Pseudo-Derandomizing Learning and Approximation

Igor Carboni Oliveira
Department of Computer Science, University of Oxford, United Kingdom.
igor.carboni.oliveira@cs.ox.ac.uk

Rahul Santhanam
Department of Computer Science, University of Oxford, United Kingdom.
rahul.santhanam@cs.ox.ac.uk

Abstract

We continue the study of *pseudo-deterministic algorithms* initiated by Gat and Goldwasser [7]. A pseudo-deterministic algorithm is a probabilistic algorithm which produces a fixed output with high probability. We explore pseudo-determinism in the settings of *learning* and *approximation*. Our goal is to simulate known randomized algorithms in these settings by pseudo-deterministic algorithms in a generic fashion – a goal we succinctly term *pseudo-derandomization*.

Learning. In the setting of learning with membership queries, we first show that randomized learning algorithms can be derandomized (resp. pseudo-derandomized) under the standard hardness assumption that E (resp. BPE) requires large Boolean circuits. Thus, despite the fact that learning is an algorithmic task that requires interaction with an oracle, standard hardness assumptions suffice to (pseudo-)derandomize it. We also *unconditionally* pseudo-derandomize any quasi-polynomial time learning algorithm for polynomial size circuits on infinitely many input lengths in sub-exponential time.

Next, we establish a generic connection between learning and derandomization in the reverse direction, by showing that deterministic (resp. pseudo-deterministic) learning algorithms for a concept class $\mathfrak{C}$ imply hitting sets against $\mathfrak{C}$ that are computable deterministically (resp. pseudo-deterministically). In particular, this suggests a new approach to constructing hitting set generators against $\mathcal{AC}^0[p]$ circuits by giving a deterministic learning algorithm for $\mathcal{AC}^0[p]$.

Approximation. Turning to approximation, we *unconditionally* pseudo-derandomize any polytime randomized approximation scheme for integer-valued functions infinitely often in subexponential time over any samplable distribution on inputs. As a corollary, we get that the $(0,1)$-Permanent has a fully pseudo-deterministic approximation scheme running in sub-exponential time infinitely often over any samplable distribution on inputs.

Finally, we investigate the notion of *approximate canonization* of Boolean circuits. We use a connection between pseudodeterministic learning and approximate canonization to show that if BPE does not have sub-exponential size circuits infinitely often, then there is a pseudo-deterministic approximate canonizer for $\mathcal{AC}^0[p]$ computable in quasi-polynomial time.

2012 ACM Subject Classification Theory of computation → Pseudorandomness and derandomization

Keywords and phrases derandomization, learning, approximation, boolean circuits

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2018.55

Related Version A full version of the paper is available at `https://eccc.weizmann.ac.il/report/2018/122/`.

Funding This work was supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant No. 615075.

Acknowledgements We thank Chris Brzuska for bringing [3] to our attention, Roei Tell for helpful discussions, and the reviewers for comments that improved the presentation.

1 Introduction

1.1 Context and Motivation

Randomness is a powerful algorithmic resource, used widely in tasks such as cryptography, distributed computing, learning, sampling and approximation. Although it often makes algorithmic tasks more efficient, randomness comes with issues. It introduces *uncertainty* – running a randomized algorithm multiple times, we cannot always expect to get the same answer. Moreover, randomized algorithms assume access to a source of independent and unbiased random bits, and this assumption is not always justified in the physical world.

Ideally, we would like to perform any efficient randomized task almost as efficiently without using randomness at all, or while using as little randomness as possible. This is the goal of *derandomization*, which has been widely studied in complexity theory. While generic derandomization is possible in many settings under widely believed circuit lower bound hypotheses, it also *implies* circuit lower bounds that are believed to be hard to establish (cf. [17, 21, 37]). Thus, while many specific randomized tasks can be derandomized, provable generic derandomization seems out of reach with our current state of knowledge.

A few years ago, Gat and Goldwasser [7] introduced the notion of *pseudo-deterministic algorithms*, motivated by applications in cryptography and distributed computing. A pseudo-deterministic algorithm is one that, on a given input, produces a fixed output with very high probability. Thus, a pseudo-deterministic algorithm is one that *looks* deterministic to an outside observer who is computationally bounded – even if such an observer were to run the algorithm multiple times, she is likely to always get the same answer.

Pseudo-deterministic algorithms have a very desirable feature possessed by deterministic algorithms, viz. little to no uncertainty in the output. Thus it is of interest to convert randomized algorithms to equivalent pseudo-deterministic ones – we term such a conversion "pseudo-derandomization".

There are several interesting examples of pseudo-derandomization known, including finding primitive roots [13] and quadratic non-residues (cf. [7]) in prime fields, finding variable settings for polynomial identity testing [7], and finding perfect matchings in bipartite graphs in parallel [9]. These pseudo-derandomization results exploit specific properties of the known randomized algorithms for these problems. The authors introduced a generic pseudo-derandomization approach for search problems in [29], and used it to give a sub-exponential pseudo-deterministic construction of primes infinitely often. This generic approach has been explored further by [16] and [10].

One limitation of the generic approach of [29] is that it seems to work only for *search problems* whose underlying relation is decidable in P or in BPP. In particular, the approach requires the ability to test in P or in BPP whether a given sequence of random choices made by a randomized algorithm is "good" or not. There are several important settings of randomized tasks where such a test is not available. We consider two such settings in this paper: *learning* and *approximation*.

1.2 Pseudoderandomization and learning

Our learning model is that of learning with membership queries, where the accuracy of the output hypothesis is measured with respect to the uniform distribution. A pseudo-deterministic learning algorithm in this model is a randomized algorithm that, when given access to a predetermined oracle, makes a fixed set of queries and outputs a fixed output

hypothesis with high probability.[1] (The pseudo-deterministic learning model is formalized in the natural way in Section A.2.)

This setting falls outside the "search problem" paradigm for a couple of different reasons: first, the algorithmic task is not self-contained but requires interaction with an oracle, and second, the test of whether an output hypothesis is good is not precise but approximate, and again requires interaction with the oracle.[2]

Pseudo-determinism is naturally a desirable property for learning algorithms. Indeed, consider a setting where Alice and Bob run the same learning program independently on the same data but wish to co-ordinate their predictions. Pseudo-determinism of the learning algorithm enables them to co-ordinate their predictions perfectly with high probability. In an alternative scenario, suppose Alice runs the learning algorithm to generate a hypothesis, and the hypothesis gets corrupted. Alice can recover the original hypothesis with high probability just by running the learning algorithm again.

The main question we ask is: can learning algorithms be derandomized or at least pseudo-derandomized in a generic fashion? Our first result is the observation that standard pseudo-random generators suffice to derandomize learning. This is somewhat surprising because standard pseudo-random generators are designed for self-contained algorithmic tasks, while learning requires interaction with an unknown oracle.

Recall that we consider randomized algorithms that learn under the uniform distribution and have membership-query access to the unknown function.

▶ **Theorem 1** (Conditional derandomization and pseudo-derandomization of learning)**.** *Let $\mathfrak{C} \subseteq \mathsf{P/poly}$ be an arbitrary circuit class, and suppose $\mathfrak{C}(s(n))$ can be learned to any constant accuracy by a randomized algorithm running in time $t(n) \geq n$.*

- *If $\mathsf{E} = \mathsf{DTIME}[2^{O(n)}]$ requires circuits of size $2^{\Omega(n)}$ on all large input lengths, then there exists a constant $c \geq 1$ such that $\mathfrak{C}$ can be deterministically learned to any constant accuracy in time at most $O(t(n)^c)$.*
- *If $\mathsf{BPE} = \mathsf{BPTIME}[2^{O(n)}]$ requires circuits of size $2^{\Omega(n)}$ on all large input lengths, then there exists a constant $c \geq 1$ such that $\mathfrak{C}(s)$ can be pseudo-deterministically learned to any constant accuracy in time at most $O(t(n)^c)$.*

The proof of conditional derandomization in Theorem 1 works as follows. Under the assumption that E requires exponential-size Boolean circuits almost everywhere, it is a standard consequence from [26, 18, 34] that there is a pseudo-random generator G computable in time $\mathrm{poly}(t(n))$ with seed length $O(\log(t(n))$ secure against circuits of size $t(n)^3$. We simulate the randomized learning algorithm using each output of the generator G as random sequence in turn to obtain hypothesis circuits $D_1 \ldots D_{\mathrm{poly}(t(n))}$, and then output the majority of these circuits as our hypothesis. To argue that this works, we show that if the simulation fails to output a correct hypothesis, there is a distinguisher for the PRG G, contrary to our assumption. The key idea is that a distinguisher can be constructed in $t(n)^3$ size by replacing the oracle in the simulation of the learning algorithm by a circuit from the class $\mathfrak{C}$ for which the simulation fails. The proof of conditional pseudo-derandomization works similarly, but we need to use an additional idea from [29]. Details are in Section 2.[3]

[1] We stress that we allow adaptive learning algorithms, and that the "canonical" set of inputs queried by the learner can depend on the target function. We do not require the order of the queries to be fixed.

