A statistical learning approach to a problem of induction

Kino Zhao
yutingz3@uci.edu

University of California, Irvine
Logic and Philosophy of Science

(Draft updated December 7, 2018)

Abstract

At its strongest, Hume's problem of induction denies the existence of any well justified assumptionless inductive inference rule. At the weakest, it challenges our ability to articulate and apply good inductive inference rules. This paper examines an analysis that is closer to the latter camp. It reviews one answer to this problem drawn from the VC theorem in statistical learning theory and argues for its inadequacy. In particular, I show that it cannot be computed, in general, whether we are in a situation where the Vapnik-Chervonenkis (VC) theorem can be applied for the purpose we want it to.

Hume's problem of induction can be analyzed in a number of different ways. At the strongest, it denies the existence of any well justified assumptionless inductive inference rule. At the weakest, it challenges our ability to articulate and apply good inductive inference rules. This paper examines an analysis that is closer to the latter camp. It reviews one answer to this problem drawing from a theorem in statistical learning theory and argues for its inadequacy.

The particular problem of induction discussed in this paper concerns what Norton (2014) calls a formal theory of induction, where "valid inductive inferences are distinguished by their conformity to universal templates" (p.673). In particular, I focus on the template that is often called *enumerative induction*. An inductive argument of this type takes observations made from a small and finite sample of cases to be indicative of features in a large and potentially infinite population. The two hundred observed swans are white, so all swans are white. Hume argues that the only reason we think a

rule like this works is because we have observed it to work in the past, resulting in a circular justification.

Nevertheless, this kind of inductive reasoning is vital to the advancement of a scientific understanding of nature. Most, if not all, of our knowledge about the world is acquired through the examination of only a limited part of the world. The scientific enterprise relies on the assumption that at least some of such inductive processes generate knowledge. With this assumption in place, a weak problem of induction asks whether we can reliably and justifiably differentiate the processes that do generate knowledge from the ones that do not. This paper discusses this weak problem of induction in the context of statistical learning theory.

Statistical learning theory is a form of supervised machine learning that has not received as much philosophical attention as it deserves. In a pioneering treatment of it, Harman and Kulkarni (2012) argue that one of the central results in statistical learning theory – the result on Vapnik-Chervonenkis (VC) dimensions – can be seen as providing a new kind of answer to a problem of induction by providing a principled way of deciding if a certain procedure of enumerative induction is reliable. The current paper aims to investigate the plausibility of their view further by connecting results about VC dimension in statistical learning with results about NIP models in the branch of logic called model theory. In particular, I argue that even if Harman and Kulkarni succeed in answering the problem of induction with the VC theorem, the problem of induction only resurfaces at a deeper level.

The paper is organized as follows: section 1 explains the relevant part of statistical learning theory, the VC theorem, and the philosophical lessons it bears. Section 2 introduces the formal connection between this theorem and model theory and proves the central theorem of this paper. Section 3 concludes with philosophical reflections about the results.

1 Statistical learning theory

The kind of problems that is relevant for our discussion of VC dimensions is often referred to as classification problems that are irreducibly stochastic. In a classification problem, each individual is designated by its k-many features such that it occupies somewhere along a k-dimensional feature space, χ. The goal is to use this information to classify potentially infinitely many such individuals into finitely many classes. To

give an example, consider making diagnoses of people according to their test results from the k tests they have taken. The algorithm we are looking for needs to condense the k-dimensional information matrix into a single diagnosis: sick or not. The algorithm can be seen as a function $f : \chi \to \{0, 1\}$, where 1 means sick and 0 means not. For reasons of simplicity, I will follow the common practice and only consider cases of binary classification.

By "irreducibly stochastic", I mean that the target function f cannot be solved analytically. This might be because the underlying process is itself stochastic – it is possible for two people with exact same measures on all tests to nevertheless differ in health condition – or because the measurements we take have ineliminable random errors. This means that even the best possible f will make some error, and so the fact that a hypothesis makes errors in its predictions does not in itself count against that hypothesis. Instead, a more reasonable goal to strive towards is to have a known, preferably tight, bound on the error rate of our chosen hypothesis.

