Complete Distributional Problems, Hard Languages, and Resource-Bounded Measure*

A. Pavan
Alan L. Selman
Department of Computer Science
University at Buffalo
Buffalo, NY 14260

February 6, 1997

Abstract

Cai and Selman [CS96] defined a modification and extension of Levin's notion of average polynomial time to arbitrary time-bounds and proved that if L is P-bi-immune, then for every polynomial-time computable distribution μ, the distributional problem (L, μ) is not polynomial on the μ-average. We prove the following consequences of the hypothesis that the p-measure of NP is not 0:

1. There is a language L that is not P-bi-immune and for every polynomial-time computable distribution μ, the distributional problem (L, μ) is not polynomial on the μ-average.
2. For every DistNP-complete distributional problem (L, μ), there exists a constant $s \geq 0$ such that $\mu(\{x \mid |x| \geq n\}) = \Omega(\frac{1}{n^s})$. That is, every DistNP-complete problem has a reasonable distribution.

1 Introduction

This paper concerns the average-time complexity of distributional problems. A distributional problem is a pair (L, μ), where L is a language over a finite alphabet Σ and μ is a distribution defined on Σ^*. Given a distributional problem, it is an important issue to either find an expected polynomial-time algorithm that solves the problem

*Research supported in part by NSF grant CCR-9400229

or to prove that such an algorithm does not exist. Levin [Lev86] provided two central notions for studying this issue. One is analogous to the class P, and provides an *easiness* notion; the other is analogous to the class of NP-complete sets, and provides a *hardness* notion. For the first, Levin defined a robust notion of what it means for an algorithm that accepts L to be polynomial on the μ-average. Using this notion, Average-P denotes the set of all distributional problems (L, μ) such that μ is computable in polynomial time and some algorithm for L is polynomial on the μ-average. Let DistNP denote the collection of all distributional problems (L, μ) such that μ is computable in polynomial time and L belongs to NP. For the second central notion, that of hardness, Levin defined reductions between distributional problems. Using these reducibilities, in the usual manner, we define a distributional problem (L, μ) to be complete for DistNP if (L, μ) belongs to DistNP and every distributional problem in DistNP is reducible to (L, μ). It is not known whether DistNP $\subseteq$ Average-P. If P $=$ NP, then DistNP $\subseteq$ Average-P, and if DistNP $\subseteq$ Average-P, then E $=$ NE [BDCGL92]. Levin showed that distributional tiling with a simple distribution is complete for DistNP, and since then, several additional DistNP-complete problems have been found [BG95, Gur91, VL88, VR92, WB95, Wan95]. However, we do not possess a catalog of DistNP-complete problems that is in any way similar to the flood-tide of NP-complete problems. Although we will explain a more immediate motivation for the work of this paper, this distinction is reason enough to analyze distributional problems for their potential completeness.

The *standard uniform distribution* on Σ^* is given by $\mu'(x) = \frac{6}{\pi^2}|x|^{-2}2^{-|x|}$. (Given a distribution μ, we let μ' denote the density function on individual strings.) In general, a polynomial-time computable distribution is *uniform* if $\mu'(x) = \rho(|x|)2^{-|x|}$, where $\sum_n \rho(n) = 1$ and there is a polynomial p such that for all n, $\rho(n) \geq 1/p(n)$. Gurevich [Gur91] defined a distribution to be *flat* if there exists a real number $\epsilon > 0$ such that for all but finitely many x, $\mu'(x) \leq 2^{-|x|^\epsilon}$. Some commonly used distributions on graphs are flat and indeed all uniform distributions are flat. Gurevich proved that no distributional problem with a flat distribution is DistNP-complete unless NEXP $=$ EXP. Assuming that NEXP and EXP are distinct classes, this result asserts that certain natural distributions do not yield complete problems. Thus, one might ask whether the reason that we know only a handful of complete distributional problems is because problems can only be complete when their distributions are unnatural. The answer is no. The distributions of known DistNP-complete problems, while not uniform, are all quite reasonable. From our results we will learn that bad distributions do not yield complete problems either—if (L, μ) is a complete problem for NP, then (under the hypothesis that the p-measure of NP is not 0) μ is a reasonable distribution. (For now we will assume that the reader is familiar with this hypothesis and return to describe it later.)

