Distributionally-Hard Languages

Lance Fortnow* A. Pavan† Alan L. Selman‡

April 23, 1999

Abstract

Define a set L to be *distributionally-hard to recognize* if for every polynomial-time computable distribution μ with infinite support, L is not recognizable in polynomial time on the μ-average. Cai and Selman [CS99] defined a modification of Levin's notion of average polynomial time and proved that every P-bi-immune language is distributionally-hard. Pavan and Selman [PS99] proved that there exist distributionally-hard sets that are not P-bi-immune if and only P contains P-printable-immune sets. We extend this characterizion to include assertions about several traditional questions about immunity, about finding witnesses for NP-machines, and about existence of one-way functions. Similarly, we address the question of whether NP contains sets that are distributionally hard. Several of our results are implications for which we cannot prove whether or not their converse holds. In nearly all such cases we provide oracles relative to which the converse fails. We use the techniques of Kolmogorov complexity to describe our oracles and to simplify the technical arguments.

1 Introduction

Levin [Lev86] was the first to advocate the general study of average-case complexity and he provided the central notions for its study. More recently, Cai and Selman [CS99] observed that Levin's definition of Average-P has limitations when applied to distributional problems with unreasonable distributions and when applied to exponential time-bounds. For example, for every language L in the complexity class E, (L,μ) is in Average-P, where μ is the flat distribution defined by $\mu'(x) = 4^{-|x|}$. However, E contains sets that are almost-everywhere complex. That is, E contains sets that require more than polynomial time to

*Department of Computer Science, University of Chicago, 1100 East 58th Street, Chicago, IL 60637. Email: fortnow@cs.uchicago.edu. Funded in part by NSF grant CCR97-32922

†Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14260. Email: aduri@cse.buffalo.edu

‡Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14260. Email: selman@cse.buffalo.edui. The author performed part of this research while visiting the Department of Computer Science, University of Chicago

recognize for all but finitely many inputs. It is unnatural to consider such a set to have average polynomial-time complexity for any distribution with infinite support, and certainly not for any flat distribution. Cai and Selman modified Levin's definition to remove these limitations. In particular, letting AVP denote the class of all distributional problems that are polynomial on the μ-average according to the definition of Cai and Selman, if a language L is almost-everywhere complex, then (L,μ) does not belong to AVP for any distribution μ. (We will provide all formal definitions in the next section.) Nevertheless, the difference between AVP and Average-P is modest: Define a distribution to be *reasonable* if there exists a constant $s > 0$ such that $\mu(\{x \mid |x| \geq n\}) = \Omega(\frac{1}{n^s})$. The reason of course is that distributions that decrease too quickly give too much weight to small instances, and for this reason are unreasonable. All distributions of natural problems in the literature, including the distributions of all known DistNP-complete problems, are reasonable. The class AVP $\subseteq$ Average-P, and Cai and Selman showed that, for all reasonable distributions μ, $(L,\mu) \in$ AVP if and only if $(L,\mu) \in$ Average-P.

It is well-known that a set L is almost-everywhere complex if and only if it is P-bi-immune [BS85]. Thus, if L is P-bi-immune, then there is no distribution μ with infinite support such that (L,μ) belongs to AVP. We say that such languages are *distributionally-hard*.

In this paper we raise and address questions about distributionally-hard sets. For example, we extensively characterize the question of whether there exist distributionally-hard sets that are not P-bi-immune. Pavan and Selman [PS99] proved that such sets exist if and only if P contains sets that are P-printable-bi-immune. We provide several other equivalent characterizations. These include assertions about the inability to compute witnesses for sets in NP and an assertion about the existence of one-way functions that are hard to invert almost everywhere.

It is easy to see that every distributionally-hard set is P-printable-bi-immune. We will show that if there exist (1) P-printable-bi-immune sets that are not distributionally-hard, then (2) there are distributionally-hard sets that are not P-bi-immune. The question of whether there exist such sets may not be forthcoming. To wit, we obtain oracles relative to which the polynomial hierarchy is infinite and sets satisfying (2) exist, and we obtain oracles relative to which the polynomial hierarchy is infinite and such sets do not exist. We obtain oracles relative to which sets satisfying (1) hold, and relative to which (1) does not hold. We obtain an oracle relative to which there exist sets that satisfy (2), but there do not exist sets that satisfy (1).

Now consider the question of whether NP contains distributionally-hard sets, for, if so, then DistNP (the class of all distributional problems (L,μ), where $L \in$ NP, and μ is polynomial-time computable) and AVP are distinct. Suppose that NP contains a P-bi-immune set L. Then, L is a distributionally-hard set in NP. We will show that $L \cap 0^* \in$ NP is P-immune, from which it follows that P contains P-printable-immune sets, and therefore that there exist sets that are distributionally-hard but not P-bi-immune. Indeed, if NP contains a P-immune tally set, then NP contains a P-immune set, which implies that P contains P-printable-immune sets. (These are straightforward observations.)

Relative to a random oracle, Lutz's hypothesis that the p-measure of NP is not zero is true [KM94]. Thus, relative to a random oracle, NP contains P-bi-immune, distributionally-hard sets [LM94]. We will see from our results that relative to an oracle constructed by Impagliazzo and Tardos [IT91], P contains P-printable-immune sets but NP does not contain P-immune tally sets. Hemaspaandra and Jha [HJ95] constructed an oracle relative to which NP contains P-immune sets but no P-immune tally sets. Here, we will construct an oracle relative to which P contains a P-printable-immune set but NP does not contain any P-immune set. Thus, relative to this oracle, there exist distributionally-hard sets that are not P-bi-immune, nevertheless, NP does not contain distributionally-hard sets. Relative to these oracles, the converses of the implications set forward in the previous paragraph do not hold.

