n-1}C_m\right|\leq\varepsilon'|U|$, and
- for any $m<n$, $(C_m,p_m)\in\mathcal{C}^*\left(U\setminus\bigcup_{m'=0}^{m-1}C_{m'}\right)$ and

$$\sum_{x\in C_m}|\theta(x)-p_m(x)|\leq\varepsilon'|C_m|.$$

For any sequence of predictors (C_m,p_m), $m=0,...,n-1$, such that all sets C_m are mutually disjoint, $\left|U\setminus\bigcup_{m=0}^{n-1}C_m\right|\leq\varepsilon'|U|$, and $(C_m,p_m)\in\mathcal{C}^*\left(U\setminus\bigcup_{m'=0}^{m-1}C_{m'}\right)$ for each $m<n$, define function $\tau_{\{C_m,p_m\}}:U\to\{0,1\}$ as

$$\tau_{\{C_m,p_m\}}(x)=1,\ \text{if } p_m(x)\geq\frac{1}{2}\text{ for any } x\in C_m\setminus(D_0\cup D_1),$$
$$\tau_{\{C_m,p_m\}}(x)=0,\ \text{if } p_m(x)<\frac{1}{2}\text{ for any } x\in C_m\setminus(D_0\cup D_1),$$
$$\tau_{\{C_m,p_m\}}(x)=0,\ \text{if } x\notin\bigcup_{m=0}^{n-1}C_m.$$

The above implies that for any model $\theta\in\mathcal{M}$, there is $\tau_{\{C_m,p_m\}}$ such that

$$\begin{aligned}&\sum_{x\in U}\left|\theta(x)-\tau_{\{C_m,p_m\}}(x)\right|\\&\leq\left|U\setminus\bigcup_{m=0}^{n-1}C_m\right|+2\sum_m\sum_{x\in C_m}|\theta(x)-p_m(x)|\\&\leq 3\varepsilon'|U|.\end{aligned}$$

Define $\mathcal{T}$ as the family of functions $\tau_{\{C_m,p_m\}}$ where $\{C_m,p_m\}$ is a sequence of predictors that satisfy the above conditions. The cardinality of $\mathcal{T}$ can be bounded by the number of such sequences of predictors:

$$|\mathcal{T}|\leq\sup_{D\subseteq U,|D|\geq\varepsilon'|U|}\mathcal{C}_D^{*\frac{1}{\gamma}}\leq 2^{\gamma\varepsilon'|U|\frac{1}{\gamma}}\leq 2^{\varepsilon'|U|}.$$

□

Lemma 15. *Suppose that $\gamma\leq\varepsilon'\leq 1$ and family of models $\mathcal{M}\subseteq\{0,1\}^X$ is (ε',γ)-decomposable on local set U. Then,*

$$\frac{1}{|U|}\log\mathcal{M}_U\leq 6\varepsilon'\log\frac{1}{3\varepsilon'}+2(1-3\varepsilon')\log\frac{1}{1-3\varepsilon'}+\varepsilon'.$$

Proof. Take family of functions $\mathcal{T}$ from the previous lemma. For any $\tau \in \mathcal{T}$, any sets $D_0, D_1 \subseteq X$, define function $\tau_{D_0,D_1} : U \to \{0,1\}$ as

$$\tau_{D_0,D_1}(x) = \tau(x) \text{ if } x \notin D_0 \cup D_1,$$
$$\tau_{D_0,D_1}(x) = 0 \text{ for any } x \in D_0 \text{ and}$$
$$\tau_{D_0,D_1}(x) = 1 \text{ for any } x \in D_1.$$

The above implies that for any model $\theta \in \mathcal{M}$, there are $D_0, D_1 \subseteq U$, $|D_0|, |D_1| \leq 3\varepsilon' |U|$ such that

$$\theta|_U = \tau_{D_0,D_1}.$$

Hence, the set $\mathcal{M}_U$ of model restrictions on U is bounded by the cardinality of the set of all functions τ_{D_0,D_1} where $|D_0|, |D_1| \leq 3\varepsilon' |U|$. By the Stirling's formula and the previous Lemma,

$$\begin{aligned}\frac{1}{|U|}\log \mathcal{M}_U &\leq \frac{1}{|U|}\log |\{\tau_{D_0,D_1} : \tau \in \mathcal{T}, |D_0|, |D_1| \leq 3\varepsilon' |U|\}| \\ &\leq \frac{1}{|U|}\log \begin{pmatrix} |U| \\ 3\varepsilon' |U| \end{pmatrix}^2 + \frac{1}{|U|}\log |\mathcal{T}| \\ &\leq 2\left(3\varepsilon' \log \frac{1}{3\varepsilon'} + (1-3\varepsilon')\log \frac{1}{1-3\varepsilon'}\right) + \varepsilon'.\end{aligned}$$

□

Definition 11. *A* S-restriction *is a family of functions* $\{\tau_g\}_{g\in G}$ *such that* $\tau_g : g \cdot S \to \{0,1\}$. *Family of models* $\mathcal{M} \subseteq \{0,1\}^X$ respects *S-restriction* $\{\tau_g\}$ *if* $\theta|_{g\cdot S} \neq \tau_g$ *for any model* $\theta \in \mathcal{M}$ *and any* $g \in G$.

If G-dimension of family $\mathcal{M}$ contains S, then there exists S-restriction respected by family $\mathcal{M}$. Conversely, if family $\mathcal{M}$ respects an S-restriction, then G-dimension of $\mathcal{M}$ contains set S.

Lemma 16. *Fix S-restriction $\{\tau_g\}$. For any $\varepsilon' \in (0,1)$, there exist $\gamma \leq \varepsilon'$ and local V such that for any permutation g, any local U so that $g \cdot V \subseteq U$ and $\gamma\varepsilon' |U| \geq 1$, if family of models $\mathcal{M}$ respects $\{\tau_g\}$, then it is (ε', γ)-decomposable on U.*

The Lemma is proved below.

Proof of Proposition 7. Fix $\varepsilon > 0$. Find $0 < \varepsilon' \leq 1$ small enough, so that

$$6\varepsilon' \log \frac{1}{3\varepsilon'} + 2(1-3\varepsilon')\log \frac{1}{1-3\varepsilon'} + \varepsilon' \leq \varepsilon. \tag{E.2}$$

The Proposition follows from the Lemmas. □

E.1. **Proof of Lemma 16.** I begin with a sketch of the main idea behind the proof. Suppose that U is local set, $D \subseteq U$ is its subset of sufficient size, $|D| \geq \varepsilon' |U|$, and, for simplicity, that $\mathcal{M}$ is a family of models that omit pattern (S, τ), where S is generic, $\mathcal{M} = \mathcal{M}(\{S, \tau\})$. I need to find a "exponentially small" set of predictors $\mathcal{C}_D^*$, such that each predictor $(C, p) \in \mathcal{C}_D^*$ has "large" support $C \subseteq U$ and for each model $\theta \in \mathcal{M}$, there is a predictor $(C, p) \in \mathcal{C}_D^*$ such that p approximates model θ on set C. Here and in what follows, "$A \subseteq B$ is large" means that proportion of the size of A to the size of B is bounded away from 0; "A is (exponentially) small" if the size of A divided by the size of B converges to 0 (exponentially quickly) when the size of B increases to infinity.