Levin's definitions concern only the distinction between polynomial on the average and super-polynomial on the average. Ben David et al. [BDCGL92] proposed a straightforward generalization and gave a definition of T on the average for an arbitrary time-bound T. However, their definition, as with Levin's, does not distinguish $T(n)$ on the average from $T(cn)$ on the the average, for any function T. Thus, consider a language $L \in \mathrm{DTIME}(4^n)$ that cannot be recognized in time 3^n *almost everywhere*; i.e., every Turing machine that accepts L requires more than 3^n steps on all but some finite number of inputs [GHS91]. According to the definition of Ben David et al. L is 2^n on the μ-average for *every* polynomial-time computable distribution μ. To avoid this, Cai and Selman [CS96] formulated a definition of T on the average that requires, for every n, that the expectation over the set $A_n = \{x \mid |x| \geq n\}$, with respect to the conditional distribution over A_n, be less than or equal to 1. The motivation and effect of this requirement is to remove dependency on any finite set of inputs. As a consequence of their definition, Cai and Selman [CS96] prove a hierarchy theorem for arbitrary average-case time-bounds that is as tight as the Hartmanis-Stearns [HS65] hierarchy theorem for worst-case deterministic time.

Consider for a moment the fundamental question of what it means for a language L to be difficult to recognize. A language that is not in P may still be easy to recognize on many input strings. In contrast, a language that is a.e. complex, or equivalently, P-bi-immune, is difficult to recognize on all but finitely many input strings. Consider the class AVP of all distributional problems (L, μ) that are polynomial on the μ-average according to the definition of Cai and Selman. Let us say that a language is *distributionally-hard to recognize* if for every polynomial-time computable distribution μ, the distributional problem $(L, \mu) \notin \mathrm{AVP}$; i.e., for every μ, no Turing machine that accepts L has a running-time that is polynomial on the μ-average. Cai and Selman [CS96] proved, as a consequence of their hierarchy theorem, that every P-bi-immune language is distributionally-hard to recognize. Here we prove, assuming that the p-measure of NP is not 0, that there exist distributionally-hard to recognize languages that are not P-bi-immune.

Recall that Average-P denotes Levin's class of distributional problems that are polynomial on the μ-average. (We will provide all formal definitions in the next section.) It is obvious from the definitions that $\mathrm{AVP} \subseteq \mathrm{Average\text{-}P}$. Define a distribution to be *reasonable* if there exists a constant $s > 0$ such that $\mu(A_n) = \Omega(\frac{1}{n^s})$. The reason of course is that distributions that decrease too quickly give too much weight to small instances, and for this reason are unreasonable. For distributions that are not reasonable, the two definitions differ. To see this, let L be a language that belongs to $\mathrm{DTIME}(2^n/n)$ but that requires more than $2^n/n^3$ time almost-everywhere. Let μ be an unreasonable distribution for which the conditional probability of the set of strings of length n is 2^{-n}. Then, (L, μ) satisfies Levin's definition and consequently belongs to the class Average-P. However, since L requires exponential time almost-everywhere, it follows that (L, μ) is not polynomial on the μ-average according to Cai and Selman. Thus, (L, μ) is not in AVP. Such distributions as this are pathological and for this reason, perhaps should not be considered. Nevertheless, if we must consider such distributions, then we contend that a language that requires more than polynomial time almost-everywhere is not polynomial time on the average for any distribution, and certainly not for a flat distribution.

Now we come to the crux of this discussion and the more immediate motivation for our next results. If (L_1, μ_1) is reducible to (L_2, μ_2), both μ_1 and μ_2 are reasonable, and (L_2, μ_2) belongs to AVP, then (L_1, μ_1) belongs to Average-P and so, by the equivalence theorem of Cai and Selman, (L_1, μ_1) belongs to AVP also. However, Belanger, Pavan, and Wang [BPW96] have proved that AVP is *not* in general closed under reductions. They have constructed a language L and distributions μ_1 and μ_2 such that μ_2 is reasonable, (L, μ_1) is reducible to (L, μ_2) (by the identity function), $(L, \mu_2) \in$ AVP, and $(L, \mu_1) \notin$ AVP. (Observe as a consequence that μ_1 is not reasonable.) One simple solution is to restrict one's attention to reasonable distributions only. This paper helps to justify this approach, for we show that we do not need to be concerned about the possibility of complete distributional problems that have unreasonable distributions: We prove that, unless the p-measure of NP is 0, every DistNP-complete distributional problem has a reasonable distribution.

2 Preliminaries

We assume that all languages are subsets of $\Sigma^* = \{0,1\}^*$ and we assume that Σ^* is ordered by standard lexicographic ordering.

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 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].) Also, 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 [CS96] 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, 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 AVP $= \bigcup_{k \geq 1}$ AVTIME(n^k). Clearly, 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 [CS96] and Gurevich [Gur91].

