Improved Learning from Kolmogorov Complexity

Halley Goldberg
Simon Fraser University, Burnaby, Canada

Valentine Kabanets
Simon Fraser University, Burnaby, Canada

Abstract

Carmosino, Impagliazzo, Kabanets, and Kolokolova (CCC, 2016) showed that the existence of natural properties in the sense of Razborov and Rudich (JCSS, 1997) implies PAC learning algorithms in the sense of Valiant (Comm. ACM, 1984), for boolean functions in $\mathsf{P/poly}$, under the uniform distribution and with membership queries. It is still an open problem to get from natural properties learning algorithms that do not rely on membership queries but rather use randomly drawn labeled examples.

Natural properties may be understood as an average-case version of MCSP, the problem of deciding the minimum size of a circuit computing a given truth-table. Problems related to MCSP include those concerning time-bounded Kolmogorov complexity. MKTP, for example, asks for the KT-complexity of a given string. KT-complexity is a relaxation of circuit size, as it does away with the requirement that a short description of a string be interpreted as a boolean circuit. In this work, under assumptions of MKTP and the related problem $\mathsf{MK^tP}$ being easy on average, we get learning algorithms for boolean functions in $\mathsf{P/poly}$ that

- work over any distribution D samplable by a family of polynomial-size circuits (given explicitly in the case of MKTP),
- only use randomly drawn labeled examples from D, and
- are agnostic (do not require the target function to belong to the hypothesis class).

Our results build upon the recent work of Hirahara and Nanashima (FOCS, 2021) who showed similar learning consequences but under a stronger assumption that NP is easy on average.

2012 ACM Subject Classification Theory of computation → Computational complexity and cryptography

Keywords and phrases learning, Kolmogorov complexity, meta-complexity, average-case complexity

Digital Object Identifier 10.4230/LIPIcs.CCC.2023.12

Acknowledgements We thank Shuichi Hirahara, Russell Impagliazzo, Zhenjian Lu, and Igor Oliveira for helpful discussions.

1 Introduction

There is a deep connection between computational learning and pseudorandomness. Loosely speaking, the goal of learning is to extract "structure" (a simple hypothesis) from a "random" environment, whereas the goal of pseudorandom constructions is to hide "structure" within a "random-looking" environment. Before mentioning any examples illustrating this antagonism between learning and pseudorandomness, let us recall the definitions of some basic learning models.

In Valiant's Probably Approximately Correct (PAC) learning model [29], a learner tries to learn an unknown concept c (say, a Boolean function) from a known class $\mathcal{C}$ of concepts, with respect to some (arbitrary) distribution D over inputs to c. The learner gets to see independently sampled labeled examples of the form $(x, c(x))$, where x is sampled by D, and needs to output (with high probability) a hypothesis h that has just tiny disagreement with c with respect to the distribution D. The agnostic PAC learning model [21] is a natural generalization of the PAC model where an unknown concept f to be learned is *not necessarily*

from the concept class $\mathcal{C}$. The learner gets to see independently sampled labeled examples of the form $(x, f(x))$, with x sampled from some distribution D, and needs to output (with high probability) a hypothesis h so that the disagreement between h and f with respect to D is very close to the disagreement between f and the concept $c_f \in \mathcal{C}$ that is closest to f with respect to D. Classical (agnostic) PAC learning model is distribution-independent in the sense that a successful PAC learning algorithm for a concept class $\mathcal{C}$ must work with respect to *any* distribution D of examples. One also considers a distribution-specific setting where a learning algorithm must work with respect to a single fixed distribution D, e.g., the uniform distribution or a polytime-samplable distribution.

Impagliazzo and Levin [18] and Blum et al. [5] (see also [24]) show that breaking *cryptographic* Pseudorandom Generators (PRGs) implies *average-case* PAC learning with respect to polytime-samplable distributions; here, rather than learning *every* concept from some concept class $\mathcal{C}$, one gets to learn a *significant fraction* of concepts from $\mathcal{C}$ under a polytime-samplable distribution over $\mathcal{C}$. In contrast, Nisan and Wigderson [25] show that breaking *complexity-theoretic* PRGs (namely, the NW generators) implies worst-case learning (of every concept in a given concept class $\mathcal{C}$) under the uniform distribution, but here the learning algorithm needs to make *membership queries* (MQs) to the concept $c \in \mathcal{C}$ it is trying to learn, i.e., the learner gets to ask the value $c(x)$ for any input x of its choosing.

Where does one get an algorithm to break a given PRG in order to get a learning algorithm? For the case of the NW PRG, Carmosino et al. [10] showed that a natural property (in the sense of Razborov and Rudich [27]) for a (sufficiently expressive) circuit class $\mathcal{C}$ yields a learning algorithm for $\mathcal{C}$ under the uniform distribution, with membership queries; this was generalized to learning with respect to polytime-samplable distributions by Binnendyk et al. [4]. Using a known natural property for the class $\mathsf{AC}^0[p]$ of constant-depth circuits with AND, OR, NOT, and mod-p counting gates (for any prime modulus p) from [27], [10] obtained a quasipolynomial-time learning algorithm for $\mathsf{AC}^0[p]$ over the uniform distribution, using membership queries. Later, [11] generalized this framework to show that one also gets *agnostic* learning algorithms from certain generalizations of natural properties. It remains an important open problem to get from a natural property a learning algorithm that uses only random labeled examples. In particular, it would be very interesting to get an efficient learning algorithm for $\mathsf{AC}^0[p]$ without membership queries, which would rule out *weak* Pseudorandom Function Generator constructions in $\mathsf{AC}^0[p]$; see [6] for a recent survey on pseudorandom functions.

A natural property for general circuits is an efficient average-case heuristic for Minimum Circuit Size Problem (MCSP) over the uniform distribution, with one-sided error. Namely, it should always accept the truth tables of Boolean functions of low circuit complexity (for a given threshold size parameter s), and should reject at least a constant fraction of all possible truth tables. MCSP is an example of a meta-complexity problem asking to estimate the circuit size of a given truth table. There are closely related meta-complexity problems for variants of time-bounded Kolmogorov complexity.

For example, MKTP (defined in [1]) asks if a given binary string x is efficiently *locally* computable (outputting bit x_i on any input i in at most t steps) by a universal Turing machine with oracle access to some short binary string d, where one seeks to minimize the sum $|d| + t$. As MCSP, MKTP asks for a description of a string x that allows one to compute x locally, any bit x_i at a time. However, such a description of x needn't be a Boolean circuit for the truth table x, the time t of an algorithm computing each x_i is explicitly taken as part of the complexity measure of x, and this reconstruction algorithm is given random access to the description string d.

We show that this extra flexibility of MKTP compared to MCSP leads to improved learning algorithms from an assumed "natural property" (or one-sided error average-case heuristic) for MKTP, where we get agnostic PAC learning algorithms over explicitly given efficiently samplable distributions. Recall that $\mathsf{SIZE}[s(n)]$ denotes the set of all n-variate Boolean functions computable by circuits of size at most $s(n)$. We have the following.

▶ **Theorem 1** (Learning from MKTP: Informal version)**.** *Suppose* MKTP *has an efficient one-sided error average-case heuristic over the uniform distribution over inputs. Then for any circuit size bound $s(n) \leq \mathsf{poly}(n)$, the concept class $\mathcal{C} = \mathsf{SIZE}[s(n)]$ is agnostic PAC-learnable in polytime with respect to any explicitly given ensemble of polysize-samplable distributions $D = \{D_n\}$.*

Here an ensemble D of distributions D_n is polysize-samplable if there is a family of polysize circuits $Samp_n$ that are samplers for D_n, i.e., the distribution of outputs of $Samp_n$ on uniformly random inputs is D_n. Explicitness of D means that a learning algorithm, when asked to learn some n-variate Boolean function, is explicitly given a description of the sampling circuit $Samp_n$ for the distribution D_n. Note that this explicitness condition for $D = \{D_n\}$ is trivially satisfied by the uniform distribution or any polytime-samplable distribution ensemble (in the latter case, one just needs a constant-size description of a polytime Turing machine that samples according to D_n, for any given n).

For the learning setting over distributions $D = \{D_n\}$ where D is polysize-samplable, but the sampling circuits $Samp_n$ are *not* explicitly given to the learning algorithm (and only their sizes are given), we can get efficient agnostic PAC learning from a "natural property" for a different Kolmogorov-complexity measure, K^t. Recall that, for any time parameter $t \in \mathbb{N}$, $\mathsf{K}^t(x)$ is defined as the length of a shortest string $d \in \{0,1\}^*$ such that a universal Turing machine with input d outputs the string x within t steps. Note that, in contrast to KT, here the time t to reconstruct a given string x is a parameter rather than part of the complexity measure of x, and there is no requirement to compute x locally. The minimization version of K^t, denoted $\mathsf{MK^tP}$, needs to decide, for a given binary string x and a size parameter s, if $\mathsf{K}^t(x) \leq s$. We have the following.

▶ **Theorem 2** (Learning from $\mathsf{MK^tP}$: Informal version)**.** *Suppose* $\mathsf{MK^tP}$ *has an efficient one-sided error average-case heuristic over the uniform distribution over inputs. Then for any circuit size bound $s(n) \leq \mathsf{poly}(n)$, the concept class $\mathcal{C} = \mathsf{SIZE}[s(n)]$ is agnostic PAC-learnable in polytime with respect to any ensemble of polysize-samplable distributions $D = \{D_n\}$.*

The conclusion of Theorem 2 is stronger than that of Theorem 1, as it does away with the requirement of explicitness of D. Though we cannot yet reach the same conclusion under average-case easiness of MKTP, we make some partial progress; we show that under *worst-case* easiness of MKTP, learning is possible without the sampling circuit explicitly given.

▶ **Theorem 3** (Learning from worst-case MKTP: Informal version)**.** *Suppose* MKTP *is decidable by an efficient randomized algorithm. Then for any circuit size bound $s(n) \leq \mathsf{poly}(n)$, the concept class $\mathcal{C} = \mathsf{SIZE}[s(n)]$ is agnostic PAC-learnable in polytime with respect to any ensemble of polysize-samplable distributions $D = \{D_n\}$.*

Below we explain our results and proof techniques in more detail.

1.1 Results

In this work, we present agnostic PAC-learners for polynomial-size circuits over efficienly samplable distributions, under assumptions of problems of time-bounded Kolmogorov complexity being easy on average. More specifically, we consider the problem of learning an

unknown target function $f\colon \{0,1\}^n \to \{0,1\}$ with respect to a concept class $\mathcal{C}$ and a class $\mathcal{D}$ of ensembles of distributions. Learnability here is *agnostic* in the sense that f does not necessarily belong to $\mathcal{C}$, and our learner is asked to learn f with error that is just a small additive ε over the disagreement between f and the closest function in $\mathcal{C}$ to f, with high probability. In this case, we say the algorithm achieves ε-agnostic learning; see Section 2.3 for more precise definitions.

We will typically take $\mathcal{C}$ to be $\mathsf{SIZE}[s(n)]$ for some function $s\colon \mathbb{N} \to \mathbb{N}$: that is, the class of functions computable by boolean circuits of size $s(n)$. Our agnostic PAC-learners have access to an *example oracle* $\mathsf{EX}(f, D)$, with each query returning an independent and identically distributed pair (x, b), where x is sampled according to the distribution D and $b = f(x)$. The *sample complexity* of the learning algorithm is the number of queries made to $\mathsf{EX}(f, D)$. Note that our learners may *not* ask membership queries of the target function f.

We will typically take $\mathcal{D}$ to be $\mathsf{Samp}[T(n)]/a(n)$ for some functions $T, a\colon \mathbb{N} \to \mathbb{N}$, i.e., the class of distributions samplable non-uniformly in time $T(n)$ and with $a(n)$ bits of advice. We consider two different kinds of access to the target distribution D. The first is *white-box* access, where the learner is explicitly given the $a(n)$ bits of advice required to sample D (as well as the parameters $T(n)$ and $a(n)$ that define the distribution class $\mathcal{D}$). In this case, we will say that $\mathcal{C}$ is *agnostic PAC-learnable over w.b.-*$\mathsf{Samp}[T(n)]/a(n)$. In the second kind of access to D, the learner is not given the advice to sample D but is given the parameters $T(n)$ and $a(n)$. In this case, we will simply say that $\mathcal{C}$ is *agnostic PAC-learnable over* $\mathsf{Samp}[T(n)]/a(n)$.[1]

We also consider two different notions of time-bounded Kolmogorov complexity. The *minimum* KT*-complexity problem*, MKTP, is the problem of deciding, given a string $x \in \{0,1\}^*$ and a parameter $s \in \mathbb{N}$, whether the KT-complexity of x is at most s. Roughly speaking, KT-complexity is the minimum $|d| + t$ such that a universal TM with oracle access to $d \in \{0,1\}^*$ can compute any individual bit of x in time $t \in \mathbb{N}$. $\mathsf{MK^tP}$ is defined analogously, where K^t-complexity is the minimum description length $|d|$ such that a universal TM U on input d outputs (the whole string) x in time t. See Section 2.2 for formal definitions of these measures of time-bounded Kolmogorov complexity and the associated decision problems.[2]

As mentioned earlier, our notion of an average-case heuristic over the uniform distribution $\mathcal{U}$ over inputs for MKTP or $\mathsf{MK^tP}$ mimics the one-sided error definition of a natural property of [27], where all yes-instances must be accepted, and a constant fraction of all instances must be rejected. Given the extreme sparsity of yes-instances of these problems over the uniform distribution, we easily get required *one-sided error* average-case heuristics for these problems from *errorless* average-case heuristics; the class of errorless randomized heuristics is denoted by AvgBPP (see Section 2.1 for the precise definition). For example, our assumption that there is an efficient errorless randomized heuristic for MKTP under the uniform distribution over inputs will be denoted by $(\mathsf{MKTP}, \mathcal{U}) \in \mathsf{AvgBPP}$.

Learning over explicitly given efficiently samplable distributions

Here we give a more formal statement of our Theorem 1.