Fix any $x^* \in S$. Say that model $\theta \in \mathcal{M}$ is "nice" if there is a "small" set $W \subseteq U$ and a "large" subset $C \subseteq D$ such that for any $x \in C$ there is a permutation g st.

- $g \cdot x^* = x$,
- for each $x' \in S \backslash \{x^*\}$, $g \cdot x' \in W$ and $\theta(g \cdot x') = \tau(x')$.

Because model θ omits pattern (S, τ), the above implies that for each $x \in C$, $\theta(x) = 1 - \tau(x^*)$. One can construct a set of predictors that approximate all "nice" models. For each "small" $W \subseteq U$, each $y : W \to Y$, let

$$C^{x^*}_{W,y;D} = \{g \cdot x^* \in D : g \cdot (S \backslash \{x^*\}) \subseteq W \text{ and } \forall_{x \in S, x \neq x^*} y(g \cdot x) = \tau(x)\}$$
$$p^{x^*}_{W,y;D} = 1 - \tau(x^*).$$

Each predictor $\left(C^{x^*}_{W,y;D}, p^{x^*}_{W,y;D}\right)$ depends on set W and on the configuration y of outcomes on set W. The number of such predictors can be bounded by

$$\begin{aligned}
&\frac{1}{|U|} \log \left| \left\{ \left(C^{x^*}_{W,y;D}, p^{x^*}_{W,y;D}\right) \right\} \right| \\
&\leq \frac{1}{|U|} \log \left[\begin{pmatrix} \text{number of "small"} \\ W \subseteq U \end{pmatrix} 2^{|W|} \right] \\
&= \frac{1}{|U|} \log \begin{pmatrix} \text{number of "small"} \\ W \subseteq U \end{pmatrix} + \frac{|W|}{|U|}.
\end{aligned}$$

If set W is sufficiently "small", the above expression is also "small".

Although predictors $\left(C^{x^*}_{W,y;D}, p^{x^*}_{W,y;D}\right)$ approximate all "nice" models, family $\mathcal{M}$ may also contain models that are not "nice." The proof in the case of these models is based on the following observation: If S is generic, U is sufficiently large, and model θ is not "nice"on large set D, then it (approximately) omits pattern $\left(S \backslash \{x^*\}, \tau|_{S \backslash \{x^*\}}\right)$. The support of such pattern is smaller than the support of the original pattern. The above procedure can be repeated. One identifies models that are "nice" with respect to the new pattern, and finds a set of predictors that approximate these models. This procedure is

repeated at most $|S|$ times. If the size of the set of predictors found at each step of the procedure is small, then the sum of these sizes is also small. This will yield the Lemma.

E.1.1. *Notation.* From now on, fix $\varepsilon' \in (0,1)$, generic $S \subseteq X$, $|S| = k$, and S-restriction $\{\tau_g\}$. The proof of the Lemma is divided into few parts.

I start with notation. An *enumeration* of S is any tuple $\bar{x}^* = (x_1^*, ..., x_k^*) \in X^k$, such that $S = \{x_1^*, ..., x_k^*\}$. Define set of k-tuples of that are obtained as permutations of enumeration $\bar{x}^*$:

$$A(\bar{x}^*) = \left\{g \cdot \bar{x}^* \in X^k : g \in G\right\}.$$

By the definition of S-restriction, there exists a function $\tau^{\bar{x}^*} : A(\bar{x}^*) \to \{0,1\}^k$ such that for any family of models $\mathcal{M} \subseteq \{0,1\}^X$, if family $\mathcal{M}$ respects S-restriction $\{\tau_g\}$, then, for any $\theta \in \mathcal{M}$ and any $\bar{x} \in A(\bar{x}^*)$, there is $l \leq k$ such that $\theta(x_l) \neq \tau_l^{\bar{x}^*}(\bar{x})$.

Take any local $U \subseteq X$ and an enumeration $\bar{x}$ of S. Define three sets of tuples:

(1) Let

$$A_U(\bar{x}^*) = A(\bar{x}^*) \cap U^k. \tag{E.3}$$

Set $A_U(\bar{x}^*)$ consists of k-tuples of elements of U that are obtained as permutations of enumeration $\bar{x}^*$. Recall that $G_U \subseteq G$ is the subgroup of all permutations that keep set U fixed (see 2.14). Group action $G_U \longmapsto U$ induces group action $G_U \longmapsto A_U(\bar{x}^*)$. Because U is local, $G_U \longmapsto A_U(\bar{x}^*)$ is transitive.

(2) For any $l \leq k$, $x \in U$, define set of k-tuples:

$$A_U^l(x; \bar{x}^*) = \{\bar{x} \in A_U(\bar{x}^*) : x_{l'} = x\}.$$

This is a set of tuples from $A_U(\bar{x}^*)$, for which the lth element is equal to x. Because $G \longmapsto A_U(\bar{x}^*)$ is transitive, it follows immediately that for any $x, x' \in U$,

$$\left|A_U^l(x; \bar{x}^*)\right| = \left|A_U^l(x'; \bar{x}^*)\right| = \frac{|A_U(\bar{x}^*)|}{|U|}. \tag{E.4}$$

(3) For any $l \leq k+1$, any model $\theta : X \to \{0,1\}$, define set of k-tuples:

$$A_U^l(\theta; \bar{x}^*) = \left\{\bar{x} \in A_U(\bar{x}^*) : \theta(x_{l'}) = \tau_{l'}^{\bar{x}^*}(\bar{x}) \text{ for any } l' \leq l-1\right\}. \tag{E.5}$$

Here, $A_U^l(\theta; \bar{x}^*)$ is a set of tuples $\bar{x} \in A_U(\bar{x}^*)$ such that the values of model θ at the first $l-1$ elements are equal to the corresponding first $l-1$ coordinates of $\tau^{\bar{x}^*}(\bar{x})$.

For any $B \subseteq X^k$, let $\mathbf{1}_B : A_U(\bar{x}^*) \to R$ be the indicator function of set B: $\mathbf{1}_B(\bar{x}) = 1$ iff $\bar{x} \in B$. For example, if $D \subseteq U$, $\mathbf{1}_{D^k}$ is an indicator of tuples whose all elements belong

to D. For any function $f : A_U(\bar{x}^*) \to \mathbf{R}$, define integral $E_{U,\bar{x}^*} f$ as in (B.1). Because $\sum_{x \in U} \mathbf{1}_{A_U^l(x,\bar{x}^*)} = \mathbf{1}$,

$$E_{U,\bar{x}^*} f = \frac{1}{A_U(\bar{x}^*)} \sum_{\bar{x} \in A_U(\bar{x}^*)} f(\bar{x}) = \sum_{x \in U} E_{U,\bar{x}^*} f \mathbf{1}_{A_U^l(x,\bar{x}^*)}. \tag{E.6}$$

E.1.2. *Indicator functions.* The next result identifies a class of subsets of local U with some useful properties. The proof of the Lemma can be found in Appendix E.3

Lemma 17. *For any $\lambda > 0$, there is a local set V such that for any permutation g and any local $U \supseteq g \cdot V$, there are an enumeration $\bar{x}^*$ of S, and sets $W_l \subseteq U, l \leq k$ such that, for each $l \leq k$, $|W_l| \leq \lambda |U|$ and*

$$|\{g \cdot x_l^* : \forall_{l' < l} g \cdot x_{l'}^* \in W_l, g \in G\}| \geq \frac{1}{2k} |U|.$$

I use the indicator sets to construct a class of indicator functions. For any $f : A_U(\bar{x}) \to \mathbf{R}$, for any permutation g, let $f_g : A_U(\bar{x}) \to \mathbf{R}$ be a function that is obtained as a g-permutation of f, i.e., $f_g(\bar{x}) = f(g \cdot \bar{x})$ for any $\bar{x} \in A_U(\bar{x})$. Say that set of functions $\mathcal{F} \subseteq (A_U(\bar{x}))^{\mathbf{R}}$ is G_U-*invariant*, if $f \in \mathcal{F}$ implies $f_g \in \mathcal{F}$.