[2] Also observe that the usual way of testing a learning hypothesis by drawing a set of random examples is not pseudo-deterministic.

[3] For simplicity, we have restricted the statement of Theorem 1 to constant-accuracy learners. As explained in Section 2, from a randomized ε-accuracy learner one can get a deterministic $O(\varepsilon)$-accuracy learner by using sufficiently strong generators (see Lemma 6).

We get some interesting corollaries from Theorem 1. Under standard hardness assumptions, both Jackson's polynomial-time learning algorithm for DNFs with membership queries [19] and the recent algorithm of [4] for $\mathcal{AC}^0[p]$ can be derandomized. Note that the randomized learner of [24] for $\mathcal{AC}^0$ has already been derandomized unconditionally by Sitharam [32].[4] Sitharam's deterministic learner exploits specific properties of $\mathcal{AC}^0$ circuits, while we are interested here in generic methods to derandomize and pseudo-derandomize learning algorithms.

Theorem 1 is conditional, but it can be used to establish an *unconditional* result for pseudo-derandomizing learning. This is in contrast to generic derandomization, which can only be done conditionally given our current knowledge of circuit lower bounds.

▶ **Theorem 2** (Unconditional pseudo-derandomization of learning)**.**
If P/poly *can be learned to any constant accuracy by a randomized algorithm running in quasi-polynomial time, then for each* $\gamma > 0$*,* P/poly *can be pseudo-deterministically learned to any constant accuracy in time* $O(2^{n^\gamma})$ *for infinitely many input lengths* n*.*

The proof of Theorem 2 proceeds in two steps. In the first step, we use a result of [28] to get circuit lower bounds for BPE from a non-trivial randomized learning for P/poly. In the second step, we apply a variant of Theorem 1 to derive an infinitely-often subexponential-time pseudo-deterministic learner using the circuit lower bounds for BPE.

The assumption in Theorem 2 is very strong; indeed, under standard cryptographic assumptions, P/poly does not have non-trivial learning algorithms. However, the proof technique of Theorem 2 works in the more general setting of *self-learners*, where a self-learner is a learning algorithm for a circuit class $\mathfrak{C}$ that produces a hypothesis in $\mathfrak{C}$ and moreover can itself be implemented in $\mathfrak{C}$. Theorem 2 is just the cleanest instantiation of this proof technique, since any learner for P/poly is automatically a self-learner. (Self-learning is a phenomenon that might be of independent interest, and we refer to Section 3 for further discussion of this concept.) The more general version of Theorem 2 presented in Section 3 shows that the same result holds for any self-learnable class that contains $\mathcal{TC}^0$ and is closed under composition. (For the interested reader, we mention that threshold gates are necessary to perform hardness amplification, a technical ingredient in our proof.)

Theorem 1 applies pseudo-random generators to the setting of learning. Our next result goes in the opposite direction, showing that derandomizing or pseudo-derandomizing learning algorithms has interesting consequences in the theory of pseudo-randomness. We say that a circuit is γ-*dense* if it accepts at least a γ-fraction of strings in $\{0,1\}^n$.

▶ **Theorem 3** (Hitting sets from deterministic and pseudo-deterministic learning)**.**
Let $\mathfrak{C} = \{\mathcal{C}_n\}$ *be an arbitrary circuit class, and assume that for every* $\varepsilon > 0$*,* $\mathfrak{C}$*-circuits of size* $s(n)$ *can be deterministically learned to accuracy* ε *in time* $T(n) \geq n$*.*

Then, for every $\gamma > 0$*, there exists a hitting set generator* $G_n \colon \{0,1\}^{\log T(n)} \to \{0,1\}^n$ *computable in time* $O(T(n))$ *against the class of* γ*-dense circuits in* $\mathcal{C}_n(s(n))$*. Similarly, if* $\mathfrak{C}$ *is pseudodeterministically learnable, there exist pseudodeterministic hitting set generators against* $\mathcal{C}_n(s(n))$ *with the same parameters.*

The proof of Theorem 3 is along the lines of the argument used to prove that a deterministic black-box PIT algorithm implies a hitting set. Suppose that there exists a deterministic learner. We run the deterministic learner with oracle the identically zero function, and output the set of queries it makes as our candidate hitting set. If the set of queries is not a hitting

[4] See also [33] for related results in the context of learnability using a linear combination of parity functions.

set, then there must be a somewhat dense function f computable in $\mathfrak{C}$ for which all the queries answer 0, just as they do for the identically zero function. But by the correctness of the learning algorithm, this would mean that f can be well-approximated by the identically zero function, which contradicts the assumption that it is somewhat dense. The consequence for pseudo-deterministic learners is shown by appropriately adapting this argument.

An interesting application of Theorem 3 is to the question of whether small hitting sets exist for $\mathcal{AC}^0[p]$ circuits. Despite much effort, no hitting sets even of sub-exponential size are known for such circuits (we refer to [5] for related results and discussion). Theorem 3 suggests an approach to this question via learning. Carmosino et al. [4] recently gave a quasi-polynomial time randomized learning algorithm for $\mathcal{AC}^0[p]$ – if this algorithm could be made deterministic, we would immediately get quasi-polynomial size hitting sets for $\mathcal{AC}^0[p]$ in quasi-polynomial time! In particular, that would imply that randomized poly-size $\mathcal{AC}^0[p]$ circuits with one-sided error can be simulated by deterministic quasi-poly size circuits. Even a pseudo-derandomization of the [4] algorithm would be interesting, as this would give somewhat efficient pseudo-deterministic hitting sets against $\mathcal{AC}^0[p]$, which is also unknown.

Theorem 3 also has consequences for non-uniform circuit lower bounds that can be derived from learning algorithms. It is known that *non-trivial* learning algorithms (i.e., those running in time $2^n/n^{\omega(1)}$) for a circuit class $\mathfrak{C}$ yield lower bounds against $\mathfrak{C}$ (cf. [23, 28, 6, 15, 36]). However, different algorithms provide different types of lower bounds. For a deterministic learner, one obtains a function in E that is hard almost everywhere [23], while for randomized learners, the hard function lives in BPE and is only hard infinitely often [28]. Interestingly, it is possible to use Theorem 3 to get something stronger from non-trivial *pseudo-deterministic* (randomized) learning algorithms: they can be used to define a function in BPE that is hard almost-everywhere for $\mathfrak{C}$.

1.3 Pseudoderandomization and approximation

We next turn our attention to a different setting, the setting of *approximation*. We are interested in integer-valued functions, i.e., functions from strings to non-negative integers, that have efficient randomized approximation schemes. The question is whether the existence of a good randomized approximation scheme generically implies the existence of a somewhat efficient pseudo-deterministic approximation scheme. (Pseudo-deterministic approximation schemes are formalized in the natural way in Section A.3.)

Note that this setting too does not conform to the "search problem" paradigm. Given a value w, it might be hard to test if the value is close to the correct value, since the correct value might be very hard to compute. Indeed, in our results, we make no assumptions about the complexity of exact computation of the integer-valued function.

Our main result here is a generic pseudo-derandomization of randomized approximation schemes; however, this pseudo-derandomization is only guaranteed to work on infinitely many input lengths with high probability over any poly-time samplable distribution of inputs.

▶ **Theorem 4** (Unconditional pseudo-derandomization of approximation). *Let $f\colon \{0,1\}^* \to \mathbb{N}$ be any function with a polynomial-time randomized approximation scheme. Then for each polynomial-time samplable sequence $\mathfrak{D}$ of distributions and for each constant $\delta > 0$, f has a pseudo-deterministic approximation scheme for infinitely many n over $\mathfrak{D}$ running in time $O(2^{n^\delta})$.*

The main idea in the proof of Theorem 4 is to exploit the uniform hardness-randomness tradeoffs used in the generic pseudo-derandomization results of [29], but adapted to this new setting. The crucial point is: how do we test efficiently that a value w is a good approximation

to the correct value? We test this simply by running the randomized approximation scheme to produce a value w' and checking if w is close to w'. This is not a deterministic polynomial-time test or indeed a bounded-error probabilistic polynomial-time test; however, we can show that it is good enough for our purposes.

As a corollary of this result and [20], we get unconditionally that the $(0,1)$-Permanent has a pseudo-deterministic approximation scheme running in sub-exponential time on infinitely many input lengths over any poly-time samplable distribution on inputs.

Finally, we consider a notion of *approximate canonization* of circuits. Canonization is a natural notion for an equivalence relation, where for each element of the set we compute a representative member of its equivalence class. Needless to say, canonization and canonical forms are fundamental notions with a variety of applications both in mathematics and computer science. We are interested in the natural equivalence relation between circuits: two circuits are equivalent if they compute the same function.