---

[1] For context, the first model is that employed in the recent work of [4], and the second model is that employed in the recent work of [16]. In the original PAC-learning framework of [29], the target distribution is allowed to be completely unknown and arbitrary.

[2] $\mathsf{MK^tP}$ has elsewhere been denoted MINKT.

▶ **Theorem 4.** *Suppose* $(\mathsf{MKTP}, \mathcal{U}) \in \mathsf{AvgBPP}$. *Then for any time-constructible function* $s\colon \mathbb{N} \to \mathbb{N}$, *polynomials* $T, a\colon \mathbb{N} \to \mathbb{N}$, *and* $\varepsilon \in (n^{-d}, 1)$ *for a constant* $d > 0$, *the concept class* $\mathsf{SIZE}[s(n)]$ *is* ε*-agnostic PAC-learnable on w.b.-*$\mathsf{Samp}[T(n)]/a(n)$

- *in time* $\mathsf{poly}(n, s(n), T(n), a(n), \varepsilon^{-1})$ *and*
- *with sample complexity* $((s(n) + n)^3 \cdot \varepsilon^{-26})^{1+o(1)}$.

Proof. The theorem follows by combining Theorem 19 and Theorem 36 below. ◀

Learning over unknown efficiently samplable distributions

Below we give a formal statement of our Theorem 2. For $\mathsf{MK^tP}$, as above, we allow errorless *randomized* heuristics.

▶ **Theorem 5.** *Suppose* $(\mathsf{MK^tP}, \mathcal{U}) \in \mathsf{AvgBPP}$. *Then for any time-constructible functions* $s, T, a\colon \mathbb{N} \to \mathbb{N}$ *and* $\varepsilon \in (0, 1)$, *the concept class* $\mathsf{SIZE}[s(n)]$ *is* ε*-agnostic learnable on* $\mathsf{Samp}[T(n)]/a(n)$

- *in time* $\mathsf{poly}(n, s(n), T(n), a(n), \varepsilon^{-1})$ *and*
- *with sample complexity* $((s(n) + a(n) + \log T(n))^3 \cdot \varepsilon^{-8})^{1+o(1)}$.

Proof. The theorem follows by combining Theorem 19 and Theorem 38 below. ◀

Finally, we give a formal statement of our Theorem 3.

▶ **Theorem 6.** *Suppose* $\mathsf{MKTP} \in \mathsf{BPP}$. *Then for any time-constructible functions* $s, T, a\colon \mathbb{N} \to \mathbb{N}$ *and* $\varepsilon \in (0, 1)$, *the concept class* $\mathsf{SIZE}[s(n)]$ *is* ε*-agnostic learnable on* $\mathsf{Samp}[T(n)]/a(n)$ *in time and sample complexity* $\mathsf{poly}(n, s(n), T(n), a(n), \varepsilon^{-1})$.

Proof. The theorem follows by combining Theorem 19 and Theorem 41 below. ◀

1.2 Techniques

All of our proofs work by way of the known reduction, due to Kothari and Livni [22], from agnostic PAC-learning to the task of *correlative RRHS-refutation*. Consider polynomials $s(n)$, $T(n)$, and $a(n)$. Given a concept class $\mathsf{SIZE}[s(n)]$, a distribution class $\mathsf{Samp}[T(n)]/a(n)$, and a tuple of labeled strings $(\langle x^{(1)}, b^{(1)}\rangle, ..., \langle x^{(m)}, b^{(m)}\rangle)$, where each $x^{(i)} \sim D_n$ for some distribution $D \in \mathsf{Samp}[T(n)]/a(n)$, a correlative RRHS-refuter R is asked to distinguish the following two cases:

- A "correlative case", in which the labels b are correlated with the output of some $s(n)$-size circuit f; that is, for each $1 \le i \le m$, independently,

$$\Pr_{x^{(i)} \sim D}\left[b^{(i)} = f(x^{(i)})\right] \ge \frac{1}{2} + \frac{\varepsilon}{2},$$

- and a "random case", in which the labels $b^{(i)}$ are sampled independently and uniformly at random.

Kothari and Livni show that if there is a probabilistic polynomial-time algorithm R satisfying the above conditions, then there is an agnostic learner for f over D. The proof of this statement essentially uses *distribution-specific* boosting algorithms for the agnostic setting, as given by Feldman [12] and Kalai and Kanade [20]. The fact that the distribution D remains the same during polynomially many boosting stages is crucial as it keeps the circuit complexity of the sampler for D polynomially bounded.

For our results, the key intuition is that the concatenated samples in the correlative case will have lower time-bounded Kolmogorov complexity than those in the random case, since the complexity of uniformly random labels $(b^{(1)}, ..., b^{(m)})$ is close to its maximum value

$m + O(1)$ with very high probability. Choosing $m = \mathsf{poly}(n)$ sufficiently larger than the circuit-complexity $s(n)$ of the target function f yields the desired gap between the two cases. In this way, a heuristic algorithm for computing time-bounded K-complexity may be used as a correlative RRHS-refuter.

A first observation is that, regardless of the version of time-bounded K-complexity available as a heuristic algorithm, it is easy to construct a correlative RRHS-refuter working over the uniform distribution. For example, suppose $(\mathsf{MKTP}, \mathcal{U}) \in \mathsf{AvgBPP}$. Let $X := (x^{(1)}, ..., x^{(m)}) \in \{0,1\}^{nm}$ and $b := (b^{(1)}, ..., b^{(m)}) \in \{0,1\}^m$. On one hand, in the correlative case, we will always have

$$\mathsf{KT}(X, b) \leq nm + \ell_s(n) + \delta \cdot m,$$

where $\ell_s(n) \leq O(s(n) \log s(n))$ is the length of an encoding of a circuit for f, and a further $\delta \cdot m$ bits for a constant $\delta < 1$ are used to encode the discrepancy between the labels b and the true outputs of f (see Lemma 34). In particular, given X, it is easy to construct b from the outputs of f (and knowing which of the labels $b^{(i)}$ are incorrect). On the other hand, in the random case, $(X, b) \sim \mathcal{U}_{nm+m}$. Since the KT-complexity of uniformly random strings is usually close to maximum, we have that with high probability,

$$\mathsf{KT}(X, b) \geq nm + m - 10.$$

It is not hard to see that our randomized heuristic for deciding KT-complexity will serve as a randomized distinguisher between these two cases.

For distributions other than uniform, the situation is less straightforward. In particular, our heuristic algorithms are only defined to work well over $\mathcal{U}$; moreover, $\mathsf{KT}(X, b)$ is not necessarily likely to be large. In recent work, Hirahara and Nanashima [16] circumvent these obstacles under the assumption that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$. In particular, they use this assumption to construct a worst-case algorithm approximating K^t within logarithmic additive error. They also use it to prove a worst-case *weak Symmetry of Information* theorem for K^t, which conditionally states that for some polynomial p, for every $X \in \{0,1\}^*$,

$$\mathsf{K}^t(X, b) \geq \mathsf{K}^{p(t)}(X) + |b| - O(\log t)$$

with high probability over a uniformly random string b. The above inequality is used in the random case of RRHS-refutation. In the correlative case, as above, it holds that

$$\mathsf{K}^t(X, b) \leq \mathsf{K}^{t'}(X) + \ell_s(n) + \delta \cdot m$$

for some time-bound $t' < t$. The authors then use the worst-case algorithm for K^t to approximate the value of $\mathsf{K}^t(X, b) - \mathsf{K}^\tau(X)$ for an appropriate choice of τ, thereby distinguishing the two cases. To overcome the technical issue of the different time bounds $p(t)$ and t' in the expressions above, they show that such differences are immaterial in the expectation over an efficiently samplable distribution:[3] for any $D \in \mathsf{Samp}[mT(n)]/a(n)$ and sufficiently large time bound t,

$$\mathop{\mathbb{E}}_{X \sim D} \left[\mathsf{K}^t(X) - \mathsf{K}(X) \right] \leq O(\log(mT(n))) + a(n). \tag{1}$$

In other words, both $\mathsf{K}^t(X)$ and $\mathsf{K}^{t'}(X)$ are likely close enough to $\mathsf{K}(X)$, and therefore close enough to each other.

[3] Note that for $x^{(i)} \sim D$, for $D \in \mathsf{Samp}[T(n)]/a(n)$, we have $X = (x^{(1)}, ..., x^{(m)}) \sim D'$, for $D' \in \mathsf{Samp}[mT(n)]/a(n)$.

1.2.1 Learning from $\mathsf{MK^tP}$

To prove our result for $\mathsf{MK^tP}$, we show that similar arguments may be carried out under a significantly weaker assumption. One issue is that [16] uses the assumption $\mathsf{DistNP} \subseteq \mathsf{AvgP}$ to achieve derandomization, as shown possible by [9]. Roughly speaking, one "encodes" a string x into a distribution $\mathsf{DP}(x)$ such that any efficient algorithm distinguishing $\mathsf{DP}(x)$ from uniform can be used to show that $\mathsf{K}^t(x)$ is small, a process that crucially relies on the derandomization of the DP reconstruction. Such derandomization is not known to hold under the assumption $\mathsf{DistNP} \subseteq \mathsf{AvgBPP}$, let alone our weaker assumption of $(\mathsf{MK^tP}, \mathcal{U}) \in \mathsf{AvgBPP}$, where $\mathsf{MK^tP}$ is not even known to be NP-hard. In our setting, compression via the DP generator gives a randomized algorithm A that, for any string X and sufficiently large $t \in \mathbb{N}$, outputs a value $\tilde{s} \in \mathbb{N}$ such that

$$\mathsf{pK}^{\mathsf{poly}(t)}(X) - O(\log t) \leq \tilde{s} \leq \mathsf{K}^t(X),$$

where $\mathsf{pK}^{\mathsf{poly}(t)}$ denotes a *probabilistic* measure of time-bounded Kolmogorov complexity. See Section 2.2 for a definition. In general, $\mathsf{pK}^t(X)$ could be much smaller that $\mathsf{K}^t(X)$, so the algorithm A does not appear very useful a priori. However, as we outline below, it turns out to be sufficient for the purposes of learning.

Another challenge in our setting is to argue for Eq. (1) above, which says that $\mathsf{K}^t(X)$ is close to $\mathsf{K}(X)$ in the expectation. In [16], the proof of that statement relies on a conditional *source-coding* theorem for K^t: if $\mathsf{DistNP} \subseteq \mathsf{AvgP}$, then for any distribution $D \in \mathsf{Samp}[mT(n)]/a(n)$ and $X \in \mathrm{supp}(D)$,

$$\mathsf{K}^{\mathsf{poly}(mT(n))}(X) \lesssim \log(1/D(X)), \tag{2}$$

where $D(X)$ denotes the probability of X under D, and "$\lesssim$" hides the term $O(\log(mT(n))) + a(n)$. Specifically, to prove Eq. (1) from this statement, one observes that

$$\begin{aligned} \mathop{\mathbb{E}}_{X\sim D}\left[\mathsf{K}^t(X)\right] &\lesssim \mathop{\mathbb{E}}_{X\sim D}\left[\log(1/D(X))\right] \\ &= H(D) \\ &\leq \mathop{\mathbb{E}}_{X\sim D}[\mathsf{K}(X)], \end{aligned}$$

where $H(D)$ denotes the Shannon entropy of the distribution D.

In our setting, without derandomization, Eq. (2) is not known to hold. Unconditionally, it is only known that with high probability over $r \sim \mathcal{U}_{\mathsf{poly}(mT(n))}$,

$$\mathsf{K}^{\mathsf{poly}(mT(n))}(X, r) \lesssim \log(1/D(X)) + |r|. \tag{3}$$

That is, source coding for K^t only holds in the presence of additional uniform randomness. A statement of this kind was originally proved in [3]. In analogy with Eq (1), we may use Eq. (3) to prove that

$$\mathbb{E}\left[\mathsf{K}^t(X, r) - \mathsf{K}(X, r)\right] \leq O(\log(mT(n))) + a(n), \tag{4}$$

for $X \sim D$ and $r \sim \mathcal{U}_{\mathsf{poly}(mT(n))}$.

We cope with the necessity of this additional randomness r by *incorporating it* into our correlative RRHS-refuter R. That is, we use the randomness of R to uniformly sample a string r, and rather than approximating $\mathsf{K}^\tau(X)$ and $\mathsf{K}^t(X, b)$, we approximate $\mathsf{K}^\tau(X, r)$ and $\mathsf{K}^t(X, b, r)$. We show the analysis of the RRHS-refutation to be unharmed by this modification.

Importantly, Eq. (4) allows us to make use of our inferior approximation algorithm A described at the beginning of this section. For any strings X and r, $\mathsf{pK}^t(X, r)$ is known to be lower-bounded by the time-unrestricted $\mathsf{K}(X, r)$. Eq. (4) then implies that the expected value of $\mathsf{K}^t(X, r) - \mathsf{pK}^t(X, r)$ will be low, for X sampled from an efficiently samplable distribution and r sampled uniformly at random. Intuitively, there is a "smoothing out" of the differences between different measures of Kolmogorov complexity in the expectation, so the correlative RRHS-refuter we construct may sometimes safely ignore such differences.

Finally, there is the issue of the Symmetry of Information theorem for K^t, which is not known to hold in the absence of derandomization. To get around this, we observe that such a statement is actually *not necessary* for our purposes. Rather, since $\mathsf{K}^t(X, b, r)$ will be close to $\mathsf{K}(X, b, r)$ with high probability over $X \sim D_n^m$, $b \sim \mathcal{U}_m$, and $r \sim \mathcal{U}_{\mathsf{poly}(mT(n))}$, we may simply apply the well-known, unconditional Symmetry of Information theorem for time-unbounded Kolmogorov complexity. This observation has the advantage of simplifying our proofs as well as painting a clearer picture of the true prerequisites of learning.

