Dense model theorems require majority

Russell Impagliazzo* and Sam McGuire†

UC San Diego, Computer Science & Engineering

August 21, 2019

Abstract

Dense model theorems, originally introduced by Green and Tao in the context of additive number theory, can be loosely interpreted as follows: If μ is an arbitrary measure over a finite universe, then either there is a dense measure ν so that μ and ν are indistinguishable (with respect to some class of tests), or there's an explanation as to why no such ν could exist. We observe two computational lower bounds on such statements.

1. There are measures μ so that any explanation of why μ fails to have a dense model can be used to compute majority. Specifically, we describe an AC^0-Turing reduction from black-box proofs of dense model theorems over $\{0,1\}^n$ with to computing majority on roughly n^c bits (with $c < 1$, depending on the parameters of the reduction) where the size of the circuit computing the reduction is inversely proportional to the distinguishing probability.
2. There are measures μ so that any dense model of μ must be correlated with a linear threshold function. The same example applies to the decomposition theorem of Trevisan, Tulsiani and Vadhan [TTV09].

As far as we know, these are the first computational lower bounds on dense model theorems to appear in the literature and both are essentially tight, as various authors have shown that thresholds suffice to both produce a model and to refute the existence of a model.

The proof of 1. is an adaptation of an approach due to Shaltiel and Viola [SV10] used in the context of hardness amplification and 2. is a simple Fourier-analytic argument. These also imply, by classical lower bounds in circuit complexity, exponential size lower bounds for AC^0 and $\mathsf{AC}^0[\oplus]$ circuits performing these two tasks. Furthermore, 1. can be used to provide simple proofs of lower bounds for boosting and hardcore lemmas.

*russell@cse.ucsd.edu

†shmcguir@eng.ucsd.edu

1 Introduction

For a finite universe U, a class $\mathcal{F}$ of bounded tests $f : U \to [0,1]$ and measure $\mu : U \to [0,1]$ over U, we say that another measure $\nu : U \to [0,1]$ is an ε-*model* of μ, for $0 < \varepsilon$, with respect to $\mathcal{F}$ if for any $f \in \mathcal{F}$,

$$|\mathbf{E}[f \cdot \nu] - \mathbf{E}[f \cdot \mu]| \leq \varepsilon;$$

that is, if μ and ν are ε-*indistinguishable* with respect to $\mathcal{F}$. Producing simple models of data turns out to be a central task in theoretical computer science and beyond. For example, the construction of pseudorandom generators can be understood as producing explicit, efficiently samplable models of the uniform distribution with respect to a class of computationally-bounded tests. Even in practice, modern machine learning employs the use of generative models to automatically produce simple models of complex data sets and the notion of distinguishability has been used to explain generalization in this setting [Risteki, Arora papers]. It therefore becomes natural to ask for generic conditions under which simple models of data exist.

Dense model theorems, originally developed in the context of additive number theory, provide a set of sufficient conditions under which μ has a model which is *dense* with respect to the underlying universe U and has enjoyed a variety of applications. The original application, for example, was the Green-Tao theorem on arithmetic progressions of primes [GT08]. There, it was used to enable an application of Szemerédi's theorem on a 'dense model' of the prime numbers with respect to tests encoding useful additive information. In computer science, it has been, for example, used in leakage-resilient cryptography [cite], differential privacy [cite] and connected to proofs of the hardcore lemma and weak graph regularity lemma [Rei+08].

To formally define such a result, we'll need a few more definitions. For a measure $\mu : U \to [0,1]$, we say that μ is δ-*dense* if $\mathbf{E}[\mu] = \delta$. Additionally, we say that μ is (δ, γ)-pseudodense with respect to $\mathcal{F}$ if

$$\mathbf{E}[f] > \delta \mathbf{E}[f \cdot \mu] - \gamma.$$

Then a dense model theorem can be defined as follows:

Definition 1. For a finite universe U and a bounded class of tests $\mathcal{F}$ of the form $f : U \to [0,1]$, any $\varepsilon, \delta > 0$, and $k = k(\varepsilon, \delta)$, $\gamma = \gamma(\varepsilon, \delta)$, an $(\varepsilon, \delta, \gamma, k)$-*dense model theorem for* $\mathcal{F}$ *over* U states the following:

Let $\mu : U \to [0,1]$ be a measure over U. Then exactly one of the following hold:

1. There's a δ-dense measure ν so that μ and ν are ε-indistinguishable by $\mathcal{F}$.
2. There are k tests $f_1, ..., f_k \in \mathcal{F}$ and a function $g : [0,1]^k \to [0,1]$ so that the map $x \mapsto g(f_1(x), ..., f_k(x))$ refutes the (δ, γ)-pseudo-density of μ.

We'll generally refer to the δ-dense measure ν as a *dense model* of ν and the function g as a *refuter of g's pseudo-density.*

There are essentially two parameter regimes for which dense model theorems are known. The first, appearing in the work of Green-Tao [GT08] and Tao-Ziegler [TZ08], has γ exponentially small in ε and δ but a computationally simple witness g. They use an iterative partitioning approach, similar in spirit to Szemerédi's proof of the graph regularity lemma.

Theorem 2 (Green-Tao-Ziegler)**.** For any finite universe U, any set of bounded functions $\mathcal{F}$ of the form $f : U \to [0,1]$ and any $\varepsilon, \delta > 0$, there's an $(\varepsilon, \delta, \gamma, k)$-dense model theorem for $\mathcal{F}$ over U with $k(\varepsilon, \delta) = \mathsf{poly}(1/\varepsilon, 1/\delta)$, $\gamma(\varepsilon, \delta) = \exp(-\mathsf{poly}(1/\varepsilon, 1/\delta))$, and $g(x_1, ..., x_k) = \prod_{i=1}^{k} x_i$.

The second setting, shown independently by Gowers [cite] and Reingold, Trevisan, Tulsiani and Vadhan [Rei+08], obtains a γ with polynomial dependence on ε and δ but uses a more complex function g. Proofs using linear programming duality or boosting are known.

Theorem 3 (RTTV, Gowers)**.** For any finite universe U, any set of bounded functions $\mathcal{F}$ of the form $f : U \to [0,1]$ and any $\varepsilon, \delta > 0$, there's an $(\varepsilon, \delta, \gamma, k)$-dense model theorem for $\mathcal{F}$ over U with $k(\varepsilon, \delta) = O(\log(1/\delta)/\varepsilon^2)$, $\gamma(\varepsilon, \delta) = O(\varepsilon\delta)$, and $g(x_1, ..., x_n) = \text{sign}(-t + \sum_{i=1}^{k} \pm x_i)$.[1]

In the latter case, it can be shown that when a dense model ν does indeed exist, it can be written in the form $\nu(x) = \sum_{i=1}^{k} w_i f_i(x)$ for weights $w_i \in \mathbb{R}$ and $f_i \in \mathcal{F}$.

We're now in a position to frame the question underlying this work — what does it mean for a dense model theorem to require majority? Answering this can be explained in terms of the following more qualitative description of a dense model theorem: either a distribution has a simple, dense model or there is an explanation as to why it doesn't. The two natural notions of computational complexity here are the the simplicity of the explanation as to why the input measure fails to have a dense model or the simplicity of the dense model itself. In both cases, we'll prove lower bounds that roughly match the known upper bounds.

Regarding the first notion of complexity, we borrow an approach from Shaltiel and Viola for proving complexity lower bounds on hardness amplification proofs, and show that any 'explanation' that the distribution of n independent biased coin tosses fails to be pseudo-dense can be used in a constant-depth oracle circuit computing majority. Specifically, let N_η be the product of n Bernoulli random variables $\mathsf{Ber}(p)$ with $p = 1/2 - \eta$. Then,

Theorem 4 (Informal)**.** Let $U = \{0,1\}^n$ and $\mathcal{F}$ be the set of functions $x \mapsto x_i$ for each $i \in [n]$. For any $\varepsilon, \delta, \gamma$ and $\eta \approx \eta$,

1. N_η is ε-distinguishable from every δ-dense measure by $\mathcal{F}$.
2. Any $g : \{0,1\}^n \to \{0,1\}$ which refutes the (δ, γ)-pseudo-density of N_η can be used in a constant-depth oracle circuit C of size $\mathsf{poly}(\delta^2/\eta^3\gamma^3)$ so that C computes the majority on inputs of length $O(1/\eta)$ where $\eta \approx \varepsilon$.

The first thing to note is the similarity to the result of Shaltiel and Viola [SV10]: they provide an AC^0-Turing reduction from majority to the coin problem — distinguishing between the biased distribution and the uniform distribution — and further show that (non-adaptive, later improved to adaptive in [GSV18]) black-box hardness amplification procedures produce algorithms for solving the coin problem. This similarity is of course no coincidence, given the relationship between the hardcore lemma (an instance of hardness amplification) and the dense model theorem. Indeed, that dense model theorems require majority (in terms of g's complexity) could be deduced as a corollary of their result if we had a reduction from hardcore lemmas to dense model theorems. While there is a reduction in the other direction — an optimal hardcore lemma implies a dense model theorem — the other direction remains unclear. We circumvent this by designing a reduction directly from dense model theorems to the coin problem.