Lemma 18. *For any $\eta > 0$, there is a local set V such that for any permutation g and any local $U \supseteq g \cdot V$, there is an enumeration $\bar{x}^*$ of S and G_U-invariant families of functions $\mathcal{F}_U^l \subseteq (A_U(\bar{x}^*))^{\mathbf{R}}, l = 1, ..., k$, such that for any $l \leq k$, for any $f \in \mathcal{F}_U^l$, any $x \in U$,*

$$E_{U,\bar{x}^*} f \geq \frac{1}{2k} \text{ and } E_{U,\bar{x}^*} f \mathbf{1}_{A_U^l(x;\bar{x}^*)} \leq \frac{1}{|U|} \text{ for each } x \in U.$$

and

$$\frac{1}{|U|} \log \left| \left\{ \mathbf{1}_{A_U^l(\theta;\bar{x})} f : l \leq k, \theta \in Y^X, f \in \mathcal{F}_U^l \right\} \right| \leq \eta.$$

Proof. Find $\lambda > 0$ so that

$$\lambda \log \frac{1}{\lambda} + (1 - \lambda) \log \frac{1}{1 - \lambda} + \lambda \leq \eta.$$

Find a local set V such that for any local $U \supseteq g \cdot V$, there are an enumeration $\bar{x}^*$ of S and sets $W_l \subseteq U$, $|W_l| \leq \lambda |U|$ such that the thesis of Lemma 17 holds. Fix $l \leq k$. Define set of tuples

$$T_A^l := \{\bar{x} \in A : \{x_1, ..., x_{l-1}\} \subseteq W_l\}.$$

Define $f_U^l : A_U(\bar{x}^*) \to R$ as

$$f_U^l(\bar{x}) = \begin{cases} \frac{|A|}{|U|} \frac{1}{\left|A_U^l(x_l;\bar{x}^*) \cap T_A^l\right|} & \text{if } \bar{x} \in T_A^l, \\ 0 & \text{if } \bar{x} \notin T_A^l. \end{cases}$$

Define also a family of functions

$$\mathcal{F}_U^l := \left\{ \left(f_U^l\right)_g : g \in G_U \right\}.$$

By definition, family $\mathcal{F}_U^l$ is G_U-invariant.

For any $x \in U$ such that $A_U^l(x;\bar{x}^*) \cap T_A^l$ is empty, it must be that

$$E_{U,\bar{x}} f_U^l \mathbf{1}_{A_U^l(x;\bar{x}^*)} = E_{U,\bar{x}} f_U^l(\bar{x}) \mathbf{1}_{A_U^l(x;\bar{x}^*)} = \frac{1}{|A_U(\bar{x})|} \sum_{g\cdot\bar{x} \in A_U^l(x;\bar{x}^*)\cap T_A^l} = 0.$$

For any $x \in U$ such that $A_U^l(x;\bar{x}^*) \cap T_A^l$ is not empty, it must be that

$$E_{U,\bar{x}} f^l \mathbf{1}_{A_U^l(x;\bar{x}^*)} = \frac{1}{|A_U(\bar{x})|} \sum_{\bar{x} \in A_U^l(x;\bar{x}^*)\cap T_A^l} \frac{|A_U(\bar{x})|}{|U|} \frac{1}{\left|A_U^l(x_l;\bar{x}^*) \cap T_A^l\right|} = \frac{1}{|U|}.$$

Therefore, for any $g \in G_U$,

$$E_{U,\bar{x}} \left(f_U^l\right)_g \mathbf{1}_{A_U^l(x;\bar{x}^*)} = E_{U,\bar{x}} f_U^l \mathbf{1}_{A_U^l(g^{-1}\cdot x;\bar{x}^*)} \leq \frac{1}{|U|},$$

and, by (**??**) and (E.6),

$$\begin{aligned} E_{U,\bar{x}} \left(f_U^l\right)_g = E_{U,\bar{x}} f_U^l &= \sum_{x\in U} E_{U,\bar{x}} f_U^l \mathbf{1}_{A_U^l(x;\bar{x}^*)} \\ &= \sum_{x\in U : A_U^l(x;\bar{x}^*)\cap T_A^l \neq \varnothing} E_{U,\bar{x}} f_U^l \mathbf{1}_{A_U^l(x;\bar{x}^*)} \\ &= \frac{\left|\left\{x \in U : A_U^l(x;\bar{x}^*) \cap T_A^l \neq \varnothing\right\}\right|}{|U|} \geq \frac{1}{2k}. \end{aligned}$$

This shows the first part of the Lemma.

By definition, if $g \cdot W_l = W_l$, then $\bar{x} \in T_A^l$ iff $g \cdot T_A^l$. Hence, $\left|A_U^l(x_l;\bar{x}^*) \cap T_A^l\right| = \left|A_U^l(g\cdot x_l;\bar{x}^*) \cap T_A^l\right|$ and

$$\left(f_U^l\right)_g = f_U^l.$$

Take any $g, g' \in G_U$ and $\theta, \theta' : X \to Y$ such that $g \cdot W_l = g' \cdot W_l$ and $\theta|_{g\cdot W^l} = \theta'|_{g'\cdot W^l}$. Then,

$$\left(f_U^l\right)_g = \left(f_U^l\right)_{g'} \text{ and } \mathbf{1}_{A_U^l(\theta;\bar{x}^*)} \left(f_U^l\right)_g = \mathbf{1}_{A_U^l(\theta';\bar{x}^*)} \left(f_U^l\right)_{g'}.$$

In other words, functions $\mathbf{1}_{A_U^l(\theta;\bar{x}^*)} \left(f_U^l\right)_g$ depend only on $g\cdot W_l$ and $\theta|_{W_l}$. There are (a) at most $\begin{pmatrix} |U| \\ |W_l| \end{pmatrix}$ ways of choosing set $g \cdot W_l^l \subseteq U$ and (b) at most $2^{|W_l|}$ ways of choosing

restriction $\theta|_{g\cdot W_l}$. Because $|W_l| \leq \lambda |U|$, an application of the Stirling formula yields:

$$\begin{aligned}
&\frac{1}{|U|}\log\left|\left\{\mathbf{1}_{A^l_U(\theta;\bar{x}^*)}f^{l*}_g : g \in G_U \text{ and } \theta : X \to \{0,1\}\right\}\right| \\
&\leq \frac{1}{|U|}\log\left[\left(\begin{array}{c}|U|\\|W|\end{array}\right)2^{|W|}\right] \\
&\leq \lambda\log\frac{1}{\lambda} + (1-\lambda)\log\frac{1}{1-\lambda} + \lambda \leq \eta.
\end{aligned}$$

