LEARNING THROUGH THEORIES

MARCIN PĘSKI*

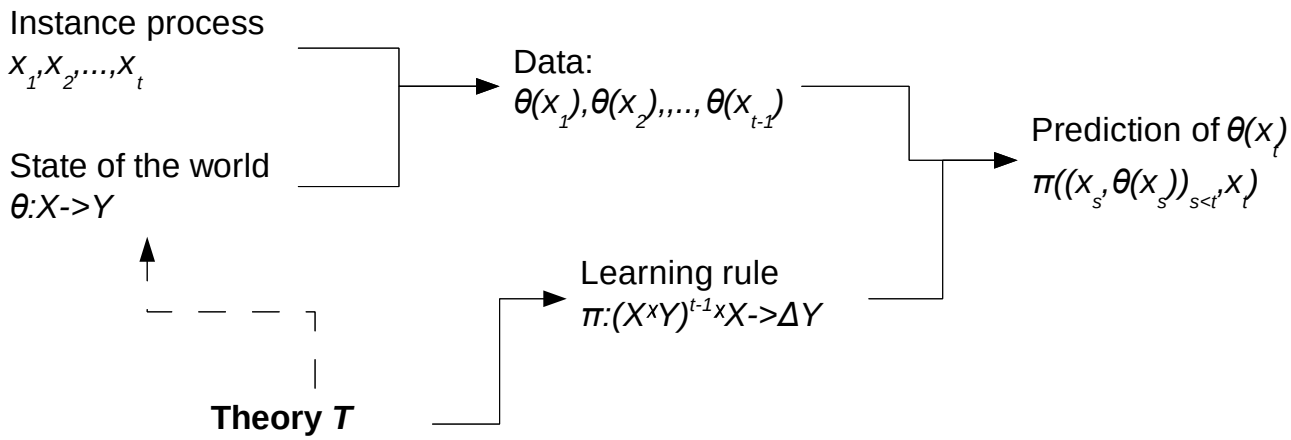
ABSTRACT. This paper examines the relationship between language and knowledge in a learning model. An agent describes the world through theories. A theory consists of universal propositions called patterns, and it is formulated in some language. We look at two characteristics of a good theory. A theory is *informative* if it ensures correct predictions. A theory is *brief* if it consists of one pattern. We offer different characterizations of informative theories. In particular, we identify 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. The results are illustrated with specific problems of reasoning under uncertainty.

1. INTRODUCTION

Both in science and in life, information is often represented in the form of theories stated in some language. Typically, theories serve two goals. First, theories are used to make predictions. For example, the theory of a consumer's choices usually contains a transitivity assumption: for all bundles, if the consumer chooses a from any bundle that includes b and b from any bundle that includes c, she must choose a from any bundle that includes a and c. Such a theory allows to deduce the choice of a from the last bundle given the observed choices in the first two bundles.

Second, theories are used to communicate information. Instead of describing the consumer's choices in all possible bundles separately, the theory of transitivity consists of one statement that applies universally to all bundles. A communicable theory must be stated briefly. Consider a teacher who wants to pass on knowledge to a student. The teacher does not have the time or other resources to discuss all the situations that the student may face in her life. Instead, the teacher will teach a general theory together with a few well-chosen examples. The student should be able to use the examples and the theory to deduce the outcomes of situations that the teacher did not mention. Because the quality of predictions may

*University of Texas at Austin, Department of Economics. E-mail: mpeski@gmail.com.edu. I am grateful to Jeroen Swinkels, Larry Samuelson, Balazs Szentes, Gil Kalai, Wojciech Olszewski, Todd Sarver, Mihai Manea, and seminar participants at Caltech, UCLA, UT Austin, RUD' 07, Yale, and Washington U. at St Louis for discussions. I am responsible for all remaining errors.



require the teacher to communicate long and complicated theories, there exists a trade-off between the two goals of the theory.

In this paper, we analyze the two goals of the theories using a learning model. In the model, a theory provides an information about the association between instances (questions, decision problems) and outcomes (answers, solutions). Let X be a space of instances. Each instance x is associated with an outcome $\theta(x)$. The association $\theta : X \to Y$, called the *state of the world,* is chosen by Nature, and an agent knows only that θ is consistent with some theory T. We assume that theory T must be expressed in some language. We formally define language as a system of symbols (relations, variables, Boolean operators, and quantifiers) and rules of combining finite tuples of symbols into formulas. Language is used to express universal formulas called *patterns.* A *theory* is any (finite or infinite) set of patterns.

In each period, the agent observes a new instance x_t makes a prediction of the outcome $\theta(x_t)$ and observes the realization of the true outcome. The prediction is formally defined as a learning rule and it is based on the entire database of past observations before period t. The choice of the learning rule is influenced by the agent's knowledge of theory T. See Figure 1.

The agent's payoff is equal to the long-run average number of periods in which his prediction was correct. The payoff depends on the instance process, the state of the world, and the learning rule. We say that a theory is *informative*, if for *some* instance process, there exists a learning rule, such that for any state θ that is consistent with the theory, the long-run average number of mistakes converges to 0. Such theories are very valuable: as long as the agent can appropriately choose the order of observations, such a theory guarantees that in the long-run the agent commits arbitrarily few mistakes.

We present different characterizations of informative theories. We show that, in a class of languages all informative (possibly infinite) theories imply theories that are simultaneously

informative and *brief*, i.e., that consist of only one pattern (Theorem 1). Additionally, we show that informative theories are separated from non-informative ones in the following sense. For each language from the class we study, there exists a constant $h > 0$ such that for each non-informative theory, the worst-case scenario average number of mistakes is equal to at least h. Constant h does not depend on the theory, only on the language.

As an illustration, we show that there exists a class of languages with important applications in which all non-trivial theories are informative. This is surprising, as the definition of informativeness is quite demanding. The language of the consumer's choices does not belong to that class. Although the (one-pattern) theory of transitivity is informative, the language allows to state non-trivial and non-informative theories.

The definition of informativeness is relatively weak: for a theory to be informative, it is enough that there exists (possibly theory-specific) an instance process and (possibly process-specific) a learning rule. Because we start with a weak definition, we are certain not to omit interesting theories. In order to test whether the definition is not too permissive, we discuss its modifications:

(1) We consider theories that lead to no mistakes in the long-run for *any* instance process. However, we argue that such a requirement is too strong as in many languages, all theories that satisfy the stronger requirement are essentially constant (see Section 4.2).
(2) One may ask whether the instance process can be chosen independently from the theory and the learning rule. Say that a class of processes satisfies the *sufficient data condition*, if for *any* informative theory, there exists a learning rule that ensures no long-run mistakes along *any* process from the class. In particular, the learning rule does not depend on the instance process, and the class of instance processes does not depend on the theory. We characterize conditions under which such classes exist (Theorem 2) and we show that, in many cases, they correspond to natural and easy-to-interpret requirements.
(3) Finally, we show that there exists a learning rule that ensures no long-run mistakes for *any* informative theory and *any* instance process from a class that satisfies the sufficient data condition (Theorem 3). We interpret the result as a possibility of induction. The proof relies on our characterization of informative theories as those that imply informative patterns.

The next two sections define language and the learning model. Section 4 illustrates the definitions and the results of the paper on examples. Section 5 characterizes theories that are informative. Section 6 proves the existence of the sufficient data condition. Section 7

discusses the existence of universal learning rule that applies to all informative theories. Section 8 discusses the tightness of various languages. We postpone the review of the literature until the last section.

2. Language

Let X be a countably infinite space of instances (decision problems, objects). For each finite tuple of elements of X, $\bar{x} = (x_1, ..., x_m)$, and each subset $S \subseteq X$, we write $\bar{x} \subseteq S$ if $\{x_1, ..., x_m\} \subseteq S$. Let $\bar{x}\hat{}\bar{x}'$ denote the concatenation finite tuples $\bar{x}$ and $\bar{x}'$.

A (relational) *language* on X is a set of relations $\mathcal{L} = \{R_i\}$[1], where $R_i \subseteq X^{k_R}$ is a k_{R_i}-ary relation on X. Language $\mathcal{L}$ is *finite*, if $|\mathcal{L}| < \infty$. We implicitly assume that each language includes the binary relation of equality '='.

Let Y be a finite space of outcomes (solutions, properties). A mapping $\theta : X \to Y$ is called a *state of the world.* Let $\mathcal{M} = Y^X$ be the space of all states equipped with the product σ-algebra.

Language $\mathcal{L}$ is used to express information about states of the world. An atomic formula is a statement of one of the following forms

$$R(\tilde{x}_1, \tilde{x}_2..., \tilde{x}_{k_R}), \tilde{\theta}(\tilde{x}) = 0, \text{ or } \tilde{\theta}(\tilde{x}) = 1,$$

where $\tilde{x}_1, , \tilde{x}_2..., \tilde{x}_k, \tilde{x}$ are variables corresponding to instances, $R \in L$ is a relation, and $\tilde{\theta}$ is variable corresponding to state. A *(free) formula* F is any combination of atomic formulas connected by Boolean operators $\vee, \wedge, \Longrightarrow, \neg$. We write $F(\tilde{x}_1, , \tilde{x}_2..., \tilde{x}_k; \theta)$ when we want to indicate all (finitely many) variables used in formula F.

For any k-tuple of instances $x_1, ..., x_k \in X$ and any state θ, let $F(x_1, ..., x_k; \theta)$ be the statement obtained from the substitution of the respective variables; such statement can be assigned with the truth value.[2] State θ *is consistent with* F if $F(x_1, ..., x_k; \theta)$ is true for any substitution of instances into variables: We write

$$\forall_{x_1, x_2..., x_k \in X} F(x_1, x_2..., x_k, \theta), \tag{2.1}$$

and we refer to (2.1) as a *(universal) pattern* F.

[1]Mathematical logic defines (first-order) language as a set of symbols of relations, functions, operators, quantifiers, etc., and rules of constructing formulas (grammar). Except for Boolean operators and quantifiers, symbols do not have any meaning. The meaning is acquired only when a formula from the language is interpreted in a particular model. In this paper, we want to compare the power of expression of concrete relational systems. For this purpose, we assume that relations $R \in \mathcal{L}$ are already defined on the space of instances X. This is a shortcut; instead, with a little bit more notation, one could follow the standard route to obtain the same interpretation of the results.

[2]If $F(.)$ is one of the atomic formulas, the definition of the truth value is immediate. For general F, the truth value is defined by induction with respect to smaller formulas.

A *theory* T is any set of patterns. Let $\mathcal{M}(T) \subseteq \mathcal{M}$ denote the set of all states θ that are consistent with all patterns in theory T. For each $A \subseteq X$, let

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

denote the set of all A-restrictions of states that are consistent with theory T. Sets $\mathcal{M}(T)$ and $\mathcal{M}(T;A))$ describe the uncertainty of the agent who knows that theory T is true. Clearly, the more states that are consistent with theory T, the larger $\mathcal{M}(T;A)$. Say that theory T *implies* theory S if $\mathcal{M}(T) \subseteq \mathcal{M}(S)$.

The above model of theory-based information is entirely standard. To see it, notice that one can think about $\mathcal{M}$ as the state space and about $\{\mathcal{M}(F), \mathcal{M}\backslash\mathcal{M}(F)\}$ as binary partitions on $\mathcal{M}$ induced by patterns F. There are two reasons for looking at languages rather than partitions. First, language is a less abstract and a more direct way of describing information. In addition, language comes with an added structure (relations, relationships between relations, and so on) that can be used in proofs and to draw comparisons.

In the rest of this section, we describe some natural properties of languages. For any pair of tuples of equal length $\bar{x}, \bar{x}' \in X^m$, we say that tuples $\bar{x}$ and $\bar{x}'$ are *analogous*, write $\bar{x} \sim \bar{x}'$, if for each k-ary relation $R \in \mathcal{L}$, for each $i_1, ..., i_k \leq m$,

$$R(x_{i_1}, ..., x_{i_k}) \Leftrightarrow R\left(x_{i'_1}, ..., x'_{i_k}\right).$$

In other words, two tuples are analogous, if they have the same description in language $\mathcal{L}$. Say that two finite sets $S, S' \subseteq X$ are analogous, $S \sim S'$, if there are enumerations $\bar{x} = (x_1, ..., x_m)$ of S and $\bar{x}' = (x'_1, ..., x'_m)$ that are analogous, $\bar{x} \sim \bar{x}'$. Also, write $\bar{x}\tilde{\subseteq}D$ (or, $S\tilde{\subseteq}D$) if there is $\bar{x}' \subseteq D$ (or, $S' \subseteq D$) such that $\bar{x}$ and $\bar{x}'$ (or, S and S') are analogous.

Language $\mathcal{L}$ is *complete* if each pair of analogous tuples $\bar{x} \sim \bar{x}'$ has analogous extensions: for each $x \in X$, there exists x' such that $\bar{x}\hat{}x \sim \bar{x}'\hat{}x'$. Thus, in complete languages, two tuples with the same relational description have the same relations with elements outside the tuples.

Language $\mathcal{L}$ is *transitive* if any two instances (1-tuples) are analogous, i.e., when the language $\mathcal{L}$ does not have any non-trivial unary relations. That means that all instances look alike a priori. All the examples of languages that are discussed in this paper are complete and all but full language from Section 4.5 are transitive.

It turns out that if the language is complete, the restrictions over the finite set depend in some sense only on the analogy class of the set. In Appendix A.1, we prove the following fundamental property.