We will use Kolmogorov complexity to describe the oracles that we construct. Because of this, both the ideas and the technical details of our proofs will be much easier than they would be otherwise.

2 Preliminaries

We assume that all languages are subsets of $\Sigma^* = \{0,1\}^*$ and we assume that Σ^* is ordered by standard lexicographic ordering. We use standard notation for complexity theoretic notions. In particular, $\mathrm{E} = \bigcup_{c>0} \mathrm{DTIME}(2^{cn})$ and $\mathrm{NE} = \bigcup_{c>0} \mathrm{NTIME}(2^{cn})$. We will use complexity class names to abbreviate descriptions of Turing machines. For example, an E-*machine* is a deterministic 2^{cn}, $c \geq 0$, time-bounded Turing machine.

A *transducer* is a Turing machine with a read-only input tape, a write-only output tape, and accepting states in the usual manner. A transducer computes a value y on an input string x if there is an accepting computation on x for which y is the final contents of the output tape. In general, such transducers compute partial, multivalued functions. PF is the set all partial functions that are computed by deterministic polynomial time-bounded transducers. A function $f \in \mathrm{PF}$ is *honest* if there is a polynomial q such that for every y in range(f), there exists x in dom(f) such that $f(x) = y$ and $|y| \leq q(|x|)$. A *one-way* function is an honest partial function in PF that cannot be inverted in polynomial time. Such functions exist if and only if $\mathrm{P} \neq \mathrm{NP}$ [Sel92]. Define an honest partial function f in PF to be *almost-always one-way* if no polynomial-time Turing machine inverts f correctly on more than a finite subset of range(f).

A *distribution function* $\mu : \{0,1\}^* \to [0,1]$ is a nondecreasing function from strings to the closed interval $[0,1]$ that converges to one. The corresponding *density function* μ' is defined by $\mu'(0) = \mu(0)$ and $\mu'(x) = \mu(x) - \mu(x-1)$. Clearly, $\mu(x) = \sum_{y \leq x} \mu'(y)$. For any subset of strings S, we will denote by $\mu(S) = \sum_{x \in S} \mu'(x)$, the probability of the event S. Define $u_n = \mu(\{x \mid |x| = n\})$. For each n, let $\mu'_n(x)$ be the conditional probability of x in $\{x \mid |x| = n\}$. That is, $\mu'_n(x) = \mu'(x)/u_n$, if $u_n > 0$, and $\mu'_n(x) = 0$ for $x \in \{x \mid |x| = n\}$, if $u_n = 0$.

A function μ from Σ^* to $[0,1]$ is *computable in polynomial time* [Ko83] if there is a polynomial time-bounded transducer M such that for every string x and every positive integer

n, $|\mu(x) - M(x, 1^n)| < \frac{1}{2^n}$. Consistent with Levin's hypothesis that natural distributions are computable in polynomial time, we restrict our attention *entirely* to such distributions. If μ is computable in polynomial time, then the density function μ' is computable in polynomial time. (The converse is false unless $P = NP$ [Gur91].) All distributions are to have infinite support—we explicitly exclude from consideration distributions μ for which $\mu'(x) = 0$ for all but a finite number of strings x. Consideration of such distributions would allow every problem to be an essentially finite problem.

Levin [Lev86] defines a function f from Σ^* to nonnegative reals to be *polynomial on* μ-*average* if there is an integer $k > 0$ such that

$$\sum_{|x| \geq 1} \mu'(x) \frac{(f(x))^{1/k}}{|x|} < \infty. \tag{1}$$

Average-P is the class of distributional problems (L, μ), where L is a language and μ is a polynomial-time computable distribution, such that L can be decided by some Turing machine M whose running time T_M is polynomial on μ-average.

For any time-constructible function T that is monotonically increasing, and hence invertible, Cai and Selman [CS99] define T on the μ-average as follows[1]: Let μ be a distribution on Σ^*, and let $W_n = \mu(\{x : |x| \geq n\})$. A function f is T on the μ-average if for all $n \geq 1$,

$$\sum_{|x| \geq n} \mu'(x) \cdot \frac{T^{-1}(f(x))}{|x|} \leq W_n. \tag{2}$$

Then, $\mathrm{AVTIME(T(n))}$ denotes the class of distributional problems (L, μ), where L is a language and μ is a polynomial-time computable distribution, such that L can be decided by some Turing machine M whose running time T_M is T on the μ-average.

Define $\mathrm{AVP} = \bigcup_{k \geq 1} \mathrm{AVTIME(n^k)}$. Clearly, $\mathrm{AVP} \subseteq$ Average-P.

A distribution μ is *reasonable* if there exists $s > 0$ such that $W_n = \Omega\left(\frac{1}{n^s}\right)$. We will require the following results of Cai and Selman [CS99] and Gurevich [Gur91].

Proposition 1 1. *If μ is a reasonable distribution, then (L, μ) belongs to* Average-P *(Levin's definition) if and only if (L, μ) belongs to* AVP *(Cai and Selman's definition).*

2. *If μ satisfies the stronger condition that there exists $s > 0$ such that $u_n = \Omega\left(\frac{1}{n^s}\right)$, then all of the following are equivalent:*

   (i) (L, μ) *belongs to* Average-P*;*

   (ii) (L, μ) *belongs to* AVP*;*

[1] Cai and Selman restricted their attention to functions that belong to Hardy's [Har24] class of logarithmico-exponential functions. We do not need to concern ourselves with this for the purpose of this paper.