Note also that the condition of booleanity on g is essentially without loss of generality: if we had a bounded test $g : [0,1]^k \to [0,1]$ which refutes the pseudo-density of a measure μ, we can convert g into a boolean-valued refuter $\tilde{g}$ so that

$$\mathbf{E}_x[g(x)] = \Pr_x[\tilde{g}(x) = 1].$$

[1] Formally, we mean that there is some $t \in \mathbb{Z}$, some sign pattern $b_1, ..., b_k \in \{\pm 1\}$ so that $\text{sign}(-t + \sum_{i=1}^{k} b_i x_i)$ satisfies the requirements of g.

To do so, let C be a randomized circuit that does the following: on input x, it computes (a dyadic approximation of) $g(x)$ and then flips a coin with bias $g(x)$ and outputs the result. Then the probability over C and x that $C(x) = 1$ is simply $\mathbf{E}_x[g(x)]$, with a small amount of error introduced by approximating $g(x)$ with finite precision. Non-uniformly, we can then average out the randomness to get a deterministic circuit. This conversion was observed before in [TTV09].

For the second notion of complexity, it is relatively straightforward to get some type of lower bound on the complexity of the model. For example, let $\mathcal{F}$ be the class of delta functions $\delta_x(y) = 1$ iff. $y = x$ for each $x \in \{0,1\}^n$. Then by taking μ to be the 1s of majority, any ε-approximation ν of μ with respect to $\mathcal{F}$ has the property that for all x, $|\nu(x) - \mu(x)| \leq \varepsilon$. Specifically, ν is an ε-approximation of μ in the ℓ_∞-norm. From here, it is known that any polynomial p which ε-approximates μ point-wise must have high degree (see, e.g., Paturi), and therefore ν must have high-degree. However, this approach doesn't explain the *structure* of g in a very useful way: even the g used in the Green-Tao proof has high approximate degree. Additionally, an ideal example would make the difference in the complexities of $\mathcal{F}$ and $g \circ \mathcal{F}$ as large as possible.

For a stronger example, we'll still choose μ to itself be a linear threshold function with integer weights but now take $\mathcal{F}$ be the set of functions $x \mapsto x_i$ for $i \in [n]$. A simple Fourier-analytic argument tells us that any model of μ must be correlated with μ, as $\mathcal{F}$, by definition, detects large-enough differences in the linear Fourier coefficients:

Theorem 5 (Informal)**.** Let $U = \{-1,1\}^n$ and $\mathcal{F}$ be the set of projections. Then for each $\delta > 0$ and ε sufficiently small, there's a measure $\mu : \{-1,1\}^n \to \{0,1\}$ so that

1. μ is a linear threshold function with integer weights. That is, $\mu(x) = \text{sign}(-\theta + \sum_{i=1}^m w_i x_i)$ for some parameters $\theta, w_1, ..., w_n \in \mathbb{Z}$.

2. μ is δ-dense (and therefore (γ, δ)-pseudodense with respect to $g \circ \mathcal{F}$ for any g).

3. If $\nu : \{-1,1\}^n \to \{0,1\}$ is a δ-dense ε-model of μ with respect to $\mathcal{F}$, then $Pr[\nu(x) \neq \mu(x)] \leq 1 - \min\{O(\varepsilon), O(\delta)\}$.

Indeed, work on the *Chow parameters problem* allows us to pick any linear threshold function as an example. Specifically, [cite, cite] show that if the linear Fourier coefficients of μ and ν are close where μ is linear threshold function, then μ and ν are actually close in ℓ_1-distance.

Trevisan, Tulsiani and Vadhan noticed that dense model theorems are a special case of a more general result, a type of *decomposition theorem*, which explains that any bounded function $f : U \to [0,1]$ can be approximated by a function g of 'constant' complexity, in the sense that f is indistinguishable from g by tests $\mathcal{F}$ and g built out of a small combination of functions in $\mathcal{F}$, the size of which is independent of $|U|$. Specifically,

Theorem 6 (Decomposition theorem, [TTV09])**.** Let U be a finite universe, $\mathcal{F}$ a set of bounded tests $f : U \to [0,1]$ and $\mu : U \to [0,1]$ a measure over U. Then there exists a measure ν with $\mathbf{E}\nu = \mathbf{E}\mu$ so that

1. ν can be built out of $k = \mathsf{poly}(1/\varepsilon)$ functions $f_1, ..., f_k \in \mathcal{F}$ using multiplication by a constant, sums, products and linear threshold functions.

2. μ and ν are ε-indistinguishable with respect to $\mathcal{F}$.

Informally, this says that simple tests can be fooled by distributions which are almost as simple, and the dense model theorem, the hardcore lemma and the weak graph regularity lemma all follow as corollaries. It is simple to see that our example also provides a lower bound on the complexity of the approximator ν, explaining that, in some cases, there is a necessary gap between the tests $\mathcal{F}$ and and distributions which fool them.

1.1 Preliminaries

Call a distribution X over $\{0,1\}^n$ δ-smooth if $\Pr[X = x] \leq (1/\delta) \cdot 2^{-n}$ for all $x \in \{0,1\}^n$. Note that this is equivalent to having min-entropy at most $n - \log(1/\delta)$. We first note the smooth distributions and dense measures are interchangable.

Lemma 7. A measure μ has density at least δ if and only if the distribution X_μ defined by

$$\Pr[X_\mu = x] = \frac{\mu(x)}{d(\mu)} \cdot 2^{-n}$$

is δ-smooth.

Proof. For the forward direction, note that $\max_x \Pr[X_\mu = x] = \frac{1}{d(\mu)2^n} \max_x \mu(x) \leq 1/d(\mu)2^n$, and so X_μ is δ-smooth when $d(\mu) \geq \delta$. In the other direction, since $\max_x \Pr[X_\mu = x] = \frac{1}{d(\mu)2^n} \max_x \mu(x) \leq \frac{1}{\delta 2^n}$ by assumption, we get $\delta \leq d(\mu)$. ■

The distribution X_μ can be sampled by rejection sampling according to μ. As the distinction will always be clear from context, we will hereafter abuse notation and avoid distinguishing between the measure μ and the distribution X_μ induced by μ.

Indeed note that f refuting the (δ, γ)-pseudo-density of N_η, in that $\Pr[f(U) = 1] \leq \delta \Pr[f(N_\eta)] - \gamma$, clearly implies that U and N_η are η-indistinguishable by f and in particular f solves the η-coin problem with advantage γ.

Let N_η denote the distribution on $\{0,1\}^n$ obtained by sampling n i.i.d. Bernoulli random variables $X_1, ..., X_n$ with $\Pr[X_i = 1] = 1/2 - \eta$ for each i.

The connection Fourier analysis is simple: for both claims, we'll be interested in the class of tests of the form $x \mapsto x_i$ which are simply the linear characters of $\mathbb{F}_2^n$. Specifically, the bias of the map $x \mapsto x_i$ when x is drawn from a measure μ is exactly the Fourier coefficients $\widehat{\mu}(i)$.

1.2 Proof overviews

1.2.1 Refuting pseudo-density requires majority

We begin with the simple observation that for any $1/2 > \eta > 0$ sufficiently large, the noise distribution N_η satisfies the promise of the dense model reduction in that N_η is distinguishable from dense distribution with significant advantage, so long as our test class contains monotone projections $x \mapsto x_i$ for each $i \in [n]$. This follows by a simple observation — made many times before — that any dense distribution over $\{0,1\}^n$ has a bit which is relatively unbiased. Then by the dense model theorem, we are left with a family of tests which refute the pseudo-density of N_η for any sufficiently large η, which we'll use to build our majority circuit.

To do so, we'll largely follow the reduction from majority to the coin problem from Shaltiel-Viola with a minor modification. Since any f solving (η, δ, γ)-coin-density problem also solves the η-coin problem with advantage γ, we might hope to employ the following construction from Shaltiel and Viola:

Lemma 8 ([SV10])**.** Suppose A solves the η-coin problem with advantage γ, where $\gamma \geq 1/\log(1/\eta)$. Then there's an AC^0 circuit with oracle access to R solving majority on inputs of length $1/\eta$.

The technicality that arises is the dependence on γ. In our context, we need to think of $\eta > \varepsilon$ and therefore γ will need to be exponentially larger than ε. However, a straightforward construction shows that the dense model theorem is actually *false* in this regime, as observed by Zhang in [Zha11].

Claim 9 ([Zha11])**.** There is no $(\varepsilon, \delta, \gamma)$-dense model reduction for any $\gamma > \varepsilon\delta$.

Proof. Let $\mu \subseteq U$ be a set of size $\delta(1-\varepsilon)|U|$ and let the class of tests solely consist of the characteristic function of μ, call it f. Then

1. There's no δ-dense ε-model of μ with respect to f. This is because $\Pr_T[f(x) = 1] \leq 1 - \varepsilon$, as any such δ-dense ν has to cover a $(1-\varepsilon)$-factor larger subset of the universe.

2. μ is $(\delta, \delta\varepsilon)$-pseudodense with respect to *any* test g

$$\Pr_U[g(x) = 1] = \delta(1-\varepsilon)\Pr_\mu[g(x) = 1] = \delta \Pr_\mu[g(x) = 1] - \delta\varepsilon \Pr_\mu[g(x) = 1] \geq \delta \Pr_S[g(x) = 1] - \delta\varepsilon$$

■

Fortunately, we have a stronger guarantee than distinguishing the noise distribution from uniform, as the acceptance probabilities are bounded away by a multiplicative constant δ.