Lemma 1. *Suppose that language $\mathcal{L}$ is complete. Then, for any theory T, any two analogous sets A and A', $|\mathcal{M}(T,A)| = |\mathcal{M}(T,A')|$.*

3. Learning

A *(deterministic) instance process* is any sequence of instances $\bar{x} = (x_1, x_2, ...)$ such that no instance is ever repeated $x_s \neq x_t$ for $s \neq t$. In each period t, the agent observes instance x_t and predicts the value of outcome $\theta(x_t)$. Because each instance is observed only once, each prediction problem is novel and non-trivial. After the prediction is made, the agent is informed about the true outcome $\theta(x_t)$. The predictions may depend on past observations from periods $s < t$. A *learning rule* $\pi : \bigcup_t (X \times Y)^{t-1} \times X \to \Delta Y$ chooses (possibly randomized) predictions after each history. The agent receives payoff 1 if her prediction is correct and 0 if not. (More general payoffs do not change any of the results.) For any instance process $\bar{x}$, any learning rule π, any state θ, let

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

This is the t-period average payoff from learning rule π given instance process $\bar{x}$ and state θ. The average payoff is never larger than 1.

Consider an agent who knows that a certain deterministic theory T holds, but does not know which state $\theta \in \mathcal{M}(T)$ is the true one. The agent chooses learning rule π in order to maximize the worst-case payoff,[3]

$$\inf_{\theta \in M(T)} U_t(\pi, \bar{x}, \theta). \tag{3.1}$$

Definition 1. *Theory T is* informative *if there is an instance process $\bar{x}$ and a learning rule π, such that*

$$\lim \inf_{t \to \infty} \inf_{\theta \in \mathrm{M}(T)} U_t(\pi, \bar{x}, \theta) = 1. \tag{3.2}$$

Definition 1 allows us to compare theories with respect to the quality of their predictions: For each informative theory, there exists a learning rule and an instance process that ensure that, in the long-run, almost all predictions are correct.

[3]The worst-case criterium has a long tradition in statistical literature. This criterium is directly used in the classic decision theory of Wald (1950), Wald (1949) and Blackwell and Girshick (1954) (see Schwarz (1994)); slightly reformulated, it is used by the learning theory of Vapnik and Chervonenkis (1971) (see Vapnik (1998) and Bousquet, Boucheron, and Lugosi (2004)). Alternatively, the worst-case criterium can be interpreted as an extreme uncertainty aversion of the decision maker Gilboa and Schmeidler (1989). 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. The robust control literature (see Hansen and Sargent (forthcoming)) uses a discounted version of formula (3.2) both in a theoretical and an empirical framework.

The definition of informativeness is relatively weak: for a theory to be informative, it is enough that there exists (possibly theory-specific) an instance process and (possibly process-specific) a learning rule. Because we start with a weak definition, we are certain not to omit theories that may reasonably lead to almost perfect predictions. On the other hand, the definition may be too weak. To fully examine its scope, we discuss three modifications:

(1) We consider theories that lead to no mistakes in the long-run for *all* instance process. Formally, say that theory T is *strongly informative*, if there is a learning rule π such that (3.2) holds for *all* instance processes x. We believe that such a requirement is too strong: in many languages, all theories that satisfy the stronger requirement are essentially constant. An example is discussed in Section 4.2 (see Lemma 2).
(2) We ask whether the instance process can be chosen independently from the theory and the learning rule. A class $\bar{X}$ instance processes satisfies *sufficient data condition*, if for any theory T that are informative, there is a learning rule π such that for all processes $\bar{x}$, (3.2) holds. If there exists such a (non-empty) class of processes $\bar{X}$, then we can switch the order of quantifiers in Definition 1: a theory is informative if and only if there exists a learning rule such that (3.2) holds for any instance process $\bar{x} \in \bar{X}$. In particular, the learning rule depends only on the theory, not on a specific instance process from the class. We show that such classes exist (Theorem 2).
(3) Say that π is an *universal learning rule* if (3.2) holds for each informative theory T, each process $\bar{x}$ from the sufficient data condition class $\bar{X}$. In Section 7, we show that an universal learning rule exists.

We assume that the agent 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 informativeness is a zero–one notion, and it divides theories into two categories: those that predict well with few mistakes and those that do not. A discounted model could allow for a refinement of the first category into theories that allow for correct predictions early and those that lead initially to many mistakes. However, at this moment, it is unclear what exactly will happen in the discounted model. This should be a subject of future research.

We assume that the instance process is a deterministic sequence. The model can be easily expanded to incorporate *probabilistic instance processes*, i.e., distributions σ over deterministic instance processes. For each probabilistic process σ, define $U_t(\pi, \sigma, \theta) = E_\omega U_t(\pi, \bar{x}, \omega)$ for probabilistic instance processes. Say that a class of probabilistic instance processes satisfies the sufficient data condition if for any theory T that is informative, there is a learning rule π such that (3.2) with σ-probability 1 for all probabilistic processes σ in the class. Thus, if class

$\bar{X}$ of deterministic processes satisfies the sufficient data condition, then $\left\{\sigma : \sigma\left(\bar{X}\right) = 1\right\}$ also satisfies the sufficient data condition.

4. Examples

In this section, we discuss examples. All examples but the last one satisfy the assumptions of the main results of the paper.

4.1. **Trivial language.** Let $\mathcal{L}_{\text{trivial}}=\varnothing$ be a language without any non-trivial relation. This is a very limited language that nevertheless allows a range of patterns to be stated. For example,

$$\forall_x \theta(x) = 0$$

is a pattern that is satisfied by only one state, $\theta \equiv 0$. On the other hand, pattern

$$\forall_{x,x'} x \neq x' \wedge \theta(x) = 1 \Longrightarrow \theta(x') = 0,$$

is satisfied by infinitely many states, in all of which, there is at most one outcome equal to 1 and all the remaining are equal to 0. We show that any non-trivial theory is informative and the set of all instance processes satisfies the sufficient data condition.

4.2. **Netflix.** Netflix is a DVD rental online company. Its customers typically do not know whether they will like a movie before they watch it. To help them, Netflix makes recommendations. The recommendations are based on rankings given by the customers in the past. To make recommendations, Netflix needs to know some universal properties of the distribution of preferences. For example, Netflix might know that:

$$\begin{gathered}\text{For any customers } c, c', \text{ any movies } m, m', \\ \text{if } c \text{ likes } m \text{ but dislikes } m', \text{ and } c' \text{ dislikes } m, \\ \text{then } c' \text{ also dislikes } m'.\end{gathered} \tag{4.1}$$

Statement (4.1) describes a situation in which there is an universal ordering of movies with respect to their quality: for each two movies m and m' , the set of customers that like m is either contained or it contains the set of customers that like m'. A similar ordering exists on the set of customers. Because statement (4.1) does not refer to any individual customer or movie, it is an example of a pattern.

We formally describe language in which pattern (4.1) can be stated. Let C be an infinite set of customers, M be an infinite set of movies, and let $X = C \times M$ be the set of instances. Each instance (c, m) is associated with an outcome $\theta(c, m) \in \{0, 1\}$, where $\theta(c, m) = 1$ if and only if customer c likes movie m.

There are two natural relations R_C and R_M on set X : for any $(c,m),(c',m') \in X$,

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

Two instances are in relation R_C if they refer to the same customer; two instances are in relation R_M if they refer to the same movie. Let $\mathcal{L}_{\text{Netflix}} = \{R_C, R_M\}$.

Any information about customers and movies that does not refer to any specific customer and movie can be expressed in language $\mathcal{L}_{\text{Netflix}}$. As an example, we show how to state (4.1). Let R be a free formula with four variables:

$$R(\tilde{x}_1,\tilde{x}_2,\tilde{x}_3,\tilde{x}_4) := \tilde{x}_1 R_C \tilde{x}_2 \wedge \tilde{x}_3 R_C \tilde{x}_4 \wedge \tilde{x}_1 R_M \tilde{x}_3 \wedge \tilde{x}_2 R_M \tilde{x}_4 \wedge \neg(\tilde{x}_1=\tilde{x}_2) \wedge \neg(x_1=x_3).$$

For example, $R(\bar{x})$ and $R(\bar{w})$ but $\neg R(\bar{z})$ for tuples $\bar{x},\bar{w}$, and $\bar{z}$ from Figure 2. (Notice that tuples $\bar{x}$ and $\bar{w}$ are analogous, but they are not analogous to $\bar{z}$.)

m	x_1		x_3
n'			
m'	x_2		x_4
n			
	c	d	c'

,

m			
n'		w_4	w_2
m'			
n		w_4	w_1
	c	d	c'

,

m	z_1		
n'		z_2	
m'			z_4
n			z_3
	c	d	c'

Figure 1. Tuples $\bar{x},\bar{w}$, and $\bar{z}$.

Then, (4.1) has the same meaning as:

$$\forall_{\bar{x}=(x_1,x_2,x_3,x_4)} \left(R(\bar{x}) \wedge \theta(x_1)=1 \wedge \theta(x_2)=\theta(x_3)=0\right) \Longrightarrow \theta(x_4)=0.$$

In other words, for each quadruple of instances $\bar{x}$ such that $R(\bar{x})$, if the outcome of the first instance is equal to 1and the outcomes of the second and the third instances are equal to 0, then the outcome of the last instance must also be equal to 0.

We will show that any non-trivial theory in language $\mathcal{L}_{\text{Netflix}}$ is informative (Theorem 1 and Lemma 6). To see why it might be so, consider one-pattern theory (4.1). Suppose that Netflix observes the preferences of customer c over a set of n_M movies and the preferences of the set of n_C customers over movie m for some n_C and n_M. See Figure 2. There are at most n_C+n_M mistakes committed while making these observations. Netflix uses the theory and initial observations to predict that customer c' will not like movie m'. In fact, Netflix can correctly predict all outcomes of customer-movie pairs denoted with '?'. The number of correct predictions is proportional to the product $n_C n_M$. For large n_C and n_M, the number

of correct predictions is of an order larger than the number of mistakes, $n_C n_M >> n_C + n_M$.

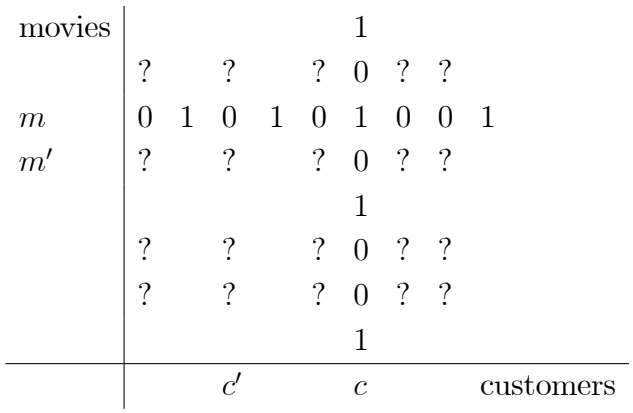


Figure 2.

The order in which observations are made has an impact on the quality of predictions. Not all instance processes lead to the same long-run payoffs. For example, consider an instance process in which no customer or movie is ever repeated. Then, pattern (4.1) will never be applied. Such an instance process does not belong to any class that satisfies the sufficient data condition.

For each instance process $\bar{x}$, each period t, define the numbers of distinct customers and movies observed until period t, $n_t^C = \#\{c_s : s \leq t\}$ and $n_t^M = \#\{m_s : s \leq t\}$. Consider a class of processes $\bar{X}_{\text{Netflix}}$ such that

$$\lim_{t\to\infty} n_t^C = \lim_{t\to\infty} n_t^M = \infty \text{ and } \sup_t \frac{n_t^C n_t^M}{t} < \infty \text{ almost surely.} \tag{4.2}$$

The first two conditions ensure that the number of distinct customers and movies is unbounded. The last condition implies that, for sufficiently high t, the average number of observations per customer is proportional to the number of observed movies and the average number of observations per movie is proportional to the number of observed customers. We show that for any informative theory, there exists a learning rule such that for any instance process $\bar{x} \in \bar{X}_{\text{Netflix}}$, (3.2) holds, i.e., $\bar{X}_{\text{Netflix}}$ satisfies the sufficient data condition.

Finally, we use the Netflix example to argue that many interesting theories are not strongly informative. For example, consider an instance process along which in each period a new movie is observed. Along such an instance process, pattern (4.1) can never be applied to make a correct prediction. In fact, the best long-run average number of correct predictions is $\frac{1}{2}$ and it can be achieved by randomized prediction that predicts 0 with probability $\frac{1}{2}$ in each period.

More generally, we show that in the language $\mathcal{L}_{\text{Netflix}}$ all strongly informative theories are essentially constant. The proof of the next result can be found in Appendix B.

Lemma 2. *Suppose that T is a theory in language $\mathcal{L}_{Netflix}$ that is strongly informative. Then, there exists k such that for each state $\theta \in \mathcal{M}(T)$, there is $y^* \in \{0,1\}$ and sets $C_0 \subseteq C, M_0 \subseteq M$ such that $|C_0|, |M_0| \leq k$, and $\theta(c,m) = y^*$ for ell $c \notin C_0$ and $m \in M_0$.*

A similar result holds for any other language discussed in this section.

4.3. **Consumer's choices.** Let X be a collection of choice sets, and let $\theta(x)$ denote the consumer's choice from choice set $x \in X$.[4] An econometrician wants to predict $\theta(x)$. He is guided by a belief that the consumer's choices are rational.

Let B be a countable set of goods. Let X be a set of ordered d-tuples of distinct elements of B. Let $Y = \{1, ..., d\}$. State $\theta : X \to Y$ assigns choices in tuples of goods: if $\theta(b_1, ..., b_d) = l$, then, faced with a choice between $b_1, ..., b_d$, the consumer chooses b_l. Define language $\mathcal{L}_{\text{choices}} = \{R_{ij}\}$ where $R_{i,j}$ is a binary relation such that for all $x, x' \in X$,

$$(b_1, ..., b_d)\, R_{ij}\, (b'_1, ..., b'_d) \text{ iff } b_i = b'_j.$$

If $d = 2$, then the rationality of the consumer's choices can be stated as two patterns:

$$\forall_{x,x'} x R_{12} x' \wedge x R_{21} x' \wedge \theta(x) = 1 \Longrightarrow \theta(x') = 2, \tag{4.3}$$

$$\forall_{x,x',x''} x R_{21} x' \wedge x R_{11} x'' \wedge x' R_{22} x'' \wedge \theta(x) = 1 \wedge \theta(x') = 1 \Longrightarrow \theta(x'') = 1. \tag{4.4}$$

Pattern (4.3) says that the consumer's choice does not depend on the order of goods in the tuple. Pattern (4.4) says that the consumer's choices are transitive: To see, notice that the pattern is satisfied if and only if there are distinct goods a, b, c such that $x = (a, b)$, $x' = (b, c)$, and $x'' = (a, c)$. Then, if the customer chooses a from bundle x, b from bundle x', then she must choose a from bundle x''.