What makes this form of statistical learning "supervised learning" is the fact that the error bound of a hypothesis is estimated using data points whose true classes are known. Throughout this paper, I will use D to denote such a dataset. D can have any cardinality, but the interesting cases are all such that D is of finite size. Recall that the feature (or attribute) space χ denotes the space of all possible individuals that D could have sampled, so that $D \subset \chi$. I understand a hypothesis to be a function $h : \chi \to \{0, 1\}$. A set of hypotheses $\mathcal{H}$ is a set composed of individual hypotheses. Usually, the hypotheses are grouped together because they share some common features, such as all being linear functions with real numbers as parameters. This observation will become more relevant later.

One obvious way of choosing a good hypothesis from $\mathcal{H}$ is to choose the one that performs the best on D. I will follow Harman and Kulkarni (2012) and call this method enumerative induction, for it bears some key similarities with Hume's description of the observation of swans. This method is inductive because it has the ampliative feature of assuming that the chosen hypothesis will keep performing well on individuals outside of D. The question we are interested in is: how do we know this? What justifies the claim that the hypothesis performs well on D will perform well outside of D too? The answer that will be examined in this section and throughout the rest of the paper is that we know this claim to be true when we are in a situation where $\mathcal{H}$ has finite VC dimension, and the VC-theorem justifies this claim.

To define the error rate of a hypothesis, recall the "ideal function" f mentioned in the introduction. Recall also that f classifies individuals from χ into $\{0, 1\}$, and f is imperfect. Nevertheless, since the process from χ to the classes is irreducibly stochastic, f is as good as we can hope for. Therefore, f will serve as our standard for the purpose of calculating the error rate of a hypothesis. Note that the hypotheses we are assessing are all from $\mathcal{H}$, our hypothesis set, but f need not be in $\mathcal{H}$.

Suppose D is of size N, and $x_1, \ldots, x_N \in D$. For each $h \in \mathcal{H}$ and $i \in [1, N]$, consider the random variable $X_i : \chi^N \to \{0, 1\}$ defined by

$$X_i\big(h(x_1, \ldots, x_N)\big) = \begin{cases} 1 & \text{if } h(x_i) \neq f(x_i), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Intuitively, $X_i = 1$ if the hypothesis we are evaluating, h, gives a different (and hence wrong) verdict on x_i than the target function f, and 0 otherwise. Assume $X_1, \ldots, X_N$ are independent, which is to say that making a mistake on one data point does not make it more or less likely for h to make a mistake on another one. This is typical if D is obtained through random sampling. Further assume $X_1, \ldots, X_N$ are identically distributed, which means that for any X_i and X_j in the sequence, $EX_i = EX_j$. This allows the error "rate" of h across multiple data points to be meaningfully computed.

Let $\overline{X} = \frac{1}{N}(\sum_{i=1}^{N} X_i)$, which is the measured mean error, and $\mu = E\overline{X}$, which is the expected mean error. I will follow Abu-Mostafa et al. (2012) in calling the former the *in-data error*, or E_{in}, and the latter *out-data error*, or E_{out}. To flesh out the relationship between these two values more clearly, we define

$$E_{in}(h) = \overline{X} = \frac{1}{N} \sum_{i=1}^{N} [\![h(\mathbf{x}_i) \neq f(\mathbf{x}_i)]\!] \tag{2}$$

$$E_{out}(h) = \mu = \mathbb{P}_N(h(\mathbf{x}) \neq f(\mathbf{x})) \tag{3}$$

Intuitively, the in-data error is the evidence we have about the performance of h, and the out-data error is the expectation that h will hold up to its performance. The amplification comes in when we claim that E_{out} is not very different from E_{in}. I will call the difference between E_{in} and E_{out} the *generalization error*.

For any single hypothesis, and for any error tolerance $\epsilon > 0$, Hoeffding (1963, p.16) proved a result called the *Hoeffding inequality* (see also Lin and Bai 2010, p. 70, and

Pons 2013, p. 205), which states that, under the assumption that the error rate for each data point is independent and identically distributed, we have (in the notations introduced above)

$$\mathbb{P}_N(|E_{in}(h) - E_{out}(h)| \geq \epsilon) \leq 2e^{-2\epsilon^2 N} \tag{4}$$

This inequation says that the probability of having a large generalization error in the assessment of a single hypothesis is bounded by $2e^{-2N\epsilon^2}$, which is a function of the size of the dataset, N, and the error tolerance ϵ.