Theorem 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*;*

(iii) *There is an integer $k > 0$ such that for all $n \geq 1$,*

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

Now consider reductions. Levin [Lev86] was the first to define polynomial-time many-one reductions on distributional problems; we will use the following form given by Gurevich [Gur91].

Let μ and ν be two distributions. Then, μ is *dominated* by ν, denoted by $\mu \preceq \nu$, if there is a polynomial p such that for all x, $\mu'(x) \leq p(|x|)\nu'(x)$. Let μ_A and μ_B be two distributions and let $f : \Sigma^* \to \Sigma^*$. Recall, for every distribution ν on Σ^*, that f induces a distribution $f(\nu)$ on Σ^* that is defined by $f(\nu)'(y) = \sum_{f(x)=y} \nu'(x)$, for all $y \in range(f)$. Then, we say that μ_A is *dominated by μ_B with respect to f*, denoted by $\mu_A \preceq_f \mu_B$, if there exists a distribution ν such that $\mu_A \preceq \nu$ and for all $y \in range(f)$, $\mu'_B(y) = f(\nu)'(y)$.

[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.

Let (A, μ_A) and (B, μ_B) be two distributional problems. Then (A, μ_A) is *many–one reducible to* (B, μ_B) *in polynomial time*, denoted by $(A, \mu_A) \leq_m^p (B, \mu_B)$, if there exists a polynomial-time computable function $f : \Sigma^* \to \Sigma^*$ such that A is many–one reducible to B via f and $\mu_A \preceq_f \mu_B$.

Gurevich [Gur91] and Wang [Wan97] provide proofs of the following properties.

Lemma 1 *1. Let (A, μ_A) and (B, μ_B) be two distributional problems such that $(A, \mu_A) \leq_m^p (B, \mu_B)$. If $(B, \mu_B) \in$ Average-P, then $(A, \mu_A) \in$ Average-P.*

2. Polynomial-time many-one reductions are transitive.

It is possible to require only that the reduction be computable in polynomial time on the average [Lev86, Gur91]: μ is *weakly dominated* by ν if there is a function g that is polynomial on the μ-average (by Levin's definition) such that for all x, $\mu'(x) \leq g(x)\nu'(x)$. (A, μ_A) is *many–one reducible to* (B, μ_B) *in average polynomial time*, denoted by $(A, \mu_A) \leq_m^{ap} (B, \mu_B)$, if there is a function f that is computable in time a polynomial on the μ_A-average (again, by Levin's definition) such that A is many–one reducible to B via f and μ_A is weakly dominated by some distribution ν such that for all x, $\mu'_B(f(x)) = f(\nu)'(f(x))$.

The analogue of Lemma 1 holds for $\leq_m^{ap}$-reductions.

Once again, if (L_1, μ_1) is reducible to (L_2, μ_2), both μ_1 and μ_2 are reasonable, and (L_2, μ_2) belongs to AVP, then (L_1, μ_1) belongs to Average-P and so, by Theorem 1, (L_1, μ_1) belongs to AVP also. However, Belanger, Pavan, and Wang [BPW96] have proved that AVP is *not* in general closed under reductions.

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).

3 Resource-bounded measure

We refer the reader to the papers of Lutz [Lut92, Lut97] for a general introduction to resource-bounded measure theory. Our exposition is brief.

Identify a language L with its characteristic sequence χ_L defined by

$$\chi_L = [s_0 \in L][s_1 \in L] \ldots,$$

where $\{s_0, s_1, \ldots\}$ is the standard ordering of strings in Σ^*, and $[s_i \in L]$ is 1 if $s_i \in L$, and 0, otherwise.

A *martingale* is a function $d : \Sigma^* \to [0, \infty)$ such that for all $w \in \Sigma^*$,

$$d(w) = \frac{d(w0) + d(w1)}{2}.$$

A martingale d *succeeds* on a language L if

$$\limsup_{n\to\infty} d(\chi_L[0..n-1]) = \infty,$$

where $\chi_L[0..n-1]$ is the initial finite subsequence of χ_L of length n.

Let the classes $p_1 = p$ and p_2, both consisting of functions $f : \Sigma^* \to \Sigma^*$, be the classes

$$\begin{aligned} p_1 &= \{f \mid f \text{ is computable in polynomial time}\} \\ p_2 &= \{f \mid f \text{ is computable in } n^{\log n^{O(1)}} \text{ time}\}. \end{aligned}$$

A martingale d is p_i-computable if there is a function $\hat{d} : N \times \{0,1\}^* \to Q$ such that $\hat{d} \in p_i$ and for all $n \in N$ and $w \in \{0,1\}^*$, $|\hat{d}(n,w) - d(w)| \leq 2^{-n}$. (This definition extends the notion of polynomial time computable distribution given in the previous section.)

