Electronic Colloquium on Computational Complexity, Report No. 179 (2021)

# Smoothed Complexity of Learning Disjunctive Normal Forms, Inverting Fourier Transforms, and Verifying Small Circuits

Tatsuie Tsukiji*

December 8, 2021

**Abstract**

This paper aims to derandomize the following problems in the smoothed analysis of Spielman and Teng. Learn Disjunctive Normal Form (DNF), invert Fourier Transforms (FT), and verify small circuits' unsatisfiability. Learning algorithms must predict a future observation from the only $m$ i.i.d. samples of a fixed but unknown joint-distribution $P(G(x), y)$ to explain an $\eta$-noisy target $P(y \neq f_\theta(G(x))) \leq \eta < 1/2$. Inverters must retrieve the hidden parameter $\theta$. The smoothed analysis can weaken the adversarial distribution $P(x, y)$ by injecting an appropriate perturbation $G$ with larger min-entropy $\mathrm{H}_\infty(G) := -\log \min_g \mathsf{Pr}[G = g]$. The previous algorithms allowed $\mathrm{H}_\infty(G) = \mathrm{poly}(n)$ for avoiding the worst-case intractability. We will derandomize them below $\mathrm{H}_\infty(G) \leq O(\log n)$ and establish **1**–**10** for planted functions (Goldreich's PRG) $f_\theta(x) = f(\theta \circ x_1, \ldots, \theta \circ x_d)$ with variables $x_i \in \{0, 1, \ldots, 2n-1\}$ plugged into $\theta \circ x_i := \theta(\lfloor x_i/2 \rfloor) \oplus x_i \in \{0,1\}$ in **1**–**4**, $\theta \circ x_i := \theta(\lfloor x_i/2 \rfloor) \cdot (-1)^{x_i} \in \mathbb{Z}_p$ of a large prime $p$ in **5** and **8**–**10**, and $\theta \circ x_i := \theta_i \cdot \lfloor x_i/2 \rfloor \cdot (-1)^{x_i}$ in **6**–**7**. **11**–**13** will verify the unsatisfiability of small circuits in the worst case analysis ($\mathrm{H}_\infty(G) = 0$). Suppose $\log n \gg \log \frac{ds}{\varepsilon} + k$. Randomly pick an example $(X, Y)$ from the observed $m$ data.

**1.** At $\mathrm{H}_\infty(G) = 0$, $\mathsf{Max}k\mathsf{CSP}$ of any $k$-variate predicate $f$ requires the sample size $\Omega(n^{(k-\epsilon)/2}) \leq m \leq \tilde{O}(n^{k/2})$ to distinguish between $|\max_\theta P(y = f_\theta(x)) - \max_\theta P'(y = f_\theta(x))| \geq \Omega(1)$ and $P(x, y) \equiv P'(x, y)$ in $n^{O(k)}$ time by given access to both samplers $P(x, y)$ and $P'(x, y)$.
**2.** At $\mathrm{H}_\infty(G) = c \log s$, the planted $s$-term DNF demands $m \geq n^{\Omega(\log s)}$ for $c < 1$, and $m \leq n^{\frac{1}{2}\log s + O(1)}$ for $c > 1$, to make $n^{O(\log s)}$-time PAC learning (even under a slight noise).
**3.** At $\mathrm{H}_\infty(G) = c \log 1/\varepsilon$, planted AND needs $m \geq n^{\Omega(\log \frac{1}{\varepsilon})}$ for $c < 1$, and $m \leq n^{\frac{1}{2}\log \frac{1}{\varepsilon} + O(1)}$ for $c > 1$, to make $(\max_\theta P(y = f_\theta(G(x)) + \varepsilon)$-accurate agnostic learning in $n^{O(\log 1/\varepsilon)}$-time.
**4.** At $\mathrm{H}_\infty(G) = O(\log s)$, the monotone DNF with expanding $s$-terms is PAC learnable from $m = n \cdot \mathrm{poly}(s)$ data with $\mathsf{Pr}[\lfloor X_i/2 \rfloor, \lfloor X_{i'}/2 \rfloor] \geq 1/n^{1+\epsilon}$ in $n \cdot \tilde{O}(s^{\log d})$ time.
**5.** At $\mathrm{H}_\infty(G) = O(\log p)$, the $k$FT $f_\theta(x) = \sum_{|w| \leq k} \hat{f}_w \prod_{i \in w} \theta \circ x_i$ over $\mathbb{Z}_p$ of $p \geq n^3$ is invertible from $m = O(n^{k+2} p)$ data with $\mathsf{Pr}[\{\lfloor X_{i_1}/2 \rfloor, \cdots, \lfloor X_{i_k}/2 \rfloor\}] \geq \Omega(1/n^k)$, $|Y| \leq r \leq p^{1/2^{k+1}}$, and $\mathsf{Pr}[Y \neq f_\theta(X) \mid \lfloor X_{i_1}/2 \rfloor, \cdots, \lfloor X_{i_k}/2 \rfloor] \ll 1/(nr)$ in $O(n^{k+3} p)$ time.
**6.** LPN and LWE over $\mathbb{Z}_p$ of $p \geq n^{\Omega(1)}$ hiding small secrets $\forall i, |\theta_i| = O(1)$ are breakable in polynomial time. **7.** $\mathrm{GapSVP}_{\tilde{O}(n^2)}$ is breakable within polynomial time.
**8.** At $\mathrm{H}_\infty(G) = O(\log n)$, any bilinear form $\sum_{i,j=1}^n \mathbf{x}_i M_{ij} \mathbf{x}_j$ with sparsity $|\{M_{ij} \in \{-1, 0, 1\} \mid M_{ij} \neq 0\}| \leq n^{2-o(1)}$ requires $\Omega(n(\log\log n)^{1-\epsilon})$ size for algebraic $\mathsf{NC}^1$ circuits over $\mathbb{Z}_p$ of $p \geq n^{o(1)}$ unless the matrix $M$ is learnable from only $m = n^{o(1)}$ data in $n^{o(1)}$ time.
**9.** At $\mathrm{H}_\infty(G) = O(n)$, any $2^{\frac{n}{2}}$ by $2^{\frac{n}{2}}$ matrix with sparsity $2^{n-n^\epsilon}$ demands $\exp(n^{\Omega(1)})$ size $\mathsf{PH}^{\mathsf{cc}}$ protocol unless it is learnable from $m = \exp(n^\epsilon)$ data in $\exp(n^\epsilon)$ time.
**10.** $\mathsf{PH}^{\mathsf{cc}} \neq \mathsf{PSPACE}^{\mathsf{cc}}$ or $\forall k, \mathsf{NP} \not\subset \mathsf{DEP}[k \log n]$. **11.** $\mathsf{VP} \neq \mathsf{VNP}$ or $\forall k, \mathsf{quasi\text{-}NP} \not\subset \mathsf{NC}^k$.
**12.** $\mathrm{PIT} \in \mathsf{DTIME}[n^{\mathrm{poly}(\log\log n)}]$ or $\forall \epsilon, \forall k, \mathsf{NTIME}[2^{n^\epsilon}] \not\subset \mathsf{SIZE}[n^k]$. **13.** $\mathsf{quasi\text{-}NP} \not\subset \mathsf{TC}^0$.

*School of Science and Engineering, Tokyo Denki University, Ishizaka, Hatoyama, Hiki, Saitama, 359-0394, Japan. Email: tsukiji@mail.dendai.ac.jp.

ISSN 1433-8092

**Keywords:** agnostic learning, Algebraic circuits, Approximaton Algorithms, boosting, Communication complexity, Correlation, cryptography, CSP, derandomization, DNF, Fourier transforms, lattice, Linear programming, circuit lower bound, Learning with Error, Matrix Rigidity, Min-entropy, natural proofs, NC, optimization, PAC learning, PIT, Polynomial Calculus, polynomial hierarchy, Refutation algorithms, Resolution, SAT, SDP, Smoothed Analysis, Statistical Analysis, sum-of-squares, SVP, Threshold Circuits, VNP, VP.

# 1 Introduction

This paper studies the computational complexity of learning a hidden parameter $\theta$ of a fixed but unknown data distribution $P_\theta(z)$. A learner aims to predict a new observation drawn from $P_\theta(z)$ at a high confidence level. The only data given to the learner is a *training dataset*[1] $\mathcal{D} = \{z(1), \ldots, z(m)\}$ composed of the i.i.d. (*independent and identically distributed*) outcomes emitted from the *unknown target distribution* $P_\theta(z)$. The worst-case analysis is the gold standard to measure the performance of algorithms learning the *target class* $\{P_\theta(z)\}_\theta$. It guarantees the algorithm's performance no matter which $\theta$ hides. Unfortunately, for many fundamental computational learning problems, worst-case analyses have revealed the existence of hard-to-compute points in the parameter space $\theta \in \mathcal{T}$. The intrinsic difficulty of the learning relied on either information theory, proof theory, or computational complexity theory. However, such a $\theta$ might be so rare that the learner living in an uncertain environment would seldom encounters it. For example, many easy-to-learn points may surround a rare hard one with a small degree of "perturbation." Then one can rarely observe a learning curve detecting the hard one. Spielman and Teng [ST04, ST09] formulated such worst-case demanding but practically easy learning situations in a *smoothed analysis* (SA). It interpolated between the worst-case $|\mathcal{G}| = 1$ and the average-case $|\mathcal{G}| = |\mathcal{Z}|$ by a more prosperous perturbation space $\mathcal{G}$ inducing a weaker adversary:

SA1: Let the adversary first choose a distribution $P_\theta(z)$.

SA2: Randomly generate a $G$ over $\mathcal{Z}$ permutation to cause a permutation $\hat{G}$ over $\mathcal{T}$.

SA3: Let the learner access the permuted distribution $P_\theta(G(z)) = P_{\hat{G}(\theta)}(z)$.

Let us review the previous smoothed analyses in computational learning theory under typical perturbations $G$ that have small quantity yet enough quality to circumvent the worst-case computational intractability and provide efficient learning algorithms.

REVIEW1: **Gaussian mixture learning** observes $z(j) \sim P_\theta(z)$ emitted from a mixture of $k$ Gaussians over $\mathbb{R}^n$ with means and covariances hidden in $\theta$ [Das99]. The worst-case analysis can estimate $\theta$ in $\mathrm{poly}(n)$ time for $k = O(1)$ [FSO06, MV10, BS15]. However, it demands an information-theoretic lower bound $\exp(k)$ of $k \geq \omega(1)$ to the number of training examples [MV10]. In a smoothed analysis, Ge, Huang, and Kakade [GHK15] gave a polynomial-time algorithm to learn $O(\sqrt{n})$ Gaussians. It disturbed a data $z$ emitted from the unknown mixture by adding a random vector $z + G$ drawn from the i.i.d. Gaussians $G_i \in \mathbb{R}$ with means $\mathbb{E}[G_i] = 0$ and variances $\mathbb{E}[G_i^2] = \epsilon^2$ for all dimensions $i \in (n] := \{1, \ldots, n\}$.

REVIEW2: **Perceptron learning** receives a dataset $\mathcal{D} = \{(x(1), y(1)), \ldots, (x(m), y(m))\} \sim P^m(x, y)$ supervised by a halfspace $y(j) = f_\theta(x(j)) = \mathrm{sgn}(\sum_{i=1}^n \theta_i x_i(j) + \theta_0)$ of $x(j) = (x_i(j))_{i=1}^n \in \mathbb{R}^n$. The famous perceptron algorithm takes $\exp(n)$ time to retrieve a hidden

[1] A dataset may contain the same data multiple times.

$\theta \in \mathbb{R}^{n+1}$ [MP17]. Under a small additive Gaussian perturbation $\big(x(j)+G, f_\theta(x(j)+G)\big)$, Blum and Dunagan [BD02] analyzed that the perceptron algorithm ran in polynomial time. Even more, the perceptron resolved *Linear Programming* (LP) as efficiently as the practical standard simplex algorithm of Spielman and Teng in the smoothed analysis [ST04].

REVIEW3: **PAC** (*Probably, Approximately, and Correctly*) learn a concept (i.e., a specific expression class) of a *Boolean function* $f : \{0,1\}^n \to \{0,1\}$ from a supervised dataset [Val84]. Elementary yet general, the most studied concept is a *Disjunctive Normal Form* (DNF) $f(\mathbf{x}) := \bigvee_{\kappa=1}^{s} \bigwedge_{i \in f_\kappa} \mathbf{x}_i \oplus f_{\kappa i}$ of given $f_\kappa \subset (d] := \{1,\dots,d\}$ and $f_{\kappa i} \in \{0,1\}$. We write it as $f \in s\text{-term}\,\mathrm{DNF}_d$, and $f \in s\text{-term}\,k\mathrm{DNF}_d$ when $\forall |f_\kappa| = k$ [Val85]. It is learnable in quasi-polynomial time under the product distribution $P(x) = \prod_{i=1}^{n}(\mu x_i + (1-\mu)(1-x_i))$ of mean $\mu \in [\epsilon, 1-\epsilon]^n$ [Ver90]. Kalai, Samorodnitsky, and Teng [KST09, Fel12] proved that DNF is polynomial-time learnable under the product distribution of mean $\mu + \hat{G}$ perturbed by the uniform random vector $\hat{G} \in [-\frac{\epsilon}{2}, \frac{\epsilon}{2}]^n$.

REVIEW4: **Low-degree Fourier inversion** is the most successful method of learning DNF over the real-number field $\mathbb{R}$: Learn $Y \approx \sum_{|w| \le k} \theta_w \prod_{i \in w}(-1)^{X_i}$ by inverting the unknown $\theta_w \approx \mathbb{E}_X\big[Y \prod_{i \in w}(-1)^{X_i}\big]$ under the *empirical distribution*, i.e., the uniform random variable $(X, Y)$ over $\mathcal{D}$ [KKL88, KM93, Man95, GKK08]. It has succeeded in quasi-polynomial time PAC learning $\mathsf{AC}^0$ [LMN93], polynomial-time PAC learning DNF with membership queries [Jac97], and even without membership queries [BMOS05]. The former two results assumed the uniform distribution $P(x) = 1/|\{0,1\}^n|$. The last one took a random walk.

REVIEW5: **Agnostic learning (empirical risk minimization)** puts virtually no assumption on a given dataset and asks to minimize $\mathrm{err}(\mathcal{D}) = \min_f \mathrm{err}_f(\mathcal{D})$ of the observed *error rate* $\mathrm{err}_f(\mathcal{D}) := \frac{1}{m}\sum_{j=1}^{m} 1[f(x(j)) \ne y(j)]$ of a hidden concept $f \in \mathcal{F}$. VC-dimension theory [BEHW89, Hau92, Vap06] promises a polynomial-size sample complexity $O(\log|\mathcal{F}|)$, but it might not provide polynomial-time learning. For example, AND := 1-term DNF took $n^{O(\sqrt{n})}$ time for agnostic learning [TT99, KKMS08], despite only $O(n)$ time in PAC [Val84].

REVIEW6: **R$k$SAT refutation** denies the existence of an assignment $\theta \in \{0,1\}^n$ satisfying the OR predicate $f_\theta(x) := \bigvee_{i=1}^{k} \theta \circ x_i = \bigvee_{i=1}^{k} \theta(\lfloor x_i/2 \rfloor) \oplus x_i$ for all constraints[2] $x \in \mathcal{U} \subset [2n)^k := \{0, 1, \dots, 2n-1\}^k$ drawn from $P(x) = \frac{1}{(2n)^k}$, i.e., disprove $\mathrm{err}(\mathcal{U}) = \min_\theta \mathrm{err}_\theta(\mathcal{U}) = 0$ [CS88, Fei02]. It has noticed a constant $\alpha_k \approx 2^k \ln 2 - (1+\ln 2)/2$ to make a sharp threshold $\forall \epsilon > 0, \lim_n \mathsf{Pr}\big[\mathrm{err}(\mathcal{U}) = 1 \mid m/n \ge \alpha_k - \epsilon\big] = 1 = \lim_n \mathsf{Pr}\big[\mathrm{err}(\mathcal{U}) = 0 \mid m/n \le \alpha_k + \epsilon\big]$ [Fri99, MPZ02, DSS15, COP16]. Moreover, its data size complexity $m = \min|\mathcal{U}|$ of efficient refutation has attained the following dichotomy. R$k$SAT refutation is polynomial-time solvable above $m \ge O(2^k n^{k/2})$ [GK01, FGK05, COGL07, FO07, COCF10, BM16, AOW15], but demanding $\exp(n^{\Omega(1)})$ proof-length or $n^{\Omega(1)}$ proof-degree below $m \le n^{(1-\epsilon)k/2}$ in the well-studied proof systems [BKPS98, AR01, BSW01, Gri01, Sch08, Tul09, BSI10, CLRS16, KMOW17, BCR20]. Feige [Fei07] refuted the 3SAT of adversarial $m = O(n^{3/2}(\log\log n)^{1/2})$ constraints efficiently under i.i.d. perturbations $x_i(j) \mapsto 2\lfloor x_i(j)/2 \rfloor + x_i(j) \oplus G_{ij}(\lfloor x_i(j)/2 \rfloor)$ by a *flipper* $G \in \{0,1\}^{nm}$ with a small mean $\mathbb{E}[G_{ij}(\lfloor x_i(j)/2 \rfloor)] = \epsilon$. Abascal, Guruswami, and Kothari [AGK21] recently generalized it to the $k$CSP refutation targetting the $f_\theta(x) = f(\theta \circ x_1, \dots, \theta \circ x_k)$ of an arbitrary $k$-variable Boolean predicate $f(\mathbf{x}_1, \dots, \mathbf{x}_k)$.

REVIEW7: **Max$k$SAT approximation** aims to measure the *empirical accuracy rate* $\mathrm{acc}(\mathcal{D}) := \max_\theta \frac{1}{m}\sum_{j=1}^{m} 1[f_\theta(x(j)) = 1] = 1 - \mathrm{err}(\mathcal{D})$ of the OR predicate $f_\theta(x) = \bigvee_{i=1}^{k} \theta \circ x_i$ on the

[2]Satisfiability problem's data $\big(x(j), y(j)\big)_{j=1}^{m}$ suppose to take the only positive labels $\forall j, y(j) = 1$.

worst-case data $\mathcal{D} \subset [2n)^k$. The $(\beta_{\texttt{cmp}}, \beta_{\texttt{snd}})$-gap approximation asks to distinguish between $\mathrm{acc}(\mathcal{D}) \leq \beta_{\texttt{snd}}$ and $\beta_{\texttt{cmp}} \leq \mathrm{acc}(\mathcal{D})$ of $0 < \beta_{\texttt{snd}} < \beta_{\texttt{cmp}} \leq 1$. The $(1, 1 - 1/2^k + \epsilon)$-gap approximation is hard on the $\mathsf{P} \neq \mathsf{NP}$ assumption [BGS98, Raz98, Hås01]. *Exponential Time Hypothesis* (ETH) conjectures that 3SAT must take $\exp(n)$ time to distinguish between $\mathrm{acc}(\mathcal{D}) = 1$ and $\neq 1$ [IP01], which obliges 3SAT to take $\exp(n/\mathrm{plog}(n))$ time even for $(1, 1-\epsilon)$-gap approximation (GapETH) [Din07, BSS08, BV19]. Under ETH, Max$k$CSP of $m \leq O(n^{k-1})$ must consume $2^{n^{1-\epsilon}}$ approximation time [FLP16, MR17]. Meanwhile, Max$k$CSP enjoys polynomial-time approximation for $m \geq \Omega(n^k)$ [AKK99]. We will study a "promise" problem to choose $P(x, y)$ from a promised class, e.g., LPNs (in Review9) with Hamming-distance noise $0, 1, 2, \ldots$ [Ale11]. R$k$CSP's refutation of [AGK21] gives rise to the promise-Max$k$CSP's approximation by only $\tilde{O}(n^{k/2})$ constraints in $n^{O(k)}$ time.

Review8: **Planted CSP** asks to invert the secret assignment[3] $\theta \in \{0,1\}^n$ planted in a predicate $f(\theta \circ x) := f(\theta \circ x_1, \ldots, \theta \circ x_d)$, which we call a *planted predicate*. Goldreich studied it for a one-way function candidate generating pseudorandom bits $f(\theta \circ x(1)) \cdots f(\theta \circ x(m))$ under the uniform $P(x) = 1/|[2n)^d|$ [Gol00, AIK06, App13, OW14, BBKK18, AL18, FPV18]. It involves several well-studied inversion problems, e.g., the planted $d$SAT by $f = \bigvee_{i=1}^{d} \mathbf{x}_i$ [BHL$^+$02, JMS07, KMZ14], the noisy $d$LIN[4] over $\mathbb{F}_2$ by $(-1)^f \approx \prod_{i=1}^{d} (-1)^{\mathbf{x}_i}$ [BFKL93, ABW10, Ale11, DMQN12, BSV19], and the planted $k$DNF by $f = \bigvee_{j=1}^{d/k} \bigwedge_{i=1}^{k} \mathbf{x}_{i+jk}$ [DSS16].

Review9: **LPN and LWE**[5] ask to invert the hidden coefficient vector $\theta \in \mathbb{Z}_q^n$ of a noisy linear equation $y(j) = \sum_{i=1}^{n} \theta_i x_i(j) + E(j)$ of a given matrix $(x_i(j))_{i,j} \in \mathbb{Z}_q^{n \times m}$ contaminated by i.i.d. errors $E(j) \in \mathbb{Z}_q$. LPN supposes the uniform random matrix with Bernoulli noise $\mathsf{Pr}[E(j) \neq 0] = \mu$ [AAB17, BCG$^+$20, CM21, JLS21], while LWE treats Gaussian error $\mathsf{Pr}[E(j)] = 1/\sqrt{2g\sigma} \cdot \boldsymbol{e}^{-E(j)^2/(2\sigma)}$ [AD97, Reg04, KS06b, BV14]. The current best attackers take $\min\big(q^{O(\frac{n}{\log n})}, 2^{\tilde{O}(\sigma^2)}\big)$ time for LPN and LWE having $q \gg (\sigma \log n)^2$ (where $\sigma = \mu q$ for LPN) [BKW03, AG11]. Remarkably, LWE enjoys a worst-case hardness guarantee even for the binary parameter $\theta \in \{0,1\}^n$ [Reg04, LM09, Pei09, BLP$^+$13]. Its security stands on lattice problems enjoying average-case hardness by assuming only the worst-case one, e.g., $\mathrm{GapSVP}_\gamma$[6] [Ajt96, Cai99, Mic02, Reg04, MR07, GPV08, GINX16].

Review10: **Matrix rigidity problem** asks to invert the unknown $\theta \in \mathbb{F}^{\sqrt{N} \times n}$ from a limited amount of data $\mathcal{D} = \{\mathcal{M}(i,j) \in \mathbb{F} \mid (i,j) \in (\sqrt{N}] \times (\sqrt{N}]\}$ of a square matrix $\mathcal{M}$ over a field $\mathbb{F}$ to satisfy $\mathsf{Pr}_{I,J}[\mathcal{M}(I,J) = \sum_{\kappa=1}^{n} \theta(I,\kappa)\mathcal{M}(\kappa,J)] \approx 1$. It tries to predict the randomly picked entry $\mathcal{M}(I,J)$ by looking at only the first $nm \leq o(N)$ entries [Val77, Raz89, Pud94, Lok01, PP06, AW17, GT18, DL19, GW20]. When $\mathcal{M}(i,j) \in \{-1, 0, 1\}$ and $-1 \neq 1$ in $\mathbb{F}$, it expresses a linear Fourier inversion problem $\mathsf{Pr}\big[Y = \sum_{i=1}^{d} \theta(\lfloor \frac{X_i}{2} \rfloor)(-1)^{X_i}\big] \approx 1$ of a randomly picked example $(X, Y) \sim \{\big((2i + \frac{1-\mathcal{M}(i,j)}{2} \mid i \in [n), \mathcal{M}(i,j) \neq 0), \mathcal{M}(\kappa, j)\big)\}_{j=1}^{m}$. A more general problem asks to invert a secret $\theta \in \mathbb{Z}_q^n$ of a "noisy" degree-$k$ Fourier transform $\mathsf{Pr}\big[Y = \sum_{|w| \leq k} \hat{f}_w \prod_{i \in w} \theta(\lfloor X_i/2 \rfloor)(-1)^{X_i}\big] \approx 1$ of the known coefficients $\hat{f}_w \in \mathbb{F}$.

Review11: **Boolean Circuit lower bounds** have brought learnability[7], and vice versa

[3] We may sometimes consider $f(\theta_1 \circ x_1, \ldots, \theta_d \circ x_d)$ with different $\theta_i$ and say that a target $f$ hides $\theta \in \{0,1\}^{dn}$.

[4] LIN: Linear equations. LIN over $\mathbb{F}_2$ (the Galois field of order 2) is the same as the planted XOR.

[5] LPN: Learning Parity with Noise. LWE: Learning With Error.

[6] $\mathrm{GapSVP}_\gamma$ poses a lattice $\mathcal{L} \subset \mathbb{Z}^n$ together with an integer $d$ and asks to distinguish between $v(\mathcal{L}) \leq d$ and $v(\mathcal{L}) \geq \gamma d$ for the hidden shortest vector's length $v(\mathcal{L}) := \min_{z \in \mathcal{L} - \{0\}} \|z\|_2$.

[7] Learnable, compressible, distinguishable and derandomizable are equivalent notions under the uniform distribution in many computational complexity frontiers [CIKK16, Wil16, OS17, SCR$^+$20].

[NW94, IW97, FK09, Wil13, OS17]. Linial, Mansor, and Nisan [LMN93] derived $\mathsf{AC}^0$'s learnability from circuit lower bounds [Ajt83, Yao83, FSS84, Hås86]. They inverted REVIEW4's Fourier coefficients from quasi-polynomially many data of a low-degree polynomial over $\mathbb{R}$ derived from the $\mathsf{AC}^0$'s circuit lower bound. Carmosino, Impagliazzo, Kabanets, and Kolokolova [CIKK16] did it on $\mathsf{AC}^0[p]$ lower bounds [Raz87, Smo87] via Nisan-Wigderson's *pseudorandom generator* (PRG) [NW94]. Murray and Williams [Wil13, Wil14a, MW19] established $\mathsf{quasi\text{-}NP} \not\subset \mathsf{ACC}$ by learning $\mathsf{ACC}$ from quasi-polynomially many data thanks to Beigel-Tarui's low-degree $\mathrm{SYM}^+$-computation[8]. The polynomial method [Bei93, Wil14b, Hop18] was a consistent mechanics to make these lower bounds work.

REVIEW12: **Algebraic circuit lower bounds** have been interplaying with derandomization and learnability [SY10, CKW11, Sap14, KS19, GKS20]. Kabanets and Impagliazzo [KI04] derandomized PIT[9]. It plugged (unproved) exponential[10] circuit size lower bounds of explicit multilinear polynomials to Nisan-Wigderson's PRG. Low-rank patrial derivatives [Nis91, NW96, KS03, KSS14, GKKS14] have brought constant-depth circuit size lower bounds [SW01, RY09, KST16, KLSS17, KS17, LST21] and multilinear formulas [SS96, Raz09], derandomized the PIT of constant-depth circuits [KMSV13, SV18, LST21], non-properly learned multilinear depth-three circuits [BBB$^+$00, KS06a], and properly learned restricted depth-three circuits [Kay12, Sin16, KS19, GKS20, GMKP20].

Let us view these previous works of learning a dataset $\mathcal{D} = \{(x(1), y(1)), \ldots, (x(m), y(m))\}$ through the lens of smoothed analysis. First, SA1 chooses an adversarial variate (marginal) distribution $P(x) = \sum_y P(x, y)$. SA2 disturbs the $P(x)$ to $P(G(x))$ by a random perturbation $G \in \mathcal{G}$ while preserving the covariate distribution $P_\theta(y|x)$ intact. SA3 generates a dataset $\mathcal{D}$ from the perturbed distribution $P(G(x))P_\theta(y|G(x))$. Its *density* is the supremum of probability mass $\rho(G) = \sup\{\Pr[G(x)|x] \mid P(x) > 0\}$ [BV06, RV07, BM12], and its *min-entropy* is $\mathrm{H}_\infty(G) = -\log \rho(G)$. In particular, the worst-case complexity assumes $\mathrm{H}_\infty(G) = 0$, while a smoothed one $\mathrm{H}_\infty(G) \leq -\log|\mathcal{G}|$ with the equality $\mathrm{H}_\infty(G) = -\log|\mathcal{G}| \Leftrightarrow \forall g, \Pr[G = g] = 1/|\mathcal{G}|$. The shift $G$ may look at the data $\{x(j)\}_{j=1}^m$ as a vector in the product space $(x(1), \ldots, x(m)) \in \mathcal{X}^m$ and perturb each $x(j)$ by different marginals. For example, the i.i.d. $m$ data from the uniform distribution over $\{0,1\}^n$ have the min-entropy $\mathrm{H}_\infty(G) \approx mn$ by the $mn$ i.i.d. flippers $G_{ij}$ disturbing the $i$th dimension of the $j$th example.

The previous works have taken the following $\mathrm{H}_\infty(G)$ to reduce the worst-case complexity in smoothed analysis. REVIEWS 1 and 2 are exponentially hard at $\mathrm{H}_\infty(G) = 0$ but polynomial-time solvable under the Gaussian perturbation $\mathrm{H}_\infty(G) = \frac{n}{2}\log\frac{1}{2g\epsilon}$. REVIEW3's DNF learning is intractable[11] at $\mathrm{H}_\infty(G) = 0$ due to the hardness of learning [DSS16]'s canonical DNF in REVIEW8, but tractable under the perturbed product distribution $\mathrm{H}_\infty(G) = \Theta(nm)$, and even under the random walk $\mathrm{H}_\infty(G) = \Theta(m \log n)$. REVIEW6's 3SAT refutation is $\mathsf{coNP}$-complete at $\mathrm{H}_\infty(G) = 0$ [Coo71], but efficiently solvable under the flipper $\mathrm{H}_\infty(G) = \tilde{\Theta}(2^{n(\frac{1}{\epsilon}\log\frac{1}{\epsilon} + (1-\frac{1}{\epsilon})\log(1-\frac{1}{\epsilon}))})$. Similarly, REVIEW7's Max$k$SAT approximation is $\mathsf{NP}$-complete at $\mathrm{H}_\infty(G) = 0$ but tractable under the dense constraints $\mathrm{H}_\infty(G) = \Theta(kn^k \log n)$. Exceptionally, REVIEW9's LPN and LWE are still intractable even for the uniform random matrice $\mathrm{H}_\infty(G) = mn \log q$.

These worst-case intractable but average-case tractable problems separate unlearnable from learnable by $\mathrm{H}_\infty(G) = 0$ versus $\mathrm{H}_\infty(G) = \mathrm{poly}(n)$. Derandomization effort might reduce this

[8] $\mathrm{SYM}^+$: Boolean functions $g(f(x))$ of a polynomial $f$ over $\mathbb{Z}$ and $g : \mathbb{Z} \to \{0,1\}$. $\mathsf{quasi\text{-}NP}$: $\mathsf{NTIME}[2^{\log^{O(1)}(n)}]$.

[9] PIT: Polynomial Identity Test asks whether a given syntactic polynomial representation is identically zero.

[10] An "explicit" polynomial must have a polynomial-size circuit (possibly nondeterministic) computation [KS19].

[11] No polynomial-time algorithm can learn DNF from the only training dataset (without membership queries).

min-entropy gap for computational complexity separation to more tight $\mathrm{H}_\infty(G) = 0$ versus $\mathrm{H}_\infty(G) \le O(\log n)$. In that case, the known average-case algorithms might learn DNF, approximate Max$k$SAT, and invert LWE from the worst-case data with a slight perturbation. Even more, it might solve REVIEW10's matrix rigidity problem in smoothed analysis, giving rise to non-uniform circuit lower bounds beyond $\mathsf{quasi\text{-}NP} \not\subset \mathsf{ACC}$. It motivates us to investigate these smoothed complexities by fixing the min-entropy somewhere[12] between $0 \le \mathrm{H}_\infty(G) \le O(\log n)$.

In this paper, we prove the following Theorems 1.1–1.13 in the asymptotic analysis on the problem's increasing magnitudes $d, k, n, p, q, s$, and $1/\varepsilon$ under dominance[13] $k+\log(ds/\varepsilon) \ll \log n$, $s/\varepsilon \le d^{O(1)}$ and $1 \ll p, q \le n^{O(1)}$. Our learnability proofs of the smoothed analysis may pick an appropriate perturbation $G$. The unlearnablility ones must endure any considerable $G$.

When the min-entropy is zero (the worst case), the promise-Max$k$CSP of REVIEW7 must have the number of constraints between $n^{(1-\epsilon)k/2} \le m \le \tilde{O}(n^{k/2})$ for efficiency.

**Theorem 1.1** (promise-Max$k$CSP, informal)**.** For any $k$-variable predicate $f$, distinguishment between $|\max_\theta P(y = f(x)) - \max_\theta P'(y = f(x))| = \Omega(1)$ and $P(x, y) \equiv P'(x, y)$ on $n^{\frac{1-\epsilon}{2}k}$ data must take $\Omega(\exp(n^\epsilon))$ time[14] by giving access to both samplers $P(x, y)$ and $P'(x, y)$. Meanwhile, $\tilde{O}(n^{k/2})$ data can distinguish them in $n^{O(k)}$ time.

When the min-entropy grows to $\log s$, the planted $s$-term DNF becomes PAC learnable.

