ABSOlutely No Free Lunches!

Gordon Belot

Abstract. This paper is concerned with learners who aim to learn patterns in infinite binary sequences: shown longer and longer initial segments of a binary sequence, they either attempt to predict whether the next bit will be a 0 or will be a 1 or they issue forecast probabilities for these events. Several variants of this problem are considered. In each case, a no-free-lunch result of the following form is established: the problem of learning is a formidably difficult one, in that no matter what method is pursued, failure is incomparably more common that success; and difficult choices must be faced in choosing a method of learning, since no approach dominates all others in its range of success. In the simplest case, the comparison of the set of situations in which a method fails and the set of situations in which it succeeds is a matter of cardinality: in other cases, it is a topological matter (meagre vs. co-meagre), or a hybrid computational-topological matter (effectively meagre vs. effectively co-meagre, in the sense of Mehlhorn (1973)).

1. Introduction

The various no-free-lunch theorems of statistical, computational, and formal learning theory offer ways to make precise the basic insight that there can be no optimal general-purpose approach to learning. These theorems come in two main forms. Some show that there are contexts in which certain approaches to learning succeed in each salient situation, but that each such approach has the same expected performance across those possible situations. Results of this kind are measure-relative: in order for expectations to be defined, a measure must be imposed on the space of situations that a learner might face—and the results in question only hold relative to some of the measures that one might impose. Results of a second kind are absolute in the sense that they do not rely upon the choice of a measure on the space of envisaged situations. Here are descriptions of two paradigmatic results of this kind.

Maybe there exists some kind of universal learner, that is, a learner who has no prior knowledge about a certain task and is ready to be challenged by any task? ... The no-free-lunch theorem states that no such universal learner exists. To be more precise, the theorem states that for binary classification prediction tasks, for every learner there exists a distribution on which it fails. ... In other words, the theorem states that no learner can succeed on all learnable tasks—every learner has tasks on which it fails while other learners succeed.

Let T be any learning machine.... [W]e will defeat the machine T. That is,
we will have constructed a regularity, depending on $T$, which is beyond the power of $T$ to extrapolate. However . . . it is always possible to build another machine which can extrapolate every regularity that $T$ can extrapolate and also extrapolate the one that $T$ can’t extrapolate. Thus, there cannot exist a cleverest learning machine: for, for every learning machine $T$, there exists a machine $T'$ which can learn everything that $T$ can learn and more besides.

We can think of such results as encapsulating two facts about the predicament of learners situated in certain contexts. (a) They face a difficult problem: no approach they might adopt succeeds across all envisaged situations. (b) Difficult choices must be made: different approaches succeed in different situations, with no approach dominating all others in its range of success.

Here we assemble some more or less elementary results, some already well-known, that combine to give no-free-lunch results of the second, absolute variety, applicable to agents attempting to learn patterns in infinite data streams of bits. Nature presents our agents with one-way infinite binary sequences one bit at a time and after each bit is revealed each agent is asked to make a prediction about the next bit. We will consider five models of learning, differing from one another as to what sort of predictions our agents are required to make or as to the criterion of success. And for each model, we will consider variants in which neither the agent nor Nature is required to follow a computable strategy, in which the agent is required to follow a computable strategy but Nature is not, and in which both the agent and Nature are required to follow computable strategies. For each variant of each model, we establish both elements required for a no-free-lunch result. (i) Difficult choices must be faced in selecting a method of learning: we will show that no approach dominates all its rivals, either by showing that for each method there is another that succeeds in a disjoint set of situations (evil twin results) or by showing that for every method there is another that succeeds in a strictly larger family of situations (better-but-no-best results). (ii) We also show our learners face a formidably difficult problem: for each of the problems we consider, there is a sense in which for any method of addressing that problem, the situations in which it fails are incomparably more common than the situations in which it succeeds (so here we go beyond the paradigm results mentioned above, which show only that for each method, there exists a situation in which it fails).

Following some preliminaries in Section 2, we investigate in Section 3 the predicament of learners who must attempt to guess, before each bit is revealed, whether it will be a 0 or a 1. We will consider two criteria of success for such next-value learners: when facing a given data stream they should eventually predict each new bit correctly (NV-learning); or the should predict each new bit correctly, except for a family of errors that has vanishing asymptotic density (weak NV-learning).

---

3Putnam (1963b, 6f.)—see also Gold (1967, Theorem I.5). As noted by Case and Smith (1983, 208): “This appears to be the earliest result indicating that there may be some difficulty with the mechanization of science.”

4A problem that, according to Li and Vitányi (2019, 6), “constitutes, perhaps, the central task of inductive reasoning and artificial intelligence.”

5For other results of the sort developed here, see Fortnow et al. (1998).

6For early investigations of such agents, see Putnam (1963a) and Gold (1967).

7The notion of NV-learning is due to Bárzdins (1972); see also Blum and Blum (1975). The notion of weak NV-learning is a near relative of the notion of coarse computability introduced in Jockusch and Schupp (2012).
arbitrary sequences, failure is incomparably more common than success in the sense that any method for predicting bits succeeds for a countable family of binary sequences and fails for an uncountable family of binary sequences. In Section 3.2 we will see that for weak NV-learning of arbitrary sequences, any method succeeds for an uncountable set of sequences and fails for an uncountable set of sequences, but the successes are always incomparably less common than the failures in a topological sense, forming a meagre set. In Section 3.3 we restrict attention to computable methods for the next-value learning of computable sequences and find that for any method, the sets of success and failures are equivalent both from the point of view of cardinality and the point of view of topology—but that the successes are nonetheless incomparably less common than the failures in the hybrid topological-computational sense (due to Mehlhorn (1973)) that they form an effectively meagre set.

Along the way we will see that the notion of weak NV-learning, while strictly more inclusive than the notion of NV-learning, is neither weaker nor stronger than two other variants of NV-learning, NV'-learning (due to Barzdins and Freivalds (1972)) and NV''-learning (due to Podnieks (1974)).

In Section 4 we turn to agents who face a data stream sampled from a probability measure chosen by Nature and who are required to issue forecast probabilities for the next bit’s being a 0 or a 1 just before it is revealed. We consider three criteria of success for agents engaged in such next chance prediction: we can ask that for any event, the probabilities that our agents assign to that event converge almost certainly to the true probability as they see larger and larger data sets (strong NC-learning); we can ask that their forecast probabilities for the next bit become arbitrarily accurate, almost certainly, in the limit of large data sets (NC-learning); or we can ask that they, almost certainly, meet the last-mentioned standard modulo a set of missteps of vanishing asymptotic frequency (weak NC-learning). For the problem of next-chance learning in the face of a data stream generated by an arbitrary measure, we see in Sections 4.1 and 4.2 that for any of our criteria of success, each method fails for an uncountable set of measures that Nature might have chosen and succeeds for an uncountable set of such measures—but that the former set is always incomparably smaller than the latter, being meagre. In 4.3 we restrict attention to computable strategies for next-chance learning in contexts in which the data stream is generated by a computable measure and find, for each of our three criteria of success, that the set of learnable measures is an effectively meagre subset of the family of computable measures. Section 5 provides a few concluding remarks.

2. Preliminaries

2.1. The Main Characters

We will be concerned below with a number of topological spaces.

(i) The space of bits, \( \mathcal{B} := \{0, 1\} \), equipped with the discrete topology (so every subset of \( \mathcal{B} \) is open).

(ii) Finite products of \( \mathcal{B} \) with itself: for each \( n \in \mathbb{N} \), the space of \( n \)-bit strings, \( \mathcal{B}^n \), equipped with the discrete topology (we count 0 as a natural number and use \( \emptyset \) to denote either the empty string of zero bits that is the sole member of \( \mathcal{B}^0 \) or the empty set, depending

---

8In the case of NV-learning, this result is due to Fortnow et al. (1998).
9For this model of learning agents, see Solomonoff (1964).
10These criteria of success derive from Blackwell and Dubins (1962), Kalai and Lehrer (1994), and Lehrer and Smorodinsky (1996). The criteria of success employed in the literature on Solomonoff induction differ in focussing on average or expected performance in the long run—on the relation between those notions and the notion of NC-learning, see Ryabko and Hutter (2007).
on context). We will think of elements of $\mathcal{B}^n$ as strings (concatenations of symbols) rather than as $n$-tuples. For $w \in \mathcal{B}^n$ and $m \leq n$ we write $w(m)$ for the $m$th bit of $w$ and write $w[m]$ for the $m$-bit initial segment of $w$. For $m, n \in \mathbb{N}$ with $m \leq n$, we have the natural projection map $\pi_{nm} : w \in \mathcal{B}^n \mapsto w[m] \in \mathcal{B}^m$.

(iii) The space of binary strings, $\mathcal{B}^* := \bigcup_{n=0}^{\infty} \mathcal{B}^n$, also equipped with the discrete topology. If $v$ and $w$ are binary strings we write $v.w$ for the string that results from concatenating $v$ and $w$ (in that order) and write $v.w^2$ for the results of concatenating $v$ with $w$ and with $w$, etc. We write $|w|$ for the number of bits in binary string $w$.

(iv) Cantor space, $\mathcal{C}$, the set of all infinite binary sequences equipped with the product topology (we take sequences to be indexed by positive natural numbers). We can characterize this topology as follows: if $w$ is an $n$-bit string, then we use $B_w$ to denote the set of sequences whose first $n$ bits are given by $w$; the set of all such $B_w$ (as $w$ ranges over $\mathcal{B}^*$) is a basis for the product topology and we call the $B_w$ basic open sets. Illustration: the set of sequences that have 0 as their second bit is an open set because it is the union of the basic open sets $B_{00}$ and $B_{01}$. For $\sigma \in \mathcal{C}$ we write $\sigma(m)$ for the $m$th bit of $\sigma$ and write $\sigma[m]$ for the $m$-bit string formed by concatenating the first $m$ bits of $\sigma$. For each $n \in \mathbb{N}$ we have the natural projection map $\pi_n : \sigma \in \mathcal{C} \mapsto \sigma[n] \in \mathcal{B}^n$. A sequence of points $\sigma_1, \sigma_2, \sigma_3, \ldots$ in Cantor space converges to $\sigma \in \mathcal{C}$ if and only if for each $k$, there exists an $N$ so that for $n \geq N$, $\sigma_n(k) = \sigma(k)$. We use $\mathfrak{B}$ to denote the $\sigma$-algebra of Borel subsets of $\mathcal{C}$. We use $\mathcal{C}$ to denote the subspace of $\mathcal{C}$ consisting of computable sequences (i.e., the $\sigma$ such that the map $k \in \mathbb{N} \mapsto \sigma[k]$ is computable).

(v) For each $k \in \mathbb{N}$, the space $\mathcal{P}_k$ of Borel probability measures on $\mathcal{B}^k$. Since $|\mathcal{B}^k| = 2^k$, we can identify any $\mu \in \mathcal{P}_k$ with a $2^k$-tuple of real numbers in the closed unit interval that sum to one. We take $\mathcal{P}_k$ to be equipped with the topology that it inherits from being embedded in this way as a closed subset of $\mathbb{R}^{2^k}$ (which we take to be equipped with its standard topology). We call $\mu \in \mathcal{P}^m$ and $\nu \in \mathcal{P}^n$ with $m \leq n$ consistent if for each subset $A$ of $\mathcal{B}^m$ we have $\mu(A) = \nu(\pi_{nm}^{-1}(A))$.

(vi) The space $\mathcal{P}$ of Borel probability measures on $\mathcal{C}$ equipped with the weak topology, which can be characterized as follows. For each binary string $w$ and each pair of numbers $p$ and $q$ in the closed unit interval with $p < q$, let

$$S_{w,p,q} := \{ \mu \in \mathcal{P} : p < \mu(B_w) < q \}.$$

The set of all such $S_{w,p,q}$ forms a sub-basis for the weak topology on $\mathcal{C}$: the open sets of the weak topology are arbitrary unions of finite intersections of these sub-bases. Under the weak topology, a sequence $\{\mu_k\}$ of measures in $\mathcal{P}$ converges to $\mu \in \mathcal{P}$ if and only if $\lim_{k \to \infty} \mu_k(B_w) = \mu(B_w)$ for each $w \in \mathcal{B}^*$.

---

11For the weak topology on spaces of measures on metric spaces, see Parthasarathy [1967, chap. II] and Billingsley [1999, §1.2]. For the special case of $\mathcal{P}$, see Kechris [1995, chap. 17] and Reimann [2008, §2.5]. Since we can specify a Borel probability measure on $\mathcal{C}$ by specifying the weight that it assigns to each binary string, by fixing an enumeration of the binary strings we can identify each $\mu \in \mathcal{P}$ with a sequence of numbers in the closed unit interval that sum to one. In this way we identify $\mathcal{P}$ with a closed subset of the Hilbert cube ($=[0,1]^\mathbb{N}$ equipped with the product topology). The weak topology is the topology that $\mathcal{P}$ inherits from this embedding.

12Each $B_w$ is a clopen subset of $\mathcal{C}$ and so is a continuity set for any measure in $\mathcal{P}$. So the Portmanteau Theorem implies that the above condition is necessary for weak convergence. And it is also sufficient, since the $B_w$ form a countable basis for $\mathcal{C}$ closed under finite intersections. See, e.g., Billingsley [1999, Theorems 2.1 and 2.2].
Below, in order to simplify notation, for $\mu$ a measure in $\mathcal{P}$ and $w$ a binary string, we will write $\mu(w)$ in place of $\mu(B_w)$. Using this notation, the Carathéodory Extension Theorem tells us that any map $\bar{\nu} : \mathcal{B}^* \rightarrow [0, 1]$ such that $\bar{\nu}(\emptyset) = 1$ and such that $\bar{\nu}(w) = \bar{\nu}(w.0) + \bar{\nu}(w.1)$ for each $w \in \mathcal{B}^*$ induces a unique $\nu \in \mathcal{P}$ such that $\nu(w) = \bar{\nu}(w)$ for all $w \in \mathcal{B}^*$.

As usual, we consider $\nu \in \mathcal{P}$ to be computable if and only if there exists a computable $F : \mathcal{B}^* \times \mathbb{N} \rightarrow \mathbb{Q}$ such that $|\nu(w) - F(w, n)| < 2^{-n}$ for all $w \in \mathcal{B}^*$ and $n \in \mathbb{N}$. We use $\mathcal{P}$ to denote subspace of $\mathcal{P}$ consisting of computable measures.

Remark 2.1. $\mathcal{B}^*$, $\mathcal{C}$, $\mathcal{P}$, each of the $\mathcal{B}^k$, and each of the $\mathcal{P}_k$ are compact, separable and completely metrizable. In $\mathcal{B}$, the other $\mathcal{B}^n$, and in $\mathcal{B}^*$ each point is isolated (i.e., for any point, the singleton set containing that point is open). There are no isolated points in $\mathcal{C}$, the $\mathcal{P}_k (k > 0)$, or $\mathcal{P}$.