It is not hard to prove that efficient canonization is impossible for even weak circuit classes such as DNFs, under standard complexity assumptions. Therefore we *relax* the notion of canonization. We still require the output of the canonizer to be the same for any two equivalent circuits, but this output need not be a circuit equivalent to the original circuit, instead it is allowed to be *close* to the original circuit over the uniform distribution on inputs.[5]

Inspired by an observation in [1], we show that efficient deterministic (resp. pseudo-deterministic) learning implies efficient deterministic (resp. pseudo-deterministic) approximate canonization. (We refer to Section A.3 for a precise definition of approximate canonization.) Using Theorem 1 and the learning algorithm in [4], we get quasi-polynomial time pseudo-deterministic approximate canonization for $\mathcal{AC}^0[p]$ circuits under a standard circuit lower bound assumption for BPE.

▶ **Theorem 5** (Approximate canonization for $\mathcal{AC}^0[p]$, Informal)**.**
Let $p \geq 2$ be a fixed prime. If BPE *requires circuits of size $2^{n^{\Omega(1)}}$ almost everywhere, then $\mathcal{AC}^0[p]$ circuits can be approximately canonized in pseudo-deterministic quasi-polynomial time.*

We leave as an open problem obtaining an unconditional version of this theorem. Another interesting research direction is the investigation of connections between approximate canonization and other meta-computational problems. In this sense, we mention that [3] provides evidence that expressive circuit classes do not admit approximate canonization. (In fact, even more relaxed notions of approximate canonization are conditionally ruled out by the results from [3], and we refer to their work for further details.)

1.4 Organization

We formally define our models and state some auxiliary results and definitions in Appendix A. Due to page constraints, the remaining of the paper discusses pseudo-deterministic learning only. Section 2 contains the proof of Theorem 1 and related results. A more general form of Theorem 2 is established in Section 3. For the proof of the other results and additional discussion, we refer to the full version of our paper.

[5] A form of approximate obfuscation is also investigated in [1], but their definition requires a much stronger correctness guarantee.

Algorithm A

Input: 1^n and oracle access to an unknown function $f \in \mathcal{C}_d(s(n))$.

1. Computes a multi-set $S_m \stackrel{\text{def}}{=} \{G_m(a) \mid a \in \{0,1\}^{\ell(m)}\}$ of m-bit strings (with multiplicities), where G_m is the pseudorandom generator with parameters as in the statement of the lemma.
2. For each $w \in S_m$, simulates D_n with oracle access to f and with its random input set to w. Let $h_w \stackrel{\text{def}}{=} D_n^f(w)$ be the hypothesis output by the learning circuit under f and w.
3. Outputs the description of a circuit $\widetilde{C}_f$ that on an input $x \in \{0,1\}^n$ computes the majority function over the multi-set $\{h_w(x) \mid w \in S_m\}$.

2 Pseudo-derandomization for randomized learning algorithms

In this section we consider the derandomization and pseudoderandomization of learning algorithms via pseudorandom generators and pseudodeterministic pseudorandom generators, respectively. For simplicity, we will mostly focus on self-learnable circuit classes, but our results can be extended to more general settings, as explained later in this section.

2.1 Derandomizing from a pseudorandom generator

We start with a technical lemma showing that standard pseudorandom generators can be used to derandomize learning algorithms.

▶ **Lemma 6** (PRG-based derandomization of learning algorithms). *Let $\mathfrak{C}$ be a circuit class closed under composition. Let $s, s' \colon \mathbb{N} \to \mathbb{N}$ be functions, where $s'(n) \geq n$. Further, let $\varepsilon, \delta > 0$ be real-valued parameters satisfying $\delta \leq \varepsilon \leq 1/100$ and possibly depending on n. Finally, assume that for each $n \geq 1$ the depth-d class $\mathcal{C}_d(s(n))$ can be (ε, δ)-learned by a (randomized) oracle $\mathcal{C}_{d'}(s'(n))$-circuit.*

There are constants $e = O(d \cdot d')$ and $k \geq 1$ for which the following holds. If there is a family of quick pseudorandom generators $G_m \colon \{0,1\}^{\ell(m)} \to \{0,1\}^m$ that ε-fool depth-e size-m $\mathfrak{C}$-circuits, for

$$m = O(s(n) \cdot s'(n) + s'(n)^k),$$

then $\mathcal{C}_d(s(n))$ can be deterministically learned to accuracy $\varepsilon' = 8\varepsilon$ in time at most $2^{O(\ell(m))} \cdot \mathsf{poly}(s'(n))$.[6]

Proof. Let $\{D_n\}_{n \geq 1}$ be the corresponding (uniform) sequence of learning circuits with fixed parameters ε and δ. We claim that the deterministic algorithm A learns every function $f \in \mathcal{C}_d(s(n))$ to accuracy ε' in time at most $2^{O(\ell(m))} \cdot \mathsf{poly}(s'(n))$.

Clearly, under our assumptions A is a deterministic algorithm that runs in time at most $2^{O(\ell(m))} \cdot \mathsf{poly}(s'(n))$. Suppose now that A fails to learn some function $f^\star \in \mathcal{C}_d[s(n)]$. In other words, the corresponding output hypothesis $\widetilde{C}_{f^\star}$ is not ε'-close to $f^\star$. We use this information to construct a *randomized* $\mathfrak{C}$-circuit B of size at most m and depth at most e that distinguishes the output of G_m from random with advantage at least ε. We then fix the randomness of B using a standard argument in order to obtain a deterministic distinguisher. This contradicts the pseudorandomness of G_m, completing the proof of the lemma.

[6] For unbounded-depth classes, the circuit depth parameters can be omitted from the statement.

Algorithm B

Input: $z \in \{0,1\}^m$ and $r \in \{0,1\}^n$.

1. Let $C_{f^\star}$ be a $\mathcal{C}_d(s(n))$-circuit that computes $f^\star$. B uses a prefix of z as the randomness of D_n, and simulates the oracle computation $D_n^{f^\star}(z)$ with $C_{f^\star}$ replacing its oracle gates.
2. Suppose h_z is the output hypothesis. B outputs 1 if and only if $h_z(r) = C_{f^\star}(r)$.

In the description of B presented next, z is a candidate string (either produced from the generator, or uniformly random), and r is a fixed string sampled according to $\boldsymbol{r} \sim \mathcal{U}_n$, a random variable representing the randomness of the distinguisher.

Since $C_{f^\star}$ has depth $\leq d$ and size $\leq s(n)$, and D_n is an oracle circuit of depth $\leq d'$ and size $\leq s'(n)$, Step 1 can be implemented by a $\mathfrak{C}$-circuit of depth at most $d \cdot d'$ and of size at most $s(n) \cdot s'(n)$. By definition, the output hypothesis of D_n is restricted to circuits in $\mathcal{C}_{d'}(s'(n))$, and h_z is an effective description of a $\mathfrak{C}$-circuit. Consequently, the evaluation $h_z(r)$ in Step 2 can be computed by a $\mathfrak{C}$-circuit of depth $O(d')$ and size $\mathsf{poly}(s'(n))$. It follows that Step 2 can be implemented by a $\mathfrak{C}$-circuit of depth no more than $O(d' + d)$ and of size no more than $O(s(n) + \mathsf{poly}(s'(n)))$. Overall, we get that B is a (randomized) $\mathfrak{C}$-circuit of depth at most e and of size at most m, where these parameters are as in the statement of the lemma.

We argue in what follows that

$$\left| \Pr_{\boldsymbol{x} \sim \mathcal{U}_m, \boldsymbol{r} \sim \mathcal{U}_n} [B(\boldsymbol{x}, \boldsymbol{r}) = 1] \; - \Pr_{\boldsymbol{y} \sim \mathcal{U}_{\ell(m)}, \boldsymbol{r} \sim \mathcal{U}_n} [B(G_m(\boldsymbol{y}), \boldsymbol{r}) = 1] \right| \; > \; \varepsilon. \tag{1}$$

Observe that this implies in particular that for some fixed choice of $r \in \{0,1\}^n$, $B_r \stackrel{\text{def}}{=} B(\cdot, r)$ is a *deterministic* $\mathfrak{C}$-circuit of no larger complexity that distinguishes $\mathcal{U}_m$ and $G_m(\mathcal{U}_{\ell(m)})$ with advantage at least ε, which completes the proof.