A set X of languages has *p_i-measure* 0 ($i = 1, 2$) if there is a p_i-computable martingale that succeeds on every language in X. A set X of languages has *p_i-measure* 1 if the complement of X has p_i-measure 0. We caution that not all sets are measurable. We assume the reader is familiar with standard set-theoretic closure properties of measure theory.

If the p-measure of a class X is 0, then the p_2-measure of X is 0. Lutz has hypothesized that neither the p-measure nor the p_2-measure of NP is 0, and from these strong hypotheses he and others have derived several consequences that do not seem to follow from weaker hypotheses [May94, LM94]. The p-measure of P is 0, and we expect that NP is quantitatively different from P. Thus, results of the form "If A, then the p_i-measure of NP is 0" provide evidence that A is false.

We will apply the following theorem of Lutz [Lut92].

Theorem 2 (Lutz) *Let $F : N \times \Sigma^* \to Q^+$ be a function such that*

1. *For all $k \in N$ and $x \in \Sigma^*$, $F(k, x)$ is computable in polynomial in $k + |x|$.*
2. *For each $k \in N$, $F_k(x) = F(k, x)$ is a martingale.*

Then, $\{A \mid$ for some $k \geq 0, F_k$ succeeds on $A\}$ has p-measure 0.

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 P-bi-immune sets is 1, and therefore, if the p-measure of NP is not 0, then NP contains a P-bi-immune set. Cai and Selman [CS96] 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.) Before turning to our main results, first we will use resource-bounded measure to make a series of simple observations that pinpoint some of the issues that we have been raising.

Remember that all distributions throughout are computable in polynomial time.

1. The set
$$S_1 = \{L \mid \text{for all } \mu, (L, \mu) \notin AVP\}$$
has p-measure 1.

   Proof. Every P-bi-immune set belongs to this set [CS96], and the p-measure of the P-bi-immune sets is 1 [May94]. □

2. The set
$$S_2 = \{L \mid \text{for all reasonable } \mu, (L, \mu) \notin AVP\}$$
has p-measure 1, because every P-bi-immune set belongs to S_2.

3. The set
$$S_3 = \{L \mid \exists \mu, (L, \mu) \in \text{AVP}\}$$
has p-measure 0, because $S_3 = \overline{S_1}$.

4. The set
$$S_4 = \{L \mid \exists \mu, (L, \mu) \in \text{Average-P}\}$$
has p-measure 1.

   Proof. The p-measure of E is 1, and it is easy to see that E is a subset of S_4: For $L \in$ E, take $\mu'(x) = 4^{-x}$. so that $u_n = 2^{-n}$. □

   Since the p-measure of P is 0, we see from items 3 and 4 that AVP is more like a feasible class than is Average-P.

5. The set
$$S_5 = \{L \mid \text{for all } \mu, (L, \mu) \notin \text{Average-P}\}$$
has p-measure 0, because $S_5 = \overline{S_4}$. In fact, $S_5 \cap \text{E} = \emptyset$.

6. The set

$$\begin{aligned} S_6 &= \{L \mid \exists \text{ reasonable } \mu, (L, \mu) \in AVP\} \\ &= \{L \mid \exists \text{ reasonable } \mu, (L, \mu) \in \text{Average-P}\} \end{aligned}$$

has p-measure 0, because $S_6 \subseteq S_3$.

From the p-measures of S_4 and S_6, we see that it is because of unreasonable, distributions that almost all sets L have a distribution μ for which $(L, \mu) \in$ Average-P.

7. The set

$$S_7 = \{L \mid \text{ for all reasonable } \mu, (L, \mu) \notin \text{Average-P}\}$$

has p-measure 1, because $S_7 = \overline{S_6}$.

8. The set

$$\begin{aligned} S_8 &= \{L \mid \text{ for all reasonable } \mu, (L, \mu) \notin \text{Average-P}, \text{ but } \exists \mu, (L, \mu) \in \text{Average-P}\} \\ &= S_7 \cap S_4 \end{aligned}$$

has p-measure 1, because S_7 and S_4 have p-measure 1.

9. The set

$$\begin{aligned} S_9 &= \{L \mid \text{ for all reasonable } \mu, (L, \mu) \notin \text{AVP}, \text{ but } \exists \mu (L, \mu) \in \text{AVP}\} \\ &= S_2 \cap S_3 \end{aligned}$$

has p-measure 0, because $S_2 \cap S_3 \subseteq S_3$, which has p-measure 0.

We do not know whether S_9 is other than the emptyset.