**Theorem 1.2** (PAC learning the planted DNF, informal)**.** Below $\mathrm{H}_\infty(G) \le (1-\epsilon)\log s$, PAC learning the planted $s$-term DNF on $n^{\Omega(\log s)}$ data must consume $\Omega(\exp(n^\epsilon))$ time. At $\mathrm{H}_\infty(G) = \log s + O(\log\log n)$, it becomes PAC learnable from $n^{\frac{1}{2}\log s + O(1)}$ data in $n^{O(\log s)}$ time.

Similarly, the agnostic learnability of the planted AND (Boolean conjunction) emerges at $\mathrm{H}_\infty(G) \approx \log(1/\varepsilon)$ to achieve the prediction accuracy $\max_\theta P(y = f(x)) + \varepsilon$.

**Theorem 1.3** (agnostically learning the planted AND, informal)**.** Below $\mathrm{H}_\infty(G) \le (1-\epsilon)\log\frac{1}{\varepsilon}$, agnostic learning the planted AND on $n^{\Omega(\log\frac{1}{\varepsilon})}$ data demands $\Omega(\exp(n^\epsilon))$ time. At $\mathrm{H}_\infty(G) = \log\frac{1}{\varepsilon} + O(\log\log n)$, it is agnostically learnable from $n^{\frac{1}{2}\log\frac{1}{\varepsilon}+O(1)}$ data in $n^{O(\log\frac{1}{\varepsilon})}$ time.

When the min-entropy goes beyond $\log s$, Theorem 1.2's data size barrier $n^{\Theta(\log s)}$ becomes breakable into a linear time of $n$ for the "monotone"[15] DNF with "expanding"[16] terms.

**Theorem 1.4** (PAC learning monotone DNF)**.** At $\mathrm{H}_\infty(G) = O(\log s)$, the planted monotone $\mathrm{DNF}_d$ with $c$-wisely $c'\log s$-expanding $s$ terms for large constants $c, c'$ is properly PAC learnable by inverting $\theta \in \{0,1\}^{dn}$ in $n \cdot \tilde{O}(s^{\log d})$ time on $n \cdot \mathrm{poly}(s)$ data with pairwisely dense attributes[17].

When the min-entropy reaches $O(\log n)$, even low-degree multi-linear polynomials may become "invertible" so properly learnable. We will investigate it for the planted *Fourier Transform* (FT) $f(\mathbf{x}) := \sum_{|w| \le k} \hat{f}_w \prod_{i \in w} \theta \circ \mathbf{x}_i$, $\theta \circ \mathbf{x}_i = \theta(\lfloor \mathbf{x}_i/2 \rfloor)(-1)^{\mathbf{x}_i}$ of REVIEW10. Our FT inversion algorithm can efficiently solve LPN and LWE with a binary secret $\theta$, so GapSVP, too.

---

[12] Our theorems (e.g., Theorem 1.1) assume $\mathrm{H}_\infty(G) = 0$ unless mentioning on $G$ nor $\mathrm{H}_\infty(G)$ in their statements.

[13] The dominance applies to only those parameters bounding the learning problem's magnitudes, say the dimension $d$ of the target concept, the number $s$ of terms in the target DNF, and the learning accuracy $\varepsilon$ to achieve.

[14] Theorems 1.1–1.3 claim $\Omega(\exp(n^\epsilon))$ lengths or $\Omega(n^\epsilon)$ degrees in several well-studied weak proof systems.

[15] In REVIEW3's terminology, $f \in \mathrm{DNF}$ is *monotone* if $i \in f_\kappa \cap f_{\kappa'} \Rightarrow f_{\kappa i} = f_{\kappa' i}$.

[16] We say that a DNF $f$ is $c$-wisely $k$-expanding if $|f_{\kappa_1} \cup \cdots \cup f_{\kappa_c}| \ge ck$ for every distinct $\kappa_1, \ldots, \kappa_c$.

[17] A random variable $X \sim [2n]^d$ has pairwisely dense attributes if $\forall (i \ne i'), \mathsf{Pr}[\lfloor X_i/2 \rfloor, \lfloor X_{i'}/2 \rfloor] \ge \Omega(\frac{1}{n^{1+\epsilon}})$.

**Theorem 1.5** (inverting degree-$k$ planted FT)**.** Let $1 \le k \le O(1)$, $n^{2+1/2^{k-1}} \ll q \in 2\mathbb{N}+1$, and $r = q^{1/2^{k+1}}$. At $\mathrm{H}_\infty(G) = O(\log(nq))$, the degree-$k$ planted FT $f$ over $\mathbb{Z}_q$ is invertible in $O(d^k n^{k+2} r^2)$ time on $O(n^{k+1}r^2)$ data of the following kind. The covariate must be as small as $|Y| \le r$. The variate must be as $k$-wisely sparse and noiseless at every location $(w,a) \in \binom{d}{k} \times [n]^k$ as $\mathsf{Pr}[\forall i \in w, \lfloor X_i/2 \rfloor = a_i] \ge \Omega(1/n^k)$ and $\mathsf{Pr}[Y \ne f(X) \mid \forall i \in w, \lfloor X_i/2 \rfloor = a_i] \ll 1/(nr)$.

**Theorem 1.6** (inverting LPN and LWE)**.** LPN and LWE over $\mathbb{Z}_p$ are breakable in polynomial time for any prime number $p \ge n^{\Omega(1)}$ and $O(1)$ size secrets $\forall i, |\theta_i| \le O(1)$ .

**Theorem 1.7** (breaking GapSVP)**.** $\mathrm{GapSVP}_{\tilde{O}(n^2)}$ is breakable in polynomial time.

Further, Theorem 1.5 can solve Review10's matrix rigidity problem and derive "natural" circuit lower bounds [RR97, Wil16, SCR+20] in the following sense. Perturb an $\sqrt{N} \times \sqrt{N}$ matrix by a shift $G$ that preserves the *density* $\rho(\mathcal{M}) := |\mathcal{M}|_{\ne 0}/N = |\{(i,j) \mid \mathcal{M}(i,j) \ne 0\}|/N$. We say that an algorithm $\mathcal{A}$ learns the matrix $\mathcal{M}$ under $G$ if $\mathcal{A}$ feeds the first $o(N^2)$ entries of the perturbed matrix $G(\mathcal{M})$ and predicts $\mathsf{Pr}_{G,I,J}[\mathcal{A}(I,J) = G(\mathcal{M})(I,J)] \approx 1$. Our natural lower bounds claim that all small-density matrices must have a large circuit size or fast learning time, so denying the existence of pseudorandom bits[18] emitted from the tiny circuits. In this sense, we will establish super-linear size lower bounds against algebraic circuits to compute quadratic polynomials over finite fields [Val77, Lok08, SY10] and communication complexity lower bounds beyond the polynomial hierarchy [BFS86, Wun12, GPW18].

**Theorem 1.8** (non-linear size lower bound)**.** At $\mathrm{H}_\infty(G) = O(\log n)$, the bilinear form of any $n \times n$ $\{-1,0,1\}$-matrix having density $n^{-o(1)}$ requires $\Omega\big(n(\log\log n)^{1-\epsilon}\big)$ size algebraic $\mathsf{NC}^1$ circuits[19] over $\mathbb{F}_p$ of any prime $p \ge n^{\Omega(1)}$, unless it is learnable in $n^{o(1)}$ time.

**Theorem 1.9** ($\mathsf{PH}^{\mathsf{cc}}$'s sub-linear depth lower bound)**.** At $\mathrm{H}_\infty(G) = O(n)$, any $2^{n/2}$ by $2^{n/2}$ $\{-1,0,1\}$-matrix of density $\exp(-n^{\Omega(1)})$ forces any $\mathsf{PH}^{\mathsf{cc}}$ protocol[20] to have depth $n^{\Omega(1)}$ unless it is learnable in $\exp(n^\epsilon)$ time.

Theorem 1.9 can derandomize the unsatisfiability of Williams's circuits [Wil13, Wil14a] to verify a short PCP [BSV14] plugged into an easy-witness lemma [MW19, CR20], yielding:

**Theorem 1.10** ($\mathsf{PH} \ne \mathsf{PSPACE}$ in the communication)**.** $\mathsf{PH}^{\mathsf{cc}} \ne \mathsf{PSPACE}^{\mathsf{cc}}$ or $\mathsf{NP} \not\subset \mathsf{DEP}[k \log n]$[21].

Similarly, we will establish new natural lower bounds to make Williams's approach succeed in the following breakthrough separations of Review11's Boolean complexity [Weg87, VL91, Pap03, AB09, Aar16] and Review12's algebraic complexity [Val79, Sap14, Wig19].

**Theorem 1.11** (deep network $\ne$ NP)**.** $\mathsf{quasi\text{-}NP} \not\subset \mathsf{TC}^0$.

**Theorem 1.12** ($\mathsf{P} \ne \mathsf{NP}$ in algebra)**.** $\mathsf{VP} \ne \mathsf{VNP}$ or $\forall k \ge 1, \mathsf{quasi\text{-}NP} \not\subset \mathsf{NC}^k$.

**Theorem 1.13** (derndomizing PIT)**.** Either PIT is solvable in deterministic $n^{\mathrm{poly}(\log\log n)}$ time, or $\forall \epsilon > 0, \forall k \ge 1, \mathsf{NTIME}[2^{n^\epsilon}] \not\subset \mathsf{SIZE}[n^k]$.

[18] We allow pseudorandom bits to be unbalanced (i.e., #1-bits $\ll$ #0-bits) by assuming a fixed structure over balanced bits, e.g., taking the $k$-wise conjunctions over balanced $nk$-bits to get unbalanced $n$ bits of density $\frac{1}{2^k}$.

[19] Algebraic circuits compute either $+$ or $\times$ of syntactic polynomials over a field.

[20] A protocol calculates $\mathcal{M}(i,j)$ by communication between the two parties knowing only $i$ or $j$.

[21] $\mathsf{DEP}[d]$ is a language class computed by a series of non-uniform binary-fanin circuits of depth $d$.

We describe these theorems more formally in Theorems 1.14–1.30 with related notations and previous works not mentioned in the Reviews. We will newly issue all of them in this paper.

**Shifts in smoothed analysis:** Let us call the SA2's perturbation $G \in \mathcal{G}$ a *shift*. It must satisfy[22] $P_\theta(G(z_i)) = P_{\hat{G}(\theta)}(z_i)$ at every $i$th dimension, as the previous works used to have. Review1's Gaussian shift causes $\hat{G}(\mu_i, \sigma_i^2) := \big(\mu_i + \mu(G(z_i)), \sigma_i^2 + \sigma^2(G(z_i))\big)$. Review3's mean shift $\hat{G}(\mu_i) = \mu_i + \hat{G}_i$ stems from the continuous data shift $G(z_i) = z_i - \hat{G}_i$ over the real-value interval $z_i \in [0,1]$ through the sigmoidal function $x_i = \big(\mathrm{sgn}(z_i - \mu_i) + 1\big)/2 \in \{0,1\}$. Review6's polarity[23] flipper induces $\hat{G}(\theta)(x_i) = \theta(\lfloor x_i/2 \rfloor) \oplus G(\lfloor x_i/2 \rfloor)$, $x_i \in [2n] := \{0, 1, \ldots, n-1\}$.

Our smoothed analysis will focus on Review8's planted functions. We will employ the most general shift satisfying both *robustness* $\{x_i \mapsto \theta \circ \big(G(x_i)\big)\}_\theta = \{x_i \mapsto \theta \circ x_i\}_\theta$ and *symmetry* $\theta \circ \big(G(x_i)\big) = \hat{G}(\theta) \circ x_i$. These two notions are equivalent, inducing a unique decomposition $G = (\Phi, \Psi)$ to an attribute permuter $\Phi \in \mathbb{S}_n^d$ and a polarity flipper $\Psi \in \{0,1\}^{dn}$ such that $G(x_i) := 2\Phi(\lfloor x_i/2 \rfloor) + \Psi(\lfloor x_i/2 \rfloor)$ and $\hat{G}(\theta)(x_i) := \theta\big(\Phi(\lfloor x_i/2 \rfloor)\big) \oplus \Psi(\lfloor x_i/2 \rfloor)$. A shift $G$ is *uniform* if it is the same over examples as $x(j) = x(j') \Rightarrow G(x(j)) = G(x(j'))$. This paper considers non-uniform shifts of the vectors $(x(j))_{j=1}^m \in \mathcal{X}^m$ unless specified as uniform.

**PAC learning the planted DNF in weak axiomatic proof systems:** Theorem 1.2's lower bound supposes the PAC learner to reside in bounded proof systems. The learner observes a training dataset $\mathcal{D}$ drawn from the unknown target distribution $P(x, y)$ and must choose a hypothesis $h$ predicting $\mathrm{err}_h(\mathcal{D}) := P(h(x) \neq y) \approx 0$ whenever $\mathrm{err}_f(\mathcal{D}) = 0$. In addition, it obliges the learner to prove $\mathrm{err}_f(\mathcal{D}) = 0 \to \mathrm{err}_h(\mathcal{D}) \approx 0$ in the following axiomatic systems. We study *resolution* (Res) [DP60, DLL62, Rob65, BSW01, MMZ$^+$01, AM20], *polynomial calculus* (PC) [CEI96, BIK$^+$96, IPS99, ABSRW02, LNSS20], *Sum-of-Squares* (SoS) [Ste74, Sho87, Nes00, GV01, Par00, Las01, Lau09, BS14, LRS15, HKP$^+$17, AH19, BHK$^+$19], LP *extended formulation* (LP) [Yan91, CLRS16, KMR17, BCR20], and *extended Frege* [CR79, Bus91, Kra95, BP98, Bus12, BBCP20]. Theorem 1.2 will measure the proof complexity of DNF's learnability on these proof systems. When the data is noisy, the learner must endure a slight amount of malicious noise $\mathrm{err}_f(\mathcal{D}) \approx 0$ [Val85, KL93, CBDF$^+$99, KLS09, ABL17, DKS18, DKK$^+$18].

Historically, PAC learning DNF in "polynomial time" had been a fundamental challenge posed by Valiant [Val84, Val85]. Unless $\mathsf{RP} \neq \mathsf{NP}$, it is hard to properly PAC learn $s$-term $k$DNF for various specific (and unspecific) $s$ and $k$ [Val84, Val85, PV88, ABF$^+$08, KS08, Fel09, GS21], where the proper learner must choose a hypothesis from the $s$-term $k$DNF or the kindred classes. The fastest "non-proper" $s$-term DNF learning time is $n^{O(n^{1/3} \log s)}$ [Bsh96, TT99, KS04]. Recently, Daniely and Shalev-Shwartz (DSS) [DLSS14, DSS16] dashed out hope for DNF's non-proper learnability as follows: Any PAC learner of the Review8's canonical planted $k$DNF with $k = \omega(1)$ must spend $n^{(1-\epsilon)k/2}$ examples unless he can refute the R$k$SAT with that many constraints. This assumption is the so-called Feige's hypothesis [Fei02, BKS13], on which many problems rely (or challenge) their average-case hardness [Ale11, DSS16, HS17, DJ19, VW21].

In this paper, we will establish the PAC learning hardness of the planted DNF as follows by bringing the Daniely and Shalev-Shwartz reduction into the weak axiomatic proof systems.

**Theorem 1.14** (hardness of learning planted $k$DNF)**.** For $k \geq 3$, PAC learning the planted $k$DNF under the uniform distribution must consume $\Omega\big(n^{\frac{1-\epsilon}{2}k}\big)$ data; otherwise, all of its SoS degree, PC degree, and Res size must be $\Omega(n^\epsilon)$, $\Omega(n^\epsilon)$, and $\Omega\big(\exp(n^\epsilon)\big)$. Similarly, the noisy planted $k$DNF

[22] We may write $g_i(z_i)$ as $g(z_i)$ or $g_i(z)$ for a function $g = (g_i)_i$ over a domain $\mathcal{Z} = \prod_i \mathcal{Z}_i$ composed of $g_i$ over $\mathcal{Z}_i$.
[23] We refer to $x_i \bmod 2 \in \{0,1\}$ and $\lfloor x_i/2 \rfloor \in [n]$ as the *polarity* and *attribute* of a variate $x_i \in [2n]$.

demands the same sample size $\Omega\big(n^{\frac{1-\epsilon}{2}k}\big)$, otherwise both SoS-degree $\Omega(n^{\epsilon})$ and LP-size $2^{\Omega(n^{\epsilon})}$.

**Theorem 1.15** (hardness of learning DNF)**.** PAC learning the planted $s$-term DNF under the uniform distribution must spend $\Omega(n^{\frac{1-\epsilon}{2}\log s})$ data; otherwise, all of its SoS degree, PC degree, and Res size must be $\Omega(n^{\epsilon})$, $\Omega(n^{\epsilon})$, and $\Omega\big(\exp(n^{\epsilon})\big)$, respectively. Similarly, the noisy planted $s$-term DNF needs the same sample size, unless SoS-degree $\Omega(n^{\epsilon})$ and LP-size $\Omega\big(\exp(n^{\epsilon})\big)$.

**Theorem 1.16** (Theorem 1.2, hardness)**.** At $\mathrm{H}_{\infty}(G) = (1-c)\log s$, $0 < c < 1$, PAC learning the planted $s$-term DNF hiding $\theta \in \{0,1\}^{dn}$ under the uniform distribution needs $\Omega(n^{\frac{c}{10-4\log c}\log s})$ data. Otherwise, both the SoS and PC degrees must be $\Omega(n^{0.06})$. Similarly, the noisy planted $s$-term DNF needs that sample size, unless SoS-degree $\Omega(n^{0.06})$ and LP-size $\Omega(\exp(n^{0.06}))$.

Furthermore, we will establish the opposite direction of the Daniely and Shalev-Shwartz reduction: The known R$k$SAT refutation algorithms can transform into PAC learning ones. Allen, O'Donnell, and Witmer [COCF10, AOW15, BM16] succeeded in a spectrally optimal R$k$SAT refutation via quadratic programming based on symmetric Grothendieck's inequality [Gro52, CW04, ABE+05, AN06]. Abascal, Guruswami, and Kothari [AGK21] did it for the malicious constraints perturbed by the random polarities of Review6. We will translate them to PAC learning algorithms working under the adversarial constraints and polarities.

**Theorem 1.17** (PAC learning planted $k$DNF)**.** For any $k \geq 2$, the planted $k$DNF hiding $\theta \in \{0,1\}^{dn}$ is distribution-free PAC learnable from $\tilde{O}(n^{\lceil k/2 \rceil})$ data in $n^{O(k)}$ time.

**Theorem 1.18** (Theorem 1.2, algorithms)**.** At $\mathrm{H}_{\infty}(G) = \log s + O(\log\log n)$, the planted $s$-term DNF of $\theta \in \{0,1\}^{dn}$ is distribution-free PAC learnable on $n^{\frac{1}{2}\log s + O(1)}$ data in $n^{O(\log s)}$ time.

In summary, in the worst-case PAC learning, the known spectral threshold $\frac{\log m}{\log n} \approx k/2$ of the R$k$SAT refutation on $m$-constraint transfers to the planted $k$DNF learning on $m$-data. In smoothed analysis, learning the planted $s$-term DNF on $n^{\Theta(\log s)}$ data becomes tractable when the min-entropy $\mathrm{H}_{\infty}(G)$ becomes comparable to the logarithm of the problem size (i.e., $\log s$):

PAC1: $\mathrm{H}_{\infty}(G) = 0$ takes $n^{O(d^{1/3}\log s)}$ learning time by the current best algorithm [KS04, RS10a].

PAC2: $\mathrm{H}_{\infty}(G) = 0$ requires $2^{\Omega(n^{\epsilon})}$ time to learn $O(n^{\frac{1-\epsilon}{2}\log s})$ data under the uniform distribution.

PAC3: $\mathrm{H}_{\infty}(G) = c\log s$ with $c < 1$ still demands sub-exponential time for $n^{\Omega(\log s)}$ data.

PAC4: $\mathrm{H}_{\infty}(G) = \log s + O(\log\log n)$ enables us to learn any $n^{\frac{1}{2}\log s + O(1)}$ data in $n^{O(\log s)}$ time.

**Agnostically learning the planted AND (a.k.a., planted Boolean conjunct) in weak axiomatic proof systems:** In Review5's agnostic model, the learner must search a hypothesis $h$ and its proof competing with $\eta = \min_f \mathrm{err}_f(\mathcal{D})$ by accuracy $\varepsilon$ to achieve $\mathrm{err}_f(\mathcal{D}) \leq \eta \to \mathrm{err}_h(\mathcal{D}) \leq \eta + \varepsilon$ for any malicious noise rate $\eta \leq 1/2 - 2\varepsilon$ [BEHW89, Hau92, KSS94, Vap06]. Unfortunately, even the AND function is already too complex to agnostically learn properly [AL88, KL93, Fel06, GR09, FGRW12, GS21] and non-properly [FK15, DSS16, DJ19].

We will translate the PAC model Theorems 1.14–1.18 to establish the following agnostic ones of leaning the planted AND, XOR, $k$AND, $k$XOR, and $k$JUNTA[24].

**Theorem 1.19** (hardness of agnostically learning planted AND)**.** For $2 \leq d \leq \log\frac{1}{\varepsilon} - O(1)$, agnostically learning the planted $\mathrm{AND}_d$ under the uniform distribution must consume $\Omega\big(n^{(1-\epsilon)d/2}\big)$ data. Otherwise, its SoS degree must be $2^{\Omega(n^{\epsilon})}$.

[24] $\mathrm{XOR} := \mathrm{XOR}_d = \{\bigoplus_{i\in w}\mathbf{x}_i \mid w \subset (d]\}$. $k\mathrm{JUNTA} := \{f_k(\mathbf{x}_i, i \in w) \mid w \subset (d], |w| \leq k, f_k : \{0,1\}^k \to \{0,1\}\}$.

**Theorem 1.20** (hardness of agnostic learning planted XOR)**.** For $2 \le d$, agnostic learning the planted $\mathrm{XOR}_d$ under the uniform distribution demands $\Omega\big(n^{(1-\epsilon)d/2}\big)$ data or $2^{\Omega(n^{\epsilon})}$ SoS degree.

**Theorem 1.21** (agnostically learning planted $k$JUNTA)**.** The planted $k$JUNTA is agnostically learnable from $\tilde{O}(n^{\lceil k/2 \rceil})$ data under any distribution in $n^{O(k)}$ time.

**Theorem 1.22** (Theorem 1.3, hardness)**.** At $\mathrm{H}_\infty(G) = c \log \frac{1}{\varepsilon}$, $c > 0$, agnostically learning the planted $\mathrm{AND}_d$ hiding $\theta \in \{0,1\}^{dn}$ under the uniform distribution must consume $\Omega\big(n^{\frac{\log(1/\varepsilon)}{10+4\log(c+1)}}\big)$ data. Otherwise, its SoS degree must be $\Omega(n^{0.06})$.

**Theorem 1.23** (Theorem 1.3, algorithms)**.** At $\mathrm{H}_\infty(G) = \log \frac{1}{\varepsilon} + O(\log \log n)$, the planted AND hiding $\theta \in \{0,1\}^{dn}$ is distribution-free agnostic learnable from $\eta$-noisy $n^{\frac{1}{2}\log\frac{1}{1-2\eta}+O(1)}$ data in $n^{O(\log \frac{1}{1-2\eta})}$ time.

In summary, agnostically learning the planted $\mathrm{AND}_d$ within accuracy $\varepsilon$ from $n^{\Theta(\log 1/\varepsilon)}$ data becomes tractable when $\mathrm{H}_\infty(G)$ reaches the learning accuracy's entropy (i.e., $\log(1/\varepsilon)$):

AGN1: $\mathrm{H}_\infty(G) = 0$ takes $n^{O(d^{1/2}\log n)}$ learning time by the current best algorithm [KKMS08].

AGN2: $\mathrm{H}_\infty(G) = 0$ requires $2^{\Omega(n^{\epsilon})}$ time to learn $O(n^{\frac{1-\epsilon}{2}\log\frac{1}{\varepsilon}})$ data under the uniform distribution.

AGN3: $\mathrm{H}_\infty(G) = c \log \frac{1}{\varepsilon}$ of $c > 0$ still demands sub-exponential time for $n^{\Omega(\log(1/\varepsilon))}$ data.

AGN4: $\mathrm{H}_\infty(G) = \log \frac{1}{\varepsilon} + O(\log\log n)$ enables us to learn any $n^{\log\frac{1}{\varepsilon}+O(1)}$ data in $n^{O(\log\frac{1}{\varepsilon})}$ time.

**Approximate Promise-Max$k$CSP in weak proof systems:** Theorems 1.19–1.21 imply the sample complexity $\Omega(n^{\frac{1-\epsilon}{2}k}) \le m \le \tilde{O}(n^{\lceil k/2 \rceil})$ of the following problem: For a predicate $f(\mathbf{x}_1, \dots, \mathbf{x}_k)$, prove $|\mathrm{acc}(P^m(x,y)) - \mathrm{acc}((P')^m(x,y))| \le \frac{3}{4}(\beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}}) \to P(x,y) \equiv P'(x,y)$ under a promise that either $\max_\theta P(y = f_\theta(x)) \ge \beta_{\mathtt{cmp}} > \beta_{\mathtt{snd}} \ge \max_\theta P'(y = f_\theta(x))$ or $P(x,y) \equiv P'(x,y)$ must hold. We call it $(\beta_{\mathtt{cmp}}, \beta_{\mathtt{snd}})$-gap (or $(\beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}})$-gap) approximation of the promise-Max$k$SAT, promise-Max$k$XOR, and promise-Max$k$CSP when $f = \bigoplus_{i=1}^k \mathbf{x}_i$, $f = \bigwedge_{i=1}^k \mathbf{x}_i$, and $f : \{0,1\}^k \to \{0,1\}$, respectively. Recently, Abascal, Guruswami, and Kothari [AGK21] established the matching upper bound $\tilde{O}(n^{k/2})$ of the Max$k$CSP under the random polarities, which brings out that of the promise-Max$k$CSP (Theorem 1.26), too. Let $\triangle := \beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}}$.

**Theorem 1.24** (Theorem 1.1, hardness)**.** Any gap ($> 4^{-k}$) approximation of promise-Max$k$SAT under the marginally uniform distribution[25] requires $\Omega(n^{\frac{1-\epsilon}{2}k})$ constraints or $\Omega(n^{\epsilon})$ SoS-degree.

**Theorem 1.25** (approximation hardness of promise-Max$k$XOR)**.** Any gap ($> 2^{-k-1}$) approximation of the promise-Max$k$XOR under a marginally uniform distribution requires $\Omega\big(n^{\frac{1-\epsilon}{2}k}\big)$ constraints unless its SoS degree is $\Omega(\exp(n^{\epsilon}))$.

**Theorem 1.26** (Theorem 1.1, algorithms)**.** The promise-Max$k$SAT is $\triangle$-gap approximable from $\tilde{O}(n^{k/2}/\triangle^5)$ constraints under any distribution in $n^{O(k)}$ time. So is the promise-Max$k$CSP from $\tilde{O}(n^{k/2}(2^k/\triangle)^5)$ constraints in $n^{O(k)}$ time, too.

**Theorem 1.27** (approximation hardness of the promise-MaxSAT in smoothed analysis)**.** At $\mathrm{H}_\infty(G) = c \log \frac{1}{\varepsilon}$ and $1 - (2\varepsilon)^{c+1} \le \beta_{\mathtt{snd}} < \beta_{\mathtt{cmp}} - 4^{-k}$, any $(\beta_{\mathtt{cmp}}, \beta_{\mathtt{snd}})$-gap approximation of the promise-MaxSAT under the marginally uniform distribution perturbed by any flipper $G$ requires $\Omega\big(n^{\frac{\log(1/\varepsilon)}{10+4\log(c+1)}}\big)$ constraints unless its SoS degree is $\Omega(n^{0.06})$.

[25] A joint-distribution $P(x,y)$ is *marginally uniform* if it does not depend on $x$ but may depend on $y$.

**Inverting monotone DNF, degree-$k$ Fourier transforms, and LWE:** When the min-entropy reaches $\mathrm{H}_\infty(G) = s^{O(1)}$, even the data-size barrier $n^{\Theta(\log s)}$ persistent through PAC 2–4 in learning the planted $s$-term DNF becomes breakable for "monotone" functions. Theorem 1.4 properly learns the monotone planted DNF in almost-linear time by inverting the unknown parameter $\theta$ in the following manner. After substituting arbitrary values but leaving a single variable $\mathbf{x}_i$ intact, a monotone function $f(\mathbf{x}_1, \dots, \mathbf{x}_d)$ collapses to always $\mathbf{x}_i$ or $\neg\mathbf{x}_i$ unless it collapses to the constants 0 or 1. Accordingly, the *correlation* $\mathbb{E}[(-1)^{X_i + f(\theta \circ X)}]$ under $\lfloor X_i/2 \rfloor = a$ could detect either $(-1)^{X_i + \theta \circ X_i} = (-1)^{\theta(a)}$ or $(-1)^{X_i + \neg\theta \circ X_i} = -(-1)^{\theta(a)}$ exclusively so that the statistical correlation analysis over the filtered dataset $\big\{(x, y) \in \mathcal{D} \mid \lfloor x_i/2 \rfloor = a\big\}$ could invert the hidden parameter $\theta(a)$. Notice that the correlation might diminish to the statistical zero if the $\mathbf{x}_i$ were a non-monotone variable of $f$. This correlation statistics gives rise to Theorem 1.4.

Similarly, suppose the target is Review10's FT: $f(\mathbf{x}) := \sum_{|w| \le k} \hat{f}_w \prod_{i \in w} \theta(\lfloor \mathbf{x}_i/2 \rfloor)(-1)^{\mathbf{x}_i}$. Observe the outcomes over the restricted data $\forall i \in w, \lfloor X_i/2 \rfloor = a_i$ on a "query" $(w, a) \in \binom{d}{k} \times \{0,1\}^w$. It collapses the target function to various subfunctions $f(\mathbf{x}_w) : \{0,1\}^w \to \mathbb{F}$, inducing the same Fourier coefficient $\sum_{x_w \in \{0,1\}^w} f(x_w) \prod_{i \in w} (-1)^{x_i} \approx 2^w \hat{f}_w \prod_{i \in w} \theta(a_i)$ independently of the different subfunctions. In this manner, the correlation analysis $\mathbb{E}[f(X)(-1)^{\sum_{i \in w} X_i}]$ over the filtered dataset $\{(x, y) \in \mathcal{D} \mid \forall i \in w, \lfloor x_i/2 \rfloor = a_i\}$ may retrieve the hidden $\prod_{i \in w} \theta(a_i)$. The correlation might vanish if $w$ were not maximal, i.e., $\exists v \supsetneq w, \hat{f}_v \neq 0$. It can invert even LWE and GapSVP due to Yao, Toda, Beigel, and Tarui's modulus amplification [Yao90, Tod91, BT94]. LPN and LWE of Review9 ask to invert the random LP instance under strictly bounded additive i.i.d. (Bernoulli or Gaussian) noise. A smoothed analysis can invert the hidden secret even under any "unbounded" additive i.i.d. noise:

**Theorem 1.28** (inverting LWE in smoothed analysis)**.** For constants $1 \le c \ll k$ and an odd prime $p \gg n^{\Omega(1)}$, the LP instance $y(j) = \sum_{i=1}^n \theta_i \cdot G(x_i(j)) + E(j)$ of any matrix $(x_i(j))_{i,j} \in [p)^{nm}$ contaminated by any i.i.d. noises $E(1), \dots, E(m) \in \mathbb{Z}_p^n$ is invertible with high confidence to retrieve the secret $\theta \in \{-c, \dots, c\}^n$ in $\mathrm{poly}(n)$ time under the following shift $G \in \{0,1\}^{nm(p-1)/2}$. It flips the matrix $x$ by $G(x_i(j)) = \lfloor x_i(j)/2 \rfloor \cdot (-1)^{x_i(j) + G(\lfloor x_i(j)/2 \rfloor)}$ such that the random column $(G(x_i(J)))_{i=1}^n$ is as $k$-wisely sparse and uniform at $\forall w \in \binom{n}{k}$ and $\forall b \in \mathbb{Z}_p$ as $\mathsf{Pr}\big[\forall i \in w, \lfloor \frac{x_i(J)}{2} \rfloor = 1\big] \ge \Omega((\frac{2}{p})^k)$ and $\mathsf{Pr}\big[\sum_{i \notin w} G(x_i(J)) + E(J) = b \mid \forall i \in w, \lfloor \frac{x_i(J)}{2} \rfloor = 1\big] \approx \frac{1}{p}$.

We should note that Theorem 1.7's GapSVP's decryption [Reg04, Pei09, BLP+13] demands $m = \mathrm{poly}(n)$ amount of data to Theorem 1.28, while the cryptographic LWE allows no larger than $m \le O(n \log p)$ data for safety [GPV08, Reg09, LPR13, Pei14, BV14, ACD+18].

**Natural circuit lower bounds in smoothed analysis:** Theorem 1.5, armed with the modulus amplification, can solve Review10's matrix rigidity and derive natural lower bounds in Theorems 1.8–1.10. A natural lower bound against a circuit class $\mathcal{F}$ entails an efficient algorithm that distinguishes between the truth table of a small $\mathcal{F}$-circuit and the uniform random one. Razborov and Rudich [RR97] proved that such lower bounds deny the existence of PRG emitting the pseudorandom bits from a small circuit in class $\mathcal{F}$. In this sense, the natural lower bounds are too weak to support cryptography.