Consider the leftmost probability in Equation 1. Since D_n learns every $f \in \mathcal{C}_d(s(n))$ to accuracy ε and with confidence parameter δ and $C_{f^\star} \equiv f^\star$, with probability at least $1 - \delta$ over $\boldsymbol{x}$, Step 2 of circuit B computes a hypothesis $h_{\boldsymbol{x}}$ that is ε-close to $f^\star$. For each fixed $x \in \{0,1\}^m$ that produces an ε-close h_x, in Step 2 circuit B accepts the input pair (x, r) with probability at least $1 - \varepsilon$ over the choice of $r \sim \boldsymbol{r}$. Consequently, using that $\delta \leq \varepsilon$, the leftmost probability in Equation 1 is at least $(1 - \delta)(1 - \varepsilon) \geq 1 - 2\varepsilon$.

It remains to upper bound the rightmost probability. Because A fails to learn $f^\star$ to accuracy ε', there is a set $T \subseteq \{0,1\}^n$ of measure at least ε' such that on every $x \in T$, $\widetilde{C}_{f^\star}(x) \neq f^\star(x)$. Consequently, for $x \in T$ at least half of the values $h_w(x)$ generated in Step 3 of A's description do not agree with $f^\star(x)$. It follows that over the choice of $\boldsymbol{y}$ and $\boldsymbol{r}$, $B(G_m(\boldsymbol{y}), \boldsymbol{r})$ rejects with probability at least $\varepsilon'/2 = 4\varepsilon$. Consequently, the rightmost probability $\leq 1 - 4\varepsilon$.

It follows from these estimates that the distinguishing probability in Equation 1 is strictly larger than ε, from which the result follows. ◀

It is important in the preceding argument for the distribution employed in the derandomization to be pseudorandom against *non-uniform* $\mathfrak{C}$-circuits.[7] First, this allows us to disregard the complexity of uniformly generating D_n in the proof that B is an appropriate distinguisher. Most importantly, we have no control over the "bad" function $f^\star$ where the derandomization

[7] Distributions that are pseudorandom against uniform algorithms were crucially employed in the pseudodeterministic construction of primes from [29].

might fail, and consequently $C_{f^\star}$ appears as a non-uniform advice in the proof of Lemma 6. Finally, r is also fixed non-uniformly when derandomizing the distinguisher.

2.2 Pseudoderandomization of learning algorithms

Similarly to Lemma 6, we now show that self-learning classes admit pseudodeterministic learners under the existence of suitable pseudodeterministic pseudorandom generators.

▶ **Lemma 7** (Pseudoderandomization via pseudodeterministic PRGs). *Let $\mathfrak{C}$ be a circuit class closed under composition. Let $s, s' \colon \mathbb{N} \to \mathbb{N}$ be functions, where $s'(n) \geq n$. Further, let $\varepsilon, \delta, \mu > 0$ be real-valued parameters satisfying $\delta \leq \varepsilon \leq 1/100$ and possibly depending on n. Finally, assume that for each $n \geq 1$ the depth-d class $\mathcal{C}_d(s(n))$ can be (ε, δ)-learned by a (randomized) oracle $\mathcal{C}_{d'}(s'(n))$-circuit.*

There are constants $e = O(d \cdot d')$ and $k \geq 1$ for which the following holds. If there is a family of quick μ-pseudodeterministic pseudorandom generators $G_m \colon \{0,1\}^{t(m)} \times \{0,1\}^{\ell(m)} \to \{0,1\}^m$ that ε-fool depth-e size-m $\mathfrak{C}$-circuits, for

$$m = O(s(n) \cdot s'(n) + s'(n)^k),$$

then $\mathcal{C}_d(s(n))$ can be $(8\varepsilon, \mu, \mu)$-pseudodeterministically learned in randomized time at most $2^{O(\ell(m))} \cdot \mathsf{poly}(s'(n))$.

Proof. We proceed as in the proof of Lemma 6, except that the corresponding derandomized algorithm A is replaced here by a pseudoderandomized algorithm A'. This procedure uses its random input $\boldsymbol{y} \in \{0,1\}^{t(m)}$ to define a candidate (deterministic) pseudorandom generator $G_m^{\boldsymbol{y}} \stackrel{\text{def}}{=} G_m(\boldsymbol{y}, \cdot) \colon \{0,1\}^{\ell(m)} \to \{0,1\}^m$. By assumption, it succeeds with probability at least $1 - \mu$, and whenever this happens, A' outputs a hypothesis $h_{\boldsymbol{y}}$ that is ε'-close to f, the unknown function, where $\varepsilon' = 8\varepsilon$ is as in Lemma 6. Consequently, A' is a (ε', μ)-learner for the class. Furthermore, with probability at least $1 - \mu$, A' constructs the same pseudorandom generator. Since the rest of its computation is deterministic, the corresponding learner will make a fixed set Q_f of queries, and generate a fixed output hypothesis h_f. This shows that A' is μ-pseudodeterministic. As the running time of A and A' are the same up to low order terms, it follows that $\mathcal{C}_d(s(n))$ can be $(8\varepsilon, \mu, \mu)$-pseudodeterministically learned in randomized time at most $2^{O(\ell(m))} \cdot \mathsf{poly}(s'(n))$. ◀

▶ Remark. As we alluded to before, Lemmas 6 and 7 hold in more generality provided that we have sufficiently strong pseudorandom generators. In particular, it is sufficient to have a generator that fools a circuit class closed under composition that is expressive enough to simulate circuits in the concept class, the learning circuit, and its hypothesis class. Consequently, existing learning algorithms can be derandomized under hardness assumptions.

▶ **Theorem 8** (Conditional learning derandomization). *Let $\mathfrak{C} \subseteq \mathsf{P/poly}$ be an arbitrary circuit class, and suppose $\mathfrak{C}(s(n))$ can be learned to any constant accuracy by a randomized algorithm running in time $t(n) \geq n$. If $\mathsf{E} = \mathsf{DTIME}[2^{O(n)}]$ requires circuits of size $2^{\Omega(n)}$ on all large input lengths, then there exists a constant $c \geq 1$ such that $\mathfrak{C}(s)$ can be deterministically learned to any constant accuracy in time at most $O(t(n)^c)$.*

Proof. Observe that a learning algorithm running in time $t(n)$ can be implemented by oracle circuits of size at most $\mathsf{poly}(t(n))$. The result is then a direct consequence of Lemma 6 and the hardness vs. randomness paradigm (Theorem 17). ◀

As a concrete example, Theorem 8 and Jackson's polynomial time learning algorithm for DNFs [19] immediately imply the following result.

▶ **Corollary 9.** *If* $\mathsf{E} = \mathsf{DTIME}[2^{O(n)}]$ *requires circuits of size* $2^{\Omega(n)}$ *on all large input lengths, then polynomial size* DNF*s can be learned to constant accuracy in deterministic polynomial time.*

The same approach provides pseudoderandomization via Lemma 7 using that a hard truth-table can be *pseudodeterministically* constructed from a weaker lower bound assumption.

▶ **Theorem 10** (Conditional learning pseudoderandomization)**.** *Let* $\mathfrak{C} \subseteq \mathsf{P/poly}$ *be an arbitrary circuit class, and suppose* $\mathfrak{C}(s(n))$ *can be learned to any constant accuracy by a randomized algorithm running in time* $t(n) \geq n$*. If* $\mathsf{BPE} = \mathsf{BPTIME}[2^n]$ *requires circuits of size* $2^{\Omega(n)}$ *on all large input lengths, then there exists a constant* $c \geq 1$ *such that* $\mathfrak{C}(s)$ *can be pseudodeterministically learned to any constant accuracy in time at most* $O(t(n)^c)$*.*

Proof. Note that, under this lower bound assumption, there exists a randomized algorithm that on input 1^ℓ, runs in time at most $2^{O(\ell)}$ and outputs with high probability the description of a fixed function $f_\ell \colon \{0,1\}^\ell \to \{0,1\}$ that requires circuits of size $2^{\Omega(\ell)}$. In other words, exponentially hard boolean functions can be pseudo-deterministically constructed in time polynomial in the size of their truth tables. The result now follows from the learning assumption, Theorem 17, and Lemma 7. ◀

For instance, thanks to the quasi-polynomial time randomized learning algorithm for $\mathcal{AC}^0[p]$ from [4], we get the following conditional result.

▶ **Corollary 11.** *If there is* $\gamma > 0$ *and a language in* $\mathsf{BPE} = \mathsf{BPTIME}[2^n]$ *that requires circuits of size* $\geq 2^{n^\gamma}$ *on all large input lengths, then* $\mathcal{AC}^0[p]$ *circuits can be learned to any constant accuracy in pseudodeterministic quasi-polynomial time.*

Proof. Simply observe that this lower bound is enough to get quasi-polynomial time (pseudo-deterministic) derandomizations using the hardness versus randomness paradigm (Theorem 17). The result follows as in the proof of Theorem 10 using the learning algorithm from [4]. ◀

3 Pseudodeterministic learners from randomized learners

Recall that $\mathcal{AC}^0$ circuits can be deterministically learned in quasi-polynomial time [33], and that $\mathcal{AC}^0[p]$ circuits are known to be learnable in randomized quasi-polynomial time [4]. In this section, we prove a general result showing that, for strong enough self-learnable circuit classes, any randomized learner running in quasi-polynomial time admits a non-trivial pseudoderandomization.