We will use this stronger assumption roughly as follows: let $w < 1/2\eta$ and suppose that A_w solves the $(w\eta, \delta, \gamma)$-coin-density problem. We'll build a randomized algorithm $\mathbf{S}_w : \{0,1\}^{1/\eta} \to \{0,1\}$ so that for $\alpha := \Pr[\mathbf{S}_w(x) = 1|\ x \text{ has weight } 1/2\eta]$, $\beta := \Pr[\mathbf{S}(x) = 1|\ x \text{ has weight } w]$, we have $\alpha \leq \delta\beta - \gamma$. By independently sampling $\mathbf{g}_w$, we'll produce a test $\tilde{S}_w : \{0,1\}^n \to \mathbb{N}$ so that $S_w(x) \leq \delta\beta m$ when x has weight $1/2\eta$ and $S_w(x) \geq \beta m$ when x is balanced (for an appropriate choice of m). Thus, a *relative-error* approximate majority circuit, implementable in AC^0, can successfully distinguish these two cases.[2] From here, we can combine S_w for each $w < 1/2\eta$ into a simple depth-2 circuit, the result of which computes majority. This approach follows Shaltiel-Viola almost exactly, with the difference being the use of a relative-error approximate majority circuit as opposed to their use of an additive-error approximate majority circuit. Since they begin with only an additive distinguishing gap, they need an additive-error approximate majority circuit, which in turn produces the undesirable dependence on γ.[3]

1.2.2 Modeling sometimes requires thresholds

Let B be the Hamming ball centered on 1^n with density δ. Any δ-dense model ν of B with respect to bit queries has the property that $|\widehat{\nu}(i) - \widehat{B}(i)| \leq 2\varepsilon$, where $\widehat{\nu}(i)$ denotes the ith linear Fourier coefficient of ν.

We would like to show that this implies that ν and B are actually *correlated*. Note that this isn't tautological because having a *model* according to $\mathcal{F}$ does not imply correlation. Specifically, we could take f to be a random function and still model it with respect to constantly-many functions $\mathcal{F}$ combined with linear threshold functions (see [TTV09]) and yet they won't be correlated. Modeling simply means that $\mathcal{F}$ can't reliably detect the differences between f and its model.

A simple claim from Birkendorf et al. tells us that this is indeed that case when ε is sufficiently small:

Claim 10. Let $B : \{-1,1\}^n \to \{-1,1\}$ be a Hamming ball,

$$\text{sign}(\sum_{i=1}^{n} x_i - t\sqrt{n})$$

[2] We'll also need to ensure that α and β aren't too small, which is dealt with in the full proof.

[3] We note that [cite] provide a reduction from approximating majority (the correlation problem) to the coin problem with better parameters when γ is a constant. Since constant-probability approximations of majority have large $\mathbb{F}_2$-degree, this yields exponential-size lower bounds for $\mathsf{AC}^0[\oplus]$ circuits. The approach does not seem applicable here, however, where the dependence on γ becomes crucial.

and $\nu : \{-1,1\}^n \to [-1,1]$ an arbitrary bounded function. Then if $\mathbf{E}B = \mathbf{E}\nu$ and for each $i \in [n]$,

$$|\widehat{B}(i) - \widehat{\nu}(i)| \leq \varepsilon$$

then

$$\mathbf{E}[|B(x) - \nu(x)|] \leq \varepsilon n$$

Applying this directly gives the desired conclusion for small ε.

Note that this claim has no dependence δ, which in this example is roughly 2^{-t^2} (thinking of B as a set). For larger ε, we can make a similar argument where now the error depends on δ:

Claim 11. Let $\nu : \{-1,1\}^n \to \{-1,1\}$ and $\mathbf{E}[\nu] = \mathbf{E}[B] = \delta$. Suppose $|\widehat{B}(i) - \widehat{\nu}(i)| \leq 2\varepsilon$ for all $i \in [n]$ and

$$\varepsilon \leq \frac{1}{\sqrt{2n}}(\sqrt{\delta} - \sqrt{\delta^2 \log(1/\delta)}).$$

Then

$$\Pr[B(x) \neq \nu(x)] \leq \delta$$

2 Refuting pseudo-density requires majority

2.1 Dense measures have nearly-unbiased bits

Claim 12. For any δ-dense distribution ν, there's some index $i \in [n]$ so that $\Pr_\nu[x_i = 1] \leq 1/2 \pm \sqrt{\delta^2 \log(1/\delta)/2n}$.

Proof. By Chang's inequality,

$$\sum_{i \leq n} \widehat{\nu}(i)^2 \leq 2\delta^2 \log(1/\delta).$$

Averaging, there's some i so that $|\widehat{\nu}(i)| \leq \sqrt{2\delta^2 \log(1/\delta)/n}$. Note that

$$\widehat{\nu}(i) = \Pr_\nu[x_i = 1] - \Pr_\nu[x_i = -1] = 2\Pr_\nu[x_i = 1] - 1,$$

and hence $\Pr_\nu[x_i = 1] = 1/2 \pm \widehat{\nu}(i)/2$, which gives us the desired conclusion. ■

We can immediately conclude the following.

Corollary 13. For any δ-dense distribution ν with $\delta < 1$, there is a test T of the form $x \mapsto x_i$ which ε-distinguishes between N_η and ν for any $1/2 > \eta \geq \varepsilon + \sqrt{\delta^2 \log(1/\delta)/2n}$.

Proof. Let i be the index from the previous claim. Then $x \mapsto x_i$ ε-distinguishes between N_η and ν for $\eta \geq \varepsilon + \sqrt{\delta^2 \log(1/\delta)/2n}$ since

$$|\Pr_{x \sim \nu}[x_i = 1] - \Pr_{x \sim N_\eta}[x_i = 1]| \geq |(\frac{1}{2} - \sqrt{\delta^2 \log(1/\delta)/2n}) - (\frac{1}{2} - \varepsilon - \sqrt{\delta^2 \log(1/\delta)/2n})| = \varepsilon.$$

■

Let $\varepsilon, \delta, \gamma$, $\eta^* = \varepsilon + \sqrt{\delta^2 \log(1/\delta)/2n}$ be fixed. The previous claim says that if the $(\varepsilon, \delta, \gamma)$-dense model theorem over $\{0,1\}^n$ with tests containing monotone projections is true, then we can apply it to the noise distribution N_η for any $\eta \leq \eta^*$ and in turn produce an algorithm for the (η, δ, γ)-coin density problem. We'll now argue oracle access to these algorithms can be used in an AC^0 circuit computing majority.

2.2 Solving majority with the coin-density problem

Suppose we have an algorithm A for the (η, δ, γ)-coin density problem for every $\eta < \eta^*$ for some fixed η^*. We proceed in steps:

1. Let $w < 1/(2\eta^*)$. We'll produce a distribution $\mathbf{S}$ over functions $s : \{0,1\}^{1/\eta^*} \to \{0,1\}$ so that $\Pr[s(y) = 1] \leq \delta \Pr[s(z) = 1] - \gamma$ for every balanced $y \in \{0,1\}^{1/\eta^*}$ and every $z \in \{0,1\}^{1/\eta^*}$ with weight w.

2. Using an approximate majority construction, we will reduce the error of the previous algorithm. Specifically, we'll produce an algorithm S so that $S(y) = 0$ whenever y is balanced and $S(z) = 1$ whenever z has weight w.

3. Repeating this construction for each $w < 1/(2\eta^*)$ to obtain algorithms S_w as above, we can see that $C(x) = \wedge_w \neg S_w(x)$ is 1 when x is balanced and 0 when x has weight less than $1/(2\eta^*)$.

4. Finally, let x^i denote x with the first i bits of x set to 0. Then $C'(x) = \vee_{i \leq 1/(2\eta^*)} C(x^i)$ is 1 if there's an i so that x^i is balanced, which characterizes majority: if x is low weight, then all of its restrictions x^i are also low weight. If x has high weight, then there's a restriction x^i which is balanced.

Once again, this approach is identical to Shaltiel and Viola [SV10], except for our error reduction step. The following simple claim from [SV10] verifies the correctness of the construction conditioned on steps 1 and 2 being correct.

Claim 14 ([SV10])**.** Suppose for each $w < 1/(2\eta^*)$, $S_w : \{0,1\}^{1/\eta^*} \to \{0,1\}$ has the property that $S_w(x) = 0$ when y is balanced and $S'(z) = 1$ whenever z has weight w. Then the depth-two oracle circuit C (with oracle access to S'_w for each $w < 1/(2\eta^*)$) of size $1/(2\eta^*)^2$ with $1/(2\eta^*)$ oracle calls defined by

$$C : x \mapsto \bigvee_{i \leq 1/(2\eta^*)} \left(\bigwedge_{w < 1/(2\eta^*)} \neg S_w(x) \right)$$

computes majority.

We'll now proceed to establishing the first two steps. First we have a randomized reduction from the coin-density problem to the promise problem of distinguishing between balanced and unbalanced strings.