Remark 2.2. $\mathcal{C}$ and the $\mathcal{P}_k (k > 0)$ of course have cardinality $\mathfrak{c}$ (the cardinality of the continuum). So does $\mathcal{P}$: since $\mathcal{P}$ is non-empty, compact, and metrizable there is a continuous map from $\mathcal{C}$ onto $\mathcal{P}$; since $\mathcal{P}$ is a non-empty, separable, and completely metrizable space without isolated points, there is an embedding of $\mathcal{C}$ into $\mathcal{P}$.

2.2. **The Meagre & the Co-Meagre**

We are going to be interested in making comparisons of size for certain subsets of $\mathcal{C}$ and $\mathcal{P}$. The most straightforward standard of comparison is cardinality: it natural to say that any uncountable set is incomparably larger than any countable set.

Below we will see examples where the set of learnable sequences or measures and the set of unlearnable sequences or measures have the same cardinality—but in which it is intuitively natural to say that the unlearnable sequences or measures are incomparably more common than the learnable sequences or measures.

The intuitive notions of size in play here correspond nicely with the topologists’ notions of meagre and co-meagre subsets of a topological space. Recall that a nowhere dense subset of a topological space is one whose closure has empty interior—or, equivalently, a subset $A$ of a topological space $X$ is nowhere dense if and only if for any non-empty open set $U \subset X$, there exists a non-empty open set $U^* \subset U$ with $A \cap U^* = \emptyset$. And recall that a meagre subset of a topological space is one that can be written as a countable union of nowhere dense sets while a co-meagre subset is one that is the complement of a meagre set.

For any topological space $X$, the class of meagre subsets of $X$ is closed under the operations of taking subsets and taking countable unions. The Baire Category Theorem tells us that in a completely metrizable space, no non-empty open set is meagre. So, in particular, no non-empty completely metrizable space has any subsets that are both meagre and co-meagre.

The results just mentioned motivate the standard practice in topology, analysis, and related mathematical fields of considering the elements of a meagre subset of a completely metrizable space to be extremely rare and the elements of the complement of such a set to be exceedingly common—so that objects that form a co-meagre set are often referred to as being typical. Illustration: one says that typical continuous functions on the unit interval are nowhere

---

13 See, e.g., Nies (2009, §1.9).
14 Kechris (1995, Theorems 4.18 and 6.2)
15 If $A \subset X$ were both meagre and co-meagre, then so would be its complement. But then $X$ could be written as a union of two meagre sets—which is impossible if no non-empty subset of $X$ is meagre.
differentiable because the nowhere differentiable functions form a co-meagre subset of the space of continuous functions under the uniform topology.

**Remark 2.3 (The Banach–Mazur Game.).** Here is an additional compelling rationale for this practice. Fix a subset $S$ of $C$. An infinite two-player game is to be played. In the first round, Player I selects a non-empty binary string $v_1$, then Player II selects a non-empty binary string $w_1$; and similarly in each subsequent round, Player I selects a non-empty binary string $v_k$, then Player II selects a non-empty binary string $w_k$. Player I wins the game if the infinite binary sequence $v_1.w_1.v_2.w_2.\ldots$ is in $S$, otherwise Player II wins. Intuitively, if Player I has a winning strategy for the the Banach–Mazur game for $S$, then $S$ must be overwhelmingly large as a subset of $X$, while if Player II has a winning strategy, then $S$ must be nigh ignorably small as a subset of $X$.

The intuitive notions of small and large subsets appealed to here correspond precisely to the notions of meagre and co-meagre subsets: Player I has a winning strategy if and only if $S$ is co-meagre in some open subset of $C$; Player II has a winning strategy if and only if $S$ is meagre as a subset of $C$.

### 3. Extrapolation

Think of Nature as having chosen a binary sequence, which is now being revealed to a learning agent one bit at a time. After each new bit is presented, the agent attempts to predict what the next value will be on the basis of the data seen so far. The agent succeeds in this task if from a certain point onwards, the predictions made match reality (almost perfectly).

**Definition 3.1 (Extrapolators).** An *extrapolator* is a function $m : B^* \to B$. We denote the set of extrapolators by $\mathcal{E}$.

**Definition 3.2 (Extrapolating Machines).** An *extrapolating machine* is a computable extrapolator—i.e., a computable function $m : w \in B^* \mapsto m(w) \in B$. We denote the set of extrapolating machines by $\mathcal{E}$.

**Definition 3.3 (NV-Learning).** Let $m$ be an extrapolator and $\sigma$ a binary sequence. We say that $m$ *NV-learns* $\sigma$ (or that $\sigma$ is *NV-learnable* by $m$) if there is an $N$ such that for all $n > N$, $m(\sigma[n]) = \sigma(n + 1)$.\[^1\][^16]

Jockusch and Schupp (2012, p. 472) remark that “In recent years, there has been a general realization that worst-case complexity measures, such as $P$, $NP$, exponential time, and just being computable, often do not give a good overall picture of the difficulty of a problem.” As an example, they observe that although there exist finitely presented groups with unsolvable word problems, in every such group the words expressing the identity have vanishing asymptotic density, when words are enumerated in lexicographic order. So the linear-time algorithm that on the input of any word guesses that that word does not express the identity would make a negligible set of errors if fed all words in lexicographic order. If we demand perfection,\[^1\][^17]

---

\[^1\]Here we have described a special version of the game adapted to $C$. A more general version makes sense in any topological space $X$ and we always have the connection between meagreness and winning strategies for Player II; the connection between co-meagreness and winning strategies for Player I requires some additional hypotheses in the general setting. See Oxtoby (1957) and Kechris (1995, §§8H and 21.C).

\[^16\]The notion of NV-learning for extrapolating machines is due to Barzdins (1972); see also Blum and Blum (1975).
then the word problem is impossibly hard—but if we can live with making mistakes a negligible fraction of the time, it is as easy as could be. This motivates Jockusch and Schupp to introduce a generalization of the notion of computability: a sequence is coarsely computable if it differs from some computable sequence in a set of bits of vanishing asymptotic density. We will consider here a notion of learning that stands to NV-learning as coarse computability stands to computability (see Remark 3.4 for more on the relation to coarse computability).

Definition 3.4 (Weak NV-Learning). We say that extrapolator $m$ weakly NV-learns binary sequence $\sigma$ (or that $\sigma$ is weakly NV-learnable by $m$) if:

$$\lim_{n \to \infty} \frac{|\{k \leq n : m(\sigma[k]) = \sigma(k + 1)\}|}{n} = 1.$$ 

For $m$ an extrapolator and $\sigma$ a binary sequence, we say that according to $m$, $n \in \mathbb{N}$ corresponds to a good bit of $\sigma$ if $m(\sigma[n]) = \sigma(n + 1)$ and corresponds to a nasty bit of $\sigma$ if $m(\sigma[n]) \neq \sigma(n + 1)$. To say that $m$ NV-learns $\sigma$ is to say that according to $m$, $\sigma$ eventually consists of nothing but good bits. To say that $m$ weakly NV-learns $\sigma$ is to say that according to $m$, although $\sigma$ may contain infinitely many nasty bits, these have vanishing limiting relative frequency. We will consider each of these two criteria of learning in turn.

3.1. NV-Learning

Officially, the job of an extrapolator is to predict the next bit on the basis of the current data set. But we can also think of an extrapolator $m$ as a means of guessing the entire data sequence on the basis of any initial segment.\(^\text{18}\)

Definition 3.5. For $m \in \mathcal{M}$ and $w \in \mathcal{B}^n$, we use $\sigma^w_m$ to denote the sequence defined as follows:

- For $k = 1, \ldots, n$, $\sigma^w_m(k) = w(k)$ (i.e., $\sigma^w_m[n] = w$).
- $\sigma^w_m(n + 1) = m(w)$;
- $\sigma^w_m(n + \ell) = m(w.\sigma^w_m(n + 1).\ldots.\sigma^w_m(n + \ell - 1))$ ($\ell = 2, 3, \ldots$).

We say that $m$ guesses $\sigma^w_m$ on input $w$.

Note that if $m \in \mathcal{E}$ then for any $w \in \mathcal{B}^*$, $\sigma^w_m \in \mathcal{E}$: on any input, an extrapolating machine guesses a computable sequence.

Trivially, there is an equivalence between the sequences NV-learned by an extrapolator and the sequences guessed by it.\(^\text{19}\)

Proposition 3.1. Extrapolator $m$ NV-learns sequence $\sigma$ if and only if $\sigma = \sigma^w_m$ for some $w \in \mathcal{B}^*$.

Proof. Suppose that $m$ NV-learns $\sigma$. Then there is an $n_0$ such that for all $n \geq n_0$, $m(\sigma[n]) = \sigma(n + 1)$. So $m$ guesses $\sigma$ on input $w = \sigma[n_0]$. Suppose, on the other hand, there is an $n_0$ such that $m$ guesses $\sigma$ on input $w = \sigma[n_0]$. Then $m$ NV-learns $\sigma$, since for all $n > n_0$, $m(\sigma[n]) = \sigma(n + 1)$.

So asking that $m$ eventually correctly predict next bits is equivalent to asking that $m$ eventually be able to answer correctly all questions about the data stream.

\(^{18}\)On this point, see, e.g., Angluin and Smith (1983 §4.2.1).

\(^{19}\)This is related to the deeper fact that $\mathcal{N}V = \mathcal{PEX}$—see Theorem 2.19 of Case and Smith (1983), where the result is attributed to private communications from van Leeuwen and from Bárzdziūš.
Proposition 3.2. For any extrapolator \( m \), the sequences NV-learnable by \( m \) form a countably infinite set dense in \( C \) while the sequences not NV-learnable by \( m \) form a dense subset of \( C \) of cardinality \( \mathfrak{c} \).

Proof. On the one hand, \( B^* \) is a countable set and the preceding proposition tells us that the map \( w \in B^* \mapsto \sigma_m^w \) has as its range the set of sequences NV-learnable by \( m \). So this set is countable. And since for any \( w \in B^* \), \( \sigma_m^w \in B_w \) the set of NV-learnable sequences is dense in \( C \) (and is therefore infinite). On the other hand, each \( B_w \) has cardinality \( \mathfrak{c} \) but contains only countably many binary sequences NV-learnable by \( m \). \( \square \)

Corollary 3.1. The set \( \{ \sigma \in C \mid \exists m \in \mathcal{E} \text{ such that } m \text{ NV-learns } \sigma \} \) is countable.

As usual, we call a sequence \( \{ \sigma_i \}_{i \in \mathbb{N}} \) of elements of \( \mathcal{C} \) uniformly computable in \( i \) if there is a computable \( f : \mathbb{N} \times \mathbb{N} \to \mathcal{B} \) such that \( f(i, j) = \sigma_i(j) \), for all \( i, j \in \mathbb{N} \).

Proposition 3.3. (a) Let \( m \in \mathcal{E} \) and \( S \) be a countable subset of \( C \). Then there is an \( m^* \in \mathcal{E} \) that NV-learns every \( \sigma \in S \) as well as everything NV-learned by \( m \). (b) Let \( m \in \mathcal{E} \) and let \( S = \{ \sigma_i \}_{i \in \mathbb{N}} \) be a family of elements of \( \mathcal{C} \) uniformly computable in \( i \). Then there is an \( m^* \in \mathcal{E} \) that NV-learns every \( \sigma \in S \) as well as everything NV-learned by \( m \).

Proof. We present the argument for (b)—essentially the same argument works for (a).

Define \( \tilde{m} \in \mathcal{E} \) as follows: on input of \( w \in B^* \), \( \tilde{m} \) finds \( K = \{ k \in \mathbb{N} \mid 1 \leq k \leq n, \sigma_k[n] = w \} \); if \( K \neq \emptyset \), then \( \tilde{m}(w) = \sigma_\ell(n + 1) \), where \( \ell \) is the least element of \( K \); otherwise, \( \tilde{m}(w) = m(w) \).

Define \( m^* \) as follows: \( m^* \) has a counter that keeps tally of how many incorrect prediction have been made in the course of processing a given data stream; in processing input \( w \in B^* \), \( m^* \) simulates \( m \) if an even number of incorrect predictions have been made and simulates \( \tilde{m} \) if an odd number have been made.

Clearly, \( m^* \) is an extrapolating machine. Suppose that \( m^* \) is shown a data stream \( \sigma \) that it does not NV-learn. Then \( m^* \) must make infinitely many incorrect predictions in processing \( \sigma \). So \( \sigma \) cannot be a sequence NV-learned by \( m \): any such sequence is guessed by \( m \) when it sees sufficiently long initial segments. Similarly, \( \sigma \) cannot be any of the \( \sigma_k \), since each of these is guessed by \( m \) when it see sufficiently long initial segments. \( \square \)

Proposition 3.4. Let \( m \) be an extrapolator and let \( S \subset C \) be the set of sequences that it NV-learns. Then there is an extrapolator \( m^! \) such that the set \( S^! \) of sequences that it NV-learns is disjoint from \( S \)—and where \( m^! \) is in \( \mathcal{E} \) if \( m \) is.

Proof. Define \( m^! \) by setting \( m^!(w) = 1 - m(w) \) for each \( w \in B^* \). \( \square \)

So we have both elements required for the sort of no-free-lunch result we seek. The problem of NV-learning is a formidably difficult one: each (computable) extrapolator fails to NV-learn incomparably more sequences that it NV-learns: the set on which it succeeds is countable (and hence meagre), so the set on which it fails is uncountable (indeed, co-meagre). And there are hard choices to be made: for any (computable) extrapolator, there is another that NV-learns sequences that the first cannot NV-learn. There is no optimal method of extrapolation.

3.2. Weak NV-Learning

If an extrapolator NV-learns a sequence, then it also weakly NV-learns it. But the converse is not true.
Example 3.1. Consider the extrapolating machine $m_1$ that outputs 1 on any input. This machine NV-learns all and only sequences that are eventually all 1’s—a countably infinite set. But $m_1$ weakly NV-learns continuum-many sequences. For, let $\sigma$ be an arbitrary binary sequence and let $\hat{\sigma}$ be the sequence defined as follows: for $n = 1, 2, \ldots$, if $k = 10^n$, then $\sigma(k) = \hat{\sigma}(n)$; otherwise, $\sigma(k) = 1$. According to $m_1$, the nasty bits of $\sigma$ have vanishing asymptotic density, so $m_1$ weakly NV-learns $\sigma$. And there are continuum-many $\sigma$ we could use as input for this construction, each determining a distinct sequence weakly NV-learned by $m_1$. Note that there is no input on which $m_1$ guesses a sequence that contains infinitely many 0’s, although it weakly NV-learns uncountably many sequences with this feature. Note also that although $m_1$ is computable, it weakly NV-learns uncountably many uncomputable sequences and weakly NV-learns sequences of arbitrary Turing degree.

Proposition 3.5. Each extrapolator weakly NV-learns a dense set of sequences of cardinality $\mathfrak{c}$ and fails to weakly NV-learn a dense set of sequences of cardinality $\mathfrak{c}$.