Not all patterns are informative. For example, pattern (4.3) alone is not informative. On the other hand, a general characterization of informative patterns from Section 8.4 implies that (4.4) is informative.

For each instance process $\bar{x}$, each period t, define the number of distinct customers and movies observed until period t, $n_t = |\{x_{i,s} : i \in \{1, ..., d\}, s \leq t\}|$. We show that the class $\bar{X}_{\text{choices}}$ of processes such that

$$\sup_t \frac{(n_t)^d}{t} < \infty, \text{ almost surely,} \tag{4.5}$$

satisfies the sufficient data condition.

[4] I am grateful to Matias Iaryczower for suggesting this example.

4.4. **Chemist.** A chemist mixes different substances in a tube and observes the type of interaction in the mixture. Some substances interact by producing salty water in a stormy reaction. The chemist would like to predict the types of interactions between pairs of substances that she has not yet observed.

Let B be a set of substances and let $\{x \subseteq B : |x| = d\}$ be the collection of d-element subsets of B. Each instance is interpreted as a mixture of d substances and is assigned an outcome $\theta(x) \in \{0 \text{ ("no reaction")}, 1 \text{ ("reaction")}\}$. Language $\mathcal{L}_{\text{choices}} = \{R_i\}$ consists of binary relations R_i for $i \leq d$ such that for all $x, x' \in X$,

$$xR_ix' \text{ iff } |x \cap x'| = k.$$

For example, suppose that $d = 2$, and define free formula with three variables:

$$R(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) := \tilde{x}_1R_1\tilde{x}_2 \wedge \tilde{x}_1R_1\tilde{x}_2 \wedge \tilde{x}_1R_1\tilde{x}_2.$$

Then, $R(x_1, x_2, x_3)$ if and only if there are distinct substances $a, b, c \in B$ such that $x_1 = \{a, b\}, x_2 = \{a, c\}$, and $x_3 = \{b, c\}$. Consider patterns:

$$\forall_{x_1,x_2,x_3} R(x_1, x_2, x_3) \wedge \theta(x_1) = 1 \wedge \theta(x_2) = 1 \Longrightarrow \theta(x_3) = 0,$$
$$\forall_{x_1,x_2,x_3} R(x_1, x_2, x_3) \wedge \theta(x_1) = 0 \wedge \theta(x_2) = 0 \Longrightarrow \theta(x_3) = 0.$$

The first pattern says that if a and b react, a and c react, then b and c don't react, and the second that if a and b don't react, a and c don't react, then b and c don't react. These two patterns jointly give rise to the acid-alkaline theory of substances: Each substance is an acid or alkaline, and two substances react if and only if they are of different types.

We show that the acid-alkaline theory is informative. Also, let $n_t = \left|\bigcup_{s \leq t} x_s\right|$. Then, the set of all instance processes such that (4.5) holds satisfies the sufficient data condition.

4.5. **Full language.** Consider a language $\mathcal{L}_X = \{R_x\}$, where, for each $x \in X$, R_x is a unary relation such that $R_x(x')$ iff $x = x'$.

We show that the main result of this paper fails for such language ($\mathcal{L}_X$ does not satisfy the assumptions of Theorem 1 below). Fix a state $\theta^* : X \to \{0, 1\}$ and consider a theory $T_{\theta^*} = \{\forall_x R_{x^*}x \Longrightarrow \theta(x) = \theta(x^*) : x^* \in X\}$. θ^* is the only state that omits all patterns in theory T_{θ^*}, $\mathcal{M}(T_{\theta^*}) = \{\theta^*\}$. Thus, theory T_{θ^*} is informative. Theory T_{θ^*} consists of infinitely many patterns. Unless θ^* is constant outside a finite set, theory T_{θ^*} does not imply any finite set of patterns that is informative.

5. Characterization of informative theories

In this section, we present three characterization of informative theories. All the proofs of the above results can be found in Appendix C.

5.1. **Entropy.** Let

$$\mathcal{E}(T;A) = \frac{1}{|A|}\log|\mathcal{M}(T;A)| \text{ for each finite } A \subseteq X, \tag{5.1}$$
$$\mathcal{E}(T) = \inf_{A\subseteq X, A \text{ is finite}} \mathcal{E}(T;A).$$

(In this paper, the logarithm always has base 2.) We refer to $\mathcal{E}(T)$ as the *entropy* of theory T and $\mathcal{E}(T;A)$ as the entropy of T over A. If the language is complete, the entropies over analogous sets are equal. The entropy measures the informational content of the theory over finite sets.

Lemma 3. *Suppose that the language is transitive and complete. Theory T is informative if and only if $\mathcal{E}(T) = 0$.*

Lemma 3 ties the informativeness of theory T to a numerical characteristic, its entropy. Notice that the definition of entropy does not depend on the choice of language. In particular, if theory T can be expressed in two (complete and transitive) languages, and it is informative in one of them, it is also informative in the other.

It is not surprising that an entropy is related to the quality of predictions, as similar results are common in the probability and information sciences. The above result may appear to be relatively strong as it says that theory is informative if there exists *some* sequence of finite sets over which the entropy disappears, and not necessarily over all such sets. The result relies heavily on the permissiveness of the definition of an informative theory, and specifically, on the fact that one can choose any instance process to determine whether a theory is informative.

5.2. **Restrictions.** Next, we show that a theory is informative, if it imposes restrictions of certain type. Recall that each theory determines the restrictions of consistent states of the world over finite sets S, $\mathcal{M}(T;S)$. The restriction sets matter the informativeness of T for two reasons. First, the smaller set $\mathcal{M}(T;S)$, the more information theory provides over outcomes of instances in S. Above, we defined entropy over S as a measure of the informational content of a theory. Here, we propose to use a simpler, qualitative measure that distinguishes theories that contain any information from the rest. In the case of binary space of outcomes, $Y = \{0,1\}$, we say that the restrictions over S are *substantive*, if there is at least one possible element excluded from set $\mathcal{M}(T;S)$. Because in the binary case set $\mathcal{M}(T;S)$ cannot be larger than $2^{|S|}$, the restrictions are substantive if $|\mathcal{M}(T;S)| < 2^{|S|}$. In this case, a theory without any substantive restrictions over any finite sets is trivial.

The case of general Y can be reduced to the binary case by mappings $\sigma : Y \to \{0,1\}$ that divide finitely many outcomes in Y into two disjoint sets. We say that a theory has *substantive* restrictions over sS if for each $\sigma : Y \to \{0,1\}$, $|\{\sigma \circ \tau : \tau \in \mathcal{M}(T;S)\}| < 2^{|S|}$.

Second, recall that in complete languages, the theory imposes analogous restrictions on analogous sets. That implies that if the restrictions over S are substantive, the restrictions over any analogous copy S' of S are substantive as well. If S has a very few analogous copies, then an observation that theory T leads to substantive restrictions over S may not matter too much. On the other hand, if S has many analogous copies, the same observation may have much stronger implications.

To capture this intuition, we need a measure of the number of analogous copies of S. It cannot be done simply by counting, because if the language is transitive and complete, each finite subset S of infinite X has infinitely many analogous copies (this follows from Lemma 10 in Appendix A). Instead, we propose to measure how common are analogous copies of S as a subsets of some finite sets. Say that S is ε-generic in finite $A \subseteq X$, if any of subset $D \subseteq A$ with at least ε-proportion of elements, , contains an analogous copy of S: $if, D \subseteq A$ and $|D| \geq \varepsilon |A|$, then $S \tilde{\subseteq} D$. Finite set S is *generic,* if for each $\varepsilon > 0$, S is ε-generic in some finite A.

The next Lemma shows that substantive restrictions over generic sets mean that the theory is informative.

Lemma 4. *Suppose that the language is transitive and complete. Then, any theory with substantive restrictions over some generic set is informative.*

The Lemma provides a sufficient condition for informativeness. The condition is easy to check as long as it is possible to characterize generic sets in a given language. It turns out that in some languages, for example $\mathcal{L}_{\text{trivial}}$ and $\mathcal{L}_{\text{Netflix}}$, all sets are generic. Section 8 contains discussion of these as well as generic sets in other languages.

5.3. **Informative and brief theories.** The converse to Lemma 4 does not always hold, as for example, there are languages like Full language with informative theories and without any generic sets. However, it turns out that when there are sufficiently many generic sets, each informative theory has to lead to substantive restrictions over at least one of them. In order to capture what it means to have sufficiently many generic sets, we have the following definition. We say that a language is *tight* if:there exists $\rho > 0$ such that for each finite A, there is $S \subseteq A$, $|S| \geq \rho |A|$ such that S is generic. Section 8 discusses the tightness of various languages.

In tight languages, the sufficient conditions from Lemma 4 are also necessary.

Theorem 1. *If language is transitive, complete and tight if and only if then theory is informative if and only if it has substantive restrictions over some generic set. In addition, there exists a constant $h_{\mathcal{L}} > 0$ such that for any theory T that is not informative, any learning*

rule π, *any instance process* $\bar{x}$, *and any period* t

$$\inf_{\theta \in \mathcal{M}(T)} U_t(\pi, \bar{x}, \theta) \leq 1 + \frac{1}{t} - h_{\mathcal{L}}.$$

The Theorem characterizes informative theories as those that generate substantive restrictions over generic sets. In the binary case of $Y = \{0,1\}$, theory T is informative if and only if it leads to at least one substantive restriction over generic set. In languages in which all sets are generic, the Theorem implies that all non-trivial theories are generic.

The second part of Theorem 1 characterizes theories that are not informative. For each tight language, there is a constant $h_{\mathcal{L}} > 0$ with the following property. For any theory T that is not informative, there is a state $\theta \in \mathcal{M}(T)$ for which the agent will commit a mistake every $1/h_{\mathcal{L}}$ periods, *no matter what learning rule she uses or what order she observes the instances*. The value of $h_{\mathcal{L}}$ depends on the language, not on the theory.

Transitivity is not essential for the result: As long as there are finitely many classes of analogy of 1-tuples (i.e., there are finitely many non-trivial unary relations), the characterization of informative theories can be divided into each of the classes separately. The role of the other assumptions is indicated in the sketch of the proof below.

In finite languages, the Theorem leads to the following corollary.

Corollary 1. *Suppose that the language is transitive, complete, tight and finite. Then, theory* T *is informative if and only if* T *implies a one-pattern theory* $\{F\}$ *that is informative.*

The Corollary says that any informative theory implies an informative theory that consists of only one pattern. If theory T consists of finitely many patterns, the Corollary is an immediate consequence of the fact that finitely many patterns can be composed into one pattern. The real bite comes when T has infinitely many patterns. By definition, pattern F is built out of finitely many atomic formulas. The Theorem says that a informative theory that is expressed using possibly infinitely many symbols can be replaced by a informative theory that is finitely expressible.

The Corollary 1 is a simple consequence of Theorem 1. We sketch the idea in the binary case $Y = \{0,1\}$. Suppose that S is a generic set such that $|\mathcal{M}(T;S)| < 2^{|S|}$. In particular, there exists a configuration of outcomes $\tau : S \to \{0,1\}$ that is omitted by all states that are consistent with theory T : for each state θ that is consistent with theory T, there exists $x \in S$ so that $\theta(x) \neq \tau(x)$. Let $\bar{x} = (x_1, ..., x_n)$ be an enumeration of elements of set S. Let $F_{\bar{x}}$ be a relational formula in language $\mathcal{L}$ such that $F_{\bar{x}}(\bar{x}')$ is true if and only if tuples $\bar{x}$ and $\bar{x}'$ are analogous. Such a formula exists because the language is finite and complete. Define formula

$$F(\tilde{x}_1, ..., \tilde{x}_n) = F_{\bar{x}}(\tilde{x}_1, ..., \tilde{x}_n) \wedge \left(\bigvee_{m=1,...,n} \theta(\tilde{x}_m) \neq \tau(x_m) \right).$$

Formula requires configuration τ to be omitted on all tuples that are analogous to tuple $\bar{x}$. Because the language is complete and by Lemma 9 from the Appendix, any state θ that is consistent with theory T must be consistent with formula F. In particular, theory T implies theory $\{F\}$. On the other hand, one-pattern theory $\{F\}$ has substantive restrictions over S, and by the above argument, it is informative.

6. Sufficient data condition

In this section, we discuss the existence of a sufficient data condition. The goal is to present a general result that encompasses the examples of sufficient data conditions from Section 4. Notice that each of these sufficient data conditions is stated in terms that are very language-specific as they refer to customers, movies, or goods. In order to present a general condition that encompasses all these examples, we will need quite abstract definitions. In order to motivate these definitions, we begin with a careful discussion of the sufficient condition in the Netflix example.

6.1. **Netflix example.**

Suppose that T is an informative theory in language $\mathcal{L}_{\text{Netflix}}$. By Lemma 3, for each $\varepsilon > 0$, there exists a finite set of instances A_ε such that $\mathcal{E}(A_\varepsilon; T) \leq \varepsilon$.