10. If the p-measure of NP is not 0, then the p-measure of the set

$$S_{10} = \{L \in \text{NP} \mid \exists \mu, (L, \mu) \notin \text{AVP})\}$$

is not 0. This follows from item 1.

Now let us focus attention on the set S_1. 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 have noted, every P-bi-immune language is distributionally-hard to recognize.

Theorem 3 *If the p-measure of* NP *is not* 0*, then there is a language L that is distributionally-hard to recognize but not* P*-bi-immune.*

That is, if the p-measure of NP is not 0, then S_1 properly includes the set of P-bi-immune languages. We begin with the following lemma that might have independent interest. A 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].

Lemma 2 *If the p-measure of* NP *is not* 0*, then there exists a set $B \in$* P *such that no infinite subset of B is* P*-printable.*

Proof. The hypothesis implies existence of a P-bi-immune set B' in NP. Since every P-printable set belongs to P, no infinite subset of B' is P-printable. Thus, by a result that Allender and Rubinstein [AR88] attribute to D. Russo, there exists a set $B \in$ P with the same property. $\square$

Proof. The proof of Theorem 3 proceeds as follows: Let A be any set that is DTIME(2^{n^3})-complex. We define $L = A \cup B$. Note that A and B are not disjoint since A is DTIME(2^{n^3})-bi-immune. Since $B \in$ P, clearly, L is not P-bi-immune. Now our goal is to prove that L is distributionally-hard to recognize. The general idea is to suppose that (L, μ) belongs to AVP, for some polynomial-time computable distribution μ, and, from this supposition, demonstrate a P-printable subset of B.

Observe that every Turing machine that recognizes L takes more than 2^{n^3} time on all but finitely many strings of $\overline{B}$. Also, recall, for any distribution μ, that $u_n = \mu(\{x \mid |x| = n\})$.

Lemma 3 *Suppose that μ is a distribution such that (L, μ) is in* AVP*. Then, there exist infinitely many n such that $u_n \neq 0$ and*

$$\mu(\{x \mid x \in \overline{B}, |x| = n\}) \leq \frac{nu_n}{2^{n^2}}.$$

Proof. We prove the claim by contradiction. Let $X_n = \{x \mid x \in \overline{B}, |x| = n\}$. Let N be a positive integer such that for all $n > N$, $u_n \neq 0$ and

$$\mu(X_n) > \frac{nu_n}{2^{n^2}}.$$

We will prove that (L, μ) is not in AVP. Let M be any Turing machine that accepts L, let T_M denote the running time of M, and assume that N is sufficiently large so that $T_M(x) > 2^{|x|^3}$ for all strings $x \in \overline{B}$, $|x| \geq N$. Let $k \geq 1$ be any positive integer.

The following inequalities demonstrate that (L, μ) does not belong to AVP.

$$
\begin{aligned}
\sum_{|x|>N} \frac{T_M^{1/k}(x)\mu'(x)}{|x|} &\geq \sum_{\substack{|x|>N \\ x\in\overline{B}}} \frac{T_M^{1/k}(x)\mu'(x)}{|x|} \\
&\geq \sum_{\substack{m>N \\ u_m\neq 0}} \sum_{\substack{|x|=m \\ x\in\overline{B}}} \frac{T_M^{1/k}(x)\mu'(x)}{|x|} \\
&> \sum_{\substack{m>N \\ u_m\neq 0}} \sum_{\substack{|x|=m \\ x\in\overline{B}}} \frac{(2^{m^3})^{1/k}\mu'(x)}{m} \\
&\geq \sum_{\substack{m>N \\ u_m\neq 0}} \frac{(2^{m^3})^{1/k}\mu(X_m)}{m} \\
&> \sum_{\substack{m>N \\ u_m\neq 0}} \frac{(2^{m^3})^{1/k}}{m}\frac{mu_m}{2^{m^2}} \\
&> \sum_{\substack{m>N \\ u_m\neq 0}} u_m = \sum_{m>N} u_m
\end{aligned}
$$

$\square$

Continuing with the proof of Theorem 3, next we show that $(L,\mu) \notin \mathrm{AVP}$, for every polynomial-time computable distribution μ. Again, by contradiction, suppose that μ is a polynomial-time computable distribution such that $(L,\mu) \in \mathrm{AVP}$.