1.2.2 Learning from MKTP

Many of the tools available in the K^t setting, such as compression via generator reconstruction yielding a worst-to-average reduction, become unavailable in the setting of KT. For this reason, we can no longer apply the framework of [16], and we obtain a model of learning that requires a stronger form of access to the target distribution in question. In this setting, we take advantage of the fact, as described above, that it is easy to learn via KT over the uniform distribution. Our goal is then to reduce the task of learning over arbitrary distributions in $\mathsf{PSamp/poly}$ to that of learning over the uniform distribution. Inspired by a recent work of Binnendyk et al. [4], we employ the *distributional inverters* of [19]. A distributional inverter for a function $g\colon \{0,1\}^* \to \{0,1\}^*$ is an algorithm that, given some $y = g(x)$, outputs a nearly uniformly random member of the set $\{z \mid g(z) = y\}$. It is already known that $(\mathsf{MKTP}, \mathcal{U}) \in \mathsf{AvgBPP}$ implies the existence of such objects for every efficiently computable g (see Section 2.4).

To construct a correlative RRHS-refuter for an arbitrary distribution $D \in \mathsf{PSamp/poly}$, we apply distributional inversion in the following way. Let I be a distributional inverter for the sampler for D, which is a polynomial-size circuit C. Recall that in the problem of correlative RRHS-refutation, we are provided labeled examples $\{(x^{(i)}, b^{(i)})\}$, where either the $b^{(i)}$s are uniformly random, or they are correlated with the outputs $f(x^{(i)})$ of the target function f. Given such pairs $\{(x^{(i)}, b^{(i)})\}$, we apply I to the first part to simulate pairs of the form $\{(r^{(i)}, b^{(i)})\}$, where the $r^{(i)}$s are now from a distribution close to uniform, and the $b^{(i)}$s are either uniformly random, or they are correlated with the outputs $f(C(r^{(i)}))$ of the target function f composed with the sampler C. Using a correlative RRHS-refuter for $f \circ C$ over the uniform distribution, we can distinguish these two cases, thereby distinguishing the two cases of the original problem over D. Because I must have oracle access to the non-uniform circuit C it is inverting, the learner will ultimately require an *explicit* description of C, so that the learner can output a circuit for f with no extra oracle gates.

1.3 Related Work

[16] proved a version of Theorem 2 under the assumption that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$. In [13], the authors adapted the learning result of [16] to the case of the randomized average-case easiness assumption that $\mathsf{DistNP} \subseteq \mathsf{AvgBPP}$, by showing that the probabilistic Kolmogorov complexity measure pK^t may be used instead of K^t, and proving (under the same average-case easiness

assumption) various results for pK^t (e.g., the existence of a randomized approximation algorithm for pK^t, and the Symmetry of Information). Using [13], it is fairly straightforward to get a learning algorithm from the assumption that $\mathsf{MpK^tP}$ (the minimization problem for pK^t) is in AvgBPP, relying on the properties of pK^t proved in [13]. However, in the present paper, we use a weaker assumption that $\mathsf{MK^tP}$ is in AvgBPP (under the uniform distribution), and avoid using any nontrivial properties of pK^t. Intuitively, the reason we are able to do so is the "smoothing out" phenomenon mentioned above: the time-bounded Kolmogorov complexity measures K^t and pK^t are close to the time-unbounded measure K, in expectation over appropriate efficiently samplable distributions.

Recall that $\mathsf{Partial\text{-}MCSP}$ is the problem of deciding, given a collection of pairs $\{(x_i, b_i)\}$, whether there is a small circuit C such that every $C(x_i) = b_i$. Ilango, Loff, and Oliveira prove that under an average-case easiness assumption about $\mathsf{Partial\text{-}MCSP}$, PAC-learning without membership queries is possible over the uniform distribution [17]. This relies on a reduction of Vadhan [28] from PAC-learning (in a distribution-*independent* sense) to the problem of "RRHS-refutation": namely, the simpler version of correlative RRHS-refutation in which the labels b_i are precisely the outputs of the target concept f applied to the samples x_i. We expect that by using the tools of this work, including correlative RRHS-refutation and distributional inversion, the statement of [17] could be extended to the agnostic setting and arbitrary efficiently samplable distributions, in the sense of our Theorem 1.

2 Preliminaries

2.1 Average-case Complexity

A *distributional problem* is a pair (L, D), where $L \subseteq \{0,1\}^*$ is a language and D is a family of distributions $\mathcal{D} = \{D_n\}_{n\in\mathbb{N}}$. We denote by $\mathcal{U}$ the family of *parameterized uniform distributions* $\{\mathcal{U}_{\langle n,t_1,\ldots,t_k\rangle}\}_{n,t_1,\ldots,t_k\in\mathbb{N}}$, where k is a constant, each $\mathcal{U}_{\langle n,t_1,\ldots,t_k\rangle} := (\mathcal{U}_n, 1^{t_1}, \ldots, 1^{t_k})$, and $\mathcal{U}_n$ is the uniform distribution over n-bit strings.[4]

▶ **Definition 7** (AvgBPP [7]). *A distributional problem (L, D) belongs to AvgBPP if there is an algorithm A and polynomial p such that, on any $n \in \mathbb{N}$, $x \in supp(D_n)$, and $\delta > 0$, $A(x; n, \delta)$ runs in time at most $p(n/\delta)$, and*

1. $\Pr_A [A(x; n, \delta) \notin \{L(x), \perp\}] \leq \frac{1}{10}$;
2. $\Pr_{x\sim D_n} [\Pr_A[A(x; n, \delta) = \perp] \geq 1/10] \leq \delta(n)$.

Such an algorithm A is called a randomized errorless heuristic scheme for (L, D).

2.2 Time-bounded Kolmogorov Complexity

▶ **Definition 8** (KT [1]). *Fix a universal oracle TM $\mathcal{U}$. For strings $x, y \in \{0,1\}^*$, the* KT*-complexity of x given y is defined as*

$$\mathsf{KT}(x \mid y) := \min_{d\in\{0,1\}^*,\, t\in\mathbb{N}} \left\{|d| + t \mid \forall\, 1 \leq i \leq N+1,\ \mathcal{U}^{d,y}(i) = x_i \text{ in at most } t \text{ steps}\right\},$$

where $x_{N+1} := \perp$, and the notation $\mathcal{U}^{d,y}$ means that $\mathcal{U}$ has random (oracle) access to strings d and y.

[4] Formally, $\langle t_1, \ldots, t_k\rangle$ denotes $\mathsf{Enc}(t_1, \ldots, t_k)$, where $\mathsf{Enc} : \mathbb{N}^* \to \mathbb{N}$ is an efficiently computable and decodable encoding function. Such an encoding function is known to exist by standard techniques; see, for example, [7].

▶ **Definition 9** (K^t). *Fix a universal deterministic TM $\mathcal{U}$. For strings $x, y \in \{0,1\}^*$ and a time bound $t \in \mathbb{N}$, the t-time-bounded Kolmogorov complexity of x given y is defined as*

$$\mathsf{K}^t(x \mid y) := \min_{k \in \mathbb{N}} \left\{ k \;\middle|\; \exists\, w \in \{0,1\}^k,\ \mathcal{U}(w, y) \text{ outputs } x \text{ within } t \text{ steps} \right\}.$$

▶ **Definition 10** (pK^t_δ). *Fix a universal deterministic TM $\mathcal{U}$. For strings $x, y \in \{0,1\}^*$, a time bound $t \in \mathbb{N}$, and $\delta \in [0,1]$, the δ-probabilistic t-time-bounded Kolmogorov complexity of x given y is defined as*

$$\mathsf{pK}^t_\delta(x \mid y) := \min_{k \in \mathbb{N}} \left\{ k \;\middle|\; \Pr_{r \sim \{0,1\}^t} \left[\exists\, w \in \{0,1\}^k,\ \ \mathcal{U}(w, y, r) \text{ outputs } x \text{ within } t \text{ steps} \right] \geq \delta \right\}.$$

▶ **Definition 11** (MKTP and $\mathsf{MK^tP}$). *We define languages*

- $\mathsf{MKTP} := \{(x, 1^s) \mid x \in \{0,1\}^*,\ s \in \mathbb{N},\ \text{and}\ \ \mathsf{KT}(x) \leq s\}$;
- $\mathsf{MK^tP} := \{(x, 1^s, 1^t) \mid x \in \{0,1\}^*,\ s, t \in \mathbb{N},\ \text{and}\ \ \mathsf{K}^t(x) \leq s\}$;

▶ **Proposition 12.** *For any string $x \in \{0,1\}^*$ and time bound $t \in \mathbb{N}$,*

$$\mathsf{pK}^t(x) \leq \mathsf{K}^t(x).$$

▶ **Proposition 13** ([13]). *There is a constant c such that, for any string $x \in \{0,1\}^*$ and time bound $t \in \mathbb{N}$,*

$$\mathsf{K}(x \mid t) \leq \mathsf{pK}^t(x) + c \log |x|.$$

▶ **Proposition 14.** *There is a constant c' such that, for any string $x \in \{0,1\}^*$ and time bound $t \in \mathbb{N}$,*

$$\mathsf{K}^t(x) \leq |x| + c'.$$

▶ **Lemma 15** (Symmetry of Information for Time-unbounded K-complexity [31]). *For every pair of strings $x \in \{0,1\}^*$ and $y \in \{0,1\}^*$,*

$$\mathsf{K}(xy) \geq \mathsf{K}(x) + \mathsf{K}(y \mid x) - O(\log |xy|).$$

2.3 Agnostic PAC-Learning and Correlative RRHS-Refutation

In the PAC-learning framework, one is asked to learn an unknown *concept*: namely, a Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ for some $n \in \mathbb{N}$. A *concept class* $\mathcal{C}$ refers to a set of such concepts, and $\mathcal{C}_n$ denotes $\mathcal{C} \cap \{f\colon \{0,1\}^n \to \{0,1\}\}$. One may ask whether $\mathcal{C}$ is PAC-learnable over a class $\mathcal{D}$ of ensembles $D = \{D_n\}_{n \geq 1}$ of distributions D_n. $\mathcal{D}_n$ denotes $\{D_n \mid D \in \mathcal{D}\}$. For a *hypothesis* $h\colon \{0,1\}^n \to \{0,1\}$, define

$$\mathsf{err}_{D_n}(h, f) = \Pr_{x \sim D_n} [h(x) \neq f(x)].$$

We also define the *minimum relative distance* between f and $\mathcal{C}$ with respect to D_n as the disagreement between f and the best-fitting hypothesis $c \in \mathcal{C}$, i.e.,

$$\mathsf{opt}_{\mathcal{C}_n, D_n, f} \;=\; \min_{c \,\in\, \mathcal{C}_n} \mathsf{err}_{D_n}(c, f).$$

Learners are provided an *example oracle* $\mathsf{EX}(f, D_n)$ such that each query returns an independently sampled pair (x, b), where $x \sim D_n$ and $b = f(x)$. We will use the term *sample complexity* to mean the number of queries made to $\mathsf{EX}(f, D_n)$.

▶ **Definition 16** (PAC learning [29]). *Let $\mathcal{C}$ be a concept class, and let D be a class of distributions. We say that $\mathcal{C}$ is PAC-learnable on $\mathcal{D}$ if there is an algorithm A with the following property. For every $n \geq 1$, $\varepsilon > 0$, $\delta > 0$, distribution $D \in \mathcal{D}_n$, and concept $f\colon \{0,1\}^n \to \{0,1\}$ belonging to $\mathcal{C}_n$,*

$$\Pr_{A,\,\mathsf{EX}(f,D)}\left[A^{\mathsf{EX}(f,D)}(n,\varepsilon,\delta) \text{ outputs a hypothesis } h \text{ such that } \mathsf{err}_D(h,f) \leq \varepsilon\right] \geq 1-\delta,$$

where the probability is over the internal randomness of A and random examples provided by $\mathsf{EX}(f,D)$.

The following definition of *agnostic* PAC learning is a generalization of the PAC learning definition above to the case where a function f to be learned is not necessarily from the concept class $\mathcal{C}$. In this case, the hypothesis h output by the learning algorithm should have an error close to the minimum relative distance between f and the concept class $\mathcal{C}$.

▶ **Definition 17** (Agnostic PAC learning [21]). *Let $\mathcal{C}$ be a concept class, and let $\mathcal{D}$ be a class of distributions. We say that $\mathcal{C}$ is ε-agnostic PAC-learnable on $\mathcal{D}$ if there is an algorithm A with the following property. For every $n \geq 1$, $\varepsilon > 0$, $\delta > 0$, distribution $D \in \mathcal{D}_n$, and concept $f\colon \{0,1\}^n \to \{0,1\}$,*

$$\Pr_{A,\,\mathsf{EX}(f,D)}\left[A^{\mathsf{EX}(f,D)}(n,\varepsilon,\delta) \text{ outputs a hypothesis } h \text{ such that } \mathsf{err}_D(h,f) \leq \mathsf{opt}_{\mathcal{C}_n,D,f}+\varepsilon\right] \geq 1-\delta.$$

▶ **Definition 18** (Correlative RRHS-Refutation). *Let $\mathcal{C}$ be a concept class, and let $D = \{D_n\}_{n\geq 1}$ be an ensemble of distributions. A randomized algorithm R is a* ε-correlative random-right-hand-side-refuter (ε-correlative RRHS-refuter) *for $\mathcal{C}$ on D with sample complexity m provided it satisfies the following. Given input parameters $n \in \mathbb{N}$ and $\varepsilon \in (0,1)$, as well as a set*

$$S = \left(\left\langle x^{(1)}, b^{(1)}\right\rangle, \ldots, \left\langle x^{(m)}, b^{(m)}\right\rangle\right)$$

of samples, where $x^{(i)} \in \{0,1\}^n$ and $b^{(i)} \in \{0,1\}$ for $i \in [m]$;

- ***Soundness:*** *Suppose the samples S are i.i.d. from a distribution D' on $\{0,1\}^n \times \{0,1\}$ such that the marginal on $\{0,1\}^n$ equals D_n, and for some $f \in \mathcal{C}_n$,*

  $$\Pr_{\left\langle x^{(i)}, b^{(i)}\right\rangle \sim D'}\left[b^{(i)} = f(x^{(i)})\right] \geq \frac{1}{2} + \frac{\varepsilon}{2}.$$

  Then,

  $$\Pr_{S,R}\left[R(n,\varepsilon,S) = \textit{correlative}\right] \geq 2/3.$$

- ***Completeness:*** *Suppose the samples S are i.i.d. with $x^{(1)}, \ldots, x^{(m)} \sim D_n$ and $b^{(1)}, \ldots, b^{(m)} \sim \mathcal{U}$. Then,*

  $$\Pr_{S,R}\left[R(n,\varepsilon,S) = \textit{random}\right] \geq 2/3.$$

Kothari and Livni [22] prove an equivalence between distribution-specific agnostic PAC learning and RRHS-refutation. We will be using the following direction from RRHS-refutation to agnostic learning.