(iii) *There is a Turing machine M that accepts L and an integer $k > 0$ such that for all $n \geq 1$,*

$$\sum_{|x|=n} \mu'(x) \frac{(T_M(x))^{1/k}}{|x|} \leq u_n. \tag{3}$$

Given any reducibility $\leq_r$, a distributional problem (L, μ) is $\leq_r$-complete for DistNP if $(L, \mu) \in$ DistNP (i.e., $L \in$ NP and μ is computable in polynomial time) and every distributional problem that belongs to DistNP is $\leq_r$ reducible to (L, μ).

Here we have given only the definitions and properties that we need for this paper; we refer the reader to the recent expositions by Impagliazzo [Imp95] and Wang [Wan97] for deeper understanding of average-case complexity.

2.1 Immunity

A language L is *immune* to a complexity class $\mathcal{C}$, or $\mathcal{C}$*-immune*, if L is infinite and no infinite subset of L belongs to $\mathcal{C}$. A language L is *bi-immune* to a complexity class $\mathcal{C}$, or $\mathcal{C}$*-bi-immune*, if L is infinite, no infinite subset of L belongs to $\mathcal{C}$, and no infinite subset of $\overline{L}$ belongs to $\mathcal{C}$. A language is DTIME($T(n)$)*-complex* if L does not belong to DTIME($T(n)$) almost everywhere; that is, every Turing machine M that accepts L runs in time greater than $T(|x|)$, for all but finitely many words x. Balcázar and Schöning [BS85] proved that for every time-constructible function T, L is DTIME($T(n)$)-complex if and only if L is bi-immune to DTIME($T(n)$).

Mayordomo [May94] proved that the p-measure of the class of DTIME(2^{n})-bi-immune sets is 1, and therefore, if the p-measure of NP is not 0, then NP contains a DTIME(2^{n})-bi-immune set. Cai and Selman [CS99] proved, for all P-bi-immune sets L and for all polynomial-time computable distributions μ, that $(L, \mu) \notin$ AVP. Thus, if NP does not have p-measure 0, then there is a language L such that for every polynomial-time computable distribution μ, the distributional problem (L, μ) belongs to DistNP but does not belong to AVP. (Independently, Schuler and Yamakami [SY95] obtained a similar result.)

We define a language L to be *distributionally-hard to recognize* if for all polynomial-time computable distributions μ, $(L, \mu) \notin$ AVP. As we noted, every P-bi-immune language is distributionally-hard to recognize.

Recall that set L is P-*printable* if there exists $k \geq 1$ such that all the elements of L up to size n can be printed by a deterministic Turing machine in time $n^k + k$ [HY84, HIS85]. Every P-printable set is sparse and belongs to P. A set A is P-*printable-immune* if A is infinite and no infinite subset of A is P-printable.

2.2 Kolmogorov Complexity

We will need several definitions and results from Kolmogorov complexity. See the book by Li and Vitányi [LV97] for an in-depth look at Kolmogorov complexity including proofs of the propositions mentioned here.

Definition 1 *Fix a universal Turing Machine Φ. For any string x in $\{0,1\}^*$, the Kolmogorov complexity of x is defined as*

$$K(x) = Min\{|p| : \Phi(p) = x\}.$$

We sometimes consider relative Kolmogorov complexity. $K(x|y)$ is the Kolmogorov complexity of x relative to a string y where we replace $\Phi(p)$ with $\Phi(p,y)$ in Definition 1. $K(x|B)$ is the Kolmogorov complexity of x relative to a set B where we replace $\Phi(p)$ with $\Phi^B(p)$ in Definition 1.

We also consider time-bounded Kolmogorov complexity, $K^t(x)$. For this definition we require that $\Phi(p)$ use at most $t(|x|)$ steps.

Proposition 2 *For every n, there is a string x of length n such that $K(x) \geq n$.*

Proposition 3 *For every $A \subseteq \{0,1\}^*$ with $|A| = m$, there is a string $x \in A$ such that $K(x) \geq \log m$.*

In fact, the following more general statement also holds.

Proposition 4 *For any positive integer c and any $A \subseteq \{0,1\}^*$ with $|A| = m$, the number of strings $x \in A$ with $K(x) \geq \log m - c$ is at least $m(1-2^{-c})$.*

Propositions 2, 3 and 4 hold even for time-bounded and/or relative to a fixed string y or set B.

We also have the following result on upper bounds of Kolmogorov complexity.

Proposition 5 *Let $A \subseteq \{0,1\}^* \times \{0,1\}^*$ be a recursively enumerable set, and define $X_y = \{x \in \{0,1\}^* | (x,y) \in A\}$ for some $y \in \{0,1\}^*$. If X_y is finite, then for every $x \in X_y$, $K(x|y) \leq \log|X_y| + c_A$ for a constant c_A depending only on A.*

3 Distributional Hardness

The following theorem is central to this paper. It completely characterizes the question of whether there exist languages that are distributionally-hard to recognize other than the P-bi-immune sets.

Theorem 1 *The following assertions are equivalent.*

1. *There exists a distributionally-hard set that is not* P*-bi-immune.*
2. *There exists a* P*-printable-bi-immune set that is not* P*-bi-immune.*
3. *There exists a set that is* P*-printable-bi-immune, not* P*-bi-immune, but whose complement is* P*-immune.*

4. P *contains a* P*-printable-immune set.*

5. NP *contains a* P*-printable-immune set.*

6. *There is an infinite set* S *in* NE *and an* NE*-machine* M *that accepts* S *such that no* E*-machine correctly computes infinitely many accepting computations of* M*.*

7. *There is an infinite tally language* L *in* NP *and an* NP*-machine* M *that accepts* L *such that no* P*-machine correctly computes infinitely many accepting computations of* M*.*

8. *There is an infinite set* S *in* NP *and an* NP*-machine* M *that accepts* S *such that no* P*-machine correctly computes infinitely many accepting computations of* M*.*

9. *Almost-always one-way functions exist.*

Theorem 1 shows that the question of whether there exist languages that are distributionally hard but not P-bi-immune is equivalent to several previously studied conjectures about immunity, about computing witnesses of nondeterministic machines, and about existence of strong one-way functions. Pavan and Selman [PS99] obtained the equivalence of items 1 and 4. The equivalence of items 4 and 5 is due to Russo [AR88]. Also, Theorem 1 strengthens a result of Hemaspaandra, Rothe, and Wechsung [HRW97], which is that item 8 implies item 4.