As opposed to the results discussed in Section 2, the next theorem is *unconditional* and does not assume the existence of pseudorandom generators. Theorem 2 is a particular case of this result.

▶ **Theorem 12** (Pseudodeterministic learners from randomized self-learners)**.** *Let* $\mathfrak{C}$ *be a circuit class that contains* $\mathcal{TC}^0$ *and is closed under compositions. Suppose that for every* $\delta, \varepsilon > 0$*,* $\mathfrak{C}(\mathsf{poly})$ *can be* (ε, δ)*-learned by (uniform)* $\mathfrak{C}$*-circuits of quasi-polynomial size. Then, for every* $\gamma > 0$ *and* $c \geq 1$*,* $\mathfrak{C}(\mathsf{poly})$ *can be* μ*-pseudodeterministically learned to accuracy* $\leq n^{-c}$ *on infinitely many input lengths by an algorithm running in time* $O(2^{n^\gamma})$*, where* $\mu = 2^{-n}$*.*

Proof. First, the assumption implies by a padding argument that for every $\varepsilon > 0$ and $k \geq 1$, there exists $k' \geq 1$ such that $\mathcal{C}(\exp((\log n)^k))$ can be learned to accuracy ε by a uniform

family $\mathfrak{D}^{(k)} = \{D_n^{(k)}\}_{n\geq 1}$ of $\mathfrak{C}$-circuits of size at most $\exp((\log n)^{k'})$.[8] We recall the following result from [23], which for convenience we state here as follows.

▶ **Lemma 13.** *There is a* PSPACE*-complete language L computable in linear space and a constant $b \geq 1$ such that the following holds. If $\mathfrak{C}(s(n))$ is learnable to error and accuracy $\leq n^{-b}$ in time at most $t(n) \geq n$, then either*

(i) *$L \notin \mathfrak{C}(s(n))$; or*

(ii) *$L \in \mathsf{BPTIME}[\mathsf{poly}(t(n))]$.*

By amplifying the success probability, we can assume that each family $\mathfrak{D}^{(k)}$ learns with confidence parameter $\delta \leq n^{-b}$, and by the result of [2], we can assume without loss of generality that the accuracy parameter is also $\leq n^{-b}$. This implies via Lemma 13 that either there exists no constant $a \geq 1$ such that $L \in \mathfrak{C}(\exp((\log n)^a))$, or for some $a' \geq 1$, we have $L \in \mathsf{BPTIME}[\exp((\log n)^{a'})]$. In the former case, since L is computable in linear space, we get that $\mathsf{BPE} \not\subseteq \mathfrak{C}[\exp((\log n)^{O(1)})]$. On the other hand, in the latter scenario, as $\mathsf{DSPACE}[s'(n)]$ can diagonalize against circuits of size $s'(n)^{\Omega(1)}$ (cf. [28, Corollary 39]), this fact together with a standard padding argument implies that $\mathsf{BPE} \not\subseteq \mathfrak{C}[\exp((\log n)^{O(1)})]$.

Let $f \in \mathsf{BPE}$ be a function that cannot be computed by quasi-polynomial size circuits from $\mathfrak{C}$ on infinitely many input lengths. We claim that the following result holds.

▶ **Lemma 14.** *Under our assumptions, there exists a function $f' \in \mathsf{BPTIME}[2^{O(n)}]$ such that for every constant $\beta \geq 1$, on infinitely many input lengths n, any $\mathfrak{C}$-circuit of size $\leq \exp((\log n)^{\beta})$ can compute f'_n with advantage at most $\exp((\log n)^{-\beta})$.*

Indeed, since $\mathcal{TC}^0 \subseteq \mathfrak{C}$, efficient worst-case to average-case reductions can be used to amplify the hardness of f (cf. [11, 14, 8]). In a bit more detail, a reduction of this form is well-known to hold for functions $f \in \mathsf{E} = \mathsf{DTIME}[2^{O(n)}]$. In order to amplify a function f in BPE, it is enough to observe that the entire truth-table of f can be computed in randomized time $2^{O(n)}$, except with negligible probability. Since the worst-case to average-case reduction acts on truth-tables, it defines with high probability a fixed function f' obtained from f.

Let $f' = \{f'_n\}_{n\geq 1}$ be given by Lemma 14, and E be a randomized algorithm running in time $2^{O(n)}$ that prints the truth-table of f'_n with probability at least $1 - 2^{-n}$. We use E together with the Nisan-Wigderson generator [26] to pseudodeterministically compute a generator against $\mathfrak{C}$. (While their result is stated with respect to general boolean circuits, it is well-known and easy to check that their construction works for any circuit class containing $\mathcal{TC}^0$.)

▶ **Theorem 15** (Corollary of Theorem 1 from [26])**.** *Let $m \leq t(m) \leq 2^m$, and suppose there is $h \in \mathsf{DTIME}[2^{O(m)}]$ such that, on infinitely many input lengths, every $\mathfrak{C}$-circuit D_m of size $\leq t(m)$ satisfies $\Pr_{\boldsymbol{x}}[D_m(\boldsymbol{x}) \neq h_m(\boldsymbol{x})] \geq 1/m$. Then there exists a constant $\lambda > 0$ and a quick pseudorandom generator $G\colon \{0,1\}^m \to \{0,1\}^{t(m^\lambda)}$ that $t(m^\lambda)$-fools $\mathcal{C}(t(m^\lambda))$-circuits on infinitely many input lengths.*

Using Theorem 15, it is possible to prove the following result.

▶ **Lemma 16.** *For every constants $c \geq 1$, $k \geq 1$, and $\gamma > 0$, there exists a function $G\colon \{0,1\}^* \times \{0,1\}^{\ell(n)} \to \{0,1\}^n$ that is a quick μ-pseudodeterministic generator that η-fools $\mathfrak{C}$-circuits of size $\leq \exp((\log n)^k)$ on infinitely many input lengths, where $\mu = 2^{-n}$, $\eta = n^{-c}$, and $\ell(n) = n^{\gamma}$.*

[8] See for instance the proof of Lemma 7 in [28].

Lemma 16 is established by a standard application of the Nisan-Wigerson generator to the family f', adapted to the pseudo-deterministic setting in the natural way.

Finally, using that $\mathfrak{C}$ is closed under composition, the existence of such generators immediately imply the statement of the theorem via an application of Lemma 7. ◀

Ideally, we would like to obtain a pseudodeterministic learner of comparable running time. However, this does not seem to be possible with these techniques. Consider for instance the more extreme case of designing a sub-exponential time pseudodeterministic learner from a sub-exponential time randomized learner. The main difficulty is that the lower bounds obtained from such a learner are not strong enough to derandomize an algorithm that runs in sub-exponential time.

Our techniques also require a strong assumption on the circuit class, namely, that it is closed under composition and able to compute threshold functions. Since there is evidence that circuit classes containing $\mathcal{TC}^0$ cannot be learned [25], it would be extremely interesting to obtain an analogue of Theorem 12 under weaker assumptions. In particular, one might be able to apply such a result to pseudoderandomize existing algorithms, such as [4].

Two remarks on the self-learnability of weak classes. These results further motivate the study of self-learning circuit classes, a direction that some might find of independent philosophical interest. In other words,

When is a circuit class $\mathfrak{C}$ learnable by algorithms that are no more powerful than $\mathfrak{C}$-circuits?

For very weak classes, this is probably impossible, given the very weak resources available to the learning algorithm, and the fact that a self-learner is in particular a proper learner. However, when $\mathfrak{C}$ becomes stronger, as in the extreme case where $\mathfrak{C} = \mathsf{P/poly}$, if learning algorithms exist then they are automatically proper learners.

It is possible to show that $\mathsf{MAJ} \circ \mathcal{AC}^0$ circuits are self-learnable by a uniform family of sub-exponential size circuits. This follows for instance from the results of [12], since the learning algorithm is based on the estimation of fourier coefficients of bounded size, and the corresponding parity computations can be simulated by randomized oracle $\mathcal{AC}^0$ circuits that output a sub-exponential size hypothesis in $\mathsf{MAJ} \circ \mathcal{AC}^0$.[9] Therefore, self-learnability is a phenomenon that is present even in constant-depth classes.

On the other hand, we do not know if $\mathcal{AC}^0$ is self-learnable by quasi-polynomial size $\mathcal{AC}^0$ circuits.[10] A natural approach here is to try to implement the LMN algorithm [24] using $\mathcal{AC}^0$ circuits, perhaps by replacing a threshold gate for an approximate majority, which is known to be computable in this class (see e.g. [35]). However, as we briefly explain next, this and other similar approaches cannot work. A self-learning algorithm of quasi-polynomial complexity for the class $\mathcal{AC}^0$ is required to output a hypothesis that is itself a quasi-polynomial size $\mathcal{AC}^0$ circuit. However, the approach in [24] is also able to learn functions that cannot be approximated by such circuits. This is an immediate consequence of [31, Theorem 3] using the connection between the influence of a boolean function and fourier concentration (cf. [27]).