▶ **Theorem 19** (Agnostic Learning from RRHS-Refutation [22]). *Let $\mathcal{C}$ be a concept class, and let $D = \{D_n\}_{n\geq 1}$ be an ensemble of distributions. If there exists an ε-correlative RRHS-refuter for $\mathcal{C}$ on D with sample complexity $m(n,\varepsilon)$ and running time $T(n,\varepsilon)$, then $\mathcal{C}$ is (2ε)-agnostic PAC-learnable over D with*

- *sample complexity $O\big(m(n,\varepsilon/2)^3 \cdot \varepsilon^{-2}\big)$, and*
- *running time $O\big(T(n,\varepsilon/2) \cdot m(n,\varepsilon/2)^2 \cdot \varepsilon^{-2}\big)$.*

The proof of the above theorem relies on *distribution-specific* boosting algorithms for the agnostic setting, such as those of Feldman [12] and Kalai and Kanade [20]. These algorithms transform a weak agnostic learner over some distribution into a strong agnostic learner over that same distribution; they work by adaptively modifying the labels of example points rather than the distributions on those points as is typically the case in boosting. Interestingly, in the agnostic setting, it is possible to accomplish this without a superpolynomial increase in the running time of the learner.

2.4 Inversion

In this section, we cover definitions of *inversion* of functions, which are the negations of corresponding definitions of the existence of one-way functions. Throughout, we take the word "function" to include *auxiliary input* functions in the sense of Ostrovsky and Wigderson, in which both function and potential inverter have access to the same non-uniform input (denoted y below) [26].

▶ **Definition 20** (Invertible functions). *Consider a function $g(y,x)$ computable uniformly in polynomial time. The function g is said to be* weakly invertible *if there is a probabilistic polynomial-time Turing machine I and a constant b such that for every $n \in \mathbb{N}$ and for every $y \in \{0,1\}^*$,*

$$\Pr_{x\sim\mathcal{U}_n}[g(y,I(y,g(y,x))) = g(y,x)] \geq \frac{1}{n^b}.$$

The function g is said to be strongly invertible *if for every constant d there is a probabilistic polynomial-time Turing machine I such that for every $n \in \mathbb{N}$ and for every $y \in \{0,1\}^*$,*

$$\Pr_{x\sim\mathcal{U}_n}[g(y,I(y,g(y,x))) = g(y,x)] \geq 1 - \frac{1}{n^d}.$$

▶ **Definition 21** (Statistical Indistinguishability). *Two probability distributions D and D' are* statistically indistinguishable within δ *if for all $T \subseteq \{0,1\}^n$,*

$$\left|\Pr_{x\sim D_n}[x \in T] - \Pr_{x\sim D'_n}[x \in T]\right| \leq \delta.$$

We denote this as $D \equiv_\delta D'$.

▶ **Definition 22** (Distributionally invertible functions). *Consider a function $g(y,x)$ computable uniformly in polynomial time. The function g is said to be* distributionally invertible *if for every constant $b > 0$ there is a probabilistic polynomial-time oracle Turing Machine I such that for every $n \in \mathbb{N}$ and $y \in \{0,1\}^*$,*

$$(x, g(y,x)) \equiv_{n^{-b}} (I(y,g(y,x)), g(y,x)),$$

*where $x \sim \mathcal{U}_n$. We refer to the machine I as an n^{-b}-*distributional inverter*.*

▶ **Lemma 23** ([30]). *If every function computable in polynomial time is weakly invertible, then every such function is strongly invertible.*

▶ **Lemma 24** ([19]). *If every function computable in polynomial time is strongly invertible, then every such function is distributionally invertible.*

▶ **Lemma 25** ([1]). *If $(\mathsf{MKTP},\mathcal{U}) \in \mathsf{AvgBPP}$, then every function computable in polynomial time is weakly invertible.*

▶ **Corollary 26.** *If $(\mathsf{MKTP},\mathcal{U}) \in \mathsf{AvgBPP}$, then every function computable in polynomial time is distributionally invertible.*

2.5 Direct Product Generator and pK^t-Compression

▶ **Definition 27** (Direct Product Generator). *For $n, k \in \mathbb{N}$, the k-wise direct product generator $\mathsf{DP}_k : \{0,1\}^n \times \{0,1\}^{nk} \to \{0,1\}^{nk+k}$ is the function defined by*

$$\mathsf{DP}_k(x; z^1, ..., z^k) = (z^1, ..., z^k; \langle x, z^1\rangle, ..., \langle x, z^k\rangle),$$

where $\langle -, -\rangle$ denotes the inner product $\langle x, y\rangle = \left(\sum_{i=1}^{|x|} x_i y_i\right) \mod 2$.

▶ **Lemma 28** (Probabilistic pK^t Reconstruction [13]). *There is a polynomial p' with the following property. For $\varepsilon > 0$, $x \in \{0,1\}^n$, $s \in \mathbb{N}$, and $k \in \mathbb{N}$ satisfying $k \le 2n$, let D be a randomized algorithm that takes an advice string β, runs in time t_D, and ε-distinguishes $\mathsf{DP}_k(x; \mathcal{U}_{nk})$ from $\mathcal{U}_{nk+k}$. Then*

$$\mathsf{pK}^{p'(t_D \cdot n/\varepsilon)}(x \mid \beta) \le k + \log p'(t_D \cdot n/\varepsilon).$$

2.6 Source Coding Theorem

The following lemma is very similar to one of [3], but with a greater probability of success on the right-hand side, which is necessary for the application in Lemma 37. For completeness, we present a slight modification of a proof due to [2], which uses hashing.

▶ **Lemma 29.** *The following holds unconditionally. There exist polynomials p and q such that for any $T, a : \mathbb{N} \to \mathbb{N}$, $n \in \mathbb{N}$, $D \in \mathsf{Samp}[T(n)]/a(n)$, and $x \in \mathsf{Supp}(D_n)$,*

$$\Pr_{r \sim \mathcal{U}_{3T(n)}}\left[\mathsf{K}^{p(T(n))}(x, r) \le \log(1/D_n(x)) + |r| + a(n) + \log p(T(n))\right] \ge 1 - \frac{1}{4T(n)},$$

where $D_n(x)$ denotes the probability of x under D_n.

Proof. Let A be a non-uniform algorithm sampling $D \in \mathsf{Samp}[T(n)]/a(n)$. That is, there is some $\alpha \in \{0,1\}^{a(n)}$ such that for any $x \in \mathrm{supp}(D_n)$,

$$\Pr_{w \sim \mathcal{U}_{T(n)}}[A(w; \alpha, 1^n)] = D_n(x).$$

Let s be the smallest integer such that $D_n(x) \ge 2^{-s}$. Define $\ell := T(n)$ and $k := \ell - s - \log(8T(n))$. Consider a universal hash function family $\mathcal{H} = \{h : \{0,1\}^\ell \to \{0,1\}^k\}$. For each $h \in \mathcal{H}$ and $w \in \{0,1\}^{T(n)}$, $h(w) = U \cdot w + v$ for some binary Toeplitz matrix U of dimension $k \times \ell$ and binary vector v of dimension k. Define a set

$$S_x := \{w \in \{0,1\}^{T(n)} \mid A(w; \alpha, 1^n) = x\}.$$

For $h \sim \mathcal{H}$, define a random variable $X := |S_x \cap h^{-1}(0^k)|$. Note that $|S_x| = D_n(x) \cdot 2^{T(n)} \ge 2^{\ell - s}$. Then $|S_x|/2^k \ge 8T(n)$, and by universality,

$$\mathsf{Var}[X] \le \mathbb{E}[X] = \frac{|S_x|}{2^k}.$$

By Chebyshev's Inequality,

$$\begin{aligned}\Pr[X = 0] &\le \Pr\left[|X - \mathbb{E}[X]| \ge \mathbb{E}[X]\right] \\ &\le \mathsf{Var}[X]/\mathbb{E}[X]^2 \\ &\le 1/8T(n).\end{aligned}$$

Now define a random variable $Y = |h^{-1}(0^k)|$, where $h \sim \mathcal{H}$. Note that $\mathbb{E}[Y] = 2^\ell/2^k = 2^{s+\log(8T(n))}$. Then by Markov's Inequality,

$$\begin{aligned}\Pr[Y \geq 2^{s+2\log(8T(n))}] &= \Pr[Y \geq 8T(n) \cdot \mathbb{E}[Y]] \\ &\leq 1/8T(n).\end{aligned}$$

By a union bound,

$$\Pr[X = 0 \text{ or } Y \geq 2^{s+2\log(8T(n))}] \leq 1/4T(n).$$

Assume $X > 0$ and $Y < 2^{s+2\log(8T(n))}$. It is possible to represent x with descriptions of the hash function h, the index of a string $w \in S_x$ in the set $h^{-1}(0^k)$, and the advice string α used in the sampler A. In particular, x may be recovered by performing Gaussian elimination to compute the set $h^{-1}(0^k)$ from the description (U, v) of h, locating w in this set, and then returning the output of $A(w; \alpha, 1^n)$. This requires $|(U, v)| < 3T(n)$ bits to describe h, at most $\log Y \leq s + 2\log(8T(n)) \leq \log(1/D_n(x)) + 1 + 2\log(8T(n))$ bits to describe the position of w in $h^{-1}(0^k)$, and $|\alpha| = a(n)$ bits to run the sampler A. Define the "random" string $r \in \{0,1\}^{3T(n)}$ as the description (U, v) of h. Overall, we have that with probability at least $1 - 1/4T(n)$ over r sampled uniformly,

$$\mathsf{K}^{p(T(n))}(x, r) \leq \log(1/D_n(x)) + |r| + a(n) + \log p(T(n))$$

for some polynomial p. ◀

3 Approximating K^t

▶ **Lemma 30** (implicit in [15]). *If* $(\mathsf{MK^tP}, \mathcal{U}) \in \mathsf{AvgBPP}$*, then there is a polynomial* p *such that the following promise problem is in* promiseBPP*:*

$$\Pi_{\mathsf{YES}} := \left\{ (x, 1^s, 1^t) \mid x \in \{0,1\}^*,\ s, t \in \mathbb{N},\ t \geq |x|,\ \textit{and}\ \mathsf{K}^t(x) \leq s \right\},$$
$$\Pi_{\mathsf{NO}} := \left\{ (x, 1^s, 1^t) \mid x \in \{0,1\}^*,\ s, t \in \mathbb{N},\ t \geq |x|,\ \textit{and}\ \mathsf{pK}^{p(t)}(x) > s + \log p(t) \right\}.$$

Proof. Let the input $(x, 1^s, 1^t)$ be given, where $x \in \{0,1\}^n$ and $s \leq n + O(1)$. Define

$$k := s + 2\log q(t), \text{ and}$$
$$s' := s + nk + \log q(t),$$

where q is a polynomial chosen later.

Let B_0 be a randomized errorless heuristic scheme for $(\mathsf{MK^tP}, \mathcal{U})$, with failure probability $1/n$. Let B be the modification of B_0 that outputs "1" whenever B_0 would output "$\perp$". Note that on yes-instances of $\mathsf{MK^tP}$, B errs with probability at most $1/10$ over its own internal randomness.

Define another algorithm B' as follows:

On input $(x, 1^s, 1^t)$, sample $z \sim \mathcal{U}_{nk}$ and then output $B(\mathsf{DP}_k(x; z), 1^{s'}, 1^{q(t)})$.

In the remainder of the proof, we argue that B' solves $(\Pi_{\mathsf{YES}}, \Pi_{\mathsf{NO}})$ correctly with high probability in the worst case.

First, suppose $(x, 1^s, 1^t) \in \Pi_{\mathsf{YES}}$. Observe that for our choice of k, given any $x \in \{0,1\}^n$ and $z \in \{0,1\}^{nk+k}$, it is possible to compute $\mathsf{DP}_k(x;z)$ in polynomial time. Thus, we let q be a polynomial such that for any $z \in \{0,1\}^{nk}$ and sufficiently large $t \in \mathbb{N}$,

$$\begin{aligned}\mathsf{K}^{q(t)}(\mathsf{DP}_k(x;z)) &\leq \mathsf{K}^t(x) + |z| + \log q(t)\\ &\leq s + |z| + \log q(t)\\ &= s'.\end{aligned}$$

Then by definition of B, for $(x, 1^s, 1^t) \in \Pi_{\mathsf{YES}}$,

$$\Pr[B'(x, 1^s, 1^t) = 1] \geq 9/10,$$

where the above probability is over the inner randomness of B and $z \sim \mathcal{U}_{nk}$.