We argue that there exists a learning rule π such that for each instance process that satisfies (4.2), the average long-run number of mistakes converges (in expectation) to 0. The idea is to divide learning into stages. The learning moves from one stage to the next one when the number of customers and/or movies crosses certain threshold. Specifically, fix a sequence of thresholds $n_k = 2^k$. We say that the learning instance process is in stage (k, l) if $n_{k-1} < n_t^C \leq n_k$ and $n_{l-1} < n_t^M \leq n_l$.

For each stage (k, l), fix an $n_k n_l$-element sets of customer-movie pairs with exactly n_k customers and n_l movies U_{kl}. In other words, U_{kl} is a product of n_k-element set of customers and n_l-element set of movies. An argument from the proof of Lemma 3 shows that one can construct a learning rule π_{kl} that ensures the expected total number of mistakes along any finite sequence of instances from set U_{kl} is not larger than the entropy of set U_{kl} multiplied by the number of elements in this set,

$$\mathcal{E}(T; U_{kl})\, n_k n_l. \tag{6.1}$$

We construct a learning rule π so that the predictions in stage (k, l) are made using a version of learning rule π_{kl} applied to a dataset of observations taken only from the same stage, and the total number of mistakes in this stage is not larger than bound (6.1). There is a (minor) difficulty: Learning rule π_{kl} is designed for set U_{kl} and U_{kl} is chosen somehow arbitrarily as there are infinitely many other $n_k n_l$-element product sets. Although the observations from stage (k, l) belong to some $n_k n_l$-element product set U, there is no guarantee

that $U = U_{kl}$. However, notice that all such sets U are analogous. We can use this fact to find a version of learning rule π_{kl} that ensures that the total number of mistakes is bounded by (6.1) along *any* finite (at most $n_k n_l$-element) sequence of instances that is *analogous* to a sequence from set U_{kl}.

The construction ensures that the average number of mistakes till period t depends on the stages $(k_1, l_1), ..., (k_m, l_m)$ visited before t and it can be bounded by

$$\left(\frac{1}{t}|U_{k_1 l_1}|\right)\mathcal{E}(T; U_{k_1 l_1}) + ... + \left(\frac{1}{t}|U_{k_m l_m}|\right)\mathcal{E}(T; U_{k_m l_m}). \tag{6.2}$$

Suppose now that the instance process satisfies condition (4.2). The first part of the condition guarantees that sets U_{kl} become larger and larger in the sense of inclusion. In particular, for sufficiently high k and l, $A_\varepsilon \tilde{\subseteq} U_{kl}$. We use the properties of set U_{kl} (namely, that set U_{kl} is local in the sense defined below) to show that $\mathcal{E}(T; U_{kl}) \leq \mathcal{E}(T; A_\varepsilon) \leq \varepsilon$. It follows that

$$\lim_m \mathcal{E}(T; U_{k_m l_m}) = 0.$$

The second condition, together with the choice of thresholds ensures that for each t,

$$\frac{1}{t}(|U_{k_1 l_1}| + ... + |U_{k_m l_m}|) \leq \frac{1}{t}2(|U_{k_m l_m}|) \leq \frac{1}{t}2(2t) = 4 \tag{6.3}$$

is uniformly bounded away from infinity. These two observations guarantee that as t goes to infinite, bound (6.2) and the average number of mistakes converges to 0.

It is instructive to explain the role of two parts of condition (4.2). The first part of condition (4.2) ensures that the stages of the learning process become larger and larger. Because they eventually contain (up to analogy) any finite set A, eventually any information contained in theory T about the restrictions of the states to A will become eventually useful.

The second part of the condition ensures that new data do not arrive too quickly. By construction, the bound on the number of mistakes committed in stage (k, l) is equal to $|U_{kl}|\,\mathcal{E}(T; U_{kl})$. Typically, a majority of these mistakes is committed in the initial periods of stage (k, l) when the number of observations in set U_{kl} is small relatively to the size of U_{kl}. (This is because the learning rule in this stage relies only on the observations from the same stage, and in order to make good predictions, the learning rule needs past data.) Even if the average number of mistakes till the beginning of stage (k, l) is small, a large number of mistakes in the initial periods of stage (k, l) may temporarily raise the average much beyond what is necessary for the long-term convergence. Inequality (6.3) ensures that this is not going to happen.

6.2. **Local sets.** Before we describe the general definition of sufficient data condition, we state an useful definition. A finite set $U \subseteq X$ is *local* if for any two tuples $\bar{x}$ and $\bar{x}'$ of elements of U such that the two tuples are analogous, for each $x \in U$, there exists $x' \in U$

such that tuples $\bar{x}\hat{}x$ and $\bar{x}'\hat{}x'$ are analogous. In other words, U is local if the restriction of the language $\mathcal{L}$ to set U is complete.[5]

We interpret local sets as finite approximations of infinite space X. The usefulness of local sets come from the fact that the entropy over a local set is an lower bound on the entropy of any of its subsets. The proof of the next result can be found in Appendix A.3.

Lemma 5. *Suppose that language $\mathcal{L}$ is transitive. For each local set U, each $A \subseteq U$, $\mathcal{E}(T;A) \geq \mathcal{E}(T;U)$.*

For example, in the trivial language $\mathcal{L}_{\text{trivial}}$, any set is local. In the language $\mathcal{L}_{\text{Netflix}}$, all products $A \times B$, where A is a finite set of customers and B is a finite set of movies are local. In the language $\mathcal{L}_{\text{choices}}$, any set of bundles of goods from some finite set of goods A is local.

6.3. **Sufficient data condition.** Suppose that $\mathcal{U}$ is a collection of local sets with the following identification property: for each finite tuple $\bar{x}$, there exists local set $U(\bar{x}) \in \mathcal{U}$ such that $\bar{x}\tilde{\subseteq}U(\bar{x})$ and for any $U \in \mathcal{U}$, if $\bar{x}\tilde{\subseteq}U$, then $U(\bar{x})\,\tilde{\subseteq}U$. In other words, for each finite tuple, there exists the smallest (up to an analogy class) local set U that contains the tuple.

Say that instance process $\bar{x} = (x_1, x_2, ...)$ is $\mathcal{U}$-adapted if two conditions are satisfied: (a) for each finite A, there exists t such that $A\tilde{\subseteq}U(\bar{x}^t)$ and (b)

$$\sup_t \frac{1}{t}\sum\nolimits_{U\in\{U(\bar{x}^s):s\leq t\}} |U| < \infty.$$

The first condition guarantees that the process encounters more and more new data and it corresponds to the first part of condition (4.2). The second condition guarantees that the new data does not arrive too quickly and it corresponds to the second part of condition (4.2).

Theorem 2. *Suppose that language $\mathcal{L}$ is transitive and complete $\mathcal{U}$ is a family of disjoint local sets. Then, the class of all $\mathcal{U}$-adapted processes satisfies the sufficient data condition.*

The proof can be found in Appendix F.

As an example, consider language $\mathcal{L}_{\text{Netflix}}$ and let $\mathcal{U}_{\text{Netflix}}$ be the collection of all product sets of 2^k customers and 2^l movies for some k and l. The above discussion demonstrates that each instance process that satisfies (4.2) is $\mathcal{U}$-adapted. In the consumer's choices example, let $\mathcal{U}_{\text{choices}}$ be a collection of sets that consist of all bundles of goods from some A, where $|A| = 2^k$. A similar argument to the one used in the case of $\mathcal{L}_{\text{Netflix}}$, shows that each instance process that satisfies (4.5) is $\mathcal{U}_{\text{choices}}$-adapted.

[5]Formally, a restriction of language $\mathcal{L}$ to set U is a collection of relations $\mathcal{L}_U = \{R \cap U^k : R \in \mathcal{L} \text{ and } R \text{ is a } k\text{-ary relation}\}$.

7. Universal learning rule

So far, we assumed that the agent knows that a certain theory is correct. The purpose of this exercise is to measure the information that is provided by a theory. In the real world, the agent faces the problem of induction: she looks for a theory by analyzing data. The question is it possible to learn theory from data? The next result gives a partial answer to the question: it constructs an universal learning rule π^* such that for *any* informative theory T, and any instance process from some sufficient data condition class, (3.2) holds. In other words, if the true state of the world θ is consistent with *some* informative theory T, learning rule π^* will predict the outcomes $\theta(x)$ correctly in the long-run, regardless what T is.

Theorem 3. *Suppose that the language is complete, transitive, finite, and tight and there exists a sufficient data condition $\bar{X}$. Then, there exists a learning rule π^* such that (3.2) holds for all instance processes $\bar{x} \in \bar{X}$, and for all theories T that are informative.*

Because there is a single learning rule that ensures (3.2) for all informative theories, Theorem 3 inverts the order of quantifiers from the definition of such theories.

Proof. The equivalence between (1), (2), and (4) from Theorem 1 shows that if theory T is informative, then it an informative one-pattern theory $\{F\}$. Additionally, the proof of Theorem 1 shows that each such pattern has the following form: Take a generic set S and find enumeration $\bar{x} = (x_1, ..., x_k)$ of S. Let $F_{\bar{x}}(\tilde{x}_1, ..., \tilde{x}_k)$ be a relational formula that defines the analogy class of tuple $\bar{x}$ (i.e., for each tuple $\bar{x}'$, $R(\bar{x}')$ if and only if $\bar{x}'$ is analogous to $\bar{x}$). Let $F_\theta(\tilde{x}_1, ..., \tilde{x}_k)$ be a formula that consists of atomic formulas $\theta(\tilde{x}_l) = y_l$, together with logical symbols. Then, $F = F_{\bar{x}} \wedge F_\theta$.

Let $\mathcal{T}$ be the set of one-pattern theories that are informative and that are constructed in such a way. Because any theory T implies a theory $T' \in \mathcal{T}$, the Theorem will be proved if there exists a learning rule π^* such that

$$\inf_{T \in \mathcal{T}} \lim \inf_{t \to \infty} \inf_{\theta \in \mathcal{M}(T)} U_t(\pi^*, \bar{x}, \theta) = 1.$$

Because Y is finite, and the set of finite tuples if countable, $\mathcal{T}$ is countable. Enumerate all theories in $\mathcal{T}$ by $T_1, T_2, ...$. By the definition of the sufficient data condition, for each k, there exists a learning rule π_k such that for each instance process $\bar{x} \in \bar{X}$,

$$\lim \inf_{t \to \infty} \inf_{\theta \in \mathcal{M}(T_k)} U_t(\pi_k, \bar{x}, \theta) = 1.$$

Each period t, let k_t be the lowest index such that theory T_k has not been contradicted by data observed before period k. Define $\pi_t^* := \pi_{k_t}$. If there exists theory $T \in \mathcal{T}$ that is true, then k_t will not grow to infinity. In other words, there is period t^*, such that for all $t > t^*$,

$k_t = k_{t^*}$. Because of the long-run payoff criterion, the payoffs from the prediction in periods before t^* do not affect the limit of payoffs and,

$$\begin{aligned}&\inf_{T\in\mathcal{T}} \lim \inf_{t\to\infty} \inf_{\theta\in\mathcal{M}(T)} U_t\left(\pi^*, \bar{x}, \theta\right)\\ &= \lim \inf_{t\to\infty} \inf_{\theta\in\mathcal{M}\left(T_{k_{t^*}}\right)} U_t\left(\pi_{k_{t^*}}, \bar{x}, \theta\right) = 1.\end{aligned}$$

□

8. Tightness of languages

This section discusses the tightness of languages from Section 4.

8.1. **Trivial language.** $\mathcal{L}_{\text{trivial}}$ is complete and transitive. Each finite subset is local and generic, and the language is trivially tight with tightness $\rho\left(\mathcal{L}_{\text{trivial}}\right) = 1$.

8.2. **Product languages.** In the Netflix example from the Introduction, the set of instances is a two-dimensional product of the set of customers and the movies, $X = C \times M$. More generally, let $X = X^1 .... \times X^d$ be a product of d sets. For each l, let $\mathcal{L}_l = \{R_l\}$ be a language on set X_l. For each $l \leq d$ and each $R \in \mathcal{L}_l$, define k_R-ary relation R^* on set X : for each $x_1, ..., x_{k_R} \in X$, let

$$R^*\left(x_1, ..., x_{k_R}\right) \Longleftrightarrow R\left(x_1^l, ..., x_{k_R}^l\right).$$

Define product language $\mathcal{L} := \bigcup_l \{R^* : R \in \mathcal{L}_l\}$. The next result shows that genericity is preserved under products. The proof can be found in Appendix 6.

Lemma 6. *If sets $S_l \subseteq X_l$ are generic in languages $\mathcal{L}_l$, then $S_1 \times ... \times S_d$ is generic in language $\mathcal{L}$.*

8.3. **Netflix.** Notice that language $\mathcal{L}_{\text{Netflix}}$ is a product of two independent copies of trivial languages. Lemma 6, the fact that any finite set is generic in language $\mathcal{L}_{\text{trivial}}$, and the fact that any subset of a generic set is also generic lead to a simple corollary.

Corollary 2. *Any finite $S \subseteq X$ is generic in language $\mathcal{L}_{Netflix}$. In particular, language $\mathcal{L}_{Netflix}$ is tight.*

8.4. **Consumer's choices.** Next, we show that language $\mathcal{L}_{\text{choices}}$ is tight. Assume that X is the set of ordered tuples of distinct elements of B.

Lemma 7. *For any finite and mutually disjoint sets $B^1, ...B^d \subseteq B$, $B^1 \times ... \times B^d$ is generic in language $\mathcal{L}_{choices}$. For any finite $A \subseteq X$, there exist finite and disjoint $B^1, ...B^d \subseteq B$ such that*

$$\left|A \cap B^1 \times ... \times B^d\right| \geq d^{-d} \left|A\right|.$$

Thus, language $\mathcal{L}_{choices}$ is tight.