[9] For the interested reader, we stress that during this implementation the empirical estimate of each fourier coefficient is not computed by the circuit, since $\mathcal{AC}^0$ circuits cannot count. The bits obtained from products of the form $\chi_S(x) \cdot f(x)$ are hard-coded directly into the final hypothesis, which can make use of a single threshold gate to compute the sign function.

[10] Note that, if this were the case, our techniques from Section 2 would provide an alternative, conceptually simpler proof of a result of [32] showing that such circuits can be learned in deterministic quasi-polynomial time.

References

1. Boaz Barak, Oded Goldreich, Russell Impagliazzo, Steven Rudich, Amit Sahai, Salil P. Vadhan, and Ke Yang. On the (im)possibility of obfuscating programs. *Journal of the ACM*, 59(2):6:1–6:48, 2012.
2. Dan Boneh and Richard J. Lipton. Amplification of weak learning under the uniform distribution. In *Conference on Computational Learning Theory (COLT)*, pages 347–351, 1993.
3. Zvika Brakerski, Christina Brzuska, and Nils Fleischhacker. On statistically secure obfuscation with approximate correctness. In *International Cryptology Conference (CRYPTO)*, pages 551–578, 2016.
4. Marco L. Carmosino, Russell Impagliazzo, Valentine Kabanets, and Antonina Kolokolova. Learning algorithms from natural proofs. In *Conference on Computational Complexity (CCC)*, pages 10:1–10:24, 2016.
5. Bill Fefferman, Ronen Shaltiel, Christopher Umans, and Emanuele Viola. On beating the hybrid argument. *Theory of Computing*, 9:809–843, 2013.
6. Lance Fortnow and Adam R. Klivans. Efficient learning algorithms yield circuit lower bounds. *Journal of Computer and System Sciences*, 75(1):27–36, 2009. doi:10.1016/j.jcss.2008.07.006.
7. Eran Gat and Shafi Goldwasser. Probabilistic search algorithms with unique answers and their cryptographic applications. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:136, 2011.
8. Oded Goldreich, Noam Nisan, and Avi Wigderson. On Yao's XOR lemma. In *Studies in Complexity and Cryptography*, pages 273–301. Springer, 2011. doi:10.1007/978-3-642-22670-0_23.
9. Shafi Goldwasser and Ofer Grossman. Bipartite perfect matching in pseudo-deterministic NC. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 87:1–87:13, 2017.
10. Shafi Goldwasser, Ofer Grossman, and Dhiraj Holden. Pseudo-deterministic proofs. *Electronic Colloquium on Computational Complexity (ECCC)*, 24:105, 2017.
11. Shafi Goldwasser, Dan Gutfreund, Alexander Healy, Tali Kaufman, and Guy N. Rothblum. Verifying and decoding in constant depth. In *Symposium on Theory of Computing (STOC)*, pages 440–449, 2007.
12. Parikshit Gopalan and Rocco A. Servedio. Learning and lower bounds for AC^0 with threshold gates. In *International Workshop on Approximation, Randomization, and Combinatorial Optimization (RANDOM-APPROX)*, pages 588–601, 2010.
13. Ofer Grossman. Finding primitive roots pseudo-deterministically. *Electronic Colloquium on Computational Complexity (ECCC)*, 22:207, 2015.
14. Dan Gutfreund and Guy N. Rothblum. The complexity of local list decoding. In *International Workshop on Approximation, Randomization and Combinatorial Optimization (RANDOM-APPROX)*, pages 455–468, 2008.
15. Ryan C. Harkins and John M. Hitchcock. Exact learning algorithms, betting games, and circuit lower bounds. *Transactions on Computation Theory (TOCT)*, 5(4):18:1–18:11, 2013. doi:10.1145/2539126.2539130.
16. Dhiraj Holden. A note on unconditional subexponential-time pseudo-deterministic algorithms for BPP search problems. *arXiv*, 2017. arXiv:1707.05808.
17. Russell Impagliazzo, Valentine Kabanets, and Avi Wigderson. In search of an easy witness: exponential time vs. probabilistic polynomial time. *Journal of Computer and System Sciences*, 65(4):672–694, 2002. doi:10.1016/S0022-0000(02)00024-7.

18 Russell Impagliazzo and Avi Wigderson. P = BPP if E requires exponential circuits: Derandomizing the XOR lemma. In *Symposium on the Theory of Computing (STOC)*, pages 220–229, 1997. doi:10.1145/258533.258590.

19 Jeffrey C. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55(3):414–440, 1997.

20 Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM*, 51(4):671–697, 2004.

21 Valentine Kabanets and Russell Impagliazzo. Derandomizing polynomial identity tests means proving circuit lower bounds. *Computational Complexity*, 13(1-2):1–46, 2004. doi:10.1007/s00037-004-0182-6.

22 Michael Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

23 Adam Klivans, Pravesh Kothari, and Igor C. Oliveira. Constructing hard functions using learning algorithms. In *Conference on Computational Complexity (CCC)*, pages 86–97, 2013.

24 Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

25 Moni Naor and Omer Reingold. Number-theoretic constructions of efficient pseudo-random functions. *Journal of the ACM*, 51(2):231–262, 2004.

26 Noam Nisan and Avi Wigderson. Hardness vs randomness. *Journal of Computer and System Sciences*, 49(2):149–167, 1994.

27 Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

28 Igor C. Oliveira and Rahul Santhanam. Conspiracies between learning algorithms, circuit lower bounds, and pseudorandomness. In *Computational Complexity Conference (CCC)*, pages 18:1–18:49, 2017.

29 Igor C. Oliveira and Rahul Santhanam. Pseudodeterministic constructions in subexponential time. In *Symposium on Theory of Computing (STOC)*, pages 665–677, 2017.

30 Alexander A. Razborov and Steven Rudich. Natural proofs. *Journal of Computer and System Sciences*, 55(1):24–35, 1997.

31 Benjamin Rossman, Rocco A. Servedio, and Li-Yang Tan. An average-case depth hierarchy theorem for boolean circuits. *CoRR*, abs/1504.03398, 2015.

32 Meera Sitharam. Pseudorandom generators and learning algorithms for AC^0. *Computational Complexity*, 5(3/4):248–266, 1995.

33 Meera Sitharam and Timothy Straney. Derandomized learning of boolean functions. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 100–115, 1997.

34 Christopher Umans. Pseudo-random generators for all hardnesses. *Journal of Computer and System Sciences*, 67(2):419–440, 2003.

35 Emanuele Viola. Randomness buys depth for approximate counting. *Computational Complexity*, 23(3):479–508, 2014.

36 Ilya Volkovich. On learning, lower bounds and (un)keeping promises. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 1027–1038, 2014. doi:10.1007/978-3-662-43948-7_85.

37 Ryan Williams. Improving exhaustive search implies superpolynomial lower bounds. *SIAM Journal on Computing*, 42(3):1218–1244, 2013. doi:10.1137/10080703X.

A Learning, Approximation, and Auxiliary Results

Let $\mathcal{F}_n$ be the set of all boolean functions $f\colon \{0,1\}^n \to \{0,1\}$ on n input variables, and $\mathfrak{F} = \bigcup_{n\geq 1} \mathcal{F}_n$ be the set of all boolean functions. We use boldface letters such as $\boldsymbol{w}$ and $\boldsymbol{x}$ to denote random variables. We say that boolean functions f and g from $\mathcal{F}_n$ are ε-close if $\Pr_{\boldsymbol{x}\sim\mathcal{U}_n}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})] \leq \varepsilon$, where $\mathcal{U}_n$ denotes the uniform distribution over $\{0,1\}^n$. We often view a string in $\{0,1\}^*$ that represents a boolean circuit D as if it were the actual circuit D, or the function that it computes.

A.1 Randomness, pseudorandomness and pseudodeterminism

We will require the notion of polynomial-time samplability of a sequence of distributions. Let $\mathfrak{D} = \{\mathcal{D}_n\}$ be a sequence of distributions, where each $\mathcal{D}_n$ is supported on $\{0,1\}^n$. We say that $\mathfrak{D}$ is polynomial-time samplable if there is a probabilistic polynomial-time algorithm B such that for each $n \in \mathbb{N}$ and each $y \in \{0,1\}^*$, $\Pr_{\boldsymbol{B}}[\boldsymbol{B}(1^n) = y] = \Pr[y \in \mathcal{D}_n]$.

We also require notions of pseudorandomness, introduced next.