Recall that every language in the class E (NE) is identifiable with a tally language in P (NP), respectively [Boo74]. This observation yields the fact that items 6 and 7 are equivalent. Similarly, NP contains a P-immune tally language if and only if NE contains an E-immune set. One does not expect existence of a P-immune set in NP to imply existence of a P-immune tally language in NP, because in general such downward separations do not hold. Hemaspaandra and Jha [HJ95] support this contention with an oracle relative to which NP contains a P-immune set but no P-immune tally language. Thus, the equivalence of items 7 and 8 is surprising and demonstrates that search problems and decision problems have different properties. The following completes the proof.

Proof. As we just noted, items 1, 4, and 5 are equivalent, and items 6 and 7 are equivalent. Items 8 and 9 are equivalent by standard techniques [Sel92].

We will prove the following cycles:

(i) $(4) \Rightarrow (7) \Rightarrow (8) \Rightarrow (4)$.

(ii) $(4) \Rightarrow (3) \Rightarrow (2) \Rightarrow (4)$.

Since item 4 appears in each cycle, this will complete our proof. The following lemmas accomplish our task.

Lemma 1 $(4) \Rightarrow (7) \Rightarrow (8) \Rightarrow (4)$.

Proof. Let $A \in \mathrm{P}$ be P-printable-immune. Define $L = \{1^n \mid \exists x[|x| = n \text{ and } x \in A\}$. Then, L is an infinite tally language in NP. Consider the NP-machine M that on input 1^n, guesses a string x of length n and accepts if $x \in A$. If some P-machine computes infinitely many accepting computations of M, then A is not P-printable immune. Thus, item 4 implies item 7. Item 7 implies item 8 trivially.

Now, for sake of completeness, we show that item 8 implies item 4. Let S be an infinite set in NP and let M be an NP-machine that accepts S such that no P-machine correctly computes infinitely many accepting computations of M, Define

$$C = \{\langle x, y\rangle \mid \ y \text{ is an accepting computation of } M \text{ on } x\ \}.$$

Then, $C \in \mathrm{P}$. Let $p(n)$ be a polynomial such that for every $\langle x, y\rangle$ in C, $|\langle x, y\rangle| \leq p(|x|)$. If C has an infinite P-printable subset A, then the following procedure computes accepting computations for infinitely many inputs of M.

input x;
print $A^{\leq p(|x|)}$;
if for some y, $\langle x, y\rangle$ is printed **then** output the first such y **else** reject.

Thus, item 8 implies item 4 □

Lemma 2 (4) $\Rightarrow$ (3) $\Rightarrow$ (2) $\Rightarrow$ (4) .

Proof. First we show that item 4 implies item 3. Let $L \in \mathrm{P}$ be P-printable immune. Let X be any set that is P-bi-immune. Define $A = X \cup L$. Since L is an infinite subset of A that belongs to P, A is not P-bi-immune. Let B be an infinite set in P. Noting that $\overline{A} \subseteq \overline{X}$, B is not a subset of $\overline{A}$, because X is P-bi-immune. Thus, $\overline{A}$ is P-immune.

Now we show that A is P-printable–bi-immune. Note that we need only to prove that $\overline{A}$ includes no infinite P-printable set. Let B denote an infinite P-printable set and suppose that B is a subset of A. B does not have a finite intersection with L because if so, then $B \cap \overline{L}$ is an infinite subset of X that belongs to P. Thus, $B \cap L$ is an infinite set. However, $B \cap L$ is P-printable, which is a contradiction because L is P-printable-immune.

Item 3 implies item 2 trivially. We need only to show that (2) implies (4). Let A be P-printable-bi-immune but not P-bi-immune. Let L be an infinite set in P so that either $L \subseteq A$ or $L \subseteq \overline{A}$. If L has an infinite P-printable subset, then so does either A or $\overline{A}$. Thus, L is P-printable-immune. □

Theorem 1 follows immediately from Lemmas 1 and 2. □

Theorem 2 *There exist oracles A and B relative to which the polynomial hierarchy is infinite; relative to A the properties listed in Theorem 1 are true, and relative to B these properties are false.*

For the proof of Theorem 2, to construct A, begin with an oracle L relative to which the polynomial hierarchy is infinite [Yao85, Hås89], and then apply techniques of Fortnow [For99], noting that the assertions of Theorem 1 are true relative to L and a random oracle. To construct B, in a similar manner, apply results of Long and Selman [LS86] to L and a sparse oracle relative to which $\mathrm{E}^{\mathrm{NP}} = \mathrm{E}$.

The equivalence of items 1 and 2 of Theorem 1 lead us to ask whether distributionally-hard and P-printable-bi-immune are equivalent. First, we note the following proposition.

Proposition 6 *Every distributionally-hard to recognize set is* P*-printable-bi-immune.*

The proof is clear: If either L or $\overline{L}$ has an infinite P-printable subset S, then define a polynomial-time computable distribution μ such that $\mu(S) = 1$. Then, it is easy to see that $(L,\mu) \in \mathrm{AVP}$.