Define an *interval* $[x_1, x_2]$ to be a finite sequence of strings in increasing order that begins with the string x_1 and ends with the string x_2. (If we identify every string with the number it represents in dyadic notation, then lexicographic order of strings and the natural ordering of the positive integers coincide.) For example, the set of all strings of length n is the interval $[0^n, 1^n]$. Given strings x_1 and x_2 such that x_1 precedes x_2, let $mid(x_1, x_2) = (x_1 + x_2)/2$. Then, $[x_1, mid(x_1, x_2)]$ contains the first $(x_2 - x_1 + 1)/2$ strings in $[x_1, x_2]$, and $[mid(x_1, x_2) + 1, x_2]$ contains the last $(x_2 - x_1 + 1)/2$ strings in $[x_1, x_2]$. We will use the following programming variables to simplify notation: Given an interval $I = [x_1, x_2]$, "Left_I" denotes the interval $[x_1, mid(x_1, x_2)]$, and "Right_I" denotes the interval $[mid(x_1, x_2) + 1, x_2]$.

We define a set T to contain at most one string of length n by the following algorithm:

Current := $[0^n, 1^n]$;
For $i = 1$ **to** n **do**
    if $\mu(\mathrm{Left}_{\mathrm{Current}}) \geq \mu(\mathrm{Right}_{\mathrm{Current}})$
        then Current := $\mathrm{Left}_{\mathrm{Current}}$ **else** Current := $\mathrm{Right}_{\mathrm{Current}}$.

The final value of Current contains exactly one string x. Put x into T if and only if $x \in B$.

Next we will prove that T is an infinite P-printable subset of B, which will complete the proof of Theorem 3. Obviously, T is a subset of B. Since μ is computable in polynomial time, $\mu(\text{Left}_{\text{Current}})$ and $\mu(\text{Right}_{\text{Current}})$ can be computed in polynomial time. Thus, T is P-printable.

We need only to show that T is an infinite set. If x is the final value of Current, $|x| = n$, then by the construction, $\mu'(x) \geq u_n/2^n$. However, by Lemma 3, there exist infinitely many n such that $u_n \neq 0$ and $\mu(X_n) \leq \frac{u_n n}{2^{n^2}}$. Thus, for all such n, $\mu'(x)$ is greater than $\mu(X_n)$. Hence, for all such n, the final value of Current belongs to B. Thus, T is an infinite set.

This completes the proof. □

Observe that Theorem 3 follows from the assumption that NP contains an immune set. The only use of the hypothesis that the p-measure of NP is not 0 is to ensure this assumption.

From the presumably stronger hypothesis that the p_2-measure of NP is not 0, we obtain the stronger result that L belongs to NP:

Corollary 1 *If the p_2-measure of* NP *is not* 0*, then there is a language $L \in$* NP *that is distributionally-hard to recognize but not* P*-bi-immune.*

Proof. From results of Mayordomo [May94], we know that if the p_2-measure of NP is not 0, then there is a set A in NP that is DTIME(2^{n^3})-bi-immune. The same hypothesis implies that the p-measure of NP is not 0, from which Lemma 2 still applies. Thus, the set $L = A \cup B$ belongs to NP. □

4 Complete Distributional Problems

In this section we show that complete distributional problems have reasonable distributions. We begin with the following lemma.

Lemma 4 *Let μ_1 be the standard uniform distribution, so that $\mu_1(\{x \mid |x| = n\}) = n^{-2}$. Let f be a polynomial-time computable reduction from (A, μ_1) to (B, μ_2), where μ_2 is not reasonable. Then, for all $k \geq 1$, there exist infinitely many strings x, such that $|f(x)|^k \leq |x|$.*

Proof. The function f many-one reduces A to B and $\mu_1 \leq_f \mu_2$. Thus, there exists a distribution ν such that $\mu_1 \preceq \nu$ and for all $y \in range(f)$, $\mu_2'(y) = f(\nu)'(y)$. It is easy to to see that ν is reasonable also.

We prove the claim by contradiction. Assume there exist positive integers k and N so that for all strings x, $|x| > N$, $|f(x)|^k > |x|$. We will prove from this assumption that μ_2 is reasonable.

Let $n > N$. Choose s such that $\nu(\{x \mid |x| \geq m\}) = \Omega(m^{-s})$, Consider the following inequalities:

$$
\begin{aligned}
\sum_{|z| \geq n^{1/k}} \mu_2'(z) &\geq \sum_{\substack{|z| \geq n^{1/k} \\ z \in f(\Sigma^*)}} \mu_2'(z) \\
&\geq \sum_{\substack{|z| \geq n^{1/k} \\ z \in f(\Sigma^*)}} \sum_{\substack{f(y)=z \\ |y| \geq n}} \nu'(y) \\
&\geq \sum_{|y| \geq n} \nu'(y) \\
&\geq 1/n^s.
\end{aligned}
$$

Thus, for all $m \geq N^{1/k}$,

$$\sum_{|z| \geq m} \mu_2'(z) \geq 1/m^{ks},$$

which proves that μ_2 is reasonable. $\square$

Theorem 4 *If there exists an $\leq_m^p$-complete distributional problem (L, μ) such that μ is not reasonable, then* NP *has p-measure* 0*.*

Proof. Let (L, μ) be an $\leq_m^p$-complete distributional problem such that μ is not reasonable. Choose $l \geq 1$ such that $L \in \mathrm{DTIME}(2^{n^l})$.