□

E.1.3. *Predictors.* Next, I construct a class of predictors $\left(C^l_{f;D}, p^l_{f,D}\right)$. Each predictor is indexed with a function $f : A_U(\bar{x}^*) \to \mathbf{R}$, subset $D \subseteq U$ and position $l \leq k$. Let

$$C^l_{f;D} = \left\{x \in D : E_{U,\bar{x}^*} f \mathbf{1}_{A^l(x;\bar{x}^*)}\mathbf{1}_{D^k} \geq \frac{\varepsilon'}{4k|U|}\gamma^{(l)}\right\} \text{ and} \tag{E.7}$$

$$p^l_{f,D}(x) = 1 - \frac{E_{U,\bar{x}^*}\mathbf{1}_{A^l(x;\bar{x})}\mathbf{1}_{D^k} f\tau^{\bar{x}^*}_l}{E_{U,\bar{x}^*}\mathbf{1}_{A^l(x;\bar{x})}\mathbf{1}_{D^k} f} \text{ for any } x \in C^l_{f,D}, \tag{E.8}$$

The definitions of predictors $\left(C^l_{f,D}, p^l_{f,D}\right)$ depend on the local set U and enumeration $\bar{x}^*$ of S. This dependence is not mentioned in order to save on notation.

For each $D \subseteq U$, define

$$\mathcal{C}^*_D = \left\{\left(C^l_{f;D}, p^l_{f,D}\right) : l \leq k, f = \mathbf{1}_{A^l_U(\theta;\bar{x})}\hat{f}, \theta \in Y^X, \hat{f} \in \mathcal{F}^l_U\right\}.$$

The next Lemma bounds the cardinality of the set of predictors.

Lemma 19. *For any $\gamma > 0$, there is a local set V such that for any permutation g and any local $U \supseteq g \cdot V$, any $l \leq k$, any $D \subseteq U, |D| \geq \varepsilon'|U|$,*

$$\log|\mathcal{C}^*_D| \leq \varepsilon'\gamma|U|,$$

where $\mathcal{C}^*_D$ is defined as above.

Proof. Let $\eta = \frac{\varepsilon'\gamma}{2}$. Find a local set V such that for any permutation g and any local $U \supseteq g \cdot V$, there is an enumeration $\bar{x}^*$ of S and G_U-invariant families of functions $\mathcal{F}^l_U \subseteq (A_U(\bar{x}^*))^{\mathbf{R}}, l = 1, ..., k$, such that for any $l \leq k$, the thesis of Lemma **??** holds. W.l.o.g., I can assume that $|U| \geq |V| \geq \frac{1}{\eta}\log k$. The result follows from the fact that

$$\begin{aligned}
\log|\mathcal{C}^*_D| &\leq \log\left|\left\{\mathbf{1}_{A^l_U(\theta;\bar{x})}f : l \leq k, \theta \in Y^X, f \in \mathcal{F}^l_U\right\}\right| + \log k \\
&\leq 2\eta|U| \leq \varepsilon'\gamma|U|.
\end{aligned}$$

□

E.1.4. *Constants.* Because set S is generic, then, for any $\varepsilon' > 0$, there is a local set U such that for any enumeration $\bar{x}$ of S, any $D \subseteq U$ st. $|D| \geq \varepsilon' |U|$, the integral $E_{U,\bar{x}} \mathbf{1}_{D^k} > 0$ is bounded away from 0. The next result strengthens this property so that it holds uniformly across all sufficiently large local Us.

Lemma 20. *There is $\delta^* > 0$ and a local set V such that for any permutation g, any local $U \supseteq g \cdot V$, any enumeration $\bar{x}^*$ of S, and any $D \subseteq U$ st. $|D| \geq \varepsilon' |U|$,*

$$E_{U,\bar{x}^*} \mathbf{1}_{D^k} \geq \delta^*.$$

The Lemma is proven in Appendix E.2. Take $\delta^* > 0$, so that thesis of Lemma 20 holds for ε'. Define constants $\gamma^{(l)}, \gamma > 0$, for any $1 \leq l \leq k+1$:

$$\begin{aligned} \gamma^{(1)} &= \delta^* \text{ and, inductively,} \\ \gamma^{(l+1)} &= \frac{(\varepsilon')^3}{50k} (1 - \varepsilon') \left(\gamma^{(l)}\right)^2 \text{ for any } l \leq k, \\ \gamma &= \frac{1}{4k} (1 - \varepsilon') \gamma^{(k)}. \end{aligned} \tag{E.9}$$

E.1.5. *Decomposability.* Let V be a local set from Lemma 20. Take any local U such that$U \supseteq g \cdot V$ for some permutation g. Suppose that family of models $\mathcal{M}$ respects S-restriction $\{\tau_g\}$. In this section, I prove that sets of predictors $\{\mathcal{C}^*_D\}_{D \subseteq U}$ decompose family $\mathcal{M}$. In order to shorten the notation, in the rest of the proof I drop subscripts referring to U, $\bar{x}$ and τ: Instead of

$$A_U(\bar{x}^*), A^l_U(x; \bar{x}^*), A^l_U(\theta; \bar{x}^*), G_U, \mathcal{F}^l_U, \tau^{\bar{x}^*}_l, E_{U,\bar{x}^*},$$

I write

$$A, A^l(x), A^l(\theta), G, \mathcal{F}^l, \tau_l, E.$$

Lemma 21. *Suppose that family of models $\mathcal{M}$ respects S-restriction $\{\tau_g\}$. For any $D \subseteq U$, $|D| \geq \varepsilon' |U|$, any model $\theta \in \mathcal{M}$, there is $l \leq k$ so that*

$$E\mathbf{1}_{A^l(\theta)} \mathbf{1}_{D^k} \geq \gamma^{(l)} \text{ and } E\mathbf{1}_{A^{l+1}(\theta)} \mathbf{1}_{D^k} \leq \gamma^{(l+1)}. \tag{E.10}$$

Proof. By definition (E.5), $A^1(\theta) = A$. By Lemma 20, $E\mathbf{1}_{A^1(\theta)} \mathbf{1}_{D^k} = E\mathbf{1}_{D^k} \geq \delta^* = \gamma^{(1)}$. On the other hand, if $\theta \in \mathcal{M}$, then $\mathbf{1}_{A^{k+1}(\theta)} \equiv 0$. Hence, $E\mathbf{1}_{A^{k+1}(\theta)} \mathbf{1}_{D^k} = 0 \leq \gamma^{(k+1)}$. The Lemma follows. □

Lemma 22. *For any $\varepsilon > 0$, any model $\theta : X \to \{0, 1\}$, and any $D \subseteq U$, if (E.10) holds for some $l \leq k$, then there is $\hat{f} \in \mathcal{F}^l$ such that*

$$\left|C^l_{f,D}\right| \geq \gamma |U| \text{ and } \sum_{x \in C^l(f;D)} \left|\theta(x) - p^l_{f;D}\right| \leq \varepsilon' \left|C^l_{f;D}\right|, \tag{E.11}$$

where $f := \mathbf{1}_{A^l_U(\theta;\bar{x})}\hat{f}$.