Once we establish a bound in the case of a single hypothesis, we can get a similar bound for finitely many such hypotheses. The reason we cannot simply apply the Hoeffding inequality to our preferred hypothesis is that it requires us to pick a hypothesis before we compute its error rate from the data. This will not help us if we need to use data to do the picking. Instead, we need to make sure *any* hypothesis we pick out will have low enough generalization error, before we can trust the method (of enumerative induction) we use to pick.

Since we assume that the error rate of one hypothesis is independent of another, the probability of any of the finitely many hypotheses we are considering having a large generalization error is just going to be the union of the probability of each one of them does. In symbolic form, suppose there are $1 \leq M < \infty$ many hypotheses in $\mathcal{H}$, then we have

$$\mathbb{P}(\max_{h\in\mathcal{H}} |E_{in}(h) - E_{out}(h)| \geq \epsilon) = \mathbb{P}(\exists h \in \mathcal{H} | E_{in}(h) - E_{out}(h)| \geq \epsilon) \leq 2Me^{-2\epsilon^2 N} \tag{5}$$

While this bound may seem "loose", it serves our purpose when we have a reasonably small M or a reasonably large N.

This simple calculation becomes tricky, however, when $\mathcal{H}$ contains infinitely many hypotheses. If we replace M with infinity, then the upper bound stops being a bound, because $2Me^{-2\epsilon^2 N}$ grows to infinity as M does. This is where the VC dimension of $\mathcal{H}$ comes to play.

To understand the role of VC dimensions, define

$$\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \{(h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N)) \ \mid \ h \in \mathcal{H}\} \tag{6}$$

which is the set of all verdicts given by $\mathcal{H}$ on dataset D. If some hypotheses agree with each other on the classification of every data point, then their verdicts would be represented by the same tuple. This means that the cardinality of the set of verdicts

may be much smaller if $\mathcal{H}$ is very homogeneous. Moreover, different datasets of the same cardinality may elicit more or fewer different verdicts from $\mathcal{H}$. Define

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1,\cdots,\mathbf{x}_N \in \chi} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)| \tag{7}$$

as the max number of different verdicts $\mathcal{H}$ can generate from any dataset of cardinality N.

If all possible classifications of D have been represented in $\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)$, then we have $m_{\mathcal{H}}(N) = 2^N$. When this happens, we say that the hypothesis set $\mathcal{H}$ *shatters* the dataset D. Define the *VC dimension of* $\mathcal{H}$ to be the maximum N such that $m_{\mathcal{H}}(N) = 2^N$. In other words, it is the maximum number N such that there exists a dataset D of size N that is shattered by $\mathcal{H}$. If $m_{\mathcal{H}}(N) = 2^N$ holds for all N, then we say the VC dimension is infinite. Let's call a hypothesis set $\mathcal{H}$ VC-learnable if it has finite VC dimension.

Very roughly, the VC dimension of a hypothesis set tracks the maximum number of hypotheses that are still distinguishable from each other with respect to their verdicts on data. This means that, if we consider any more hypotheses, some of them will always agree with some others on all of the classifications they give to all possible data points, and so if one has low generalization error, the others will, too. The VC generalization bound is given as follows (Abu-Mostafa et al., 2012, p.53)

$$\mathbb{P}_N[\![\big(E_{out}(h) - E_{in}(h)\big) \leq \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}]\!] \geq 1 - \delta \tag{8}$$

where δ is the uncertainty tolerance. If $\mathcal{H}$ has an infinite VC dimension, then no such upper bound can be found. Notice that, holding everything else equal, increasing N brings the right-hand side down, which means that increasing data size allows us to make a better estimate of E_{out} with the same uncertainty tolerance. One can further show that

$$\lim_{N\to\infty} \mathbb{P}_N(\max_{h\in\mathcal{H}} |E_{in}(h) - E_{out}(h)| = 0) > 1 - \delta \tag{9}$$

for all $\delta > 0$. This means that, when $\mathcal{H}$ is either finite or has finite VC dimension, we can justifiably claim enumerative induction to be a reliable process that can pick out a good hypothesis from $\mathcal{H}$.