Theorems 1.8 demonstrates a natural super-linear lower bound to learn the quadratic polynomials. Historically, algebraic circuits [Val79, SY10] have enjoyed explicit lower bounds, e.g., super-linear lower bounds of degree-$\omega(1)$ polynomials on the general circuits [Str73, BS83], super-polynomial lower bounds of permanent and determinant on multilinear formulas [Raz06, Raz09], cubic lower bounds on formula size based on Nechiporuk's argument [Nec66, Kal85], $\tilde{\Omega}(n^{2.5})$ lower bounds on depth-4 circuits [Sha17, GST20], and super-polynomial lower bounds on constant-depth circuits [SS96, Raz10, LST21]. However, super-linear lower bounds of constant-degree (e.g., quadratic) polynomials against $\mathsf{NC}^1$ circuits are still unknown. Valiant's seminal work [Val77] has already presented them for rigid matrices, although their explicit construction is not yet known [Lok08, AW17]. Theorem 1.5 can supply a learning algorithm to it and derive Theorem 1.8. Baur-Strassen's partial derivates [BS83] translates a lower bound of a matrix $\mathcal{M}$ to a lower bound of the bilinear form $\sum_{i,j} \mathbf{x}_i \mathcal{M}(i,j)\mathbf{x}_j$.

Theorems 1.9 establishes a natural sub-linear depth lower bound to learn $\mathsf{PH}^{\mathsf{cc}}$, the communication complexity class[26] corresponding to the polynomial hierarchy. Structural communication complexity [BFS86, Wun12, GPW18] has succeeded in separating primitive complexity classes [27] , e.g., $\mathsf{BPP}^{\mathsf{cc}} \not\subset (\mathsf{P}^{\mathsf{NP}})^{\mathsf{cc}}$ [PSS14], $(\mathsf{P}^{\mathsf{MA}})^{\mathsf{cc}} \not\subset \mathsf{UPP}^{\mathsf{cc}}$ [RS10b, CM17], $\mathsf{MA}^{\mathsf{cc}} \not\subset (\mathsf{ZPP}^{\mathsf{NP}[1]})^{\mathsf{cc}}$ [GPW18], $\mathsf{AM}^{\mathsf{cc}} \cap \mathsf{coAM}^{\mathsf{cc}} \not\subset \mathsf{UPP}^{\mathsf{cc}}$ [Kla11, BCH+19]. However, no explicit lower bounds are known for $\mathsf{PH}^{\mathsf{cc}}$ and even a much smaller $\mathsf{AM}^{\mathsf{cc}} \cap \mathsf{coAM}^{\mathsf{cc}}$ [GPW18]. Razborov [Raz89] presented super-$\mathsf{PH}^{\mathsf{cc}}$ lower bounds of rigid matrices. Again, Theorem 1.5's learning algorithm turns Razborov's lower bounds to those of the $h$-alternating protocols of $2^{n/2} \times 2^{n/2}$ matrices:

**Theorem 1.29** (Theorem 1.9)**.** Let $\log n \ll d \ll n^{\epsilon/h}$. At $\log\big(\mathrm{H}_\infty(G)\big) = O(n)$, any $\{-1,0,1\}$-matrix of density $\Omega(2^{-d^{2h+4}})$ demands depth $d$ for $\mathsf{PH}^{\mathsf{cc}}_h$ unless it is learnable in $O(2^{d^{2h+4}})$ time.

Theorem 1.5's learning algorithm derives even Theorem 1.10, separating either $\mathsf{PSPACE}$ from $\mathsf{PH}$ in communication complexity or quasi-$\mathsf{NP}$ from parallel-$\mathsf{P}$ in circuit complexity. The former is a fundamental open problem in communication complexity classes [BFS86, GPW18], matrix rigidity [Wun12], margin complexity of data classifiers (e.g., support vector machine) [LS09], and graph complexity [PRS88, Juk12]. The latter is a lower bound beyond the class[28] $\mathsf{NC}$ containing cryptographic primitives [GGM86, KV94, Kha95, IN96]. Theorem 1.10 is a fruit of Williams's algorithmic approach [Wil13, Wil14a]. It is a reduction from the uniform time unary language hierarchy [Žák83] to the unsatisfiability of a small depth circuit through Ben-Sassen and Viola's short PCP [BSGH+06, BSV14] armed with an easy witness lemma for circuit depth [NW96, CR20] derived from Sudan, Trevisan, and Vadhalan's PRG [STV01]. Theorem 1.5 can solve this circuit unsatisfiability problem as follows. Let CMD (Connected Matrix Determinant) be an explicit language in $\mathsf{PSPACE}^{\mathsf{cc}}$, computing the modulo-2 determinant of the connected matrix $\mathcal{M}$, i.e., $\mathcal{M}(i,j) \in \{0,1\}$ and $i-j \geq 2 \Rightarrow M(i,j) = 0$.

**Theorem 1.30** (Theorem 1.10)**.** CMD $\notin \mathsf{PH}^{\mathsf{cc}}$ or $\mathsf{quasi\text{-}NP} \not\subset \mathsf{quasi\text{-}NC}^k$.

**Natural circuit lower bounds in worst-case analysis:** We will provide the new natural lower bounds of Theorems 1.11–1.13. Previously, Boolean circuits size has enjoyed explicit lower bounds, e.g., $5n$ lower bound for unrestricted circuit model [Blu83, IM02], exponential lower bounds for monotone circuits [Raz85, AB87], $\mathsf{AC}^0$ [Ajt83, FSS84, Yao85, Hås86], and $\mathsf{AC}^0[p]$ [Raz87, Smo87]. After 30 years of silence, Murray and Williams broke this $\mathsf{AC}^0[p]$ lower bound barrier, establishing $\mathsf{quasi\text{-}NP} \not\subset \mathsf{ACC}$ [Wil13, Wil14a, MW19].

Theorem 1.11 is another fruit of Williams's program obtained by providing a new worst-case learning algorithm of $\mathsf{TC}^0$. As far as we know, this is the first explicit ($\mathsf{quasi\text{-}NP}$) lower bound against the class $\mathsf{TC}^0 = \mathsf{AC}^0[\mathrm{SYM}]$[29] executing the basic arithmetic operations [Weg87, HAB02, Vol16], PRG [KL01, NR04, BPR12, AR16], cryptographic primitives [Kha95, BGI+12, AGS21],

[26] $\mathcal{F}^{\mathsf{cc}}$ denotes the two-party communication correspondence of a structural complexity class $\mathcal{F}$.

[27] $\mathsf{BPP}$, $\mathsf{ZPP}$, and $\mathsf{UPP}$ are probabilistic polynomial-time computations with bounded, zero, and unbounded errors. $\mathsf{AM}/\mathsf{MA}$ are those with bounded error to verify a proof that may/never depend on the verifier's randomness.

[28] quasi-$\mathcal{F}$ is a class of problems (circuits) $\mathcal{F}$ with the magnitude of time (size) $2^{(\log n)^{O(1)}}$.

[29] $\mathsf{AC}^0[\mathrm{SYM}]$ consists of the constant-depth circuits arming all symmetric gates of unbonded fan-in.

and even deep neural networks [Dan17, Sha18, VS20, MYSSS21, VRPS21]. Previously, the constant-depth MOD$[m]$ circuits have succeeded in efficiently simulating OR [BBR94] and even MAJ by a composite number $m$ of $O(\log n)$ distinct primes [Tsa96, BGL06, OSS19, CW21]. Yao, Beigel, and Tarui simulated $\mathsf{AC}^0[m]$ by $\mathrm{SYM}^+ = \mathrm{SYM} \circ \mathrm{AND}_d$ of quasi-polynomially large degree $d$ [Yao90, BT94]. Our new learning algorithm will do it even for the depth-$h$ $\mathsf{TC}^0$:

**Lemma 1.31** ($\mathsf{TC}^0 \subset \mathrm{SYM}^+$)**.** $\mathsf{TC}^0_h \subset \mathrm{SYM}^+[\mathtt{deg}\colon (c\log n)^{2^h}, \mathtt{norm}\colon \exp((c\log n)^{2^h})]$.

Williams's program brings out Theorem 9.12, too. Raz's elusive function approach [Raz10, SY10] can supply a natural lower bound of small algebraic circuits. It can learn a small sum of multi-linearized bilinear forms from a limited amount of data, so a succinct algebraic circuit as well since Raz's multi-linearization can transform the latter to the former [Raz13]. Theorem 1.13 is a by-product of Kabanets-Impagliazzo's derandomization [KI04] in Review12.

**Lemma 1.32** (learning elusive bilinear functions)**.** Any sum of $s$ ($\ll \sqrt{n}$) set-multilinearized bilinear forms over $\mathbb{F}$ is exactly learnable from $O(s^2 n)$ data and $O(s^2 n \log|\mathbb{F}|)$ guess bits.

**Organization:** As in the title, this paper splits into three parts, learning DNF until Section 7.2, inverting Fourier transforms in Sections 7.3–8, and proving natural lower bounds in Section 9. Technically speaking, combinatorial optimization analysis (for upper and lower bounds) ends in Section 6, statistical correlation analysis in Sections 7–8, and purely number-theoretic and algebraic analyses in Section 9 (Section 9 has nothing to do with the smoothed analysis in the other sections). The reader can go immediately to Section 7.3 if interested in LWE inversion and to Section 9 for circuit lower bounds to separate $\mathsf{quasi\text{-}NP} \not\subset \mathsf{TC}^0$ and $\mathsf{VP} \neq \mathsf{VNP}$.

# 2 Preliminaries

This paper measures the computational complexities by the problem's magnitudes $n, d, k, p, q$ $s, t, 1/\varepsilon, 1/\delta$ under dominance $k + t + \log \frac{ds}{\varepsilon\delta} \ll \log n$, $\frac{s}{\varepsilon\delta} \leq d^{O(1)}$ and $1 \ll p, q \leq n^{O(1)}$. Our upper bound proof will exhibit only sketchy algorithms that any standard assembler language compatible with the Turing machine can compile, say the RAM program [AHU74]. See any computational complexity textbook for details, say [AB09, O'D14, Wig19].

**Numbers:** As usual, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{F}$ are the non-negative integers (i.e., natural numbers), the integer ring, the rational-number field, the real-number field, and any (finite or infinite) field, respectively. Write the ceil $\lfloor a \rfloor = \max\{i \in \mathbb{Z} \mid i \leq a\}$ and floor $\lceil a \rceil = \min\{i \in \mathbb{Z} \mid i \geq a\}$ of $a \in \mathbb{R}$. Let $\mathbb{Z}_q := \{\lceil (1-q)/2 \rceil, \ldots, 0, \ldots, \lceil (q-1)/2 \rceil\}$ be the integer ring modulo $q$ represented by the $q$ integers nearest to zero. Let $a \bmod q := b \in \mathbb{Z}_q$ with $a - b \in q\mathbb{Z}$. For $a, b \in \mathbb{Z}$, define $a = b \bmod m \Leftrightarrow a - b \in m\mathbb{Z}$, and $a \oplus b = (a + b \bmod 2) \in \{0, 1\}$.

**Sets:** Define $[n) := \{0, 1, \ldots, n-1\}$, $(n] := \{1, 2, \ldots, n\}$, and $[n] := \{0, 1, 2, \ldots, n\}$. In more general, for integers $m < n$, $[m, n) := \{m, m+1, \cdots n-1\}$, $[m, n) := \{m, m+1, \cdots, n-1\}$, and $[m, n] := \{m, m+1, \cdots, n\}$. We sometimes abbreviate $\{a\}$ as $a$. For sets $\mathcal{S}$ and $\mathcal{T}$, write their disjoint union by $\mathcal{S} \sqcup \mathcal{T}$, a difference $\mathcal{S} \backslash \mathcal{T} = \{a \in \mathcal{S} \mid a \notin \mathcal{T}\}$, the complement $\mathcal{S}^c = \mathcal{U} \backslash \mathcal{S}$ for the (predetermined) universal set $\mathcal{U} \supset \mathcal{S}$, a power $2^{\mathcal{S}} = \{\mathcal{T} : \mathcal{T} \subset \mathcal{S}\} \cong \{0, 1\}^{\mathcal{S}} \cong \{\varphi : \mathcal{S} \to \{0, 1\}\}$, a *functional* $\mathcal{T}^{\mathcal{S}} \cong \{\varphi : \mathcal{S} \to \mathcal{T}\}$, a combination $\binom{\mathcal{S}}{k} = \{\mathcal{T} \subset \mathcal{S} : |\mathcal{T}| = k\}$, and cartesian products $\mathcal{S} \times \mathcal{T} = \{(a, b) : a \in S, b \in \mathcal{T}\}$, $\mathcal{S}^n = \prod_{i=1}^n \mathcal{S} = \{(a_1, \ldots, a_n) \mid a_i \in \mathcal{S}\}$ ($\mathcal{S}^0 = \{\mathtt{null}\}$), and $\mathcal{S}^* = \bigsqcup_{n=0}^{\infty} \mathcal{S}^n$. We call $v \in \mathcal{S}^n$ an $\mathcal{S}$-vector (or sequence) of length $n$. Specific vectors are $a^n := (a, \ldots, a)$ and $\mathbf{1}_i := (0, \ldots, 0, 1, 0, \ldots, 0)$ of 1 at the $i$th component.

We write $v \subset w$ $v$ when a permuted $v$ occurs in $w$ as $\exists i_1, \ldots, i_{|v|}, \forall j, v_j = w_{i_j}$. A binary vector may represent the binary number $\{0,1\}^n \ni v = \sum_{i=1}^n v_i 2^{i-1}$. The binomial coefficient $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ is identical with a combination $\binom{\mathcal{S}}{k} = \{v \subset \mathcal{S} \mid |v| = k\}$ of some order-$n$ set $\mathcal{S}$. The binomial sum up to $k < n/2$ is close to the largest term $\binom{n}{k} \leq \sum_{\kappa=0}^k \binom{n}{\kappa} \leq \binom{n}{k}\frac{n-k}{n-2k}$.

**Functions:** As usual, $\log a$ and $\ln a$ are the logarithms of $a > 0$ of base 2 and $\boldsymbol{e} = 2.718\cdots$ (the natural logarithm). Denote the range $\mathrm{rng}(f) := f(\mathcal{X}) = \{f(x) \mid x \in \mathcal{X}\}$ and the domain (support) $\mathrm{dom}(f) := \mathrm{supp}(f) = f^{-1}(\mathcal{Y}) = \{f(y) \mid y \in \mathcal{Y}\}$. For $f_i : \mathcal{X}_i \to \mathcal{Y}$, $f = (f_i)_{i=1}^d$, $x \in \mathcal{X} = \prod_{i=1}^d \mathcal{X}_i$ and $w \subset (d]$, write $\mathcal{X}_w = \prod_{i\in w} \mathcal{X}_i$, $x_w := (x_i)_{i\in w}$, $f(x) = (f(x_i))_{i=1}^d$ and $f(x_w) = f_w(x) = (f(x_i))_{i\in w}$, say $\lfloor x/2 \rfloor_w = \lfloor x_w/2 \rfloor = \lfloor x_i/2 \rfloor_{i\in w}$ for $x \in [2n)^d$.

**Logics:** Propositional calculus of Boolean predicates writes the truth values $0 :=$ FALSE, $1 :=$ TRUE, the implication $\phi \to \psi := \neg\phi \vee \psi$, and equivalence $\phi \equiv \psi := (\phi \to \psi) \wedge (\psi \to \phi)$. Write $\{a \mid \phi(a)\}$ and $f(a|\phi(a))$ for the subset and subfunction induced by the condition that $\phi(a)$ is TRUE. The indicator function $1[\phi] \in \{0,1\}$ takes one if $\phi$ is TRUE.

**Graphs:** A graph is a pair $(\mathcal{V}, \mathcal{E})$ of a variable set $\mathcal{V}$ and an edge set $\mathcal{E} \subset \binom{\mathcal{V}}{2}$. Subsets $\mathcal{V}' \subset \mathcal{V}$ and $\mathcal{E}' \subset \mathcal{E}$ induce the subgraphs $(\mathcal{V}', \mathcal{E}[\mathcal{V}'])$ and $(\mathcal{V}[\mathcal{E}'], \mathcal{E}')$ of $\mathcal{E}[\mathcal{V}'] = \{e \in \mathcal{E} \mid e \cap \mathcal{V}' \neq \emptyset\})$ and $\mathcal{V}[\mathcal{E}'] = \{v \in \mathcal{V} \mid v \in \exists e \in \mathcal{E}'\}$. It is bipartite if the vertex set divides into two non-empty parts $\mathcal{V} = \mathcal{I} \sqcup \mathcal{J}$ between which the edges span, i.e., $\mathcal{E} \subset \mathcal{I} \times \mathcal{J}$.

**Algebras:** $\mathbb{S}_n = \mathbb{S}(\mathcal{S})$ is the permutation group over a set $\mathcal{S}$ of cardinality $n$, say $\mathcal{S} = [n)$. $\mathbb{F}_q$ is the finite Galois field of order $q$, identical with $\mathbb{F}_q \cong \mathbb{Z}_q$ as rings. $\mathbb{F}_q^* = \mathbb{F}_q \backslash \{0\}$ and $\mathbb{Z}_q^* = \{a \in \mathbb{Z}_q \mid a \text{ is coprime with } q\}$ are the groups of invertible elements. The $n$-variate polynomial ring over $\mathbb{F}$ is $\mathbb{F}[\mathbf{x}_1, \ldots, \mathbf{x}_n] = \left\{\sum_{w\in\mathbb{N}^n} a_w \prod_{i\in w} \mathbf{x}_i^{w_i} \mid a_w \in \mathbb{F}\right\}$ having $\mathbb{F}$-linear summation and multiplication. The multilinear one is a quotient ring $\{f \in \mathbb{F}[\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}] \mid \forall i, \mathbf{x}_i^2 = \mathbf{x}_i\} \cong \{f = \sum_{w\subset[n)} a_w \prod_{i\in w} \mathbf{x}_i \mid a_w \in \mathbb{F}\}$. A polynomial $f$'s degree is $\mathbf{deg}(f) = \max\{\sum_i w_i \mid a_w \neq 0\}$, and the norm is $\mathbf{norm}(f) = \sum_w |a_w|$. It is *homogeneous* of degree-$k$ if $\sum_{i=1}^n w_i \neq k \Rightarrow a_w = 0$. *Fundamental theorem of algebra:* Any degree-$d$ single-variable polynomial over an algebraically closed field must have exactly $d$ zeros. *Fermat's little theorem:* $\forall a \in \mathbb{Z}_q, a^{|\mathbb{Z}_q^*|} = 1$. *Chinese remainder theorem:* If $q_1, \ldots, q_n$ are coprime, $\mathbb{Z}_{\prod_{i=1}^n q_i} \cong \prod_{i=1}^n \mathbb{Z}_{q_i}$ via $a \leftrightarrow (a \bmod q_i)_{i=1}^n$.

**Matrices:** An square matrix $\mathcal{M}$ is degenerate (non-singular, invertible) if it prohibits a non-trivial linear relation, i.e., $a \neq 0 \Rightarrow \sum_{i,j} a_i \mathcal{M}_{ij} \neq 0$. The $\mathcal{M}$'s rank measures the maximum size of a non-degenerate submatrix $\mathrm{rank}(\mathcal{M}) = \max\{|\mathcal{I}| = |\mathcal{J}| \mid (\mathcal{M}(i,j))_{i\in\mathcal{I}, j\in\mathcal{J}} \text{ is non-degenerate}\}$. We write the $(i,j)$-entry $\mathcal{M}_{ij} = \mathcal{M}_{i,j} = \mathcal{M}(i,j) = \mathcal{M}_i(j)$, the $i$th row $\mathcal{M}_i = (\mathcal{M}_{ij})_j$, the $j$th column $\mathcal{M}^j = \mathcal{M}(j) = (\mathcal{M}_{ij})_i$, $\mathcal{M}_\mathcal{I} = (\mathcal{M}_i)_{i\in\mathcal{I}}$, and $\mathcal{M}^\mathcal{J} = \mathcal{M}(\mathcal{J}) = (\mathcal{M}(j))_{j\in\mathcal{J}}$. We measure $\mathcal{M}_{\neq 0} = \{(i,j) \mid \mathcal{M}_{ij} \neq 0\}$, $|\mathcal{M}|_{\neq 0} = |\mathcal{M}_{\neq 0}|$, and call $\frac{|\mathcal{M}|_{\neq 0}}{nm}$ the density of an $n$ by $m$ matrix $\mathcal{M}$.

**Random variables:** A capital letter $X$ denotes a random variable of an outcome $x \in \mathcal{X}$ generated by a probability mass function $\mathsf{Pr}[X] = \mathsf{Pr}_X[X = x]$. Write $X \sim P(x)$ for $\forall x, \mathsf{Pr}[X = x] = P(x)$ and $\mathsf{Pr}[X|X']$ for $\mathsf{Pr}_{X,X'}[X = x|X' = x'] = \mathsf{Pr}[X = x, X' = x']/\mathsf{Pr}[X' = x']$. Also, $X \sim \mathcal{X}$ is the uniform random variable $X \sim \mathsf{Pr}[X] = 1/|\mathcal{X}|$. Random variables $X_i$ are independent if $\mathsf{Pr}[(X_i)_i] = \prod_i \mathsf{Pr}[X_i]$, and mutually independent if $\mathsf{Pr}[X_i, X_{i'}] = \mathsf{Pr}[X_i]\mathsf{Pr}[X_{i'}]$ for all $i \neq i'$, written as $X_i \perp X_{i'}$. Their mixture is $X = \sum_i \rho_i X_i$ by a proportion $\sum_i \rho_i = 1$ obeying to $\mathsf{Pr}[X] = \sum_i \rho_i \mathsf{Pr}[X_i]$. A random variable $X$ is explicit if $X$'s sampler runs in $\mathrm{plog}|X|$ time (or $O(\log|X|)$ space), i.e., there is a $\mathrm{plog}|X|$ time (or $O(\log|X|)$ space) computable deterministic function $X : \mathcal{Z} \to \mathcal{X}$ to have $\mathsf{Pr}_X[X] = \mathsf{Pr}_Z[X(Z)]$. An event (a random Boolean predicate) $E$ occurs (becomes TRUE) with *confidence* $1-\delta$ if $\mathsf{Pr}[E] \geq 1-\delta$, or equivalently, with *significance* $\delta$ if $\mathsf{Pr}[\neg E] \leq \delta$. We say that $E$ happens with high confidence (low significance) if $\mathbb{E}[E] \geq 1-O(\delta)$. A *union bound* guarantees $\mathsf{Pr}[E \vee E'] \leq \mathsf{Pr}[E] + \mathsf{Pr}[E']$, which we will use without mentioning. A PRG $G : \{0,1\}^n \to \{0,1\}^m$ is secure against $t$ time if no probabilistic $t$-time algorithm $\mathcal{A}$ can distinguish between $G(U_n)$ and the genuinely random $U_m \sim \{0,1\}^m$ by accuracy $\mathsf{Pr}[\mathcal{A}(G(U_n)) \neq \mathcal{A}(U_m)] \geq \Omega(1)$.

**Measures:** Denote by $c, c', \tilde{c}, \ldots$ positive constants. Let $\epsilon$ be a positive constant sufficiently close to zero, while $(\varepsilon, \delta) = (\varepsilon_n, \delta_n)$ are positive variables diminishing to zero. For $a, b \in \mathbb{R}$, write $a \approx b \Leftrightarrow |a-b| < \epsilon$, and $a \ll b \Leftrightarrow a/b \leq \epsilon$ by $\epsilon$ taken in context. $|\mathcal{S}|$, $|X|$ and $|v|$ are polymorphic notions denoting the number of elements in a set $\mathcal{S}$, the support size $|\{x \mid \mathsf{Pr}[X = x] > 0\}|$ of a random variable $X$, and the length of a sequence (dimension of a vector) $v$, respectively. The real vector's $\ell_k$-norm is $\|v\|_k := (\sum_i |v_i|^k)^{1/k}$. The *statistical distance* between two random variables $X$ and $X'$ over the support $\mathcal{X}$ is $d_{\mathtt{st}}(X, X') = \frac{1}{2}\sum_{x\in\mathcal{X}}\big|\mathsf{Pr}[X = x] - \mathsf{Pr}[X' = x]\big|$. It is equal to the *minimum coupling distance* $d_{\mathtt{st}}(X, X') = \min_{(\tilde{X},\tilde{X}')\sim\mathsf{Pr}[X]\times\mathsf{Pr}[X']} \mathsf{Pr}[\tilde{X} \neq \tilde{X}']$, so $|\mathbb{E}[g(X)] - \mathbb{E}[g(X')]| \leq d_{\mathtt{st}}(X, X') \max|g(X)|$ for any function $g : \mathcal{X} \to \mathbb{R}$.

**Asymptotics:** For non-decreasing sequences $a_n, b_n : \mathbb{N} \to \mathbb{R}$ starting from $a_0 = b_0 = 1$, we write $a_n = \Theta(b_n) \Leftrightarrow 0 < \lim_{n\to\infty} a_n/b_n < \infty$, $a_n = O(b_n) \Leftrightarrow \lim_{n\to\infty} a_n/b_n < \infty$, $a_n = \Omega(b_n) \Leftrightarrow 0 < \lim_{n\to\infty} a_n/b_n$, $a_n = o(b_n) \Leftrightarrow \lim_{n\to\infty} a_n/b_n = 0$, and $a_n = \omega(b_n) \Leftrightarrow \lim_{n\to\infty} a_n/b_n = \infty$. Let $\mathrm{poly}(a_n) := \{b_n \mid \exists c > 1, \lim_{n\to\infty} b_n/a_n^c \to 1\}$, $\mathrm{plog}(a_n) := \{b_n \mid \exists c > 1, \lim_{n\to\infty} b_n/\log^c(a_n) \to 1\}$, $\mathrm{qpoly}(a_n) := \{b_n \mid \exists c > 1, b_n/2^{\log^c(a_n)} \to 1\}$, and $\exp(a_n) = \{b_n \mid \exists c > 1, \lim_{n\to\infty} b_n/c^{a_n} = 1\}$. Denote $\tilde{O}(a_n) = O\big(a_n\,\mathrm{plog}(a_n)\big)$. *Polynomial* growth means $\mathrm{poly}(n)$, *quasi-polynomial* $\mathrm{qpoly}(n)$, *exponential* $\exp(n)$, *sub-exponential* $\exp(n^\epsilon)$, *quasi-linear* $\tilde{O}(n)$, *linear* $\Theta(n)$, *sub-linear* $\Theta(n^\epsilon)$, *poly-logarithmic* $\mathrm{plog}(n)$, *logarithmic* $\Theta(\log n)$, and *constant* $O(1)$. For any sufficiently large scale $n$, $O(1) \ll \Theta(\log n) \ll \mathrm{plog}(n) \ll \Theta(n^\epsilon) \ll \Theta(n) \ll \mathrm{poly}(n) \ll \mathrm{qpoly}(n) \ll \exp(n^\epsilon) \ll \exp(n)$.

**Computational Complexity:** The complexity of a computational problem is the necessary and sufficient amount of resource for the modern computer to solve it. The time and space are the numbers of steps and memory size. It supposes an ideal mathematical machine, called *deterministic Turing machine*, whose mechanics the modern computer has inherited. It has an ultimate performance solving any constant-size problem in a moment to measure the asymptotic behavior of the problem scaled by $n$. $\mathsf{P}$ is the class of polynomial-time solvable problems (languages in $\{0,1\}^*$ or functions from $\{0,1\}^*$ to $\{0,1\}^*$). A computational problem is *tractable* or *efficiently* solvable if it belongs to $\mathsf{P} = \mathsf{DTIME}[\mathrm{poly}(n)]$, i.e., a computer can solve a given $n$-bit instance within $\mathrm{poly}(n)$ time. $\mathsf{NP}$ is the class of efficiently verifiable problems, i.e., a computer can verify a given proof of a given instance in polynomial time. For example, CSP $\in \mathsf{NP}$ asserts that a polynomial-time algorithm can ascertain whether or not a presented proof (assignment) satisfies a given instance (constraints). The class $\mathsf{quasi\text{-}NP}$ is the same as $\mathsf{NP}$ but allows $\mathrm{qpoly}(n)$ complexities for proof length and verification time. The class $\mathsf{coNP} = \{\mathcal{L} \subset \{0,1\}^* : \mathcal{L}^c \in \mathsf{NP}\}$ argues the efficient verification of the *refutation* $x \notin \mathcal{L}$. A language $\mathcal{L}$ is $\mathcal{F}$*-hard* if it can efficiently solve any $\mathcal{M} \in \mathcal{F}$ by simulation, i.e., $\forall \mathcal{M} \in \mathcal{F}, \exists f \in \mathsf{P}, \forall x,\ x \in \mathcal{M} \Leftrightarrow f(x) \in \mathcal{L}$. An $\mathcal{F}$*-complete* problem is an $\mathcal{F}$-hard problem belonging to $\mathcal{F}$. Randomized algorithms can observe the fair coin flippings, defining the complexity classes $\mathsf{DTIME}[t]$, $\mathsf{DSPACE}[s]$, $\mathsf{RTIME}[t]$, and $\mathsf{RSPACE}[s]$ of the problems solvable by deterministic/randomized algorithms within $t/s$ time/space.

**Circuit complexity:** A circuit is a *Directed Acyclic Graph* (DAG) labeling each $k$-fan-in node, called a gate, by a $k$-ary function. Every gate receives inputs from the in-coming edges and conveys the function's output to the outgoing edges. The *size* of a circuit is the number of edges. Its *depth* is the maximum path length, and the depth of a node is the maximum path length from the root to that node. $\mathsf{SIZE}[s(n)]$ and $\mathsf{DEP}[d(n)]$ are the classes of size $s$ and depth $d$ circuits has fan-in 2 Boolean gates and computing Boolean functions $\{0,1\}^n \to \{0,1\}$, respectively. $\mathsf{AC}^0$ consists of those languages admitting a non-uniform computation by a series polynomial-size, constant-depth, and unbounded fan-in circuits consisting of AND and OR gates, having the bottom nodes labeled by the $2n$ literals $\mathbf{x}_i$ and $\neg\mathbf{x}_i$. $\mathsf{AC}^0[p]$ is the same as $\mathsf{AC}^0$ but having $\mathrm{MOD}[p] = \mathrm{MOD}_p = 1[\sum_i x_i \not\equiv 0 \bmod p]$ gates for a fixed prime $p$, and $\mathsf{ACC} := \bigcup_{m\geq 2}\mathsf{AC}^0[m]$. In more general, SYM is the class of all symmetric functions, and $\mathsf{AC}^0[\mathcal{G}]$ can use any unbonded-fanin symmetric gates of types in $\mathcal{G} \subset \mathrm{SYM}$, e.g., $\mathsf{AC}^0[m] = \mathsf{AC}^0[\mathrm{MOD}[m]]$. A symmetric function is representable by a set of the adequate numbers of ones in the input bits, say the $\mathrm{parity}_n \cong [n] \cap (2\mathbb{N}+1)$. The classes $\mathsf{AC}^0_h$ and $\mathsf{quasi\text{-}AC}^0_h$ are the depth-$h$ $\mathsf{AC}^0$ of size $\mathrm{poly}(n)$ and $\mathrm{qpoly}(n)$. Similarly, $\mathsf{NC}^k$ and $\mathsf{quasi\text{-}NC}^k$ are the classes of binary fain-in Boolean circuits of (size, depth) $= \big(\mathrm{poly}(n), O(\log^k n)\big)$, $\big(2^{(\log n)^k}, O((\log n)^k)\big)$. By definition, $\mathsf{AC}^0 \subset \mathsf{AC}^0[p] \subset \mathsf{ACC} \subset \mathsf{NC}^1 \subset \mathsf{NC}^2 \subset \cdots$. An algebraic circuit over $\mathbb{F}$ is a DAG havingg unbounded $+$-gates, binary $\times$-gates[30] , and $\mathbb{F}$-coefficient edges to compute syntactic polynomials in $\mathbb{F}[\mathbf{x}_1,\ldots,\mathbf{x}_n]$ at the gates. Its $\times$-depth is the maximum number of $\times$-gates on a path. It is homogeneous/multi-linear if all gates compute homogeneous/multi-linear polynomials.

**Communication complexity:** A communication protocol is a binary $\{\mathrm{AND},\mathrm{OR}\}$-tree to compute a function $f(x,y) : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ by labeling to each leaf node $w$ either $1[(x,y) \in \mathcal{I}_w \times \mathcal{J}_w]$ or its negation for some $\mathcal{I}_w, \mathcal{J}_w \subset \{0,1\}^n$. $\mathsf{DEP}^{\mathsf{cc}}[d]$ is the class of depth-$d$ protocols. Its subclass $\mathsf{PH}^{\mathsf{cc}}_h[d] \subset \mathsf{DEP}^{\mathsf{cc}}[d]$ has those protocols of all root-to-leaf paths switching at most $(h-1)$ times between AND and OR gates, and $\mathsf{PH}^{\mathsf{cc}}_h[d] = \bigcup_{h\geq 1}\mathsf{PH}^{\mathsf{cc}}_h[d]$.

## 2.1 A Learning Model

Our learning model extends the worst-case standards with proof-theoretic refutation attached.

**Definition 2.1** (learning in smoothed analysis)**.** Learn a target class $\mathcal{F}$ by a hypothesis class $\mathcal{H}$ from $\eta$-noisy[31] data $\mathcal{D}$ under a shift $G$ in a proof system $\mathcal{Q}$ in the following manner.

*Device:* Fix efficient embeddings of the classes $\mathcal{F} \subset \mathcal{H} \subset \{0,1\}^\ell$.

*Shift:* Randomly pick a shift $G \in \mathcal{G}$.

*Sufficiently many examples:* Draw a dataset $\mathcal{D} \sim \big(P(G(x))P(y \mid G(x))\big)^m$ of size $m \gg \varepsilon^{-2}\big(\ell + \log\frac{1}{\delta}\big)$.