Now, for any set L, we have the following implications:

$$\begin{aligned} & L \text{ is P-bi-immune} \\ & \Rightarrow \quad (4) \\ & L \text{ is distributionally-hard} \\ & \Rightarrow \quad (5) \\ & L \text{ is P-printable-bi-immune.} \end{aligned}$$

To study our question further, consider the following hypothesis:

Hypothesis 1 *There exists a* P*-printable-bi-immune-set that is not distributionally hard.*

Item 1 of Theorem 1 asserts that the implication 4 does not collapse. Hypothesis 1 asserts that the implication 5 does not collapse. Hypothesis 1 implies item 2 of Theorem 1. Thus, by Theorem 1, if 4 collapses, then 5 collapses. Next we show that there is an oracle relative to which Hypothesis 1 is true, and there is an oracle relative to which the assertions in Theorem 1 hold but Hypothesis 1 is false.

Theorem 3 *There exists an oracle relative to which Hypothesis 1 is true.*

Proof. We let μ be the standard uniform distribution, i.e., $\mu'(x) = \frac{1}{|x|^2 2^{|x|}}$. Fix a set C in DTIME($n^{\log n}$) that is P-printable-bi-immune by the usual diagonalization argument. We will create A such that C is still P^A-printable-bi-immune yet (C,μ) is in AVP^A.

We define R^n inductively by $R^0 = \emptyset$ and, for $n \geq 1$, R^n is the set of strings x in Σ^n such that

$$K^{2^n}(x \mid R^0 \cup R^1 \cup \cdots \cup R^{n-1}) \geq n/2.$$

Let R be the union of the R_n. We define A as follows:

$$A = \{\langle x, 0\rangle \mid x \in R - C\} \cup \{\langle x, 1\rangle \mid x \in R \cap C\}$$

Consider the following algorithm P^A for C on input x:

1. If $\langle x, 0\rangle$ is in A then reject.

2. If $\langle x, 1\rangle$ is in A then accept.

3. Otherwise simulate the unrelativized DTIME($n^{\log n}$) algorithm for C.

Clearly from the construction the algorithm P^A accepts the language C. Since μ is a reasonable distribution, by Proposition 1 we need only show that (C, μ) is in Average-P^A.

By Proposition 4 there are at most $2^{n/2}$ strings with $K(x) \leq n/2$ so $|\Sigma^n - R^n| \leq 2^{n/2}$. Consider

$$\begin{aligned}
\sum_{|x|\geq 1} \mu'(x) \frac{(T_P(x))^{1/2}}{|x|} &= \sum_{n\geq 1} \sum_{x\in\Sigma^n} \frac{1}{n^2 2^n} \frac{T_P(x)^{1/2}}{n} \\
&= \sum_{n\geq 1} \frac{1}{n^3 2^n} \sum_{x\in\Sigma^n} T_P(x)^{1/2} \\
&= \sum_{n\geq 1} \frac{1}{n^3 2^n} \Big(\sum_{x\in R^n} T_P(x)^{1/2} + \sum_{x\in\Sigma^n - R^n} T_P(x)^{1/2} \Big) \\
&\leq \sum_{n\geq 1} \frac{1}{n^3 2^n} \Big(\sum_{x\in R^n} n^{1/2} + \sum_{x\in\Sigma^n - R^n} (n^{\log n})^{1/2} \Big) \\
&\leq \sum_{n\geq 1} \frac{1}{n^3 2^n} (2^n n^{1/2} + 2^{n/2} (n^{\log n})^{1/2}) \\
&\leq \sum_{n\geq 1} \frac{1}{n^2} \\
&= O(1)
\end{aligned}$$

Suppose D is a P^A-printable set. We will show that D is also P-printable. Thus since C was P-printable immune it will also be P^A-printable immune.

Let M^A be a machine that on input 1^n outputs the strings of length n of D in time n^k.

Inductively compute R^0, R^1, …, $R^{10k\log n}$ by simulating all of the small programs. This takes time polynomial in n. We can also compute C on all strings of length up to $10k \log n$ in time polynomial in n.

Simulate $M^A(1^n)$ and whenever it asks a query $\langle y, 0\rangle$ we answer yes if $|y| \leq 10k \log n$ and $y \in R - C$, and if M asks a query $\langle y, 1\rangle$ we answer yes if $|y| \leq 10k \log n$ and $y \in R \cap C$. If $|y| > 10k \log n$, then answer no.

This simulation of M^A will produce the correct answer unless $M^A(1^n)$ queries some $\langle y, i\rangle$ for $|y| > 10k \log n$ and y in R. We will show this cannot happen for large n. Suppose it does and consider the first such y queried. Let $m = |y| > 10k \log n$.

Note that

$$K^{2^m}(y \mid R^0 \cup R^1 \cup \cdots \cup R^{m-1}) \leq (k + O(1)) \log n$$

since $2^m > n^k$ and we can give a description of y by the index of the query made by M^A since the answers to all previous queries can be found in $R^0 \cup R^1 \cup \cdots \cup R^{m-1}$. Since $(k + O(1)) \log n < m/2$ (for large n) we have a contradiction. □

Theorem 4 *There exists an oracle relative to which the assertions in Theorem 1 hold but Hypothesis 1 is false.*

Proof. Assume P = PSPACE. We can remove this assumption by first relativizing to TQBF. Let t_i be the double towers defined as follows: $t_0 = 1$ and $t_{i+1} = 2^{2^{t_i}}$ for every $i \geq 0$. For each i pick a Kolmogorov random string z_i of length t_i. Our oracle B will consist of the $z_i's$ and the encoding of the $z_i's$ at length $t_i^{\log t_i}$, i.e., for each i and j we put $\langle 1^{t_i^{\log t_i}}, j \rangle$ in the oracle if the jth bit of z_i is one.