Claim 15 ([SV10])**.** Let $1 \leq w < 1/(2\eta^*)$ and suppose A solves the $(1/2 - w\eta^*, \delta, \gamma)$-coin density problem (so the bias of each coin in the noisy distribution is w/η^*). Then there's a randomized test $\mathbf{S} : \{0,1\}^{1/\eta^*} \to \{0,1\}$ with the following property: Let $\alpha = \Pr[\mathbf{S}(x) = 1 \mid x \text{ is balanced}]$ and let $\beta = \Pr[\mathbf{S}(x) = 1 \mid x \text{ has weight } w]$. Then $\alpha \leq \delta\beta - \gamma$.

Proof. Let $\mathbf{i}_1, ..., \mathbf{i}_n$ denote independent random variables which are uniform over the co-ordinates $[1/\eta^*]$. Then define $\mathbf{S}(x) \equiv A(x_{\mathbf{i}_1}, ..., x_{\mathbf{i}_n})$. For balanced y, $y_{\mathbf{i}_1}, ..., y_{\mathbf{i}_n}$ is distributed as N_0, and for z with weight w, $z_{\mathbf{i}_1}, ..., z_{\mathbf{i}_n}$ is distributed as $N_{1/2 - w\eta^*}$. The conclusion follows since A solves the $(1/2 - w\eta^*, \delta, \gamma)$-coin problem. ■

Note that since $w\eta^* \leq 1/2 - \eta^*$ and we have solutions to the coin problem for each $\eta < \eta^*$, the assumption from the previous claim is satisfied for any choice of $1 \leq w < 1/(2\eta^*)$. This step is what restricts the input size to be $1/\eta^*$. More generally, for a fixed length ℓ, we could pick any weights w so that $w/l \leq 1/2 - \eta^*$ i.e. so that difference in weights is at least η^*. This approach is outlined in [Srikanth's coin problem paper] by picking $\ell \approx (1/\eta^*)^2$ meaning that w can range up to roughly $\ell/2 - \sqrt{\ell}$. This can

be used to *approximate* majority (that is, the correlation problem) since most strings with weight at most $\ell/2$ lie outside of the interval $[\ell/2 - \sqrt{\ell}, \ell/2)$. For applications to circuit lower bounds, this yields stronger parameters.

Does this work in our case? I think we can actually employ the same idea, but we still need to do the error reduction, whose complexity is going to be $\mathsf{poly}(1/\gamma)$, and we'd still get only an approximation of majority. Still, I think this will give us better explicit lower bounds as the input length is significantly bigger.

Next, we perform a non-uniform error reduction to make $\mathbf{S}$ deterministic using two steps. First, we reduce to the problem of distinguishing strings in $\{0,1\}^m$ of weight roughly $\delta\beta m$ from βm for an appropriate choice of m. Second, we solve this distinguishing problem in AC^0 using a relative-error approximate majority construction.

Claim 16 (Shaltiel-Viola)**.** Let $\mathbf{S}$ be from Claim 12. For any $\Delta > 0$, there are $m = O(\frac{1}{\eta^*\Delta^2\alpha})$ tests $s_1, ..., s_m$ with the following properties:

1. When $x \in \{0,1\}^{1/\eta^*}$ is balanced, $\sum_{i\in[m]} s_i(x) \leq (\delta\beta - \gamma)m(1+\Delta)$
2. When $x \in \{0,1\}^{1/\eta^*}$ has weight $w\eta^*$, $\sum_{i\in[m]} s_i(x) \geq \beta m(1-\Delta)$.

Proof. Suppose y is balanced and z has weight $w\eta^*$ and consider drawing m samples $S_1, ..., S_m$ from $\mathbf{S}$. By the Chernoff bound,

$$\Pr\left[\sum_{i\in[m]} S_i(y) \geq (\delta\beta-\gamma)m(1+\Delta)\right] \leq \Pr\left[\sum_{i\in[m]} S_i(y) \geq \alpha m(1+\Delta)\right] \leq \exp\left(-\frac{\alpha m\Delta^2}{2}\right) \tag{1}$$

$$\Pr\left[\sum_{i\in[m]} S_i(z) \leq \beta m(1-\Delta)\right] \leq \exp\left(-\frac{\beta m\Delta^2}{2}\right) \leq \exp\left(-\frac{\alpha m\Delta^2}{2}\right). \tag{2}$$

both of which are bounded by $2^{-1/\eta^*}$ by taking $m \geq O(1/(\eta^*\Delta^2\alpha))$. The union bound therefore tells us that the probability over our samples that every balanced string's sum is small and every unbalanced string's sum is large is more than 0, meaning in particular that some choice of samples $s_1, ..., s_m$ realizes the desired properties. ■

Let $\Theta = \frac{\delta\beta}{\delta\beta-\gamma}$ and let $\Delta = \frac{\Theta-1}{\Theta+1} = \frac{\gamma}{2\delta\beta-\gamma}$. Doing so means that $(\delta\beta-\gamma)m(1+\Delta) = \delta\beta m(1-\Delta)$ and in particular our we'll distinguish weight $k = \beta m(1-\Delta)$ strings from weight δk strings, setting up the relative-error approximate majority problem.[4] We will solve the relative-error approximate majority problem with pairwise independent hashing, essentially following the construction of Stockemeyer's theorem for counting NP witnesses from [Sto83].

The following is a standard concentration bound for pairwise independent hash functions.

Lemma 17. Let $\mathcal{H}$ be a family of pairwise independent hash functions $h : [m] \to [r]$. Then for any $\lambda > 0$ and $T \subseteq [m]$,

$$\Pr_{h\sim\mathcal{H}}[||h^{-1}(0)\cap T| - \frac{|T|}{r}| \geq \lambda\frac{|T|}{r}] \leq \frac{r}{\lambda^2|T|}.$$

In particular, when $|T| \geq 4r/\lambda^2$,

$$\Pr_{h\sim\mathcal{H}}[(1-\lambda)\frac{|T|}{r} \leq |h^{-1}(0)\cap T| \leq (1+\lambda)\frac{|T|}{r}] \geq 3/4.$$

[4] Note also that when γ is exponentially small in $1/\delta$ (taking, temporarily, $\beta = O(1)$), then Δ is exponentially small in $1/\delta$. This means that m would need to be exponentially large, which explains why dense model theorems with small γ, like Tao-Ziegler, would require an exponentially large oracle circuit to compute majority.

Proof. Let $C_x = \mathbf{I}[h(x) = 0]$ and $X_T = \sum_{x \in T} C_x$. By pairwise independence, we can write $\mathrm{Var}[X_T] = \sum_i \mathrm{Var}[C_x] \leq |T|$, as C_x are boolean. The conclusion follows by applying Chebyshev's inequality. ■

We can apply this directly to distinguish between high weight and low weight strings by hashing and counting the number of collisions.

Claim 18. Let $r = \delta k/16$, $\mathcal{H}$ a family of pairwise independent hash functions $h : [m] \to [r]$, and $S \subseteq [m]$. Then

1. If $|S| = k$, then
$$\Pr_{h \sim \mathcal{H}}[|h^{-1}(0) \cap T| \in [\frac{16}{\delta} - \frac{8}{\sqrt{\delta}}, \frac{16}{\delta} + \frac{8}{\sqrt{\delta}}]] \geq 3/4,$$
2. If $|S| = \delta k$, then
$$\Pr_{h \sim \mathcal{H}}[|h^{-1}(0) \cap T| \notin [8, 24]] \leq 1/4.$$

For these tests to be mutually exclusive, we require that $24 < \frac{16}{\delta} - \frac{8}{\sqrt{\delta}}$ or that $\delta < 4/9$, excluding, for example, the case where $\delta = \frac{k-1}{k}$ which is known to be intractable for AC^0 circuits (see, e.g., Hastad's thesis). Note that implementing this only requires counting up to a constant, as we're not concerned with the behavior of the test when the input doesn't satisfy the promise of being weight k or weight δk.

Since we're hashing the input co-ordinates, this test can be easily implemented in AC^0. Constructions of pairwise independent families $h : \{0,1\}^{\log m} \to \{0,1\}^{\log r}$ with size $|\mathcal{H}| = m \cdot r$ are known by taking, e.g. linear functions over finite fields. Specifically, we can consider $m = 2^s$ and $r = 2^t$, and then consider the first t bits of the function $h_{a,b}(x) = a \cdot x + b$, thinking of $a, b, x \in \mathbb{F}_{2^s}$. Since $s = \log m$, we can perform finite field arithmetic with an AC^0 circuit polynomial size in m. We then apply the hash function to each co-ordinate i so that the ith input bit is 1, and count the number of co-ordinates which hash to 0.

TO-DO: specify the depth of computing h

We therefore get the following claim:

Claim 19. There's a depth-d =?, randomized AC^0 circuit of size $\mathsf{poly}(m)$ that distinguishes between weight δk and weight k strings in $\{0,1\}^m$ with probability at least 3/4.

From here, there's two approaches towards derandomization. The first approach is to reduce the error our construction from 1/4 to $O(1/(\eta^*)^2)$ from which we can get a circuit which approximates majority on some large constant fraction of inputs. This still suffices for $\mathsf{AC}^0[\oplus]$ and AC^0 lower bounds.