Let $f_0, f_1, \ldots$ be a standard enumeration of the polynomial-time computable functions. Let ν be the standard uniform distribution. Let S belong to NP and let f_k be a $\leq_m^p$-reduction from (S, ν) to (L, μ). By Lemma 4, there exist infinitely many strings x such that $|f_k(x)|^l \leq |x|$.

We define the martingale F_k as follows; the essential idea is to bet only when $|f(s_n)|^l \leq |s_n|$: Define $F_k(\lambda) = 1$, where λ is the empty word. Let $n \geq 1$ and by induction hypothesis assume F_k is defined on all word z of length n. Recall that s_n is the $n+1$-st string in the standard ordering of Σ^*, and note that $|s_n| = \log(n)$.

If $|f_k(s_n)|^l > |s_n|$,
  then define $F_k(z1) = F_k(z0) = F_k(z)$.
If $|f_k(s_n)|^l \leq |s_n|$,
  then if $f_k(s_n) \in L$,
    then define $F_k(z1) = 2F_k(z)$ and $F_k(z0) = 0$,
  else if $f_k(s_n) \notin L$, then define $F_k(z1) = 0$ and $F_k(z0) = 2F_k(z)$.

To apply Theorem 2, define $F(k,n) = F_k(n)$. For each k, F_k is a martingale. Since there exist infinitely many strings x such that $|f_k(x)|^l \leq |x|$, F_k doubles in value on infinitely many partial characteristic sequences of S. Thus, F_k succeeds on S. Finally, the complexity of F_k is determined by the complexity of deciding whether $f_k(s_n) \in L$. Since $L \in \mathrm{DTIME}(2^{n^l})$, the latter is given by $2^{|f_k(s_n)|^l} \leq 2^{|s_n|} \leq n$. Thus, $F(k,z)$ is computable in a polynomial in $k + |z|$. Therefore, by Theorem 2, $\mathrm{NP} = \{A \mid \text{ for some } k \geq 0, F_k \text{ succeeds on } A\}$ has p-measure 0. □

Theorem 5 *If there exists an $\leq_m^{ap}$-complete distributional problem (L,μ) such that μ is not reasonable, then* NP *has p-measure* 0.

Proof. Define the distribution μ_1 by $\mu_1'(0^n) = n^{-2}$, for all $n \geq 1$, and $\mu_1'(x) = 0$, for all $x \notin \{0\}^*$. For all n, $\mu_1(x \mid |x| = n\}) = n^{-2}$. So, by definition, μ_1 is a reasonable distribution. Let $S \in \mathrm{NP}$; choose $l \geq 1$ such that $S \in \mathrm{DTIME}(2^{(n^l)})$. Let f be a function that is computable in time a polynomial on the μ_1-average and that $\leq_m^{ap}$-reduces (S,μ_1) to (L,μ). By Theorem 1, there is a Turing machine M that computes f whose running-time T_M, for some integer $j \geq 1$, satisfies the following inequality, for all $n \geq 1$:

$$\sum_{|x|=n} \frac{T_M(x)^{1/j}}{n} \mu_1'(x) \leq n^{-2}.$$

Thus,

$$\frac{T_M(0^n)^{1/j}}{n} n^{-2} \leq n^{-2},$$

from which it follows that $T_M(0^n) \leq n^j$, for all n. Thus, the restriction of f to $\{0\}^*$ is polynomial-time computable. Let $f_0, f_1, \ldots$ be a standard enumeration of the polynomial-time computable functions define on $\{0\}^*$ and choose $k \geq 1$ such that $f = f_k$.

Similar to Lemma 4, our first task is to demonstrate that for all $s \geq 1$, there exist infinitely many $n \geq 1$ such that $|f(0^n)|^s \leq n$. Let ν weakly dominate μ_1 so that for all strings $y \in range(f)$, $\mu'(y) = f(\nu)'(y)$. There is a function g that is polynomial on the μ_1-average so that for all x, $\mu_1'(x) \leq g(x)\nu'(x)$. As in the previous paragraph, since μ_1 is reasonable, there exists $j \geq 1$ such that for all $n \geq 1$,

$$\sum_{|x|=n} \frac{g(x)^{1/j}}{n} \mu_1'(x) \leq n^{-2},$$

from which, as above, $g(0^n) \leq n^j$. Then,

$$\begin{aligned} \nu(\{x \mid |x| = n\}) &= \sum_{|x|=n} \nu'(x) \\ &\geq \sum_{|x|=n} \mu_1'/g(x) \\ &\geq (n^{-2})(n^{-j}). \end{aligned}$$

It follows readily that ν is reasonable also. Now the proof of our task proceeds exactly as does the proof of Theorem 4.