Proof. Let $m$ be an extrapolator, $w$ an $n$-bit binary string, and $\hat{\sigma}$ an arbitrary sequence. We construct sequences $\sigma^*$ and $\sigma^\dagger$ as follows:
- For $k = 1, \ldots, n$, $\sigma^*(k) = \sigma^\dagger(k) = w(k)$.
- For $k = n + 10^\ell$ ($\ell = 1, 2, \ldots$), $\sigma^*(k) = \sigma^\dagger(k) = \hat{\sigma}(\ell)$.
- For all other $k$, $\sigma^*(k) = m(\sigma^*(1), \sigma^*(2), \ldots, \sigma^*(k - 1))$ and $\sigma^\dagger(k) = 1 - \sigma^*(k)$.

According to $m$, any nasty (good) bits in $\sigma^*$ ($\sigma^\dagger$) occur with indices of the form $n + 10^\ell$. So $m$ weakly NV-learns $\sigma^*$ and fails to weakly NV-learn $\sigma^\dagger$. By varying $w$, we obtain weakly NV-learnable and not weakly NV-learnable sequences in each basic open set of $\mathcal{C}$. And by varying $\sigma$ we obtain continuum-many sequences of each type.

So for any extrapolator, there are continuum-many sequences that it can weakly NV-learn and continuum-many sequences that it cannot weakly NV-learn. But, intuitively, there is a sense in which it is much more difficult to construct a sequence weakly NV-learnable by a given extrapolator than it is to construct a sequence that is not weakly NV-learnable by that extrapolator. Consider again the extrapolator $m_1$ that outputs 1 on any input. In order to construct a sequence that this extrapolator weakly NV-learns, you begin with the all 1’s sequence, then sprinkle in some 0’s, subject to the constraint that the set of indices of the slots containing 0’s has vanishing asymptotic density in $\mathbb{N}$. In order to construct a sequence that this extrapolator can’t weakly NV-learn, you begin with the all 1’s sequence and sprinkle in as many 0’s as you like, just being careful to make sure that the set of indices of the slots containing 0’s doesn’t have vanishing asymptotic density. The latter task, is intuitively easier: e.g., because there are a lot more densities not equal to zero than equal to zero. This intuition is borne out by the following result.

Proposition 3.6. Let $m$ be any extrapolator. The sequences weakly NV-learnable by $m$ form a meagre subset of $\mathcal{C}$.

Proof. Let us say that binary string $w$ is wicked according to $m$ if at least half of the bits of $w$ are nasty according to $m$. For each $n \in \mathbb{N}$, let $A_n$ be the set of sequences that do not have at least $n$ initial segments that are wicked according to $m$.

We claim that each $A_n$ is nowhere dense. To establish this, it suffices to show that for any binary string $w$, there is another, $w^*$, depending on $n$ and $w$, such that $w^*$ extends $w$ and $B_{w^*} \cap A_n = \emptyset$. To this end, let $w$ be a string and let $w^*$ be the result of extending $w$ by $|w|$ bits that are nasty according to $m$, then tacking on $n$ more nasty bits. Every sequence in
$B_{w^*}$ then has at least $n$ initial segments that are wicked according to $m$.

So $A := \bigcup_{n=1}^{\infty} A_n$ is a meagre subset of $\mathcal{C}$. And any sequence $\sigma$ weakly NV-learnable by $m$ must be in $A$—for otherwise, $\sigma$ would have the feature that for each $k$, it contained at least $k$ initial segments wicked according to $m$, which would mean that the asymptotic relative frequency of nasty bits in $\sigma$ could not vanish. So the set of sequences weakly NV-learnable by $m$, being a subset of a meagre set, is meagre.

\[ \square \]

**Corollary 3.2.** The set $\{\sigma \in \mathcal{C} \mid \exists m \in \mathcal{E} \text{ such that } m \text{ weakly NV-learns } \sigma \}$ is meagre in $\mathcal{C}$.\(^{20}\)

So the problem of weakly NV-learning sequences is formidable. And difficult choices must be made in the face of this intractability—there can be no optimal extrapolator for weak NV-learning.

We have the following better-but-no-best result.\(^2\)

**Proposition 3.7.** (a) Let $m \in \mathcal{E}$ and let $S$ be a countable subset of $\mathcal{C}$. Then there is an $m^* \in \mathcal{C}$ that NV-learns every $\sigma \in S$ and also weakly NV-learns everything that $m$ does. (b) Let $m \in \mathcal{E}$ and let $S = \{\sigma_i\}_{i \in \mathbb{N}}$ be a family of elements of $\mathcal{C}$ that is uniformly computable in $i$. Then there is an $m^* \in \mathcal{C}$ that NV-learns every $\sigma \in S$ and also weakly NV-learns everything that $m$ does.

**Proof.** We present the argument for (b)—essentially the same argument works for (a).

Define $m^*$ as follows: on input of $w \in \mathcal{B}^n$, $m^*$ finds $K = \{k \in \mathbb{N} \mid 1 \leq k \leq \log_2 n, \sigma_k[m] = w\}$; if $K \neq \emptyset$, then $m^*(w) = \sigma_e(n+1)$, where $e$ is the least element of $K$; otherwise, $m^*(w) = m(w)$. Clearly, $m^*$ is an extrapolating machine and NV-learns each $\sigma_k$ (since each $\sigma_k$ has an initial segment on which $m^*$ guesses $\sigma_k$). And if $\sigma \in \mathcal{C}$ is weakly NV-learned by $m$ then it is also weakly NV-learned by $m^*$: in processing the first $2^n$ bits of $\sigma$, $m^*$ can disagree with $m$ at most $n$ times; so the asymptotic density of bits on which $m^*$ and $m$ disagree in processing $\sigma$ is zero.

We also have the usual sort of evil-twin result.

**Proposition 3.8.** Let $m$ be an extrapolator and let $S \subset \mathcal{C}$ be the set of sequences that it weakly NV-learns. Then there is an extrapolator $m^\dagger$ such that the set $S^\dagger$ of sequences that it weakly NV-learns is disjoint from $S$—and where $m^\dagger$ is in $\mathcal{E}$ of $m$ is.

**Proof.** Define $m^\dagger$ by setting $m^\dagger(w) = 1 - m(w)$ for each $w \in \mathcal{B}^*$. According to either $m$ or $m^\dagger$, in any sequence that the other weakly NV-learns, the good bits have asymptotic density zero. \(\square\)

### 3.3. Extrapolation of Computable Sequences

While it is plausible that every method of learning implementable by a natural or artificial learning agent is computable, the data streams that our agents face may or may not be computable.\(^2\)

Still, there are many settings in which we can be confident that our agents face computable data streams. So let us specialize to the setting in which computable extrapolators attempt to (weakly) NV-learn computable sequences and see how the landscape surveyed above is transformed.

\(\footnotesize{20}\)Remark \footnotesize{below will show that this strengthens the observation of Jockusch and Schupp (2012, p. 438) that the set of coarsely computable sequences is meagre in $\mathcal{C}$.}

\(\footnotesize{2}\)Thanks here to Tom Sterkenburg and to an anonymous referee for helpful suggestions.

\(\footnotesize{23}\)Unless, that is, physical reality itself is fundamentally computational in nature—for a range of views of this topic, see the papers collected in Zenil (2013).
We use \( \mathcal{N}\mathcal{V}(m) \) to denote:

\[
\{ S \subset \mathcal{C} \mid \exists m \in \mathcal{C} \text{ with } S \subseteq \mathcal{N}\mathcal{V}(m) \}.
\]

We likewise use \( \mathcal{N}\mathcal{V}^w(m) \) to denote the set of computable sequences weakly NV-learned by an extrapolating machine \( m \) and use \( \mathcal{N}\mathcal{V}^w \) to denote:

\[
\{ S \subset \mathcal{C} \mid \exists m \in \mathcal{C} \text{ with } S \subseteq \mathcal{N}\mathcal{V}^w(m) \}.
\]

**Proposition 3.9.** \( \mathcal{N}\mathcal{V} \) is a proper subset of \( \mathcal{N}\mathcal{V}^w \).

**Proof.** Clearly \( \mathcal{N}\mathcal{V} \subseteq \mathcal{N}\mathcal{V}^w \). We give an example of a set in \( \mathcal{N}\mathcal{V}^w - \mathcal{N}\mathcal{V} \).

Consider again the extrapolating machine \( m_1 \) of Example 3.1 above that outputs 1 on every input. Let \( U = \mathcal{N}\mathcal{V}^w(m_1) \), the set of computable binary sequences in which 0’s have vanishing asymptotic density. We are going to show that \( U \) is not in \( \mathcal{N}\mathcal{V} \).

Suppose that there is an extrapolating machine \( m \) that NV-learns each sequence in \( U \). Notice that for any \( w \in \mathcal{B}^* \), the sequence \( w.1^w \) is in \( U \)— so for sufficiently large \( \ell \in \mathbb{N} \), we must have \( m(w.1^\ell) = 1 \). Let \( \sigma \in \mathcal{C} \) be the sequence of the form \( 1^{n_1}.0.1^{n_2}.0.1^{n_3}.0 \ldots \) where each \( n_j \) is chosen to be the smallest \( n \) larger than \( 2^j \) such that \( m(1^{n_1}.0.1^{n_2}.0.1^{n_3}.0 \ldots 1^{n_j-1}.0.1^n) = 1 \). Then \( \sigma \in U \): since \( m \) is computable, so is \( \sigma \); and by construction, 0’s occur with vanishing asymptotic density in \( \sigma \). But \( m \) does not NV-learn \( \sigma \), since \( \sigma \) contains infinitely many 0’s each of which \( m \) predicts will be a 1. \( \square \)

**Remark 3.1.** We mention two of the most fundamental variations on \( \mathcal{N}\mathcal{V} \). A partial extrapolating machine is a partial recursive function \( m : \mathcal{B}^* \rightarrow \mathcal{B} \). We say that a partial extrapolating machine \( m \) \( \mathcal{N}\mathcal{V}^m \)-extrapolates \( \sigma \in \mathcal{C} \) if: (i) \( m(\sigma[k]) \) is defined for all \( k \in \mathbb{N} \); and (ii) \( \exists N \in \mathbb{N} \) such that for all \( n > N \), \( m(\sigma[n]) = \sigma(n+1) \).\(^23\) We write \( \mathcal{N}\mathcal{V}^\prime \) for the set of \( S \subset \mathcal{C} \) such that there is a partial extrapolating machine that \( \mathcal{N}\mathcal{V}^\prime \)-extrapolates each \( \sigma \in S \).

We say that a partial extrapolating machine \( m \) \( \mathcal{N}\mathcal{V}^m \)-extrapolates \( \sigma \in \mathcal{C} \) if: \( \exists N \in \mathbb{N} \) such that for all \( n > N \), \( m(\sigma[n]) \) is defined and equal to \( \sigma(n+1) \).\(^24\) We write \( \mathcal{N}\mathcal{V}^\prime \prime \) for the set of \( S \subset \mathcal{C} \) such that there is a partial extrapolating machine that \( \mathcal{N}\mathcal{V}^\prime \prime \)-extrapolates each \( \sigma \in S \).

Obviously, \( \mathcal{N}\mathcal{V} \subseteq \mathcal{N}\mathcal{V}^\prime \subseteq \mathcal{N}\mathcal{V}^\prime \prime \). In fact, \( \mathcal{N}\mathcal{V} \subset \mathcal{N}\mathcal{V}^\prime \subset \mathcal{N}\mathcal{V}^\prime \prime \).\(^25\)

The proof of Proposition 3.9 above carries over essentially unchanged (except that dove-tailing is required) to show that \( \mathcal{N}\mathcal{V}^w \) is not contained in \( \mathcal{N}\mathcal{V}^\prime \) (let alone in \( \mathcal{N}\mathcal{V}^\prime \prime \)). We will see below in Remark 3.3 that \( \mathcal{N}\mathcal{V}^w \) does not contain \( \mathcal{N}\mathcal{V}^\prime \) (let alone \( \mathcal{N}\mathcal{V}^\prime \prime \)).

Via Propositions 3.1 and 3.2, we know that every extrapolating machine NV-learns a countably infinite subset of \( \mathcal{C} \). But there there can be no best extrapolating machine: Propositions 3.4 and 3.8 tells us each extrapolation machine has an evil twin that (weakly) NV-learns a disjoint set of computable sequences; and Propositions 3.3(b) and 3.7(b) tell us that each extrapolating machine is dominated by another that (weakly) NV-learns everything it can while also NV-learning every member of a uniformly computable family of elements of \( \mathcal{C} \).

In this setting, what comparative judgements can we make about the sets of computable sequences that a given extrapolating machine (weakly) NV-learns and doesn’t (weakly) NV-learn?

\(^23\)This notion is due to Barzdins and Freivalds (1972).

\(^24\)This notion is due to Podnieks (1974).

\(^25\)That \( \mathcal{N}\mathcal{V} \subset \mathcal{N}\mathcal{V}^\prime \) is due to Barzdins and Freivalds (1972); that \( \mathcal{N}\mathcal{V}^\prime \subset \mathcal{N}\mathcal{V}^\prime \prime \) is due to Podnieks (1974). See Case and Smith (1983, Corollary 2.29, Corollary 2.31, Theorem 3.1, and Theorem 3.5).
Proposition 3.10. For any \( m \in \mathcal{C} \), following are dense subsets of \( \mathcal{C} \):
(a) the set of computable sequences NV-learnable by \( m \);
(b) the set of computable sequences not NV-learnable by \( m \);
(c) the set of computable sequences weakly NV-learnable by \( m \);
(d) the set of computable sequences not weakly NV-learnable by \( m \).

Straightforward adaptations of the proofs of Propositions 3.2 and 3.5 yield that (a) and (d) are dense. And (b) and (c) are super-sets of (d) and of (a), respectively.

It follows that the set of computable sequences (weakly) NV-learned by an extrapolating machine \( m \) and the set of computable sequences not (weakly) NV-learned by an extrapolating machine \( m \) are both countably infinite subsets of \( \mathcal{C} \)—so we have parity at the level of cardinality. A classical result implies that we also have parity at the level of topology.

Proposition 3.11 (Sierpiński). Any two countable dense subsets of \( \mathcal{C} \) are homeomorphic.

Proof. See, e.g., Dasgupta (2014, chapter 17).

So we can say: for any computable method of extrapolating computable sequences, failure and success are equally common—and difficult choices must be made in selecting a computable method of extrapolation, since no method dominates all its rivals in its range of success.

But, intuitively, we ought to be able to say something stronger. After all, Bārzdins and Freivalds (1972) showed that if \( S \) is a set of total recursive functions, then following are equivalent:
(i) each member of \( S \) is NV-learnable; (ii) \( S \) is a subclass of a recursively enumerable set of total recursive functions; (iii) \( S \) is a subclass of an abstract complexity class.

So only very special subsets of \( \mathcal{C} \) are NV-learnable—which means that generic subsets should not be in \( \mathcal{N} \).