Proof. Because $\mathcal{F}^l$ is G-invariant, by the second part of Lemma 8, there is $\hat{f} \in \mathcal{F}^l$ such that

$$E\mathbf{1}_{A^l(\theta)}\mathbf{1}_{D^k}\hat{f} \geq \frac{1}{2}E\hat{f}E\mathbf{1}_{A^l(\theta)}\mathbf{1}_{D^k} \text{ and}$$
$$E\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}\hat{f} \leq 3\frac{E\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}}{E\mathbf{1}_{A^l(\theta)}\mathbf{1}_{D^k}}E\mathbf{1}_{A^l(\theta)}\mathbf{1}_{D^k}\hat{f}.$$

Take

$$f = \mathbf{1}_{A^l_U(\theta)}\hat{f}.$$

Because of (E.10),

$$f = \mathbf{1}_{A^l(\theta)}f, \tag{E.12}$$
$$Ef\mathbf{1}_{A^l(x)} \leq \frac{1}{|U|} \text{ for each } x \in U \tag{E.13}$$
$$E\mathbf{1}_{D^k}f \geq \frac{1}{4k}\gamma^{(l)} \tag{E.14}$$
$$E\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}f \leq 3\frac{\gamma^{(l+1)}}{\gamma^{(l)}}E\mathbf{1}_{D^k}f. \tag{E.15}$$

By (E.13) and the definition of set $C^l_{f;D}$

$$\begin{aligned} E\mathbf{1}_{D^k}f &= \sum_{x\in C^l_{f;D}} E\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k}f + \sum_{x\notin C^l_{f;D}} Ef\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k} \\ &\leq \left|C^l_{f;D}\right|\frac{1}{|U|} + \frac{1}{4k}\varepsilon'\gamma^{(l)}. \end{aligned}$$

Therefore, by (E.14),

$$\begin{aligned} \frac{1}{|U|}\left|C^l_{f;D}\right| &\geq Ef\mathbf{1}_{D^k} - \frac{1}{4k}\varepsilon'\gamma^l \geq (1-\varepsilon')\,Ef\mathbf{1}_{D^k} \\ &\geq \frac{1}{4k}(1-\varepsilon')\,\gamma^l \geq \gamma. \end{aligned} \tag{E.16}$$

This shows the first part of (E.11).

Note that for any $\bar{x} \in A^l(\theta)$, $\bar{x} \in A^{l+1}(\theta)$ if and only if $\theta(x_l) = \tau_l^{\bar{x}^*}(\bar{x})$. Hence,

$$\begin{aligned} &\left|(\theta(x)-1)\,E\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k}f + E\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k}f\tau_l\right| \\ &= \sum_{\bar{x}\in A}\frac{1}{|A|}\mathbf{1}_{A^l(x)}(\bar{x})\left|\tau_l(\bar{x}) - (1-\theta(x))\right|\mathbf{1}_{D^k}(\bar{x})\,\mathbf{1}_{A^l(\theta)}(\bar{x})\,f(\bar{x}) \\ &= E\mathbf{1}_{A^l(x)}\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k}f \end{aligned}$$

and, by (E.12),

$$\begin{aligned}
&\sum_{x\in C^l_{f,D}} \left|\theta(x) - p^l_{f,D}(x)\right| \\
&= \sum_{x\in C^l_{f;D}} \frac{\left|(\theta(x)-1)\, E\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k} f - E\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k} f\tau_l\right|}{E\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k} f} \\
&= \sum_{x\in C^l_{f;D}} \frac{E\mathbf{1}_{A^l(x)}\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k} f}{E\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k} f}.
\end{aligned}$$

Suppose that the second part of (E.11) does not hold. Then, there is subset $B \subseteq C^l_{f;D}$ such that

$$|B| \geq \frac{\varepsilon'}{2}\left|C^l_{f;D}\right| \geq \frac{\varepsilon'}{2}(1-\varepsilon')\,|U|\, E\mathbf{1}_{D^k} f$$

(the second inequality comes from (E.16)), and for each $x \in B$,

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

But then,

$$\begin{aligned}
E\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k} f &= \sum_{x\in U} E\mathbf{1}_{A^l(x)}\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k} f \\
&\geq \sum_{x\in B} E\mathbf{1}_{A^l(x)}\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k} f \\
&\geq \frac{\varepsilon'}{2}\sum_{x\in B} E\mathbf{1}_{A^l(x)}\mathbf{1}_{D^k} f \\
&\geq \frac{(\varepsilon')^2}{8k}\gamma^{(l)}\frac{|B|}{|U|} \geq \frac{(\varepsilon')^3}{16k}(1-\varepsilon')\,\gamma^{(l)} Ef\mathbf{1}_{D^k}.
\end{aligned}$$

(The third inequality is a consequence of the definition of set $C^l_{f;D}$.) By inequality (E.15),

$$3\frac{\gamma^{(l+1)}}{\gamma^{(l)}} E\mathbf{1}_{D^k} f \geq E\mathbf{1}_{A^{l+1}(\theta)}\mathbf{1}_{D^k} f \geq \frac{(\varepsilon')^3}{16k}(1-\varepsilon')\,\gamma^{(l)} Ef\mathbf{1}_{D^k}.$$

This leads to

$$\gamma^{(l+1)} \geq \frac{(\varepsilon')^3}{48k}(1-\varepsilon')\left(\gamma^{(l)}\right)^2,$$

which contradicts the definition of constant $\gamma^{(l+1)}$. The contradiction shows the second part of (E.11). □

E.1.6. *Proof of Lemma 16.* Find local V such that the theses of Lemmas 19 and 20 hold. Such a local set V exists, because G is locally generated and for any finite sets S and S' there is a local V and permutations g, g' such that $V \supseteq g \cdot S \cup g' \cdot S'$. Lemma 16 is a corollary to Lemmas 19, 21, and 22.

E.2. **Proof of Lemma 20.** Fix generic S and $\varepsilon' > 0$. By definition, there is a local V such that for any subset $D \subseteq V, |D| \geq (\varepsilon')^2 |V|$,

$$E_{V,\bar{x}} \mathbf{1}_{D^k} \geq \frac{1}{|A_V(\bar{x})|} > 0.$$

I will show that for any local $U \supseteq V$ and any $D \subseteq U, |D| \geq \varepsilon' |U|$,

$$E_{U,\bar{x}} \mathbf{1}_{D^k} \geq \delta^* := \frac{1}{2}\varepsilon' \frac{1}{|A_V(\bar{x})|} > 0.$$

Let $G_U = \{g \in G : g \cdot U = U\}$. Because U is local and $G \longmapsto X$ is transitive, $G_U \longmapsto U$ is a transitive group action. By the first part of Lemma 8, for any subset $D \subseteq U$,

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

For any $D \subseteq U$, define

$$\alpha(D) = \frac{\left|\left\{g \in G_U : |D \cap g \cdot V| \geq (\varepsilon')^2 |V|\right\}\right|}{|G_U|}.$$