A less efficient approach can give us a perfect derandomization: since $|\mathcal{H}|$ being polynomial in m allows us to derandomize the construction by an exhaustive search and then taking an *additive* approximate majority circuit, distinguishing between relative weights of 1/4 and 3/4 , for which efficient AC^0 constructions are known (e.g. Ajtai [Ajt83] or Viola [Vio09]). This increases the depth of the construction by 3.

We could also reduce the error from 1/4 to roughly $(\binom{m}{k} + \binom{m}{\delta k})^{-1} = O(m^{-k})$ and then union bound over the randomness, but then the hashing test itself seems to be hard for AC^0.

Theorem 20 ([Ajt83], [Vio09])**.** There exists a family of polynomial-size AC^0 circuits of depth 3 distinguishing between relative weight 1/4 strings and relative weight 3/4 strings.

As explained above, we can apply such an approximate majority circuit to the output of the tests $s_1, ..., s_m$ from Claim 14 to distinguish between weight $w\eta^*$ and weight $1/2\eta^*$ strings in $\{0,1\}^{1/\eta^*}$, can then be used in steps 3. and 4. to complete the reduction.

We're now left with computing m. Recall that $\Delta = \frac{\gamma}{2\delta\beta - \gamma}$, $\eta^* = \varepsilon + \sqrt{\frac{\delta^2 \log 1/\delta)}{2n}}$.

$$m = O(\frac{(\delta\beta - \gamma)^2}{\eta^* \gamma^2 \alpha}).$$

Of course, α could be small (or even 0) which in our case could be problematic. We can fix this by replacing the randomized test $\mathbf{S}$ with the test $\mathbf{S}'$ which outputs $\mathbf{S}(x)$ with probability $1 - \sigma$ and outputs 1 with probability σ. The new acceptance probabilities $\alpha' = \alpha + \sigma$ and $\beta' + \sigma$ will continue to satisfy the pseudo-density relation so long as

$$\alpha + \sigma \leq \delta(\beta + \sigma) - \gamma,$$

or $\sigma \leq (\delta\beta - \gamma - \alpha)/(1 - \delta)$, meaning we can assume with no loss of generality that the pseudo-density inequality is satisfied with equality with no affect on the previous analysis. This gives us:

$$m = O(\frac{(\delta\beta' - \gamma)}{\eta^* \gamma^2}),$$

where $\beta' = \beta + (\delta\beta - \gamma - \alpha)/(1 - \delta) = (\beta - \gamma - \alpha)/(1 - \delta)$ is the new acceptance probability for unbalanced strings by setting σ.

Finally, we can take β' to be a constant by taking the 'OR' of sufficiently-many independent copies of $\mathbf{S}'$.

Claim 21. There's a randomized test $\mathbf{S}''$ so that $\alpha'' \leq \delta\beta'' - \gamma$ where $\beta'' = \Pr[\mathbf{S}''(x) = 1 \mid x \text{ has weight } w] = O(1)$ and $\alpha'' \geq \alpha' = \Pr[\mathbf{S}'(x) = 1 \mid x \text{ is balanced}]$. Moreover, $\mathbf{S}''$ is the OR of $O(\delta/\gamma)$ copies of $\mathbf{S}'$.

Proof. Draw p independent tests $s_1, ..., s_p$ from $\mathbf{S}'$. Then if x has weight $w\eta^*$, then by inclusion-exclusion we have

$$\Pr_{s_1,...,s_p}[\bigvee_{i=1}^{p} s_i(x) = 1] \geq p\beta'(1 - \beta')$$

which is a constant when $p = O(1/\beta')$. By construction of $\mathbf{S}'$, we had $\beta' = (\gamma + \alpha)/\delta$. ■

We can therefore estimate $m = O(\frac{\delta}{\eta^* \gamma^2})$. To compute the total size of the circuit, we have $O((1/\eta^*)^2)$ calls to $\mathbf{S}''$ from steps 3. and 4. and $O(\delta/\gamma)$ calls to $\mathbf{S}$ for each call to $\mathbf{S}''$. Moreover, $\mathbf{S}''$ is derandomized using a depth-6, $\mathsf{poly}(m)$-size AC^0 circuit. Putting this together yields:

Claim 22. Suppose for each $\eta \geq \eta^*$, there's an algorithm A_η solving the (η, δ, γ)-coin-density problem on inputs of length n. Then there's a size-s, depth-d AC^0 circuit using oracle access to the A_η's computing majority on inputs of length $1/\eta^*$, where $s = \mathsf{poly}(\frac{\delta^2}{(\eta^*)^3 \gamma^3})$ and $\eta^* = \varepsilon + \sqrt{\frac{\delta^2 \log(1/\delta)}{2n}}$. $d =?$

3 Measures whose models require thresholds

We will once again take our tests $\mathcal{F}$ to to be bit queries, but we'll think of our universe as $U = \{-1, 1\}^n$. In this setting, it's easy to see the following claim:

Claim 23. Let $\mu : \{-1, 1\}^n \to [-1, 1]$ be a measure and $\nu : \{-1, 1\}^n \to [-1, 1]$ be an ε-model of μ with respect to bit queries $x \mapsto x_i$. Then for all $i \in [n]$,

$$\widehat{\nu}(i) = \widehat{\mu}(i) \pm 2\varepsilon$$

Now we'll choose μ to be the extremal case of $\widehat{\mu}(i)$ for subsets of $\{0,1\}^n$ with a fixed density. Our example B_δ will be the 'heaviest' subset of $\{-1,1\}^n$ with density δ. Specifically, note that, by the Chernoff bound, threshold functions of the form $\mathsf{Thr}_t(x) = \mathrm{sign}(-t\sqrt{n} + \sum_{i=1}^n x_i)$ have density $2^{\Theta(-t^2)}$. Then we'll take $B_\delta = \mathsf{Thr}_t$ with $t = \sqrt{n \log(1/\delta)}$. We can now see that B_δ is indeed the extremal case of $\widehat{\mu}(i)$:

Claim 24. For each $i \in [n]$,

$$\widehat{B_\delta}(i) \approx \sqrt{2\delta^2 \log(1/\delta)/n}$$

Proof. **To-do...** ■

A simple Fourier-analytic argument explains that, since $B \coloneqq B_\delta$ is a linear threshold function with integer weights, any ν whose linear Fourier coefficients are close to B's is actually close to B in distance.

Claim 25 (Birkendorf et al.)**.** Suppose $|\widehat{B}(i) - \widehat{\nu}(i)| \leq 2\varepsilon$ for all $i \in [n]$. Then

$$\Pr[B(x) \neq \nu(x)] \leq 2\varepsilon n$$

Proof. B is a threshold function and so $B = \mathrm{sign}(b)$ for $b(x) = -t + \sum_{i=1} x_i$. Note that by definition $\widehat{b}(i) = 1$ for all $i \in [n]$ and $\widehat{b}(S) = 0$ for $|S| > 1$. Then on the one hand, (by Plancherel's theorem)

$$\mathbf{E}[\nu \cdot b] = \sum_S \widehat{\nu}(S)\widehat{b}(S) = \delta \mathbf{E}[b] + \sum_{i=1} \widehat{\nu}(i).$$

On the other hand, (again by Plancherel's)

$$\mathbf{E}[B \cdot b] = \sum_S \widehat{B}(S)\widehat{b}(S) = \delta \mathbf{E}[b] + \sum_i \widehat{B}(i) \leq \delta \mathbf{E}[b] + \sum_i (\widehat{\nu}(i) + \varepsilon)$$

This means that

$$\mathbf{E}[B \cdot b] - \mathbf{E}[\nu \cdot b] = \mathbf{E}[(B - \nu) \cdot b] \leq \varepsilon n$$

Now note that $B(x) \cdot b(x) = |b(x)|$ and $|\nu(x) \cdot b(x)| = |b(x)|$, and so whenever $B(x) \neq \nu(x)$, it means that ν and b have different signs and so that the contribution to $\mathbf{E}[(B - \nu) \cdot b]$ is positive. Moreover, since $|b(x)| \geq 1$, the contribution is always at least 2^{-n}, finishing the proof. ■

Note that this holds independent of the density of the approximation ν. Using the density, we can get a similar argument where the error now depends on δ. When δ is large, this is less restrictive in terms of ε.

Claim 26. Let $\nu : \{-1,1\}^n \to \{-1,1\}$ and $\mathbf{E}[\nu] = \mathbf{E}[B] = \delta$. Suppose $|\widehat{B}(i) - \widehat{\nu}(i)| \leq 2\varepsilon$ for all $i \in [n]$ and

$$\varepsilon \leq \frac{1}{\sqrt{2n}}(\sqrt{\delta} - \sqrt{\delta^2 \log(1/\delta)}).$$

Then

$$\Pr[B(x) \neq \nu(x)] \leq \delta$$

Proof. Assume towards a contradiction that $\mathrm{dist}(\nu, \overline{B}) \geq \delta/2$, where $\overline{B} = 1 - B$. If we reach a contradiction then we can conclude the proof as follows: otherwise, $\mathrm{dist}(\nu, B) \geq \delta/2$. Then, since B is δ-dense, ν incurs at most $\delta/2$ error on B and, since ν is δ-dense, ν incurs at most $\delta/2$ error on $\overline{B}$, which gives us the desired claim.