Indeed, there is a natural hybrid computational-topological notion of that underwrites the conclusion that failure is incomparably more common than success for computable extrapolation of computable sequences. Mehlhorn (1973) introduced the important notion of an effectively meagre subset of the set of total recursive functions. We specialize this apparatus to the \( \mathcal{C} \).

By way of motivation, note that in any topological space \( X \) with basis of open sets \( \mathcal{W} \), a subset \( A \) is nowhere dense if and only if for every non-empty \( U \in \mathcal{W} \) there is a non-empty \( U^* \in \mathcal{W} \) with \( U^* \subseteq U \) such that \( A \cap U^* = \emptyset \). So a subset \( A \subseteq \mathcal{C} \) is nowhere dense if and only if there is a function \( f : \mathcal{B}^* \to \mathcal{B}^* \) such that for each binary string \( w \): (i) \( f(w) \) extends \( w \); and (ii) \( A \cap B_{f(w)} = \emptyset \). And \( A \subseteq \mathcal{C} \) is meagre if and only if there is a function \( F : \mathbb{N} \times \mathcal{B}^* \to \mathcal{B}^* \) such that: (i) for each \( n \in \mathbb{N} \) there is an \( A_n \subseteq \mathcal{C} \) such that \( f_n = F(n, \cdot) \) is a witness to the fact that \( A_n \) is nowhere dense in \( \mathcal{C} \); and (ii) \( A = \bigcup_{n \in \mathbb{N}} A_n \).

Definition 3.6 (Mehlhorn). Let \( A \) be a subset of \( \mathcal{C} \) and let \( f : \mathcal{B}^* \to \mathcal{B}^* \) be a computable function. Then \( A \) is effectively nowhere dense via \( f \) if for each \( w \in \mathcal{B}^* \):
(i) \( f(w) \) extends \( w \);
(ii) \( A \cap B_{f(w)} = \emptyset \).

See also Blum and Blum (1975, 127), who attribute the complexity-theoretic condition independently to Adleman. As Blum and Blum remark, this result shows “in essence, that the extrapolable sequences are the ones that can be computed rapidly.”
Definition 3.7 (Mehlhorn). A subset $A$ of $\mathcal{C}$ is effectively meagre if there is a computable function $F : \mathbb{N} \times \mathbb{B}^* \to \mathbb{B}^*$ such that:

i) for each $n \in \mathbb{N}$, there is an $A_n \subset \mathcal{C}$ such that $A_n$ is effectively nowhere dense via $f_n = F(n, \cdot)$;
ii) $A \bigcup_{n \in \mathbb{N}} A_n$.

The complement in $\mathcal{C}$ of an effectively meagre subset of $\mathcal{C}$ is called effectively co-meagre.

Proposition 3.12 (Mehlhorn). The family of effectively meagre subsets of $\mathcal{C}$ is closed under the following operations:

a) taking subsets;
b) taking finite unions;
c) taking effective unions.

Proof. The first claim is immediate from the definition and the second follows from the third. So suppose that $M$ is a subset of $\mathcal{C}$ such that there exist a computable $H : \mathbb{N} \times \mathbb{N} \times \mathbb{B}^*$ and a decomposition $M = \bigcup N_i$, such that for each $k \in \mathbb{N}$, $H(k, \cdot, \cdot)$ is a witness to the fact that $N_k$ is effectively meagre. There exists, then, for each $i \in \mathbb{N}$, a decomposition $N_i = \bigcup N_{ij}$ such that each $N_{ij}$ is effectively nowhere dense in virtue of $H(i, j, \cdot)$. Fix a computable bijection $\pi : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ and let $p_1$ and $p_2$ be the computable components of the inverse of $\pi$ (so that $\pi(p_1(k), p_2(k)) = k$ for all $k \in \mathbb{N}$). Set $M_k := N_{p_1(k), p_2(k)}$ and for each $w \in \mathbb{B}^*$, set $F(k, t) := H(p_1(k), p_2(k), w)$. Then $F : \mathbb{N} \times \mathbb{B}^*$ is computable, $M = \bigcup M_k$, and each $M_k$ is effectively nowhere dense in virtue of $F(k, \cdot)$. So $M$ is effectively meagre.

Crucially, the set of effectively meagre subsets of $\mathcal{C}$ is not closed under arbitrary countable unions due to an effective analog of the Baire Category Theorem.

Proposition 3.13 (Mehlhorn). Let $w$ be a binary string. Then $B_w \cap \mathcal{C}$ is not effectively meagre.

Proof. Let $M = \bigcup M_k$ be an effectively meagre set with witness $F : \mathbb{N} \times \mathbb{B}^*$. We construct strings $w_0, w_1, \ldots$ inductively: $w_0 := w$; and $w_{k+1} = F(k, w_k).0$. By construction, each $w_k$ is a proper initial segment of $w_{k+1}$. Let $\sigma = \lim_{n \to \infty} w_n$. Then $\sigma \in B_w \cap \mathcal{C}$. But for each $k$, $\sigma \notin M_k$ (since $\sigma$ begins with $w_{k+1}$), so $\sigma \notin M$.

In light of these results, it is natural to think of elements of effectively meagre subsets of $\mathcal{C}$ as being incomparably less common than elements of effectively co-meagre subsets of $\mathcal{C}$, even when the meagre and co-meagre sets in question are both dense as subsets of $\mathcal{C}$.

Remark 3.2. A further reason for this standard practice (due to Lisagor 1981): a subset $S$ of $\mathcal{C}$ is effectively meagre if and only if when the Banach–Mazur game (described in Remark 3.2 above) is played for $S$, Player II has a winning strategy that is computable. Another reason (due, again, to Mehlhorn 1973): each abstract complexity class is effectively meagre as a subset of the family of total recursive functions.

Example 3.2 (Self-Describing Sequences). Fix an enumeration $M_1, M_2, \ldots$ of the Turing machines, with associated acceptable programming system $\phi_1, \phi_2, \ldots$ (so that $\phi_k$ is the partial recursive function computed by $M_k$). Following Blum and Blum (1975, 132), we call a sequence $\sigma \in \mathcal{C}$ self-describing if $\sigma$ has an initial segment of the form $1^0\sigma$ and is computed by $M_k$. As Blum and Blum note the set $S_1$ of self-describing sequences is non-trivial: it follows from the Recursion Theorem that each computable binary sequence is a finite variant of a
self-describing sequence—so there are arbitrarily complex sequences in $S_1$.

Fortnow et al. (1998) observe that $S_1$ is not effectively meagre. For, consider any computable strategy $\beta : \mathcal{B}^n \rightarrow \mathcal{B}^n$ that Player II could use to play the Banach–Mazur game for $S_1$. For each $k \in \mathbb{N}$, let $\alpha_k$ be the following strategy that Player I might adopt: on the first turn, play $1^{k0}$; on all subsequent turns, play 0. The assumption that Player I plays strategy $\alpha_k$ and Player II plays strategy $\beta$ determines a unique sequence $\sigma_k \in \mathcal{C}$. The map $F : (k, \ell) \in \mathbb{N}^2 \mapsto \sigma_k(\ell) \in \mathcal{B}$ is computable. So by the Recursion Theorem, there is a $k_0 \in \mathbb{N}$ such that $\sigma_k(0)$ is computed by $M_{k_0}$. That is: there exists a strategy (namely, $\alpha_{k_0}$) via which Player I can defeat $\beta$. So $S_1$ is not in effectively meagre.

**Proposition 3.14.** Let $m$ be an extrapolating machine. $\mathcal{N}V^w(m)$ (the set of computable sequences weakly NV-learnable by $m$) is an effectively meagre subset of $\mathcal{C}$.

**Proof.** A straightforward adaptation of the proof of Proposition 3.6, appealing to the fact that when $m$ is computable, the map $(n, w) \mapsto w^* \notin A_n$ used there is computable. \(\square\)

**Corollary 3.3 (Fortnow et al. (1998)).** Let $m$ be an extrapolating machine. $\mathcal{N}V(m)$ (the set of computable sequences NV-learnable by $m$) is an effectively meagre subset of $\mathcal{C}$.

So there is a natural sense in which, for any computable extrapolator $m$, among computable sequences, those (weakly) NV-learnable by $m$ are incomparably less common than those not (weakly) NV-learnable by $m$. The problem of (weakly) NV-learning computable sequences is formidably difficult.

**Corollary 3.4.** No extrapolating machine can (weakly) NV-learn each self-describing sequence.

**Remark 3.3.** It is illuminating to situate these results with respect to a couple of results of Fortnow et al. (1998). Fortnow et al. are concerned with the problem of identifying computable binary sequences—so they specialize the standard notions of the theory of inductive identification of functions to the case of sequences.

i) They show via an effective Banach–Mazur argument, that any $S \in \mathcal{P}\mathcal{E}\mathcal{X}$ is effectively meagre.

ii) They observe that the set of self-describing functions (see Example 3.2 above) is in $\mathcal{E}\mathcal{X}_0$. So any identification class that contains $\mathcal{E}\mathcal{X}_0$ has members that are not effectively meagre.

They remark: “Since virtually every inference class is either a subset of $\mathcal{P}\mathcal{E}\mathcal{X}$ or a superset of $\mathcal{E}\mathcal{X}_0$ the results here settle virtually all open questions that could be raised.”

Contact can be made with the present approach by recalling that $\mathcal{P}\mathcal{E}\mathcal{X} = \mathcal{N}V$ and that $\mathcal{E}\mathcal{X}_0 \subset \mathcal{N}V$. So the first result of Fortnow et al. noted above is our Corollary 3.3: the set of sequences NV-learnable by an extrapolating machine is effectively meagre. And since every set in $\mathcal{N}V^w$ is effectively meagre, neither $\mathcal{N}V'$ nor $\mathcal{N}V''$ is a subset of $\mathcal{N}V^w$. In $\mathcal{N}V^w$ we have an example of a natural inference class that is neither a subset of $\mathcal{P}\mathcal{E}\mathcal{X}$ nor a superset of $\mathcal{E}\mathcal{X}_0$.

---

27In this discussion, I will assume that the reader is familiar with the standard notions and notation of work on inductive learning—see, e.g., the classic survey Case and Smith (1983).

28Consider a learner who is silent until a data set of the form $1^k0$ is seen and who from then on assumes that the data stream is being generated by $M_k$.

29Fortnow et al. (1998, 145).

30See Case and Smith (1983): that $\mathcal{P}\mathcal{E}\mathcal{X} = \mathcal{N}V$ is their Theorem 2.19 (attributed to private communications from van Leeuwen and Bárzdins); that $\mathcal{E}\mathcal{X}_0 \subset \mathcal{N}V'$ is their Theorem 2.28.
In the present setting, in which computable extrapolators (i.e., extrapolating machines) attempt to learn computable sequences, we find that generalizing our basic model by allowing merely partially defined extrapolating machines allows us to crash through a size barrier in a way that loosening our criterion of success by allowing infinitely many errors in the sense of weak NV-learning does not—since every set in $\mathcal{N\mathcal{V}}$ or in $\mathcal{N\mathcal{V}}^w$ is effectively meagre, whereas this is not the case for every set in $\mathcal{N\mathcal{V}}$ or $\mathcal{N\mathcal{V}}^\omega$. This is the reverse of what we find if we challenge (possibly computable) extrapolators to NV-learn arbitrary sequences. In that setting, in the basic model every learner masters only countably many sequences. And this is unchanged if we countenance merely partially defined learners. But if we loosen our criterion of success to weak NV-learnability we crash through a cardinality barrier, as each learner weakly NV-learns uncountably many sequences.

**Remark 3.4.** Coarse computability implies computable weak NV-learnability. If $\sigma \in \mathcal{C}$ differs from $\sigma^* \in \mathcal{C}$ only in bits of vanishing asymptotic density, then the extrapolating machine that assumes it is being shown $\sigma^*$ on any input weakly NV-learns $\sigma$.

Weak NV-learnability does not imply coarse computability. Let $\sigma_0$ be uncomputable. Construct a sequence $\sigma_1$ as follows: begin with two copies of the first bit of $\sigma_0$, followed by four copies of the second bit of $\sigma_0$, ... followed by $2^k$ copies of the $k$th bit of $\sigma_0$, .... Suppose that $\sigma_1$ is coarsely computable. Then there must be a computable sequence $\sigma_2$ that differs from $\sigma_1$ only in a set of bits of vanishing asymptotic density. Define a new sequence $\sigma_3$ as follows: make the first bit of $\sigma_3$ a 0 if at least one of the first two bits of $\sigma_2$ is a 0, otherwise make it a 1; make the second bit of $\sigma_3$ a 0 if at least two of the next four bits of $\sigma_2$ are 0, otherwise make it a 1; ...; make the $k$th bit of $\sigma_3$ a 0 if at least $2^{k-1}$ of the next $2^k$ bits of $\sigma_2$ are 0, otherwise make it a 1; .... Since $\sigma_2$ is computable (by assumption), so is $\sigma_3$. But $\sigma_3$ is a finite variant of $\sigma_0$ and so must be uncomputable. So there can be no such $\sigma_2$: $\sigma_1$ is not coarsely computable. But $\sigma_1$ is weakly NV-learned by the extrapolating machine that predicts the first bit will be a 1 then subsequently predicts that each bit will be the same as the last bit seen.

4. Forecasting

So far we have set our learners the problem of recognizing which binary sequence is being revealed in the data stream—where such recognition consists in becoming good at predicting future bits. In effect, we are picturing that in generating new bits, Nature simply consults a lookup table chosen in advance and that the learner’s job is to attempt to guess which of the possible such tables is being used (or, in the case of weak learning, to attempt to come close to guessing the right table, in a certain sense).

We might instead picture a different sort of procedure. Suppose that what Nature has chosen in advance is not a sequence but, rather, a Borel probability measure $\lambda$ on $\mathcal{C}$ and that the learner’s data stream is randomly sampled from $\lambda$. So we now picture Nature as being equipped with a complete set of biased coins and an instruction manual that says which coin to toss to generate the next bit, given the bits that have been generated so far. To mention just some of the tamest possibilities: Nature may have chosen a Bernoulli measure, in which case the instruction will be to use the same coin to generate each new bit; or Nature may have chosen a measure corresponding to a Markov chain, in which case the coin chosen to generate a new bit will depend only on some fixed finite number of immediately preceding

\[31\text{Thanks to an anonymous referee for pointing this out.}\]
Definition 4.1 (Sources). A source is a Borel probability measure \( \lambda \) on \( C \).

In what follows, we will think of Nature as having chosen a source from which our learner’s data stream is sampled. Recall for \( w \in B^* \), we write \( \lambda(w) \) in place of \( \lambda(B_w) \). Similarly, for \( s = 0,1 \) and \( w \in B^n \), we will write \( \lambda(s \mid w) \) for the conditional probability \( \lambda \) gives for the \((n + 1)\)st bit to be \( s \) given that the first \( n \) bits were given by \( w \).