Now suppose $(x, 1^s, 1^t) \in \Pi_{\mathsf{NO}}$. For a contradiction, suppose

$$\Pr[B'(x, 1^s, 1^t) = 1] = \Pr_{B,z}[B(\mathsf{DP}_k(x;z), 1^{s'}, 1^{q(t)}) = 1] > 1/4. \tag{5}$$

By a counting argument, for randomly selected $w \sim \mathcal{U}_{nk+k}$,

$$\Pr_w\left[\mathsf{K}^{q(t)}(w) \leq s'\right] \leq \frac{2^{s'}}{2^{nk+k}} = \frac{1}{q(t)}.$$

Then by definition of B,

$$\begin{aligned}\Pr_{B,w}\left[B(w, 1^{s'}, 1^{q(t)}) = 1\right] &= \frac{1}{10} + \frac{1}{n} + \frac{1}{q(t)}\\ &< 1/8.\end{aligned} \tag{6}$$

Comparing Equations (5) and (6), we see that $B(-, 1^{s'}, 1^{q(t)})$ (1/8)-distinguishes $\mathsf{DP}_k(x;\mathcal{U}_{nk})$ from $\mathcal{U}_{nk+k}$. Then by Lemma 28, for some polynomial p',

$$\begin{aligned}\mathsf{pK}^{p'(t)}(x) &\leq k + O(\log t)\\ &= s + O(\log t).\end{aligned}$$

In other words, for an appropriate choice of the polynomial p in the statement of the lemma, $(x, 1^s, 1^t)$ is *not* in Π_{NO}. This gives a contradiction. We conclude that for $(x, 1^s, 1^t) \in \Pi_{\mathsf{NO}}$,

$$\Pr\left[B'(x, 1^s, 1^t) = 1\right] \leq 1/4. \qquad \blacktriangleleft$$

▶ **Lemma 31** ([15]). *If $(\mathsf{MK^tP}, \mathcal{U}) \in \mathsf{AvgBPP}$, then there exists a polynomial p and a randomized algorithm A that on input $(x, 1^t)$, where $x \in \{0,1\}^n$ and $t \in \mathbb{N}$, runs in time $\mathsf{poly}(n,t)$ and with probability at least $1 - 2^{-n}$ outputs an integer $\widetilde{s}$ such that*

$$\mathsf{pK}^{t^c}(x) - \log p(t) \;\leq\; \widetilde{s} \;\leq\; \mathsf{K}^t(x).$$

Proof. Consider the polynomial-time randomized algorithm B' that solves the promise problem from Lemma 30. By standard success amplification, we may assume that the error of B' is at most 2^{-2n} on inputs satisfying the promise. Algorithm A runs B' on $(x, 1^s, 1^t)$ for $s = 1, 2, \ldots, n + \log n$, and outputs the first $\widetilde{s}$ such that $B'(x, 1^{\widetilde{s}}, 1^t) = 1$. If B' never accepts, A simply outputs $n + \log n$.

On one hand, if $s = \mathsf{K}^t(x)$, then $(x, 1^s, 1^t) \in \Pi_{\mathsf{YES}}$, so $\Pr[B'(x, 1^s, 1^t) = 1] \geq 1 - 2^{-2n}$. On the other, if $s < \mathsf{pK}^{p(t)}(x) - \log p(t)$, then $(x, 1^s, 1^t) \in \Pi_{\mathsf{NO}}$, so $\Pr[B'(x, 1^s, 1^t) = 1] \leq 2^{-2n}$.

By a union bound, with probability at least $1 - 2^{-n}$, $\widetilde{s}$ has the desired property. $\blacktriangleleft$

4 Agnostic Learning from Heuristics for K-complexity

In what follows, for a distribution D and $m \in \mathbb{N}$, D^m will denote the distribution $(x^{(1)}, ..., x^{(m)})$ where $x^{(i)} \sim D$ for $i \in [m]$. Moreover, $\ell_s(n) \leq O(s(n) \log s(n))$ will denote the number of bits needed to encode a function $f \in \mathsf{SIZE}[s(n)]$.

▶ **Lemma 32** ([16]). *There exists a polynomial t' such that for any $m \geq n \in \mathbb{N}$, string $b \in \{0,1\}^m$, function $f : \{0,1\}^n \to \{0,1\}$, $X = (x^{(1)}, ..., x^{(m)}) \in (\{0,1\}^n)^m$, and $\delta \in (0,1)$ satisfying*

$$\left|\{i \in [m] \mid b_i = f(x^{(i)})\}\right| \geq (1/2 + \delta) \cdot m,$$

we have that for any $r \in \{0,1\}^*$,

$$\mathsf{K}^{t'(m)}(b \mid X, r) \leq \ell_s(n) + \left(1 - 2\delta^2\right) \cdot m.$$

Proof. Given X, we can compute $f(x^{(1)}), \ldots, f(x^{(m)})$ in time $\mathsf{poly}(m \cdot \ell_s(n))$ using the encoding of f, which requires $\ell_s(n)$ bits. Note that b and $f(x^{(1)}), \ldots, f(x^{(m)})$ disagree on at most $(1/2 - \delta) \cdot m$ coordinates. So to recover b, it suffices to encode the string $e \in \{0,1\}^m$ such that $e_i = 1$ iff $f(x^{(i)}) \neq b_i$. We will show that $\mathsf{K}^{\mathsf{poly}(m)}(e) \leq (1 - 2\delta^2) \cdot m$, which will conclude the proof of the lemma.

Note that e has hamming weight at most $m' = (1/2 - \delta) \cdot m$. Every m'-size subset of an m-size set can be represented using $\log_2 \binom{m}{m'}$ bits, via the combinatorial number system, with both encoding and decoding algorithms running in time polynomial in m (see, e.g., [14] for details). Using standard inequalities for binomial coefficients and the binary entropy function H_2, we get

$$\begin{aligned}\log_2 \binom{m}{m'} &\leq \log_2 2^{H_2(m'/m) \cdot m} \\ &= H_2\left(1/2 - \delta\right) \cdot m \\ &\leq \left(1 - 2\delta^2\right) \cdot m,\end{aligned}$$

as required. ◀

We will also need a lemma similar to the above for the case of KT: that is, bounding the KT-complexity of the labels b in the case that they correlate with a function f. Lemma 32 is insufficient as-is, since the time bound $t'(m)$ would render $\mathsf{KT}(b)$ trivial. To overcome this issue, we use an encoding scheme from Golovnev et al. for strings of bounded hamming weight.

▶ **Lemma 33** ([14]). *For some $m, m' \in \mathbb{N}$ and $e \in \{0,1\}^m$, suppose e has hamming weight at most m'. Then there is a string e' of length at most $\log \binom{m}{m'} + m^{3/4}$ such that for all $1 \leq i \leq m$, e_i can be computed with random access to e' in time $m^{2/3}$.*

▶ **Lemma 34.** *For any $m, n \in \mathbb{N}$, string $b \in \{0,1\}^m$, function $f : \{0,1\}^n \to \{0,1\}$, $X = (x^{(1)}, ..., x^{(m)}) \in (\{0,1\}^n)^m$, $r \in \{0,1\}^*$, and $\delta \in (0,1)$ satisfying*

$$\left|\{i \in [m] \mid b_i = f(x^{(i)})\}\right| \geq (1/2 + \delta) \cdot m,$$

we have that

$$\mathsf{KT}(X, b, r) \leq \mathsf{KT}(X, f(x^{(1)}), ..., f(x^{(m)}), r) + \left(1 - 2\delta^2\right) \cdot m + 2m^{3/4}.$$

Proof. It is clear that any bit of X or r can be computed in time and description size upper-bounded by $\mathsf{KT}(X, f(x^{(1)}), ..., f(x^{(m)}), r)$. To compute a bit b_i of b, for $i \in [m]$, we observe the following. As in Lemma 32, let $e \in \{0,1\}^m$ be such that $e_i = 1$ iff $f(x^{(i)}) \neq b_i$. Then b_i is $f(x^{(i)}) \oplus e_i$. Note that the the hamming weight of e is at most $m' := (1/2 - \delta) \cdot m$. Applying Lemma 33, e_i may be computed in time at most $m^{2/3}$ from a description e' of length at most

$$\log \binom{m}{m'} + m^{3/4}.$$

Arguing as in Lemma 32, we upper-bound the above by $(1 - 2\delta^2) \cdot m + m^{3/4}$.

To compute a bit b_i, we first use time and description size $\mathsf{KT}(X, f(x^{(1)}), ..., f(x^{(m)}), r)$ to obtain the corresponding $f(x^{(i)})$. Then, given $f(x^{(i)})$, b_i may be computed in time at most $m^{2/3} + O(1)$ from a description of e' of size at most $\left(1 - 2\delta^2\right) \cdot m + m^{3/4}$. This concludes the proof. ◀

4.1 Learning over the Uniform Distribution from MKTP

Here, we construct a correlative RRHS-refuter, working over distributions that are statistically close to uniform, under the assumption that MKTP is easy on average. In the next section, we will reduce the case of arbitrary efficiently samplable distributions to this case.

▶ **Theorem 35.** *If* $(\mathsf{MKTP}, \mathcal{U}) \in \mathsf{AvgBPP}$*, then for any time-constructible function* $s\colon \mathbb{N} \to \mathbb{N}$*, constants* $c, \zeta > 0$*, and any family of distributions* D *such that* $D_n \equiv_{n^{-c}} \mathcal{U}_n$*, there is an* ε*-correlative RRHS-refuter for* $\mathsf{SIZE}[s(n)]$ *under* D_n *taking parameters* $n \in \mathbb{N}$ *and* $\varepsilon \in (0,1)$ *with sample complexity*

$$m(n, \varepsilon) := \left(\frac{s(n) + n}{\varepsilon^8}\right)^{1+\zeta}$$

and running time $\mathsf{poly}\left(n, \varepsilon^{-1}, s(n)\right)$.

Proof. Let A_0 be a randomized errorless heuristic scheme for $(\mathsf{MKTP}, \mathcal{U})$ with failure probability $1/n$. Let A be the algorithm that simulates A_0 and outputs "correlative" whenever it would output "1" or "$\perp$", and "random" whenever it would output "0". Note that on yes-instances of MKTP, A errs with probability at most $1/10$ over its own internal randomness.

The (correlative) RRHS-refuter R. On input $n \in \mathbb{N}$, $\varepsilon > 0$, and a set

$$S = \left(\left\langle x^{(1)}, b^{(1)} \right\rangle, \ldots, \left\langle x^{(m)}, b^{(m)} \right\rangle\right)$$

of samples, let $X := (x^{(1)}, \ldots, x^{(m)})$ and $b := (b^{(1)}, \ldots, b^{(m)})$. R is defined as follows.

1. Compute $\theta := mn + (1 - \varepsilon^2/16) \cdot m$.
2. Evaluate $A((X, b), 1^\theta)$. Output "correlative" if A accepts, and output "random" otherwise.

Correlative Case (Soundness). Suppose the labeled examples in S are sampled i.i.d from some distribution D' on $\{0,1\}^n \times \{0,1\}$, whose marginal on $\{0,1\}^n$ is given by D_n, and there exists $f \in \mathsf{SIZE}[s(n)]$ such that

$$\Pr_{\left\langle x^{(i)}, b^{(i)} \right\rangle \sim D'} \left[b^{(i)} = f(x^{(i)})\right] \geq \frac{1}{2} + \frac{\varepsilon}{2}.$$

In this case, by a Chernoff bound, the probability over $S \sim (D')^m$ that

$$\left|\{i \in [m] \mid b_i = f(x^{(i)})\}\right| < (1/2 + \varepsilon/4) \cdot m$$

is at most $\exp(-2m(\varepsilon/4)^2) \leq o(1)$. So with probability $1 - o(1)$, the conditions of Lemma 34 are met. Now observe that

$$\mathsf{KT}(X, f(x^{(1)}), ..., f(x^{(m)})) \leq mn + 2\ell_s(n) + 2n,$$

using an mn-bit description of X to obtain any bit of X in constant time, along with an $\ell_s(n)$-bit description of a circuit computing f to obtain any bit $f(x^{(i)})$ from X in time at most $\ell_s(n) + 2n$. This, along with Lemma 34 (with r the empty string), implies that

$$\mathsf{KT}(X, b) \leq mn + \left(1 - \varepsilon^2/8\right) \cdot m + 2\ell_s(n) + 2n + 2m^{3/4}. \tag{7}$$

Finally, by our choice of $m = \omega\left((s(n) + n) \cdot \varepsilon^{-8}\right)$,

$$m > \frac{32\left(\ell_s(n) + n + m^{3/4}\right)}{\varepsilon^2};$$

re-written and combined with Eq. (7),

$$\begin{aligned}\mathsf{KT}(X, b) &\leq mn + \left(1 - \varepsilon^2/8\right) \cdot m + 2\ell_s(n) + 2n + 2m^{3/4} \\ &< mn + \left(1 - \varepsilon^2/16\right) \cdot m \\ &= \theta.\end{aligned}$$

By definition of A, R will output "correlative" with probability at least $9/10 - o(1) > 2/3$.

Random Case (Completeness). Suppose (X, b) is sampled from the distribution $(D_n^m, \mathcal{U}_m)$. Note that for $X \sim \mathcal{U}_n^m$ and $b \sim \mathcal{U}_m$, it holds that

$$\Pr_{X,b}[\mathsf{KT}(X, b) > mn + m - 10] \geq 9/10.$$

Then by the definition of statistical distance, with probability at least $9/10 - o(1)$ over $X \sim D_n^m$ and $b \sim \mathcal{U}_m$,

$$\begin{aligned}\mathsf{KT}(X, b) &> mn + m - 10 \\ &> \theta.\end{aligned}$$

In other words, $((X, b), 1^\theta) \notin \mathsf{MKTP}$.

Now, since the failure probability of our heuristic A is at most $1/10 + 1/n$ over the uniform distribution, the definition of statistical distance implies that its failure probability is at most $1/10 + o(1)$ over the distribution $(D_n^m, \mathcal{U}_m)$.

Overall, by a union bound, R outputs "random" with probability at least $4/5 - o(1) > 2/3$. ◀

4.2 Learning over PSAMP/poly from MKTP

In this section, we generalize the previous theorem to give correlative RRHS-refuters working over arbitrary efficiently samplable distributions. In particular, we reduce to the case of a nearly-uniform distribution by inverting the circuit that samples our given target distribution.

This requires *distributional inversion* as defined by Impagliazzo and Luby [19], which is possible under the assumption of MKTP being easy on average.[5]

▶ **Theorem 36.** *Suppose* $(\mathsf{MKTP}, \mathcal{U}) \in \mathsf{AvgBPP}$. *Consider any time-constructible function* $s : \mathbb{N} \to \mathbb{N}$, *polynomials* $T, a : \mathbb{N} \to \mathbb{N}$, *constant* $\zeta > 0$, *and* $\varepsilon \in (n^{-d}, 1)$ *for a constant* $d > 0$. *Let* $D = \{D_n\}_{n\in\mathbb{N}}$ *be a family of distributions such that each* D_n *is samplable in time* $T(n)$ *with* $a(n)$ *bits of non-uniform advice* α_n. *There is an algorithm which, given* α_n *and parameters* $n \in \mathbb{N}$ *and* ε, *is an* ε*-correlative RRHS-refuter for* $\mathsf{SIZE}[s(n)]$ *under* D_n. *This RRHS-refuter has sample complexity*

$$m(n, \varepsilon) := \left(\frac{s(n) + n}{\varepsilon^8}\right)^{1+\zeta}$$

and running time $\mathsf{poly}\big(n, T(n), a(n), s(n), \varepsilon^{-1}\big)$.