*Verifiable hypothesis:* Choose a hypothesis $h$ and its proof $\xi \in \mathcal{Q}$ with confidence $1 - O(\delta)$ to verify

$$(\eta + c\varepsilon)\textbf{-learning: } \exists f \in \mathcal{F}, \mathrm{err}_f(\mathcal{D}) \leq \eta \to P(y \neq h(x)) \leq \eta + c\varepsilon.$$

We say that $\mathcal{F}$ is learnable from $m$ data in $t$ learning time and $t'$ prediction time if a probabilistic algorithm receives $m$ data, runs in $t$ time, and outputs a function $h \in \mathsf{DTIME}[t']$ (or $h \in \mathsf{RTIME}[t']$). It defines the worst-case learning by $\mathrm{H}_\infty(G) = 0$, the proper one by $\mathcal{H} = \mathcal{F}$,

[30] If a degree-$k$ polynomial has $\{+,\times,\div\}$-circuits of size $s$ then it has $\{+,\times\}$-ones of size $\mathrm{poly}(s,k,n)$ [Str73, HY11].
[31] The noise must be below $\eta + c\varepsilon \leq 1/2 - \Omega(\varepsilon)$ to make the $(\eta+c\varepsilon)$-learning possible (even in the agnostic model).

the exact one by $\forall x, h(x) = f(x)$, the uniform-distribution one by $P(x) = 1/|\mathcal{X}|$, the marginally uniform-distribution one by $P(x, y) = P(y)/|\mathcal{X}|$, and the empirical one by $P(x(j)) = 1/m$. The PAC model (Review3) requires the clean ($\eta = 0$) or $\varepsilon$-noisy ($\eta = \varepsilon$) data, while the agnostic model (Review5) puts no assumption on $\eta$. The *white $\eta$-noise* injects the independent random classification error $\forall x, \forall y, P(f(x) \neq y \mid x) \leq \eta$ [AL88, Kea98, BFKV98, KS05], on which the PAC learner must achieve $P\big(y \neq h(x)\big) \leq c\varepsilon$, while the agnostic one $P(y \neq h(x)) \leq \eta + c\varepsilon$.

In unbounded proof systems, say the extended Frege, hypothesis's verification is automatic: The learner may choose a hypothesis $h$ together with its computational history $\xi \in \{0,1\}^*$ [CR79, Bus12]. Our learnability theorems will usually adopt this unrestricted proof system but sometimes bound it among SoS, LP, PC, and Res.

## 2.2 Shifts

SA2's shift $(g(x), \hat{g}(\theta))$ consists of the following permutations $g \in \mathbb{S}([2n))$ and $\hat{g} \in \mathbb{S}(\{0,1\}^n)$.

**Lemma 2.2** (symmetry $\equiv$ robustness)**.** The following four assertions are equivalent.

Shift1: *Robustness:* $\{x \mapsto \theta \circ g(x)\}_\theta = \{x \mapsto \theta \circ x\}_\theta$.

Shift2: *Symmetry:* $\forall x, \theta \circ g(x) = \hat{g}(\theta) \circ x$.

Shift3: $!\exists \phi \in \mathbb{S}_n, !\exists \psi \in \{0,1\}^n, g(x) = 2\phi(\lfloor x/2 \rfloor) + \psi(\lfloor x/2 \rfloor) \oplus x$.

Shift4: $!\exists \phi \in \mathbb{S}_n, !\exists \psi \in \{0,1\}^n, \hat{g}(\theta) = \theta(\phi) \oplus \psi$.

*Proof.* We will demonstrate the following implications. Shift1 $\Rightarrow^1$ $\lfloor x/2 \rfloor \overset{\phi}{\mapsto} \lfloor g(x)/2 \rfloor$ is awell-defined injective mapping $\Rightarrow^2$ Shift3 $\Rightarrow^3$ $\theta \circ g(x) = \theta(\phi(\lfloor x/2 \rfloor)) \oplus \psi(\lfloor x/2 \rfloor) \oplus x$ $\Rightarrow^4$ Shift4 $\Rightarrow^5$ Shift2 $\Rightarrow^6$ Shift1.

$\Rightarrow^1$: If $\phi$ is not well-defined, the robustness breaks down by $\lfloor x/2 \rfloor = \lfloor y/2 \rfloor \wedge \lfloor g(x)/2 \rfloor \neq \lfloor g(y)/2 \rfloor \wedge \theta \circ g(x) \neq \theta \circ g(y) \Rightarrow \big(x \mapsto \theta \circ g(x)\big) \in \{x \mapsto \theta \circ g(x)\}_\theta \setminus \{x \mapsto \theta \circ x\}_\theta$. Also, if $\phi$ is not injective, then $\lfloor x/2 \rfloor \neq \lfloor y/2 \rfloor \wedge \lfloor g(x)/2 \rfloor = \lfloor g(y)/2 \rfloor \wedge \theta \circ x \neq \theta \circ y \Rightarrow \big(x \mapsto \theta \circ g(x)\big) \in \{x \mapsto \theta \circ x\}_\theta \setminus \{x \mapsto \theta \circ g(x)\}_\theta$.

$\Rightarrow^2$: The permutation $\phi$ over $[n)$ induces $x \mapsto (g(\lfloor x/2 \rfloor), g(x) \bmod 2) := (\phi(\lfloor x/2 \rfloor), \psi(\lfloor x/2 \rfloor))$.

$\Rightarrow^3$: Review6 has defined $\circ$ as $\theta \circ \big(2\phi(\lfloor x/2 \rfloor) + \psi(\lfloor x/2 \rfloor) \oplus x\big) = \theta(\phi(\lfloor x/2 \rfloor)) \oplus \psi(\lfloor x/2 \rfloor) \oplus x$.

$\Rightarrow^4$: Suppose SA3's distribution $P_\theta(g(x)) = P_{\hat{g}(\theta)}(x)$ is an injection $g(x) \neq g(x') \Rightarrow P(g(x)) \neq P(g(x'))$. It forces $\forall x, \hat{g}(\theta) \circ x = \theta \circ g(x) = \theta(\phi(\lfloor x/2 \rfloor)) \oplus \psi(\lfloor x/2 \rfloor) \oplus x$, i.e., $\hat{g}(\theta) = \theta(\phi) \oplus \psi$.

$\Rightarrow^5$: Shift4 and Shift3 assert $\hat{g}(\theta) \circ x = \theta(\phi(\lfloor x/2 \rfloor)) \oplus \psi(\lfloor x/2 \rfloor) \oplus x = \theta \circ g(x)$.

$\Rightarrow^6$: $\{x \mapsto \theta \circ g(x)\}_\theta = \{x \mapsto \hat{g}(\theta) \circ x\}_\theta = \{x \mapsto \theta \circ x\}_\theta$ since $\hat{g}$ is a permutation over $\mathcal{T}$. □

## 2.3 Concentration Bounds

A random variable $X \in \mathbb{R}$ can derive sharper concentrations around the average $\mu = \mathbb{E}[X]$ from higher moment analyses (see any textbook of the probabilistic method, say [AS98]).

**Lemma 2.3** (momental concentration bounds)**.** For any random variable $X$ and any $0 < \gamma \leq 1$,

$(a, b)$-*slice*, $(\min, \max)$-*bound*: $a \leq \mathbb{E}[X \mid a \leq X \leq b] \leq b$. In particular, $\min X \leq \mathbb{E}[X] \leq \max X$.

$$\textit{Markov's inequality:}\ \mathsf{Pr}[X \geq 0] = 1 \Rightarrow \mathsf{Pr}[X/\mathbb{E}[X] \geq 1/\gamma] \leq \gamma.$$

$$\textit{Chebyshev's inequality:}\ \mathsf{Pr}\Big[|X - \mathbb{E}[X]| \geq \sqrt{\mathbb{E}[(X-\mathbb{E}[X])^2]/\gamma}\Big] \leq \gamma.$$

For the i.i.d. data analysis, Chernoff-Hoeffding Bounds [Che52, Hoe63] guarantees an exponentially fast convergence to the hitting rate.

**Lemma 2.4** (i.i.d. data's concentration)**.** For a sum $X = \sum_i X_i$ and the average $\mu(X) = \sum_i \mathbb{E}[X_i]$ of i.i.d. variables $X_i$ within range $X_i \in \{0,1\}$ for CB and $a \leq X_i \leq b$ for HB,

$$\textit{Chernoff Bound (CB):}\ \mathsf{Pr}[X/\mu(X) \geq 1+\gamma] < \boldsymbol{e}^{-\frac{\gamma^2}{2+\gamma}\mu(X)} \text{ for all } \gamma \geq 0.$$

$$\textit{Chernoff Bound below average:}\ \mathsf{Pr}[X/\mu(X) \leq 1-\gamma] < \boldsymbol{e}^{-\frac{\gamma^2}{2}\mu(X)} \text{ for all } 0 \leq \gamma \leq 1.$$

$$\textit{Hoeffding Bound (HB):}\ \mathsf{Pr}[|X/\mu(X) - 1| \geq \gamma] < 2\boldsymbol{e}^{-2\gamma^2\frac{\mu^2(X)}{(b-a)^2 n}} \text{ for all } 0 \leq \gamma \leq 1.$$

We apply LLL to measure the probability of "dependent" events happening simultaneously.

**Lemma 2.5** (Lovás's Local Lemma [EL73])**.** For probabilistic events $E_i$ and $0 \leq \gamma_i < 1$,

$$\textit{LLL:}\quad \forall i, \mathsf{Pr}[\neg E_i] \leq \gamma_i \textstyle\prod_{E_{i'} \not\perp E_i}(1-\gamma_{i'}) \Rightarrow \mathsf{Pr}[\bigwedge_{i=1}^n E_i] \geq \prod_{i=1}^n (1-\gamma_i).$$

## 2.4 $k$-wise independence

When pseudorandom $n$ bits look random at every local $k$-bits, they are *$k$-wisely independent.*

**Definition 2.6** (local independence)**.** Let $(w,x) \in \binom{n}{k} \times \{0,1\}^w$. A random bit-sequence $X$ is:

$$\textit{Perfectly } k\textit{-independent:}\ \forall w, \forall x, \mathsf{Pr}[\forall i \in w, X_i = x_i] = 2^{-k}.$$

$$\varepsilon\textit{-away } k\textit{-independent:}\ \forall w, \textstyle\sum_x \big|\mathsf{Pr}[\forall i \in w, X_i = x_i] - 2^{-k}\big| < \varepsilon.$$

$$\varepsilon\textit{-biased } k\textit{-independent:}\ (\forall v, 0 < |v| \leq k), \big|\mathbb{E}\big[\textstyle\prod_{i\in v}(-1)^{X_i}\big]\big| < \varepsilon.$$

$$\varepsilon\textit{-approximate } k\textit{-independent:}\ \forall w, \forall x, \big|\mathsf{Pr}[\forall i \in w, X_i = x_i] - 2^{-k}\big| < \varepsilon.$$

$$k\textit{-universal:}\ \forall w, \forall x, \mathsf{Pr}[\forall i \in w, X_i = x_i] > 0.$$

Their relative strength (with [references]) are as follows: Perfectly $k$-independent [ABI86, Lub86, CG89] $\Rightarrow$ $\varepsilon$-away $k$-independent [NN93] $\Rightarrow$ $\varepsilon$-biased $k$-independent [CGH$^+$85, Vaz86] $\Rightarrow$ $\varepsilon$-approximate $k$-independent [NN93] $\Rightarrow$ $k$-universal [KS73, CKMZ83, Alo86, ABN$^+$92]. A converse holds from the $\varepsilon$-bias to $\varepsilon$-away independence [Vaz86].

**Lemma 2.7** (from bias to away)**.** If a random bit sequence is $\varepsilon$-biased $k$-independent, then it is $\varepsilon\sqrt{2^k - 1}$-away $k$-independent.

This paper considers several variations of $k$-independence over a finite alphabet space $\mathcal{S}$.

**Definition 2.8** (local independence)**.** Let $(w,x) \in \binom{n}{k} \times \mathcal{X}_w$. A random vector $X \in \prod_{i=1}^n \mathcal{X}_i$ is:

$$k\textit{-wisely } \rho\textit{-dense:}\ \forall w, \forall x, \mathsf{Pr}[\forall i \in w, X_i = x_i] \leq 1/(|\mathcal{X}_w|\rho).$$

$$k\textit{-wisely } (\mu,\alpha)\textit{-sparse:}\ \forall w, \forall x, \mathsf{Pr}[X_w = x_w] > \alpha\mu^k.$$

*k-wisely* $(\mu,\alpha)$*-cover:* $\forall x\in\mathcal{X}_{(k)}, \mathsf{Pr}[\exists w, X_w\subset x] > \alpha\mu^k$.

*ε-away k-independent:* $\forall w, \sum_x \big|\mathsf{Pr}[\forall i\in w, X_i=x_i] - 1/|\mathcal{X}_w|\big| < \varepsilon$.

*$(h_w,\delta)$-hashed ε-away k-independent:* For a functional hash $\{h_w : \mathcal{S}^{w^c}\to\mathbb{N}\}_w$ of $w^c=(n)-w$,

$$\forall w,\forall\xi, \big(\mathsf{Pr}[h_w(X_{w^c})=\xi] > \delta \Rightarrow \textstyle\sum_x \big|\mathsf{Pr}[\forall i\in w, X_i=x_i \mid h_w(X_{w^c})=\xi] - 1/|\mathcal{X}_w|\big| < \varepsilon\big).$$

The probabilistic methods [Erd59, Erd61] can provide small $k$-wisely independent probability spaces of almost matching size to the counting argument's lower bounds.

**Lemma 2.9** ($k$-wisely universal and 1/2-dense, probabilistic)**.** There is a $k$-wisely universal and 1/2-dense random $n$ bit sequence $X$ of cardinality $|X| = O(k2^k\log n)$.

*Proof.* The random $m$ i.i.d. sampling $X(j)\sim\{0,1\}^n$ provides a desired one by a non-zero probability of chance. Lemma 2.4's Chernoff bound parameter $\gamma=1$ guarantees the data size $m = 3\cdot 2^k\big(\ln(\binom{n}{k}2^k)+O(1)\big)$ to gain a probabilistic existence

$$\text{\textit{Probabilistic method}}: \mathsf{Pr}\Big[\exists w\in\binom{n}{k}, \exists x\in\{0,1\}^k, \neg(0<\mathsf{Pr}[\forall i\in w, X_i(J)=x_i] < 2/2^k)\Big] < \binom{n}{k}2^k\big((1-2^{-k})^m + e^{-\frac{1}{3}\frac{m}{2^k}}\big) \ll 1.$$ □

**Lemma 2.10** (biased $k$-independence, probabilistic)**.** There is an $\varepsilon$-biased $k$-independent random $n$-bit sequence of cardinality $O((k/\varepsilon^2)\log n)$.

*Proof.* When $k<n/2$, Lemma 2.9's probabilistic method on $m=\frac{6}{\varepsilon^2}\big(\ln(\binom{n}{k}\frac{n-k}{n-2k})+O(1)\big)$ samples (due to $\sum_{\ell=1}^k\binom{n}{\ell}\le\binom{n}{k}\frac{n-k}{n-2k}$) and CB parameter $\gamma=\varepsilon$ demonstrates

$$\mathsf{Pr}\Big[1\le\exists\ell\le k, \exists w\in\binom{n}{\ell}, \big|\mathbb{E}[\textstyle\prod_{i\in w}(-1)^{X_i(J)}]\big|\ge\varepsilon\Big] < \textstyle\sum_{\ell=1}^k\binom{n}{\ell}\big(e^{-\frac{\gamma^2}{2+\gamma}\frac{m}{2}} + e^{-\frac{\gamma^2}{2}\frac{m}{2}}\big) \ll 1.$$

When $k\ge n/2$, take $m=\frac{6}{\varepsilon^2}\big(\ln(2^n)+O(1)\big)$ and apply $\sum_{\ell=1}^k\binom{n}{\ell}\le 2^n$ instead of $\le\binom{n}{k}\frac{n-k}{n-2k}$. □

**Lemma 2.11** (hashed $k$-independence, probabilistic)**.** There is an $(h_w,\delta)$-hashed $\varepsilon$-away $k$-independent random $n$-bit sequence of cardinality $\frac{6\cdot 2^k}{\varepsilon^2\delta}\ln(\max_w|h_w(\mathcal{X}_{w^c})|)+O(k\log n)$.

*Proof.* Lemma 2.10's probabilistic method on $m=\frac{6}{\varepsilon'^2\delta}\big(\ln(\binom{n}{k}^2\frac{n-k}{n-2k}\max_w|h_w(\mathcal{X}_{w^c})|)+O(1)\big)$ samples and CB of $\gamma=\varepsilon'$ produces a hashed $\varepsilon'$-biased $k$-independent sequence. Lemma 2.7 of bias $\varepsilon'=\varepsilon/\sqrt{2^k-1}$ transforms it to the claimed $\varepsilon$-away one:

$$\text{\textit{Probabilistic method}}: \mathsf{Pr}\left[\begin{matrix}\big(\exists w\in\binom{n}{k}, \exists\xi\in h_w(\mathcal{X}_{w^c}), \mathsf{Pr}[h_w(X_{w^c})=\xi]\ge\delta\big), (\emptyset\ne\exists v\subset w),\\ \big|\mathbb{E}[\prod_{i\in v}(-1)^{X_i}\mid h_w(X_{w^c})=\xi]\big|\ge\varepsilon'\end{matrix}\right] < \max_w|h_w(\mathcal{X}_{w^c})|\cdot\binom{n}{k}\textstyle\sum_{\ell=1}^k\binom{n}{\ell}\big(e^{-\frac{\gamma^2}{2+\gamma}\frac{\delta m}{2}}+e^{-\frac{\gamma^2}{2}\frac{\delta m}{2}}\big)\ll 1.$$ □

**Theorem 2.12** (limited independence [SSS95])**.** For a sum $X=\sum_i X_i$ of the real numbers $0\le X_i\le 1$ of an $\varepsilon$-away $k$-wise independent random vector $X$ with the average $\mu(X)=\mathbb{E}[X]$,

$$\text{\textit{Limited independence}}: \mathbb{P}\Big[\frac{|X-\mu(X)|}{\mu(X)}\ge\gamma+\varepsilon\Big] < e^{-\lfloor k/2\rfloor} \text{ for any } 0\le\gamma\le 1 \text{ and any } k\le\gamma^2e^{-1/3}\mu(X),$$

$$\text{\textit{Limited independence of short tail}}: \mathbb{P}\Big[\frac{|X-\mu(X)|}{\mu(X)}\ge\gamma+\varepsilon\Big] < e^{-\lfloor k/2\rfloor} \text{ for any } \gamma\ge 1 \text{ and any } k\le\gamma e^{-1/3}\mu(X).$$

*Proof.* Schmidt, Siegel and Srinivasan [SSS95] proved them for $\varepsilon=0$ on the $k$th moment inequality $\mathsf{Pr}[|\tilde{X}-\mu(\tilde{X})|\ge 1/\gamma]\le\gamma^k\mathbb{E}[|\tilde{X}-\mu(\tilde{X})|^k]$ of $0<\gamma\le 1$ for the sum $\tilde{X}=\sum_i\tilde{X}_i$ of perfectly $k$-independent $\tilde{X}_i$. The claimed inequalities generalize them to an $\varepsilon$-away $k$-wise independent $X$ on the differential bound $|\mathbb{E}[|X-\mu(X)|^k]-\mathbb{E}[|\tilde{X}-\mu(\tilde{X})|^k]|\le\varepsilon$. □

### 2.5 Explicit $k$-independence

Perfect $k$-independence has an explicit construction of cardinality $O(n^{k/2})$ [ABI86, CG89], and a matching lower bound $\Omega(n^{\lfloor k/2 \rfloor})$ of any $k$-independent one[32] [CGH+85, AGM03]. Weaker $k$-independence enjoys polynomial-size explicit constructions for $k = \log n$ based on graph expanders [LPS86, NN93] and three different algebraic structures [AGHP92]. One of [AGHP92] is the inner product $X_i = \langle A^i, B \rangle := \sum_{\nu=0}^{e-1} A^i_\nu B_\nu \bmod 2$ of the uniform random $A, B \sim \mathbb{F}_{2^e}$. The fundamental theorem of algebra assures that the vector $(X_i)_{i=0}^{n-1}$ is $(n/2^e)$-biased $n$-independent.

**Theorem 2.13** (weak $k$-independence, explicit [AGHP92, NN93])**.** There are explicit constructions of $\varepsilon$-approximate $k$-independent $n$-bits of cardinality $(\frac{k \log n}{2\varepsilon})^2$, $\varepsilon$-away $k$-independent ones of cardinality $2^k(\frac{k \log n}{2\varepsilon})^2$, and $\varepsilon$-biased $n$-independent ones of cardinality $(\frac{n}{\varepsilon \log(n/\varepsilon)})^2$. They are computable in quasi-linear time of the logarithm of their cardinalities.

Circuit lower bounds in Theorems 1.8–1.10 must employ an "explicit" shift for their smoothed analysis. This section will provide it, beginning from its building block, a small construction of a random *splitter*: A $t$-coloring $\Psi \in [t]^{nt}$ splits the $nt$ nodes if $\forall \ell \in [t]$, $|\Psi^{-1}(\ell)| = n$.

**Lemma 2.14** ($k$-independent $t$-splitter, explicit)**.** For $k, t \geq 2$ with $\log t \in \mathbb{N}$, let $\varepsilon = n^{-1/3}$ and $\varepsilon_{\mathtt{spl}} = (2kt + 2 + \varepsilon)\varepsilon$. There is an explicit construction of $\varepsilon_{\mathtt{spl}}$-away $k$-independent $t$-splitter $\Psi \in [t]^{nt}$ with $|\Psi| = nt^{k+1}\big(\frac{k \log t}{2\varepsilon^2} \log(nt \log t)\big)^2$.

*Proof.* Let $\Psi \in [t]^{nt} \cong \{0,1\}^{nt \log t}$ be a perfectly $k \log t$-independent bit sequence. It is a random $t$-coloring to split the $nt$ nodes into the $t$ parts of equal size $n$ in expectation $\forall \ell \in [t], \mathbb{E}[|\Psi^{-1}(\ell)|] = n$ and variance $(\sum_{x \in [nt]} \mathbb{E}[1[\Psi(x) = \ell] - 1/t])^2 = \sum_x (\mathsf{Pr}[\Psi(x) = \ell] - 1/t^2) = (n/t)(1 - 1/t)$. Although it is not precisely $t$-splitting, Chebyshev's inequality of $\gamma = \frac{\varepsilon}{t}$ gives

$$\underset{t\text{-splitting}}{\mathit{Almost}}: \quad \mathsf{Pr}\big[\forall \ell \in [t], \big||\Psi^{-1}(\ell)| - n\big| < \sqrt{n/(\gamma t) \cdot (1 - 1/t)} < \sqrt{n/\varepsilon}\big] \geq 1 - \gamma t = 1 - \varepsilon.$$

To get an exact splitter, execute $\Psi$ "sequentially" until some color gets exactly $n$ nodes, and stop there. The almost $t$-splitting may leave $t\sqrt{n/\varepsilon}$ (or less) uncolored ones, so color them appropriately to get an exact $t$-splitter $\hat{\Psi}$. If this sequential coloring $\hat{\Psi}$ starts from a randomly picked node, and runs sequentially and circularly (the next to the last node is the first one), the probabilistic distance between $\Psi(x)$ and $\hat{\Psi}(x)$ at a location $w \in \binom{nt}{k}$ is only:

$$\begin{aligned}\underset{\mathit{distance}}{\mathit{Probabilistic}}: \mathsf{Pr}[\exists x \in w, \Psi(x) \neq \hat{\Psi}(x)] &\leq \mathsf{Pr}[\neg(\text{almost } t\text{-splitting})] + \mathsf{Pr}\left[\begin{matrix}\hat{\Psi} \text{ may leave} \\ \text{some node in } w \text{ uncolored}\end{matrix}\right] \\ &\leq \varepsilon + \frac{kt\sqrt{n/\varepsilon}}{n} = (1 + kt)\varepsilon.\end{aligned}$$

We cannot explicitly construct a perfectly $k$-independent $\Psi$ within the claimed size. However, Lemma 2.13 provides an explicit $\tilde{\Psi} \in [t]^{nt}$ of size $|\tilde{\Psi}| \leq t^k\big(\frac{k \log t}{2\varepsilon^2} \log(nt \log t)\big)^2$ having statistical distance $d_{\mathtt{st}}(\Psi(x), \tilde{\Psi}(x)) \leq \varepsilon^2/2$, yielding a $t$-splitter $\hat{\tilde{\Psi}}$ of the claimed size $|\tilde{\Psi}| \cdot nt$ by factoring $nt$ to pick up the start node of the sequential coloring. Markov's inequality parameter $\gamma = \varepsilon$ applies to the expected difference $\sum_{\ell \in [t]} \mathbb{E}[\big||\Psi^{-1}(\ell)| - |\tilde{\Psi}^{-1}(\ell)|\big|] \leq \varepsilon^2 n$ and bounds

$$\begin{aligned}&\mathsf{Pr}\big[\exists \ell \in [t], \big||\tilde{\Psi}^{-1}(\ell)| - n\big| > \sqrt{n/\varepsilon} + \varepsilon^2 n/\gamma\big] \\ &\quad \leq \mathsf{Pr}\big[\exists \ell, \big||\Psi^{-1}(\ell)| - n\big| > \sqrt{n/\varepsilon}\big] + \mathsf{Pr}\big[\exists \ell, \big||\Psi^{-1}(\ell)| - |\tilde{\Psi}^{-1}(\ell)|\big| > \varepsilon^2 n/\gamma = \varepsilon n\big] \leq \varepsilon + \varepsilon.\end{aligned}$$

[32] Any random $n$-bit vector $X$ having statistical distance $d_{\mathtt{st}}(X, X') < 1/2$ from some perfectly $k$-independent $n$-bits $X'$ must have $|X| \geq n^{k/2}/(2k^k)$ [AGM03], although $\forall w \in \binom{n}{k}, d_{\mathtt{st}}(X_w, X'_w) \leq \varepsilon/2$ by definition.

The probabilistic distance analysis on $\Pr[\exists x \in w, \hat{\tilde{\Psi}}(x) \neq \tilde{\Psi}(x)]$ derives the claimed deviation

$$\underset{\text{distance}}{\text{Statistical}}: 2d_{\mathtt{st}}(\Psi_w, \hat{\tilde{\Psi}}_w) \le 2d_{\mathtt{st}}(\Psi_w, \tilde{\Psi}_w) + 2d_{\mathtt{st}}(\tilde{\Psi}_w, \hat{\tilde{\Psi}}_w) \le \varepsilon^2 + \Pr[\exists x \in w, \hat{\tilde{\Psi}}(x) \neq \tilde{\Psi}(x)]$$
$$\le \varepsilon^2 + 2\varepsilon + \frac{kt(\sqrt{n/\varepsilon} + \varepsilon n)}{n} \le (\varepsilon + 2 + 2kt)\varepsilon. \qquad \square$$

Theorems 1.8–1.10 want an explicitly defined $k$-wisely independent permutation over $[N)$. Our construction will color $[N)$ by Lemma 2.14's $k$-splitter $\Psi$ and permute the $\ell$-color nodes $\{x \in [N) \mid \Psi(x) = \ell\}$ by the bits $\langle A^{(i+j)k+\ell}, B\rangle$ of the modulo-$k$ remainder $\ell \in [k)$. We call it a DFT-shift since $(A^{(i+j)k+\ell})_{i,j}$ induces a *Discrete Fourier Transform* over $\mathbb{F}_{2^e}$.

**Definition 2.15** (DFT-shift)**.** Let $N := 2^{n+\log k}$ for even $n$ and $\log k$. Let $(\iota, \kappa) \in \{\mathtt{i}, \mathtt{o}\} \times \{\mathtt{r}, \mathtt{c}\}$.

*$k$-splitters:* Lemma 2.14 provides four i.i.d. $\varepsilon_{\mathtt{spl}}$-away $2k$-independent $k$-splitters $\Psi_{\iota,\kappa}$ of $[\sqrt{N})$.

*DFT-bits:* Let $\Phi_\ell(i,j) := \langle A^{(i+j)k+\ell}, B\rangle$ and $\Phi_\ell(z) := \big(\sum_{i \in z} \Phi_\ell(i,j)\big)_{j \in [n)} : \mathbb{Z}_2^n \to \mathbb{Z}_2^n$, $\mathbb{Z}_2^n \cong 2^{[n)}$.

*Linear order:* Over $\{x = (x_{\mathtt{r}}, x_{\mathtt{c}}) \in [\sqrt{N}) \times [\sqrt{N}) \mid \Psi_\iota(x) := (\Psi_{\iota,\mathtt{r}}(x_{\mathtt{r}}) + \Psi_{\iota,\mathtt{c}}(x_{\mathtt{c}})) \bmod k = \ell\}$, introduce a linear order $\#_\iota(x) := (\#_{\iota,\mathtt{r}}(x_{\mathtt{r}}), \#_{\iota,\mathtt{c}}(x_{\mathtt{c}}), \Psi_{\iota,\mathtt{c}}(x_{\mathtt{c}})) \in [\frac{\sqrt{N}}{k}) \times [\frac{\sqrt{N}}{k}) \times [k) \cong \mathbb{Z}_2^n$ via $\#_{\iota,\kappa}(x) := |\{x'_\kappa < x_\kappa \mid \Psi_{\iota,\kappa}(x'_\kappa) = \Psi_{\iota,\kappa}(x_\kappa)\}|$.

*DFT-shift:* Define $\Phi(x) = y \Leftrightarrow \Psi_{\mathtt{i}}(x) = \Psi_{\mathtt{o}}(y) = \ell \wedge \#_{\mathtt{o}}(y) = \Phi_\ell(\#_{\mathtt{i}}(x))$.

**Lemma 2.16** ($k$-wise independent permutation)**.** For $(z_\ell \neq z'_\ell)_{\ell=0}^{k-1} \in (\mathbb{Z}_2^n \times \mathbb{Z}_2^n)^k$,

*Permutation:* $\Pr[\text{All } \Phi_\ell \text{ are non-singular linear maps}] \ge 1 - \frac{k^2 n 2^{2n+1}}{2^e}$.

*$\varepsilon$-approximate $k$-independence:* $\Pr\big[\big|\Pr[\forall \ell, \Phi_\ell(z_\ell) = \Phi_\ell(z'_\ell)] - 2^{-kn}\big| \le \varepsilon\big] \ge 1 - \frac{2k^2 n}{\varepsilon 2^e}$.