How should a learner proceed in the setting where the data stream is given by a probabilistic source? In the setting of Section 3, where we were thinking of new bits as being generated by a deterministic process, we asked learners to choose an extrapolator that would allow them to definitively predict at each stage what the next bit would be, given the data seen so far. That approach would be suboptimal in the present setting: if Nature is using the fair coin measure (the Bernoulli measure of bias \(.5\)) to generate the data stream, then (with probability one) no extrapolator will do better (or worse) than random in its predictions of the next bit—but the fact that Nature is using this procedure seems like a paradigm example of the sort of thing that we ought to be able to detect by looking at data. This will be possible if we ask agents to choose a forecasting procedure that allows them to issue a forecast probability before each bit is revealed, rather than choosing an extrapolator that at each stage issues definitive predictions regarding the next bit.

A natural way to encode such a strategy for learning would be as a forecasting function: a map \( \hat{\mu} : B^* \times B \to (0,1) \) with the feature that for all \( w \in B^* \), \( \hat{\mu}(w, 1) + \hat{\mu}(w, 0) = 1 \) (here \( \hat{\mu}(w,s) \) is to be interpreted as the forecast probability that \( \hat{\mu} \) issues for the next bit to be \( s \) \((s = 0,1)\), conditional on data set \( w \)). In fact, it is more convenient to employ a slightly different representation. Note that any \( \hat{\mu} \) of the above form induces, for each \( n \), a probability measure \( \mu_n \) on \( B^n \). Further, for any \( \hat{\mu} \) and \( \mu \), \( m \leq n \), \( \mu_m \) and \( \mu_n \) will be consistent. So by the Kolmogorov Consistency Theorem, \( \hat{\mu} \) induces a measure \( \mu \) on \( C \) (with \( \hat{\mu} \) computable if and only if \( \mu \) is).

Example 4.1. Define \( \hat{\mu} : B^* \times B \to (0,1) \) as follows: if \( w \in B^n \) contains \( k \) 1’s, then \( \hat{\mu}(w, 1) = \frac{k+1}{n+2} \) and \( \hat{\mu}(w, 0) = \frac{n-k+1}{n+2} \). This map satisfies the condition that for all \( w \in B^* \), \( \hat{\mu}(w, 1) + \hat{\mu}(w, 0) = 1 \) and the corresponding measure is the Laplace–Bayes prior, the Lebesgue–uniform mixture of the Bernoulli measures.

Not all measures in \( P \) correspond in this way to such \( \hat{\mu} \): a measure \( \mu \in P \) is associated in this way with a \( \hat{\mu} \) of the above form if and only if it is a measure of full support (i.e., it assigns positive weight to each open set—or, equivalently, to each basic open set \( B_w \)).

Definition 4.2 (Forecasters). A forecaster is a Borel probability measure on \( C \) of full support. We denote the family of forecasters by \( F \).

Definition 4.3 (Forecasting Machines). A forecasting machine is a computable Borel probability measure on \( C \) of full support. We denote the family of forecasting machines by \( \mathcal{F} \).

The aim of a forecaster is to give faithful estimates of the chance of the next bit’s being a 0 or a 1, given the data seen so far. We are going to distinguish three criteria for the

---

32See item \( \square \) of Section 3 for the relevant notion of consistency.

33For a treatment adapted to the special case of measures on \( C \), see Baez-Duarte (1971). For a general treatment, see, e.g., Purushotham (1965) chap. V.
successful next-chance learning.\textsuperscript{34} The most restrictive one, due to Blackwell and Dubins (1962), requires that the forecaster eventually offer answers arbitrarily similar to those of the source concerning any (measurable) question that might be asked about the data stream.\textsuperscript{35} The intermediate one, due to Kalai and Lehrer (1994), requires that the forecaster’s probabilistic predictions concerning the next bit eventually approach the true values arbitrarily closely.\textsuperscript{36} The least restrictive one, due to Lehrer and Smorodinsky (1996), relaxes this last requirement by allowing errors, so long as they eventually become arbitrarily rare.

\textit{Definition 4.4} (Strong NC-Learning). We say that forecaster $\mu$ \textit{strongly NC-learns} source $\lambda$ (or that $\lambda$ is \textit{strongly NC-learnable} by $\mu$) if with $\lambda$-probability 1 the data stream $\sigma \in \mathcal{C}$ satisfies:

$$\lim_{n \to \infty} \sup_{A \in \mathfrak{B}} |\mu(A | \sigma[n]) - \lambda(A | \sigma[n])| = 0$$

(recall that $\mathfrak{B}$ denotes the family of Borel subsets of $\mathcal{C}$).

\textit{Definition 4.5} (NC-Learning). We say that forecaster $\mu$ \textit{NC-learns} source $\lambda$ (or that $\lambda$ is \textit{NC-learnable} by $\mu$) if with $\lambda$-probability 1 the data stream $\sigma \in \mathcal{C}$ satisfies

$$\lim_{n \to \infty} \mu(s | \sigma[n]) - \lambda(s | \sigma[n]) = 0 \quad s = 0, 1.$$  

\textit{Definition 4.6} (Weak NC-Learning). We say that forecaster $\mu$ \textit{weakly NC-learns} source $\lambda$ (or that $\lambda$ is \textit{weakly NC-learnable} by $\mu$) if with $\lambda$-probability 1, the data stream $\sigma \in \mathcal{C}$ satisfies

$$\lim_{n \to \infty} \mu(s | \sigma[n]) - \lambda(s | \sigma[n]) = 0 \quad s = 0, 1$$

for some $K \subset \mathbb{N}$ with asymptotic density one.

\textit{Example 4.2}. If $\mu$ is a forecaster, then $\mu$ is also a source and it is immediate that $\mu$ strongly NC-learns, NC-learns, and weakly NC-learns $\mu$.

\textit{Proposition 4.1} (Kalai and Lehrer (1994), Lehrer and Smorodinsky (1996)). For any source $\lambda$ and any forecaster $\mu$, strong NC-learnability of $\lambda$ by $\mu$ implies NC-learnability (but not conversely) and NC-learnability of $\lambda$ by $\mu$ implies weak NC-learnability (but not conversely).

\textit{Proof}. Strong NC-learnability implies NC-learnability: in the definition of strong NC-learnability, for each $n$ take $A$ to be the event of the $(n+1)$st bit being a 1. To see that the converse is not true, consider the family $\{\lambda_p | p \in [0, 1]\}$ of Bernoulli measures and let $\mu$ be the Laplace–Bayes prior. It is a basic fact about $\mu$ that it is statistically consistent for the problem of identifying the bias of a coin from knowledge of outcomes of a sequence of tosses.\textsuperscript{37} It follows that the forecaster $\mu$ NC-learns each $\lambda_p$. But $\mu$ does not strongly NC-learn any $\lambda_p$: let $E_p$ be the event that the limiting relative frequency of 1’s in the data stream is $p$; then $\lambda_p(E_p) = 1$ but $\mu(E_p) = 0$; so for any $w$ for which $\lambda_p(w) \neq 0$, $|\mu(E_p | w) - \lambda_p(E_p | w)| = 1$.

\textsuperscript{34}Investigation of inductive learning as next-chance learning traces back to Solomonoff (1964). Several criteria of success are prevalent in the literature on Solomonoff induction—see Solomonoff (1978), Hutter (2007), Li and Vitányi (2019, §5.2). But these differ from those considered below in their focus on average or expected performance.

\textsuperscript{35}Note that in the deterministic setting of Section 3 above, the distinction between eventually becoming good at answering all questions and eventually becoming good at predicting the next bit collapsed—recall Proposition 3.1 above.

\textsuperscript{36}For relations between this criterion of success and those alluded to in fn. 34 above, see Ryabko and Hutter (2007).

\textsuperscript{37}See, e.g., Freedman (1963).
Clearly, NC-learnability implies weak NC-learnability. To see that the converse is not true, take \( \mu \) to be the fair coin measure and take \( \lambda \) to be the source that generates bits as follows: for \( k = 10^m \), \( s_k \) is the \( m \)th bit in the binary expansion of \( \pi \); all other \( s_k \) are generated by flipping a fair coin. The forecaster \( \mu \) weakly NC-learns this \( \lambda \) but does not NC-learn it, since there are large discrepancies between the forecast probabilities and the true probabilities at arbitrarily late times.

We are going to see that relative to each of these three criteria, the problem of next-chance learning is formidable difficult and involves hard choices.

4.1. Strong NC-learning

A famous result and its converse give a necessary and sufficient condition for a source to be strongly NC-learnable by a forecaster.

**Proposition 4.2** [Blackwell and Dubins 1962]. If source \( \lambda \) is absolutely continuous with respect to forecaster \( \mu \) (i.e., \( \lambda(A) > 0 \) implies \( \mu(A) > 0 \) for all \( A \in \mathcal{B} \)), then \( \mu \) strongly NC-learns \( \lambda \).

**Proposition 4.3** [Lehrer and Smorodinsky 1996]. If forecaster \( \mu \) strongly NC-learns source \( \lambda \), then \( \lambda \) is absolutely continuous with respect to \( \mu \).

**Proposition 4.4.** Let \( \mu \) be a forecaster. The sources strongly NC-learnable by \( \mu \) form a dense subset of \( \mathcal{P} \) of cardinality \( c \).

**Proof.** Let \( w_1, \ldots, w_n \) be binary strings such that \( \mathcal{C} \) is a disjoint union of the \( B_{w_k} \). And let \( p_1, \ldots, p_n \in (0,1) \) with \( \sum_{k=1}^n p_k = 1 \). Since the \( B_{w_k} \) partition \( \mathcal{C} \), each \( w \in \mathcal{B} \) is either one of the \( w_k \), or a proper prefix of some of the \( w_k \), or a proper extension of one of the \( w_k \). Define a map \( \tilde{\lambda} : \mathcal{B}^* \to [0,1] \) as follows:

(a) If \( w = w_k \) for some \( k \), then \( \tilde{\lambda}(w) = p_k \).

(b) If \( w \) is a prefix of \( w_{j_1}, \ldots, w_{j_\ell} \), then \( \tilde{\lambda}(w) = \sum_{k=1}^\ell p_{j_k} \).

(c) If \( w \) is of the form \( w_k v \) for some binary string \( v \), then \( \tilde{\lambda}(w) = p_k \cdot \mu(v | w_k) \).

It is immediate that \( \tilde{\lambda}(\emptyset) = 1 \) and that \( \tilde{\lambda}(w) = \tilde{\lambda}(w.0) + \tilde{\lambda}(w.1) \) for each \( w \in \mathcal{B}^* \). So by the Carathéodory Extension Theorem, \( \tilde{\lambda} \) extends uniquely to a measure \( \lambda \in \mathcal{C} \) such that \( \lambda w = \tilde{\lambda}(w) \) for each \( w \in \mathcal{B}^* \).

The source \( \lambda \) is strongly NC-learnable by \( \mu \). For suppose that \( A \) is a Borel subset of \( \mathcal{C} \) with \( \lambda(A) > 0 \). Since

\[
\lambda(A) = \sum_{k=1}^n \lambda(A | w_k) \lambda(w_k),
\]

there must be some \( 1 \leq \ell \leq n \) such that \( \lambda(A | w_\ell) > 0 \). And, since \( p_\ell \cdot \lambda(A | w_\ell) = \mu(A | w_\ell) \), \( \mu(A | w_\ell) \) must likewise be positive. Since

\[
\mu(A) = \sum_{k=1}^n \mu(A | w_k) \mu(w_k)
\]

and since \( \mu(w_\ell) > 0 \) (\( \mu \) being a forecaster) this tells us that \( \mu(A) > 0 \). So \( \lambda \) is absolutely continuous with respect to \( \mu \) and Proposition 4.2 tells us that \( \mu \) strongly NC-learns \( \lambda \). And since the \( w_k \) and the \( p_k \) can be chosen arbitrarily, we construct in this way continuum-many such sources in any finite intersection of sub-basic open sets of \( \mathcal{P} \). \( \square \)
The next result follows from the stronger Proposition 4.9 below, but we include it here in order to indicate an independent route to establishing it.

**Proposition 4.5** (Noguchi (2015)). For any \( \mu \in \mathcal{F} \), the set \( S(\mu) \subset \mathcal{P} \) of sources strongly NC-learned by \( \mu \) is meagre in \( \mathcal{P} \).

**Proof.** A classical result tells us that for any measure in \( \mathcal{P} \), there is some meagre subset of \( \mathcal{C} \) to which that measure assigns probability 1.\(^{38}\) And Proposition 1 of Dekel and Feinberg (2006) tells us that for any meagre subset of \( \mathcal{C} \), the set of probability measures that assign that set positive probability is meagre in \( \mathcal{P} \).\(^{39}\) So let \( A \) be a meagre subset of \( \mathcal{C} \) such that \( \mu|_A = 1 \) and let \( P_A \subset \mathcal{P} \) be the, necessarily meagre, subset of measures that assigns \( A \) positive probability. By Proposition 4.3, if \( \lambda \in S(\mu) \), then \( \lambda \in P_A \). So \( S(\mu) \), being a subset of a meagre set, is meagre. \( \square \)

**Corollary 4.1.** The set \( \{ \lambda \in \mathcal{P} \mid \exists \mu \in \mathfrak{F} \text{ such that } \mu \text{ strongly NC-learns } \lambda \} \) is meagre in \( \mathcal{P} \).

**Proposition 4.6.** For any \( \mu \in \mathcal{F} \), the set \( J(\mu) \subset \mathcal{P} \) of sources strongly NC-learn at least one source in common with \( \mu \) is meagre.

**Proof.** Let \( N(\mu) \) be the set of \( \nu \in \mathcal{P} \) such that there is no \( \lambda \in \mathcal{P} \) that is absolutely continuous with respect to both \( \mu \) and \( \nu \). Note that \( J(\mu) \subset N(\mu) \). So it suffices to show that \( N(\mu) \) and \( \mathcal{F} \) are both co-meagre subsets of \( \mathcal{P} \).

Let \( A \) and \( P_A \) be as in the proof of the preceding proposition. Consider an arbitrary \( \nu \in \mathcal{P} \). If there is a \( \lambda \in \mathcal{P} \) that is absolutely continuous with respect to both \( \mu \) and \( \nu \), then \( \lambda \) must assign the complement of \( A \) zero probability (since \( \mu \) does), which means that \( \nu \) must assign \( A \) positive probability (since \( \lambda \) does). So the complement of \( N(\mu) \) is meagre, being a subset of the meagre set \( P_A \).

And \( \mathcal{F} \) is co-meagre: the forecasters form a dense \( G_\delta \) subset of \( \mathcal{P} \); and in any completely metrizable space (such as \( \mathcal{P} \)), any dense \( G_\delta \) subset is co-meagre.\(^{40}\) \( \square \)

**Corollary 4.2.** For any forecaster, there is another, such that the sets of sources strongly NC-learned by the two forecasters are disjoint.