Proof. By Corollary 26, every function $g(y, x)$ computable in polynomial time is distributionally invertible. In particular, let I be a $\varepsilon/4$-distributional inverter for the function g that evaluates the Boolean circuit y on the input string x. Let $\{C_n\}_{n\in\mathbb{N}}$ be the family of circuits that sample D. In particular, each C_n applies the $T(n)$-time sampler for D_n along with the advice α_n. By the definition of distributional inversion (Definition 22), we have that for all sufficiently large $n \in \mathbb{N}$,

$$(I(C_n, C_n(w)), C_n(w)) \equiv_{\varepsilon/4} (w, C_n(w)), \tag{8}$$

where $w \sim \mathcal{U}_\ell$, $\ell \le a(n)$, and I runs in time $\mathsf{poly}(T(n), a(n))$.

Given labeled samples of the form (x, b), where $x \sim D_n = C_n(\mathcal{U}_\ell)$, one may apply I to the first part to simulate labeled samples of the form (r', b), where $r' \in \{0,1\}^\ell$. Specifically, $r' \sim D'_\ell$, where D'_ℓ is the distribution $I(C_n, C_n(\mathcal{U}_\ell))$ sampled by the circuit $C'_\ell(-) := I(C_n, C_n(-))$. By Eq. (8), $D'_\ell \equiv_{\varepsilon/4} \mathcal{U}_\ell$.

We will reduce to the case of a nearly-uniform distribution: namely, the case of Theorem 35. Consider a target function f computable in $\mathsf{SIZE}[s(n)]$. By Theorem 35, since D' is statistically close to uniform, there is a correlative RRHS-refuter R' for $f \circ C_n$ over D' with parameter $\varepsilon' := \varepsilon/2$ that has sample complexity $m = \big((s(n) + n)/\varepsilon^8\big)^{1+\zeta}$ and running time $\mathsf{poly}(n, s(n), \varepsilon^{-1})$. To get a correlative RRHS-refuter R for f over D, we simply return the output of this R' on the simulated examples (r', b). Note that R takes time $\mathsf{poly}(n, T(n), a(n), s(n), \varepsilon^{-1})$ overall.

We now argue that in the "random" case of the original problem, R will output "random" with high probability, and likewise for the "correlative" case. In the random case, the labels b are simply sampled from the uniform distribution $\mathcal{U}$, so R will output "random" with probability at least $2/3$, by the correctness of R'. In the correlative case, b is such that

$$\Pr_{x\sim D_n}[b = f(x)] \ge \frac{1}{2} + \frac{\varepsilon}{2}. \tag{9}$$

We would now like to show that the above probability is not too much smaller when x is sampled from $C_n(D'_\ell)$ rather than $D_n = C_n(\mathcal{U}_\ell)$. Define a set

$$T := \{(r, x) \mid x = C_n(r)\}$$

[5] Similar ideas are employed in the work of Binnendyk et al. [4], which shows that PAC-learning with membership queries over arbitrary efficiently samplable distributions is possible under the existence of natural properties.

and note that samples from the distribution $(r, C_n(r))$, for $r \sim \mathcal{U}_\ell$, belong to T with probability 1. By the property of distributional inversion, ie. Eq. (8), samples from the distribution $(I(C_n, C_n(r)), C_n(r)) = (C'_\ell(r), C_n(r))$, for $r \sim \mathcal{U}_\ell$, belong to the set T with probability at least $1 - \varepsilon/4$. Whenever this holds, by definition of T, we have that $C_n(C'_\ell(r)) = C_n(r)$. Particularly, $f(C_n(r')) = f(x)$, for $r' = C'_\ell(r)$ and $x = C_n(r)$. Then by a union bound with Eq. (9), in the correlative case of the original problem,

$$\Pr_{r' \sim D'_\ell}[b = f(C_n(r'))] \geq \frac{1}{2} + \frac{\varepsilon}{2} - \frac{\varepsilon}{4} = \frac{1}{2} + \frac{\varepsilon'}{2}.$$

Thus, R will output "correlative" with probability at least $2/3$, by the correctness of R'. This completes the proof of the theorem. ◀

4.3 Learning from MK$^{\mathsf{t}}$P

The following lemma is similar to one from [16], but accounts for a uniformly random string $r \sim \mathcal{U}_{3mT(n)}$, which is essential given Lemma 29. This lemma states that in the expectation, over an efficiently samplable distribution (along with the uniformly random string r), the K^t-complexity of a string is close to its time-unbounded K-complexity. Note that the lemma from [16] holds under the assumption that $\mathsf{DistNP} \subseteq \mathsf{AvgP}$ whereas this one holds unconditionally.

▶ **Lemma 37.** *There exists a polynomial $p_1 \colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ such that for any $T, a \colon \mathbb{N} \to \mathbb{N}$ and $n, m \in \mathbb{N}$, the following holds unconditionally. Let $D_n \in \mathsf{Samp}[T(n)]/a(n)$. For every $t \geq p_1(T(n), m)$, $X \sim D_n^m$, and $r \sim \mathcal{U}_{3mT(n)}$,*

$$\mathop{\mathbb{E}}_{X,r}[\mathsf{K}^t(X, r) - \mathsf{K}(X, r)] \leq a(n) + O(\log m + \log T(n)).$$

Proof. Let p_1 be the polynomial p in Lemma 29. Note that for $D_n \in \mathsf{Samp}[T(n)]/a(n)$, we have $D_n^m \in \mathsf{Samp}[m \cdot T(n)]/a(n)$. For every $t \geq p_1(T(n), m)$, for $X \sim D_n^m$ and $r \sim \mathcal{U}_{3mT(n)}$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X,r}\left[\mathsf{K}^t(X, r)\right] &\leq \mathop{\mathbb{E}}_{X,r}\left[\mathsf{K}^{p_1(T(n),m)}(X, r)\right] \\
&\leq \frac{1}{4mT(n)} \cdot (mn + 3mT(n) + O(\log mn)) && \text{(Proposition 14)} \\
&\qquad + \mathop{\mathbb{E}}_{X}[\log(1/D_n^m(X))] + |r| + a(n) + O(\log(m) + \log T(n)) \\
& && \text{(Lemma 29)} \\
&\leq H(D_n^m) + |r| + a(n) + O(\log(m) + \log T(n)) \\
&\leq \mathop{\mathbb{E}}_{X,r}[\mathsf{K}(X) + \mathsf{K}(r \mid X)] + a(n) + O(\log(m) + \log T(n)) \\
&\leq \mathop{\mathbb{E}}_{X,r}[\mathsf{K}(X, r)] + a(n) + O(\log(m) + \log T(n)), && \text{(Time-unbounded S.o.I.)}
\end{aligned}$$

where the second last inequality uses the fact that for any distribution D, the Shannon entropy $H(D)$ is at most $\mathbb{E}[\mathsf{K}(x)]$ for $x \sim D$ (see [23, Theorem 8.1.1]), as well as a counting argument showing that $\mathbb{E}_{X,r}[\mathsf{K}(r \mid X)] \geq |r| - 3$.

Rearranging the above, we get

$$\mathop{\mathbb{E}}_{X,r}[\mathsf{K}^t(X, r) - \mathsf{K}(X, r)] \leq a(n) + O(\log m + \log T(n))$$

as desired. ◀

▶ **Theorem 38.** *If* $(\mathsf{MK^tP}, \mathcal{U}) \in \mathsf{AvgBPP}$*, then for any time-constructible functions* $s, T, a\colon \mathbb{N} \to \mathbb{N}$*, any* $\varepsilon \in (0,1)$*, and any constant* $\zeta > 0$*, there is an* ε*-correlative RRHS-refuter for* $\mathsf{SIZE}[s(n)]$ *under* $\mathsf{Samp}[T(n)]/a(n)$ *taking parameters* $n \in \mathbb{N}$ *and* $\varepsilon \in (0,1)$ *with sample complexity*

$$m := \left(\frac{s(n) + a(n) + \log T(n)}{\varepsilon^2}\right)^{1+\zeta}$$

and running time $\mathsf{poly}\big(n, \varepsilon^{-1}, T(n), a(n), s(n)\big)$.

Proof. The proof closely follows that of [16, Theorem 8].

The (correlative) RRHS-refuter R. On input $n \in \mathbb{N}$, $\varepsilon > 0$, and a set

$$S = \left(\left\langle x^{(1)}, b^{(1)}\right\rangle, \ldots, \left\langle x^{(m)}, b^{(m)}\right\rangle\right)$$

of samples, let $X := (x^{(1)}, \ldots, x^{(m)})$ and $b := (b^{(1)}, \ldots, b^{(m)})$. R is defined as follows.

1. Compute $t := p_1(T(n), m)$, where p_1 is the polynomial from Lemma 37. Also compute $t' := t'(p(t))$, where t' is the polynomial from Lemma 32 and p is the polynomial from Lemma 31.
2. Sample $r \sim \mathcal{U}_{3mT(n)}$.
3. Compute

$$\beta := A\left((X, r), 1^t\right) \text{ and}$$
$$\beta' := A\left((X, b, r), 1^{t'}\right),$$

   where A is the randomized algorithm from Lemma 31.
4. Output "correlative" if $\beta' - \beta \le \theta$, where $\theta = \left(1 - \frac{\varepsilon^2}{16}\right) m$, and output "random" otherwise.

We now argue for the correctness of R. Consider any distribution $D_n \in \mathsf{Samp}[T(n)]/a(n)$.

Correlative Case (Soundness). Suppose the samples S are i.i.d. from a distribution D' on $\{0,1\}^n \times \{0,1\}$ such that the marginal on $\{0,1\}^n$ equals D_n, and there exists $f \in \mathsf{SIZE}[s(n)]$ such that

$$\Pr_{\langle x^{(i)}, b^{(i)}\rangle \sim D'}\left[b^{(i)} = f(x^{(i)})\right] \ge \frac{1}{2} + \frac{\varepsilon}{2}.$$

Chernoff bounds imply that

$$\left|\{i \in [m] \mid b_i = f(x^{(i)})\}\right| \ge (1/2 + \varepsilon/4) \cdot m$$

holds with probability at least $1 - \exp(-2m(\varepsilon/4)^2)$ over the choice of samples $S \sim (D')^m$, in which case the conditions of Lemma 32 are met.

Now, suppose that in Step 3 of R, β and β' output by the algorithm A are good approximations in terms of Lemma 31, which happens with probability at least $1 - o(1)$. Moreover, by Lemma 37,

$$\begin{aligned}\mathop{\mathbb{E}}_{X,r}\left[\mathsf{K}^t(X,r) - \mathsf{pK}^{p(t)}(X,r)\right] &\le \mathop{\mathbb{E}}_{X,r}\left[\mathsf{K}^t(X,r) - \mathsf{K}(X,r)\right] && \text{(Prop. 13)}\\ &\le a(n) + O(\log(mT(n))).\end{aligned}$$

Applying Markov's inequality, with probability at least $3/4$, there is a constant c such that

$$\mathsf{K}^t(X,r) - \mathsf{pK}^{p(t)}(X,r) \le c \cdot (a(n) + \log(mT(n))). \tag{10}$$

Thus, by a union bound, with probability at least $3/4 - o(1) > 2/3$ over the samples $S \sim (D')^m$ and the internal randomness of R,

$$\begin{aligned}
\beta' - \beta &\le \mathsf{K}^{t'}(X,b,r) - \mathsf{pK}^{p(t)}(X,r) + \log p(t) && (\beta' \text{ and } \beta \text{ are good approximations}) \\
&\le \left(\mathsf{K}^t(X,r) - \mathsf{pK}^{p(t)}(X,r)\right) + \log p(t) + \ell_s(n) + \left(1 - \varepsilon^2/8\right) \cdot m && \text{(Lemma 32)} \\
&\le m \cdot \left(1 - \varepsilon^2/8\right) + c \cdot (a(n) + \log(mT(n))) + \ell_s(n) && \text{(Eq. (10))} \\
&< \theta.
\end{aligned}$$

For the last inequality, observe that by our choice of $m = \omega((s(n) + \log T(n) + a(n)) \cdot \varepsilon^{-2})$,

$$m > 16 \cdot \left(\frac{c \cdot (a(n) + \log m + \log T(n)) + \ell_s(n)}{\varepsilon^2}\right);$$

re-written,

$$\begin{aligned}
m \cdot \left(1 - \varepsilon^2/8\right) + c \cdot (a(n) + \log(mT(n))) + \ell_s(n) &< m \cdot \left(1 - \varepsilon^2/16\right) \\
&= \theta.
\end{aligned}$$

Thus, R will output "correlative".

Random Case (Completeness). Suppose the labels b_i are sampled from $\mathcal{U}$. For $X \sim D_n^m$, $r \sim \mathcal{U}_{3mT(n)}$, and $b \sim \mathcal{U}^m$, we get by Lemma 37 and Markov's inequality, that, with probability at least $3/4$ over X, r,

$$\mathsf{K}^t(X,r) - K(X,r) \le 4(a(n) + O(\log mT(n))). \tag{11}$$

Since β' and β are good estimates with high probability, we get that, with probability at least $3/4 - o(1)$ over X, r, b and the internal randomness of A,

$$\begin{aligned}
\beta' - \beta &\ge \mathsf{pK}^{p(t')}(X,b,r) - \mathsf{K}^t(X,r) - O(\log(mT(n))) && (\beta', \beta \text{ good w.h.p.}) \\
&\ge \mathsf{K}(X,b,r) - \mathsf{K}^t(X,r) - O(\log(mT(n))) && \text{(Prop. 13)} \\
&\ge \mathsf{K}(X,r) + \mathsf{K}(b \mid X,r) - \mathsf{K}^t(X,r) - O(\log(mT(n))) && \text{(Lemma 15)} \\
&= m - \left(\mathsf{K}^t(X,r) - \mathsf{K}(X,r)\right) - O(\log(mT(n))) && (b \sim \mathcal{U}^m) \\
&\ge m - 4\,(a(n) + O(\log(mT(n)))) && \text{(Eq. (11))} \\
&> \theta,
\end{aligned}$$

and hence R outputs "random". ◀

4.4 Learning from Worst-case Easiness of MKTP

In this section, we show that if MKTP is easy for efficient randomized algorithms in the *worst case*, then it is possible to PAC learn *without* white-box access to the target distribution.