The proof can be found in Appendix G.2. We discuss some examples of generic and non-generic sets. To fix attention, assume that $d = 2$. We show that set

$$S = \{(a, b), (b, c), (a, c)\}$$

is not generic. Indeed, it is easy to see that for any finite A, there are finite and disjoint sets of goods B^1 and B^2 such that $|A \cap B^1 \times B^2| \geq \frac{1}{4}|A|$. But neither S nor any of its analogous copies can be represented as a subset of the product of two disjoint sets of goods B^1 and B^2. The contradiction shows that S is not generic.

On the other hand, take

$$S' = \{(a_1, b_1), (a_2, b_1), (a_1, b_2), (a_2, b_2)\}$$

for some a_1, b_1, a_2, b_2. Then, $S' = \{a_1, a_2\} \times \{b_1, b_2\}$ is generic by Lemma 7.

Let F_{trans} denote transitivity pattern (4.4). We show that theory $\{F_{\text{trans}}\}$ is informative. Notice that $\mathcal{E}(\{F_{\text{trans}}\}, S) < 1$. Unfortunately, Lemma 16 cannot be applied, because S is not generic. On the other hand, notice that transitivity F_{trans} implies pattern F

$$\forall_{x_1,x_1',x_2,x_2'} x_1 R_{11} x_1' \wedge x_2 R_{11} x_2' \wedge x_1 R_{22} x_2 \wedge x_1' R_{22} x_2' \wedge x_1 \neq x_2 \wedge x_1 \neq x_1'$$
$$\wedge \theta(x_1) = 2 \wedge \theta(x_1') = 1 \wedge \theta(x_2) = 1 \Longrightarrow \theta(x_2') = 1.$$

To see it notice that the pattern is satisfied only if there exists distinct goods a, b, c, d such that

$$x_1 = (a, c), x_1' = (a, d), x_2 = (b, c), x_2' = (b, d).$$

By transitivity, if the customer chooses b from bundle x_2 and c from bundle x_1, then she must choose b from bundle $x = (a, b)$. In addition, if she chooses a from x_1', then she must choose b from bundle x_2'.

Then, $\mathcal{E}(\{F\}, S') < 1$. Because S' is generic, theories $\{F_{\text{trans}}\}$ and $\{F\}$ are informative.

8.5. **Chemist.** Here, we show the tightness of language $\mathcal{L}_{\text{chemist}}$. For any mutually disjoint finite sets $B^1, ...B^d \subseteq B$, define:

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

Suppose that sets B^j are disjoint. Then, $X\left(B^1, ..., B^d\right)$ contains d-element subsets $x \subseteq B, |x| = d$, such that x contains exactly one element of B^j, $|x \cap B^j| = 1$ for each $j = 1, .., d$.

Lemma 8. *$X\left(B^1, ..., B^d\right)$ is generic in language $\mathcal{L}_{choices}$ for any finite and disjoint $B^1, ...B^d \subseteq B$, $B^l \cap B^{l'} = \varnothing$ for $l \neq l$. For any finite $A \subseteq X$, there exist finite and disjoint $B^1, ...B^d \subseteq B$ such that*

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

Thus, language $\mathcal{L}_{chemist}$ is tight.

The proof of Lemma 8 is analogous to the proof of Lemma 7, and, therefore, skipped. As an example, we show that the acid-alkaline theory is informative. The acid-alkaline theory implies pattern F :

$$\forall_{x_1,x_1',x_2,x_2'} x_1 R_1 x_1' \wedge x_2 R_1 x_2' \wedge x_1 R_1 x_2 \wedge x_1' R_1 x_2' \wedge x_1' \neq x_2$$
$$\wedge \theta(x_1) = 1 \wedge \theta(x_1') = 1 \wedge \theta(x_2) = 1 \Longrightarrow \theta(x_2') = 0.$$

Take any 4-tuple x_1, x_1', x_2, x_2' that satisfies the first part of the pattern. Then,

$$x_1 = (a_1, b_1), x_1' = (a_2, b_1), x_2 = (a_1, b_2), x_2' = (a_2, b_2).$$

for some a_1, b_1, a_2, b_2. Pattern F 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 $\{x_1, x_1', x_2, x_2'\} = X(\{a_1, a_2\}, \{b_1, b_2\})$ is generic by Lemma 8, Theorem 1 implies that $\{F\}$ and the acid-alkaline theory are informative.

8.6. **Full language $\mathcal{L}_X$.** In language $\mathcal{L}_X$, any two sets A and A' are analogous if and only if $A = A'$. It follows that language $\mathcal{L}_X$ cannot be tight.

9. Review of literature

This paper is closely related to the statistical *VC-theory* and to the computer-science model of *probability approximately correct learning* (*PAC*); see Vapnik (1998). Let X be a set of instances, M be a set of functions $\theta : X \to \{0, 1\}$, and μ be a distribution on X. Say that M is *PAC-learnable* for distribution μ if there exist prediction functions $\pi_t : (X \times Y)^{t-1} \times X \to \Delta Y$ such that for each $\theta \in C$,

$$\lim_{t\to\infty} E_\mu \pi_t \left((x_s, \theta(x_s))_{s<t}, x_t \right) (\theta(x_t)) = 1,$$

where $x_1, ..., x_t$ are drawn i.i.d. from distribution μ. The classic result of the PAC-theory is the Vapnik-Chervonenkis Theorem that says that set M is PAC-learnable for *all* distributions $\mu \in \Delta X$ if and only if it has a finite *VC-dimension*, i.e., there exists constant k such that for each finite set $A \subseteq X$, if $|A| \leq k$, then the cardinality of the set of restrictions $M(A) = \left\{\theta_{|A} : \theta \in M\right\}$ is strictly smaller than 2^k.

We briefly describe the main differences. The first one is somewhat technical: we do not use a probability distribution μ over instances. If X is infinite and discrete, it is difficult (or impossible, if X is transitive) to point to a "right" distribution. For example, it is difficult to point to a natural distribution in the Netflix example.

Notice that the role of a distribution in the PAC model to generate instances; that role is played in our model by instance processes. And so, the PAC-learnability of set $\mathcal{M}(T)$ for all distributions corresponds to the definition of strongly informative theories defined in Section. Lemma 2 can be interpreted as a (very simple) corollary of the VC-Theorem in the case when set $\mathcal{M}$ is equal to the set of states of theories expressed in language $\mathcal{L}_{\text{Netflix}}$.

The second difference between the VC-theorem and Theorem 1 is that the latter has a weaker hypothesis and a weaker thesis. On one hand, Theorem 1 establishes the learnability only for some processes, not for all. On the other hand, statement (1) of Theorem 1 is weaker than finite VC-dimension: the statement says that there exists a generic S such that the cardinality of the set of restrictions $M(T;S)$ is strictly smaller than $2^{|S|}$. In other words, instead of restrictions on all finite k-element subsets, statement (1) imposes restrictions only on set S (and its analogous copies).

Both PAC-learning and Theorem 1 heavily rely on the bounds on entropy obtained from Sauer-Shelah's Lemma. In specific cases, these bounds are far from sharp. For example, with $\mathcal{L}_{\text{choices}}$, the problem can be restated as the *omitted subgraph problem,* which is studied by the graph-theoretic literature (for example, 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)). With $\mathcal{L}_{\text{Netflix}}$, much sharper bounds on the entropy can be obtained as a corollary to Lemma 6.3 from Alon, Fischer, and Newman (2005). Note that better bounds could not improve the qualitative claims of Theorem 1; however, they may lead to a (wider) sufficient data conditions. There are two open questions. Are there any better general bounds? Second, what are the best bounds possible in specific cases?

Kalai (2003) and Salant (2007) study the informativeness of specific theories in specific languages (the transitivity of consumer's choices in the former case, and the majority voting in the latter). These two papers very closely follow the PAC model; in particular, they assume that the sets of instances are finite and the distribution μ is uniform. By focusing on specific examples, these papers are able to obtain sharper bounds on the number of possible mistakes in finite samples.

More broadly, the current paper is related to the economic literature on languages (for an excellent review, see Lipman (2003)). Rubinstein (1996) (also Chapter 1 of Rubinstein (2000)) notices that natural languages often contain certain binary relations (for example, "is bigger that"). He asks which properties distinguish binary relations observed in the natural languages from other relations. Among others, he argues that binary relations should be easy to describe: One should be able to learn it from a universal proposition (theory, in my terminology) aided by a few examples. (In a similar spirit, Blume (2004) postulates that a

natural language, understood as an assignment of meanings to utterances, should be simple to learn.) There exist some rough correspondences between those definitions and the current paper: In the terminology of the current paper, a binary relation corresponds to a state θ, and the ease of describibility corresponds to informativeness. Rubinstein looks for relations that are optimal with respect to such a criterion. Here, the goal is to characterize the set of all informative (i.e., "easy to describe") theories.

References

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.

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.

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.

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.

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

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.

——— (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. COMPLETE LANGUAGES

In this appendix, we present some general results about complete languages.

A.1. **Fundamental property.** The next result shows that if the language is complete, then theory T imposes the same type of restrictions over analogous tuples or sets.

Lemma 9. *Suppose that language $\mathcal{L}$ is complete. Then, for each theory T, any two analogous tuples $\bar{x}, \bar{x}' \in X^t$, each $y_1, ..., y_t$, if there exists $\theta \in \mathcal{M}(T)$ such that $\theta(x_s) = y_s, s \leq t$, then there exists $\theta' \in \mathcal{M}(T)$ such that $\theta'(x_s) = y_s, s \leq t$.*

Proof. Let $\bar{x} = (x_1, ..., x_t, x_{t+1}, ...)$ and $\bar{x} = \left(x'_1, ..., x'_t, x'_{t+1}, ...\right)$ be enumerations of set X such that for each m,, tuples $(x_1, ..., x_m)$ and $(x'_1, ..., x'_m)$ are analogous. Such enumerations exist, because X is countably infinite, and because of the extension property of complete languages.

Suppose that there exists $\theta \in \mathcal{M}(T)$ such that $\theta(x_s) = y_s, s \leq t$. Define $\theta' \in \mathcal{M}$ so that $\theta'(x'_m) = \theta(x_m)$ for each m. We show that $\theta' \in \mathcal{M}(T)$. Indeed, take any universal pattern $F \in T$ such that formula F has k free variables. For any tuple $\left(x'_{m_1}, ..., x'_{m_k}\right) \in X^k$, the truth value of statement $F\left(x'_{m_1}, ..., x'_{m_k}; \theta'\right)$ is the same as the truth value of statement $F(x_{m_1}, ..., x_{m_k}; \theta)$. Because the latter statement is true (since $\theta \in \mathcal{M}(T) \subseteq \mathcal{M}\{F\}$), it must be that $F\left(x'_{m_1}, ..., x'_{m_k}; \theta'\right)$ is true. In particular, state θ' is consistent with F. Because any pattern could have been chosen, it must be that $\theta' \in \mathcal{M}(T)$. □

Proof of Lemma 1. The result follows from Lemma 9. □

A.2. **Transitive languages.**

Lemma 10. *Suppose that language $\mathcal{L}$ is complete and transitive and X is infinite. Then, for each two finite sets $A, B \subseteq X$, there exists A' such that A and A' are analogous and A' and B are disjoint.*

Proof. Let G be the set of all bijections $g : X \to X$ such that for each tuple $\bar{x} = (x_1, ..., x_k)$, tuples $\bar{x}$ and $g(\bar{x}) = (g(x_1), ..., g(x_k))$ are analogous. Because the language is complete, by the same argument as in Lemma 9, we can show that two tuples $\bar{x}$ and $\bar{x}'$ are analogous if and only if there exists $g \in G$ such that $\bar{x}' = g(\bar{x})$. In particular, if the language is transitive, then for each x and x', there exists $g \in G$ such that $g(x) = x'$.

The proof proceeds by induction on the size of set A. If $|A| = 1$, then the claim follows from the fact that the language is transitive and any two elements are analogous. Suppose that the claim holds for all finite B and all A such that $|A| \leq k - 1$. Take any A and B such that $|A| \leq k - 1$.

By the repeated application of the inductive claim, we can find infinitely many $g_1, g_2, ...$ such that $g_n \cdot A \cap \left(B \cup \bigcup_{m<n} g_m \cdot A\right) = \varnothing$. Suppose that there is $a \notin A$ such that $g_n \cdot a \in B$ for each n. Because B is finite, there exists b such that set $N_0 = \{n : g_n \cdot a = b\}$ has infinite cardinality. Because of transitivity, there exists $g \in G$ such that $g \cdot a \notin B$. Because B is finite and sets $g_n \cdot A$ are disjoint, there is $n \in N_0$ such that $gg_n \cdot A \cap B = \varnothing$. This ends the proof of the inductive claim. □

A.3. **Entropy and transitive languages.** Fix theory T. Suppose that $U \subseteq X$ is a local set. Let G be the set of bijections $g : U \to U$ such that for each tuple $\bar{x} \subseteq U$, tuples $\bar{x} = (x_1, ..., x_k)$ and $(g(x_1), ..., g(x_k))$ are analogous. Then, G contains identity bijection, it is closed with respect to compositions (for each $g, g' \in G$, $g \circ g' \in G$) and inverses (for each $g \in G$, $g^{-1} \in G$). Moreover, if language $\mathcal{L}$ is transitive and U is local, then for any two x and $x' \in U$, there exists $g \in G$ such that $g(x) = x'$.

For any subset $A \subseteq U$, let Y^A be the set of functions from A to Y. For any $\tau_A \in Y^A$, let $[\tau_A] = \{\tau \in \mathcal{M}(T;U) : \tau|_A = \tau_A\}$ be the set of functions $\tau \in Y^U$ which restrictions to A are equal to τ_A. For any subset $A \subseteq U$, define

$$\mathcal{E}_A := -\sum_{\tau \in Y^U} \frac{1}{\mathcal{M}(T;U)} \log \frac{|[\tau|_A]|}{\mathcal{M}(T;U)}.$$

Lemma 11. $\frac{1}{|U|}\mathcal{E}_U = \mathcal{E}(T;U)$. *For each* $A \subseteq U$, $\frac{1}{|A|}\mathcal{E}_A \leq \mathcal{E}(A;T)$. *For each* $A \subseteq U$, *and for each* $g \in G$, $\mathcal{E}_A = \mathcal{E}_{g(A)}$.