Consider the set C of the z_i's. Clearly C is in P^B. Suppose there is a P^B-printable infinite subset of B. There is some n^k-time machine M^B that for arbitrarily large i on input 1^{t_i} will print z_i. Without loss of generality, assume that M^B queries z_i before it prints it.

Consider such i such that $\log t_i >> k$. For these i's, M^B will be unable to read the encoding of z_i at length $t_i^{\log t_i}$. We can describe z_i by i, the index of the first query M^B makes to z_i and the z_j's for $j < i$. This whole description is $O(\log t_i)$ bits which contradicts the randomness of z_i for large i. Thus, C is P^B-printable-immune.

Next, let E be a set that is not distributionally hard relative to B. We will show that E is not P-printable-bi-immune.

By definition of distributionally hard, there must be a Turing machine M^B that accepts E, a polynomial-time computable (relative to B) distribution μ of infinite support, and a constant k such that for all n,

$$\sum_{|x| \geq n} \mu'(x) \frac{(T_M(x))^{1/k}}{|x|} \leq W_n. \tag{6}$$

Observe that if there is an infinite P^B-printable set D such that M^B runs in polynomial time on the elements of D, then either E or $\overline{E}$ contains an infinite P^B-printable subset and thus E is not P^B-printable bi-immune.

We call a length n an *easy* length if for some i, $t_i^{\log t_i} \leq n^{5k} < t_{i+1}$. Suppose that M^B runs in time n^{5k} for infinitely many strings of easy lengths. For these easy lengths we can in polynomial-time relative to B construct all the strings of B that M^B can see by querying the encodings of the z_j for $j \leq i$. We can then use the P = PSPACE assumption to print an infinite set D of the lexicographically first strings of easy lengths on which M^B runs in time n^{5k}. Thus, as we just observed, in this case E is not P^B-printable bi-immune.

For the rest of the proof we assume M^B runs in time greater than n^{5k} for all but finitely strings of easy length n.

Fix a noneasy length n, that is, for some i, $t_i \le n^{5k} < t_i^{\log t_i}$. Let S_n be the set of strings x of length n such that $M^B(x)$ queries z_i within n^{5k} steps.

Lemma 3 *The weight of S_n is small; i.e., $\mu(S_n) \le u_n/2$.*

Proof. First some notation. Consider N^B, the polynomial-time machine that computes μ. Let $B^{<t_i}$ represent the strings of B of length less than t_i. Note that these strings can be queried in time polynomial in n. Let μ^* represent the function computed by $N^{B^{<t_i}}$ and μ^w be the function computed by $\mu^{B^{<t_i}\cup\{w\}}$. Note that for strings x of length n we have $\mu(x) = \mu^{z_i}(x)$.

Also note that for all x of length n, $\mu^*(x) \le \mu^*(x+1)$. Otherwise, pick the smallest such counterexample and we have a short description of z_i since $\mu^{z_i}(x) \le \mu^{z_i}(x+1)$.

Define T_w as the set of strings x of length n such that either

1. $M(x)$ queries w within n^{5k} steps, or
2. $\mu^*(x)$ queries w or
3. $\mu^*(x+1)$ queries w.

Note that $S_n \subseteq T_{z_i}$.

Consider the sum of all of the $\mu^*(T_w)$.

$$\sum_{w\in\Sigma^n} \mu^*(T_w) = \sum_{w\in\Sigma^n}\sum_{x\in T_w} (\mu^*(x+1)-\mu^*(x))$$

$$= \sum_{x\in\Sigma^n}\sum_{w:x\in T_w} (\mu^*(x+1)-\mu^*(x)) \le n^{O(1)} \sum_{x\in\Sigma^n} (\mu^*(x+1)-\mu^*(x)) = n^{O(1)}u_n.$$

The last equality holds since $\mu^*(0^n)$ and $\mu^*(0^{n+1})$ cannot query z_i or we would have a short description of z_i, and

$$u_n = \mu(0^{n+1}) - \mu(0^n) = \mu^*(0^{n+1}) - \mu^*(0^n) = \sum_{x\in\Sigma^n} (\mu^*(x+1)-\mu^*(x)).$$

Now suppose $\mu^{z_i}(S_n) \ge u_n/2$. We then have $\mu^{z_i}(T_{z_i}) \ge u_n/2$. We claim $\mu^{z_i}(T_{z_i}) = \mu^*(T_{z_i})$: Consider a maximal interval $I = \{x, x+1, \ldots, x+q\}$ in T_{z_i}. Note $\mu^*(x)$ does not query z_i or $x-1$ is in T_{z_i}. Also $\mu^*(x+q+1)$ does not query z_i or $x+q+1$ is in T_{z_i}. So $\mu^{z_i}(I) = \mu^{z_i}(x+q+1) - \mu^{z_i}(x) = \mu^*(x+q+1) - \mu^*(x) = \mu^*(I)$. Since T_{z_i} can be partitioned into a set of maximal intervals, we have that $\mu^*(T_{z_i}) = \mu^{z_i}(T_{z_i})$. (Note that 0^n and 1^n are not in T_{z_i} or we would have a short description for z_i.)

There are only a polynomial number of w such that $\mu^*(T_w) \ge u_n/2$ so if z_i was one of these we could give a short description of z_i. □

Now we will show that if E is P^B-printable-bi-immune, then for all but finitely many n, either $u_n = 0$ or

$$\sum_{|x|=n} \mu'(x)\frac{(T_M(x))^{1/k}}{|x|} \ge u_n. \tag{7}$$

This immediately contradicts Equation (6) for all but finitely many n.