Suppose that $|D| \geq \varepsilon' |U|$. Then, by (E.17),

$$\varepsilon' \leq \frac{|D|}{|U|} = \frac{1}{|G|} \sum_{g \in G_U} \frac{|D \cap g \cdot V|}{|V|} \leq \alpha(D) + (\varepsilon')^2 (1 - \alpha(D));$$

hence,

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

Because U is local, group action $G_U \longmapsto U$ induces transitive group action $G_U \longmapsto A_U(\bar{x})$. Take any $D \subseteq U$, such that $|D| \geq \varepsilon' |U|$. By the first part of Lemma 8,

$$\frac{|A_D(\bar{x})|}{|A_U(\bar{x})|} \frac{|A_V(\bar{x})|}{|A_U(\bar{x})|} = \frac{1}{|G|} \sum_{g \in G_U} \frac{|A_{D \cap g \cdot V}(\bar{x})|}{|A_U(\bar{x})|}.$$

By definition, for any set $D \subseteq X$ such that $|D \cap g \cdot V| \geq (\varepsilon')^2 |V|$,

$$\frac{|A_{D \cap g \cdot V}(\bar{x})|}{|A_V(\bar{x})|} \geq \frac{1}{|A_V(\bar{x})|}. \tag{E.17}$$

Hence,

$$\frac{|A_D(\bar{x})|}{|A_U(\bar{x})|} = \frac{1}{|G|}\sum_{g\in G_U}\frac{|A_{D\cap g\cdot V}(\bar{x})|}{|A_V(\bar{x})|} \geq \alpha(D)\frac{1}{|A_V(\bar{x})|} \geq \frac{1}{2}\varepsilon'\delta(V,\bar{x},\varepsilon).$$

E.3. **Proof of Lemma 17.** Fix generic S, $k = |S|$. In the results that follow, phrase "S is λ-generic for local U" for some $\lambda > 0$ means that U is a local set, such that for any subset $D \subseteq U, |D| \geq \lambda|U|$, there is a permutation $g \in G$, such that $g \cdot S \subseteq U$. I also assume that $S \subseteq U$.

Lemma 23. *For any local U, for which S is λ-generic, there are $x^* \in S$ and subsets $W, T \subseteq U$ such that*

$$|W| \leq \lambda|U| \text{ and } |T| \geq \frac{1}{2k}|U|,$$

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

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

(a) for any permutation $g \in G$, $g \cdot S \nsubseteq W$ and

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

There is at least one such a set and $|W| < \lambda|U|$ because S is λ-generic for U. For any $x^* \in S$, define sets $T_{x^*} \subseteq U$ such that for any $x \in T_{x^*}$, there is a permutation $g \in G$ such that $g \cdot x^* = x$ and $g \cdot (S\setminus\{x^*\}) \subseteq W$. Then, $\bigcup_{x^*} T_{x^*} = U\setminus W$, and there is $x^* \in S$ such that

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

□

Lemma 24. *For any local U, for which S is λ-generic, there are an enumeration $\bar{x}^*$ of S and sets $W^l, T^l, l \leq k$ such that for any $l \leq k$*

$$\left|W^l\right| \leq \lambda|U| \text{ and } \left|T^l\right| \geq \frac{1}{2k}|U|, \tag{E.18}$$

and

$$T^l \subseteq \left\{g \cdot x_l^* : g \in G_U, g \cdot \left\{x_1^*, ..., x_{l-1}^*\right\} \subseteq W^l\right\}. \tag{E.19}$$

Proof. Note that if S is λ-generic for U, then any subset $S' \subseteq S$ is also λ-generic for U. The Lemma is a corollary to this observation and to Lemma 23. Enumeration $\bar{x}^*$ is constructed by induction. Let $\bar{x}^k$ be any enumeration of S. By Lemma 23, there is an enumeration $\bar{x}^{k*} = \left(x_1^{k*}, ..., x_{k-1}^{k*}, x_k^*\right)$ and sets W^k, T^k that satisfy the thesis of the Lemma. Next, consider $S^{k-1} = S\setminus\{x_k^*\}$. Then, S^{k-1} is λ-generic for U. By Lemma 23,

there is an enumeration $\bar{x}^{(k-1)*} = \left(x_1^{(k-1)*}, ..., x_{k-2}^{(k-1)*}, x_{k-1}^*\right)$ of S^{k-1} and sets W^{k-1}, T^{k-1} such that (E.18) and (E.19) hold for $l = k-1$.A repetition of this argument for $l = k-2, k-3, ..., 1$ proves the lemma the Lemma. □

I use the above results to finish the proof of Lemma 17

Proof of Lemma 17. By Lemma 20, there is a local V such that for any permutation g and any local $U \supseteq g \cdot V$, S is λ-generic for U. I can assume w.l.o.g. that $S \subseteq U$ (this is because the group action is locally generated.) The result follows from Lemma 24. □

Appendix F. Proofs of Section 5

F.1. **Proof of Lemma 4.** The proof is divided into three parts.

F.1.1. *Product of local sets is local.* I consider only the product of two groups, $d = 2$. The general result follows by induction on d. Let U_j be local sets under group actions $G_j \longmapsto X_j$ 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 jth coordinate:

$$S_j = \{x_j : (x_j, x_{-j}) \in S\}.$$

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.

F.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 $\bar{x}^{*j} = \left(x_1^{*j}, ..., x_{k_j}^{*j}\right)$ be an enumeration of S_j, i.e. $\left\{x_1^{*j}, ..., x_{k_j}^{*j}\right\} = S_j$.

For any j, any local set $U_j \subseteq X_j$, any $\varepsilon > 0$, let $A_{U_j}(\bar{x}^{*j}) \subseteq U^{k_j}$ be defined as in (E.3). Let G_{U_j} be as in (**??**). Let $E_{U_j, \bar{x}^{*j}}$ be an integral defined in (B.1).

Choose local set $U_1 \subseteq X_1$ so that for any subset $D_1 \subseteq U_1, |D_1| \geq \frac{\varepsilon}{2}|U_1|$, there is a permutation g so that $g \cdot S_1 \subseteq D_1$. Hence, $E_{U_1, \bar{x}^{*1}} \mathbf{1}_{D_1^k} > 0$, and let

$$\phi_1 := \inf_{D_1 \subseteq U_1, |D_1| \geq \frac{\varepsilon}{2}|U_1|,} E_{U_1, \bar{x}^{*1}} \mathbf{1}_{D_1^k}.$$

Choose local set $U_2 \subseteq X_2$ so that for any subset $D_2 \subseteq U_2, |D_1| \geq \frac{\varepsilon}{4}\phi |U_1|$, there is a permutation g so that $g \cdot S_2 \subseteq D_2$. Hence,

$$\phi_2 := \inf_{D_2 \subseteq U_2, |D_2| \geq \frac{\varepsilon}{4}\phi |U_2|,} E_{U_2, \bar{x}^{*2}} \mathbf{1}_{D_2^k} > 0.$$