The following lemma is analogous to the source-coding lemma for K^t, Lemma 29, but with some modifications to allow for KT-compression in the case that we have many independent samples from the distribution D_n.

▶ **Lemma 39.** *For some constant $d \in \mathbb{N}$, the following holds unconditionally. For any $T, a : \mathbb{N} \to \mathbb{N}$, $m, n \in \mathbb{N}$, distribution $D \in \mathsf{Samp}[T(n)]/a(n)$, and string $X = (x^{(1)}, ..., x^{(m)}) \in \mathsf{Supp}(D_n^m)$,*

$$\mathsf{KT}(X, r) \leq \log(1/D_n^m(X)) + |r| + a(n) + d \cdot m^{3/4} \cdot T(n)^3$$

holds with probability at least $1 - \frac{1}{6mT(n)}$ over $r \sim \mathcal{U}_{4mT(n)}$.

Moreover, for any $s : \mathbb{N} \to \mathbb{N}$ and function $f \in \mathsf{SIZE}[s(n)]$,

$$\mathsf{KT}(X, f(x^{(1)}), ..., f(x^{(m)}), r) \leq \log(1/D_n^m(X)) + 2\ell_s(n) + |r| + a(n) + d \cdot m^{3/4} \cdot T(n)^3$$

holds with probability at least $1 - \frac{1}{6mT(n)}$ over $r \sim \mathcal{U}_{4mT(n)}$.

Proof. The proof is quite similar to that of Lemma 29, with some modifications (namely, the partitioning of $[m]$) to get the bound for KT-complexity. Let A be a non-uniform algorithm sampling $D \in \mathsf{Samp}[T(n)]/a(n)$. That is, there is some $\alpha \in \{0,1\}^{a(n)}$ such that for any $x \in \mathrm{supp}(D_n)$,

$$\Pr_{w \sim \mathcal{U}_{T(n)}}[A(w; \alpha, 1^n)] = D_n(x).$$

Consider any $X = (x^{(1)}, ..., x^{(m)}) \in \mathsf{Supp}(D_n^m)$. For N and L chosen later, we will partition $[m]$ into N blocks $b_1, ..., b_N$, each of size at most L. For every block b_j, let s_j be the largest integer such that $D_n^L(x^{(j_1)}, ..., x^{(j_L)}) \leq 2^{-s_j}$, where $b_j = \{j_1, ..., j_L\}$, and let $s = \sum_{j \in [N]} s_j$. For each b_j, consider a universal hash function family $\mathcal{H}_j = \{h : \{0,1\}^{L \cdot T(n)} \to \{0,1\}^{k_j}\}$, where $k_j = L \cdot T(n) - s_j - \log(12m^2T(n)) - 1$. As in Lemma 29, we represent hash functions with Toeplitz matrices.

For each block b_j, define a set

$$S_j := \{(w_1, ..., w_L) \in (\{0,1\}^{T(n)})^L \mid \forall l \in [L],\ A(w_l; \alpha, 1^n) = x^{(j_l)}\},$$

where j_l denotes the l^{th} element of b_j. For each $j \in [N]$, define a random variable $X_j := |S_j \cap h^{-1}(0^{k_j})|$, where $h \sim \mathcal{H}_j$. Arguing as in Lemma 29,

$$\Pr[X_j = 0] \leq \frac{1}{12m^2T(n)}.$$

Now, for each $j \in [N]$, define a random variable $Y_j = |h^{-1}(0^{k_j})|$, where $h \sim \mathcal{H}_j$. Arguing as in Lemma 29,

$$\Pr[Y_j \geq 2^{s_j + 2\log(12m^2T(n)) + 1}] \leq \frac{1}{12m^2T(n)}.$$

By a union bound, with probability at least $1 - 1/6mT(n)$, we have that for every $j \in [N]$, $X_j \neq 0$ and $Y_j < 2^{s_j + 2\log(12m^2T(n)) + 1}$.

Assume the above holds. It is possible to obtain any bit of a substring $x^{(i)}$ of X, for i in some block b_j, from the description of the hash function h sampled from $\mathcal{H}_j$, the index of a string $(w_1, ..., w_L) \in S_j$ in the set $h^{-1}(0^{k_j})$, and the advice string α used in the sampler A. In particular, $x^{(i)}$ may be recovered by performing Gaussian elimination to compute the set $h^{-1}(0^{k_j})$ from the description of h, locating w_l in this set such that i is the l^{th} element of b_j, and then returning the desired bit of $A(w_l; \alpha, 1^n) = x^{(i)}$. Given h, this requires at most $\log Y_j \leq s_j + 2\log(12m^2T(n)) + 1$ bits to describe the position of $(w_1, ..., w_L)$ in $h^{-1}(0^{k_j})$ and $|\alpha| = a(n)$ bits to run the sampler A. Define the "random" string $r_j \in \{0,1\}^{3L \cdot T(n)}$ as the description of $h \sim \mathcal{H}_j$. So, a description working for *any* block (and therefore any bit of X) is of length

$$\sum_{j\in[N]} \left[s_j + 2\log(12m^2T(n)) + 1\right] + a(n) \le s + N \cdot (2\log(12m^2T(n)) + 1) + a(n)$$

given randomness $r = (r_1, ..., r_N)$ of length $3L \cdot N \cdot T(n)$. The amount of time required is dominated by the Gaussian elimination step, at most $O((L \cdot T(n))^3)$.

To obtain some bit $f(x^{(i)})$, one may apply the above procedure to obtain $x^{(i)}$ and then apply an $\ell_s(n)$-bit description of a circuit computing f, taking additional time at most $\ell_s(n)$.

Overall, with probability at least $1 - 1/6mT(n)$ over r, we have that

$$\mathsf{KT}(X, r) \le s + |r| + a(n) + N \cdot O(\log(mT(n))) + O((L \cdot T(n))^3)$$

and

$$\mathsf{KT}(X, f(x^{(1)}), ..., f(x^{(m)}), r) \le s + |r| + 2\ell_s(n) + a(n) + N \cdot O(\log(mT(n))) + O((L \cdot T(n))^3).$$

The lemma follows by setting $L = m^{1/4}$ and $N = \lceil m^{3/4} \rceil$. ◀

The following lemma is analogous to Lemma 37, showing that KT and K complexities are somewhat close in the expectation over efficiently sampled strings.

▶ **Lemma 40.** *For any $T, a \colon \mathbb{N} \to \mathbb{N}$, $n, m \in \mathbb{N}$, $D_n \in \mathsf{Samp}[T(n)]/a(n)$, $X \sim D_n^m$, $b \sim \mathcal{U}_m$, and $r \sim \mathcal{U}_{4mT(n)}$,*

$$\mathop{\mathbb{E}}_{X,b,r}[\mathsf{KT}(X, b, r) - \mathsf{K}(X, b, r)] \le a(n) + 2d \cdot m^{3/4} \cdot T(n)^3,$$

where d is the constant from Lemma 39.

Moreover, for any function $f \in \mathsf{SIZE}[s(n)]$,

$$\mathop{\mathbb{E}}_{X,r}[\mathsf{KT}(X, f(x^{(1)}), ..., f(x^{(1)}), r) - \mathsf{K}(X, f(x^{(1)}), ..., f(x^{(1)}), r)] \le a(n) + 2\ell_s(n) + 2d \cdot m^{3/4} \cdot T(n)^3.$$

Proof. The proof closely follows that of Lemma 37.

$$\begin{aligned}
\mathop{\mathbb{E}}_{X,b,r}[\mathsf{KT}(X, b, r)] &\le \mathop{\mathbb{E}}_{X,r}[\mathsf{KT}(X, r)] + |b| + \log m \\
&\le \frac{1}{6mT(n)} \cdot (mn + 4mT(n) + m + O(\log mn)) \\
&\quad + \mathop{\mathbb{E}}_{X}[\log(1/D_n^m(X))] + |r| + a(n) + d \cdot m^{3/4} \cdot T(n)^3 + |b| + \log m \qquad \text{(Lemma 39)} \\
&\le H(D_n^m) + |r| + a(n) + d \cdot m^{3/4} \cdot T(n)^3 + |b| + O(\log m) \\
&\le \mathop{\mathbb{E}}_{X,b,r}[\mathsf{K}(X) + \mathsf{K}(b \mid X) + \mathsf{K}(r \mid b, X)] + a(n) + d \cdot m^{3/4} \cdot T(n)^3 + O(\log m) \\
&\le \mathop{\mathbb{E}}_{X,b,r}[\mathsf{K}(X, b, r)] + a(n) + 2d \cdot m^{3/4} \cdot T(n)^3. \qquad \text{(Time-unbounded S.o.I.)}
\end{aligned}$$

Rearranging the above, we get

$$\mathop{\mathbb{E}}_{X,b,r}[\mathsf{KT}(X, b, r) - \mathsf{K}(X, b, r)] \le a(n) + 2d \cdot m^{3/4} \cdot T(n)^3$$

as desired.

The proof of the "moreover" part of the lemma is very similar. It follows by applying the "moreover" part of Lemma 39 in the second line, and in the last line using the simple fact that $\mathsf{K}(X, r) \le \mathsf{K}(X, f(x^{(1)}), ..., f(x^{(1)}), r)$. ◀

▶ **Theorem 41.** *If* $\mathsf{MKTP} \in \mathsf{BPP}$*, then for any time-constructible functions* $s, T, a\colon \mathbb{N} \to \mathbb{N}$*, and any* $\varepsilon \in (0,1)$*, there is an* ε*-correlative RRHS-refuter for* $\mathsf{SIZE}[s(n)]$ *under* $\mathsf{Samp}[T(n)]/a(n)$ *taking parameters* $n \in \mathbb{N}$ *and* $\varepsilon \in (0,1)$ *with sample complexity*

$$m := \left(\frac{s(n) + a(n) + T(n)^{12}}{\varepsilon^8}\right)^{12}$$

and running time $\mathsf{poly}\left(n, \varepsilon^{-1}, T(n), a(n), s(n)\right)$.

Proof. Let A_0 be the assumed randomized algorithm for MKTP, and let A be the randomized poly-time "search" algorithm that on input y runs $A_0(y, 1^s)$ for $s = 1, ..., |y| + \log|y|$ and outputs the smallest s on which A accepts. It is not hard to see, using standard techniques, that A can be made to correctly compute $\mathsf{KT}(y)$ with probability $1 - 2^{-|y|}$.

The (correlative) RRHS-refuter R. On input $n \in \mathbb{N}$, $\varepsilon > 0$, and a set

$$S = \left(\left\langle x^{(1)}, b^{(1)}\right\rangle, \ldots, \left\langle x^{(m)}, b^{(m)}\right\rangle\right)$$

of samples, let

$$k := \left(\frac{s(n) + a(n) + T(n)^{12}}{\varepsilon^8}\right)^2.$$

Partition the $m = k^6$ samples S into k^5 sets, each containing k samples. Denote these sets S_i, for $i \in [k]$. Then partition each S_i into two equally sized sets,

$$S_i^0 = \left(\left\langle x_i^{(1)}, b_i^{(1)}\right\rangle, \ldots, \left\langle x_i^{(k/2)}, b_i^{(k/2)}\right\rangle\right) \text{ and } S_i^1 = \left(\left\langle x_i^{(k/2+1)}, b_i^{(k/2+1)}\right\rangle, \ldots, \left\langle x_i^{(k)}, b_i^{(k)}\right\rangle\right).$$

Let $Z_i := (x_i^{(1)}, ..., x_i^{(k/2)})$, $X_i := (x_i^{(k/2+1)}, \ldots, x_i^{(k)})$ and $b_i := (b_i^{(k/2+1)}, \ldots, b_i^{(k)})$.

R is defined as follows. We repeat the following k^5 times: once on each set of samples S_i. For simplicity, we omit the subscripts i: denote $(\langle x^{(1)}, b^{(1)}\rangle, \ldots, \langle x^{(k)}, b^{(k)}\rangle) := S_i$, $Z := Z_i$, $X := X_i$, and $b := b_i$.

1. Sample $r \sim \mathcal{U}_{2kT(n)}$.
2. Sample $u \sim \mathcal{U}_{k/2}$, and using the first half of the samples Z, compute

   $$\gamma_i := A(Z, u, r).$$

3. Using the second half of the samples X along with their given labels b, compute

   $$\beta_i := A\left(X, b, r\right).$$

4. Let $w_i = \gamma_i - \beta_i$.
5. At the end, after k^5 repetitions of the above, take the sum

   $$w = \sum_{i \in [k^5]} w_i.$$

   Let d be the constant from Lemma 39. Output "correlative" if $w \geq k^5 \cdot \theta$, where $\theta = 2 \cdot \left(a(n) + 4d \cdot k^{3/4} \cdot T(n)^3\right)$, and output "random" otherwise.

We begin by showing that the expected value of γ_i is roughly $H(D_n^{k/2}) + k/2 + |r|$. On one hand, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z,u,r,A}[\gamma_i] &\leq \mathop{\mathbb{E}}_{Z,u,r}[\mathsf{KT}(Z,u,r)] + O(1) && \text{(definition of } A\text{)} \\
&\leq \mathbb{E}[\mathsf{K}(Z,u,r)] + a(n) + 2d \cdot k^{3/4} \cdot T(n)^3 + O(1) && \text{(Lemma 40)} \\
&\leq \Big(H(D_n^{k/2}) + k/2 + |r|\Big) + a(n) + 3d \cdot k^{3/4} \cdot T(n)^3, && (12)
\end{aligned}$$

where the last line follows by a counting argument and the fact that $\mathbb{E}[\mathsf{K}(Z)] \leq H(D_n^{k/2})$.