What makes this theorem especially powerful is not just that it shows how the error rates converge in the limit, but also that the convergence is uniform. What is

practically useful for statisticians is not so much that, if we have infinite data, we can figure out the true error rate of our hypothesis, but that, as soon as we know how many data points we have and the VC dimension of $\mathcal{H}$, we know precisely how confident we should be of our estimation of the error rate.

In what sense does this theorem answer a problem of induction? According to the analysis in Harman and Kulkarni (2012), this theorem defines precise conditions (i.e., ones where $\mathcal{H}$ has finite VC dimension) under which a particular inductive method (i.e., supervised learning in classification problems) is reliable. To the extent that we are concerned with the "easy" problem – the practical problem – of induction, the VC theorem does seem to provide a kind of answer we are looking for. In the next section, I challenge the applicability of this answer. In particular, I show that we can never know in general if we are in a situation where the above answer is applicable.

2 Finiteness of VC dimensions is uncomputable

A preliminary observation about the finiteness requirement is that we do not have a good grasp of what it means. What is the difference between these two sets of hypotheses such that one has finite VC dimension and the other does not? To put this point more concretely, we know that polynomial functions with arbitrarily high degrees have finite VC dimension, whereas the set of formulas with the sine function has infinite VC dimension. What is the difference between them? If we have a problem that can be reasonably formulated as polynomials or with a sine function, do we have good principled reasons why we should formulate it in one way rather than another?

Surprisingly, model theory in logic might help shed light on this question. It turns out that the concept of NIP theories corresponds to the class of hypothesis sets with finite VC dimensions. A theorem provably equivalent to the VC theorem was independently proved by the model theorist Shelah about these NIP theories and the corresponding NIP models. This connection was first recognized by Laskowski (1992). Interestingly, with the real numbers as their underlying domains, models with the usual plus and multiplication signs are NIP, whereas adding the sine curve makes them not NIP. This suggests that we can ask the same questions we would like to ask about our statistical hypothesis sets in model theory, which has a richer structure that is better understood independently.

In the previous section we discussed how the idea of "distinguishable hypotheses"

is important for the VC theorem. If a hypothesis set has finite VC dimension, we can think of it as having finitely many *distinguishable* hypotheses, even if it in fact has infinitely many. Intuitively speaking, if our dataset is "large enough" that not every combination of verdicts is representable with our hypotheses, then we can talk about which hypothesis is truly better than its competitors, as opposed to accidentally matching the specific data points. Having finite VC dimension ensures that there exist finite datasets that are "large enough". If a hypothesis set has finite VC dimension, let us call the set *VC-learnable*.

The corresponding concept in model theory relies on the same idea of distinguishability. Intuitively, if a formula is NIP – has the not-independent property – then there exists a natural number n such that no set larger than that number can be defined using this formula. A model is NIP just in case all of its formulas are (a formal definition is presented in Appendix A; for more formal details, see Simon, 2015).

We can then treat each hypothesis set as a formula defined on some domain. Laskowski (1992) shows that a hypothesis set is VC-learnable just in case the corresponding formula is NIP. What makes this correspondence especially useful is that model theorists have devoted a lot of efforts into determining which model is NIP. Once we know of a model that it's NIP, we also know that any hypothesis sets formulated using the language and domain of this model are VC-learnable.

For example, there is a group of models called *o-minimal*, which roughly means that all the definable subsets of the domain are finite unions of simple topological shapes like intervals and boxes. It suffices for our purposes to note that all o-minimal models are NIP (van den Dries, 1998, p. 90). As it happens, the real numbers with just addition and multiplication are o-minimal (van den Dries, 1998, p. 37). This means that any hypothesis set consisted of addition, multiplication, and the real numbers are going to have finite VC dimension. Similarly, the real numbers with addition, multiplication, and exponentiation is also o-minimal (Wilkie, 1996). This means that all sets of polynomials are VC-learnable.

As alluded to already, the real numbers with the sine function added are not NIP. This is roughly because, with the sine function, we can define copies of the integers using the set $\{x \in \mathbb{R} : \sin(x) = 0\}$, which allows us to define all of second-order arithmetic, and second-order arithmetic allows coding of arbitrary finite sets. As expected, this is reflected in statistical learning theory by the fact that the set of sine functions has infinite VC dimension, and so is not VC-learnable.