As usual, we call a sequence \( \{ \lambda_i \}_{i \in \mathbb{N}} \) of elements of \( \mathfrak{F} \) *uniformly computable in* \( i \) if there is a computable \( F : \mathbb{N} \times \mathcal{B}^* \times \mathbb{N} \to \mathbb{Q} \) such that \( |\lambda_i(w) - F(i, w, n)| \leq 2^{-n} \), for all \( i, n \in \mathbb{N} \) and \( w \in \mathcal{B}^* \).

**Proposition 4.7.** (a) Let \( \mu \in \mathcal{F} \) and let \( S \) be a countable subset of \( \mathcal{P} \). Then there is a \( \mu^* \in \mathcal{F} \) that strongly NC-learns every source in \( S \) as well as every source strongly NC-learned by \( \mu \). (b) Let \( \mu \in \mathfrak{F} \) and let \( S = \{ \lambda_i \}_{i \in \mathbb{N}} \) be a sequence of measures in \( \mathfrak{F} \) that is uniformly computable in \( i \). Then there is a \( \mu^* \in \mathfrak{F} \) that strongly NC-learns every source in \( S \) as well as every source strongly NC-learned by \( \mu \).

---

38Szpilrajn (1934, Théorème 1) shows that any non-atomic Borel probability measure on a separable metric space assigns measure 0 to some co-meagre set. Marczewski and Sikorski (1949, fn. 3) observe that this result implies that in a separable metric space without isolated points, every Borel probability measure assigns probability 0 to some co-meagre set (in fact, the hypothesis of separability can be dropped—see, e.g., Zindulka 1999, Corollary 3.7). Note that Szpilrajn = Marczewski.

39This is a special case of a result of Koumoullis (1996).

40See Dubins and Freedman (1964, §3.13) and Oxtoby (1980, Theorem 9.2).
Proof. For part (a), enumerate the members of
\[ S = \{ \lambda_1, \lambda_2, \ldots \} \] and set
\[ \mu^{\ast} = \frac{1}{2} \mu + \frac{1}{2} \sum_{k=1}^{\infty} \frac{1}{2^k} \lambda_k. \]
Each \( \lambda_k \) is absolutely continuous with respect to \( \mu \), so by Proposition 4.2, \( \mu^{\ast} \) strongly NC-learns every source in \( S \). And if \( \nu \) is a source strongly NC-learned by \( \mu \), then by Proposition 4.3, \( \nu \) must be absolutely continuous with respect to \( \mu \) and hence also with respect to \( \mu^{\ast} \)—so by Proposition 4.2, \( \mu^{\ast} \) strongly NC-learns \( \nu \).

For part (b), we can proceed in the same way. The only thing to check is that if \( \mu \in \mathcal{P} \) is computable and \( \{ \lambda_i \} \subset \mathcal{P} \) is uniformly computable in \( i \), then the measure \( \mu^{\ast} \) as defined above is also computable. To this end, suppose that \( F_0 : \mathcal{B}^* \times \mathbb{N} \to \mathbb{Q} \) and \( F : \mathbb{N} \times \mathcal{B}^n \times \mathbb{N} \to \mathbb{Q} \) are computable, with \(|\mu(w) - F_0(w, n)| \leq 2^{-n}\) and \(|\lambda_i(w) - F(i, w, n)| \leq 2^{-n}\), for all \( w \in \mathcal{B}^* \) and \( i, n \in \mathbb{N} \).

We define \( F^{\ast} : \mathcal{B}^* \times \mathbb{N} \to \mathbb{Q} \) as follows:
\[
F^{\ast}(n, w) := \frac{1}{2} F_0(w, n + 2) + \sum_{k=1}^{n+1} \frac{1}{2^{k+1}} F(k, w, 2n)
\]
Then for any given \( w \in \mathcal{B}^* \) and \( n \in \mathbb{N} \), we define \( \alpha, \beta, \gamma \in \mathbb{R} \):
\[
\alpha := \frac{1}{2} (\mu(w) - F_0(w, n + 1))
\]
\[
\beta := \sum_{k=1}^{n+1} \frac{1}{2^{k+1}} (\lambda_k(w) - F(k, w, 2n))
\]
\[
\gamma := \sum_{k=n+2}^{\infty} \frac{1}{2^{k+1}} \lambda_k(w).
\]
Note that each of \(|\alpha|, |\beta|, \) and \(|\gamma|\) is no greater than \(2^{-(n+2)}\). In the case of \(|\alpha|\), this follows from what we know about \( F_0 \). For \(|\beta|\), we have:
\[
|\beta| \leq \sum_{k=1}^{n+1} \frac{1}{2^{k+1}} |\lambda_k(w) - F(k, w, 2n)|
\]
\[
< \sum_{k=1}^{n+1} \frac{1}{4} |\lambda_k(w) - F(k, w, 2n)|
\]
\[
< \sum_{k=1}^{n+1} \frac{1}{4} 2^{-2n}
\]
\[
= \frac{n + 1}{2^{-n}} 2^{-(n+2)},
\]
and for any \( n \geq 1 \) we have that \( n + 1 \leq 2^n \). And since for each \( k \) we have \( 0 \leq \lambda_k(w) \leq 1 \), we have that \(|\gamma| \leq \sum_{k=n+2}^{\infty} \frac{1}{2^{k+1}} \).

Now, \( F^{\ast} \) is computable and we have:
\[
|\mu^{\ast}(w) - F^{\ast}(w, n)| = |\alpha + \beta + \gamma|
\]
\[
\leq |\alpha| + |\beta| + |\gamma|
\]
\[
< 2^{-n}.
\]
So $\mu^*$ is computable. □

Thus we have a no-free-lunch result for strong NC-learning: every forecaster strongly NC-learns an uncountable and dense but meagre set of sources; for every (computable) forecaster there is another (computable) forecaster that strongly NC-learns everything it does, plus a further countably infinite set of sources; and for every forecaster there is another that strongly NC-learns a disjoint set of sources (indeed, typical forecasters have this feature).

4.2. NC-LEARNING AND WEAK NC-LEARNING

Proposition 4.5 above tells us that each forecaster strongly NC-learns a dense and uncountable but meagre set of sources. This implies that the sets of sources NC-learned and weakly NC-learned by any forecaster are also dense and uncountable. Strong NC-learning is, intuitively, a much more restrictive notion than NC-learning: being able to accurately answer all questions about the data stream, including questions about the infinite future, is much a much more demanding standard than being able to accurately estimate the chances for the next bit.\footnote{Indeed, there is a sense in which strong NC-learning implies rapid NC-learning, and a sense in which the converse implication holds—see \cite{SandroniSmorodinsky1999} Propositions 2 and 3.}

Similarly, NC-learning is, intuitively, a much more restrictive notion than weak NC-learning: we saw in Section 3 above that weakening NV-learning by allowing an infinite number of errors (so long as they were of asymptotic density zero) made a marked difference to the size of the set of sequences that a given extrapolator could learn—any extrapolator NV-learns a countable set of sequences but weakly NV-learns an uncountable set of sequences. So it is not obvious that the set of sources (weakly) NC-learned by a given forecaster should always be meagre.\footnote{\cite{Noguchi2015}, after discussing the results cited in fn. 41 above, remarks that: “These results lead us to conjecture that, in general, a merged set (of probability measures) \[i.e.,\] a set of sources strongly NC-learned by a given forecaster] may be much smaller than a weakly merged set [i.e, a set of sources NC-learned by a given forecaster].” He then goes on to observe that each forecaster strongly NC-learns a meagre set of sources—so it is natural to read him as conjecturing that the set of sources NC-learned by a forecaster need not be meagre.}

Not obvious—but, as we will see, nonetheless true.

We begin by introducing a basis $W$ for the weak topology on $\mathcal{P}$. First, for each $k \in \mathbb{N}$, we fix a metric on $\mathcal{P}_k$ compatible with its topology: we take the distance between $\lambda, \mu \in \mathcal{P}_k$ to be given by:

$$d(\mu, \lambda) := \max_{w \in B^k} |\mu(w) - \lambda(w)|.$$ 

In terms of our identification of $\mathcal{P}_k$ with a closed subset of $\mathbb{R}^{2^k}$, this is the metric induced by the $\ell^\infty$ norm on $\mathbb{R}^{2^k}$. For $\mu \in \mathcal{P}_k$ and $\varepsilon > 0$, we write $B(\mu, \varepsilon)$ for the open metric ball of radius $\varepsilon$ centred at $\mu$:

$$B(\mu, \varepsilon) := \{\lambda \in \mathcal{P}_k \mid d(\mu, \lambda) < \varepsilon\}.$$ 

We call $B(\mu, \varepsilon) \subset \mathcal{P}_k$ rational if $\varepsilon \in \mathbb{Q}$ and $\mu(w) \in \mathbb{Q}$ for each $w \in \mathcal{B}_k$.

For each $k \in \mathbb{N}$, let $\Pi_k : \mathcal{P} \to \mathcal{P}_k$ be the restriction map: for $\mu \in \mathcal{P}$, $\Pi_k(\mu)$ is the measure in $\mathcal{P}_k$ such that $\Pi_k(\mu)(w) = \mu(w)$ for each $w \in \mathcal{B}_k$. We now take $W$ to comprise the inverse images under the $\Pi_k$ of the open rational metric balls in the various $\mathcal{P}_k$:

$$W := \{W = \Pi_k^{-1}(B(\mu, \varepsilon)) \mid k \in \mathbb{N}, \mu \in \mathcal{P}_k, \mu(w) \in \mathbb{Q} \forall w \in \mathcal{B}_k, \varepsilon > 0, \varepsilon \in \mathbb{Q}\}.$$ 

Proposition 4.8. $W$ is a basis for the weak topology on $\mathcal{P}$. 
Proof. It suffices to show: (i) that each $W \in \mathcal{W}$ is open; and (ii) that for any open set $U \subset \mathcal{P}$ and for any $\nu \in U$, there is a $W \in \mathcal{W}$ with $\nu \in W \subset U$.

(i) Fix $W \in \mathcal{W}$ of the form $W = \Pi_k^{-1}(B(\mu, \varepsilon))$. Let $w_1, \ldots, w_{2^k}$ be an enumeration of the $k$-bit strings. And for each $1 \leq j \leq 2^k$, let $p_j := \max\{0, \mu(w_j) - \varepsilon\}$ and $q_j := \min\{1, \mu(w_j) + \varepsilon\}$. Then we have:

$$W = \left\{ \lambda \in \mathcal{P} \mid \max_{1 \leq j \leq 2^k} |\mu(w_j) - \lambda(w_j)| < \varepsilon \right\}$$

where each $S_{w_j,p_j,q_j}$ is a sub-basic open subset of $\mathcal{P}$ (as in item (vi) of Section 2.1 above). So $W$ is an open subset of $\mathcal{P}$.

(ii) It suffices to consider an open set $U \subset \mathcal{P}$ that is a finite intersection of sub-basic open sets. Let $S_{w_1,p_1,q_1}, \ldots, S_{w_n,p_n,q_n}$ be arbitrary sub-basic open subsets of $\mathcal{P}$ and suppose that $U := \bigcap_{k=1}^n S_{w_k,p_k,q_k} \neq \emptyset$. Let $N = \max\{|w_1|, \ldots, |w_n|\}$ and let $\nu \in U$. Note that each $\Pi_N(S_{w_j,p_j,q_j})$ is an open subset of $\mathcal{P}_N$: each condition of the form $p_j < \Pi_N(\mu)(w_j) < q_j$ just imposes an inequality on (sums of) differences of coordinate relative to our identification of $\mathcal{P}_N$ with a subset of $\mathbb{R}^{2^N}$. So we can find a rational open metric ball $B$ contained in $\Pi_N(U)$ with $\Pi_N(\nu) \in B$. Letting $W := \Pi_N^{-1}(B) \in \mathcal{W}$, we have $\nu \in W \subset U$. $\square$

Proposition 4.9. Let $\mu$ be a forecaster. The sources weakly NC-learnable by $\mu$ form a meagre subset of $\mathcal{P}$.

Proof. For any source $\lambda$ and $k \in \mathbb{N}$ let us say that $(\mu, \lambda)$ considers $k$ bad if for each $w \in \mathcal{B}^k$ we have

$$|\mu(s \mid w) - \lambda(s \mid w)| \geq \frac{1}{5} \quad s = 0, 1.$$ 

And let us say that $(\mu, \lambda)$ considers $k$ super-bad if $(\mu, \lambda)$ considers more than half of the $m \leq k$ to be bad. And, by extension, for any subset $S \subset \mathcal{P}$, let us say that $(\mu, S)$ considers $k \in \mathbb{N}$ (super-)bad if $(\mu, \lambda)$ does for each $\lambda \in S$.

For each $n \in \mathbb{N}$, let $F_n$ be the set of $\lambda \in \mathcal{P}$ such that $(\mu, \lambda)$ considers at least $n$ natural numbers to be super-bad. And let $A_n$ be the complement of $F_n$ in $\mathcal{P}$. Note that $\mu$ cannot NC-learn $\lambda$ if there are infinitely many $k \in \mathbb{N}$ that $(\mu, \lambda)$ considers bad and that $\mu$ cannot weakly NC-learn $\lambda$ if there are infinitely many $k \in \mathbb{N}$ that $(\mu, \lambda)$ considers super-bad. So if $\mu$ weakly NC-learns $\lambda$, then $(\mu, \lambda)$ can consider only finitely many natural numbers to be super-bad, which means that there will be an $N$ such that $\lambda \notin F_N$, which implies that $\lambda \in A := \bigcup_{n \in \mathbb{N}} A_n$. So in order to establish our proposition, it suffices to show that each $A_n$ is nowhere dense in $\mathcal{P}$.

The first step is to suppose that we are given a set $W \in \mathcal{W}$ of the form $\Pi_k^{-1}(B(\lambda, \varepsilon))$ and to show how to find $W_1 \in \mathcal{W}$ of the form $W_1 = \Pi_{k+1}^{-1}(B(\lambda_1, \varepsilon_1))$, such that $W_1 \subset W$ and $(\mu, W_1)$ considers $k + 1$ to be bad.

First, we select $\lambda_1$. For each $w \in \mathcal{B}^k$, if $\mu(w.0) \geq \mu(w.1)$, we set

$$\lambda_1(w.0) = \frac{1}{10} \cdot \lambda(w) \quad \text{and} \quad \lambda_1(w.1) = \frac{9}{10} \cdot \lambda(w);$$

otherwise we set

$$\lambda_1(w.0) = \frac{9}{10} \cdot \lambda(w) \quad \text{and} \quad \lambda_1(w.1) = \frac{1}{10} \cdot \lambda(w).$$
This gives us a well-defined \( \lambda_1 \in \mathcal{P}_{k+1} \) that assigns rational values to each string in \( \mathcal{B}^{k+1} \).