For sufficiently large easy n, we have already shown that $T_M(x) > n^{5k}$ for all x of length n. Equation (7) follows for these lengths.

Let D' be the set of x of a noneasy length n such that $T_{M^B}(x)$ halts in less than n^{5k} steps without querying z_i for the appropriate i. Let D be the lexicographically first string x of noneasy length such that $T_{M^{B-\{z_i\}}}(x)$ halts in n^{5k} steps. Note that by the P=PSPACE assumption D is P^B-Printable. Also note that $D \subseteq D'$ (at least for large n) or we have a short description of z_i. Finally note that D' infinite implies D infinite.

If D' is infinite, then E is not P^B-printable-bi-immune. Suppose that D' is finite. Consider a noneasy n such that D' has no strings of length n. If x is an input such that $T_M(x)$ uses at most n^{5k} steps, then $T_M(x)$ must query z_i, so x is in S_n.

We then have

$$
\begin{aligned}
\sum_{|x|=n} \mu'(x)\frac{(T_M(x))^{1/k}}{|x|} &\geq \sum_{x\in\Sigma^n - S_n} \mu'(x)\frac{(T_M(x))^{1/k}}{|x|} \\
&\geq \sum_{x\in\Sigma^n - S_n} \mu'(x)\frac{(n^{5k})^{1/k}}{n} \\
&\geq \sum_{x\in\Sigma^n - S_n} 2\mu'(x) \\
&= 2\mu(\Sigma^n - S_n) \\
&= 2(u_n - \mu(S_n))
\end{aligned}
$$

Equation (7) now follows from Lemma 3. □

Now we turn to the question of whether NP contains distributionally-hard to recognize sets. The following for the most part are easy to prove and essentially known [PS99].

Theorem 5 *For any language L in* NP,

$$
\begin{aligned}
&\textit{L is P-bi-immune} \\
&\Rightarrow && (8) \\
&\textit{L is distributionally-hard} \\
&\Rightarrow && (9) \\
&L\cap 0^* \textit{ is P-immune} \\
&\Leftrightarrow && (10) \\
&L\cap 0^* \textit{ is P-printable-immune} \\
&\Rightarrow && (11) \\
&\textit{The assertions of Theorem 1 all hold.}
\end{aligned}
$$

The equivalence (10) holds because a tally language is P-immune if and only if it is P-printable immune. Then, it is immediate that item 4 of Theorem 1 holds, which proves that (11) holds.

It follows immediately from Theorem 5 that NP has distributionally-hard sets relative to a random oracle (because NP has P-bi-immune sets), and that, relative to an oracle for which the conditions of Theorem 1 fails, NP does not have distributionally-hard sets.

If NP contains P-bi-immune sets, then NP contains sets having various combinations of immunity properties. For example, we have the following result.

Theorem 6 *If* NP *contains a* P*-printable-immune set, then* NP *contains sets that are*

1. P*-printable-immune and not* P*-immune, and*
2. P*-printable-bi-immune and not* P*-bi-immune.*

If the p-measure of NP is not 0, then NP contains a P-bi-immune set. The obvious corollaries hold, of which, the following are the most interesting.

Corollary 1 **(i)** *If the p-measure of* NP *is not* 0*, then there is a language that is distributionally-hard to recognize but not* P*-bi-immune.*

(ii) *If the p_2-measure of* NP *is not* 0*, then* NP *contains a language that is distributionally-hard to recognize but not* P*-bi-immune.*

(iii) *If the p-measure of* NP *is not* 0*, then* NP *contains sets that are*

1. P*-printable-immune and not* P*-immune, and*
2. P*-printable-bi-immune and not* P*-bi-immune.*

Parts (i) and (ii) are due to Pavan and Selman [PS99]. (The proof of Corollary 1(ii) is not obvious; it follows from a close analysis the proof of Theorem 1.) In light of these results, it is interesting to ask the following question: If NP contains a P-printable-immune set, does NP contain a P-immune set? Let us consider this question in the context of the following obvious implications:

$$\text{NP has a P-immune tally set} \tag{12}$$
$$\Rightarrow$$
$$\text{NP has a P-immune set} \tag{13}$$
$$\Rightarrow$$
$$\text{NP has a P-printable-immune set.} \tag{14}$$

Item (14) is one of the equivalent assertions in Theorem 1. Item 6 of Theorem 1 was studied by Impagliazzo and Tardos [IT91], who obtained an oracle relative to which NE = E and this assertion holds. Thus, there is an oracle relative to which (14) holds and (12) does not hold.[2] Hemaspaandra and Jha [HJ95] constructed an oracle relative to which (13) holds and (12) fails. However, it remains an open question as to whether there is an oracle relative to which (13) holds and NE = E. The following result completes our study of these assertions.

Theorem 7 *There is an oracle relative to which* NP *has a* P*-printable-immune set but* NP *has no* P*-immune set.*

Proof. The proof builds on techniques of Hemaspaandra and Jha [HJ95]. Let t_i be the towers defined as such: $t_0 = 1$ and $t_{i+1} = 2^{t_i}$ for every $i \geq 0$. For each i, pick a Kolmogorovly random string y_i of length t_i. Let B be set consisting of all the pairs $\langle 0, y_i \rangle$. Let $M_1, \ldots$ be an enumeration of nondeterministic oracle Turing machines where M_i uses time at most n^i.

Initially set $A = B$. For each i do the following:

Unmark all of the j.
S: Pick a pair $\langle x, j \rangle$ that minimizes $|\langle 1, j, x, 1^{|x|^j} \rangle|$ such that $M_j^A(x)$ accepts with $t_i \leq |\langle 1, j, x, 1^{|x|^j} \rangle| < t_{i+1}$ and an unmarked $j \leq i$. If no such x exists then go to next i.
Mark j and put $\langle 1, j, x, 1^{|x|^j} \rangle$ in A.
Go to S.