Another important observation from model theoretic investigations on NIP theory is that there seem to be no easy test for when an expansion of the real numbers is NIP. Although the relationship between the NIP property and properties like o-minimal and stable (a set of structures that are not o-minimal but are NIP) is well-researched and understood, there is no uniform way of telling where exactly a model lies (see, e.g., Miller, 2005[1]).

The statistical learning community echoes this difficulty with the observation that "it is not possible to obtain the analytic estimates of the VC dimension in most cases" (Shao et al., 2000; also see Vapnik et al., 1994). Recall that the VC dimension decides how big a dataset is "big enough". If the view is that enumerative induction is a reliable method when we are confident (i.e., low δ) that its assessment of hypotheses generalizes (i.e., low ϵ) and the VC theorem is supposed to guarantee this, then our inability to analytically solve the VC dimension of a given hypothesis set seems deeply handicapping.

To make the matter worse, it turns out that even knowing when we do have finite VC dimension is not a straightforward task, as witnessed by the following theorem, whose proof is given in Appendix A

Theorem 1. *The set $\{\varphi(x,y) : \varphi(x,y)$ is $NIP\}$, where $\varphi(x,y)$ is formulated in the language of arithmetic with addition and multiplication, is not decidable. In particular, this set computes $\emptyset^{(2)}$, the second Turing jump of the empty set.*

What this theorem tells us is that, in general, there is no effective procedure we can follow that can tell us, for any 2-place formula $\varphi(x,y)$, if it's NIP. With Laskowski's result, this means that we cannot compute, in general, if a given hypothesis set is VC-learnable either.

The specific way in which the set of all NIP formulas is uncomputable is significant also. For some time now, philosophers who study knowledge and learning from a formal perspective have placed a lot of emphasis on learning in the limit. Kelly (1996, p.52), for example, argues that the concept of knowledge (as opposed to, say, mere belief) implies that the method of generating such beliefs is stable in the limit. He then argues that the best way to formalize the notion of "stability in the limit" is to understand it as computable in the limit. Relatedly, a venerable tradition of formal learning

[1]Technically, Miller is interested in dichotomy theorems which establish either that an expansion of the reals is o-minimal or that it defines second-order arithmetic. As mentioned before, the former suffices for being NIP, and the latter suffices for being not NIP.

theory following Gold (1967) has explored extensively the conditions under which a noncomputable sequence may or may not be approximated by a computable sequence making only finitely many mistakes (cf. Osherson et al., 1986; Jain et al., 1999). From this perspective, it seems we might still be able to claim knowledge of what is or isn't knowable if we can compute the set of NIP formulas in the limit. Unfortunately, this latter task cannot be accomplished. This is because that, in order for a sequence to be approximable in the limit by another sequence, it cannot be harder than the first Turing jump of the sequence used to approximate it (Soare, 1987, p.57; see also Kelly, 1996, p.280). This means that something that is at least as hard as the second Turing jump cannot be approximated by a computable sequence.

To recapitulate the dialectic so far: an easy problem of induction asks us to identify and then justify the conditions under which a given ampliative method is reliable. The VC theorem gives one answer: supervised statistical learning from data is reliable just in case the hypothesis set has finite VC dimension. However, it turns out that we cannot, in general, decide if a hypothesis set is VC-learnable.

Can we judge our $\mathcal{H}$ on a case-by-case basis? Once we fix an $\mathcal{H}$, we can usually tell if it has finite VC dimension, and we can develop methods of empirically estimating its VC dimension using multiple datasets with varying sizes. However, this seems to just push the same problem to a deeper level. The problem that a method "sometimes is reliable, sometimes isn't", is solved by specifying a condition under which it always is reliable. But the problem that the condition "sometimes occurs, sometimes doesn't" seems to have no simple solution. In fact, the above theorem says that the latter problem has no solution.