We now select \( \varepsilon_1 > 0 \) as follows: we choose \( m \) large enough so that \( \varepsilon_1 = 10^{-m} \) is small enough so that for any \( \lambda' \in B(\lambda_1, \varepsilon_1) \), for each \( w \in \mathcal{B}^k \), if \( \mu(w.0) \geq \mu(w.1) \), then

\[
\lambda'(w.1) > \frac{8}{10} \cdot \lambda(w),
\]

and if \( \mu(w.0) < \mu(w.1) \) then,

\[
\lambda'(w.0) > \frac{8}{10} \cdot \lambda(w).
\]

This process can be iterated. In particular, if we are given \( W \in \mathcal{W} \) of the form \( \Pi_{k+1}^{-1}(B(\lambda, \varepsilon)) \), we can run the process once to construct \( W_1 \in \mathcal{W} \) with \( W_1 \subset W \) such that \((\mu, W_1)\) considers \( k + 1 \) bad; applying the process again (now with \( W_1 \) if place of \( W \)) yields a \( W_2 \in \mathcal{W} \) with \( W_2 \subset W_1 \) such that \((\mu, W_2)\) considers \( k + 2 \) bad and so on.

So if we are given \( W \in \mathcal{W} \) of the form \( \Pi_{k+1}^{-1}(B(\lambda, \varepsilon)) \), we can run the process \( k + n \) times to yield \( W^* := W_{k+n} \in \mathcal{W} \) such that \( W^* \subset W \) and \((\mu, W^*)\) considers at least \( n \) numbers to be super-bad, so that \( W^* \cap A_n = \emptyset \). So \( A_n \) is nowhere dense in \( \mathcal{P} \).

**Corollary 4.3.** The set \( \{ \lambda \in \mathcal{P} \mid \exists \mu \in \mathcal{F} \text{ such that } \mu \text{ weakly NC-learns } \lambda \} \) is meagre in \( \mathcal{P} \).

It of course follows that the set of sources (strongly) NC-learnable by a given forecaster are likewise meagre—and that set of all sources collectively (strongly) learnable by forecasting machines are likewise meagre.

**Remark 4.1.** The set of sources not even weakly NC-learnable by a given forecaster \( \mu \) is a co-meagre subset of \( \mathcal{P} \), and so is uncountable. A variant on the proof of the above proposition shows that even if the continuum hypothesis fails, this set has the cardinality of the continuum. Fix a metric on \( \mathcal{P} \) compatible with the weak topology, such as the Prokhorov metric.\footnote{For details see, e.g., Billingsley (1999, 72 f.).} And let us amend the iterative procedure of the proof of the preceding proposition so that the diameter of \( W_{k+1} \) relative to this metric is no more than half of the diameter of \( W_k \). Then if we are given \( W \in \mathcal{W} \) and repeatedly apply our revised iterative procedure, we will construct a sequence \( W_1, W_2, \ldots \) of nested \( \mathcal{W} \)-sets such that \( \bigcap_{k=1}^{\infty} W_k \) contains a single source, which is not even weakly NC-learnable by \( \mu \). There are continuum-many distinct \( W \) we could use to initiate this procedure—these determine continuum-many distinct sources not even weakly NC-learnable by \( \mu \).

**Proposition 4.10.** Let \( \mu \) be a forecaster. Then there is a second forecaster \( \mu^\dagger \) such that the sets of sources weakly NC-learned by \( \mu \) and by \( \mu^\dagger \) are disjoint. If \( \mu \) is computable, we can take \( \mu^\dagger \) to be likewise computable.

**Proof.** Let \( \mu \) be given. We construct a map \( \nu : \mathcal{B}^* \rightarrow [0, 1] \) inductively as follows:

a) \( \nu(\emptyset) = 1 \);

b) Supposing that \( \nu(w) \) is given, we define \( \nu(w.0) \) and \( \nu(w.1) \) as follows:
   i) if \( \mu(0 \mid w) \leq \mu(1 \mid w) \), then \( \nu(w.0) = 9/10 \cdot \nu(w) \) and \( \nu(w.1) = 1/10 \cdot \nu(w) \);
   ii) if \( \mu(0 \mid w) > \mu(1 \mid w) \), then \( \nu(w.0) = 1/10 \cdot \nu(w) \) and \( \nu(w.1) = 9/10 \cdot \nu(w) \).

Clearly, for any \( w \in \mathcal{B}^* \), \( \nu(w) = \nu(w.0) + \nu(w.1) \). So by the Carathéodory Extension Theorem, \( \nu \) extends to a unique Borel probability measure on \( \mathcal{C} \), which we take as our \( \mu^\dagger \).
For any non-empty \( w \in B^* \), \( |\mu(1 \mid w) - \mu^f(1 \mid w)| \geq \frac{2}{3} \). So for any \( \lambda \in \mathcal{P} \), any \( \sigma \in \mathcal{C} \), and \( n \in \mathbb{N} \) we have:

\[
\max\{|\mu(1 \mid \sigma[n]) - \lambda(1 \mid \sigma[n])|, \ |\mu^f(1 \mid \sigma[n]) - \lambda(1 \mid \sigma[n])|\} \geq \frac{1}{5}.
\]

So there can be no \( \lambda \in \mathcal{P} \) such that for every \( \sigma \) in a set of \( \lambda \)-measure one, there is a set \( K \) of natural numbers of asymptotic density 1, such that for sufficiently large \( n \in K \), \( \mu(1 \mid \sigma[n]) \) and \( \mu^f(1 \mid \sigma[n]) \) are both arbitrarily close to \( \lambda(1 \mid \sigma[n]) \)—i.e., there is no source \( \lambda \) that is weakly NC-learned by both \( \mu \) and \( \mu^f \).

Of course, it follows that \( \mu \) and \( \mu^f \) also (strongly) NC-learn disjoint sets of sources.

So we have no-free-lunch theorems for (weak) NC-learning: each forecaster, computable or not, (weakly) NC-learns an uncountable and dense but meagre set of sources; and for each (computable) forecaster there is another that (weakly) NC-learns a disjoint set of sources.

**Remark 4.2.** Lehrer and Smorodinsky (1996, Corollary 6) show that if \( \mu \in \mathcal{F} \) NC-learns \( \lambda \in \mathcal{P} \), then any nontrivial mixture of \( \mu \) with any \( \nu \in \mathcal{P} \) weakly NC-learns \( \lambda \). Ryabko and Hutter (2007, Proposition 10) show that this result is sharp: they give an example of of measures \( \mu, \nu, \) and \( \lambda \) where \( \mu \) NC-learns \( \lambda \) but any non-trivial mixture of \( \mu \) and \( \nu \) merely weakly NC-learns \( \lambda \). So there is no prospect of using the strategy of the proof of Proposition 4.7 above to prove an analogous result for NC-learning.

### 4.3. Forecasting of Computable Sources

Let us now specialize to problem of (strong, weak) NC-learning for computable forecasters facing data streams generated by computable sources.

For \( \mu \in \mathcal{F} \), we denote by \( \mathcal{NC}(\mu) \) the set of computable sources that are NC-learned by \( \mu \).

We use \( \mathcal{NC} \) to denote:

\[
\{ S \subset \mathcal{P} \mid \exists \mu \in \mathcal{F} \text{ with } S \subseteq \mathcal{NC}(\mu) \}.
\]

Let us likewise use \( \mathcal{NC}^s(\mu) \) and \( \mathcal{NC}^w(\mu) \) to denote the set of computable sources strongly NC-learned and weakly NC-learned by the forecasting machine \( \mu \) and use \( \mathcal{NC}^s \) and \( \mathcal{NC}^w \) to denote the class of subsets of \( \mathcal{P} \) that can be strongly/weakly NC-learned by some forecasting machine. It is immediate from the definitions that \( \mathcal{NC}^s \subseteq \mathcal{NC} \) and that \( \mathcal{NC} \subseteq \mathcal{NC}^w \). A variant on the proof of Proposition 3.9 shows that the latter containment is proper.

**Proposition 4.11.** \( \mathcal{NC} \subseteq \mathcal{NC}^w \).

**Proof.** Let \( V \) be the subset of \( \mathcal{P} \) consisting of \( \delta \)-function measures concentrated on computable binary sequences in which 0’s have vanishing asymptotic density. Let \( \mu \in \mathcal{F} \) be the measure that on input of any \( w \in B^n \), considers the chance of seeing a 0 next to be \( 2^{-n-1} \). We have \( \nu \in \mathcal{NC}^w(\mu) \). But suppose that \( \nu \in \mathcal{NC}(\nu) \) for some \( \nu \in \mathcal{F} \). We define \( \sigma \in \mathcal{C} \) as follows: \( \sigma \) is of the form \( 1^n0.1^n2.0.1^n3.0\ldots \) where where each \( n_j \) is chosen to be the smallest \( n \) larger than \( 2^j \) such that \( \nu(1^n0.1^n2.0\ldots1^{n_j-1}0.1^n) > .9 \) (such \( n_j \) must exist, since by assumption \( \nu \mathcal{NC} \)-learns each delta-function measure concentrated on a sequence containing only finitely many 0’s). The delta-function measure concentrated on \( \sigma \) is in \( V \). But \( \nu \notin \mathcal{NC}(\nu) \), since when fed \( \sigma \), there are infinitely many occasions on which \( \nu \) issues forecast probabilities for seeing a 0 next of less than 1, when the true chance is 1. \( \square \)

**Proposition 4.12.** For any \( \mu \in \mathcal{F} \), the following are dense subsets of \( \mathcal{P} \):

(a) \( \mathcal{NC}^s(\mu) \).
(b) The complement of $N\mathcal{C}^*(\mu)$ in $\mathfrak{P}$.
(c) $N\mathcal{C}(\mu)$.
(d) The complement of $N\mathcal{C}(\mu)$ in $\mathfrak{P}$.
(e) $N\mathcal{C}^w(\mu)$.
(f) The complement of $N\mathcal{C}^w(\mu)$ in $\mathfrak{P}$.

Proof. The claim concerning (a) follows via straightforward adaptation of the proof of Proposition 4.4, while that of (f) follows from Proposition 4.15 below. The other sets listed are supersets of (a) or (f). □

For any of our senses of probabilistic learning, for any computable forecaster, that forecaster succeeds in learning a countable infinity of computable sources and fails to learn a countable infinity of computable sources. So we have parity between the learnable and the unlearnable at the level of cardinality. And this parity again persists at the level of topology: a version of Sierpiński’s Theorem tells us that, up to homeomorphism, there is only one countable metrizable topological space without isolated points—see Neumann (1985, §2).

But it is straightforward to extend Mehlhorn’s notion of an effectively meagre set to the context of $\mathfrak{P}$, with the elements of the basis $\mathcal{W}$ for $\mathcal{P}$ of Section 4.2 playing the role that the basic open sets $B_w$ played in our discussion of effectively meagre subsets of $\mathcal{C}$. And, as in the case of next-value learning, we find that the for our species of next-chance learning, this notion allows us to isolate a sense in which failure is incomparably more common than success.

Recall that elements of $\mathcal{W}$ are specified by specifying a natural number $k$, a rational-valued measure $\mu \in \mathcal{P}_k$ (which is determined in turn by specifying the values that it assigns each $w \in \mathcal{B}_k$), and a rational $\varepsilon > 0$. So the elements of $\mathcal{W}$ can be effectively represented by binary strings. In the following definition we take such a coding scheme to be fixed.

**Definition 4.7.** Let $A$ be a subset of $\mathfrak{P}$. Let $f : \mathcal{W} \to \mathcal{W}$ be a computable function. Then $A$ is **effectively nowhere dense via $f$** if for each $W \in \mathcal{W}$:

i) $f(W) \subset W$;

ii) $A \cap f(W) = \emptyset$.

**Definition 4.8.** A subset $A$ of $\mathfrak{P}$ is **effectively nowhere dense** if there is a computable $F : \mathbb{N} \times \mathcal{W} \to \mathcal{W}$ such that:

i) for each $n \in \mathbb{N}$ there is an $A_n$ that is effectively nowhere dense via $F(n, \cdot) : \mathcal{W} \to \mathcal{W}$;

ii) $A = \bigcup A_n$.

The complement of an effectively meagre subset of $\mathfrak{P}$ is **effectively co-meagre**.

The proofs of Propositions 3.12 and 3.13 are easily adapted to yield:

**Proposition 4.13.** The set of effectively meagre subsets of $\mathfrak{P}$ is closed under the following operations:

(i) taking subsets;

(ii) taking finite unions;

(iii) taking effective unions.

**Proposition 4.14.** For any $W \in \mathcal{W}$, the set $W \cap \mathfrak{P}$ is not effectively meagre.

So it is again natural to consider the elements of effectively meagre subsets of $\mathfrak{P}$ to be incomparably less common than the elements of effectively co-meagre subsets of $\mathfrak{P}$.
Proposition 4.15. For any $\mu \in \mathfrak{F}$, $NC^w(\mu)$ is an effectively meagre subset of $\mathfrak{P}$.

Proof. A straightforward adaptation of the proof of Proposition 4.9, appealing to the fact that when $\mu$ is computable, the map $(n,W) \mapsto W^* \notin A_n$ used there is computable. \qed

So there is a natural sense in which, for any computable forecaster $\mu$, among computable sources, those weakly NC-learnable by $\mu$ are incomparably less common than those not NC-weakly learnable by $\mu$. And, a fortiori, those computable sources (strongly) NC-learnable by $\mu$ are incomparably less common than those not (strongly) NC-learnable by $\mu$. Learning in this setting is formidably difficult. And hard choices must be made: the proof of Proposition 4.10 above shows that each $\mu \in \mathfrak{F}$ has an evil twin $\mu^1 \in \mathfrak{F}$ such that the two weakly NC-learn disjoint sets of measures.

Remark 4.3. In Remark 3.3 above, we saw above that liberalizing our notion of NV-learning of computable sequences by allowing merely partial recursive extrapolators made an interesting difference: while every set in $N\mathcal{V}$ is effectively meagre, this is not true of every set in $N\mathcal{V}'$ (let alone $N\mathcal{V}''$).

So it is natural to consider enlarging the class of next-chance learners beyond the set of computable forecasters. For definiteness, let us consider the possibility of generalizing our notion of strong NC-learning. Let $X$ be a class of objects that includes $\mathfrak{F}$ as a subset and for which the definition of strong NC-learning above makes sense when we quantify over members of $X$ rather than measures in $\mathcal{F}$. For $\mu \in X$, let $NC^*_X(\mu)$ be the set of computable sources learned in this generalized sense by $\mu$ and let $S \subset \mathfrak{P}$ count as a member of $NC^*_X$ just in case there is some $\mu \in X$ that learns every member of $S$ in this sense. A trivial example: $\mathfrak{P} \in NC^*_\mathcal{F}$, since $\mathcal{F}$, unlike $\mathfrak{F}$, includes a measure $\mu$ that is a linear superposition of all the computable measures (so that each computable measure is absolutely continuous with respect to $\mu$). We are in search of more interesting examples.