To see the contradiction, note that

$$4\,\mathrm{dist}(\nu, \overline{B}) = \mathbf{E}[(\nu - \overline{B})^2] = \sum_S (\widehat{\nu}(S) - \widehat{\overline{B}}(S))^2 \geq \sum_{i=1}^n (\widehat{\nu}(i) - \widehat{\overline{B}}(i))^2,$$

and so

$$\sum_{i=1}^n (\widehat{\nu}(i) - \widehat{\overline{B}}(i))^2 \leq 2\delta.$$

For simplicity, write $\beta = \overline{B}(i) = -\sqrt{2\delta^2 \log(1/\delta)/n}$. Then by averaging, there's some $i \in [n]$ so that

$$|\widehat{\nu}(i) - \beta| \leq \sqrt{2\delta/n}.$$

Since $\widehat{\nu}(i) \leq \beta$ would already be a contradiction of the fact that ν is a model for B, we get

$$\widehat{\nu}(i) \leq \beta + \sqrt{2\delta/n}$$

which is a contradiction (remember B's bias is $-\beta$) so long as

$$\beta + \sqrt{2\delta/n} \leq -\beta - 2\varepsilon, \text{ or } \varepsilon \leq \frac{1}{\sqrt{2n}}(\sqrt{\delta} - \sqrt{\delta^2 \log(1/\delta)}),$$

■

Implicit in this claim is an argument that closeness in function space implies closeness of the linear Fourier coefficients. This argument has appeared before in [O'Donnell, Servedio, chow parameters problem].

Russell mentioned another claim if we look at $B_{\delta'}$ and B_δ for $\delta' > \delta$, then most of ν needs to lie on the correct side of $B_{\delta'}$. This appears to be the same argument: if I lied too much on the wrong side of $B_{\delta'}$, I would be too close to $\overline{B_{\delta'}}$ so that my bias would be too small

Proof. Consider $B_{\delta'}$ and B_δ and a δ-dense ε-model ν of B_δ. Suppose that $\mathrm{dist}(\nu, \overline{B_{\delta'}}) \geq \gamma$ for some γ. Then

$$\sum_{i=1}^n (\widehat{\nu}(i) - \widehat{\overline{B_{\delta'}}}(i))^2 \leq 4\gamma$$

and so some co-ordinate has

$$|\widehat{\nu}(i) - \widehat{\overline{B_{\delta'}}}(i)| \leq \sqrt{4\gamma/n},$$

meaning that $\widehat{\nu}(i) \leq \sqrt{4\gamma/n} + \sqrt{2\delta'^2 \log(1/\delta')/n}$. Supposing that this is a contradiction for our choice of ε, it tells us that $\mathrm{dist}(\nu, B_{\delta'}) \geq \gamma$, which we can use to compute the distance to B_δ...

This doesn't use the density of ν... ■

Claim 27. Dense model lower bound

Proof. Let δ be arbitrary and $\varepsilon \leq \frac{1}{\sqrt{2n}}(\sqrt{\delta} - \sqrt{\delta^2 \log(1/\delta)})$. Then any δ-dense, ε-model ν of B_δ has the property that

$$\Pr[\nu(x) \neq B_\delta(x)] \leq \min\{\varepsilon n, \delta\}.$$

■

It is easy to see that the same example applies to the decomposition lemma of Trevisan, Tulsiani and Vadhan [TTV09].

4 Applications

4.1 Explicit circuit lower bounds

Recall the following worst-case lower bounds from circuit complexity:

Theorem 28 ([Raz87], [Smo87], [Has86])**.** 1. Any depth-d $\mathsf{AC}^0[\oplus]$ circuit computing majority has size $2^{\Omega(n^{1/2(d-1)})}$.

2. Any depth-d AC^0 circuit computing majority has size $2^{\Omega(n^{1/(d-1)})}$.

Given the $\mathsf{poly}(\frac{\delta^2}{\gamma^3\eta^{*3}})$ bound on the size of an AC^0 circuit computing majority with oracle access to a pseudo-density refuter, we can prove explicit circuit lower bounds on pseudo-density refuters. Suppose we substitute the oracle for an explicit $\mathcal{C} \in \{\mathsf{AC}^0, \mathsf{AC}^0[\oplus]\}$ circuit which implements a dense-model reduction as above. If it has size s and depth d, then we obtain a $\mathcal{C}$ circuit computing majority of depth $d+6$ and size $O(s \cdot \frac{\delta^2}{\gamma^3\eta^3})$. We therefore get the following corollaries:

Additionally, we have correlation bounds on linear threshold functions for AC^0 circuits and $\mathsf{AC}^0[\oplus]$:

Theorem 29. 1. $\mathbb{F}_2$-degree bounds on linear threshold functions

2. AC^0-correlation bounds on linear threshold functions

Most of the references just look at majority, so let's look at B_δ instead as a function of δ. Of course, if δ is very far from $1/2$, then it will be easy because the function will be sparse.

4.2 Simple proofs of lower bounds for hardcore lemmas and boosting

We'll present simple proofs that hardcore lemmas and boosting require majority by reduction from dense model theorems. Both of these results can be deduced from [SV10], but our proofs are arguably simpler, since they avoid having to a construct a solution to the coin problem from a hardness amplification proof.

4.2.1 Hardcore lemmas require majority

The first application we present leverages the connection between dense model theorems and hardcore lemmas due to Impagliazzo in [Imp19].[5] Recall that the hardcore lemma, originally appearing in [Imp95], is an approach to hardness amplification and can be used to, for example, prove Yao's XOR lemma. It states that if we have a function f which mildly hard for circuits of size s, then there's a dense set of the inputs on which f is extremely hard for circuits of size s' slightly smaller than s. We'll define a hardcore lemma in the contrapositive: using that f is very mildly approximated by a circuit of s' over any dense set of the inputs, we'll build, in a black-box fashion, a good approximation of f over the uniform distribution by a circuit of size s slightly larger than s'.

Before stating the lemma, we'll fix some notation. A function f γ-approximates a function g over a distribution D if $\Pr_{x\sim D}[f(x) = g(x)] \geq \gamma$. We'll informally refer to a strong approximation as something like a γ-approximation when $\gamma = 1 - \delta$, thinking of δ as small. Similarly, a weak approximation will refer to a $(1/2 + \varepsilon)$-approximation, again thinking of ε as small. A function f is γ-hard over D for a class of tests $\mathcal{T}$ if $\Pr_{x\sim D}[T(x) = f(x)] \leq 1 - \gamma$ for any $T \in \mathcal{T}$.

[5] The connection had been observed before — see, for example, [TTV09], [Rei+08] — but [Imp19] presents an explicit reduction.

Definition 30 (Black-box hardcore lemma)**.** A (ε, δ, k)-black-box hardcore lemma reduction over the universe U with respect to tests $\mathcal{T}$ is a map from a function $f : U \to \{0,1\}$ to an oracle circuit $\mathsf{HC}(f) : U \to \{0,1\}$ so that the following holds: if for every 2δ-dense measure μ over U, there's a function $t \in \mathcal{T}$ that $(1/2+\varepsilon)$-approximates f over μ, then there exist k functions $T_1, ..., T_k$ so that computing $\mathsf{HC}(f)$ with oracle access to $t_1, ..., t_k$ is a $(1-\delta)$-approximation of f over the uniform distribution.

We have the luxury of choosing 2δ as the density parameter, which is optimal, due to the more refined boosting-type argument of Holenstein [Hol05] in which Holenstein presents and explicit $(\varepsilon, \delta, O(\frac{1}{\delta\varepsilon^2}))$ hardcore lemma reduction.

Theorem 31 ([Imp19], informal)**.** If there's a (ε, δ, k)-black-box hardcore lemma reduction over $\{0,1\}^n$ for tests $\mathcal{T}$, then there's a $(\varepsilon', \delta', \varepsilon\delta, k)$-black-box dense model reduction over $\{0,1\}^n$ for tests $\mathcal{T}$.

In order to state our results more explicitly, we'll need a slightly refined notion of black-box hardcore lemma. Namely, fix a *ground distribution* ρ over U. Define the density of a measure $\mu : U \to [0,1]$ relative to ρ as the weighted sum $d_\rho(\mu) = \mathbf{E}_{x\sim\rho}[\mu(x)]$. Similarly, the distribution induced by μ over the ground distribution ρ is given by the mass function $p(x) = \frac{\rho(x)\mu(x)}{d_\rho(\mu)}$. A black-box hardcore lemma relative to ρ is then a black-box translation from $(1/2+\varepsilon)$-approximations of f for any μ with $d_\rho(\mu) \geq 2\delta$ into a $(1-\delta)$-approximation of f over ρ.

Consider a set $D : U \to [0,1]$ which is distinguishable by $\mathcal{T}$ from every dense measure over U. We want to exhibit find a function that refutes D's pseudo-density.

The most obvious approach is to use D's characteristic function, which could be a somewhat easy function for $\mathcal{T}$, as $\mathcal{T}$ distinguishes D. Applying this directly, could be a problem: suppose μ is a set of size $\delta_0|U|$ and D is a subset of size $|\mu|/3$. Then the characteristic function of μ distinguishes between D and μ with advantage $2/3$. But since D is small inside of μ, the probability that $\mathbf{1}_\mu = D$ with inputs drawn from μ is $1/3$.