Such local sets exist 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

$$D_1^A \subseteq A_{U_1}\left(\bar{x}^{*1}\right) \times U_2, \ D_{12}^A \subseteq A_{U_1}\left(\bar{x}^{*1}\right) \times A_{U_2}\left(\bar{x}^{*2}\right)$$

as follows:

$$\begin{aligned} D_1^A &= \left\{\left(\bar{x}^1, x^2\right) : \left(x_l^1, x^2\right) \in D \text{ for any } l \leq k_1\right\} \text{ and} \\ D_{12}^A &= \left\{\left(x_1^1, ..., x_{k_1}^1, x_1^2, ..., x_{k_2}^2\right) : \text{for any } l \leq k_2, \left(x_1^1, ..., x_{k_1}^1, x_l^2\right) \in D_1^A\right\} \\ &= \left\{\left(x_1^1, ..., x_{k_1}^1, x_1^2, ..., x_{k_2}^2\right) : \text{for any } l_1 \leq k_1, l_2 \leq k_2, \left(x_{l_1}^1, x_{l_2}^2\right) \in D\right\}. \end{aligned}$$

Suppose that $|D| \geq \varepsilon |U_1| |U_2|$. I will show that D_{12}^A is not empty. Since D_{12}^A is contained in $A_{U_1}(\bar{x}^{*1}) \times A_{U_2}(\bar{x}^{*2})$, 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 $x_2 \in U_2$, define

$$\alpha_2(x_2) = \frac{\left|\{x_1 : (x_1, x_2) \in D\}\right|}{|U_1|}.$$

Then,

$$\begin{aligned} |D| &\leq \left\{x_2 : \alpha_2(x_2) \geq \frac{\varepsilon}{2}\right\} |U_1| + \frac{\varepsilon}{2} |U_1| |U_2|, \text{ and} \\ \frac{|D|}{|U_1| |U_2|} - \frac{\varepsilon}{2} &\leq \frac{\left\{x_2 : \alpha_2(x_2) \geq \frac{\varepsilon}{2}\right\}}{|U_2|}. \end{aligned}$$

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

$$\alpha_1(\bar{x}_1) = \frac{\left|\left\{x_2 : (\bar{x}^1, x^2) \in D_1^A\right\}\right|}{|U_2|}.$$

Then,

$$\begin{aligned} \left|D_1^A\right| &\leq \left\{\bar{x}_1 : \alpha_1(x_1) \geq \frac{\varepsilon}{4}\phi_1\right\} |U_2| + \frac{\varepsilon}{4}\phi_1 \left|A_{U_1}\left(\bar{x}^{*1}\right)\right| |U_2|, \text{ and} \\ \frac{\left|D_1^A\right|}{|U_2| \left|A_{U_1}(\bar{x}^{*1})\right|} - \frac{\varepsilon}{4}\phi_1 &\leq \frac{\left\{\bar{x}_1 : \alpha_1(\bar{x}_1) \geq \frac{\varepsilon}{4}\phi_1\right\}}{\left|A_{U_1}(\bar{x}^{*1})\right|}. \end{aligned}$$

Observe that

$$\begin{aligned}\left|D_1^A\right| &= \sum_{x_2\in U_2}\left|\left\{\left(\bar{x}^1,x^2\right): \text{for any } l\leq k_1, \left(x_l^1,x^2\right)\in D\right\}\right| \\ &\geq \sum_{x_2\in U_2,\alpha_2(x_2)\geq\frac{\varepsilon}{2}}\left|\left|\left\{\bar{x}^1: x_l^1\in U_1\cap\left\{x_1:(x_1,x_2)\in D\right\} \text{ for any } l\leq k_1\right\}\right|\right| \\ &\geq \sum_{x_2\in U_2,\alpha_2(x_2)\geq\frac{\varepsilon}{2}}\phi_1\left|A_{U^1}\left(\bar{x}^{*1}\right)\right| \\ &\geq \left(\frac{|D|}{|U_1|\,|U_2|}-\frac{\varepsilon}{2}\right)\phi_1\left|A_{U^1}\left(\bar{x}^{*1}\right)\right||U_2|\end{aligned}$$

Similarly,

$$\begin{aligned}|D_{12}| &= \sum_{\bar{x}^1\in A_{U_1}(\bar{x}^{*1})}\left|\left\{\left(\bar{x}^1,\bar{x}^2\right): \text{for any } l\leq k_2, \left(\bar{x}^1,x_l^2\right)\in D_1^A\right\}\right| \\ &\geq \sum_{\bar{x}^1\in A_{U_1}(\bar{x}^{*1}),\alpha_1(\bar{x}^1)\geq\frac{\varepsilon}{4}\phi_1}\left|\left\{\bar{x}^2: x_l^2\in U_2\cap\left\{x_1:(\bar{x}_1,x_2)\in D_1^A\right\} \text{ for any } l\leq k_1\right\}\right| \\ &\geq \sum_{\bar{x}^1\in A_{U_1}(\bar{x}^{*1}),\alpha_1(\bar{x}^1)\geq\frac{\varepsilon}{4}\phi_1}\phi_2\left|A_{U^2}\left(\bar{x}^{*2}\right)\right| \\ &\geq \left(\frac{\left|D_1^A\right|}{|U_2|\,\left|A_{U_1}\left(\bar{x}^{*1}\right)\right|}-\frac{\varepsilon}{4}\phi_1\right)\phi_2\left|A_{U^2}\left(\bar{x}^{*2}\right)\right|\left|A_{U_1}\left(\bar{x}^{*1}\right)\right|\end{aligned}$$

By the above,

$$\begin{aligned}\frac{\left|D_1^A\right|}{|U_2|\,\left|A_{U_1}\left(\bar{x}^{*1}\right)\right|} &\geq \frac{1}{|U_2|\,\left|A_{U_1}\left(\bar{x}^{*1}\right)\right|}\left(\frac{|D|}{|U_1|\,|U_2|}-\frac{\varepsilon}{2}\right)\phi_1\left|A_{U^1}\left(\bar{x}^{*1}\right)\right||U_2| \\ &\geq \left(\frac{|D|}{|U_1|\,|U_2|}-\frac{\varepsilon}{2}\right)\phi_1\geq\left(\varepsilon-\frac{\varepsilon}{2}\right)\phi_1=\frac{\varepsilon}{2}\phi_1>\frac{\varepsilon}{4}\phi_1.\end{aligned}$$

Thus,

$$|D_{12}|\geq\frac{\varepsilon}{4}\phi_1\phi_2\left|A_{U^2}\left(\bar{x}^{*2}\right)\right|\left|A_{U_1}\left(\bar{x}^{*1}\right)\right|>0$$

and D_{12} is not empty.

F.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 $A_j\subseteq X_j$, there is generic $S_j\subseteq A_j$, $|S_j|\geq\delta_j|A_j|$. Take any finite $A\subseteq X_1\times X_2$. I show that it contains a generic subset with at least $\delta_1\delta_2|A|$ elements.

Indeed, there are local sets $U_j\in X_j$ such that $A\subseteq U_1\times U_2$. By the above, $U_1\times U_2$ is local, and group action $G_{U_1\times U_2}\longmapsto U_1\times U_2$ 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 the first part of Lemma 8, there is a permutation $g \in G_{U_1 \times U_2}$ such that

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

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.