Recall that the Carathéodory Extension Theorem allows us to identify probability measures on $\mathcal{C}$ with maps $\mu : \mathcal{B}^* \to [0,1]$ satisfying $\mu(\emptyset) = 1$ and $\mu(w) = \mu(w,0) + \mu(w,1)$ for all $w \in \mathcal{B}^*$, so that $\mu \in \mathcal{P}$ is s computable if and only if there exists a computable $F : \mathcal{B}^* \times \mathbb{N} \to \mathbb{Q}$ such that $|\mu(w) - F(w,n)| < 2^{-n}$ for all $w \in \mathcal{B}^*$ and $n \in \mathbb{N}$.

One way to loosen our restriction to computable measures would to countenance computable semi-measures: computable maps $\mu : \mathcal{B}^* \to [0,1]$ satisfying $\mu(\emptyset) \leq 1$ and $\mu(w) \geq \mu(w,0) + \mu(w,1)$ for all $w \in \mathcal{B}^*$. But this gets us nowhere. Normalizing a computable semi-measure yields a computable measure—and any source strongly learned by the semi-measure will be learned by its normalization.\footnote{See the discussion of Solomonoff normalization in \cite{LiVitanyi2019} §4.5.3}

Another way to loosen our restrictions would be to allow merely semi-computable measures, where $\mu : \mathcal{B}^* \to [0,1]$ is lower semi-computable if and only if there exists a partial recursive $\phi : \mathcal{B}^* \times \mathbb{N} \to [0,1]$ with $\mu(w) = \lim_{\ell \to \infty} \phi(w,\ell)$ and $\phi(w,k) \leq \phi(w,k+1)$, for all $w \in \mathcal{B}^*$ and $k \in \mathbb{N}$. But this again leads nowhere: every lower semi-computable measure on $\mathcal{C}$ is computable.\footnote{\cite{LiVitanyi2019} Lemma 4.5.1.}

However, if we allow arbitrary lower semi-computable semi-measures to count as next-chance learners, then we make spectacular gains: Solomonoff (1978) constructed lower semi-computable semi-measures that strongly NC-learn every computable source. So if we let $\text{Sol}$ be the class of lower semi-computable semi-measures, then $NC_{sd}$ is the power set of $\mathfrak{P}$. But this vast gain in learning power comes comes at a price: although Solomonoff’s
learners are lower semi-computable as maps that associate probabilities with binary strings, they are merely limit computable as maps that take as input a binary string and give as output the conditional probability for the next bit to be a 0 or a 1 (the ratio of two lower semi-computable numbers need not be lower semi-computable). Indeed, Sterkenburg (2019) shows that if \( X \) contains only objects that induce semi-computable conditional probability maps, then \( \mathcal{P} \notin \mathcal{NC}_X^s \). Question: is it possible to generalize Proposition 4.15 to show that for such \( X \), every set in \( \mathcal{NC}_X^s \) is effectively meagre?

5. Discussion

Over the course of the last century, it became widely understood that successful inductive learning is possible only against a background of biases that favour some hypotheses over others. No-free-lunch results substantiate this insight. If we don’t presuppose anything about the binary sequence being revealed to us, then we face a formidable difficulty problem: no matter what approach to learning we adopt, the situations that we might face, those in which we fail are incomparably more common than those in which we succeed. And no approach dominates all rivals in its range of success: for any approach, there are others that succeed in situations in which the given one fails; indeed, for any approach, there is another that succeeds in a disjoint set of situations. To adopt an approach to learning is to make a bet about what the world is like.

No-free-lunch results place upper bounds on our reasonable ambitions. Suppose that one is interested in the question: Why should someone interested in arriving at the truth proceed inductively (expecting the future to be like the past) rather than counter-inductively? Consider how this question looks in the simplest of our contexts, in which an agent being shown a binary sequence bit by bit aims to eventually be able to correctly predict each new bit on the basis of the bits seen so far. Here each method of learning succeeds on a countable dense subspace of the space of binary sequences. And all such subspaces are isomorphic (Sierpiński’s Theorem again). So unless we impose more structure on our problem, we have parity between the set of possibilities in which an inductive extrapolator \( m \) is successful and the set of possibilities in which counter-inductive extrapolator \( m' \) is successful.

Of course, the results developed above presuppose that we are operating in an austere setting—one in which we countenance arbitrary (computable) data streams or data streams generated by sampling from arbitrary (computable) probability measures. In more tightly constrained settings, learning becomes tractable—e.g., if one knows that the data stream is generated by a Bernoulli measure, then it is a straightforward task to use the data to successfully estimate the relevant parameter. But this observation illustrates rather than undercuts the perspective of the preceding paragraphs, making the point that although a universal learning algorithm is an impossibility, learning becomes possible when sufficiently strong presuppositions are in play. Of course, one would ultimately like to know more about

---

46See Leike and Hutter (2015) and Sterkenburg (2019).
47See, e.g., Jeffreys (1933, 524 f.), Kuhn (1963, 3 ff.), Chomsky (1965, §1.8), and Hempel (1966, §2.3). Arguably, this theme can be found already in Leibniz—see Item 6 in Loemker (1969).
48That is one of their uses. Of the results canvassed in Section II above, Putnam’s was designed to expose a serious flaw in objective Bayesian approaches in the tradition of Carnap (1943) while that of Shalev–Shwartz and Ben–David was devised to provide an elegant motivation for the definition of VC-dimension.
49Some formal learning theorists take the view that extrapolators prone to mind-changes are to be eschewed (thereby adding further structure to our sort of problem) and establish, in some contexts, a link between counter-inductive behaviour and mind changes—see, e.g., Kelly (2011) and Lin (2018).
where the boundaries lie of the class of learning problems in which failure is typical and of the class of learning problems in which success is typically achievable.

The results developed above are absolute in the sense that they do not presuppose the choice of a privileged measure on Cantor space or on the space of probability measures on Cantor space. But most of them do depend on the choice of topology. For the results concerning (weak) learning of sequences by extrapolators, this is not very worrying. In the vast majority of applications in statistics, economics, and computer science, the space of binary sequences is equipped with the product topology. And with good reason: this topology can be thought of as the topology of point-wise convergence and motivated by thinking of binary sequences as encoding real numbers in the usual way. The situation is not quite as straightforward with the space of probability measures on Cantor space. Certainly, the weak topology is extremely natural—but it is only one of several natural options. So it is natural to wonder whether the intractability of our learning problems would hold under other reasonable choices of topology.

References

Angluin, Dana and Carl Smith (1983) “Inductive Inference: Theory and Methods.” ACM Computing Surveys 15: 237–269.

Báez-Duarte, Luis (1970) “C(X)* and Kolmogorov’s Consistency Theorem for Cantor Spaces.” Studies in Applied Mathematics 49: 401–403.

Bārzdžiņš, Jānis (1972) “Prognostication of Automata and Functions.” In C.V. Frieman (ed.), Information Processing ’71, North–Holland, volume 1, 81–84.

Bārzdžiņš, Jānis and Rūsiņš Freivalds (1972) “On the Prediction of General Recursive Functions.” Soviet Mathematics Doklady 13: 1224–1228.

Billingsley, Patrick (1999) Convergence of Probability Measures, second edition. Wiley.

Blackwell, David and Lester Dubins (1962) “Merging of Opinions with Increasing Information.” The Annals of Mathematical Statistics 33: 882–886.

Blum, Lenore and Manuel Blum (1975) “Toward a Mathematical Theory of Inductive Inference.” Information and Control 28: 125–155.

Carnap, Rudolf (1945) “On Inductive Logic.” Philosophy of Science 12: 72–97.

Case, John and Carl Smith (1983) “Comparison of Identification Criteria for Machine Inductive Inference.” Theoretical Computer Science 25: 193–220.

Chomsky, Noam (1965) Aspects of the Theory of Syntax. MIT Press.

Dasgupta, Abhijit (2014) Set Theory: With an Introduction to Real Point Sets. Springer.

Dekel, Eddie and Yossi Feinberg (2006) “Non-Bayesian Testing of a Stochastic Prediction.” The Review of Economic Studies 73: 893–906.

Dubins, Lester and David Freedman (1964) “Measurable Sets of Measures.” Pacific Journal of Mathematics 14: 1211–1222.

Friedman, Lance, Rūsiņš Freivalds, William Gasarch, Martin Kummer, Stuart Kurtz, Carl Smith, and Frank Stephan (1998) “On the Relative Sizes of Learnable Sets.” Theoretical Computer Science 197: 139–156.

Gold, E. Mark (1967) “Language Identification in the Limit.” Information and Control 10: 447–474.

See Fortnow et al. (1998) for some results of this kind for the problem of identification of sequences.
Hempel, Carl G. (1966) *Philosophy of Natural Science*. Prentice-Hall.
Ho, Yu-Chi and David Pepyne (2002) “Simple Explanation of the No-Free-Lunch Theorem and its Implications.” *Journal of Optimization Theory and Applications* 115: 549–570.
Hutter, Marcus (2007) “On Universal Prediction and Bayesian Confirmation.” *Theoretical Computer Science* 384: 33–48.
Jeffreys, Harold (1933) “Probability, Statistics, and the Theory of Errors.” *Proceedings of the Royal Society of London. Series A* 140: 523–535.
Jockusch, Carl and Paul Schupp (2012) “Generic Computability, Turing Degrees, and Asymptotic Density.” *Journal of the London Mathematical Society* 85: 472–490.
Kalai, Ehud and Ehud Lehrer (1994) “Weak and Strong Merging of Opinions.” *Journal of Mathematical Economics* 23: 73–86.
Kechris, Alexander (1995) *Classical Descriptive Set Theory*. Springer.
Kelly, Kevin (2011) “Simplicity, Truth, and Probability.” In Prasanta Bandyopadhyay and Malcolm Forster (eds.), *Philosophy of Statistics*. Elsevier, 983–1024.
Koumoullis, George (1996) “Baire Category in Spaces of Measures.” *Advances in Mathematics* 124: 1–24.
Kuhn, Thomas S. (1963) *The Structure of Scientific Revolutions*. University of Chicago Press.
Lehrer, Ehud and Rann Smorodinsky (1996) “Merging and Learning.” In Thomas Ferguson, Lloyd Shapley, and James MacQueen (eds.), *Statistics, Probability and Game Theory: Papers in honor of David Blackwell*. Institute of Mathematical Statistics, 147–168.
Leike, Jan and Marcus Hutter (2015) “On the Computability of Solomonoff Induction and Knowledge-Seeking.” In Kamalika Chaudhuri, Claudio Gentile, and Sandra Zilles (eds.), *International Conference on Algorithmic Learning Theory*, Springer, 364–378.
Li, Ming and Paul Vitányi (2019) *An Introduction to Kolmogorov Complexity and its Applications*, fourth edition. Springer.
Lin, Hanti (2018) “Modes of Convergence to the Truth: Steps Toward a Better Epistemology of Induction.” Forthcoming in *The Review of Symbolic Logic*.
Lisagor, L.R. (1981) “The Banach-Mazur Game.” *Mathematics of the USSR-Sbornik* 38: 201–216.
Loemker, Leroy (ed.) (1969) *Gottfried Wilhelm Leibniz: Philosophical Papers and Letters*, second edition. D. Reidel.
Marczewski, Edward and Roman Sikorski (1949) “Remarks on Measure and Category.” *Colloquium Mathematicum* 2: 13–19.
Mehlhorn, Kurt (1973) “On the Size of Sets of Computable Functions.” In Ronald Book, Allan Borodin, Forbes Lewis, Amar Mukhopadhyay, Arnold Rosenberg, Raymond Strong, and Jeffrey Ullman (eds.), *14th Annual Symposium on Switching and Automata Theory*, IEEE Computer Society Publications Office, 190–196.
Neumann, Peter (1985) “Automorphisms of the Rational World.” *Journal of the London Mathematical Society* 2: 439–448.
Nies, André (2009) *Computability and Randomness*. Oxford University Press.
Noguchi, Yuichi (2015) “Merging with a Set of Probability Measures: A Characterization.” *Theoretical Economics* 10: 411–444.
Oxtoby, John (1957) “The Banach-Mazur Game and Banach Category Theorem.” In Melvin Dresher, Albert Tucker, and Philip Wolfe (eds.), *Contributions to the Theory of Games*, volume III. Princeton University Press, 159–163.
Oxtoby, John (1980) Measure and Category: A Survey of the Analogies between Topological and Measure Spaces, second edition. Springer.

Parthasarathy, Kalyanapuram (1967) Probability Measures on Metric Spaces. Academic Press.

Podnieks, Kārlis (1974) “Comparing Various Concepts of Function Prediction.” Theory of Algorithms and Programs 210: 68–81 (in Russian).

Putnam, Hilary (1963a) “‘Degree of Confirmation’ and Inductive Logic.” In Paul Schilpp (ed.), The Philosophy of Rudolf Carnap. Open Court, 761–783.

Putnam, Hilary (1963b) Probability and Confirmation. US Information Agency.

Reimann, Jan (2008) “Effectively Closed Sets of Measures and Randomness.” Annals of Pure and Applied Logic 156: 170–182.

Ryabko, Daniil and Marcus Hutter (2007) “On Sequence Prediction for Arbitrary Measures.” In 2007 IEEE International Symposium on Information Theory, IEEE, 2346–2350.

Sandroni, Alvaro and Rann Smorodinsky (1999) “The Speed of Rational Learning.” International Journal of Game Theory 28: 199–210.

Shalev-Shwartz, Shai and Shai Ben-David (2014) Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Solomonoff, Ray (1964) “A Formal Theory of Inductive Inference. I.” Information and Control 7: 1–22.

Solomonoff, Ray (1978) “Complexity-Based Induction Systems: Comparisons and Convergence Theorems.” IEEE Transactions on Information Theory 24: 422–432.

Sterkenburg, Tom (2019) “Putnam’s Diagonal Argument and the Impossibility of a Universal Learning Machine.” Erkenntnis 84: 633–656.

Szpilrajn, Edward (1934) “Remarques sur les fonctions complètement additives d’ensemble et sur les ensembles jouissant de la propriété de Baire.” Fundamenta Mathematicae 22: 303–311.

von Luxburg, Ulrike and Bernhard Schölkopf (2011) “Statistical Learning Theory: Models, Concepts, and Results.” In Dov Gabbay, Stephan Hartmann, and John Woods (eds.), Inductive Logic. Elsevier, 651–706.

Wolpert, David (2002) “The Supervised Learning No-Free-Lunch Theorems.” In Rajkumar Roy, Mario Köppen, Seppo Ovaska, Takeshi Furuhashi, and Frank Hoffmann (eds.), Soft Computing and Industry. Springer, 25–42.

Wolpert, David and William Macready (1997) “No Free Lunch Theorems for Optimization.” IEEE Transactions on Evolutionary Computation 1: 67–82.

Zenil, Hector (ed.) (2013) A Computable Universe: Understanding and Exploring Nature as Computation. World Scientific.

Zindulka, Ondřej (1999) “Killing Residual Measures.” Journal of Applied Analysis 5: 223–238.

Department of Philosophy, University of Michigan belot@umich.edu