On the other hand,

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z,u,r,A}[\gamma_i] &\geq \mathop{\mathbb{E}}_{Z,u,r}[\mathsf{KT}(Z,u,r)] - O(1) && \text{(definition of } A\text{)} \\
&\geq \mathbb{E}[\mathsf{K}(Z,u,r)] - O(1) \\
&\geq \mathbb{E}[\mathsf{K}(Z) + \mathsf{K}(u \mid Z) + \mathsf{K}(r \mid Z,u)] - O(\log(kn)) && \text{(symmetry of information)} \\
&\geq \Big(H(D_n^{k/2}) + k/2 + |r|\Big) - O(\log(kn)), && (13)
\end{aligned}$$

where in the last line we use that $\mathbb{E}[\mathsf{K}(Z)] \geq H(D_n^{k/2}) - O(\log(kn))$ [23, Theorem 8.1.1].

Correlative Case (Soundness). Suppose the samples S are i.i.d. from a distribution D' on $\{0,1\}^n \times \{0,1\}$ such that the marginal on $\{0,1\}^n$ equals D_n, and there exists $f \in \mathsf{SIZE}[s(n)]$ such that

$$\Pr_{\langle x^{(j)}, b^{(j)} \rangle \sim D'}\left[b^{(j)} = f(x^{(j)})\right] \geq \frac{1}{2} + \frac{\varepsilon}{2}.$$

Chernoff bounds imply that

$$\left|\{j \in \{k/2+1, ..., k\} \mid b_i = f(x^{(j)})\}\right| \geq (1/2 + \varepsilon/4) \cdot k/2$$

holds with probability at least $1 - \exp(-k(\varepsilon/4)^2/8)$ over the choice of samples S_i^1, in which case the conditions of Lemma 34 are met. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X,b,r,A}[\beta_i] &\leq \mathop{\mathbb{E}}_{X,b,r}[\mathsf{KT}(X,b,r)] + O(1) \\
&\leq \mathbb{E}[\mathsf{KT}(X, f(x^{(k/2+1)}), ..., f(x^{(k)}), r)] + (1 - \varepsilon^2/8) \cdot k/2 + 2k^{3/4} && \text{(Lemma 34)} \\
&\leq \mathbb{E}[\mathsf{K}(X, f(x^{(k/2+1)}), ..., f(x^{(k)}), r)] + (1 - \varepsilon^2/8) \cdot k/2 + a(n) + 2\ell_s(n) + 3d \cdot k^{3/4} \cdot T(n)^3 \\
&\leq H(D_n^{k/2}) + (1 - \varepsilon^2/8) \cdot k/2 + |r| + a(n) + 3\ell_s(n) + 3d \cdot k^{3/4} \cdot T(n)^3,
\end{aligned}$$

where the second-last line uses Lemma 40, and the last line uses the observation that

$$\mathsf{K}(f(x^{(k/2+1)}), ..., f(x^{(k)}) \mid X) \leq \ell_s(n).$$

Combining the above with Eq. (13), we have

$$\begin{aligned}
\mathbb{E}[\gamma_i - \beta_i] &\geq \frac{\varepsilon^2}{16} \cdot k - \Big(a(n) + 3\ell_s(n) + O(k^{3/4} T(n)^3)\Big) \\
&\geq 2\theta,
\end{aligned}$$

by our choices of k and θ.

After k^5 trials of the above, we have $\mathbb{E}[w] \geq 2k^5\theta$. By Hoeffding's inequality, with probability at least $1 - 2^{-k}$, it holds that $|2k^5\theta - w| \leq k^5\theta$, and so

$$w \geq k^5\theta,$$

in which case R will output "correlative".

Random Case (Completeness). Suppose the labels b_j are sampled from $\mathcal{U}$. Arguing as in Eq. (13),

$$\mathop{\mathbb{E}}_{X,b,r,A}[\beta_i] \geq \left(H(D_n^{k/2}) + k/2 + |r|\right) - O(\log(kn)).$$

Combining the above with Eq. (12),

$$\begin{aligned}\mathbb{E}[\gamma_i - \beta_i] &\leq a(n) + 4d \cdot k^{3/4} \cdot T(n)^3 \\ &= \theta/2.\end{aligned}$$

After k^5 trials of the above, we have $\mathbb{E}[w] \leq k^5\theta/2$. By Hoeffding's inequality, with probability at least $1 - 2^{-k}$, it holds that $|k^5\theta/2 - w| < k^5\theta/2$, and so

$$w < k^5\theta,$$

in which case R will output "random". ◀

5 Open questions

We showed that "natural propeties" for more expressive Kolmogorov-complexity relatives of MCSP such as MKTP and $\mathsf{MK^tP}$ allow one to cross the divide between learning algorithms with membership queries and those without. An obvious disadvantage of relying on more expressive Kolmogorov measures rather than MCSP is that it is difficult to get meaningful circuit class restrictions when talking about MKTP or $\mathsf{MK^tP}$, and utilize the known circuit lower bound proofs for these restricted circuit classes in order to derive a learning algorithm. Can one use our understanding of $\mathsf{AC}^0[2]$ circuit lower bounds (e.g., the known natural property for $\mathsf{AC}^0[2]$) to get an RRHS-refuter for $\mathsf{AC}^0[2]$ on uniform distribution? This question is also very interesting from the point of view of cryptography in the context of efficient constructions of weak PRFs; see, e.g., [8] for more discussion on this direction.

Another question is whether it is possible to bridge the gap between the assumptions used in our two main theorems. More precisely, is it possible to get an agnostic PAC learning algorithm over any *not necessarily explicitly given* polysize samplable distribution ensemble D from a one-sided average-case heuristic for MKTP rather than $\mathsf{MK^tP}$?

References

1 Eric Allender, Harry Buhrman, Michal Koucký, Dieter van Melkebeek, and Detlef Ronneburger. Power from random strings. *SIAM J. Comput.*, 35(6):1467–1493, 2006. doi:10.1137/050628994.

2 Eric Allender, Joshua A. Grochow, Dieter van Melkebeek, Cristopher Moore, and Andrew Morgan. Minimum circuit size, graph isomorphism, and related problems. *SIAM J. Comput.*, 47(4):1339–1372, 2018. doi:10.1137/17M1157970.

3 Luis Filipe Coelho Antunes and Lance Fortnow. Worst-case running times for average-case algorithms. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity, CCC 2009, Paris, France, 15-18 July 2009*, pages 298–303. IEEE Computer Society, 2009. doi:10.1109/CCC.2009.12.

4 Eric Binnendyk, Marco Carmosino, Antonina Kolokolova, R. Ramyaa, and Manuel Sabin. Learning with distributional inverters. In Sanjoy Dasgupta and Nika Haghtalab, editors, *International Conference on Algorithmic Learning Theory, 29-1 April 2022, Paris, France*, volume 167 of *Proceedings of Machine Learning Research*, pages 90–106. PMLR, 2022. URL: https://proceedings.mlr.press/v167/binnendyk22a.html.

5 Avrim Blum, Merrick L. Furst, Michael J. Kearns, and Richard J. Lipton. Cryptographic primitives based on hard learning problems. In Douglas R. Stinson, editor, *Advances in Cryptology – CRYPTO '93, 13th Annual International Cryptology Conference, Santa Barbara, California, USA, August 22-26, 1993, Proceedings*, volume 773 of *Lecture Notes in Computer Science*, pages 278–291. Springer, 1993. doi:10.1007/3-540-48329-2_24.

6 Andrej Bogdanov and Alon Rosen. Pseudorandom functions: Three decades later. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography*, pages 79–158. Springer International Publishing, 2017. doi:10.1007/978-3-319-57048-8_3.

7 Andrej Bogdanov and Luca Trevisan. Average-case complexity. *Found. Trends Theor. Comput. Sci.*, 2(1), 2006. doi:10.1561/0400000004.

8 Dan Boneh, Yuval Ishai, Alain Passelègue, Amit Sahai, and David J. Wu. Exploring crypto dark matter: New simple PRF candidates and their applications. In Amos Beimel and Stefan Dziembowski, editors, *Theory of Cryptography – 16th International Conference, TCC 2018, Panaji, India, November 11-14, 2018, Proceedings, Part II*, volume 11240 of *Lecture Notes in Computer Science*, pages 699–729. Springer, 2018. doi:10.1007/978-3-030-03810-6_25.

9 Harry Buhrman, Lance Fortnow, and Aduri Pavan. Some results on derandomization. *Theory Comput. Syst.*, 38(2):211–227, 2005. doi:10.1007/s00224-004-1194-y.

10 Marco L. Carmosino, Russell Impagliazzo, Valentine Kabanets, and Antonina Kolokolova. Learning algorithms from natural proofs. In Ran Raz, editor, *31st Conference on Computational Complexity, CCC 2016, May 29 to June 1, 2016, Tokyo, Japan*, volume 50 of *LIPIcs*, pages 10:1–10:24. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2016. doi:10.4230/LIPIcs.CCC.2016.10.

11 Marco L. Carmosino, Russell Impagliazzo, Valentine Kabanets, and Antonina Kolokolova. Agnostic learning from tolerant natural proofs. In Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA*, volume 81 of *LIPIcs*, pages 35:1–35:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi:10.4230/LIPIcs.APPROX-RANDOM.2017.35.

12 Vitaly Feldman. Distribution-specific agnostic boosting. In Andrew Chi-Chih Yao, editor, *Innovations in Computer Science – ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*, pages 241–250. Tsinghua University Press, 2010. URL: http://conference.iiis.tsinghua.edu.cn/ICS2010/content/papers/20.html.

13 Halley Goldberg, Valentine Kabanets, Zhenjian Lu, and Igor Carboni Oliveira. Probabilistic kolmogorov complexity with applications to average-case complexity. In Shachar Lovett, editor, *37th Computational Complexity Conference, CCC 2022, July 20-23, 2022, Philadelphia, PA, USA*, volume 234 of *LIPIcs*, pages 16:1–16:60. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. doi:10.4230/LIPIcs.CCC.2022.16.

14 Alexander Golovnev, Rahul Ilango, Russell Impagliazzo, Valentine Kabanets, Antonina Kolokolova, and Avishay Tal. $AC^0[p]$ lower bounds against MCSP via the coin problem. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*, volume 132 of *LIPIcs*, pages 66:1–66:15. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.ICALP.2019.66.

15 Shuichi Hirahara. Non-black-box worst-case to average-case reductions within NP. In Mikkel Thorup, editor, *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 247–258. IEEE Computer Society, 2018. doi:10.1109/FOCS.2018.00032.

16 Shuichi Hirahara and Mikito Nanashima. On worst-case learning in relativized heuristica. In *62nd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2021, Denver, CO, USA, February 7-10, 2022*, pages 751–758. IEEE, 2021. doi:10.1109/FOCS52979.2021.00078.

17 Rahul Ilango, Bruno Loff, and Igor Carboni Oliveira. Np-hardness of circuit minimization for multi-output functions. In Shubhangi Saraf, editor, *35th Computational Complexity Conference, CCC 2020, July 28-31, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 169 of *LIPIcs*, pages 22:1–22:36. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.CCC.2020.22.

18 Russell Impagliazzo and Leonid A. Levin. No better ways to generate hard NP instances than picking uniformly at random. In *31st Annual Symposium on Foundations of Computer Science, St. Louis, Missouri, USA, October 22-24, 1990, Volume II*, pages 812–821. IEEE Computer Society, 1990. doi:10.1109/FSCS.1990.89604.

19 Russell Impagliazzo and Michael Luby. One-way functions are essential for complexity based cryptography (extended abstract). In *30th Annual Symposium on Foundations of Computer Science, Research Triangle Park, North Carolina, USA, 30 October – 1 November 1989*, pages 230–235. IEEE Computer Society, 1989. doi:10.1109/SFCS.1989.63483.

20 Adam Kalai and Varun Kanade. Potential-based agnostic boosting. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 880–888. Curran Associates, Inc., 2009. URL: https://proceedings.neurips.cc/paper/2009/hash/13f9896df61279c928f19721878fac41-Abstract.html.

21 Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Mach. Learn.*, 17(2-3):115–141, 1994. doi:10.1007/BF00993468.

22 Pravesh K. Kothari and Roi Livni. Improper learning by refuting. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 55:1–55:10. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.ITCS.2018.55.

23 Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications, 4th Edition*. Texts in Computer Science. Springer, 2019. doi:10.1007/978-3-030-11298-1.

24 Moni Naor and Eylon Yogev. Bloom filters in adversarial environments. *ACM Trans. Algorithms*, 15(3):35:1–35:30, 2019. doi:10.1145/3306193.

25 Noam Nisan and Avi Wigderson. Hardness vs randomness. *J. Comput. Syst. Sci.*, 49(2):149–167, 1994. doi:10.1016/S0022-0000(05)80043-1.

26 Rafail Ostrovsky and Avi Wigderson. One-way fuctions are essential for non-trivial zero-knowledge. In *Second Israel Symposium on Theory of Computing Systems, ISTCS 1993, Natanya, Israel, June 7-9, 1993, Proceedings*, pages 3–17. IEEE Computer Society, 1993. doi:10.1109/ISTCS.1993.253489.

27 Alexander A. Razborov and Steven Rudich. Natural proofs. *J. Comput. Syst. Sci.*, 55(1):24–35, 1997. doi:10.1006/jcss.1997.1494.

28 Salil P. Vadhan. On learning vs. refutation. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 1835–1848. PMLR, 2017. URL: http://proceedings.mlr.press/v65/vadhan17a.html.

29 Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984. doi:10.1145/1968.1972.

30 Andrew Chi-Chih Yao. Theory and applications of trapdoor functions (extended abstract). In *23rd Annual Symposium on Foundations of Computer Science, Chicago, Illinois, USA, 3-5 November 1982*, pages 80–91. IEEE Computer Society, 1982. doi:10.1109/SFCS.1982.45.

31 Alexander K. Zvonkin and Leonid A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83, 1970.