Pseudorandom generators and hitting set generators. Let $\mathcal{D}_m$ be a probability distribution supported over $\{0,1\}^m$. We say that $\mathcal{D}_m$ is (η, s)-*pseudorandom* for a circuit class $\mathcal{C} \subseteq \mathcal{F}_m$ if for each size-s circuit $g \in \mathcal{C}(s)$,

$$\left| \Pr_{\boldsymbol{x}\sim\mathcal{U}_m}[g(\boldsymbol{x}) = 1] - \Pr_{\boldsymbol{y}\sim\mathcal{D}_m}[g(\boldsymbol{y}) = 1] \right| < \eta.$$

In other words, the circuit g is η-*fooled* by $\mathcal{D}_m$. This definition extends to an ensemble $\mathfrak{D} = \{\mathcal{D}_m\}_{m\geq 1}$ of distributions, by requiring the condition above to hold for every $m \geq 1$. Moreover, we say that a function $G_m\colon \{0,1\}^{\ell(m)} \to \{0,1\}^m$ is an η-*pseudorandom generator* for a class $\mathcal{C}$ if the induced distribution $G_m(\mathcal{U}_{\ell(m)})$ is (η, m)-pseudorandom for $\mathcal{C}$. Equivalently, the induced distribution η-fools every size-m $\mathcal{C}$-circuit over m-input variables. The function $\ell(m)$ computes the *seed length* of the generator G_m.

We also consider the weaker notion of hitting sets. We say that a set $\mathcal{H}_m \subseteq \{0,1\}^m$ is an (η, s)-*hitting set* for $\mathcal{C}$ if for each size-s circuit $g \in \mathcal{C}(s)$ such that $\Pr_{\boldsymbol{x}\sim\mathcal{U}_m}[g(\boldsymbol{x}) = 1] \geq \eta$, we have $g^{-1}(1) \cap \mathcal{H}_n \neq \emptyset$. This definition extends to ensembles of sets in the natural way. Similarly, a function $H_m\colon \{0,1\}^{\ell(m)} \to \{0,1\}^m$ is an η-*hitting set* for $\mathcal{C}$ if the induced set $H_m(\{0,1\}^{\ell(m)}) \subseteq \{0,1\}^m$ is an (η, m)-hitting set for $\mathcal{C}$. (Note that the support of a pseudorandom distribution is a hitting set with the same parameter η.)

We say that a pseudorandom generator or a hitting set generator is *quick* if it can be computed in time $2^{O(\ell(m))}$, where $\ell(m)$ is the corresponding seed length.

The following result will be useful.

▶ **Theorem 17** ([34]). *Given a function $f\colon \{0,1\}^{\log \ell} \to \{0,1\}$ of circuit complexity at least s, it is possible to construct a pseudorandom generator $G\colon \{0,1\}^{O(\log \ell)} \to \{0,1\}^m$ that $(1/m)$-fools size m circuits, where $m = s^{\Omega(1)}$. Moreover, G can be computed in time $\ell^{O(1)}$ given the description of the truth table of f.*

Pseudodeterministic pseudorandommness. We will make use of pseudorandom distributions $\mathcal{D}_m$ and hitting sets $\mathcal{H}_m$ that are constructed *pseudodeterministically*. For our purposes, we define the relevant concepts as follows. Let $G_m\colon \{0,1\}^{t(m)} \times \{0,1\}^{\ell(m)} \to \{0,1\}^m$. We say that G_m is a μ-pseudodeterministic (η, m)-pseudorandom generator for $\mathcal{C}$ if there is an (η, m)-pseudorandom generator $G^\star_m\colon \{0,1\}^{\ell(m)} \to \{0,1\}^m$ for $\mathcal{C}$ such that

$$\Pr_{\boldsymbol{a}\sim\mathcal{U}_t}[G(\boldsymbol{a}, \cdot) \equiv G^\star_m(\cdot)] \;\geq\; 1 - \mu,$$

where the "≡" symbol represents identity among functions. A μ-pseudodeterministic hitting set generator $H_m\colon \{0,1\}^{t(m)} \times \{0,1\}^{\ell(m)} \to \{0,1\}^m$ is defined analogously. Analogously, we say that a pseudodeterministic pseudorandom or hitting set generator is *quick* if it can be computed in time $2^{O(\ell(m))}$, where $\ell(m)$ is the seed length.

Note that the pseudodeterministic parameter μ of a quick pseudorandom or hitting set generator can be boosted by standard techniques. Indeed, since quick generators can tolerate a running time overhead of $2^{O(\ell(n))}$, one can always design a new generator that uses a larger random string, samples independent copies of the initial pseudodeterministic generator, and behaves as the most common generator among the induced generators provided by these samples.

A.2 Learning

Learning algorithms. We consider randomized learning algorithms under the uniform distribution that can make membership queries to the unknown function. We formalize such algorithms next.

Fix a class of functions $\mathfrak{C} \subseteq \mathfrak{F}$, often referred to as the *concept class*. For convenience, we write $\mathfrak{C} = \{\mathcal{C}_n\}_{n\geq 1}$, where $\mathcal{C}_n \subseteq \mathcal{F}_n$. A randomized algorithm A (ε,δ)-*learns* a class $\mathfrak{C}$ if for every $n \geq 1$ and for each $f \in \mathcal{C}_n$, when given oracle access to f and access to inputs 1^n, $\varepsilon > 0$ (accuracy), and $\delta > 0$ (confidence), A outputs the description of a boolean circuit D such that

$$\Pr_{\boldsymbol{w}}[\boldsymbol{D} = A^f(1^n,\varepsilon,\delta,\boldsymbol{w}) \text{ is } \varepsilon\text{-close to } f\,] \geq 1-\delta.$$

Here $\boldsymbol{w} \in \{0,1\}^*$ is a uniformly random boolean string representing the randomness of A, and $\boldsymbol{D} = A^f(1^n, w, \varepsilon, \delta)$ is a random variable denoting the (representation of the) circuit output by A over these inputs and with oracle access to f. For convenience, we might omit some input parameters when discussing the computation of A.[11]

While many of our results hold in a more general setting, for simplicity we will focus on the learnability of classes of boolean circuits. Therefore, $\mathcal{C}_n$ will always denote a class of the form $\mathcal{C}(s(n))$, where $\mathcal{C} \in \{\mathcal{AC}^0, \mathcal{AC}^0[p], \mathcal{TC}^0, \text{etc.}\}$, and $s(n)$ is an upper bound on circuit size complexity. When there is no risk of confusion, we might write $\mathcal{C}_d$ to restrict the class to circuits of depth at most d. The worst-case running time of the learning algorithm A over the choice of $f \in \mathcal{C}(s(n))$ and of its internal random string w is measured by the function $t_A(n, s, 1/\varepsilon, 1/\delta)$.

Pseudodeterministic learning. A randomized algorithm A $(\varepsilon,\delta,\gamma)$-*pseudodeterministically learns* a class $\mathfrak{C} = \{\mathcal{C}_n\}$ if A (ε,δ)-learns this class, and moreover for every $n \geq 1$ and $f \in \mathcal{C}_n$ there is a fixed set of queries $Q_f \subseteq \{0,1\}^n$ and a fixed string D_f representing a boolean circuit such that

$$\Pr_{\boldsymbol{w}}[A^f(1^n,\boldsymbol{w}) \text{ queries } f \text{ exactly over } Q_f \text{ and } A^f(1^n,\boldsymbol{w}) = D_f] \geq 1-\gamma.$$

In other words, with high probability the learner makes the same set of queries and outputs the same boolean circuit (representation) as its hypothesis. We say in this case that A is a γ-*pseudodeterministic* learning algorithm.

[11] It is well-known that the confidence parameter δ can be made arbitrarily small (cf.[22]). It is also known how to boost the accuracy parameter ε if the concept class satisfies a certain closure property (see e.g. [2]).

We would like to stress that it makes sense to consider variants of this notion where only the set of queries is pseudodeterministic (*query-pseudodet. learner*), or where only the output hypothesis is pseudodeterministic (*hypothesis-pseudodet. learner*). For instance, if A is a pseudodeterministic learner, running several independent copies of A and outputting the most common hypothesis will boost the initial hypothesis-pseudodeterminism parameter, but the resulting learner will be no longer query-pseudodeterministic.