*Proof.* Let $\varphi_\ell(\mathbf{x}|w) := \sum_{i \in w} \sum_{j \in [n)} \mathbf{x}^{(i+j)k+\ell}$ of $w \subset [e)$, so $\sum_{j \in [n)} \Phi_\ell(z)_j = \langle \varphi_\ell(A|z), B\rangle$. The fundamental theorem of algebra over $\mathbb{F}_{2^e}$ on $\mathbb{E}[(-1)^{\langle \varphi_\ell(a|z), B\rangle} \mid \varphi_\ell(a|z) \neq 0] = 0$ promises

$$\textit{DFT-bits are unbiased:} \quad \big|\mathbb{E}[(-1)^{\langle \varphi_\ell(A|w), B\rangle}]\big| \le \mathbf{deg}\big(\varphi_\ell(\mathbf{x}|w)\big)/2^e.$$

**Permutation:** Ler $z \circ z' = \sum_{(i,j) \in z \times z'} \mathbf{1}_{i+j} \bmod 2$ of $z, z' \subset [n)$. The unbiased DFT-bits can estimate the inner products of Fourier character functions[33] $\chi_\ell(z, z') := \prod_{i \in z} \prod_{j \in z'} (-1)^{\Phi_\ell(i,j)}$ over the uniform random vector $Z \subset [n)$:

$$\begin{aligned}
\mathbb{E}[\chi_\ell(z_\ell, Z) \cdot \chi_\ell(z'_\ell, Z)] &= \mathbb{E}[\textstyle\prod_{j \in Z} (-1)^{\Phi_\ell(z_\ell)_j + \Phi_\ell(z'_\ell)_j} \mid \Phi_\ell(z_\ell) = \Phi_\ell(z'_\ell)]\Pr[\Phi_\ell(z_\ell) = \Phi_\ell(z'_\ell)] \\
&\quad + \mathbb{E}[\textstyle\prod_{j \in Z} (-1)^{\Phi_\ell(z_\ell)_j + \Phi_\ell(z'_\ell)_j} \mid \Phi_\ell(z_\ell) \neq \Phi_\ell(z'_\ell)]\Pr[\Phi_\ell(z_\ell) \neq \Phi_\ell(z'_\ell)] \\
&= \Pr_A[\Phi_\ell(z_\ell) = \Phi_\ell(z'_\ell)] + 0, \\
\big|\mathbb{E}[\chi_\ell(z_\ell, Z) \cdot \chi_\ell(z'_\ell, Z)] - 2^{-n}\big| &= \big|\mathbb{E}[(-1)^{\langle \varphi_\ell(A|z_\ell \circ Z) + \varphi_\ell(A|z'_\ell \circ Z), B\rangle} \mid Z \neq \emptyset]\Pr[Z \neq \emptyset]\big| \\
&\le \mathbf{deg}(\varphi_\ell(\mathbf{x} \mid z_\ell \circ Z) + \varphi_\ell(\mathbf{x} \mid z'_\ell \circ Z))/2^e < 2kn/2^e.
\end{aligned}$$

These inner product's $z'_\ell = \emptyset$ case assures that all $\Phi_\ell$ must be non-singular. Some $\Phi_\ell$'s singularity derives a contradiction on Markov's inequality of $\gamma = 2k^2n/2^e \cdot 2^{2n}$ on $\mu(A, B, Z) := \big|\Pr[\Phi_\ell(Z) = 0^n] - 2^{-n}\big|$'s average analysis:

$$\textstyle\sum_\ell \mathbb{E}_Z\big[\mu(A, B, Z) \mid Z \neq \emptyset\big] = (2^n - 1)^{-1} \sum_\ell \sum_{z_\ell \neq \emptyset} \mu(A, B, Z) \le \frac{(2^n - 1)k}{2^n - 1} \cdot 2kn/2^e$$

[33] By definition, $\Phi_\ell(\emptyset) = 0^n$ and $\chi_\ell(z, \emptyset) = 1$.

$$\Rightarrow \mathsf{Pr}_{A,B}\big[\textstyle\sum_\ell \mathbb{E}_Z[\mu(A,B,Z) \mid Z \neq \emptyset] \leq 2k^2n/(2^e\gamma) = 2^{-2n}\big] \geq 1-\gamma$$
$$\Rightarrow \tfrac{1}{2^n-1} - \tfrac{1}{2^n} \leq \textstyle\sum_\ell \mathbb{E}_Z\big[\big|\mathsf{Pr}[\Phi_\ell(Z) = 0^n] - \tfrac{1}{2^n}\big| \mid Z \neq \emptyset\big] \leq 2^{-2n}.$$

**$k$-independence:** Similarly, the inner products of the i.i.d. $k$ tuples $(Z_\ell)_{\ell=0}^{k-1} \subset [n)^k$ yields

$$\begin{aligned}
&\mathbb{E}[\textstyle\prod_\ell \chi_\ell(z_\ell, Z_\ell)\cdot \chi_\ell(z'_\ell, Z_\ell)] \\
&\quad = \mathbb{E}[\textstyle\prod_\ell\prod_{j\in Z_\ell}(-1)^{\Phi_\ell(z_\ell)_j + \Phi_\ell(z'_\ell)_j} \mid \forall \ell, \Phi_\ell(z_\ell) = \Phi_\ell(z'_\ell)]\mathsf{Pr}[\forall\ell, \Phi_\ell(z_\ell) = \Phi_\ell(z'_\ell)] \\
&\qquad + \mathbb{E}[\textstyle\prod_\ell\prod_{j\in Z_\ell}(-1)^{\Phi_\ell(z_\ell)_j + \Phi_\ell(z'_\ell)_j} \mid \exists \ell, \Phi_\ell(z_\ell) \neq \Phi_\ell(z'_\ell)]\mathsf{Pr}[\exists\ell, \Phi_\ell(z_\ell) \neq \Phi_\ell(z'_\ell)] \\
&\quad = \mathsf{Pr}[\forall\ell, \Phi_\ell(z_\ell) = \Phi_\ell(z'_\ell)] + 0, \\
&\big|\mathbb{E}[\textstyle\prod_\ell \chi_\ell(z_\ell, Z_\ell)\cdot \chi_\ell(z'_\ell, Z_\ell)] - 2^{-kn}\big| \\
&\quad \leq \mathbb{E}\big[(-1)^{\langle \sum_{\ell\in[k)} \varphi_\ell(A|z_\ell \circ Z_\ell) + \varphi_\ell(A|z'_\ell\circ Z_\ell), B\rangle} \mid \exists \ell, Z_\ell \neq \emptyset\big]\mathsf{Pr}[\exists \ell, Z_\ell \neq \emptyset] \leq 2kn/2^e.
\end{aligned}$$

Markov's inequality parameter $\gamma = (2k^2n)/(\varepsilon 2^e)$ on this expectation bound deduces

$$\mathsf{Pr}\big[\textstyle\sum_\ell \big|\mathsf{Pr}[\forall \ell, \Phi_\ell(z_\ell) = \Phi_\ell(z'_\ell)] - 2^{-kn}\big| \leq 2k^2n/(2^e\gamma) = \varepsilon\big] \geq 1-\gamma. \qquad \square$$

**Theorem 2.17** (DFT-shift)**.** Given any $\sqrt{N}$ by $\sqrt{N}$ matrix $\mathcal{M}$ of density $1/(k(\delta N)^{\frac{1}{2k}}) \ll \mu = \frac{|\mathcal{M}|_{\neq 0}}{N}$, $\mathcal{I} \subset [\sqrt{N})$, and $w \in \binom{[\sqrt{N})}{k}$. Let $\mu_{\mathtt{dns}} \approx \mu|\mathcal{I}|$ and $\mu_{\mathtt{cvr}} \approx k!\binom{|\mathcal{M}|_{\neq 0}}{k}/N^k$. If $k^2n2^{2kn} \ll 2^e\delta$, Definition 2.15's DFT-shift $\Phi$ permutes $\mathcal{M}$ as $\mathcal{M}\circ\Phi(i,j) := \mathcal{M}(\Phi(i,j))$ on the random $J \sim [\sqrt{N}) \cong \{0,1\}^{(n+\log k)/2}$ with high confidence in the following manner.

*Inversion:* $\Phi^{-1}(y)$ is computable in $O(n^2)$ time, once having all $\Psi_{\iota,\kappa}(x_\kappa)$ of $(\iota,\kappa,x_\kappa) \in \{\mathtt{i},\mathtt{o}\}\times\{\mathtt{r},\mathtt{c}\}\times[\sqrt{N})$ in $\tilde{O}(\sqrt{N})$ time, and all linear mappings $\Phi_\ell$ of $\ell\in[k)$ in $\tilde{O}(ekn^2)$ time.

*Permutation:* $\Phi$ is a permutation.

*Uniform density:* $\mathbb{E}[\,\big||(\mathcal{I},J)\cap(\mathcal{M}\circ\Phi)_{\neq 0}| - \mu_{\mathtt{dns}}\big|] \ll \mu_{\mathtt{dns}}\sqrt{1+4\varepsilon_{\mathtt{spl}}}$.

*$k$-cover:* $\big|\mathsf{Pr}[(w,J)\subset(\mathcal{M}\circ\Phi)_{\neq 0}, |\Psi_{\mathtt{o},\mathtt{r}}(w)| = k] - \mu_{\mathtt{cvr}}\big| \ll \mu_{\mathtt{cvr}}\sqrt{1+3\varepsilon_{\mathtt{spl}}}$.

*Proof.* **Inversion:** The $\Psi$'s coloring induces Definition 2.15's linear order $\#_\iota(x)$ over the $\ell$-monotone nodes $\{x\in[N) \mid \Psi_\iota(x) = \ell\} \cong [2^n)$. Computing the $n\times n$ $\mathbb{F}_2$-matrices $\Phi_\ell$ and inverting them for all $\ell\in[k)$ takes only $\tilde{O}(ekn^2)$ time to execute the $\mathbb{F}_{2^e}$-powers $A^{(i+j)k+\ell}$ of all $(i,j,\ell)\in[n)\times[n)\times[k)$ [SS71, Sch77]. The DFT-shift and its inversion conduct these operations.

**Permutation:** $\Phi$ is a permutation if so are all $\Phi_\ell$, whose confidence level Lemma 2.16 guarantees.

**Uniform-density and $k$-cover:** Suppose that the four $\Psi_{\iota,\kappa}$ are perfectly $2k$-independent $k$-splitters of $[\sqrt{N}]$. Let $\varepsilon_1 := \frac{\delta}{2^n}$, $\varepsilon_2 := \frac{\varepsilon_1}{2^n}$, $\varepsilon_k := \frac{\delta}{2^{kn}}$, and $\varepsilon_{2k} := \frac{\varepsilon_k}{2^{kn}}$. Lemma 2.16 on $k^2n2^{2kn} \ll 2^e\delta$ has provided those $\varepsilon$-approx $t$-independent permutations of $(\varepsilon,t)\in\{(\varepsilon_1,1),(\varepsilon_1,2),(\varepsilon_k,k),(\varepsilon_{2k},2k)\}$ with high confidence. For $y_\lambda, x_\lambda\in[N)$, $v_\lambda\in\binom{N}{k}$, and $j_\lambda\in[\sqrt{N})$, let

$$\begin{aligned}
&E(x,y) := 1[\forall\lambda, \Phi(x_\lambda) = y_\lambda \mid \forall\lambda, \Psi_{\mathtt{i}}(x_\lambda) = \Psi_{\mathtt{o}}(y_\lambda)] \text{ for } x = (x_\lambda)_{\lambda\in\Lambda} \text{ and } y = (y_\lambda)_{\lambda\in\Lambda}, \\
&E(v,j) := 1[\forall\lambda, \Phi(v_\lambda) = (w,j_\lambda) \mid \forall\lambda, |\Psi_{\mathtt{i}}(v_\lambda)| = |\Psi_{\mathtt{o},\mathtt{r}}(w)| = k] \text{ for } v = (v_\lambda)_{\lambda\in\Lambda},\ j = (j_\lambda)_{\lambda\in\Lambda}, \\
&\overline{E}(x,y) := E(x,y) - 2^{-|\Lambda|n}, \quad \overline{E}(v,j) := E(v,j) - 2^{-|\Lambda|kn}.
\end{aligned}$$

Since $\Phi$ is a permutation, $x \neq x' \Leftrightarrow y \neq y'$ under $E(x,y) = E(x',y') = 1$. Similarly, $v \cap v' = \emptyset$ $\Leftrightarrow j \neq j'$ under $E(v,j) = E(v',j') = 1$. Let $\sigma_{\mathtt{dns}}^2 \approx 3\delta\mu_{\mathtt{dns}}^2$ and $\sigma_{\mathtt{cvr}}^2 \approx 3\delta\mu_{\mathtt{cvr}}^2$. Lemma 2.16's $k$-independent $\Phi$ calculates the first two moments of the $k$-splitting and $k$-covering claims:

*Uniform-density's average:* $\left|\mathbb{E}_J[|(\mathcal{I},J) \cap (\mathcal{M}\circ\Phi)_{\neq 0}|] - \mu_{\mathtt{dns}}\right|$

$$= \left|\left(\sum_{x\in\mathcal{M}_{\neq 0}}\sum_{y\in(\mathcal{I},J)}\mathsf{Pr}[\Phi(x)=y \mid \Psi_{\mathtt{i}}(x)=\Psi_{\mathtt{o}}(y)]\mathsf{Pr}[\Psi_{\mathtt{i}}(x)=\Psi_{\mathtt{o}}(y)]\right) - \mu_{\mathtt{dns}}\right|$$

$$= \tfrac{1}{k}\left|\sum_x\sum_y \mathbb{E}[\overline{E}(x,y)]\right| \leq \tfrac{\varepsilon_1}{k}|\mathcal{M}|_{\neq 0}|\mathcal{I}| = \delta\mu_{\mathtt{dns}}.$$

*Uniform-density's variance:* $\left|\mathbb{E}_J[|(\mathcal{I},J) \cap (\mathcal{M}\circ\Phi)_{\neq 0}|] - \mu_{\mathtt{dns}}\right|^2$

$$= \frac{1}{k^2N}\left|\sum_{x\in\mathcal{M}_{\neq 0}}\sum_{y\in\mathcal{I}\times[\sqrt{N}]}\mathbb{E}[\overline{E}(x,y)^2] + \sum_{(x,y)\neq(x',y')}\mathbb{E}[\overline{E}(x,y)\overline{E}(z',y')]\right|$$

$$= \frac{1}{k^2N}\left|\begin{array}{c}\sum_{(x,y)}\left(\mathbb{E}[E(x,y)](1-2^{-n}) - 2^{-n}\mathbb{E}[\overline{E}(x,y)]\right)\\ +\sum_{(x,y)\neq(x',y')}\left(\mathbb{E}[\overline{E}\big((x,x'),(y,y')\big)] - 2^{-n}(\mathbb{E}[\overline{E}(x,y)] + \mathbb{E}[\overline{E}(x',y')])\right)\end{array}\right|$$

$$\leq \frac{1}{k^2N}\begin{pmatrix}|\mathcal{M}|_{\neq 0}|\mathcal{I}|\sqrt{N}(\varepsilon_1 + 2^{-n})(1-2^{-n})\\ +|\mathcal{M}|_{\neq 0}^2|\mathcal{I}|^2N(\varepsilon_2 + 2^{1-n}\varepsilon_1)\end{pmatrix} := \sigma_{\mathtt{dns}}^2. \quad \left(\because \tfrac{|\mathcal{M}|_{\neq 0}|\mathcal{I}|}{k^2\sqrt{N}2^n} = \tfrac{\mu_{\mathtt{dns}}}{k\sqrt{N}} \ll \sigma_{\mathtt{dns}}^2.\right)$$

*$k$-cover's average:* $\left|\mathsf{Pr}[(w,J) \subset (\mathcal{M}\circ\Phi)_{\neq 0}, |\Psi_{\mathtt{o,r}}(w)| = k] - \mu_{\mathtt{cvr}}\right|$

$$= \left|\tfrac{k!}{k^k}\left(\sum_{v\in\binom{\mathcal{M}_{\neq 0}}{k}}\mathsf{Pr}[\Phi(v) = (w,J) \mid |\Psi_{\mathtt{i}}(v)| = |\Psi_{\mathtt{o,r}}(w)| = k] - 2^{-kn}\right)\right|$$

$$= \tfrac{k!}{k^k}\left|\sum_{v\in\binom{\mathcal{M}_{\neq 0}}{k}}\mathbb{E}[\overline{E}(v,J)]\right| \leq \tfrac{k!}{k^k}\binom{|\mathcal{M}|_{\neq 0}}{k}\cdot\varepsilon_k = \delta\mu_{\mathtt{cvr}}.$$

*$k$-cover's variance:* $\left|\mathsf{Pr}_J[\Phi^{-1}(w,J) \subset \mathcal{M}_{\neq 0} \mid |\Psi_{\mathtt{o,r}}(w)| = k] - \mu_{\mathtt{cvr}}\right|^2$

$$= \frac{1}{N}\left(\frac{k!}{k^k}\right)^2\left|\begin{array}{c}\sum_{v\in\binom{\mathcal{M}_{\neq 0}}{k}}\sum_{j\in[\sqrt{N}]}\left(\mathbb{E}[E(v,j)](1-2^{-kn}) - 2^{-kn}\mathbb{E}[\overline{E}(v,j)]\right)+\\ \sum_{v\cap v'=\emptyset, j\neq j'}\left(\mathbb{E}[\overline{E}\big((v,v'),(j,j')\big)] - 2^{-kn}(\mathbb{E}[\overline{E}(v,j)] + \mathbb{E}[\overline{E}(v',j')])\right)\end{array}\right|$$

$$\leq \frac{1}{N}\left(\frac{k!}{k^k}\right)^2\begin{pmatrix}\binom{|\mathcal{M}|_{\neq 0}}{k}\sqrt{N}(\varepsilon_k + 2^{-kn})(1-2^{-kn})\\ +\binom{|\mathcal{M}|_{\neq 0}}{k}\binom{|\mathcal{M}|_{\neq 0}-k}{k}N\big(\varepsilon_{2k} + 2^{1-kn}\varepsilon_k\big)\end{pmatrix} := \sigma_{\mathtt{cvr}}^2. \quad \left(\because \tfrac{\binom{|\mathcal{M}|_{\neq 0}}{k}(k!)^2}{\sqrt{N}N^kk^k} = \tfrac{k!\mu_{\mathtt{cvr}}}{\sqrt{N}k^k} \ll \sigma_{\mathtt{cvr}}^2.\right)$$

Chebyshev's inequality of $\gamma \ll \delta^{-1/2}$ applies to these moments and establishes the claimed concentrations. It must replace $\Psi_{\iota,\kappa}$ with Lemma 2.14's $\varepsilon_{\mathtt{spl}}$-away $2k$-independent $k$-splitters $\tilde{\Psi}_{\iota,\kappa}$ so that $\mu_\lambda$ with $\mu_\lambda(1 \pm O(\varepsilon_{\mathtt{spl}}))$ and $\sigma_\lambda$ with $\sigma_\lambda(1 + O(\varepsilon_{\mathtt{spl}}))$ for $\lambda = \mathtt{dns}, \mathtt{cvr}$ by ratios

$$\frac{\mathsf{Pr}[\tilde{\Psi}_{\mathtt{i}}(x) = \tilde{\Psi}_{\mathtt{o}}(y), \tilde{\Psi}_{\mathtt{i}}(x') = \tilde{\Psi}_{\mathtt{o}}(y')]}{\mathsf{Pr}[\Psi_{\mathtt{i}}(x) = \Psi_{\mathtt{o}}(y), \Psi_{\mathtt{i}}(x') = \Psi_{\mathtt{o}}(y')]} \leq 1 + \sum_{\kappa\in\{\mathtt{r},\mathtt{c}\}}2\begin{pmatrix}d_{\mathtt{st}}\big(\Psi_{\mathtt{i},\kappa}(x_\kappa, x'_\kappa), \tilde{\Psi}_{\mathtt{i},\kappa}(x_\kappa, x'_\kappa)\big)+\\ d_{\mathtt{st}}\big(\Psi_{\mathtt{o},\kappa}(y_\kappa, y'_\kappa), \tilde{\Psi}_{\mathtt{o},\kappa}(y_\kappa, y'_\kappa)\big)\end{pmatrix} \leq 1 + 4\varepsilon_{\mathtt{spl}},$$

$$\frac{\mathsf{Pr}[|\tilde{\Psi}_{\mathtt{i}}(v)| = |\tilde{\Psi}_{\mathtt{i}}(v')| = |\tilde{\Psi}_{\mathtt{o,r}}(w)| = k]}{\mathsf{Pr}[|\Psi_{\mathtt{i}}(v)| = |\Psi_{\mathtt{i}}(v')| = |\Psi_{\mathtt{o,r}}(w)| = k]} \leq 1 + 2\begin{pmatrix}\sum_{\kappa\in\{\mathtt{r},\mathtt{c}\}}d_{\mathtt{st}}\big(\Psi_{\mathtt{i},\kappa}(v_\kappa, v'_\kappa), \tilde{\Psi}_{\mathtt{i},\kappa}(v_\kappa, v'_\kappa)\big)\\ +d_{\mathtt{st}}\big(\Psi_{\mathtt{o,r}}(w), \tilde{\Psi}_{\mathtt{o,r}}(w)\big)\end{pmatrix} \leq 1 + 3\varepsilon_{\mathtt{spl}}.$$

□

# 3 Learning versus Refutation

The DSS reduction revealed that learning is equivalent to refuting on polynomial time computation by allowing *False Negative Error* (FNE) and possibly rejecting some satisfiable instances [DSS16, Vad17, KL18]. This section will extend it from the worst-case to smoothed analysis in the (usual) No FNE refutation [DLL62, CS88, CEI96, GK01, Fei02, App16, FPV18, BBKK18].

**Definition 3.1** (refutation in smoothed analysis)**.** A randomized algorithm $\mathcal{A}$ refutes $\mathcal{F}$ if it distinguishes between the training dataset $\mathcal{D} \sim \big(P(G(x))P(y|G(x))\big)^m$ with noise $\eta \le 1/2 - \Theta(\varepsilon)$ and the random-label $\mathcal{U} \sim (P'(x) \cdot \frac{1}{2})^m$ drawn from an arbitrary variate distribution $P'(x)$:

$$\textit{$\eta$-noisy refutation: } \mathsf{Pr}_{\mathcal{D},\mathcal{U}} \left[ \begin{array}{c} \exists f \in \mathcal{F}, \mathrm{err}_f(\mathcal{D}) \le \eta \Rightarrow \\ \mathsf{Pr}_{\mathcal{A}}[\mathcal{A}(\mathcal{U}) = \mathtt{refute}] \approx 1 \wedge \mathsf{Pr}_{\mathcal{A}}[\mathcal{A}(\mathcal{D}) = \mathtt{refute}] = 0 \end{array} \right] \ge 1 - O(\delta).$$

A reduction from refutability to learnability is immediate. The previous reductions from learning to refuting transformed any refutation algorithm on $m$ constraints into a weak learner, then boosted it to $O(\varepsilon)$-learner by spending $\tilde{O}(m^c)$ examples for $c \ge 3$. However, they lacked *Uniform Generalization Error Bounds* (UGEB) so that each new prediction might claim a new training dataset. This section will compensate for UGEB to them. We will adopt a *smooth boosting* [Imp95, DW$^+$00, Ser03, Hat06] to realize an $\tilde{O}(m^2)$-data reduction. It can endure even malicious noise since it never puts too much weight on any single example.

**Lemma 3.2** (learner to refuter)**.** Any $(\eta + c\varepsilon)$-learner with $\eta + c\varepsilon \le 1/2 - \epsilon\varepsilon$ in Definition 2.1 must be Definition 3.1's $\eta$-noisy refuter.

*Proof.* Let the given learner feed Definition 3.1's dataset $\mathcal{D}' \in \{\mathcal{D}, \mathcal{U}\}$, and choose a hypothesis $h = h(\mathcal{D}')$ to verify $(\mathrm{err}(\mathcal{D}') + c\varepsilon)$-learning with high confidence. Let the learner refute $\mathcal{D}'$ if and only if getting a proof of $\mathrm{err}(\mathcal{D}') > \eta$. Supply to the learner Definition 2.1's sufficiently many examples $m \gg \varepsilon^{-2}(\log|\mathcal{H}| + \log\frac{1}{\delta})$. Lemma 2.4's Chernoff bound of $\gamma = \frac{1/2 - (\eta + c\varepsilon)}{1/2}$ guarantees

$$\textit{UGEB:} \quad P\big(\mathrm{err}_h(\mathcal{U}) > \eta + c\varepsilon\big) \ge 1 - |\mathcal{H}|\boldsymbol{e}^{-\frac{\gamma^2}{2}\cdot\frac{1}{2}m} \ge 1 - o(\delta).$$

With high confidence, Definition 2.1's $(\eta + c\varepsilon)$-learner can get a prof of $\mathrm{err}(\mathcal{U}) > \eta$, but can never that of $\mathrm{err}(\mathcal{D}) > \eta$, realizing Definition 3.1's $\eta$-noisy refutation. $\square$

**Theorem 3.3** (smooth boosting [Ser03])**.** `SmoothBoost` repeats producing distributions $P_\nu(x, y)$ over a given dataset $\mathcal{D}$ and receiving $h_\nu \in [-1, 1]^{\mathcal{D}}$ for $\nu_0 \le \frac{2}{\varepsilon\alpha^2(1-\alpha^{1/2})}$ times. Finally, it outputs their majority vote $h = (\mathrm{sgn}(\frac{1}{\nu_0}\sum_{\nu=1}^{\nu_0} h_\nu) + 1)/2$. It weights and performs over $\mathcal{D}$ as follows.

*Counting:* $N_\nu(x, y) = N_{\nu-1}(x, y) + (-1)^y h_\nu(x) - \alpha/(2 + \alpha)$.

*Weighting:* $P_{\nu+1}(x, y) \propto 1[N_\nu(x, y) < 0] + (1 - \alpha)^{N_\nu(x,y)/2} \cdot 1[N_\nu(x, y) \ge 0]$.

*Boosting:* $\forall\nu, \mathbb{E}_{(X_\nu, Y_\nu)\sim P_\nu(x,y)}[\,|(-1)^{Y_\nu} - h_\nu(X_\nu)|/2] \le 1/2 - \alpha \Rightarrow \mathsf{Pr}_{(X,Y)\sim\mathcal{D}}[h(X) \ne Y] \le \varepsilon$.

*Smoothness:* $\forall\nu, P_\nu(x, y) \le 1/(\varepsilon|\mathcal{D}|)$.

**Theorem 3.4** (refutation to PAC learning)**.** Let $\delta_{3.4} := \frac{\varepsilon\delta}{m^4 \log^3 m \log\frac{1}{\varepsilon\delta}}$. If noise-free $\mathcal{F}$ is refutable with significance $O(\delta_{3.4})$ from $m$ data in $t$ time, $\mathcal{F}$ is PAC learnable from $m^2/\varepsilon \cdot O\big(\log\frac{m}{\varepsilon\delta}\log\frac{1}{\delta}\big)$ data in $t \cdot m^4/\varepsilon \cdot O\big(\log^3 m \log\frac{1}{\varepsilon\delta}\big)$ time, given free access to $P(x)$.

*Proof.* **Yao's reduction on binary search:** Let $\mathcal{A}$ be Definition 3.1's refutation algorithm. Suppose $m = 2^{\log m}$. Let $\alpha \approx \frac{1}{m}$. Let $(X, Y) \sim \mathcal{D}$ and $(X', Y') \sim \mathcal{D}'$ be the training and test datasets of size $m$, respectively. Let $U \sim \{0, 1\}^m$ be the i.i.d. random $m$ labels. Write $i_j = \lfloor i/2^{j-1}\rfloor - 2\lfloor i/2^j\rfloor$ (the $j$th bit of $i$). For $i \in [m]$ and $b \in \{0, 1\}^*$ with $|b| \le \log m$, define $Z_{b,i} = Z'_{b,i} := (X_i, Y_i)$ if $1 \le \exists j \le |b|$, $i = b \bmod 2^{|b|-1} \wedge i_j = 0 \ne 1 = b_j$; $Z_{b,i} = Z'_{b,i} := (X'_i, U_i)$ if $1 \le \exists j \le |b|$, $i = b \bmod 2^{|b|-1} \wedge i_j = 1 \ne 0 = b_j$; $(Z_{b,i}, Z'_{b,i}) := \big((X_i, Y_i), (X'_i, U_i)\big)$ otherwise. Let $\mathcal{D}_b =$

$(Z_{b,i})_{i=0}^{m-1}$ and $\mathcal{D}'_b = (Z'_{b,i})_{i=0}^{m-1}$. Let $\mathcal{A}_b = 1[\mathcal{A}\ \texttt{refutes}\ \mathcal{D}'_b] - 1[\mathcal{A}\ \texttt{refutetes}\ \mathcal{D}_b]$. It parses the given refutation gap $\mathbb{E}[\mathcal{A}_{\texttt{null}}] \approx 1$ into the $m$ pieces by $\mathbb{E}[\mathcal{A}_b] = \mathbb{E}[\mathcal{A}_{0b}] + \mathbb{E}[\mathcal{A}_{1b}]$, promising $\mathbb{E}[\mathcal{A}_{b_0}] \geq \alpha$ for some $b_0 \in \{0,1\}^{\log m}$. Let[34] $\hat{\mathcal{A}}(x,y) := 1[\mathcal{A}\ \texttt{refutes}\ (\mathcal{D}_{b_0} \backslash Z_{b_0,b_0}) \sqcup (x,y)\ ]$. The binary search version of Yao's reduction gives rise to a weak learner $\hat{\mathcal{A}}(x) := \hat{\mathcal{A}}(x,1) - \hat{\mathcal{A}}(x,0)$:

$$
\begin{aligned}
\alpha \leq \mathbb{E}[\mathcal{A}_{b_0}] &= \mathbb{E}[\hat{\mathcal{A}}(X',U) - \hat{\mathcal{A}}(X,Y)] = \mathbb{E}\big[(1/2)\big(\hat{\mathcal{A}}(X',Y'\oplus 1) + \hat{\mathcal{A}}(X',Y')\big) - \hat{\mathcal{A}}(X,Y)\big] \\
&= \mathbb{E}\big[(1/2)\big(\hat{\mathcal{A}}(X',Y'\oplus 1) - \hat{\mathcal{A}}(X',Y')\big) + \hat{\mathcal{A}}(X',Y') - \hat{\mathcal{A}}(X,Y)\big] \\
&= (1/2)\mathbb{E}[\hat{\mathcal{A}}(X')(-1)^{Y'}] + \mathbb{E}[\hat{\mathcal{A}}(X',Y')] - \mathbb{E}[\hat{\mathcal{A}}(X,Y)] \\
\Rightarrow \textit{Advantage: } & \mathbb{E}[\hat{\mathcal{A}}(X')(-1)^{Y'}] \geq 2\alpha + 2(\mathbb{E}[\hat{\mathcal{A}}(X,Y)] - \mathbb{E}[\hat{\mathcal{A}}(X',Y')]).
\end{aligned}
$$

**Weak learning:** Let $\nu_0 \approx \frac{2}{\varepsilon\alpha^2}$, $\kappa_0 \gg (\frac{\log m}{\alpha})^2 \log \frac{\nu_0 \log m}{\delta}$, $\tilde{m} \gg (\frac{1}{\alpha})^2 \log \frac{\nu_0}{\delta}$ and $\tilde{m}' \gg \frac{\tilde{m}}{\varepsilon} \log \frac{1}{\delta}$. Sample $\mathcal{D} \sim P^{\tilde{m}}(x, f(x))$, $\mathcal{D}' \sim P^{\tilde{m}'}(x, f(x))$ with $\mathcal{D} \perp \mathcal{D}'$ and fix them. Subsample $\mathcal{D}_\nu = (X_i, Y_i)_{i=0}^{m-1} \sim (P_\nu \circ \mathcal{D})^m$ and $(X'_i, Y'_i)_{i=0}^{m-1} \sim (P_\nu \circ \mathcal{D}')^m$ of Theorem 3.3's weighting $(P_\nu \circ \mathcal{D})(x,y) = P_\nu(x, y \mid (x,y) \in \mathcal{D})$, and feed them to Yao's reduction. It transforms a given refuter $\mathcal{A}$ to an advantageous weak learner $\hat{\mathcal{A}}$ through binary searching a path $b$ reaching to $b_0$ by induction on $|b| = 0, 1, \ldots, \log m - 1$ in the following manner. Draw the i.i.d. $\kappa_0$ subsamples $\{(\mathcal{D}_{\nu,\kappa}, \mathcal{D}'_{\nu,\kappa})\}_{\kappa=1}^{\kappa_0}$, feed them to $\mathcal{A}_b$ with $\mathbb{E}[\mathcal{A}_b] \geq (1 - \epsilon - \frac{\epsilon'|b|}{\log m})\frac{1}{2^{|b|}}$, and detect $b' \in \{0b, 1b\}$ to preserve $\mathbb{E}[\mathcal{A}_{b'}] \geq (1 - \epsilon - \frac{\epsilon'|b'|}{\log m})\frac{1}{2^{\ell(b')}}$. Chernoff bound parameters $\mu = (\mathbb{E}[\mathcal{A}_b] + 1)/2$ and $\gamma = \frac{\epsilon}{2^{|b|}\mu \log m}$ guarantees the successful detections in all $(\nu, |b|)$ with significance $\nu_0 \log m \cdot \boldsymbol{e}^{-\gamma^2/2 \cdot \mu\kappa_0} = o(\delta)$. In addition, $\forall \nu, |\mathbb{E}[\hat{\mathcal{A}}(X,Y)] - \mathbb{E}[\hat{\mathcal{A}}(X',Y')]| \leq 2\epsilon\alpha$ since $\mathcal{D}$ and $\mathcal{D}'$ stem from the same target $P(x, f(x))$. CB of $\mu = (\mathbb{E}[\hat{\mathcal{A}}(X,Y)] + 1)/2 = (\mathbb{E}[\hat{\mathcal{A}}(X',Y')] + 1)/2$ and $\gamma = \epsilon\alpha/\mu$ guarantees it with significance $\nu_0 \cdot O(\boldsymbol{e}^{-\frac{\gamma^2}{2+\gamma} \cdot \mu\tilde{m}}) = o(\delta)$, deriving weak learning of advantage $\forall \nu, \mathbb{E}[\hat{\mathcal{A}}(X')(-1)^{Y'}] \geq 2(1-\epsilon)\alpha$.

**Boosting:** Theorem 3.3 takes the majority vote of these $H_\nu = \hat{\mathcal{A}}$ depending on $\{(\mathcal{D}_{\nu,\kappa}, \mathcal{D}'_{\nu,\kappa})\}_{\nu,\kappa}$ to get an $\varepsilon$-learner $H(x)$ over the test dataset $x \in \mathcal{D}'$. It consults only $\mathcal{D}$'s data's labels but never to $\mathcal{D}'$'s ones, so applying Chernoff bound parameter $\gamma = 1$ on $|\mathcal{H}| \leq |\{0,1\}^{\tilde{m}}|$ promises

$$
\textit{UGEB: } \mathsf{Pr}[P(y \neq H(x)) \geq 2\varepsilon] \leq |\mathcal{H}|\boldsymbol{e}^{-\gamma/3 \cdot \varepsilon\tilde{m}'} < o(\delta).
$$

The number of refutation calls is no more than $\nu_0\kappa_0 \log m$, so the learning time is $\nu_0\kappa_0 \log m \cdot O(t)$. All refutation calls may succeed with significance $\nu_0\kappa_0 \log m \cdot O\big(\delta_{3.4}\big) = O(\delta)$. For every new prediction, the learner must access $P(x)$ and refresh $Z'$ in searching $b_0$ of Yao's reduction. □

**Theorem 3.5** (refutation to noisy PAC learning)**.** If $\eta$-noisy $\mathcal{F}$ is refutable, then $\varepsilon\eta$-noisy $\mathcal{F}$ is PAC learnable in the same way as Theorem 3.4.

*Proof.* Theorem 3.3's smoothness for $\mathrm{err}_f(\mathcal{D}) \leq \varepsilon\eta$ guarantees $\mathrm{err}_f(\mathcal{D}_\nu) \leq \eta$. Definition 3.1's $\eta$-noisy refutation promises $\mathbb{E}[\mathcal{A}_{\texttt{null}}] \approx 1$ in Theorem 3.4's Yao's reduction on binary search. It reduces Theorem 3.5 to 3.4. □

**Theorem 3.6** (refutation to PAC learning in smoothed analysis)**.** If noise-free $\mathcal{F}$ is refutable with significance $O(\delta_{3.4}^2/\delta)$, $\mathcal{F}$ is PAC learnable under any shift in the same way as Theorem 3.4.