Consider any M_j such that $L(M_j^A)$ is infinite. The set $\{x \mid \langle 1, j, x, 1^{|x|^j} \rangle\}$ is an infinite subset of $L(M_j^A)$ in P^A. The set B is in P^A. Suppose f is a polynomial-time computable function such that $f^A(1^{t_i}) = \langle 0, y_i \rangle$. This gives us a short description of y_i since A only has $O(i^2)$ strings of length at most polynomial in t_i. Thus f could only have this property for finitely many y_i and B is P^A-printable immune. □

In particular, neither the assertions listed in Theorem 1 (14) nor the existence of P-immune sets in NP (13) appear to be strong enough to ensure that NP contains distributionally-hard sets. We leave as open the questions of whether (12) implies that NP contains distributionally-hard sets, and whether existence of distributionally-hard sets in NP implies that NP contains P-bi-immune sets.

References

[AR88] E. Allender and R. Rubinstein. P-printable sets. *SIAM Journal on Computing*, 17(6):1193–1202, 1988.

[Boo74] R. Book. Tally languages and complexity classes. *Information and Control*, 26:186–193, 1974.

[2] Actually, the property studied by Impagliazzo and Tardos is somewhat weaker. Nevertheless, our claim is correct, as is our attribution.

[BS85] J. Balcázar and U. Schöning. Bi-immune sets for complexity classes. *Mathematical Systems Theory*, 18(1):1–10, June 1985.

[CS99] J-Y. Cai and A. Selman. Fine separation of average time complexity classes. *SIAM Journal on Computing*, 28(4):1310–1325, 1999.

[For99] L. Fortnow. Relativized worlds with an infinite hierarchy. *Information Processing Letters*, 69(6):309–313, 1999.

[Gur91] Y. Gurevich. Average case completeness. *Journal of Computer and System Sciences*, 42:346–398, 1991.

[Har24] G. Hardy. *Orders of Infinity, The 'infinitärcalcül' of Paul du Bois-Reymond*, volume 12 of *Cambridge Tracts in Mathematics and Mathematical Physics*. Cambridge University Press, London, 2nd edition, 1924.

[Hås89] J. Håstad. Almost optimal lower bounds for small depth circuits. In S. Micali, editor, *Randomness and Computation*, volume 5 of *Advances in Computing Research*, pages 143–170. JAI Press, Greenwich, 1989.

[HIS85] J. Hartmanis, N. Immerman, and V. Sewelson. Sparse sets in NP-P: EXPTIME versus NEXPTIME. *Information and Control*, 65:158–181, 1985.

[HJ95] L. Hemaspaandra and S. Jha. Defying upward and downward separation. *Information and Computation*, 121(1):1–13, 1995.

[HRW97] L. Hemaspaandra, J. Rothe, and G. Wechsung. Easy sets and hard certificate schemes. *Acta Informatica*, 34(11):859–879, 1997.

[HY84] J. Hartmanis and Y. Yesha. Computation times of NP sets of different densities. *Theoretical Computer Science*, 34:17–32, 1984.

[Imp95] R. Impagliazzo. A personal view of average-case complexity. In *Proceedings of the Tenth Annual IEEE Conference on Structure in Complexity Theory*, pages 134–147, 1995.

[IT91] R. Impagliazzo and G. Tardos. Search versus decision in super-polynomial time. In *Proc. IEEE Foundations of Computer Science*, pages 222–227, 1991.

[KM94] S. Kautz and P. Miltersen. Relative to a random oracle, NP is not small. In *Proceedings of the Ninth Annual IEEE Structure in Complexity Theory Conference*, pages 162–174, 1994.

[Ko83] K. Ko. On self-reducibility and weak P-selectivity. *Journal of Computer and System Sciences*, 26:209–211, 1983.

[Lev86] L. Levin. Average case complete problems. *SIAM Journal on Computing*, 15:285–286, 1986.

[LM94] J. Lutz and E. Mayordomo. Cook versus Karp-Levin: Separating completeness notions if NP is not small. In *Proceedings of the Eleventh Annual Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science*, volume 755, pages 415–426, Berlin, 1994. Springer-Verlag.

[LS86] T. Long and A. Selman. Relativizing complexity classes with sparse oracles. *Journal of the ACM*, 33(3):618–627, 1986.

[LV97] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Graduate Texts in Computer Science. Springer, New York, second edition, 1997.

[May94] E. Mayordomo. Almost every set in exponential time is P-bi-immune. *Theoretical Computer Science*, 136:487–506, 1994.

[PS99] A. Pavan and A. Selman. Complete distributional problems, hard languages, and resource-bounded measure. *Theoretical Computer Science*, 1999. To appear.

[Sel92] A. Selman. A survey of one-way functions in complexity theory. *Mathematical Systems Theory*, 25:203–221, 1992.

[SY95] R. Schuler and T. Yamakami. Sets computable in polynomial time on the average. In *Proceedings of the First Annual International Computing and Combinatorics Conference, Lecture Notes in Computer Science*, volume 959, pages 650–661. Springer-Verlag, Berlin, 1995.

[Wan97] J. Wang. Average-case computational complexity theory. In L. Hemaspaandra and A. Selman, editors, *Complexity Theory Retrospective II*, pages 295–328. Springer-Verlag, 1997.

[Yao85] A. Yao. Separating the polynomial-time hierarchy by oracles. In *Proceedings of the 26th IEEE Symposium on Foundations of Computer Science*, pages 1–10, 1985.