Again, recall that s_n is the $n+1$-st string in the standard ordering of Σ^* and that $|s_n| = \log(n)$. Now we know that there exist infinitely many strings $s_n \in \{0\}^*$ such that $|f_k(s_n)| \leq |s_n|$.

Now we define the martingale F_k that succeeds on S. This time the idea is to bet only when $|f_k(s_n)|^l \leq |s_n|$ and $s_n \in \{0\}^*$:

If $|f_k(s_n)|^l > |s_n|$ or $s_n \notin \{0\}^*$,
then define $F_k(z1) = F_k(z0) = F_k(z)$.
If $|f_k(s_n)|^l \leq |s_n|$ and $s_n \in \{0\}^*$,
then if $f_k(s_n) \in L$,
then define $F_k(z1) = 2F_k(z)$ and $F_k(z0) = 0$,
else if $f_k(s_n) \notin L$, then define $F_k(z1) = 0$ and $F_k(z0) = 2F_k(z)$.

Then, as in the proof of Theorem 4, using Theorem 2, we conclude that NP has p-measure 0. $\square$

5 Acknowledgments

The proof of Lemma 4 is due to Jin-Yi Cai. Our original assertion of this result was more restrictive and the proof was more complicated. We thank Jin-Yi and the other members of the University at Buffalo's complexity theory research group for their helpful comments. Also, the authors benefited from correspondence with Jack Lutz and D. Sivakumar.

References

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

[BDCGL92] S. Ben-David, B. Chor, O. Goldreich, and M. Luby. On the theory of average case complexity. *J. of Computer and System Sciences*, 44(2):193–219, 1992.

[BG95] A. Blass and Y. Gurevich. Matrix transformation is complete for the average case. *SIAM J. Comput.*, 24:3–29, 1995.

[BPW96] J. Belanger, A. Pavan, and J. Wang. Reductions do no preserve fast convergence rates in average time. Technical Report 96-21, University at Buffalo, Department of Computer Science, 226 Bell Hall, Buffalo, NY 14260, 1996.

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

[CS96] J-Y. Cai and A. Selman. Fine separation of average time complexity classes. In *Procs. of the Thirteenth Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science*, volume 1046, pages 307–318. Springer-Verlag, 1996.

[GHS91] J. Geske, D. Huynh, and J. Seiferas. A note on almost-everywhere-complex sets and separating deterministic-time-complexity classes. *Inf. and Comput.*, 92(1):97–104, 1991.

[Gur91] Y. Gurevich. Average case completeness. *J. 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.

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

[HS65] J. Hartmanis and R. Stearns. On the computational complexity of algorithms. *Trans. Amer. Math. Soc.*, 117:285–306, 1965.

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

[Ko83] K. Ko. On the definition of some complexity classes of real numbers *Math. Systems Theory*, 16:95–109, 1983.

[Lev86] L. Levin. Average case complete problems. *SIAM J. of Comput.*, 15:285–286, 1986.

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

[Lut92] J. Lutz. Almost everywhere high nonuniform complexity. *Journal of Computer and System Sciences*, 44:220–258, 1992.

[Lut97] J. Lutz. The quantitative structure of exponential time. In L. Hemaspaandra and A. Selman, editors, *Complexity Theory Retrospective II*. Springer-Verlag, New York, 1997. In preparation.

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

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

[VL88] R. Venkatesan and L. Levin. Random instances of a graph coloring problem are hard. In *Procs. of the Twentieth Annual ACM Symposium on Theory of Computing*, pages 217–222, 1988.

[VR92] R. Venkatesan and S. Rajagopalan. Average case intractability of diophantine and matrix problems. In *Procs. of the Twenty Fourth Annual ACM Symposium on Theory of Computing*, pages 632–642, 1992.

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

[Wan95] J. Wang. Average-case completeness of a word problem for groups. In *Procs. of the Twenty-Seventh ACM Symposium on Theory of Computing*, pages 325–334, 1995.

[WB95] J. Wang and J. Belanger. On the NP-isomorphism problem with respect to random instances. *J. of Computer and System Sciences*, 50:151–164, 1995.