*Proof.* Definition 3.1 assumes that the refutations called in Theorem 3.4's boosting attain the significance levels no larger than $O(\delta_{3.4}\delta)$ on average under a random shift $G$. Markov's inequality parameter $\gamma = \delta$ bounds the significance of picking a correct $G$ over all these refutations by $\nu_0\kappa_0 \log m \cdot O(\delta_{3.4}^2)/(\gamma\delta) = O(\delta_{3.4})$ with high confidence, reducing Theorem 3.6 to 3.4. □

[34] $(\mathcal{D}_{b_0} \backslash Z_{b_0,b_0}) \sqcup (x,y) = \big(Z_{b_0,0}, \ldots, Z_{b_0,b_0-1}, (x,y), Z_{b_0,b_0+1}, \ldots, Z_{b_0,m-1}\big) = \big(Z'_{b_0,0}, \ldots, Z'_{b_0,b_0-1}, (x,y), Z'_{b_0,b_0+1}, \ldots, Z'_{b_0,m-1}\big)$.

**Theorem 3.7** (refutation to noisy PAC learning in smoothed analysis)**.** If $\eta$-noisy $\mathcal{F}$ is refutable with significance $O(\delta_{3.4}^2/\delta)$, $\varepsilon\eta$-noisy $\mathcal{F}$ is as PAC learnable under any shift as in Theorem 3.4.

*Proof.* A reduction to Theorem 3.5, as Theorem 3.6 to 3.4. □

# 4 Proof Theoretic Hardness of PAC Learning DNF

The DSS of REVIEW8 [DSS16] has reduced R$k$SAT-refutation to planted $k$DNF-learning. This section will extend their worst-case reduction to smoothed ones and establish PAC2 and PAC3.

**Proof ideas of PAC2 and PAC3 on SoS degree:** As mentioned in REVIEW8, what they have proved is the hardness of refuting $\exists\theta \in \{0,1\}^n, \mathrm{err}_{f(\theta\circ x)}(\mathcal{U}) = 0$ for the uniformly random data $\mathcal{U} \subset [2n]^{d/k} \times \{0,1\}$ and the canonical DNF expression $f = \bigvee_{j=1}^{d/k} \bigwedge_{i=1}^{k} \mathbf{x}_{i+jk}$. Divide $\mathcal{U} = \mathcal{P} \sqcup \mathcal{N}$ and observe that $\forall\theta, \mathrm{err}_{f(\theta\circ x)}(\mathcal{P}) = 0$ over $\mathcal{P} = \{(x,y) \in \mathcal{D} \mid y = 1\}$ with high confidence when $2^n(1-2^{-k})^{d/k} \approx 0$, say $d \approx k2^k n \ln 2$. Also, $\forall h, \mathrm{err}_h(\mathcal{U}) \approx 1/2$ since Definition 2.1 supplies sufficiently many examples. Definition 2.1's $\eta = 0$ case obliges the PAC learner to prove $\forall\theta, \mathrm{err}_{f(\theta\circ x)}(\mathcal{N}) > 0$, or equivalently, $\bigwedge_{(x,0)\in\mathcal{N}} \bigwedge_{j=1}^{d/k} \bigvee_{i=1}^{k} (\theta \circ x_{i+jk} \oplus 1)$ is unsatisfiable. This worst-case reduction from refutation to learning is extensible to a smoothed analysis under any polarity flipper $G$ of min-entropy $\mathrm{H}_\infty(G) = (1-c)k$, $0 < c < 1$. It reduces learning the canonical DNF to proving $\bigwedge_g \bigwedge_{(x,0)\in\mathcal{N}} \bigwedge_{j=1}^{d/k} \bigvee_{i=1}^{k} (\hat{g}(\theta) \circ x_{i+jk} \oplus 1)$ as unsatisfiable. Kothari, Mori, O'Donnell, and Witmer [KMOW17] proved this refutation's hardness in the following manner. For every $(j,(x,0)) \in [d/k] \times \mathcal{N}$, a linear algebra (Lemma 4.12) on $|\mathcal{S}_j| > 2^k - 2^{(1-c)k}$ guarantees that the local solution space $\mathcal{S}_j := \{0,1\}^k \backslash \{(g(\lfloor x_{i+jk}/2 \rfloor) \oplus x_{i+jk} \oplus 1)_{i=1}^{k}\}_g$ must contain a $(t-1)$-uniform subspace (Definition 4.8) for $t = \Omega(k)$. Then, any degree-$n^\epsilon$ SoS proof may "think" the shifted $k$CNF satisfiable. Consequently, PAC learning DNF requires SoS degree $\Omega(n^\epsilon)$ even under the smoothed analysis of min-entropy $\mathrm{H}_\infty(G) = (1-c)k$.

## 4.1 SoS Lower Bounds

*Sum-of-Squares* (SoS), known by Hilbert's 17th problem [Pfi76], can prove non-negativity and even positivity of low-degree multi-linear polynomials "efficiently" [Sho87, Par00, GV01, Las01].

**Definition 4.1** (SoS proof)**.** Let $\mathbb{Q}_{\mathrm{D}}[\mathbf{x}] = \{f(\mathbf{x}) \in \mathbb{Q}[\mathbf{x}_1,\ldots,\mathbf{x}_n]/\{\forall i, \mathbf{x}_i^2 = \mathbf{x}_i\} \mid \mathbf{deg}(f) \le \mathrm{D}\}$.

$$\textit{Non-negativity proof degree:}\quad \mathbf{deg}_{\mathrm{SoS}}[f(\mathbf{x}) \ge 0] = \min\{\mathrm{D} \mid \exists f_i \in \mathbb{Q}_{\mathrm{D}/2}[\mathbf{x}_1,\ldots,\mathbf{x}_n], f = \textstyle\sum_i f_i^2\}.$$
$$\textit{Positivity proof degree:}\quad \mathbf{deg}_{\mathrm{SoS}}[f(\mathbf{x}) > 0] = \min\{\mathrm{D} \mid \exists\epsilon > 0, \mathbf{deg}_{\mathrm{SoS}}[f \ge \epsilon] \le \mathrm{D}\}.$$

As far as we know, the SoS degree is currently the most promising proof complexity for measuring the computational hardness of RCSP refutation. It has provided not only the state-of-the-art algorithms of R$k$SAT [GK01, FO05], R$k$CSP [COCF10, RRS17] and $t$-uniform RCSP[35] [AOW15, AGK21] but also the matching lower bounds of R$k$SAT, R$k$XOR [Gri01, Sch08, BM16], 2-uniform RCSP [Tul09, BCK15] and $t$-uniform RCSP [KMOW17]. This subsection will transfer the SoS degree lower bound of [KMOW17] to PAC learning hardness results.

**Definition 4.2.** The *unsatisfiability rate* of an assignment $\theta \in \{0,1\}^n$ to $\psi = (x_{i+jk})_{(i,j)\in(k]\times(m]}$ $\in k\mathrm{CNF}_n^m \cong [2n]^{km}$ is $\mathrm{unsat}_\psi(\theta) := \frac{1}{m}\sum_{j=1}^{m}\prod_{i=1}^{k} \theta \circ x_{i+jk} \oplus 1$ at $\mathbf{x} = \theta$ of $\mathrm{unsat}_\psi(\mathbf{x}) \in \mathbb{Q}_k[\mathbf{x}]$.

[35] $t$-uniform RCSP is RCSP of the $t$-uniform predicates supporting a $t$-uniform random variable in Definition 4.8.

**Theorem 4.3** (SoS hardness of RSAT refutation [KMOW17]). Any sub-linear degree SoS proof is hard to refute the uniform random $k$CNF expression $\Psi \in k\mathrm{CNF}_n^m$ with $k \geq 3$ as follows:

$$\text{SoS } \textit{hardness of } \mathrm{R}k\mathrm{SAT}\text{:}\ \Pr\big[\mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_{\Psi}(\mathbf{x}) > 0] \geq \tfrac{n}{\Delta^{2/(k-2)}\log\Delta}\big] \geq 1 - \epsilon'^k \text{ for } \Delta := m/n.$$

**Theorem 4.4** (Theorems 1.14 and 1.15 for SoS degree[36]). For $3 \leq k \leq \log \frac{s}{\log s \log n}$, PAC learning the canonical planted DNF class $\{\bigvee_{j=1}^{s}\bigwedge_{i=1}^{k}\theta \circ x_{i+jk} \mid \theta \in \{0,1\}^n\}$ under the uniform distribution requires either sample size $\Omega(n^{\frac{1-\epsilon}{2}k})$ or SoS degree $\Omega(n^{\epsilon})$.

*Proof.* Suppose the sample size $2m \approx n^{(1-\epsilon)k/2}$ and prove the SoS degree $\geq \mathrm{D} := n^{\epsilon}$ for the random constraint $\mathcal{U} \sim (k\mathrm{CNF}_n^s \times \{0,1\})^m$. Theorem 3.2's UGEB has demonstrated $\forall h, \mathrm{err}_h(\mathcal{U}) \approx 1/2$, so Definition 2.1 asks to prove $\mathrm{err}_\theta(\mathcal{U}) > 0$. We will suppose $\mathbf{deg}_{\mathrm{SoS}}[\mathrm{err}_\theta(\mathcal{U}) > 0] < \mathrm{D}$ and derive a contradiction to Theorem 4.3's SoS hardness of R$k$CSP in the following manner.

Divide the data into the positive and negative ones $\mathcal{U} = \mathcal{P} \sqcup \mathcal{N}$, and accordingly decompose

$$\mathcal{PN}\ \textit{decomposition:}\quad \mathrm{err}_\theta(\mathcal{U}) = \frac{|\mathcal{P}|}{|\mathcal{P}| + |\mathcal{N}|}\mathrm{err}_\theta(\mathcal{P}) + \frac{|\mathcal{N}|}{|\mathcal{P}| + |\mathcal{N}|}\mathrm{err}_\theta(\mathcal{N}).$$

The i.i.d. random polarities $X_{i+jk}$ mod 2 of $(X,1) \sim \mathcal{P}$ must have, under $\log(1/\delta) \ll \log n$,

$$\begin{aligned}\textit{No FPE:}\quad \Pr[\mathrm{err}_\theta(\mathcal{P}) > 0] &= \Pr\big[\exists (X,1) \in \mathcal{P}, \forall j \in (s], \exists i \in (k], \theta(\lfloor X_{i+jk}/2 \rfloor) \oplus X_{i+jk} = 0\big] \\ &< |\mathcal{P}|(1 - 1/2^k)^s < (1 + o(1))n^{(1-\epsilon)k/2}\boldsymbol{e}^{-s/2^k} \leq o(\delta).\end{aligned}$$

It implies $\mathbf{deg}_{\mathrm{SoS}}[\mathrm{err}_\theta(\mathcal{N}) > 0] < \mathrm{D}$, or equivalently $\mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_\Psi(\mathbf{x}) > 0] < \mathrm{D}$ for the random constraint $\Psi \in k\mathrm{CNF}_n^{s|\mathcal{N}|}$, which contradicts to Theorem 4.3, under $k + \log s \ll \log n$, by

$$\begin{aligned}\textit{Sub-linear degree:}\ \Delta &= (s/n)|\mathcal{N}| \approx sn^{(k-2)(1-\epsilon)/2-\epsilon} \\ \Rightarrow\quad \mathrm{D} &\geq \frac{n}{\Delta^{2/(k-2)}\log\Delta} > (n^{\epsilon(1+\frac{2}{k-2})})/(s^{\frac{2}{k-2}}(\log s + k\log n)) \gg n^{\epsilon}.\end{aligned}$$

□

**Theorem 4.5** (Theorem 1.14 and 1.15 for SoS degree under noise[37]). For $3 \leq k \leq \log \frac{s}{\log(1/\varepsilon)}$, PAC learning the $\varepsilon$-noisy canonical planted DNF is PAC learnable as Theorem 4.4.

*Proof.* The $\varepsilon$-noisy model asks to negate $\mathbf{deg}_{\mathrm{SoS}}[\mathrm{err}_\theta(\mathcal{D}) > \varepsilon] < n^{\epsilon}$. It rewrites Theorem 4.4's No FPE proof by Chernoff bound of $\gamma = \frac{\varepsilon}{(1-1/2^k)^s} - 1$ on $(1 - \frac{1}{2^k})^s \leq \varepsilon^{\log \boldsymbol{e}} \ll \varepsilon$ as follows:

$$\begin{aligned}\textit{Small FPE:}\ \Pr[\mathrm{err}_\theta(\mathcal{P}) > \varepsilon] &= \Pr\big[\varepsilon < \tfrac{1}{|\mathcal{P}|}\textstyle\sum_{(X,1)\sim\mathcal{P}} 1[\forall j \in (s], \exists i \in (k], \theta(\lfloor X_{i+jk}/2 \rfloor) \oplus X_{i+jk} = 0]\big] \\ &< \boldsymbol{e}^{-\frac{\gamma}{3}(1-1/2^k)^s|\mathcal{P}|} < \boldsymbol{e}^{-(\frac{1}{3}-o(1))(\varepsilon - \varepsilon^{\log \boldsymbol{e}})m} = o(\delta).\end{aligned}$$

Theorem 4.4's sub-linear degree analysis has shown the claimed SoS degree lower bound. □

In summary, the worst-case learning hardnesses Theorems 1.14 and 1.15 on SoS degree are fruits of the worst-case RSAT refutation hardness Theorem 4.3. Similarly, the smoothed-case hardness Theorem 1.16 will stand on the following smoothed-case RSAT refutation hardness.

**Theorem 4.6** (SoS hardness of RSAT refutation in the smoothed analysis [this paper]). Any sub-linear degree SoS proof is hard to refute the uniform random expression $\Psi \sim k\mathrm{CNF}_n^m$ of $m \leq n^{\frac{ck}{10-4\log c}}$ shifted by any flipper space of size $|\mathcal{G}| \leq 2^{(1-\epsilon)k}$ as follows:

$$\begin{array}{c}\text{SoS } \textit{hardness of } \mathrm{R}k\mathrm{SAT} \\ \textit{in smoothed analysis}\end{array}\text{:}\ \Pr\big[\mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_{\wedge_{g\in\mathcal{G}}g(\Psi)}(\mathbf{x}) > 0] \geq n^{0.06}\big] \geq 1 - \epsilon'^k.$$

[36] Set $k = \log \frac{s}{\log s \log n}$ in Theorems 4.15, 4.16, 4.27, and 4.30.

[37] Set $k = \log \frac{s}{\log(1/\varepsilon)}$ in Theorems 4.5, 4.17, 4.18, 4.21, and 4.22.

Previously, Molloy and Salavatipour [Mit02, MS07] provided a detailed map of the resolution refutation complexities under the uniform random solution spaces (mentioned in the proof ideas) $\mathcal{S}_j \subset \{0,1\}^k$ in terms of the co-cardinality $2^k - |\mathcal{S}_j|$. Meanwhile, Theorem 4.6 allows even malicious $\mathcal{S}_j$. It is a gift from a pretty general CSP refutation lower bound on SoS proof of degree guaranteed by only an "expanding" property of factor graphs [KMOW17].

**Definition 4.7** (graphical CSP)**.** A *factor graph* is a bipartite graph $(\mathcal{I} \sqcup \mathcal{J}, \mathcal{E})$ between a variable $i \in \mathcal{I}$ and a constraint $j \in \mathcal{J}$. It takes solution spaces $\mathcal{S}_j \subset \{0,1\}^{\mathcal{E}[j]}$ and presents a graphical CSP instance $\mathcal{G} = (\mathcal{I} \sqcup \mathcal{J}, \mathcal{E}, \mathcal{S})$ of density $\Delta := |\mathcal{J}|/|\mathcal{I}|$ to minimize $\mathrm{unsat}_{\mathcal{G}}(\theta) = \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} 1\big[(\theta(i))_{i\in\mathcal{E}[j]} \notin \mathcal{S}_j\big]$.

**Definition 4.8** (uniformity of solution space)**.** The *uniformity* of a space $\mathcal{S}_j \subset \{0,1\}^k$ is the maximum dimension $t \le k$ for $\mathcal{S}$ to support a $t$-uniform random variable $X$ as

$$\mathrm{unif}(\mathcal{S}_j) := \max\big\{0 \le t \le k \mid \exists X \in \mathcal{S}_j, \forall w \in \tbinom{k}{t}, \forall x \in \{0,1\}^t, \mathsf{Pr}[\forall i \in w, X_i = x_i] = 2^{-t}\big\}.$$

**Definition 4.9** (expansion)**.** Fix any $\zeta = o(1)$. A $k$CSP instance $\mathcal{G} = (\mathcal{I} \sqcup \mathcal{J}, \mathcal{E}, \mathcal{S})$ must have $k$-regular bipartite edges $\mathcal{E} \in \mathcal{I}^{k|\mathcal{J}|}$ and solution spaces $\mathcal{S}_j \subset \{0,1\}^k$. It is *random* if the edge set $\mathcal{E}$ is the uniform random variable, and D-*expanding* if any edge-induced subgraph $(u \sqcup v, w) \subset \mathcal{G} = (\mathcal{I} \sqcup \mathcal{J}, \mathcal{E})$ with at most $|v| \le \mathrm{D}$ constraints must satisfy

$$\text{D-}\textit{expanding:}\quad |u| \ge |w| - (1/2 - \zeta)|v| - (1/2)\textstyle\sum_{j\in v}\mathrm{unif}(\mathcal{S}_j).$$

**Lemma 4.10** (R$k$CSP is expanding [KMOW17])**.** For $3 \le t = \Omega(k)$ and $\mathrm{D} \ll \frac{|\mathcal{I}|}{k\Delta^{2/(t-2-2\zeta)}}$, any $k$CSP instance $\mathcal{G}$ to meet $\forall v \subset \mathcal{J}, \sum_{j\in v}\mathrm{unif}(\mathcal{S}_j) \ge (t-1)|v|$ must be

$$\text{R}k\text{CSP } \textit{is expanding:}\quad \mathsf{Pr}_{\mathcal{E}}[\mathcal{G} \text{ is D-expanding}] \ge 1 - \epsilon^k \text{ for the uniform random edge set } \mathcal{E}.$$

*Proof.* The uniform random $\mathcal{E} \sim \mathcal{I}^{k|\mathcal{J}|}$ assures Definition 4.9's D-expanding with significance

$$\begin{aligned}
&\mathsf{Pr}_{\mathcal{E}}\big[\exists w \subset \mathcal{E}\colon v = \mathcal{J}[w], u = \mathcal{I}[w], |v| \le \mathrm{D}, |u| \le k|v|, |u| + (t/2-\zeta)|v| \le |w| \le k|v|\big]\\
&\le \textstyle\sum_{|v|,|u|,|w|}\binom{|\mathcal{J}|}{|v|}\binom{|\mathcal{I}|}{|u|}\binom{|w|-1}{|v|-1}\max\limits_{(|w[j]|)_{j\in v}}\prod_{j\in v}\mathsf{Pr}_{\mathcal{E}_j\in\binom{|\mathcal{I}|}{|w[j]|}}[\mathcal{E}_j \subset u]\\
&< \textstyle\sum_{|v|,|u|,|w|}(\frac{e|\mathcal{J}|}{|v|})^{|v|}(\frac{e|\mathcal{I}|}{|u|})^{|u|}(\frac{e|w|}{|v|})^{|v|}(\frac{|u|}{|\mathcal{I}|})^{|w|}\\
&\le \textstyle\sum_{|v|,|u|,|w|}\Big(\boldsymbol{e}^{2+\frac{|u|}{|v|}}\frac{|u||w|}{|v|^2}(\frac{|u|}{|\mathcal{I}|})^{\frac{t}{2}-\zeta-1}\Delta\Big)^{|v|} \quad (\because |u| + (t/2-\zeta)|v| \le |w|)\\
&\le \textstyle\sum_{|u|,|w|}\sum_{|v|}\Big(k^2\boldsymbol{e}^{2+k}\big(\mathrm{D}\cdot k\Delta^{\frac{2}{t-2-2\zeta}}/|\mathcal{I}|\big)^{\frac{t}{2}-\zeta-1}\Big)^{|v|} \quad (\because |u| \le k|v|, |w| \le k|v|)\\
&\overset{\star}{<} 2k^4\boldsymbol{e}^{2+k}\big(\mathrm{D}\cdot k\Delta^{\frac{2}{t-2-2\zeta}}/|\mathcal{I}|\big)^{\frac{t}{2}-\zeta-1} < \epsilon^k, \quad (\because \mathrm{D} \ll \tfrac{|\mathcal{I}|}{k\Delta^{2/(t-2-2\zeta)}})
\end{aligned}$$

where $\overset{\star}{<}$ bounds the geometric sum by its start term $k^2\boldsymbol{e}^{2+k}\big(k\mathrm{D}\Delta^{\frac{2}{t-2-2\zeta}}/|\mathcal{I}|\big)^{\frac{t}{2}-\zeta-1} = o(1)$. ☐

**Theorem 4.11** (SoS hardness of expanding CSP's refutation [KMOW17])**.** Any low degree SoS proof is hard to refute any $d$-expanding CSP instance $\mathcal{G}$ of $\max_\theta \mathrm{val}_\theta(\mathcal{G}) < 1$ and $\forall j, |\mathcal{E}[j]| \le \zeta\mathrm{D}$:

$$\text{SoS } \textit{hardness of expanding } \text{CSP:}\ \mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_{\mathcal{G}}(\mathbf{x}) > 0] \ge \zeta\mathrm{D}/3.$$

**Lemma 4.12.** For any set $\mathcal{S}_j \subset \{0,1\}^k$ and any integers $1 \le t \le r \le k$,

$$(\tbinom{k}{t} < 2^{r-t}) \wedge (2^k - 2^{k-r} < |\mathcal{S}_j|) \Rightarrow \mathrm{unif}(\mathcal{S}_j) \ge t.$$

*Proof.* Randomly generate a $k \times r$ matrix $\mathcal{M} \sim \mathbb{F}_2^{k\times r}$. Then, all of its $t \times r$ sub-matrices happen to have the full rank $t$ with a probability of at least $1 - 2^{-r}2^t\binom{k}{t} > 0$. The probabilistic method provides such a matrix $\mathcal{M}$. Divide the $k$-dimensional linear space $\mathbb{F}_2^k$ by this $\mathcal{M}$ to make the $2^{k-r}$ (or more if $\mathcal{M}$ is degenerate) cosets. Shifting the same linear kernel yields these disjoint affine subspaces of $\mathbb{F}_2^k$ obtained. Then, the pigeon-hole principle over $|\mathbb{F}_2^k - \mathcal{S}_j| < 2^{k-r}$ can pick a coset disjoint from $\mathbb{F}_2^k - \mathcal{S}_j$. It gives a desired $t$-uniform random variable supported by $\mathcal{S}_j$. □

**Theorem 4.13.** Let $t_{4.13} := \frac{ck}{1+\log \boldsymbol{e}+1.725\log((1+\log \boldsymbol{e})/c)} \geq 3$ and $\mathrm{D}_{4.13} := \frac{3\zeta n}{\Delta^{2/(t_{4.13}-2-2\zeta)}} \geq k/\zeta$ for $0 < c < 1$. Any $k$CSP instance $\mathcal{G}$ with $\forall j, |\mathcal{S}_j| \geq 2^k - 2^{(1-c)k}$ under the uniform random $\mathcal{E}$ has

$$\underset{\textit{of graphical}\ \mathrm{R}k\mathrm{CSP}}{\textit{The hardness}}:\ \Pr_{\mathcal{E}\sim\mathcal{I}^{k|\mathcal{J}|}}\big[\mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_{\mathcal{G}}(\mathbf{x})] > 0\big] \geq \zeta\mathrm{D}_{4.13}/3\big] \geq 1-\epsilon^k.$$

*Proof.* Since $|\mathcal{S}_j| \geq 2^k - 2^{k-(r-1)}$ for $r = \lfloor ck \rfloor$, Theorem 4.12 of $t = t_{4.13}$ shows $\mathrm{unif}(\mathcal{S}_j) \geq t-1$:

$$1 + \log \boldsymbol{e} + 1.725\log\left((1+\log \boldsymbol{e})/c\right) < c\left((1+\log\boldsymbol{e})/c\right)^{1.725} \quad \Rightarrow$$
$$2^{(t-1)-(r-1)}\binom{k}{t-1} < 2^{t-ck}(\tfrac{ek}{t})^t = 2^{(1+\log \boldsymbol{e}+\log(k/t)-ck/t)t} < 2^{(1+\log\boldsymbol{e}+1.725\log\frac{1+\log \boldsymbol{e}}{c}-\frac{ck}{t})t} = 1.$$

Consequently, Theorem 4.10's R$k$CSP's expansion has revealed $\Pr_{\mathcal{E}}\big[\mathcal{G}$ is $\mathrm{D}_{4.13}$-expanding$\big] \geq 1-\epsilon^k$ for $\mathrm{D}_{4.13} \ll |\mathcal{I}|/(k\Delta^{2/(t_{4.13}-2-2\zeta)})$, so that Theorem 4.11 with $|\mathcal{E}[j]| \leq k \leq \zeta\mathrm{D}_{4.13}$ demonstrates $\mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_{\mathcal{G}}(\mathbf{x}) > 0] \geq \zeta\mathrm{D}_{4.13}/3$ with confidence $1-\epsilon^k$. □

**Theorem 4.14** (Theorem 4.6)**.** Any low-degree SoS proof is hard to refute the uniform random $k$CNF expression $\mathit{\Psi} \sim k\mathrm{CNF}_n^m$ shifted by any flipper of size $|\mathcal{G}| \leq 2^{(1-c)k}$ for $0 < c < 1$:

$$\text{SoS } \textit{hardness in smoothed analysis:}\ \Pr\big[\mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_{\wedge_g g(\mathit{\Psi})}(\mathbf{x}) > 0\big] \geq \zeta\mathrm{D}_{4.13}/3\big] \geq 1-\epsilon^k.$$

*Proof.* Rewrite $\mathrm{unsat}_{\wedge_g g(\mathit{\Psi})}(\theta) = \mathrm{unsat}_{\mathcal{G}}(\theta)$ by a CSP $\mathcal{G}$ corresponding to $\mathit{\Psi} = (x_{i+jk}) \in k\mathrm{CSP}_n^m$:

$$\mathcal{I} = [n), \mathcal{J} = (m], \mathcal{E}(\mathcal{G}) = \{(\lfloor x_{i+jk}/2\rfloor, j) \mid i \in (k], j \in \mathcal{J}\},$$
$$\mathcal{S}_j = \{0,1\}^{\mathcal{I}[j]} \backslash \big\{(g(\lfloor x_{i+jk}/2\rfloor) \oplus x_{i+jk} \oplus 1)_{i=1}^k \mid \Pr[G = g] > 0\big\},$$
$$\mathrm{unsat}_{\mathcal{G}}(\theta) = \frac{1}{m}\sum_g \Pr[G=g]\sum_{j=1}^m \bigwedge_{i=1}^k \theta(g(\lfloor x_{i+jk}/2\rfloor) \oplus x_{i+jk} \oplus 1).$$

Since $|\mathcal{S}_j| \geq 2^k - |\mathcal{G}| \geq 2^k - 2^{(1-c)k}$, Theorem 4.14 reduces to 4.13 and derives 4.6 by taking[38]

$$(\epsilon, \mathrm{D}, t, m) = (0.066, \mathrm{D}_{4.13}, t_{4.13}, n^{\frac{ck}{10-4\log c}})$$
$$\Rightarrow 2(1+\log\boldsymbol{e}+1.725\log\left((1+\log\boldsymbol{e})/c\right)) < (1-\epsilon)(10+4\log(1/c))$$
$$\Rightarrow \textit{Sub-linear degree: } \zeta\mathrm{D}/3 = \frac{\zeta^2 n}{k\Delta^{2/(t-2-o(1))}} > \frac{\zeta^2 n/k}{\left(m^{\frac{(1-\epsilon)(10-4\log c)}{ck}} \cdot n^{-\frac{t}{2}}\right)^{\frac{2}{t}\cdot\frac{1}{1-(2+o(1))/t}}} = \frac{\zeta^2 n/k}{n^{\frac{1-\epsilon-2/t}{1-(2+o(1))/t}}} > n^{0.06}.$$

□

**Theorem 4.15** (Theorem 1.16 for SoS degree under flipper)**.** For $3 \leq k \leq \log\frac{s}{\log s \log n}$ and $0 < c < 1$, PAC learning the canonical planted DNF $\{\bigvee_{j=1}^s \bigwedge_{i=1}^k \theta \circ x_{i+jk} \mid \theta \in \{0,1\}^n\}$ under the uniform distribution shifted by any flipper $G$ of $\mathrm{H}_\infty(G) = (1-c)k$ must take either sample size $\Omega(n^{(1-\epsilon)t_{4.13}/2})$ or SoS proof of degree $\Omega(n^\epsilon)$.

[38]This sub-linear degree analysis will deduce not only Theorem 4.6 but also Theorem 1.16 from Theorems 4.15–4.18 and 4.22, and Theorem 1.22 from Theorems 6.11 and 6.12, too.

*Proof.* Adjust Theorem 4.4's one to $\mathrm{H}_\infty(G) = (1-c)k$. No FPE analysis changes therein to

$$\begin{aligned}&\mathsf{Pr}\big[\exists g, \exists (g(X),1) \in \mathcal{P}, \forall j \in (s], \exists i \in (k], \theta(\lfloor X_{i+jk}/2 \rfloor) \oplus X_{i+jk} \oplus g(\lfloor X_{i+jk}/2 \rfloor) = 0\big] \\ &< |\mathcal{G}||\mathcal{P}|(1-1/2^k)^s < 2^{(1-c)k}(1+o(1))n^{(1-\epsilon)k/2}\boldsymbol{e}^{-s/2^k} \le o(\delta).\end{aligned}$$

Since $\mathsf{Pr}\big[\mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_{G(\Psi)}(\mathbf{x})] > 0\big] \ge \Omega(\delta) \Rightarrow \exists g, \mathrm{unsat}_{g(\Psi)}(\mathbf{x}) > 0 \Leftrightarrow \mathrm{unsat}_{\bigwedge_g g(\Psi)}(\mathbf{x}) > 0$, Theorem 4.4's sub-linear degree analysis at $(t, \mathrm{D}) = (t_{4.13}, \mathrm{D}_{4.13})$ derives a contradiction to 4.14:

$$\textit{Sub-linear degree: } \zeta\mathrm{D}/3 = \frac{\zeta^2 n/3}{k\Delta^{2/(t-2-o(1))}} \ge \frac{\zeta^2 n/3}{k\big(n^{(1-\epsilon)t/2\cdot 2/t}(s/n)^{2/t}\big)^{\frac{1}{1-(2+o(1))/t}}} \gg n^{\epsilon}.$$

□

**Theorem 4.16** (Theorem 1.16 for SoS degree under flipper and noise)**.** The $\varepsilon$-noisy canonical planted DNF is PAC learnable in the same way as Theorem 4.15.

*Proof.* Adjust Theorem 4.5's proof to get the small FPE by

$$\begin{aligned}&\mathsf{Pr}\big[\varepsilon < \tfrac{1}{|\mathcal{P}|}\textstyle\sum_{(X,1)\sim\mathcal{P}} 1[\exists g, \forall j \in (s], \exists i \in (k], \theta(\lfloor X_{i+jk}/2 \rfloor) \oplus X_{i+jk} \oplus g(\lfloor X_{i+jk}/2 \rfloor) = 0]\big] \\ &< \boldsymbol{e}^{-\frac{\gamma}{3}(1-1/2^k)^s|\mathcal{P}|} < 2^{(1-c)k}\boldsymbol{e}^{-(\frac{1}{3}-o(1))(\varepsilon-\varepsilon^{\log \boldsymbol{e}})m} = o(\delta).\end{aligned}$$

Sub-linear degree analysis is the same as Theorem 4.15, which contradicts 4.14. □

Theorems 4.15 and 4.16's smoothed analysis under a flipper $G \in \{0,1\}^{dn}$ are extensible to Lemma 2.2's general shift $G = (\Phi_i, \Psi_i)_{i=1}^d \in (\mathbb{S}_n \times \{0,1\}^n)^d$ for learning a planted function $f_d(\theta_1 \circ x_1, \ldots, \theta_d \circ x_d)$ hiding an assignment $\theta \in \{0,1\}^{dn}$.

**Theorem 4.17** (Theorem 1.16[39] for SoS degree)**.** For $3 \le k \le \log \frac{s}{\log(1/\varepsilon)}$ and $0 < c < 1$, PAC learning the canonical planted DNF class $\{\bigvee_{j=1}^s \bigwedge_{i=1}^k \theta_{i+jk} \circ x_{i+jk} \mid \theta \in \{0,1\}^{ksn}\}$ under the uniform distribution perturbed by any shift $G \in (\mathbb{S}_n \times \{0,1\}^n)^d$ of $\mathrm{H}_\infty(G) = (1-c)k$ requires either sample size $\Omega\big((n/4^{(1-c)k})^{(1-\epsilon)t_{4.13}/2}\big)$ or SoS degree $\Omega\big((n/4^{(1-c)k})^{\epsilon}\big)$.