Proof. The first observation is immediate. Let $p \in \Delta\mathcal{M}(T;A)$ be any probability distribution over functions $\tau_A \in \mathcal{M}(T;A)$. Let $p_0 \in \Delta\mathcal{M}(T;A)$ be the uniform probability distribution. By standard results,

$$\begin{aligned}\sum_{\tau_A \in Y^A} p(\tau_A) \log p(\tau_A) &\geq \sum_{\tau_A \in Y^A} p_0(\tau_A) \log p_0(\tau_A) \\ &= -\log|\mathcal{M}(T;A)| = -|A|\,\mathcal{E}(A;T).\end{aligned}$$

Take $p(\tau_A) := \frac{|[\tau_A]|}{\mathcal{M}(T;U)}$.

The last part of the Lemma follows from the following observation: For any $\tau \in \mathcal{M}(T;U)$, and each $g \in G$, $\tau_g \in \mathcal{M}(T;U)$, where $\tau_g \in Y^U$ is a function defined as $\tau_g(x) = \tau(g \cdot x)$ for each $x \in U$. Then, for each $\tau \in \mathcal{M}(T;U)$, and each $g \in G$, $\tau_g \in \mathcal{M}(T;U)$. □

Lemma 12. *For any* $A, B \subseteq U$,

$$\mathcal{E}_{A\cup B} - \mathcal{E}_A \leq \mathcal{E}_B - \mathcal{E}_{A\cap B}.$$

Proof. Observe that

$$\mathcal{E}_{A\cup B} - \mathcal{E}_A = -\sum_{\tau \in Y^U} \frac{1}{\mathcal{M}(T;U)} \log \frac{|[\tau|_{A\cup B}]|}{|[\tau|_A]|} \text{ and}$$

$$\mathcal{E}_B - \mathcal{E}_{A\cap B} = -\sum_{\tau \in Y^U} \frac{1}{\mathcal{M}(T;U)} \log \frac{|[\tau|_B]|}{|[\tau|_{A\cap B}]|}.$$

Because the logarithm function is concave, Jensen's inequality implies that

$$\begin{aligned}&\mathcal{E}_{A\cup B}-\mathcal{E}_A-[\mathcal{E}_B-\mathcal{E}_{A\cap B}]\\&=\sum_{\tau\in Y^U}\frac{1}{\mathcal{M}(T;U)}\log\frac{|[\tau|_A]|\,|[\tau|_B]|}{|[\tau|_{A\cap B}]|\,|[\tau|_{A\cup B}]|}\leq\log\left(\sum_{\tau\in Y^U}\frac{1}{\mathcal{M}(T;U)}\frac{|[\tau|_A]|\,|[\tau|_B]|}{|[\tau|_{A\cap B}]|\,|[\tau|_{A\cup B}]|}\right).\end{aligned}$$

The Lemma follows from the fact that

$$\begin{aligned}&\sum_{\tau\in Y^U}\frac{1}{\mathcal{M}(T;U)}\frac{|[\tau|_A]|\,|[\tau|_B]|}{|[\tau|_{A\cap B}]|\,|[\tau|_{A\cup B}]|}\\&=\sum_{\tau_{A\cap B}\in Y^{A\cap B}}\frac{|[\tau_{A\cap B}]|}{\mathcal{M}(T;U)}\sum_{\tau_{A\cup B}\in Y^{A\cup B},\text{ st. }\tau_{A\cup B}|_{A\cap B}=\tau_{A\cap B}}\frac{|[\tau_{A\cup B}|_B]|\,|[\tau_{A\cup B}|_A]|}{|[\tau_{A\cap B}]|\,|[\tau_{A\cap B}]|}\\&=\sum_{\tau_{A\cap B}\in Y^{A\cap B}}\frac{|[\tau_{A\cap B}]|}{\mathcal{M}(T;U)}\left(\sum_{\tau^A\in Y^A\text{ st.}\tau^A|_{A\cap B}=\tau_{A\cap B}}\frac{\left|\left[\tau^A\right]\right|}{|[\tau_{A\cap B}]|}\right)\left(\sum_{\tau^B\in Y^B\text{ st.}\tau^B|_{A\cap B}=\tau_{A\cap B}}\frac{\left|\left[\tau^B\right]\right|}{|[\tau_{A\cap B}]|}\right)\\&=\sum_{\tau_{A\cap B}\in Y^{A\cap B}}\frac{|[\tau_{A\cap B}]|}{\mathcal{M}(T;U)}=1.\end{aligned}$$

□

Lemma 13. *For any $A\subseteq U$,*

$$\frac{1}{|A|}\mathcal{E}_A\geq\frac{1}{|U|}\mathcal{E}_U.$$

Proof. Choose $A^*\subseteq U$ so that (a) $A^*\in\arg\max_A\frac{1}{|A|}\mathcal{E}_A$ and (b) for each $A\supseteq A^*$, $A\neq A^*$, $\frac{1}{|A|}\mathcal{E}_A<\frac{1}{|A^*|}\mathcal{E}_{A^*}$. Such a set exists because U is finite. If $A^*=U$, then the Lemma holds. Suppose not and $A^*\neq U$. Since $G\longmapsto U$ is transitive, there is $g\in G$, so that $g(A^*)\neq A^*$. By Lemma 12,

$$\mathcal{E}_{A^*\cup g(A^*)}\leq\mathcal{E}_{A^*}+\mathcal{E}_{g(A^*)}-\mathcal{E}_{A^*\cap g(A^*)}.$$

By Lemma 11,

$$\begin{aligned}\frac{\mathcal{E}_{A^*\cup g(A^*)}}{|A^*\cup g(A^*)|}&\leq\frac{\mathcal{E}_{A^*}+\mathcal{E}_{g(A^*)}-\mathcal{E}_{A^*\cap g(A^*)}}{|A^*|+|g(A^*)|-|A^*\cap g(A^*)|}\\&=\frac{\mathcal{E}_{A^*}}{|A^*|}+\frac{\left(\frac{\mathcal{E}_{A^*}}{|A^*|}-\frac{\mathcal{E}_{A^*\cap g(A^*)}}{|A^*\cap g(A^*)|}\right)\cdot|A^*\cap g(A^*)|}{2|A^*|-|A^*\cap g(A^*)|}.\end{aligned}$$

Because of (a),

$$\frac{\mathcal{E}_{A^*}}{|A^*|}\leq\frac{\mathcal{E}_{A^*\cap g(A^*)}}{|A^*\cap g(A^*)|}.$$

However, this implies that

$$\frac{\mathcal{E}_{A^*\cup g(A^*)}}{|A^*\cup g(A^*)|}\leq\frac{\mathcal{E}A^*}{|A^*|},$$

which leads to a contradiction with (b). □

Proof of Lemma 5. The Lemma follows from Lemmas 11 and 13. □

Appendix B. Proof of Lemma 2

Let $\bar{x} = (x_1, x_2, ...)$ be an instance process such that for each $m \neq n$, neither $x_m R_C x_n$ nor $x_m R_M x_n$. Let $A_m = \{x_1, ..., x_m\}$. If (3.2) holds for instance processes $\bar{x}$ and some learning rule π, then it must be that $|\mathcal{M}(T; A_k)| < 2^k$ for some k. In particular, there are $y_1, ..., y_k \in \{0, 1\}$ such that for each $\theta \in \mathcal{M}(T)$, if $\theta(x_l) = y_l$ for each $l < k$, then $\theta(x_k) = y_k$.

Let $A_n^y(\theta) = \{m \leq n : \theta(x_m) = y\}$. We show that for each $n \geq 2k$, for each $\theta \in \mathcal{M}(T)$, either $|A_n^0(\theta)| \leq k$ or $|A_n^1(\theta)| \leq k$. If not, there exists a k-tuple $(x_{m_1}, ..., x_{m_k})$ of distinct elements of A_n such that $\theta(x_{m_l}) = y_l$ for each $l < k$ and $\theta(x_{m_k}) \neq y_k$. Notice that tuple $(x_1, ..., x_k)$ is analogous to tuple $(x_{m_1}, ..., x_{m_k})$. Because language $\mathcal{L}_{\text{Netflix}}$ is complete, and because of Lemma 9 from the Appendix, there is $\theta' \in \mathcal{M}(T)$ so that $\theta(x_l) = y_l$ for each $l < k$ and $\theta(x_k) \neq y_k$. A contradiction.

Suppose that there exists $\theta \in \mathcal{M}(T)$ such that for some y, there are $2k+2$-element set of customers $\{c_1, ..., c_{2k+2}\}$ and movies $\{m_1, ..., m_{2k+2}\}$ such that $\theta(c_i, m_i) = y$ for each $i \leq k+1$ and $\theta(c_i, m_i) \neq y$ for each $i \geq k+2$. Because tuples $(x_1, ..., x_{2k+2})$ and $((c_1, m_1), ..., (c_{2k+2}, m_{2k+2}))$ are analogous, by Lemma 9, there exists $\theta' \in \mathcal{M}(T)$ such that $\theta'(x_i) = y$ for each $i \leq k+1$ and $\theta'(x_i) \neq y$ for each $k+2 \leq i \leq 2k+2$. This leads to a contradiction with the above observation. THe contradiction demonstrates the Lemma.

Appendix C. Proofs of Section 5

C.1. **Entropy characterization.** We start with two preliminary results. The proof of the next Lemma can be found in Appendix D.

Lemma 14. *Suppose that language $\mathcal{L}$ is transitive and complete. For each theory T, if $\mathcal{E}(T) = 0$, then theory T is informative.*

For each $w \in (0, 1]$, define $h(w)$ as the unique solution to equation

$$w = h(w)(\log|Y| + 2 - \log h(w)).$$

For each $w > 0$, $h(w) > 0$ is well-defined, continuous, and increasing in w.[6]

The next Lemma shows that for any learning rule, the average number of mistakes committed by the learning rule over any set A is not smaller than $h(\mathcal{E}(T))$. The proof can be found in Appendix E.

[6]Consider function $w(h) = h\left(\log|Y| + 2 + \log\frac{1}{h}\right)$. Then, $w(h) = 0$, $w(.)$ is strictly increasing for any $h \leq 1$, and $w(1) \geq 1$. Thus, there exists an increasing function $h : [0, 1] \to R$ such that for each $w \in [0, 1]$, $w(h(w)) = w$.

Lemma 15. *Suppose that language $\mathcal{L}$ is transitive and complete. For each theory T, any instance process $\bar{x}$, learning rule π, and any period t*

$$\inf_{\theta\in\mathcal{M}(T)} U_t\left(\pi,\bar{x},\theta\right)\leq 1+\frac{1}{t}-h\left(\mathcal{E}\left(T\right)\right). \tag{C.2}$$

The two results imply Lemma 3

C.2. **Sauer-Shelah's Lemma.** Consider first the binary case $Y=\{0,1\}$. First, we show that a non-trivial restriction over generic sets leads to much stronger restrictions over sufficiently larger but finite sets.

Lemma 16. *Suppose that $Y=\{0,1\}$. For any generic S, any $\varepsilon>0$, there is finite $U_\varepsilon\subseteq X$ such that for any theory T, if $\mathcal{E}\left(T;S\right)<1$, then $\mathcal{E}\left(T;U_\varepsilon\right)\leq\varepsilon$. In particular, $\mathcal{E}\left(T\right)=0$ and T is informative.*

The proof of Lemma 16 is an application of the famous Sauer-Shelah Lemma. Suppose that A is a finite set and $\mathcal{M}\subseteq\{0,1\}^A$ is a set of binary functions on A. For any $D\subseteq A$, let $\mathcal{M}_D:=\{\tau|_D:\tau\in\mathcal{M}\}$ be the set of restrictions to set D. Notice that $|\mathcal{M}_D|\leq 2^{|D|}$.

Lemma 17 (Sauer (1972), Shelah (1972), Vapnik and Chervonenkis (1971))**.** *Take any $\mathcal{M}\subseteq\{0,1\}^A$ and suppose that there exists k such that $|\mathcal{M}_D|<2^{|D|}$ for any $D\subseteq A$, $|D|\geq k$.*[7] *Then, $|\mathcal{M}|\leq\sum_{l=1}^{k}\left(\begin{array}{c}|A|\\ l\end{array}\right)$.*[8]

Proof of Lemma 16. We prove Lemma 16. Fix $\varepsilon>0$ and find $\lambda>0$ small enough so that

$$\lambda\log\lambda+(1-\lambda)\log\left(1-\lambda\right)\leq\frac{1}{2}\varepsilon.$$

Find U_ε large enough such that $S\tilde{\subseteq}U_\varepsilon$ for any $D\subseteq U_\varepsilon$, $|D|\geq\lambda\left|U_\varepsilon\right|$. Because $\mathcal{M}\left(T;S\right)<2^{|S|}$, it must be that $\mathcal{M}\left(T;D\right)<2^{|D|}$ for each $D\subseteq U_\varepsilon,|D|\geq\lambda\left|U_\varepsilon\right|$. By Sauer-Shelah's Lemma,

$$\begin{aligned}\mathcal{M}\left(T;U\right)&\leq\sum_{l=1}^{\lambda|U_\varepsilon|}\left(\begin{array}{c}|U_\varepsilon|\\ l\end{array}\right)\leq 2\left(\begin{array}{c}|U_\varepsilon|\\ \lambda|A|\end{array}\right)\\&\leq 2^{2(\lambda\log\lambda+(1-\lambda)\log(1-\lambda))|A|}\leq 2^{\varepsilon|U_\varepsilon|}.\end{aligned}$$

(The second inequality is satisfied for any $\lambda\leq\frac{1}{4}$, and the third is an application of Stirling's inequality.) □

[7] The smallest such k is known as a *VC-dimension* of set $\mathcal{M}$; see Vapnik (1998).