3 Conclusion

A reasonable conclusion to draw from the discussions we've had so far, I think, is that the VC theorem still does not give us the kind of robust reliability we need to answer a question with some scope of philosophical generality. As is typical of answers people give to problems of induction, as soon as a rule is formulated, a question arises concerning its applicability. Similarly, what started out as a concern over the robustness of the method of enumerative induction turns into a concern over the robustness of the identifiable condition (i.e., the VC-learnable condition) under which enumerative induction is justified to be reliable.

A related question concerns the distinction, if there is one, between the cases where $\mathcal{H}$ has infinite VC dimension and cases where it has a VC dimension so large that it's impractical for us to make use of it. There is a sense in which the case of an infinite VC dimension fails *in principle*, whereas the case of a very large VC dimension only fails in *practice*. However, it is often impossible to analytically solve the VC dimension of a hypothesis set even if we do know that it's VC-learnable. Together with the result that we cannot test if a case is VC-learnable *in principle*, it seems to suggest that any information we might gain from the distinction between failing in principle and failing in practice will not be very informative, since we often can't tell which case we are in.

The philosophical difficulties discussed above raise an interesting question of how the practitioners view the same obstacle. Perhaps the way out is to accept a 'piecemeal' solution after all. It seems that when the VC dimension is small, we can often know both that it is finite, and that it is small. Theorists have also developed ways of estimating VC dimension using multiple datasets (see, e.g., Vapnik et al., 1994 and Shao et al., 2000). It seems that, as it often happens, philosophical problems are much more manageable when we do not look for principled solutions.

Acknowledgement

I would like to express my gratitude towards Sean Walsh for his supervision, as well as towards the participants in the 2016 Logic Seminar and attendees of the Society for Exact Philosophy 2017 meeting for their valuable feedback and discussion.

Appendix A

This appendix presents the proof of Theorem 1. I will follow the definition of NIP formulas given by Simon (2015) as follows (with notations changed to match preceding text)

> Let $\varphi(x;y)$ be a partitioned formula. We say that a set A of $|x|$-tuples is *shattered* by $\varphi(x;y)$ if we can find a family $(b_I : I \subseteq A)$ of $|y|$-tuples such that
>
> $$M \models \varphi(a;b_I) \iff a \in I, \quad \text{for all } a \in A$$

A formula $\varphi(x;y)$ is NIP if no infinite set of $|x|$-tuples is shattered by it.

Following notations from Soare (1987), let W_e to be the domain of the e-th partial recursive function and $Fin = \{e : W_e < \omega\}$.

Lemma Given e, define the following formula in the language of arithmetic

$$\begin{aligned}\theta_e(x, y) =& \exists l > x \; \exists \text{ enumeration } c_1, \ldots c_{2^l}, \text{ first } 2^l \text{ elements of } W_e \\ & \wedge \exists |\sigma| = l \text{ with } y = c_\sigma \wedge \sigma(x) = 1\end{aligned}$$

Then $e \in Fin$ iff θ_e is NIP.

Proof. ($\Rightarrow$) Suppose $e \in Fin$. The claim is: there is finite number N such that $|W_e| \leq 2^N$, and for all n, if a set A with cardinality n is shattered by θ_e, then $n \leq N$.

In particular, we show that the claim holds for N being the size of W_e. For the sake of contradiction, suppose there is A, with size n, shattered by θ_e, and $n > N$.

Let $A = \{a_1, \ldots, a_n\}$, $\{b_I : I \subset \{a_1, \ldots, a_n\}\}$, such that $\theta_e(a_i, b_I)$ iff $a_i \in I$.

Without loss of generality, let $a_n \geq n - 1$, and $I = \{a_n\}$. Then $a_n \in I$, and $\theta_e(a_n, b_I)$. This means that $\exists l > a_n \geq n - 1$ with the first 2^l many elements of W_e enumerated. Recall that the reductio hypothesis states $n > N$. This means that $|W_e| \geq 2^l > 2^{n-1} \geq 2^N$. This contradicts the original assumption that $|W_e| \leq 2^N$.

($\Leftarrow$) To show the contrapositive of this direction, suppose $e \notin Fin$, $|W_e| = \omega$. The claim is: θ_e is IP. Namely, $\forall N \; \exists n \geq N$, with some set A of cardinality n that is shattered by θ_e.