*Proof.* A reduction to Theorem 4.15. Force SA1's adversary to choose the hidden parameter $\theta_\iota \in \{0,1\}^n$, $\iota = i + jk$, in the following manner. Let $\mathcal{O}_\iota(a) = \{\phi_i(a) \mid (\phi_\iota, \psi_\iota) \in \mathcal{O}\}$ be the orbit permuting an attribute $a \in [n)$, $\mathcal{O}_\iota^{-1} \circ \mathcal{O}_\iota(a) = \{\phi_\iota'^{-1}(\phi_\iota(a)) \mid (\phi_\iota, \psi_\iota), (\phi_\iota', \psi_\iota') \in \mathcal{O}\}$, and $\mathcal{A}_\iota \subset [n)$ be a maximal attribute set of these orbits with $\mathcal{O}(a) \ne \mathcal{O}(a') \Rightarrow \mathcal{O}_\iota(a) \cap \mathcal{O}_\iota(a') = \emptyset$. Since $\bigsqcup_{\mathcal{O}_\iota(a) \in \mathcal{A}_\iota} \mathcal{O}_\iota^{-1} \circ \mathcal{O}_\iota[a] \supset [n)$, $|\mathcal{A}_\iota| \ge n/|\mathcal{O}|^2 := n'$. Bound the adversary's choice of the hidden $\theta \in \{0,1\}^{dn}$ to make $\theta_\iota \circ x_\iota$ invariant modulo these orbits $\mathcal{O}_\iota(a)$, i.e., $\forall i, \lfloor x_\iota/2 \rfloor \in \mathcal{O}_\iota(a) \Rightarrow \theta_\iota(x_\iota) = \theta_\iota(a) \oplus x$. Further, the adversary must choose $\theta$ from $\#_\iota(a) = \#_{\iota'}(a') \Rightarrow \theta_\iota(a) = \theta_{\iota'}(a')$, where $\#_\iota$ is a linear order over $\mathcal{A}_\iota$. Then, learning under $G$ reduces to learning under the induced flipper over $\prod_{i=1}^d (\mathcal{A}_\iota \times \{0,1\})$. It replaces $n$ with $n' = n/|\mathcal{O}|^2$, $2m$ with $2m' \approx n'^{(1-\epsilon)t/2}$ of $t = t_{4.13}$, and $\mathrm{D}$ with $\mathrm{D}' = 3\zeta n'/(sm'/n')^{2/(t-2-2\zeta)}$. Still, Theorem 4.15's sub-linear degree analysis derives a contradiction to Theorem 4.14:

$$\textit{Sub-linear degree: } \zeta\mathrm{D}'/3 = 3\zeta n'/(sm'/n')^{2/(t-2-2\zeta)} \gg n'^{\epsilon}.$$

□

**Theorem 4.18** (Theorem 1.16 for SoS degree under noise)**.** For $3 \le k \le \log \frac{s}{\log(1/\varepsilon)}$ and $0 < c < 1$, PAC learning the $\varepsilon$-noisy canonical planted DNF class $\{\bigvee_{j=1}^s \bigwedge_{i=1}^k \theta_{i+jk} \circ x_{i+jk} \mid \theta \in \{0,1\}^{ksn}\}$ under the uniform distribution perturbed by any shift $G \in (\mathbb{S}_n \times \{0,1\}^n)^d$ of $\mathrm{H}_\infty(G) = (1-c)k$ requires either sample size $\Omega\big((n/4^{(1-c)k})^{(1-\epsilon)t_{4.13}/2}\big)$ or SoS degree $\Omega\big((n/4^{(1-c)k})^{\epsilon}\big)$.

*Proof.* A reduction to Theorem 4.16 as 4.17 to 4.15. □

[39] Set $(k, n', 2m', d', \epsilon) = (\log \frac{s}{\log(1/\varepsilon)}, \frac{n}{4^{(1-c)k}}, n'^{(1-\epsilon)t_{4.13}/2}, 3\zeta n'/(sm'/n')^{2/(t_{4.13}-2-2\zeta)}, 0.065) \Rightarrow n'^{\epsilon} > n^{0.06}$.

## 4.2 General LP Lower Bounds

Linear Programming is the most popular approach taken in industrial applications of optimization. It enjoys polynomial-time algorithms [Kha80, Kar84] and is a practically excellent solver over the decades with reason, the simplex algorithm with polynomial-time smoothed complexity [Dan51, ST04]. Moreover, Sherali-Adams LP hierarchy can solve CSP [OS18, HST20] and refute RCSP [OS18] as efficiently as SoS hierarchy, even matching to the known SDP lower bound [CMM09]. Worst-case LP relaxation size lower bounds hold for not only specific lift and project schemes, e.g., Lovás-Schrijver [LS91, ABLT06, STT07, AAT11, TW13] and Sherali-Adams [SA90, CMM09, BGMT12, OW14, ALN16], but also the general LP hierarchy [Yan91, CLRS16, KMR17]. Recently, Brown-Cohen and Raghavendra [BCR20] have established "average-case" sub-exponential size lower bounds of RCSP on the general LP.

**Definition 4.19** (LP proof)**.** A *lift* $\varphi$ of a function $f(\theta) : \mathcal{S} \to \mathbb{R}$ are embeddings $\varphi(f), \varphi(\theta) \in \mathbb{R}^s$ to a higher dimensional metric space[40] $\mathbb{R}^e$. Let $\mathcal{P} \subset \mathbb{R}^e$ be a polytope $\mathcal{P} = \{x \in \mathbb{R}^e \mid Ax \le b\}$.

$$\text{LP \textit{proof size:}}\ \ \mathbf{size}_{\mathrm{LP}}[f(\mathbf{x}) > 0] = \min\left\{ e \;\middle|\; \begin{matrix} \exists\varphi, \exists\mathcal{P}, \forall\theta \in \mathcal{S}, f(\theta) = \langle\varphi(f), \varphi(\theta)\rangle \wedge \\ \varphi(\mathcal{S}) \subset \mathcal{P} \ \wedge\ \min\limits_{x\in\mathcal{P}} \langle\varphi(f), x\rangle > 0 \end{matrix} \right\}.$$

**Theorem 4.20** (LP hardness of RSAT refutation [BCR20])**.** Suppose $\mathcal{G} = (\mathcal{I} \sqcup \mathcal{J}, \mathcal{E}, \mathcal{S})$ with $\log(1/\varepsilon) \ll \log|\mathcal{J}|$ has solution spaces $\mathcal{S}_j \subset \{0,1\}^k$ with $\forall j \in \mathcal{J}, \mathrm{unif}(\mathcal{S}_j) \ge t - 1 \ge 2$. Any sub-exponential size LP proof cannot refute any such CSP instance $\mathcal{G}$ with the uniform random bipartite edge span $\mathcal{E} \sim \mathcal{I}^{k|\mathcal{J}|}$ as follows:

$$\textit{Expansion:}\quad \Pr\big[\mathbf{size}_{\mathrm{LP}}[\mathrm{unsat}_{\mathcal{G}}(\mathbf{x}) > \varepsilon] \ge \exp\big(\big(\tfrac{|\mathcal{I}|^{(t-2)/2}}{\Delta}\big)^{2(1-\epsilon')/k}\big)\big] \ge 1 - o(1).$$

**Theorem 4.21** (Theorems 1.14 and 1.15 for LP size under noise)**.** For $3 \le k \le \log\frac{s}{\log(1/\varepsilon)}$, PAC learning the $\varepsilon$-noisy canonical planted DNF class $\{\bigvee_{j=1}^{s}\bigwedge_{i=1}^{k}\theta \circ x_{i+jk} \mid \theta \in \{0,1\}^n\}$ under the uniform distribution requires either sample size $\Omega(n^{(1-\epsilon)k/2})$ or LP-size $\Omega(\exp(n^{\epsilon}))$.

*Proof.* To follow Theorem 4.5's proof, assume $\mathbf{size}_{\mathrm{LP}}[\mathrm{err}_{\theta}(\mathcal{D}) > \varepsilon] \le \exp(n^{\epsilon})$. The small FPE gives $\mathbf{size}_{\mathrm{LP}}[\mathrm{unsat}_{\Psi}(\mathbf{x}) > \varepsilon/2] \le \exp(n^{\epsilon})$. Take $2m \approx n^{(1-\epsilon')k/2}$, $t = k$, $\epsilon = \epsilon'(1-\epsilon')$, and derive a contradiction to Theorem 4.20 by replacing Theorem 4.4's sub-linear degree analysis with

$$\textit{Sub-exp size:}\ \mathbf{size}_{\mathrm{LP}}[\mathrm{unsat}_{\Psi}(\mathbf{x}) > \tfrac{\varepsilon}{2}] \ge \exp\big(\big(\tfrac{2n^{k/2}}{n^{(1-\epsilon')k/2}}\big)^{\frac{2(1-\epsilon')}{k}}\big) = \exp\big(2^{\frac{(\epsilon'-1)}{k}} n^{\epsilon'(1-\epsilon')}\big) = \exp(n^{\epsilon}).\quad\square$$

**Theorem 4.22** (Theorem 1.16 for LP size under noise)**.** For $3 \le k \le \log\frac{s}{\log(1/\varepsilon)}$ and $0 < c < 1$, PAC learning the $\varepsilon$-noisy canonical planted DNF class $\{\bigvee_{j=1}^{s}\bigwedge_{i=1}^{k}\theta \circ x_{i+jk} \mid \theta \in \{0,1\}^{ksn}\}$ under the uniform distribution perturbed by any shift $G \in (\mathbb{S}_n \times \{0,1\}^n)^d$ of $\mathrm{H}_\infty(G) = (1-c)k$ requires either sample size $\Omega\big((n/4^{(1-c)k})^{(1-\epsilon)t_{4.13}/2}\big)$ or LP-size $\Omega\big(\exp\big((n/4^{(1-c)k})^{\epsilon}\big)\big)$.

*Proof.* A reduction to 4.21, like 4.18 to 4.16. □

## 4.3 Lower Bounds on Resolution and Polynomial Calculus

Resolution (Res) and Polynomial calculus (PC) are the most studied propositional and algebraic proof systems in the fields of automated theorem proving and proof complexity lower bounds

[40]The inner product $\langle a, b\rangle = \sum_{i=1}^{m} a_i b_i$ induces the metric into the vector field $\mathbb{R}^e$.

[BP98, Nor15]. PC may contain the twin variables[41] to simulate Res for stronger lower bounds on width and space [ABSRW02]. They have provided not only the most popular SAT solvers [DP60, DLL62, CEI96, BJS97, MS99, MMZ$^+$01] but also the first breakthrough of proving RSAT refutation hardness made in Res [CS88, BP96, BKPS98, BSW01] and PC [AR01, BSI10].

**Definition 4.23** (resolution proof)**.** For disjunctive constraints $\xi_j \in \mathcal{F} = \{\bigvee_{i\in w} \mathbf{x}\circ i, w \subset [2n)\}$ over the $n$ Boolean indeterminates $\{\mathbf{x}(i)\}_{i\in[n)}$ with $\mathbf{x}\circ i := \mathbf{x}(\lfloor i/2 \rfloor) \oplus i$ for $i \in [2n)$,

$$\textit{Resolution proof size:}\ \mathbf{size}_{\mathrm{Res}}(\wedge_{j\in(m]}\xi_j \not\equiv 1) = \min\left\{ \mathrm{S} \mid \begin{array}{c} \exists\{\xi_j\}_{j=m+1}^{\mathrm{S}}, \xi_e = 0, \forall j > m, \exists i \in [n), \exists \kappa < j, \exists \kappa' < j, \exists \xi' \in \mathcal{F}, \\ \xi_\kappa = \xi \vee \mathbf{x}\circ(2i) \text{ and } \xi_{\kappa'} = \xi \vee \mathbf{x}\circ(2i-1),\ \text{or } \xi_j = \xi_\kappa \vee \xi' \end{array} \right\}.$$

**Definition 4.24** (PC proof)**.** For low-degree multi-linear polynomial constraints $\xi_j \in \mathbb{Q}_{\mathrm{D}}[\mathbf{x}]$,

$$\text{PC } \textit{proof degree:}\ \mathbf{deg}_{\mathrm{PC}}(\wedge_{j=1}^m 1[\xi_j = 0] \not\equiv 1) := \min\left\{ \mathrm{D} \mid \begin{array}{c} \exists e, \exists\{\xi_j\}_{m=m+1}^{e}, \xi_e = 1 \wedge \forall j > m, \exists i \in [n), \exists \kappa < j, \exists \kappa' < j, \exists a \in \mathbb{Q}, \\ \xi_j \in \{\xi_\kappa + a\xi_{\kappa'},\ \xi_\kappa \cdot \mathbf{x}\circ(2i),\ \mathbf{x}\circ(2i) + \mathbf{x}\circ(2i-1) - 1\} \end{array} \right\}.$$

**Theorem 4.25** (Res hardness of RSAT refutation [BSW01] )**.** Any sub-exponential size Res proof is hard to refute the uniform random $\Psi \sim k\mathrm{CNF}_n^m$ with $k \ge 3$ and $\Delta = o(n^{\frac{k-2}{2}})$ as follows:

$$\mathsf{Pr}_{\Psi\sim k\mathrm{CNF}_n^m}\big[\mathbf{size}_{\mathrm{Res}}[\mathrm{unsat}_\Psi(\mathbf{x}) > 0] \ge \exp(\frac{n}{\Delta^{2/(k-2)}\log\Delta})\big] \ge 1 - o(1).$$

**Theorem 4.26** (PC hardness of RSAT refutation [AR01, BSI10] )**.** Any sub-exponential size PC proof is hard to refute the uniform random $k$CNF $\Psi \sim k\mathrm{CNF}_n^m$ with $k \ge 3$ and $\Delta = o(n^{\frac{k-2}{2}})$:

$$\mathsf{Pr}_{\Psi\sim k\mathrm{CNF}_n^m}\big[\mathbf{deg}_{\mathrm{PC}}[\mathrm{unsat}_\Psi(\mathbf{x}) > 0] \ge \Omega\big(\frac{n}{\Delta^{2/(k-2)}\log\Delta}\big)\big] \ge 1 - o(1).$$

**Theorem 4.27** (Theorems 1.14 and 1.15 for Res size and PC degree)**.** For $3 \le k \le \log\frac{s}{\log s \log n}$, PAC learning the canonical planted DNF $\{\bigvee_{j=1}^s \bigwedge_{i=1}^k \theta\circ x_{i+jk} \mid \theta \in \{0,1\}^n\}$ under the uniform distribution requires sample size $\Omega(n^{(1-\epsilon)k/2})$ unless Res-size is $\Omega(\exp(n^\epsilon))$ and PC-degree $\Omega(n^\epsilon)$.

*Proof.* The same with Theorem 4.4's one but applying Theorems 4.25 and 4.26 for the sub-linear degree analysis to derive contradictions to Res size and PC degree lower bounds, respectively, instead of Theorem 4.3. □

Berkholz [Ber18] showed that SoS could simulate PC over the Boolean variables without blowing up degree and size, although neither non-Boolean SoS [GV01], Nullstellensatz [BOCIP02], nor Sherali-Adams LP [Ber18] can do it.

**Theorem 4.28** (PC to SoS [Ber18])**.** Any PCR proof of $\mathbf{deg}_{\mathrm{PC}}[\mathrm{unsat}_\Psi(\mathbf{x}) > 0] \le \mathrm{D}$ is rewritable to an SoS proof of $\mathbf{deg}_{\mathrm{SoS}}[\mathrm{unsat}_\Psi(\mathbf{x}) > 0] \le 2\mathrm{D}$ in polynomial time.

**Theorem 4.29** (PC hardness of RSAT refutation in smoothed analysis)**.** Any low-degree PC proof is hard to refute the uniform random $k$CNF expression $\Psi \sim k\mathrm{CNF}_n^m$ shifted by any flipper space of size $|\mathcal{G}| \le 2^{(1-c)k}$ for $0 < c < 1$ as follows:

$$\mathsf{Pr}\big[\mathbf{deg}_{\mathrm{PC}}[\mathrm{unsat}_{\wedge_g g(\Psi)}(\mathbf{x}) > 0\big] \ge \zeta\mathrm{D}_{4.13}/3\big] \ge 1 - \epsilon'^k.$$

[41] The twin variable of $\mathbf{x}_i$ is another formal variable $\bar{\mathbf{x}}_i$ with the complementary axiom $\mathbf{x}_i + \bar{\mathbf{x}}_i - 1 = 0$.

*Proof.* A reduction to Theorem 4.14 via 4.28. □

**Theorem 4.30** (Theorem 1.16 for PC degree)**.** For $3 \le k \le \log \frac{s}{\log s \log n}$ and $0 < c < 1$, PAC learning the canonical planted DNF $\{\bigvee_{j=1}^{s} \bigwedge_{i=1}^{k} \theta_{i+jk} \circ x_{i+jk} \mid \theta \in \{0,1\}^{ksn}\}$ under the uniform distribution perturbed by any shift of min-entropy $\mathrm{H}_\infty(G) = (1-c)k$ requires either sample size $\Omega\big((n/\frac{n}{4^{(1-c)k}})^{(1-\epsilon)t_{4.13}/2}\big)$ or PCR degree $\Omega\big((\frac{n}{4^{(1-c)k}})^{\epsilon}\big)$.

*Proof.* The same with Theorem 4.17's one but applying Theorem 4.29 instead of 4.14. □

# 5 PAC Learning DNF in Smoothed Analysis

The previous section established PAC2 and PAC3, the unlearnability of the planted $s$-term DNF from $n^{\Theta(\log s)}$ data when the min-entropy is below the problem size $\log s$. This section will demonstrate PAC1 and PAC4 for the learnability when the min-entropy goes beyond $\log s$.

**PAC1:** Let us begin by reviewing the current best worst-case DNF learning algorithm.

**Theorem 5.1** (computational complexity of LP [Kar84, Vai90])**.** Any LP with $n$ variables and $m$ constraints is solvable to $\ell$-bit precision in deterministic $O((m+n)^{1.5} n\ell)$ time.

**Theorem 5.2** (threshold degree of planted $s$-term DNF [KS04])**.** Polynomial threshold functions of degree $\mathrm{D} = O(d^{1/3} \log s)$ can express any planted $s$-term DNF $f(\mathbf{x}_1, \ldots, \mathbf{x}_d)$ by

$$\textit{Threshold polynomial of DNF:}\ (-1)^{f(\mathbf{x}_1,\ldots,\mathbf{x}_d)} = \operatorname{sgn}\big(\textstyle\sum_w a_w (-1)^{\sum_{i\in w} \mathbf{x}_i}\big),\ a_w \in \mathbb{Q}, w \subset [n], |w| \le \mathrm{D}.$$

**Theorem 5.3** (PAC learning DNF [KS04])**.** The planted $s$-term $\mathrm{DNF}_d$ hiding $\theta \in \{0,1\}^{dn}$ is PAC learnable in deterministic $n^{O(d^{1/3} \log s)}$ time.

*Proof.* Solve an LP instance $\forall j \in (m), (-1)^{y(j)} = \operatorname{sgn}\big(\sum_w a_w (-1)^{\sum_{i\in w} \theta_i \circ x_i(j)}\big)$ of Theorem 5.2's threshold polynomial of DNF. Inside the sgn is a linear function of at most $n' := \sum_{k=0}^{d} n^k \binom{d}{k}$ variables $\mathbf{x}_{w,a} = (-1)^{\sum_{i\in w} \theta_i(a_i)} \in \{1,-1\}$ for $(w,a) \in \binom{d}{k} \times n^k$, $k \le d$. Hence, Theorem 5.1's LP algorithm can find a solution by $O\big(\log(n'/\varepsilon)\big)$-bit precision in $O(m^{1.5} n' \log(n'/\varepsilon))$ time. Since this hypothesis has bit-length $O(n' \log(n'/\varepsilon))$, an $n^{O(d^{1/3} \log s)}$ amount of data assures Definition 2.1's $O(\varepsilon)$-learning with significance $2^{O(n' \log(n'/\varepsilon))}(1-\varepsilon)^m = o(\delta)$. □

**PAC4:** We will translate the known efficient R$k$SAT refutation [COCF10, AOW15, BM16] and its derandomization [Fei07, AOW15, Wit17, AGK21] of Review6 into $k$DNF learning. They are SDP algorithms [Kha80, Ans00, NN94, LSW15, JLSW20, JKL+20] to solve *Grothendieck Inequality* (GIE) and find refutation certificates.

**Theorem 5.4** (GIE [Gro52])**.** There is a universal constant $c_{\mathtt{g}} \le \frac{g}{2\ln(1+\sqrt{2})} < 1.8$ for any $n$ by $n$ matrix $\mathcal{M}$ over $\mathbb{R}$, $u_1, \ldots, u_n, v_1, \ldots, v_n \in \mathbb{R}^{2n}$, and $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}$,

$$\textit{Grothendieck Inequality (GIE):}\ \max_{\|u_i\|,\|v_j\|\le 1} \textstyle\sum_{i,j} \mathcal{M}_{ij} \langle u_i, v_j\rangle \le c_{\mathtt{g}} \max_{|x_i|,|y_j|\le 1} \sum_{i,j} \mathcal{M}_{ij} x_i y_j.$$

$$\textit{Symmetric GIE:}\ \max_{\|v_i\|,\|v_j\|\le 1} \textstyle\sum_{i,j} \mathcal{M}_{ij} \langle v_i, v_j\rangle \le c_{\mathtt{g}} \max_{|x_i|,|x_j|\le 1} \sum_{i,j} \mathcal{M}_{ij} x_i x_j \text{ if } \forall \mathcal{M}_{ii} = 0.$$

**Theorem 5.5** (computational complexity of SDP [JKL+20])**.** SDP with variable size $n \times n$ and $m$ constraints is solvable within precision $\varepsilon$ in time[42] $\tilde{O}(\sqrt{n}(mn^2 + m^\omega + n^\omega)\log(1/\varepsilon))$. It is $t_{\mathtt{sdp}}(n) := \tilde{O}(n^{3.5})$ when $m = O(n)$.

[42] $\omega$ is the exponent in the matrix multiplication complexity. The current best is $\omega = 2.472\cdots$ [Str69, AW21].

Coja-Oghlan, Cooper, and Freize [COCF10, AOW15] reduced Max$k$CSP's "average-case" approximation to planted $k$XOR's "strong" refutation: Prove $\mathrm{acc}(\{(x(j),y(j))\}_{j=1}^m) \le 1/2+\varepsilon$ for the parity predicate $y(j) = \bigoplus_{i=1}^k \theta \circ x_i(j)$ and the i.i.d. random constraints $(x(j),y(j)) \in [2n)^k \times \{0,1\}$. Furthermore, the refutation proof on the malicious constraints would yield the planted $k$DNF's PAC learnability. Recently, Abascal, Gurusuwami, and Kothari [Fei07, AOW15, Wit17, AGK21] succeeded in derandomizing $x(j)$ ($y(j)$ is still random) in the following manner.

**Theorem 5.6** (strongly refuting planted $k$XOR [AGK21])**.** The following refutation's proof enjoys a witness computable by SDP in $t_{\mathtt{sdp}}(N^2)\cdot O(n)$ time with confidence $1-\frac{1}{N}$ from any $m=\sum_{i=1}^n |\mathcal{D}_i| = N\sqrt{n}\cdot O(\frac{\log^3 N}{\varepsilon^5})$ data of $\mathcal{D}_i \subset [2N)^2 \times \{0,1\}$ having the i.i.d. random $m$ labels:

$$\textit{Strong refutation:}\quad \max_{\mathbf{z}\in\{0,1\}^N, z'\in\{0,1\}^n} \frac{1}{m}\sum_{i=1}^n \sum_{(x,y)\in\mathcal{D}_i} 1[(\mathbf{z}\circ x_1)\oplus(\mathbf{z}\circ x_2) = z'_i \oplus y] \le 1/2+\varepsilon.$$

**Theorem 5.7** (refuting planted $k$DNF)**.** For $k\ge 2$, the planted $s$-term $k$DNF is refutable by $n^{k/2}\cdot O\big(s^5(k\log n)^3\big)$ data in $t_{\mathtt{sdp}}(n^k)\cdot O(2^k n)$ time.

*Proof.* Given the data $\mathcal{D} = \{(x(j),y(j))\}_{j=1}^m$, our algorithm measures the bias of a term $f = \bigwedge_{i=1}^k \theta\circ \mathbf{x}_i$ in the target planted $k$DNF function through a lens of Fourier coefficients:

$$\begin{aligned}\textit{Bias measurement:}\ \mathrm{bias}_f(\mathcal{D}) &:= \tfrac{1}{m}\textstyle\sum_{j=1}^m (-1)^{y(j)+1} 1[f(x(j))=1]\\ &= \tfrac{1}{2^k m}\textstyle\sum_{j=1}^m (-1)^{y(j)+1}\prod_{i=1}^k \sum_{a_i\in[n)}\big((-1)^{\theta(a_i)+x_i(j)+1}+1\big)1[\lfloor x_i(j)/2\rfloor = a_i]\\ &= \textstyle\sum_{w\subset(k]}\sum_{a\in[n)^w} \hat{\mathcal{M}}_w(a)\prod_{i\in w}(-1)^{\theta(a_i)},\end{aligned}$$

$$\textit{Fourier coefficients:}\ \hat{\mathcal{M}}_w(a) = \tfrac{1}{m}\textstyle\sum_{j:\lfloor x_w(j)/2\rfloor = a}(-1)^{y(j)+1}\prod_{i\in w}(-1)^{x_i(j)+1}.$$

Let $1\le \kappa = \lfloor |v|/2\rfloor$ or $\lfloor |w|/2\rfloor \le \lfloor k/2\rfloor$. Lift and project the bias maximization problem $\max_\theta \mathrm{bias}_f(\mathcal{D})$ over $\theta\in\{0,1\}^n$ to the following QPs (Quadratic Programming) over $\mathbf{z} = (\mathbf{z}_a)_{a\in[n)^\kappa} \in \{-1,1\}^{n^\kappa}$ of $\mathbf{z}_a = \prod_{i=1}^\kappa (-1)^{\theta(a_i)}$ and $z' = (z'_b)_{b\in[n)}$ of $z'_b = (-1)^{\theta(b)} \in \{-1,1\}$ to bound $\mathrm{bias}_f(\mathcal{D}) \le \mathrm{val}(\mathcal{D})$:

$$\mathsf{QP}\ \textit{by lift and project:}\ \mathrm{val}(\mathcal{D}) := \textstyle\sum_{w\subset(k],|w|\in 2\mathbb{Z}} \max_{\mathbf{z}} \mathcal{M}_w(\mathbf{z}) + \sum_{v\subset(k],|v|\in 2\mathbb{Z}+1}\max_{\mathbf{z},z'}\mathcal{M}_v(\mathbf{z},z'),$$

$$\textit{Even}\ \mathsf{QP}\textit{:}\ \mathcal{M}_w(\mathbf{z}) := \textstyle\sum_{a\in[n)^\kappa}\sum_{a'\in[n)^\kappa}\hat{\mathcal{M}}_w(aa')\mathbf{z}_a\mathbf{z}_{a'}\ \text{for}\ |w| = 2\kappa,$$

$$\textit{Odd}\ \mathsf{QP}\textit{:}\ \mathcal{M}_v(\mathbf{z},z') := \textstyle\sum_{a\in[n)^\kappa}\sum_{a'b\in[n)^{\kappa+1}}\hat{\mathcal{M}}_v\big(aa'b\big)\mathbf{z}_a\mathbf{z}_{a'}z'_b\ \text{for}\ |v| = 2\kappa+1.$$

Solve all these maximization problems and distinguish $\mathcal{D}$ by measuring $\mathrm{val}(\mathcal{D})$.

**Completeness:** Take a threshold $\beta \approx 1/(2s)$ as follows. We may assume $|\mathbb{E}[(-1)^Y]| \le \epsilon\beta$. Otherwise, the constant function is already Definition 3.1's refuter to distinguish between $\big|\mathbb{E}[(-1)^Y]\big| \ge \epsilon\beta$ and $\big|\mathbb{E}[(-1)^{Y'}]\big| < \epsilon\beta$ for the random-label data[43] $(X',Y')\sim\mathcal{U}$. It promises the complete data $\mathrm{err}_f(\mathcal{D}) = 0$ to gain an advantage by choosing the heaviest $f$ from the $s$ terms:

$$\textit{Completeness:}\ \mathrm{bias}_f(\mathcal{D}) = \mathsf{Pr}[f(X) = Y = 1] \ge \tfrac{1}{2s} - \tfrac{1}{2}|\mathbb{E}[(-1)^Y]| := \beta.$$

**Soundness:** Take the sample size $m \gg n^\kappa\cdot\sqrt{n'}\cdot\big(k\log n\big)^3/\beta^5$, $N = n^\kappa$ and $n' = n$ (resp. 1) for odd QPs (resp. even QPs). Theorem 5.6 bounds $\mathrm{bias}_f(\mathcal{U})$ with significance $1/N$:

$$\textit{Soundness:}\ \mathrm{bias}_f(\mathcal{U}) \le \mathrm{val}(\mathcal{U}) \le \tfrac{1}{2^k}\Big(\textstyle\sum_w \max_{\mathbf{z}}\mathcal{M}_w(\mathbf{z}) + \sum_v \max_{\mathbf{z},z'}\mathcal{M}_v\big(\mathbf{z},z'\big)\Big) \ll \beta.$$

[43] Chernoff bound parameter $\gamma = \frac{\epsilon\beta/2}{1/2}$ guarantees a confidence level $\mathsf{Pr}[|\mathbb{E}[(-1)^{Y'}]| \ge \epsilon\beta] \le 2e^{-\gamma^2/3\cdot m/2} \ll o(\delta)$.

**Computational complexity:** Theorem 5.6 solves both even and odd QPs and provides a certificate of $\mathrm{val}(\mathcal{U}) \ll \beta$ for the soundness data $\mathcal{U}$. The overall confidence level is $1 - 2^k/N = 1 - o(\delta)$ to succeed in Definition 3.1's refutation of $\mathcal{D}' \in \{\mathcal{D}, \mathcal{U}\}$ only when getting a certificate of $\mathrm{val}(\mathcal{D}') \le \frac{\beta}{2}$ from $m$ data in $t_{\mathtt{sdp}}(n^k) \cdot O(n) \cdot 2^k$ time. $\square$

Theorem 5.7's refutation algorithm can PAC learn the planted DNF under the malicious label $y(j)$ (instead of the random label assumption of Theorem 5.7). Grothendieck inequality can do it by $\tilde{O}(n^{\lceil k/2 \rceil})$ data, so losing a $\sqrt{n}$ factor in the odd $k$ case. Moreover, the refutation's SDP solution is too long to make a PAC hypothesis. Charikar and Wirth [GW95, Meg01, CW04] rounded Theorem 5.4's symmetric GIE solution in over $\mathbb{R}$ to a binary one over $\{-1, 1\}$.

**Theorem 5.8** (rounding symmetric GIE [CW04])**.** Any QP: $\max_x \left| \sum_{i=1}^N \sum_{j=1}^N \mathcal{M}_{ij} x_i x_j \right|$ with $\forall \mathcal{M}_{ii} = 0$ over $x \in \{-1, 1\}^N$ is approximable by ratio[44] $\gamma_{\mathtt{g}} := \Omega(1/\log N)$ in $t_{\mathtt{sdp}}(N^2)$ time.

**Theorem 5.9** (Theorem 1.17)**.** For $k \ge 2$, planted $s$-term $k$DNF is PAC learnable from $n^{\lceil k/2 \rceil} \cdot O(2^k k (ks \log n)^2/\varepsilon^2)$ data in $t_{\mathtt{sdp}}(n^k) \cdot O\big(2^k n \ (ks \log n)^2/\varepsilon\big)$ learning time.