[8] Here, $\left(\begin{array}{c}k\\ l\end{array}\right)$ is the Newton binomial.

Next, we can turn to the proof of Lemma 4 in the general case of finite Y. The idea is to reduce the characterization of theories to the case of the binary space of outcomes Y. For any $\sigma : Y \to \{0,1\}$, any state $\theta : X \to Y$, define a binary σ-representation of state θ as $\theta_\sigma(x) := (\sigma \circ \theta)(x)$. For any finite $A \subseteq X$, let $\mathcal{M}_\sigma(T;A) := \{\theta_\sigma|_A : \theta \in \mathcal{M}(T)\}$. Define $\mathcal{E}_\sigma(T;A) := \frac{1}{|A|}\mathcal{M}_\sigma(T;A)$. Suppose that there exists generic $S \subseteq X$ such that for any $\sigma : Y \to \{0,1\}$, $\mathcal{E}_\sigma(T;S) < 1$. Take any $\varepsilon > 0$ and find finite U_ε from Lemma 16. Then,

$$\begin{aligned}\mathcal{E}(T) \leq \mathcal{E}(T;U_\varepsilon) &= \frac{1}{|U|}\log|\mathcal{M}(T;U_\varepsilon)| \\ &\leq \frac{1}{|U_\varepsilon|}\log\prod_{\sigma:Y\to\{0,1\}}|\mathcal{M}_\sigma(T;U_\varepsilon)| \leq \sum_\sigma \mathcal{E}_\sigma(T;U_\varepsilon) \leq 2^{|Y|}\varepsilon,\end{aligned}$$

and $\mathcal{E}(T) = 0$.

C.3. **Proof of Theorem 1.** For the first part of the Theorem, given Lemma 4, it is enough to show that if a theory is informative, then it has substantive restrictions over some generic set S. On the contrary, suppose that for each generic $S \subseteq X$, there exists $\sigma_S : Y \to \{0,1\}$ so that $\mathcal{E}_{\sigma_S}(T;S) = 1$. Because the language is tight, each finite $A \subseteq X$ has a generic subset $S \subseteq A$ such that $|S| \geq \rho(\mathcal{L})|A|$ where $\rho(\mathcal{L}) > 0$. Then,

$$\begin{aligned}\mathcal{E}(T) \geq \mathcal{E}(T;A) = \frac{1}{|A|}\log\mathcal{M}(T;A) &\geq \frac{1}{|A|}\log\mathcal{M}(T;S) \\ \geq \log\mathcal{M}_{\sigma_S}(T;S) = \frac{|S|}{|A|}\mathcal{E}_{\sigma_S}(T;S) &\geq \rho(\mathcal{L}) > 0.\end{aligned}$$

By Lemma 15, T is not informative.

The second part of the Theorem follows from inequality (C.2).

C.4. **Proof of Corollary 1.** Suppose that the language is complete, transitive, tight and finite. We show that theory T is informative only if it implies an informative theory that consists of one pattern. The other direction is immediate.

Suppose that T is informative. By the Theorem, there exists generic S such that for any $\sigma : Y \to \{0,1\}$ there exists $\tau_\sigma : S \to \{0,1\}$ such that $\tau_\sigma \notin \mathcal{M}_\sigma(T;A)$. Let $\bar{x} \in X^k$ be a k-tuple such that $k = |S|$ and $S = \{x_1, ..., x_k\}$.

Recall that '$\neg$' denotes the logical negation, and '$\neg\neg$' denotes the double negation, i.e., the truth. For any relation $R \in \mathcal{L}$, and any assignment $i : \{1, ..., k_R\} \to \{1, ..., k\}$, define logical symbol

$$(-)^{R,i} := \left\{ \begin{array}{ll} \neg\neg, & \text{if } R\left(x_{i(1)}, ..., x_{i(k_R)}\right), \\ \neg, & \text{if } \neg R\left(x_{i(1)}, ..., x_{i(k_R)}\right). \end{array} \right\}$$

Define free formula $F_{\bar{x}}\left(\tilde{x}_1,...,\tilde{x}_k\right)$ as

$$F_{\bar{x}} := \bigwedge_{R\in\mathcal{L}i:\{1,...,k_R\}\to\{1,...,k\}} \bigwedge (-)^{R,i} R\left(\tilde{x}_{i(1)},...,\tilde{x}_{i(k_R)}\right).$$

Because language $\mathcal{L}$ is finite, $F_{\bar{x}}$ is a well-defined, and $F_{\bar{x}}\left(\bar{x}'\right)$ is true if and only if $\bar{x}'$ is analogous to $\bar{x}$.

For each $\sigma : Y \to \{0,1\}$, define free formula $F_\sigma\left(\tilde{x}_1,...,\tilde{x}_k\right)$ as

$$F_\sigma := \bigvee_{i=1,....,k} \left(\bigwedge_{y\in\sigma^{-1}(\tau_\sigma(i))} \theta\left(x_i\right) \neq y \right)$$

If $F_\sigma\left(x_1,...,x_k\right)$ is true, then $\left(\sigma\circ\theta\right)|_S \neq \tau_\sigma$.

Define free formula $F := F_{\bar{x}} \wedge \bigwedge_\sigma F_\sigma$. Then, if θ is consistent with theory T, then it also is consistent with pattern F. Hence, T implies pattern F. Moreover, for any $\sigma : Y \to \{0,1\}$, $\mathcal{E}_\sigma\left(\{F\};S\right) < 1$, and $\{F\}$ is informative.

Appendix D. Proof of Lemma 14

Lemma 18. *Fix theory T. For each finite $A \subseteq X$, there exists a learning rule π_A such that for any sequence $x_1,...,x_m \in A$*

$$\inf_{\theta\in \mathrm{M}(T)} \sum_{t<m} \pi\left(\left(x_s,\theta\left(x_s\right)\right)_{s\leq m}, x_{s+1}\right)\left(\theta\left(x_{s+1}\right)\right) \geq m - \mathcal{E}\left(T;A\right)|A|.$$

Proof. Fix finite $A \subseteq X$. For all tuples $\bar{x} \in A^k, \bar{y} \in Y^k$, Define learning rule π_A: for each $\bar{x} \in A^k, \bar{y} \in Y^k$, define set

$$\mathcal{M}\left(T;A,\bar{x},\bar{y}\right) = \left\{\tau \in \mathcal{M}\left(T;A\right) : \tau\left(x_l\right) = y_l, l \leq k\right\},$$

and for each $x \in A$, let the prediction $\pi_A\left(\left(x_l,y_l\right)_{l\leq k}, x\right)$ assign a probability 1 to set

$$\arg\max_{y\in Y} \frac{\left|\mathcal{M}\left(T;A,\bar{x}\hat{}x,\bar{y}\hat{}y\right)\right|}{\left|\mathcal{M}\left(T;A,\bar{x},\bar{y}\right)\right|}.$$

Then, for each y so that $\pi_A\left(\left(x_l,y_l\right)_{l\leq k}, y\right) < 1$,

$$\left|\mathcal{M}\left(T;A,\bar{x}\hat{}x,\bar{y}\hat{}y\right)\right| \leq \frac{1}{2}\left|\mathcal{M}\left(T;A,\bar{x},\bar{y}\right)\right|. \tag{D.1}$$

Take any state $\theta \in \mathcal{M}(T)$ and a finite sequence $x_1, ..., x_m \in A$. W.l.o.g. assume that $m \leq |A|$ and all the elements of the sequence are distinct. Then,

$$\begin{aligned}
&\sum_{t<m} \pi\left(\left(x_s, \theta\left(x_s\right)\right)_{s\leq m}, x_{s+1}\right)\left(\theta\left(x\right)\right) \\
&\geq m - \left|\left\{t < m : \pi\left(\left(x_s, \theta\left(x_s\right)\right)_{s\leq m}, x_{s+1}\right)\left(\theta\left(x_{s+1}\right)\right) < 1\right\}\right| \\
&\geq m - \mathcal{E}\left(T; A\right)|A|,
\end{aligned}$$

where the last equality follows from the definition of entropy and the bound (D.1). □

Suppose that $\mathcal{E}(T) = 0$. For each $n > 0$, there exists finite $A_n \subseteq X$ such that $\mathcal{E}(T; A_n) \leq \frac{1}{n}|A|$. Let

$$m_n = \left\lceil \frac{|A_{n+1}|}{|A_n|} \right\rceil \text{ and } M_n = \sum\nolimits_{n'<n} k_{n'}.$$

Using Lemma 10, find disjoint sets A^m such that for each n and $M_n < m \leq M_{n+1}$, A^m is analogous to A_n. For each m, find a learning rule π_{A^m} from Lemma 18. Let

$$t_m = \sum\nolimits_{m'<m} |A^m| + 1.$$

Find an instance process $\bar{x}$ such that for each m, $\left(x_{t_m}, ..., x_{t_{m+1}-1}\right)$ is an enumeration of A^m. Define learning rule π so that for each m, and $t_m \leq t < t_{m-1}$,

$$\pi\left(\left(x_s, y_s\right)_{s<t}, x_t\right) = \pi_{A^m}\left(\left(x_s, y_s\right)_{t_m\leq s<t}, x_t\right).$$

Then, for each $\theta \in \mathcal{M}(T)$,

$$U_t\left(\pi, \bar{x}, \theta\right) \geq 1 - \frac{1}{t}\sum\nolimits_{m'\leq m} \left|A^{m'}\right| \mathcal{E}\left(T; A^{m'}\right).$$

The result follows from the fact that $\sum_{m'\leq m} \left|A^{m'}\right| \leq 2t$ and that $\lim_{m\to\infty} \mathcal{E}\left(T; A^{m'}\right) = 0$.

Appendix E. Proof of Lemma 15

E.1. **Combinatorial result.** The following combinatorial result will be useful. Define function $f : R \times \{1, 2, ...\} \to R$:

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

Lemma 19. *Function $f(p, m)$ is continuous, convex, and increasing in p, and for any $p > 0$, any natural $r \leq m$,*

$$f\left(r, m\right) \leq \left(r + 1\right)|Y|^r \binom{m}{r}. \tag{E.1}$$

By definition, $f(p,1)$ is continuous, convex, and increasing in p. When $m>1$, the continuity, convexity, and monotonicity of $f(p,m)$ follows from the inductive argument.

First, we show that for m,r,

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

Indeed, this is true for $m=1$ by the definition of f. Suppose that the above is true for m. For any r,

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

In particular, this implies that

$$f(m,m)\leq (m+1)\,|Y|^m. \tag{E.2}$$

Notice also that, by definition, for any m

$$f(0,m)=1. \tag{E.3}$$

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

$$f(r,r+k)\leq (r+1)\,|Y|^r \left(\begin{array}{c} r+k \\ r \end{array}\right). \tag{E.4}$$

By (E.2), the statement is true for $k=0$ and any r. Suppose that the statement is true for k and any r. Then,

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

where the inequality follows from the inductive step and (E.3). Suppose now that (E.4) is true for $k+1$ and $r\geq 1$. Then

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

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.

E.2. **Proof of Lemma 15.** Fix a theory T, an instance process $\bar{x}$, a learning rule l, and number of periods t. Define $A_t = \{x_1, ..., x_t\}$ and

- the set of state restrictions that are considered as plausible by the agent, who observed that $\theta(x_1) = y_1, ..., \theta(y_{t-m}) = y_{t-m}$:

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

  Of course,

$$|\mathcal{M}(y_1, ..., y_{t-m})| = \sum_y |\mathcal{M}(y_1, ..., y_{t-m}, y)|,$$

- the worst-case sum of payoffs obtained by learning rule π in periods $s = t-m+1, ..., t$, given that the state restriction belongs to set $\mathcal{M}(y_1, ..., y_{t-m})$:

$$u_m(y_1, ..., y_{t-m}) = \inf_{\tau \in \mathcal{M}(y_1,...,y_{t-m})} \sum_{s=t-m+1}^{t} \pi\left((x_{s'}, \tau(x_{s'}))_{s'<s}, x_s\right)(\tau(x_s)).$$

  Notice that u_m is characterized by a recursive formula:

$$u_m(y_1, ..., y_{t-m}) = \inf_y \left[\pi\left((x_s, y_s)_{s\leq t-m}, x_{t-m+1}\right)(y) + u_{m-1}(y_1, ..., y_{t-m}, y)\right], \tag{E.5}$$

- 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

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

  with a convention that $M_m(u) = 0$ for any $u > m$. Notice that $M_m(u)$ is (weakly) decreasing in u.