To fix this, we want to build a new universe U' out of U that contains a blown-up copy of D, so as to build a function f which encodes D that is relatively unbiased. Specifically, consider a new universe

$$U' = \{(x,1) : x \in D\} \cup \{(x,-1) : x \in U\}.$$

The *ρ-mixture distribution* on U' is defined as follows: with probability ρ, sample x from D and output $(x,1)$ and with probability $1-\rho$, sample x from U and output $(x,-1)$. For a class $\mathcal{T}$ of tests over U, we'll extended each $T \in \mathcal{T}$ to a function over U' as $T(x,b) = T(x)$. Now we can explicitly state the theorem:

Theorem 32. Let $\mathcal{T}$ be a class of tests over U containing the constants and closed under negation and suppose HC is an $(\varepsilon, \delta-\lambda)$-hardcore lemma over the ρ-mixture distribution on U' with respect to $\mathcal{T}$ (extending $\mathcal{T}$ to U' as above). Then there's a $(\varepsilon', \delta', \gamma)$-dense model theorem over $\mathcal{T}$ with $\varepsilon' = O(\varepsilon + \lambda/\delta)$, $\delta' = \frac{\lambda}{1-\delta}$, and $\gamma = \frac{\delta}{1-\delta}$ whenever $\delta \leq 2\varepsilon\lambda + \lambda/\delta - 2\varepsilon\lambda^2/\delta$.

For completeness, we'll reproduce the proof from [Imp19] in the Appendix. This reduction allows us to readily conclude the following:

Corollary 33. Any (ε, δ)-black-box hardcore set lemma reduction HC requires majority. Specifically, if HC is an $(\varepsilon, \delta-\lambda)$-black-box hardcore setlemma reduction, then there's an AC^0 circuit of size $\mathsf{poly}(\frac{\delta'}{\eta^3\gamma^2})$ and oracle access to HC computing majority on $1/\eta$ bits where $\delta' = \frac{\lambda}{1-\delta}$, $\gamma = \frac{\delta}{1-\delta}$ and $\eta = \varepsilon + \frac{\delta'}{\gamma} + \sqrt{\frac{\log(1/\delta')}{2(n-1)}}$.

[LTW11] had previously observed an exponential $\mathsf{AC}^0[\oplus]$ lower bound on HC by observing that HC can be used to approximate majority. This result is slightly different, as it gives an AC^0 Turing reduction from majority to any proof of the hard-core lemma (which in turn implies exponential $\mathsf{AC}^0[\oplus]$ lower bounds, see below). Since the hardcore lemma can be used to prove a hardness amplification result, it can also be deduced as a corollary of the Shaltiel-Viola lower bound [SV10].

4.2.2 Boosting requires majority

As originally observed by Klivans and Servedio in [KS03], hardcore set lemma reductions are essentially 'boosting' algorithms, developed in the machine learning community (Freund and Schapire [FS97]). Recall that a boosting algorithm is a generic procedure for taking a family of weak learners and producing a single strong learner.

Definition 34 (Black-box boosting algorithm)**.** For parameters $0 < \varepsilon, \delta$ and set $\Delta \subseteq \mathbb{R}^{2^n}$ of distributions over $\{0,1\}^n$, an (ε, δ)-*black-box boosting algorithm* with weak learners in $\mathcal{T}$ consists of a *family of weak learners* $\mathsf{WL} : \mathcal{F}_n \times \mathbb{R}^{2^n} \to \mathcal{T}$, a *distribution-generating mechanism* $\mathsf{Mech} : \mathcal{F}_n \times \mathcal{T} \to \mathbb{R}^{2^n}$ and a k-*query combiner* $\mathsf{Boost} \in \mathcal{F}_n$ which is an oracle circuit. The tuple $(\mathsf{WL}, \mathsf{Mech}, \mathsf{Boost})$ is a k-query, (η, γ)-boosting algorithm for Δ if the following stipulations hold for any $f \in \mathcal{F}_n$:

1. For any distribution $\mu \in \Delta$, $\mathsf{Mech}(f, \mathsf{WL}(f, \mu)) \in \Delta$.
2. If for any distribution $\mu \in \Delta$, $\mathsf{WL}(f, \mu)$ is an $(1/2 + \varepsilon)$-approximation of f on μ, then there is a distribution $\mu_0 \in \Gamma$ so that $\mathsf{Boost}(x)$ is a $(1 - \delta)$-approximation of f over the uniform distribution when computed with oracle access to the weak learners $h_0 = \mathsf{WL}(f, \mu_0)$, $h_1 = \mathsf{WL}(f, \mathsf{Mech}(f, h_i))$, ..., $h_k = \mathsf{WL}(f, \mathsf{Mech}(f, h_{k-1}))$.

A k-query combiner Boost is (ε, δ)-*correct* for Δ if for any family of weak learners WL, there's some distribution-generating mechanism Mech so that $(\mathsf{WL}, \mathsf{Mech}, \mathsf{Boost})$ is a k-query (ε, δ)-boosting algorithm for Δ. When a k-query combiner Boost which is correct for the distributions Δ_δ induced by δ-dense measures, we say that Boost is k-smooth.

That such algorithms do indeed exist is due to Freund and Schapire. Specifically, they take Mech to be the familiar multiplicative weights update procedure and take $\mathsf{Boost} = \mathsf{MAJ}$. Once a proof of correctness for boosting is established, it's relatively straightforward to see the following

Theorem 35. Let $(\mathsf{WL}, \mathsf{Mech}, \mathsf{Boost})$ be a 2δ-smooth, (ε, δ)-boosting algorithm with weak learners in $\mathcal{T}$ and Boost a k-ary combiner. Then there's a (ε, δ, k)-hardcore set lemma reduction for $\mathcal{T}$.

The translation is essentially immediate: the reduction HL simply implements the boosting algorithm on input f. If f has weak approximations over 2δ-dense distributions, then the hypothesis of the second component in the definition of $(\mathsf{WL}, \mathsf{Mech}, \mathsf{Boost})$ is satisfied, meaning that Boost with oracle access to the corresponding weak learners actually computes a strong approximation of f over the uniform distribution.

Corollary 36. Let Boost be a combiner for an (ε, δ)-boosting algorithm. Then there's an AC^0 circuit of size $\mathsf{poly}(\frac{\delta'}{\eta^3\gamma^2})$ and oracle access to Boost computing majority on $1/\eta$ bits, where δ', η and γ are set as in Corollary 24.

5 Acknowledgements

References

[Ajt83] M. Ajtai. "Σ^1_1-Formulae on Finite Structures". In: *Annals of Pure and Applied Logic* 24.1 (1983), p. 1. DOI: 10.1016/0168-0072(83)90038-6.

[FS97] Yoav Freund and Robert E Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: http://www.sciencedirect.com/science/article/pii/S002200009791504X.

[GSV18] Aryeh Grinberg, Ronen Shaltiel, and Emanuele Viola. "Indistinguishability by Adaptive Procedures with Advice, and Lower Bounds on Hardness Amplification Proofs". In: *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*. 2018, pp. 956–966. DOI: 10.1109/FOCS.2018.00094. URL: https://doi.org/10.1109/FOCS.2018.00094.

[GT08] Ben Green and Terence Tao. "The Primes Contain Arbitrarily Long Arithmetic Progressions". In: *Annals of Mathematics* 167.2 (2008), pp. 481–547. ISSN: 0003486X. URL: http://www.jstor.org/stable/40345354.

[Has86] John Hastad. "Almost optimal lower bounds for small depth circuits". In: *Proceedings of the eighteenth annual ACM symposium on Theory of computing*. Citeseer. 1986, pp. 6–20.

[Hol05] Thomas Holenstein. "Key Agreement from Weak Bit Agreement". In: *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*. STOC '05. Baltimore, MD, USA: ACM, 2005, pp. 664–673. ISBN: 1-58113-960-8. DOI: 10.1145/1060590.1060689. URL: http://doi.acm.org/10.1145/1060590.1060689.

[Imp19] Russell Impagliazzo. "Connections between pseudo-randomness and machine learning: boosting, dense models, and regularity". 2019.

[Imp95] R. Impagliazzo. "Hard-core Distributions for Somewhat Hard Problems". In: *Proceedings of the 36th Annual Symposium on Foundations of Computer Science*. FOCS '95. Washington, DC, USA: IEEE Computer Society, 1995, pp. 538–. ISBN: 0-8186-7183-1. URL: http://dl.acm.org/citation.cfm?id=795662.796290.

[KS03] Adam R. Klivans and Rocco A. Servedio. "Boosting and Hard-Core Set Construction". In: *Mach. Learn.* 51.3 (June 2003), pp. 217–238. ISSN: 0885-6125. DOI: 10.1023/A:1022949332276. URL: https://doi.org/10.1023/A:1022949332276.

[LTW11] Chi-Jen Lu, Shi-Chun Tsai, and Hsin-Lung Wu. "Complexity of Hard-Core Set Proofs". In: *computational complexity* 20.1 (Mar. 2011), pp. 145–171. ISSN: 1420-8954. DOI: 10.1007/s00037-011-0003-7. URL: https://doi.org/10.1007/s00037-011-0003-7.