F.2. **Proof of Proposition 5.** The proof of genericity of $X\left(B^1, ..., B^d\right)$ relates properties of language *Ordered Graph*$_d$ to language *Product*$_d$ from Example 4. Let $X^* = B^d$ be a product of d-copies of B. Then, $X \subseteq X^*$.

Fix $m_1, ..., m_l \in \mathbf{N}$. For any $\varepsilon > 0$, there is a finite set $U_\varepsilon \subseteq B$, $|U_\varepsilon| \geq \frac{2d^2}{\varepsilon}$, $l \leq d$ such that for any $D \subseteq U_\varepsilon^d$, $|D| \geq \frac{\varepsilon}{2} \left|U_\varepsilon^d\right|$, there are subsets $C^l \subseteq B$, $\left|C^l\right| = m_l$, $l \leq d$ such that

$$C_1 \times .... \times C_l \subseteq D.$$

Indeed, this is simply a restatement of the fact that any finite set is generic under *Product*$_d$ group action (Corollary 2) together with an observation that the local set from the definition of genericity can be taken of any large size.

For any finite set $U \subseteq B$, recall the set $X(U) \subseteq X$ of all ordered tuples of distinct elements of B' in (2.15). This is a local set under group action *Ordered Graph*$_d$. Moreover,

$$X(U) = U^d \cap X,$$

where U^d is a product of d copies of U. Also, notice that

$$\left|U^d \cap X\right| \geq |U \times ... \times U| - d \sum_{l \leq d} \left|U^{d-1}\right|.$$

In particular, if $U_\varepsilon \subseteq B$ is as above, then $|U_\varepsilon| \geq \frac{2d^2}{\varepsilon}$, and

$$|X(U)| \geq \left(1 - \frac{d^2}{\frac{2d^2}{\varepsilon}}\right) \left|U^d\right| = \left(1 - \frac{\varepsilon}{2}\right) \left|U^d\right|.$$

Finally, the above remarks can be put together. Let $B^1, ...B^d \subseteq B$ be mutually disjoint finite sets and let $m^l = \left|B^l\right|$, $l \leq d$. Find set $U_\varepsilon \subseteq B$ with the properties described above. Take any subset $D \subseteq X(U_\varepsilon)$ such that $|D| \geq \varepsilon |X(U_\varepsilon)|$. Then,

$$|D| \geq \varepsilon |X(U_\varepsilon)| \geq \varepsilon \left(1 - \frac{\varepsilon}{2}\right) \left|U^d\right|,$$

and there are sets $C^1, ..., C^d$, $\left|C^l\right| = m^l, l \leq d$ such that

$$C^1 \times ... \times C^d \subseteq D.$$

Since $D \subseteq X$, all tuples from $C^1 \times ... \times C^d$ must consist of distinct elements of B. This implies that sets $C^1, ..., C^d$ are mutually disjoint. Moreover,

$$X\left(C^1, ..., C^d\right) = C^1 \times ... \times C^d \cap X \subseteq D$$

and sets $X\left(C^1, ..., C^d\right)$ and $X\left(B^1, ..., B^d\right)$ are permutations of each other under group action *Ordered Graph*$_d$. (Indeed, one can find a permutation $g : B \to B$ such that for any $l \leq d$, $g \cdot C^l = B^l$.) Because $X(U_\varepsilon)$ is local, this implies that $X\left(B^1, ..., B^d\right)$ is generic.

To show that language *Ordered Graph*$_d$ is tight, take any finite $A \subseteq X$ and find finite $U \subseteq B$ such that $A \subseteq X(U)$. Let $m = |U|$ and let $m^1 + ... + m^d = m$ be a finite sequence of natural numbers such that $m^l \in \left\{\left\lceil \frac{m}{d} \right\rceil - 1, \left\lceil \frac{m}{d} \right\rceil\right\}$ for each l. Find mutually disjoint subsets $B^l \subseteq U$ so that $\left|B^l\right| = m_l$. Then,

$$\frac{\left|X\left(B^1, ..., B^d\right)\right|}{|X(U)|} \geq \frac{m_1}{m} ... \frac{m_d}{m} \approx \left(\frac{1}{d}\right)^d.$$

By the first part of Lemma 8, there is a permutation g such that $g \cdot X(U) = X(U)$ and

$$\left|A \cap X\left(B^1, ..., B^d\right)\right| \geq \frac{\left|X\left(B^1, ..., B^d\right)\right|}{|X(U)|} |A| \geq \left(\frac{1}{d}\right)^d.$$

This shows that *Ordered Graph*$_d$ is tight and $t(G) \geq \left(\frac{1}{d}\right)^d$.

F.3. **Proof of Lemma 6.** Notice that $G' \longmapsto X$' is transitive. Indeed, $G \longmapsto X$ is transitive; thus, for any $x'_1, x'_2 \in X$, any $x_j \in p_X^{-1}\left(x'_j\right)$, $j = 1, 2$, there is a permutation $g \in G$, such that $g \cdot x_1 = x_2$. By (5.1), $p_G(g) \cdot x'_1 = x'_2$.

Let $U \subseteq X$ be local under $G \longmapsto X$. Take any subsets $A'_1, A'_2 \subseteq p_X(U)$ such that there is a permutation $g' \in G'$ so that $g' \cdot A'_1 = A'_2$ Let $A_j = p_X^{-1}\left(A'_j\right) \cap U, j = 1, 2$. Take any permutation $g \in p_G^{-1}(g')$. By (5.1), $g \cdot A_1 = A_2$. Because U is local, there is a permutation g_U such that $g_U \cdot A_1 = A_2$ and $g_U \cdot U = U$. By (5.1), $p_G(g_U) \cdot p_X(U) = p_X(U)$ and $p_G(g_U) \cdot A'_1 = A'_2$. This implies that $p_X(U)$ is local.

Let $U \subseteq X$ be a local set. Because $G_U \longmapsto U$ is transitive, for any $x'_1, x'_2 \in p_X(U)$, there is a permutation $g_U \in G_U$ such that $p_G(g_U) \cdot x'_1 = x'_2$. By (5.1),

$$p_X\left(g_U \cdot \left(p_X^{-1}(x'_1) \cap U\right)\right) \subseteq p_X^{-1}(x'_2) \cap U.$$

Because the above is true for any $x'_1, x'_2 \in p_X(U)$, the inclusion can be replaced by equality. This implies that for any subset $D' \subseteq p_X(U)$,

$$\frac{|D'|}{|p_X(U')|} = \frac{\left|p_X^{-1}(D) \cap U\right|}{|U|}.$$

Suppose that S is generic under $G \longmapsto X$. Then, for each $\delta > 0$, there is a local set U such that for each $D \subseteq U, |D| \geq \delta |U|$, there is a permutation g so that $g \cdot S \subseteq D$. Take any subset $D' \subseteq p_X(U)$ such that $|D'| \geq \delta |p_X(U)|$. By the above, $\left|p_X^{-1}(D') \cap U\right| \geq \delta |U|$. Hence, there is a permutation g so that $g \cdot S \subseteq p_X^{-1}(D') \cap U$. By (5.1),

$$p_G(g) \cdot p_X(S) \subseteq D'.$$

This shows that $p_X(S)$ is generic.

University of Chicago, Department of Economics
E-mail address: mpeski@uchicago.edu