Lemma 20. *$M_m(u) \leq f(m-u, m)$ for any u and any $m = 1, ..., t$.*

Proof. By recursive formula (E.5),

$$u_m(y_1, ..., y_{t-m}, y) - \pi\left((x_s, y_s)_{s\leq t-m}, x_{t-m+1}\right)(y) \leq u_{m-1}(y_1, ..., y_{t-m}, y),$$

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

$$\begin{aligned}|\mathcal{M}(y_1, ..., y_{t-m}, y)| &\leq M_{m-1}(u_{m-1}(y_1, ..., y_{t-m}, y)) \\ &\leq M_{m-1}\left(u_m(y_1, ..., y_{t-m}, y) - \pi\left((x_s, y_s)_{s\leq t-m}, x_{t-m+1}\right)(y)\right).\end{aligned}$$

This leads to a recursive bound,

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

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

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

Suppose that the inductive step holds for $m-1$. Then, due to (E.6) applied at $m-1$,

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

Since function $f(x, m-1)$ 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,

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

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

Let $u^* := u_t(\varnothing) = \inf_{\theta \in \mathcal{M}(T)} tU_t(\pi, \bar{x}, \theta)$ be the worst case sum of payoffs from predictions made between periods 1 and t. Then, $\mathcal{M}(T; A_t) = M_t(u^*)$. By Lemmas 19 and 20,

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

Using Stirling's formula, one obtains that

$$\log \begin{pmatrix} t \\ \lceil t-u^* \rceil \end{pmatrix} \leq \lceil t-u^* \rceil \left(\log \frac{t}{\lceil t-u^* \rceil} + 1 \right).$$

Also, $\log(\lceil t-u^* \rceil + 1) \leq \lceil t-u^* \rceil$. Hence,

$$\frac{1}{t}\mathcal{E}(T; A_t) \leq \frac{\lceil t-u^* \rceil}{t} \left(\log |Y| + 2 - \log \frac{\lceil t-u^* \rceil}{t} \right).$$

Because h is increasing, this shows that

$$1 - \frac{u^*}{t} + \frac{1}{t} \geq \frac{\lceil t-u^* \rceil}{t} \geq h\left(\frac{1}{t}\mathcal{E}(T; A_t) \right).$$

This ends the proof of the Lemma.

Appendix F. Proof of Theorem 2

We describe the proof. Fix an informative theory T. Using Lemma 18 in Appendix D, for each finite set $U \subseteq X$, we can find a learning rule π_U such that the number of mistakes committed when predicting the outcomes of instances from set U is not larger than $|U| \mathcal{E}(T;U)$, no matter what is the order of instances.

If set U is local, then we can construct a learning rule $\tilde{\pi}_U$ such that for any finite tuple $(x_1,...,x_m) \tilde{\subseteq} U$, the average number of mistakes committed when predicting the outcomes of instances $x_1,...,x_m$ is not larger than $|U| \mathcal{E}(T;U)$. First notice that because U is local, for all analogous tuples $\bar{x} \sim \bar{u}$ and each instance x, if $\bar{x}\hat{}x \tilde{\subseteq} U$ and $\bar{u} \subseteq U$, then there exists an instance $u \in U$ such that $\bar{x}\hat{}x \sim \bar{u}\hat{}u$. Second, we can use the first observation to find a tuple $\bar{u}(\bar{x}) \subseteq U$ for each tuple $\bar{x} \tilde{\subseteq} U$ so that for each x, if $\bar{x}\hat{}x \tilde{\subseteq} U$, then there exists $u \in U$ such that $\bar{u}(\bar{x}\hat{}x) = \bar{u}(\bar{x})\hat{}u$. Finally, for each $(x_1,...,x_m) \tilde{\subseteq} U$, any $y_1,...,y_{m-1} \in Y$, let

$$\tilde{\pi}_U\left((x_s,y_s)_{s<m},x_m\right) = \pi_U\left((u_s,y_s)_{s<m},u_m\right),$$

where $\bar{u}(x_1,...,x_m) = (u_1,....,u_m)$. The fact that restrictions are analogous (Lemma 9 from Appendix A.1) implies that the learning rule $\tilde{\pi}_U$ has the required property.

Suppose that $\mathcal{U}$ is a family of local sets with the identification property as described before Theorem 2. Then, for each $t<s$, $\bar{x}^t \tilde{\subseteq} U(\bar{x}^s)$, which implies that $U(\bar{x}^t) \tilde{\subseteq} U(\bar{x}^s)$. Moreover, if $U(x^s) \sim U(\bar{x}^t)$, then $U\left(x^{s'}\right) \sim U(\bar{x}^t)$ for each $s \leq s' \leq t$.

By Lemma 3, for each $\varepsilon > 0$, there exists finite A_ε such that the entropy over A_ε is not larger than ε. If an instance process is $\mathcal{U}$-adapted, then for all sufficiently high $t \geq t_\varepsilon$, set A_ε is contained (up to analogy) in $U(\bar{x}^t)$, $A_\varepsilon \tilde{\subseteq} U(\bar{x}^t)$. Lemma 5 shows that the entropy of any theory T over any local set U is an lower bound on the entropy over its subsets $A \subseteq U$. Hence, $\mathcal{E}(T;U(\bar{x}^t)) \leq \varepsilon$.

We construct learning rule π. For each finite tuple $(x_1,...,x_t)$, find the smallest $t_0 \leq t$ such that $U(x_1,...,x_{t_0}) \sim U(x_1,...,x_t)$. For each $y_1,...,y_{t-1}$, let

$$\pi\left((x_s,y_s)_{s<t},x_t\right) = \pi_{U(x_1,...,x_t)}\left((x_s,y_s)_{t_0\leq t<t},x_t\right).$$

Then, the number of mistakes committed by π during periods $t_0 \leq s \leq t$ is not larger than $|U(x_1,...,x_t)| \mathcal{E}(T;U(x_1,...,x_t))$. It follows that the average number of mistakes committed till period t is not larger than

$$\frac{1}{t}\sum\nolimits_{U\in\{U(x_1,...,x_s):s\leq t\}} |U| \mathcal{E}(T;U).$$

If the instance process is $\mathcal{U}$-adapted, then the average number of mistakes converges to 0.

Appendix G. Proofs of Section 8

G.1. **Proof of Lemma 6.** It is sufficient to prove the result for $d = 2$. Fix $\varepsilon > 0$. For each $j = 1, 2$, 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. $\left\{x_1^{*j}, ..., x_{k_j}^{*j}\right\} = S_j$.

Find finite $U_1 \subseteq X_1$ so that $S_1 \tilde{\subseteq} D_1$ for any subset $D_1 \subseteq U_1, |D_1| \geq \frac{\varepsilon}{2} |U_1|$. Define the set of tuples

$$A_1 = \left\{\bar{x}^1 : \bar{x}^1 \text{ is analogous to } \bar{x}^{*1} \text{ and } \bar{x}^1 \subseteq U_1\right\}.$$

Then,

$$\phi_1 := \inf_{D_1 \subseteq U_1, |D_1| \geq \frac{\varepsilon}{2}|U_1|,} \frac{\left\{\bar{x}^1 \in A_1 \cap D_1^k\right\}}{|A_1|} > 0.$$

Find finite $U_2 \subseteq X_2$ so that $S_2 \tilde{\subseteq} D_2$ for any subset $D_2 \subseteq U_2, |D_1| \geq \frac{\varepsilon}{4}\phi |U_1|$. Define the set of tuples

$$A_2 = \left\{\bar{x}^2 : \bar{x}^2 \text{ is analogous to } \bar{x}^{*2} \text{ and } \bar{x}^2 \subseteq U_2\right\}.$$

Then,

$$\phi_2 := \inf_{D_2 \subseteq U_2, |D_2| \geq \frac{\varepsilon}{4}\phi |U_2|,} \frac{\left\{\bar{x}^1 \in A_2 \cap D_2^k\right\}}{|A_2|} > 0.$$

Such sets U_1 and U_2 exist because S_1 and S_2 are generic. Define $U = U_1 \times U_2$.

Take any $D \subseteq U$, $|D| \geq \varepsilon |U| = \varepsilon |U_1| |U_2|$. Define sets $D_1^A \subseteq A_1 \times U_2$ and $D_{12}^A \subseteq A_1 \times A_2$ as follows:

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

We will show that D_{12}^A is not empty, and there exists $\left(x_1^1, ..., x_{k_1}^1, x_1^2, ..., x_{k_2}^2\right) \in D_{12}^A$. Define $S_j' := \left\{x_1^i, ..., x_{k_j}^j\right\} \subseteq U_j$. Then, $S_1' \times S_2'$ is analogous to $S_1 \times S_2$ and $S_1' \times S_2' \subseteq D$. Thus, $S_1 \times S_2 \tilde{\subseteq} D$, and $S_1 \times S_2$ is generic.

For each $x_2 \in U_2$, define

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

Because $|D| \geq \varepsilon |U_1| |U_2|$,

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

For any $\bar{x}_1 \in A_1$, define

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

Then,

$$\left|D_1^A\right| \leq \left\{\bar{x}_1 : \alpha_1(x_1) \geq \frac{\varepsilon}{4}\phi_1\right\}|U_2| + \frac{\varepsilon}{4}\phi_1 |A_1||U_2|, \text{ and}$$
$$\frac{\left|D_1^A\right|}{|U_2||A_1|} - \frac{\varepsilon}{4}\phi_1 \leq \frac{\left\{\bar{x}_1 : \alpha_1(\bar{x}_1) \geq \frac{\varepsilon}{4}\phi_1\right\}}{|A_1|}.$$

Observe that

$$\begin{aligned}
\left|D_1^A\right| &= \sum_{x_2 \in U_2} \left|\left\{\left(\bar{x}^1, x^2\right) : \text{for any } l \leq k_1, \left(x_l^1, x^2\right) \in D\right\}\right| \\
&\geq \sum_{x_2 \in U_2, \alpha_2(x_2) \geq \frac{\varepsilon}{2}} \left|\left\{\bar{x}^1 : x_l^1 \in U_1 \cap \{x_1 : (x_1, x_2) \in D\} \text{ for any } l \leq k_1\right\}\right| \\
&\geq \sum_{x_2 \in U_2, \alpha_2(x_2) \geq \frac{\varepsilon}{2}} \phi_1 |A_1| = \phi_1 |A_1| \left|\left\{x_2 : \alpha_2(x_2) \geq \frac{\varepsilon}{2}\right\}\right| \\
&\geq \left(\frac{|D|}{|U_1||U_2|} - \frac{\varepsilon}{2}\right)\phi_1 |A_1||U_2|
\end{aligned}$$

Similarly,

$$\begin{aligned}
|D_{12}| &= \sum_{\bar{x}^1 \in A_{U_1}(\bar{x}^{*1})} \left|\left\{\left(\bar{x}^1, \bar{x}^2\right) : \text{for any } l \leq k_2, \left(\bar{x}^1, x_l^2\right) \in D_1^A\right\}\right| \\
&\geq \sum_{\bar{x}^1 \in A_{U_1}(\bar{x}^{*1}), \alpha_1(\bar{x}^1) \geq \frac{\varepsilon}{4}\phi_1} \left|\left\{\bar{x}^2 : x_l^2 \in U_2 \cap \left\{x_1 : (\bar{x}_1, x_2) \in D_1^A\right\} \text{ for any } l \leq k_1\right\}\right| \\
&\geq \sum_{\bar{x}^1 \in A_{U_1}(\bar{x}^{*1}), \alpha_1(\bar{x}^1) \geq \frac{\varepsilon}{4}\phi_1} \phi_2 |A_2| \\
&\geq \left(\frac{\left|D_1^A\right|}{|U_2||A_1|} - \frac{\varepsilon}{4}\phi_1\right)\phi_2 |A_2||A_1|
\end{aligned}$$

By the above,

$$\begin{aligned}
\frac{\left|D_1^A\right|}{|U_2||A_1|} &\geq \frac{1}{|U_2||A_1|}\left(\frac{|D|}{|U_1||U_2|} - \frac{\varepsilon}{2}\right)\phi_1 |A_1||U_2| \\
&\geq \left(\frac{|D|}{|U_1||U_2|} - \frac{\varepsilon}{2}\right)\phi_1 \geq \left(\varepsilon - \frac{\varepsilon}{2}\right)\phi_1 = \frac{\varepsilon}{2}\phi_1 > \frac{\varepsilon}{4}\phi_1.
\end{aligned}$$

Thus,

$$|D_{12}| \geq \frac{\varepsilon}{4}\phi_1\phi_2 |A_2||A_1| > 0$$

and D_{12} is not empty.

G.2. **Proof of Lemma 7.** Take any set $S = B^1 \times ... \times B^d$, where B^i are finite and disjoint subsets of B. We show that S is generic. Take any infinite and disjoint sets $\hat{B}^1, ..., \hat{B}^d \subseteq B$ such that $B^j \subseteq \hat{B}^j$ for $j \leq d$. Let $\hat{X} = \hat{B}^1 \times ... \times \tilde{B}^d$. Consider language $\mathcal{L}_{\text{choices}}$ restricted to $\hat{X}$. It is easy to notice that such language $\mathcal{L}_{\text{choices}}$ is isomorphic to the product of trivial languages on sets $\hat{B}^j, j \leq d$ (see Section 8.2). By Corollary 2, set $S \subseteq \hat{X}$ is generic in the product language. Because of isomorphism, set S is also generic in language $\mathcal{L}_{\text{choices}}$.

Next, we show that for any finite $A \subseteq X$, there is a generic $S \subseteq A$ such that $|S| \geq d^{-d} |A|$. Let $B_1 \subseteq B_2 \subseteq ... \subseteq B$ be an increasing sequence of subsets of B such that $|B_n| = nd$. Let $U_n \subseteq X$ consists of all tuples of elements of B_n. Then, sequence U_n is increasing and there exists n high enough, so that $A \tilde{\subseteq} U_n$. Let $B_n^1, ..., B_n^d \subseteq B_n$ be disjoint subsets of B_n such that $|B_n^j| = d$ for each j. Let $S^* = B_n^1 \times ... \times B_n^d \subseteq U_n$ and observe that

$$\frac{|S^*|}{|U_n|} \geq d^{-d}. \tag{G.1}$$

We show that there exists an analogous copy S of S^* such that $|S \cap A| \geq d^{-d} |A|$ (notice that, by the above, S^*, hence S and any of its subsets are generic.) Indeed, let Π be a collection of all subsets of U_n that are analogous to S. Consider a uniform distribution on Π. Because of (G.1), for each x, the probability that x belongs to randomly chosen set $S \in \Pi$ is not smaller than d^{-d}. It follows, that there exists at least one realization of S such that $|S \cap A| \geq d^{-d} |A|$.

University of Texas at Austin, Department of Economics
E-mail address: mpeski@gmail.com