Take an arbitrary $n \geq N$. Let $A = \{0, \ldots, n - 1\}$. Let b_σ's be the first 2^n elements of W_e, as σ ranges over finite strings of length n. Since σ is a string, we say $a \in \sigma \Leftrightarrow \sigma(a) = 1$.

We need to show that $\theta_e(a, b_\sigma) \Leftrightarrow \sigma(a) = 1$.

The left to right direction is trivial, since it is part of $\theta_e(a, b_\sigma)$ to state that $\sigma(a) = 1$.

To show the right to left direction, note that since $|W_e| = \omega$, there definitely exists an initial segment of 2^n many elements of W_e, and $n > a$ for all $a \in A$. This satisfies the first conjunct. To satisfy the second conjunct of θ_e, recall that we defined our enumeration to be such that $|\sigma| = n$ with σ being identified with every number $\leq 2^n$. This means that an enumeration of $c_1 \ldots c_{2^n}$ includes all c_σ with $|\sigma| = n$. Define $b_\sigma = c_\sigma$, and we are guaranteed that b_σ is in the enumeration, and $|\sigma| = n$. Finally, the last conjunct of θ_e is satisfied by supposition.

□

Theorem. *The set* $\{\varphi(x,y) : \varphi(x,y)$ *is* $NIP\}$*, where* $\varphi(x,y)$ *is formulated in the language of arithmetic with addition and multiplication, is not decidable. In particular, this set computes* $\emptyset^{(2)}$*, the second Turing jump of the empty set.*

Proof. Suppose not, then for any formula $\varphi(x,y)$, we can decide if it's NIP. This means that, for any e, we can decide if $\theta_e(x,y)$ as defined in the lemma above is NIP. By lemma, $\theta_e(x,y)$ is NIP just in case $e \in Fin$. If we could decide the former, we would be able to decide the set Fin. But by Soare (1987, p.66, Theorem 3.2), Fin is Σ_2-complete, and so computes $\emptyset^{(2)}$, the second Turing jump of the empty set, and hence is not computable. □

Bibliography

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook Singapore.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Harman, G. and Kulkarni, S. (2012). *Reliable reasoning: Induction and statistical learning theory*. MIT Press.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Jain, S., Osherson, D. N., Royer, J., and Sharma, A. (1999). *Systems that learn: an introduction to learning theory*. MIT press.

Kelly, K. T. (1996). *The logic of reliable inquiry*. Oxford University Press.

Laskowski, M. C. (1992). Vapnik-Chervonenkis classes of definable sets. *Journal of the London Mathematical Society*, 45(2):377–384.

Lin, Z. and Bai, Z. (2010). *Probability inequalities*. Science Press Beijing, Beijing; Springer, Heidelberg.

Miller, C. (2005). Tameness in Expansions of the Real Field. In *Logic Colloquium '01*, volume 20 of *Lecture Notes in Logic*, pages 281–316. Associaton for Symbolic Logic, Urbana, IL.

Norton, J. D. (2014). A material dissolution of the problem of induction. *Synthese*, 191(4):671–690.

Osherson, D. N., Stob, M., and Weinstein, S. (1986). *Systems that learn: An introduction to learning theory for cognitive and computer scientists.* The MIT Press.

Pons, O. (2013). *Inequalities in analysis and probability.* World Scientific.

Shao, X., Cherkassky, V., and Li, W. (2000). Measuring the VC-dimension using optimized experimental design. *Neural computation*, 12(8):1969–1986.

Simon, P. (2015). *A Guide to NIP Theories.* Lecture Notes in Logic. Cambridge University Press, Cambridge.

Soare, R. I. (1987). *Recursively Enumerable Sets and Degrees.* Perspectives in Mathematical Logic. Springer, Berlin.

van den Dries, L. (1998). *Tame Topology and O-Minimal Structures*, volume 248 of *London Mathematical Society Lecture Note Series.* Cambridge University Press, Cambridge.

Vapnik, V., Levin, E., and Le Cun, Y. (1994). Measuring the VC-dimension of a learning machine. *Neural computation*, 6(5):851–876.

Wilkie, A. J. (1996). Model completeness results for expansions of the ordered field of real numbers by restricted pfaffian functions and the exponential function. *Journal of the American Mathematical Society*, 9(4):1051–1094.