[Raz87] A. A. Razborov. "Lower bounds on the size of bounded depth circuits over a complete basis with logical addition". In: *Mathematical notes of the Academy of Sciences of the USSR* 41.4 (Apr. 1987), pp. 333–338. ISSN: 1573-8876. DOI: 10.1007/BF01137685. URL: https://doi.org/10.1007/BF01137685.

[Rei+08] Omer Reingold et al. "Dense Subsets of Pseudorandom Sets". In: *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*. 2008, pp. 76–85. DOI: 10.1109/FOCS.2008.38. URL: https://doi.org/10.1109/FOCS.2008.38.

[Smo87] Roman Smolensky. "Algebraic Methods in the Theory of Lower Bounds for Boolean Circuit Complexity". In: *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987, New York, New York, USA*. 1987, pp. 77–82. DOI: 10.1145/28395.28404. URL: https://doi.org/10.1145/28395.28404.

[Sto83] Larry Stockmeyer. "The Complexity of Approximate Counting". In: *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*. STOC '83. New York, NY, USA: ACM, 1983, pp. 118–126. ISBN: 0-89791-099-0. DOI: 10.1145/800061.808740. URL: http://doi.acm.org/10.1145/800061.808740.

[SV10] R. Shaltiel and E. Viola. "Hardness Amplification Proofs Require Majority". In: *SIAM Journal on Computing* 39.7 (2010), pp. 3122–3154. DOI: 10.1137/080735096. eprint: https://doi.org/10.1137/080735096. URL: https://doi.org/10.1137/080735096.

[TTV09] Luca Trevisan, Madhur Tulsiani, and Salil P. Vadhan. "Regularity, Boosting, and Efficiently Simulating Every High-Entropy Distribution". In: *Proceedings of the 24th Annual IEEE Conference on Computational Complexity, CCC 2009, Paris, France, 15-18 July 2009*. 2009, pp. 126–136. DOI: 10.1109/CCC.2009.41. URL: https://doi.org/10.1109/CCC.2009.41.

[TZ08] Terence Tao and Tamar Ziegler. "The primes contain arbitrarily long polynomial progressions". In: *Acta Math.* 201.2 (2008), pp. 213–305. DOI: 10.1007/s11511-008-0032-5. URL: https://doi.org/10.1007/s11511-008-0032-5.

[Vio09] Emanuele Viola. "On Approximate Majority and Probabilistic Time". In: *Comput. Complex.* 18.3 (Oct. 2009), pp. 337–375. ISSN: 1016-3328. DOI: 10.1007/s00037-009-0267-3. URL: http://dx.doi.org/10.1007/s00037-009-0267-3.

[Zha11] Jiapeng Zhang. "On the query complexity for Showing Dense Model". In: *Electronic Colloquium on Computational Complexity (ECCC)* 18 (2011), p. 38. URL: http://eccc.hpi-web.de/report/2011/038.

A Deducing the dense model theorem from the hardcore lemma

Theorem 37. Let $\mathcal{T}$ be a class of tests over U containing the constants and closed under negation and suppose HC is an $(\varepsilon, \delta - \lambda)$-hardcore lemma over the ρ-mixture distribution on U' with respect to $\mathcal{T}$ (extending $\mathcal{T}$ to U' as above). Then there's a $(\varepsilon', \delta', \gamma)$-dense model theorem over $\mathcal{T}$ with $\varepsilon' = O(\varepsilon + \lambda/\delta)$, $\delta' = \frac{\lambda}{1-\delta}$, and $\gamma = \frac{\delta}{1-\delta}$ whenever $\delta \leq 2\varepsilon\lambda + \lambda/\delta - 2\varepsilon\lambda^2/\delta$.

Define the function $f : U \times \{-1, 1\} \to \{-1, 1\}$ to be 'indicator' function of D: $f(x, b) = b$. The following claim establishes the basic correspondence between hardness of f in U' versus pseudodensity of D in U.

Claim 38. Let $0 < \beta < \alpha < 1$ and $0 < \rho < 1$ be arbitrary. f is λ-hard for a class $\mathcal{T}$ of tests (over U, extended to U' as above) over the ρ-mixture distribution iff. D is $(\frac{\lambda-\rho}{1-\rho}, \frac{\rho}{1-\rho})$-pseudodense for $\mathcal{T}$.

Proof. Fix some $T \in \mathcal{T}$ so that $\Pr_{(x,b)\sim U'}[T(x) = f(x, b)] \geq 1 - \lambda$. By definition of the ρ-mixture distribution on U', we have

$$\begin{aligned}
\Pr_{(x,b)\sim U'}[T(x) = f(x, b)] &= \rho \Pr_{x\sim D}[T(x) = 1] + (1 - \rho) \Pr_{x\sim U}[T(x) = -1] \\
&= \rho \Pr_{x\sim D}[T(x) = 1] + (1 - \rho)(1 - \Pr_{x\sim U}[T(x) = 1]) \\
&\geq 1 - \lambda
\end{aligned}$$

Rearranging gives us

$$\frac{\rho}{1 - \rho} \Pr_{x\sim D}[T(x) = 1] + \frac{\lambda - \rho}{1 - \rho} \geq \Pr_{x\sim U}[T(x) = 1].$$

The other direction is similar. ∎

Now suppose we have an (ε, δ, k)-hardcore lemma reduction HC. By the above, it suffices to prove that f is *not* δ-hard over the ρ-mixture distribution for an appropriate ρ. We'll do so by contradiction: suppose that f is δ-hard for the class of functions of the form $\mathsf{HC}(T_1, ..., T_k)$ where $T_1, ..., T_k \in \mathcal{T}$ and HC is our

hardcore lemma reduction. By definition of the hardcore reduction, this implies that the hypothesis fails: that is, it implies the existence of a 2δ-dense distribution μ' on which f is $(1/2-\varepsilon)$-hard for $\mathcal{T}$. Then we make the following claim.

Claim 39. Let μ' be a $2(\delta-\lambda)$-dense measure on the ρ-mixture of U' which is $(1/2-\varepsilon)$-hard for $\mathcal{T}$ and let $\mu(x) = \mu'(x,1)$ denote a measure on U. Then

1. μ is $\frac{\delta-\lambda}{\rho}(1-2\varepsilon)$-dense, and
2. By picking $\rho = \delta$, the distribution induced by μ is *statistically* indistinguishable from D with advantage at most $2\varepsilon + \lambda/\delta - 2\varepsilon\lambda/\delta$.

Proof. Since $\mathcal{T}$ contains the constant functions and for any $T \in \mathcal{T}$, $\Pr_{(x,b)\sim\mu'}[T(x) = b] \leq 1/2 + \varepsilon$, it follows that the probability that $b = 1$ when drawn from μ' is at least $1/2 - \varepsilon$ (otherwise, the function $T(x) = -1$ would approximate f). On the other hand, we can compute the probability that $b = 1$ as

$$\Pr_{(x,b)\sim\mu'}[b = 1] = \rho\frac{d(\mu)}{d(\mu')},$$

since the distribution induced by the measure μ' has probability mass function $p((x,b)) = \rho\frac{\mu'(x,b)}{d(\mu')}$. Thus

$$\begin{aligned}\rho d(\mu) &\geq (1/2-\varepsilon)d(\mu') \geq (1/2-\varepsilon)2(\delta-\lambda)\\ d(\mu) &\geq (1-2\varepsilon)\frac{\delta-\lambda}{\rho}.\end{aligned}$$

For the second claim, note that μ is supported on D and the uniform distribution on D has density 1. ■

The theorem then follows: since we're assuming that D is *computationally* distinguishable (a stronger assumption than statistical distinguishability, as the distinguishing test is encoded by function in $\mathcal{T}$) from dense distributions, we can conclude conclude that f is not δ-hard. This implies that the approximation of f is indeed a refuter of D's pseudodensity.

To move from a dense model theorem over sets to a dense model theorem over measures, fix a measure $D : U \to [0,1]$ which is distinguishable from every δ'-dense measure over U. We will sample from the distribution over U induced by D to obtain a set and then apply the dense model theorem for sets to obtain a refuter of the pseudo-density. By concentration bounds, such a refuter will also refute the pseudo-density of D (with high probability).

Lemma 40. Let $S \subseteq U$ be a set of m samples from (the distribution induced by) D and suppose that

$$\Pr_{x\sim U}[T(x) = 1] \leq \delta' \Pr_{x\sim S}[T(x) = 1] - \gamma.$$

Then with probability at least $1 - \exp\big(-O(m\iota^2)\big)$ over our choice of S, T also refutes the $(\delta', \gamma - \iota)$-pseudodensity of D.

Proof. Let $X_1, ..., X_m \sim D$ denote our random variables for members of S. Then

$$\mathbf{E}_S[|\{T(x) = 1 : x \in S\}|] = \sum_{i=1}^{m} \Pr_{X_i\sim D}[T(X_i) = 1] = m \cdot \Pr_{X\sim D}[T(X) = 1].$$

as $X_1, ..., X_m$ are i.i.d. By the Chernoff-Hoeffding bound, the probability over S that $\Pr_{x\sim S}[T(x) = 1] - \Pr_{x\sim D}[T(x) = 1] \geq \iota$ is bounded from above $\exp\big(-O(m\iota^2)\big)$. ■