The circuit complexity of learning algorithms. It is crucial in our investigations to consider a notion of complexity for learning algorithms that is more refined than running time. We measure instead the *circuit complexity* of learning algorithms. In other words, we specify a learning algorithm by a sequence $\{D_n\}_{n\geq 1}$ of *multi-output oracle* circuits D_n that have access to $\boldsymbol{w}$, ε, and δ, and whose *oracle queries* are answered according to the unknown function $f \in \mathcal{C}_n$. (In particular, the main input string of the oracle circuit is the random string, and in many cases we will fix ε and δ in advance.) The output bits of D_n encode a circuit describing the output hypothesis.[12]

In the case of learning circuits that are less powerful than general circuits (such as $\mathcal{AC}^0$, $\mathcal{TC}^0$, etc.), we further restrict the output hypothesis of the learner. We say that a class $\mathfrak{C}$ is learnable by $\mathfrak{D}$-circuits if the sequence $\{D_n\}$ consists of circuits from $\mathfrak{D}$, and moreover the output string is an *effective encoding* of a $\mathfrak{D}$-circuit. The meaning of effective description is that it should be possible for $\mathfrak{D}$-circuits to interpret the output string as the description of a $\mathfrak{D}$-circuit, and to efficiently evaluate computations given this description. We will not be explicit about such encodings, and simply note that they exist for the typical circuit classes investigated in our work.[13]

For definiteness, we briefly discuss a notion of *uniformity* for such sequence of learning circuits. In *learning upper bounds*, we assume that the sequence can be generated from 1^n by a deterministic algorithm that runs in time polynomial in the size of the circuits. We will not discuss *circuit lower bounds for learning* in this paper, but in such a context it is also natural to consider *non-uniform* sequences of learning circuits (see e.g. [28, Section 4]).

We say that a circuit class $\mathfrak{C}$ is *self-learnable* if it can be learned by a sequence of $\mathfrak{C}$-circuits, typically of quasi-polynomial size. This is an informal working definition, since for instance we do not specify the dependence on the parameters ε and δ. We will leave such details to the formal statement of our results, where we will often assume that these learning parameters are sufficiently small constants, and allow the class $\mathcal{C}_d(n^k)$ to be learned by $\mathcal{C}_{d'}(n^{(\log n)^{k'}})$-circuits (multi-output and with oracle gates).

We assume that functions related to algorithmic parameters such as time bounds, circuit size, learning accuracy, etc. are sufficiently constructive, in the sense that they do not affect the asymptotic complexity of our reductions whenever an algorithm needs to compute one of these functions.

A.3 Approximation

Approximation schemes. We define notions of approximation for computing integer-valued functions. An *integer-valued function* is a function from strings to non-negative integers,

[12] In the case of *deterministic* learning circuits, we remark that each D_n has access to the constant input bits 0 and 1, and one can think of its "input string" as the first batch of answers provided by the oracle queries.

[13] For bounded-depth circuit classes, we tolerate a constant-factor depth blow-up during the evaluation if this is necessary from the choice of encoding.

i.e., from $\{0,1\}^*$ to $\mathbb{N}$. We say that an integer-valued function f has a *polynomial-time randomized approximation scheme* (PRAS) if for each rational number $\epsilon > 0$ there is a probabilistic polynomial-time machine M, which given any string x as input, outputs an integer $M(x)$ (which might depend on the random choices of M) such that with probability $1 - 2^{-\Omega(|x|)}$ over the random choices of M, we have that $(1-\epsilon)f(x) \leq M(x) \leq (1+\epsilon)f(x)$. We say that f has an *fully polynomial-time randomized approximation scheme* (FPRAS) if there is a probabilistic machine M, which given a string x and a rational number ϵ (in some prespecified format) as input, runs in time $\text{poly}(|x|, 1/\epsilon)$ and outputs a number $M(x)$ such that with probability $1 - 2^{-\Omega(|x|)}$ over the random choices of M, we have that $(1-\epsilon)f(x) \leq M(x) \leq (1+\epsilon)f(x)$.

An example of an integer-valued function is the permanent of a $(0,1)$-matrix, when the matrix is represented as a bitstring. By the celebrated result of Jerrum, Sinclair and Vigoda [20], this integer-valued function has an FPRAS.

We will be interested in converting randomized approximation schemes to *pseudo-deterministic* ones, where with high probability the algorithm outputs a fixed number that is a good approximation to the correct value. Given a time function $T\colon \mathbb{N} \to \mathbb{N}$, we say that an integer-valued function f has a *pseudo-deterministic approximation scheme* (PDAS) running in time T if for each rational number $\epsilon > 0$ there is a function $g\colon \{0,1\}^* \to \mathbb{N}$ and a probabilistic machine M, which given a string x as input, runs in time $T(|x|)$ and outputs $g(x)$ with probability $1-2^{-\Omega(|x|)}$, and moreover we have that $(1-\epsilon)f(x) \leq g(x) \leq (1+\epsilon)f(x)$. A PDAS running in polynomial time is called a PPDAS. We say that an integer-valued function f has a *fully pseudo-deterministic approximation scheme* (FPDAS) running in time T if there is a function $g\colon \{0,1\}^* \times \mathbb{Q}^+ \to \mathbb{N}$ and a probabilistic machine M, which given a string x and a rational number ϵ (in some prespecified format) as input, runs in time $T(|x|, 1/\epsilon)$ and outputs $g(x,\epsilon)$ with probability $1 - 2^{-\Omega(|x|)}$, and moreover we have that $(1-\epsilon)f(x) \leq g(x,\epsilon) \leq (1+\epsilon)f(x)$. An FPDAS running in polynomial time is called a PFPDAS.

Note that a *d*eterministic approximation scheme running in time T is a special case of a PDAS running in time T where the machine M uses no randomness, and similarly a *fully deterministic approximation scheme* running in time T is a special case of an FPDAS running in time T where the machine M uses no randomness.

We also need more relaxed notions of pseudo-deterministic approximation schemes which are not guaranteed to work for all inputs. An *infinitely-often pseudo-deterministic approximation scheme* (i.o.PDAS) is only guaranteed to be pseudo-deterministic and output a correct approximation for infinitely many input lengths (rather than all of them). The notion of *infinitely-often fully pseudo-deterministic approximation scheme* (i.o.FPDAS) is defined analogously.

Finally, given a samplable distribution $\mathfrak{D} = \{\mathcal{D}_n\}$, an i.o.PDAS over $\mathfrak{D}$ is only guaranteed to be pseudo-deterministic and output a correct approximation with probability $1 - 1/n^{\omega(1)}$ over inputs sampled according to $\mathcal{D}_n$, for infinitely many n. Again, the notion of i.o.FPDAS over $\mathfrak{D}$ is defined analogously.

Canonization and approximate canonization. Next we define notions of *canonization* and *approximate canonization* for circuit classes. Let $\mathfrak{C}$ be a circuit class and $s\colon \mathbb{N} \to \mathbb{N}$ be a size function. Given a time function $T\colon \mathbb{N} \to \mathbb{N}$, we say that $\mathfrak{C}(s(n))$ has deterministic (resp. pseudo-deterministic) canonization in time T if there is a deterministic (resp. 1/3-pseudo-deterministic) Turing machine M such that (i) M operates in time $T(n)$ when given as input any circuit C in $\mathfrak{C}(s(n))$, (ii) for any circuit C in $\mathfrak{C}(s(n))$, $M(C)$ is a Boolean circuit

on n variables that is *equivalent* to C, i.e., computes the same Boolean function as C, and (*iii*) for any two equivalent circuits C and C' in $\mathfrak{C}(s(n))$, $M(C) = M(C')$. Note that a pseudo-deterministic canonization algorithm is allowed to output an arbitrary circuit with probability at most $1/3$.

We relax the notion of canonization by requiring the output to only be *approximately* equivalent to the input. Given a parameter $\epsilon > 0$, we say that $\mathfrak{C}(s(n))$ has deterministic (resp. pseudo-deterministic) ϵ-approximate canonization in time T if there is a deterministic (resp. 1/3-pseudo-deterministic) Turing machine M such that (*i*) M operates in time $T(n)$ when given as input any circuit C in $\mathfrak{C}(s(n))$, (*ii*) for any circuit C in $\mathfrak{C}(s(n))$, $M(C)$ is a Boolean circuit on n variables that is an ϵ-approximation to C, i.e., disagrees with C on at most an ϵ-fraction of inputs of length n, and (*iii*) for any two equivalent circuits C and C' in $\mathfrak{C}(s(n))$, $M(C) = M(C')$.

Finally, we stress that in the definition of canonization and approximate canonization it is important that the size bound $s(n)$ is fixed before the formalization of the problem. Indeed, by using an alternative definition that simply postulates that on an arbitrary circuit C from the class $\mathfrak{C}$ the machine M must output (say) in polynomial time on the description length of C an equivalent canonical circuit $M(C)$, one can easily use M to define a *natural property* useful against $\mathfrak{C}$ (in the sense of [30]).[14] However, as far as we know, there might be circuit classes that admit approximate canonization in the original sense introduced above, but do not admit natural properties. The first definition is therefore preferred.

[14] Given a truth-table f, construct an exponential size $\mathfrak{C}$-circuit C for f. Since M is a canonizer and has to run in polynomial time on every circuit D equivalent to C, one can infer from its output on C the approximate $\mathfrak{C}$-circuit complexity of f.