*Proof.* Theorem 5.8 with $N = n^\kappa$ of $\kappa = \lfloor |v|/2 \rfloor$ or $\lfloor |w|/2 \rfloor$ approximates Theorem 5.7's QP by lift-and-project to get even-QP's $\mathcal{M}_w$'s rounded solutions $z(w) = (z_a(w))_{a \in [n]^\kappa}$. Theorem 5.8's QP requires removing the trace $\sum_a \hat{\mathcal{M}}_w(aa) \mathbf{z}_a \mathbf{z}_a = \sum_a \hat{\mathcal{M}}_w(aa)$. Theorem 5.6 divides odd-QP's $\mathcal{M}_v$ into a sum over $b \in [n]$ of $\mathcal{M}_{v,b}(\mathbf{z}) = (-1)^{\theta(b)} \mathcal{M}_{v \setminus \{i\}}(\mathbf{z})$ on $\mathcal{D}_b = \{(x(j), y(j)) \in \mathcal{D} \mid \lfloor x_i(j)/2 \rfloor = b\}$. Theorem 5.8 provides $\mathcal{M}_{v,b}$'s rounded solutions $z(v,b) = (z_a(v,b))_{a \in [n]^\kappa}$, too. These QP's solutions induce a hypothesis function $h : [2n)^k \to \mathbb{Q}$ to bound $\mathrm{bias}_f(\mathcal{D}) \le \mathrm{bias}_h(\mathcal{D}) := \mathbb{E}[(-1)^Y h(X)]$ over the empirical data $(X, Y) \in \{(x(j), y(j))\}_{j=1}^m$:

$$
\begin{aligned}
h_w(x|a) &:= \textstyle\prod_{i \in w} (-1)^{x_i + 1} 1[\lfloor x_w/2 \rfloor = a], \\
h_w(x) &:= \textstyle\sum_{(a \ne a') \in [n)^\kappa \times [n)^\kappa} h_w(x|aa') z_a(w) z_{a'}(w), \quad g_w(x) := \sum_{a \in [n)^\kappa} h_w(x|aa), \\
h_{v,b}(x) &:= 1[\lfloor x_i/2 \rfloor = b] \textstyle\sum_{(a \ne a') \in [n)^\kappa \times [n)^\kappa} h_{v \setminus \{i\}}(x|aa') z_a(v,b) z_{a'}(v,b), \\
g_{v,b}(x) &:= 1[\lfloor x_i/2 \rfloor = b] \textstyle\sum_{a \in [n)^\kappa} h_{v \setminus \{i\}}(x|aa), \\
\textit{Weak hypothesis: } h(x) &:= \tfrac{1}{2^k} \textstyle\sum_{w \subset (k], |w| \in 2\mathbb{Z}} h_w(x) + \tfrac{1}{2^k} \sum_{u \subset (k], |v| \in 2\mathbb{Z}+1} \sum_{b \in [n)} h_{v,b}(x), \\
\textit{Trace: } g(x) &:= \tfrac{1}{2^k} \textstyle\sum_{w \subset (k], |w| \in 2\mathbb{Z}} g_w(x) + \tfrac{1}{2^k} \sum_{v \subset (k], |v| \in 2\mathbb{Z}+1} \sum_{b \in [n)} g_{v,b}(x). \\
\mathrm{bias}_f(\mathcal{D}) - \mathrm{bias}_g(\mathcal{D}) &\le \mathrm{val}(\mathcal{D}) - \mathrm{bias}_g(\mathcal{D}) \\
&= \textstyle\sum_{\substack{w \subset (k], \\ |w| \in 2\mathbb{Z}}} \max_z |\mathcal{M}_w(z)| + \sum_{\substack{v \subset (k], \\ |v| \in 2\mathbb{Z}+1}} \sum_{b \in [n)} \max_z |\mathcal{M}_{v,b}(z)| \\
&\le \tfrac{1}{\gamma_{\mathtt{g}}} \textstyle\sum_{w \subset (k], |w| \in 2\mathbb{Z}} \mathrm{bias}_{h_w}(\mathcal{D}) + \tfrac{1}{\gamma_{\mathtt{g}}} \sum_{v \subset (k], |v| \in 2\mathbb{Z}+1} \sum_{b \in [n)} \mathrm{bias}_{h_{v,b}}(\mathcal{D}) \\
&= \tfrac{1}{\gamma_{\mathtt{g}}} \mathrm{bias}_h(\mathcal{D}) \text{ by Theorem 5.8's ratio } \gamma_{\mathtt{g}} := \tfrac{\Omega(1)}{\log(n^\kappa)}.
\end{aligned}
$$

**Boosting:** Theorem 5.7's completeness proof has shown $\mathrm{bias}_h(\mathcal{D}) \ge \gamma_{\mathtt{g}}(\mathrm{bias}_f(\mathcal{D}) - \mathrm{bias}_g(\mathcal{D})) \ge \gamma_{\mathtt{g}}(\beta - \mathrm{bias}_g(\mathcal{D}))$ for $\beta \approx \frac{1}{2s}$. Theorem 3.3's `SmoothBoost` turns this weak hypothesis $h(x) = h_\nu(x)$ feeding $\mathcal{D} = \mathcal{D}_\nu \sim (P_\nu \circ \mathcal{D})^*$ to an $\varepsilon$-accurate hypothesis in the following manner. First of all, we may assume $|\mathrm{bias}_g(\mathcal{D}_\nu)| \le \epsilon\beta$. Otherwise, `SmoothBoost` can feed $g(x)$ or $-g(x)$ for a weak predictor. Take $\nu_0 \approx \frac{2}{\varepsilon((1-\epsilon)\beta\gamma_{\mathtt{g}})^2}$ and $m \gg n^{\lceil k/2 \rceil} \cdot 2^k k (\frac{ks \log n}{\varepsilon})^2$. It is much larger than the

[44] Charikar and Wirth's $\Omega(1/\log n)$ approximation ratio is best possible [ABE+05, AMMN06, AN06].

logarithm of the hypothesis size $|\{h_\nu\}_\nu| \le \prod_w |\mathrm{rng}(z(w))| \cdot \prod_{u,b} |\mathrm{rng}(z(u,b))| \le 2^{\sum_{\kappa=1}^{\lfloor k/2 \rfloor} \binom{k}{2\kappa} n^\kappa} \cdot 2^{n \sum_{\kappa=1}^{\lfloor k/2 \rfloor} \binom{k}{2\kappa+1} n^\kappa}$, so the final majority vote enjoys UGEB by Chernoff bound parameter $\gamma = 1$:

$$\textit{UGEB: } \prod_{\nu \in [\nu_0]} |\{h_\nu\}_\nu| \cdot \boldsymbol{e}^{-\frac{1}{3}\cdot \varepsilon m} \le 2^{\nu_0 \sum_{\kappa=1}^{\lfloor k/2 \rfloor} \binom{k}{2\kappa} n^\kappa} \cdot 2^{\nu_0 n \sum_{\kappa=1}^{\lfloor k/2 \rfloor} \binom{k}{2\kappa+1} n^\kappa} \cdot \boldsymbol{e}^{-\frac{1}{3}\cdot \varepsilon m} = o(\delta).$$

The overall learning time is $\nu_0(\sum_w t_{\mathtt{sdp}}(n^{|w|}) + \sum_{v,b} t_{\mathtt{sdp}}(n^{|v|})) \le \nu_0 \cdot t_{\mathtt{sdp}}(n^k) \cdot O(2^k n)$. □

**Theorem 5.10** (PAC Learning planted $s$-term $k$DNF with white noise)**.** The planted $s$-term $k$DNF with white $\eta$-noise is PAC learnable from $n^{\lceil \frac{k}{2} \rceil} \cdot O((\frac{ks \log n}{\varepsilon(1-2\eta)})^2)$ data in $t_{\mathtt{sdp}}(n^k) \cdot O(\frac{2^k n}{\varepsilon}(\frac{ks \log n}{1-2\eta})^2)$ time.

*Proof.* The white $\eta$-noise replaces $\beta \approx \frac{1}{2s}$ to $\beta \approx \frac{1-2\eta}{2s}$. It changes Theorem 5.9's boosting's $\nu_0$ in accordance, proving the claimed sample size and learning time complexities. □

Verbeurgt [Ver90] reduced DNF learning to $k$DNF learning under the uniform distribution. Verbeurgt's reduction is extensible to an arbitrary distribution in smoothed analysis.

**Lemma 5.11** (DNF to $k$DNF in the smoothed analysis [Ver90])**.** Learning a planted $s$-term DNF expression $f$ under any $k$-wisely $\rho$-dense flipper $G$ reduces to learning its degree-$k$ sub-formula $\tilde{f}$ obtained by removing all terms longer than $k$:

$$\textit{No FPE:}\quad f(\mathbf{x}) = 0 \Rightarrow \tilde{f}(\mathbf{x}) = 0.$$

$$\textit{Recall:}\quad \mathsf{Pr}_G\big[f(G(x)) = 1, \tilde{f}(G(x)) = 0\big] \le s/(2^{k+1}\rho).$$

*Proof.* If $f(\mathbf{x})$ is false, so are all its terms, hence so is $\tilde{f}(\mathbf{x})$, implying No FPE. The $k$-wise $\rho$-dense shift $G$ bounds the recall of REVIEW3's DNF's term $f_\kappa \cong \bigwedge_{i \in f_\kappa} \mathbf{x}_i \oplus f_{\kappa i}$ as

$$\begin{aligned}
&\mathsf{Pr}_G[f(G(x)) \ne \tilde{f}(G(x))] = \mathsf{Pr}_G[f(G(x)) = 1 \wedge \tilde{f}_\theta(G(x)) = 0] \\
&\le \mathsf{Pr}_G[\exists \kappa \in (s], |f_\kappa| \ge k+1, f_\kappa(G(x)) = 1] \\
&= \mathsf{Pr}_G[\exists \kappa \in (s], |f_\kappa| \ge k+1, \forall i \in f_\kappa,\ G(\lfloor x_i/2 \rfloor) = \theta(\lfloor x_i/2 \rfloor) \oplus x_i \oplus f_{\kappa i} \oplus 1] \le s/(2^k \rho).
\end{aligned}$$ □

**Theorem 5.12** (Theorem 1.18[45])**.** The planted $s$-term DNF is PAC learnable from any $n^{\lceil k/2 \rceil} \cdot O\big((\frac{k^2 s \log n}{\varepsilon})^2\big)$ data in $t_{\mathtt{sdp}}(n^k) \cdot O\big(\frac{2^k n (ks \log n)^2}{\varepsilon}\big)$ time under any $k$-wisely $\frac{s}{2^k \delta}$-dense uniform flipper.

*Proof.* Let Theorem 5.9's proof target only Lemma 5.11's short terms in choosing Theorem 5.7's completeness's $f$ with significant $\mathrm{bias}_f(\mathcal{D})$. Theorem 5.11's recall guarantees $\mathrm{bias}_f(\mathcal{D}) \ge \big(1 - \epsilon - \frac{s}{2^k \rho \gamma}\big)/(2s) = \big(1 - \epsilon - \epsilon\big)/(2s)$ for the assumed density $\rho = \frac{s}{2^k \delta}$ by Markov's inequality parameter $\gamma = \delta/\epsilon$ with significance $O(\gamma)$. Hence, Theorem 5.12 reduces to 5.9. □

**Theorem 5.13** (PAC learning planted $s$-term DNF with white noise)**.** The planted $s$-term DNF with white $\eta$-noise is PAC learnable from any $n^{\lceil (k+1)/2 \rceil} \cdot O((\frac{ks \log n}{\varepsilon(1-2\eta)})^2)$ data in $t_{\mathtt{sdp}}(n^k) \cdot O(\frac{2^k n}{\varepsilon}(\frac{ks \log n}{1-2\eta})^2)$ learning time under any $k$-wisely $\frac{s}{2^k \delta(1-2\eta)}$-dense uniform flipper.

*Proof.* By reducing to Theorem 5.12 in the same way as Theorem 5.10 to 5.9. □

[45] Set $k = \log \frac{2s}{\delta}$ and $\frac{1}{\delta} = O(1)$. Take Lemma 2.9's $\frac{1}{2}$-dense $dn$-bit flipper of cardinality $O(2^k k \log(dn))$.

# 6 Smoothed Complexity of Agnostic Learning AND functions

This section translates the so-far obtained PAC theorems in smoothed analysis to the corresponding agnostic ones, i.e., PAC 1–4 to AGN 1–4. Let us begin from AGN1 to review the current best agnostic algorithm of learning planted $\mathrm{AND}_d$. It owes to Kalai, Klivans, Mansour, and Servedio [KOS04, KKMS08, BOW10], adopting $\ell_1$-norm regression to $\Omega(\sqrt{d})$-degree approximation of $\mathrm{AND}_d = \{f(\mathbf{x}) := \bigwedge_{i\in f} \mathbf{x}_i \oplus f_i \mid f \subset (d], f_i \in \{0,1\}\}$ [Pat92, NS94, TT99, KKMS08].

**Theorem 6.1** (polynomial degree of AND.)**.** The $\mathrm{AND}_d$ functions enjoy a low-degree point-wise approximation $\forall \mathbf{x} \in \{0,1\}^n, \left|(-1)^{\bigwedge_{i=1}^d \mathbf{x}_i} - f_d(\mathbf{x})\right| \le \varepsilon$ by $f_d(\mathbf{x}) \in \mathbb{Q}[\mathbf{x}]$ of degree $O(d^{1/2}\log\frac{1}{\varepsilon})$.

**Theorem 6.2** ([KKMS08])**.** The planted $\mathrm{AND}_d$ is agnostically learnable from $\eta$-noisy data in deterministic $n^{O(d^{1/2}\log(n/(1-2\eta)))}$ time.

*Proof.* Apply Theorem 6.1 to $\mathrm{err}(\mathcal{D}) \le \eta$ of the target $\bigwedge_{i=1}^d \mathbf{x}_i$ function, giving a rational polynomial $f_d$ of degree $\mathrm{D} = O(d^{1/2}\log\frac{1}{\varepsilon})$ to bound $\frac{1}{m}\sum_{j=1}^m |f_d(\theta \circ x(j)) - (-1)^{y(j)}| \le \eta + \varepsilon$. Theorem 5.1 can solve this LP with $n' = \sum_{k=0}^{\mathrm{D}} n^k \binom{d}{k}$ variables in $t = O\big(m^{1.5} n' \log(n'/\varepsilon)\big)$ time by $O(\log(n'/\varepsilon))$-bit precision. The $\ell_1$-norm regression chooses a hypothesis $h = (\mathrm{sgn}(f_d(\theta \circ x) - t) + 1)/2$ for an appropriate threshold $t \in [-1,1]$ to become a weak empirical learner achieving $\mathrm{err}_h(\mathcal{D}) \le \eta + \varepsilon + o(\varepsilon)$ [KKMS08]. Sufficiently many examples $m = O\big(\varepsilon^2/\eta \cdot n' \log(n'/\varepsilon)\big)$ turn this weak learner of description length $O(n' \log(n'/\varepsilon))$ to an actual one $P(y \ne h(x)) \le \eta + \varepsilon$ by Chernoff bound parameter $\gamma = \varepsilon/\eta$ with significance:

$$\textit{UGEB:}\quad 2^{O(n'\log(n'/\varepsilon))} \cdot \big(\boldsymbol{e}^{-\gamma^2/(2+\gamma)\cdot\eta|\mathcal{D}|} \cdot 1[\eta > \varepsilon] + \boldsymbol{e}^{-\gamma/3\cdot\eta|\mathcal{D}|} \cdot 1[0 < \eta \le \varepsilon]\big) = o(\delta'). \qquad \square$$

## 6.1 Agnostic Learning versus Refutation

Theorem 3.4's reduction from refutation to PAC learning is extensible to agnostic one by cooperating with agnostic boosting [BDLM01, KS05, KK09, Fel10].

**Theorem 6.3** (agnostic boosting [Fel10])**.** If $\eta'$-noisy $\mathcal{F}$ is $(1/2-\alpha)$-learnable with significance $\delta'$ for $\eta \le \forall\eta' \le 1/2 - \varepsilon$, then it is $(\eta + 2\varepsilon)$-learnable with significance $O(\delta'/\alpha^2)$ under the same variate distribution $P(x)$ by calling the $(1/2-\alpha)$-learner for $c_{6.3}/\alpha^2$ times. If the $(1/2-\alpha)$-learner runs in $t$ time, then the $(\eta' + 2\varepsilon)$-learner in $O(t/\alpha^2 + 1/\varepsilon^2)$ time.

**Theorem 6.4** (noisy refutation to agnostic learning)**.** Let $\delta_{6.4} := \frac{\delta}{m^4 \log^3 m \log\frac{m}{\delta}}$. If $\eta'$-noisy $\mathcal{F}$ is refutable for any $\eta \le \eta' \le 1/2 - \varepsilon$ with significance $O(\delta_{6.4})$ from $m$ data in $t$ time, $\eta$-noisy $\mathcal{F}$ is agnostic learnable from $m^2 \cdot O(\log\frac{m}{\delta}\log\frac{1}{\delta})$ data in $m^4\, t \cdot O\big(\log^3 m \log\frac{m}{\delta}\big) + O(\frac{1}{\varepsilon^2})$ learning time.

*Proof.* Theorem 3.4's weak learning can provide Theorem 6.3's agnostic booster a weak-learner performing well under the same variate (but possibly different covariate) distribution with the unknown target. For $\alpha \approx \frac{1}{m}$, $\nu_0 = c_{6.3}/\alpha^2$, $\kappa_0 \gg (\frac{\log m}{\alpha})^2 \log\frac{\nu_0 \log m}{\delta}$, $\tilde{m} \gg (\frac{1}{\alpha})^2 \log\frac{\nu_0}{\delta}$ and $\tilde{m}' \gg \frac{\tilde{m}}{\varepsilon}\log\frac{1}{\delta}$, Theorem 3.4's boosting on the agnostic booster spends $\tilde{m}'$ data, runs in $\nu_0\kappa_0 \log m \cdot O(t/\alpha^2) + O(1/\varepsilon^2)$ time, and succeed with significance level $\nu_0\kappa_0 \log m \cdot O(\delta_{6.4}) = O(\delta)$. $\square$

**Theorem 6.5** (noisy refutation to agnostic learning in smoothed analysis)**.** If $\eta'$-noisy $\mathcal{F}$ is refutable for any $\eta \le \eta' \le 1/2 - \varepsilon$ with significance $O(\delta_{6.4}^2/\delta)$, $\eta$-noisy $\mathcal{F}$ is agnostic learnable under any shift in the same way as Theorem 6.4.

*Proof.* It reduces to Theorem 6.4, as Theorem 3.6 to 3.4. $\square$

## 6.2 Proof Theoretic Hardness of Agnostic Learning AND functions

Section 4 relied on Theorems 4.3 and 4.6 of PAC learning hardness. Similarly, the current section will depend on Theorem 6.8 below of agnostic learning hardness. It is an extension of Theorem 4.6 for weak refutation to a strong one.

**Definition 6.6** (bounded expansion). A CSP instance $\mathcal{G} = (\mathcal{I} \sqcup \mathcal{J}, \mathcal{E})$ is $r$-bounded $(\textsc{d}, t)$-expanding if the number of edge-induced $(d, t)$-expanding subgraphs are bounded by $r$:

$$\underset{(\textsc{d},t)\text{-expansion}}{\textit{r-bounded}}: \left|\left\{(u \sqcup v, w) \mid \begin{matrix} \emptyset \neq w \subset \mathcal{E}, u \sqcup v = \mathcal{E}[w], (\forall j, j \in u \Rightarrow |w[j]| \geq t), \\ |u| \leq \textsc{d}, |v| \leq |w| - (t/2 - \zeta)|u| - (t-1)/2 \end{matrix}\right\}\right| \leq r.$$

**Lemma 6.7** (RCSP is bounded expanding [KMOW17]). For $3 \leq t = \Omega(k)$ and $\textsc{d} \ll \frac{|\mathcal{I}|}{k\Delta^{2/(t-2-2\zeta)}}$, any $k$CSP instance $\mathcal{G}$ of the uniform random $\mathcal{E}$ and density $\Delta \gg 1$ must be

$$\textit{r-bounded } (\textsc{d}, t)\textit{-expanding: } \mathsf{Pr}\big[\ \mathcal{G} \text{ is } |\mathcal{I}|^{\frac{1}{2}+\zeta}\Delta\text{-bounded } (\textsc{d}, t)\text{-expanding}\big] \geq 1 - \epsilon'^k.$$

*Proof.* Theorem 4.10's analysis can count the expanding subgraphs:

$$\begin{aligned}
&\textstyle\sum_{\emptyset \neq w \subset \mathcal{E}} \mathsf{Pr}_{\mathcal{E}}\big[v = \mathcal{J}[w], u = \mathcal{I}[w], |v| \leq \textsc{d}, |u| \leq k|v|, |u| + (\frac{t}{2} - \zeta)|v| - \frac{t-1}{2} \leq |w| \leq k|v|\big] \\
&< \textstyle\sum_{|v|,|u|,|w|} \big(\boldsymbol{e}^{2+\frac{|u|}{|v|}} (\frac{|u||w|}{|v|^2}) (\frac{|u|}{|\mathcal{I}|})^{\frac{t}{2}-\zeta-1} \Delta\big)^{|v|} (\frac{|u|}{|\mathcal{I}|})^{-\frac{t-1}{2}} \\
&< \textstyle\sum_{|v|,|u|,|w|} \big(k^2 \boldsymbol{e}^{2+k} (\frac{|u|}{|\mathcal{I}|})^{\frac{t}{2}-\zeta-1} \Delta\big)^{|v|-1} \cdot k^2 \boldsymbol{e}^{2+k} (\frac{|u|}{|\mathcal{I}|})^{-\zeta-\frac{1}{2}} \Delta \\
&\overset{\star}{<} \textstyle\sum_{|v|,|w|} \sum_{|v| \geq 2} \boldsymbol{e}^{2+k} k^2 |v|^2 |\mathcal{I}|^{\frac{1}{2}+\zeta} \Delta \cdot \big(k^2 \boldsymbol{e}^{2+k} \big(k\textsc{d}\Delta^{\frac{2}{t-2-2\zeta}} / |\mathcal{I}|\big)^{\frac{t}{2}-\zeta-1}\big)^{|v|-1} \\
&< 4k^6 \boldsymbol{e}^{4+2k} |\mathcal{I}|^{\frac{1}{2}+\zeta} \Delta \big(k\textsc{d}\Delta^{\frac{2}{t-2-2\zeta}} / |\mathcal{I}|\big)^{\frac{t}{2}-\zeta-1} = \epsilon'^k |\mathcal{I}|^{\frac{1}{2}+\zeta} \Delta.
\end{aligned}$$

The right-hand side of $\overset{\star}{<}$ does not count $|v| = 1$ since the case $|v| + (\frac{t}{2} - \zeta)|v| - \frac{t-1}{2} - |w| = |v| + (t/2 - \zeta) \cdot 1 - \frac{t-1}{2} - |v| = 1/2 - \zeta > 0$ never happens in Definition 6.6's expansion. Markov's inequality parameter $\gamma = \epsilon'^k$ on this expectation derives Lemma 6.7's bounded expansion. $\square$

**Theorem 6.8** (SoS hardness of bounded-expanding CSP's refutation [KMOW17]). For any $r$-bounded $(\textsc{d}, t)$-expanding CSP instance $\mathcal{G}$ with $\forall j \in \mathcal{J}, |\mathcal{I}[j]| \leq \zeta\textsc{d}$, and any integers $2 \leq t-1 \leq t'$, there exists $\mathcal{J}' \subset \mathcal{J}$ with $|\mathcal{J}'| \approx |\mathcal{J}|$ such that for any $t'$-uniform variable $X_j \in \{0,1\}^{\mathcal{I}[j]}$,

$$\underset{\textit{on bounded expansion}}{\text{SoS } \textit{hardness}}: \mathbf{deg}_{\text{SoS}}\big[\text{unsat}_{\mathcal{G}}(\mathbf{x}) > \tfrac{1}{|\mathcal{J}|} \textstyle\sum_{j \in \mathcal{J}'} \mathsf{Pr}_{X_j}[X_j \notin \mathcal{S}_j] + \frac{|\mathcal{J}| - |\mathcal{J}'|}{|\mathcal{J}|}\big] \geq \frac{\zeta\textsc{d}}{3}.$$

**Theorem 6.9** (Theorem 1.19). For $2 \leq d \leq \log(1/\varepsilon) - O(1)$ and $0 \leq \eta \leq 1/2 - O(\varepsilon)$, agnostic learning the $\eta$-noisy canonical planted AND class $\{\bigwedge_{i=1}^d \theta \circ x_i \mid \theta \in \{0,1\}^n\}$ under the uniform distribution demands either sample size $\Omega\big(n^{(1-\epsilon)d/2}\big)$ or SoS degree $\Omega(n^\epsilon)$.

*Proof.* Remake Theorem 4.15's proof to derive a contradiction to Theorem 6.8's SoS hardness from the assumption $\mathbf{deg}_{\text{SoS}}[\text{err}_\theta(\mathcal{D}) > \eta] < \zeta\textsc{d}/3 := n^\epsilon$. Let us learn a joint-distribution $P(x, f(x))$ having the uniform variate $P(x) = 1/(2n)^d$ and the white-$\tilde{\eta}$-noisy covariate:

$$\textit{White noisy constraint sampler: } \tilde{\eta}P(x) \otimes |x, 0\rangle + \tilde{\eta}P(x) \otimes |x, 1\rangle + (1 - 2\tilde{\eta})P(x, f(x))|x, f(x)\rangle$$
$$\text{of } f(x) = \textstyle\bigwedge_{i=1}^d \theta \circ x_i \text{ and } \tilde{\eta} := \eta + (c + \epsilon)\varepsilon \leq \frac{1}{2} - \Omega(\varepsilon).$$

This mixture draws a data $(X_j, Y_j) \sim \mathcal{D}$ by first throwing the $(\tilde{\eta} : \tilde{\eta} : 1 - 2\tilde{\eta})$-biased dice $B_j \in \{0, 1, 2\}$ and then sampling the example from $P(x) \otimes |x, 0\rangle$, $P(x) \otimes |x, 1\rangle$ and $P(x, f(x))|x, f(x)\rangle$

when $B_j = 0,1,2$, respectively. Lemma 3.2's UGEB has shown by $\Pr[\mathrm{err}_\theta(\mathcal{D}) \le \eta + c\varepsilon] < |\{0,1\}^n|\boldsymbol{e}^{-\gamma^2/2\cdot\tilde{\eta}m} < o(\delta)$, so Definition 2.1 obliges the SoS learner to prove $\mathrm{err}_\theta(\mathcal{D}) > \eta$. Similarly, the hitting sets $\mathcal{J}_b := \{j \mid B_j = b\}$ must have cardinality $\forall b, |\frac{|\mathcal{J}_b|}{m} - \tilde{\eta}| \le \epsilon\varepsilon\tilde{\eta}$ with significance $2\boldsymbol{e}^{-\frac{\gamma^2}{3}\cdot\tilde{\eta}m} = o(\delta)$ by Chernoff bounds of $\gamma = \epsilon\varepsilon/\tilde{\eta}$. The $\mathcal{J}_b$ with $b = 0,1$ induce CSP instances $\mathcal{G}_b = \big(\mathcal{I} \sqcup \mathcal{J}_b, \mathcal{E}, \mathcal{S}_b\big)$ of the uniformity $t = d$:

*Factor Graph*: $\mathcal{I} = [n)$ and $\mathcal{E} = \{(j, \lfloor x_i(j)/2 \rfloor) \mid i \in (d], j \in \mathcal{J}_b\}$, $\mathcal{J}_b' \subset \mathcal{J}_b$ for $|\mathcal{J}_b'| \ge |\mathcal{J}_b| - n^{-\frac{1}{2}+\zeta}|\mathcal{J}_b|$.

*Solution spaces*: $\mathcal{S}_{1,j} = \{(x_i(j) \oplus 1)_{i=1}^d\}$ and $\mathcal{S}_{0,j} = \{0,1\}^{\mathcal{I}[j]} - \mathcal{S}_{1,j}$.

*Unif -ormity*: Take $(d-1)$-uniform variable $X_{b,j}$ with $\Pr[X_{0,j} \in \mathcal{S}_{0,j}] = 1$ and $\Pr[X_{1,j} \in \mathcal{S}_{1,j}] = \frac{1}{2^{d-1}}$.

These CSP instances $\mathcal{G}_b$ appeal $\frac{|\mathcal{I}|}{k(\frac{|\mathcal{J}_b|}{n})^{2/t-2-2\zeta}} \ge \frac{n}{kn^{((d-2)(1-\epsilon)/2-\epsilon)(2/d-2-2\zeta)}} \gg \mathrm{D}$ to Lemma 6.7's SoS hardness of bounded expansion, yielding a contradiction:

$$\forall b \in \{0,1\}, \zeta\mathrm{D}/3 \le \mathbf{deg}_{\mathrm{SoS}}\big[\mathrm{unsat}_{\mathcal{G}_b}(\mathbf{x}) > \tfrac{1}{|\mathcal{J}_b|}\textstyle\sum_{j\in\mathcal{J}_b'} \Pr_{X_{b,j}}[X_{b,j} \notin \mathcal{S}_{b,j}] + \frac{|\mathcal{J}_b| - |\mathcal{J}_b'|}{|\mathcal{J}_b|}\big] \Rightarrow$$

$$\zeta\mathrm{D}/3 \le \mathbf{deg}_{\mathrm{SoS}}\big[\mathrm{err}_\theta(\mathcal{D}) = \textstyle\sum_{b=0}^1 \frac{|\mathcal{J}_b|}{m}\mathrm{unsat}_{\mathcal{G}_b}(\mathbf{x}) > \sum_{b=0}^1\big(\frac{\sum_{j\in\mathcal{J}_b'}\Pr[X_{b,j}\notin\mathcal{S}_{b,j}]}{m} + \frac{|\mathcal{J}_b| - |\mathcal{J}_b'|}{m}\big)\big]$$

$$\le \mathbf{deg}_{\mathrm{SoS}}\big[\mathrm{err}_\theta(\mathcal{D}) > \tilde{\eta}(1 - \tfrac{1}{2^{d-1}}) + 2\tilde{\eta}n^{-\frac{1}{2}+\zeta} + 2\epsilon\varepsilon\tilde{\eta}\big] \overset{\star}{\le} \mathbf{deg}_{\mathrm{SoS}}\big[\mathrm{err}_\theta(\mathcal{D}) > \eta\big] < \tfrac{\zeta\mathrm{D}}{3}.$$

$$\overset{\star}{\le}: d \le \log\tfrac{1}{\varepsilon} - O(1) \Rightarrow \tilde{\eta}(1 - \tfrac{1}{2^{d-1}}) + 2\tilde{\eta}n^{-1/2+\zeta} + 2\epsilon\varepsilon\tilde{\eta} < \eta.$$

□

**Theorem 6.10** (Theorem 1.20)**.** For $d \ge 2$ and $0 \le \eta \le 1/2 - O(\varepsilon)$, agnostic learning the $\eta$-noisy canonical parity function class $\{\bigoplus_{i=1}^d \theta \circ x_i \mid \theta \in \{0,1\}^n\}$ under the uniform distribution demands either sample size $\Omega\big(n^{(1-\epsilon)d/2}\big)$ or SoS degree $\Omega(n^\epsilon)$.

*Proof.* As in Theorem 6.9, take CSP instances $\mathcal{G}_b = (\mathcal{I} \sqcup \mathcal{J}_b, \mathcal{E}, \mathcal{S}_b)$ of the $(d-1)$-uniform random variable $X_{b,j} \in \mathcal{S}_{b,j} = \{x \in \{0,1\}^d \mid \bigoplus_{i=1}^d x_i = b \oplus \bigoplus_{i=1}^d x_i(j)\}$, yielding

$$\zeta\mathrm{D}/3 \le \mathbf{deg}_{\mathrm{SoS}}\big[\mathrm{err}_\theta(\mathcal{D}) > 2\tilde{\eta}n^{-1/2+\zeta} + 2\epsilon\varepsilon\tilde{\eta}\big] \overset{\star}{\le} \mathbf{deg}_{\mathrm{SoS}}\big[\mathrm{err}_\theta(\mathcal{D}) > \eta\big] < \zeta\mathrm{D}/3,$$

$$\text{where } \overset{\star}{\le} \text{ by } \eta(1 + (c+\epsilon)\varepsilon/\eta)(2n^{-1/2+\zeta} + 2\epsilon\varepsilon) \ll \eta.$$

□

**Theorem 6.11** (Theorem 1.22[46] for AND function under flippers)**.** For $0 < c < 1$, let $t_{6.11} := \frac{cd}{1+\log\boldsymbol{e}+1.725\log((1+\log\boldsymbol{e})/c)} \ge 3$ (i.e., $t_{6.11} = t_{4.13}(k\leftarrow d)$). For $0 < c < 1$, $2 \le d \le \frac{1}{c}\log\frac{1}{\varepsilon} - O(1)$ and $\Omega(1) \le \eta \le 1/2 - O(\varepsilon)$, agnostic learning the $\eta$-noisy planted AND class $\{\bigwedge_{i=1}^d \theta \circ x_i \mid \theta \in \{0,1\}^n\}$ under the uniform distribution shifted by any flipper $G$ of $\mathrm{H}_\infty(G) = (1-c)d$ requires either sample size $\Omega\big(n^{(1-\epsilon)t_{6.11}/2}\big)$ or SoS proof of degree $\Omega(n^\epsilon)$.

*Proof.* Adjust Theorem 6.9's argument to take the shifted solution space as in Theorem 4.14, i.e.,

*Solution spaces*: $\mathcal{S}_{1,j} = \big\{\big(g(\lfloor\frac{x_i(j)}{2}\rfloor) \oplus x_i(j) \oplus 1\big)_{i=1}^d \mid \Pr[G = g] > 0\big\}$ and $\mathcal{S}_{0,j} = \{0,1\}^{\mathcal{I}[j]} \backslash \mathcal{S}_{0,j}$,

*Unif -ormity*: $\Pr[X_{0,j} \in \mathcal{S}_{0,j}] = 1$ and $\Pr[X_{1,j} \in \mathcal{S}_{1,j}] \ge \max\{\Pr[X \in \mathcal{S}_{1,j}] \mid X \text{ is } t\text{-uniform}\} \ge 1 - \frac{1}{2^{cd}}$.

[46] By $d = \frac{1}{c}\log(1/\varepsilon) - O(1)$ and replacing $\frac{1}{c} - 1 \mapsto c$.

Let $t = t_{6.11}$ of Theorem 4.14. Since $|\mathcal{S}_{1,j}| = 2^{(1-c)d}$ and Lemma 4.12's cosets disjointly cover $\mathcal{S}_{1,j}$, Lemma 4.12 presents $t$-uniform random variables $X_{b,j}$, deriving a contradiction to Theorem 6.8:

$$\tfrac{\zeta_{\rm D}}{3} \le \mathbf{deg}_{\rm SoS}\big[\mathrm{err}_{\mathcal{D}}(\mathbf{x}) > \tilde{\eta}(1 - \tfrac{1}{2^{cd}}) + 2\tilde{\eta}n^{-\frac{1}{2}+\zeta} + 2\epsilon\varepsilon\tilde{\eta}\big] \le \mathbf{deg}_{\rm SoS}\big[\mathrm{err}_{\mathcal{D}}(\mathbf{x}) > \eta\big] < \tfrac{\zeta_{\rm D}}{3}. \qquad \square$$

**Theorem 6.12** (Theorem 1.22)**.** For $0 < c < 1$, $2 \le d \le \frac{1}{c}\log(1/\varepsilon) - O(1)$, and $\Omega(1) \le \eta \le 1/2 - O(\varepsilon)$, agnostic learning $\eta$-noisy, agnostic learning the $\eta$-noisy canonical planted AND $\{\bigwedge_{i=1}^{d} \theta \circ x_i \mid \theta \in \{0,1\}^{dn}\}$ under the uniform distribution perturbed by any shift of $\mathrm{H}_\infty(G) = (1-c)d$ demands either sample size $\Omega\big((\frac{n}{4^{(1-c)d}})^{(1-\epsilon)t_{6.11}/2}\big)$ or SoS proof of degree $\Omega\big((\frac{n}{4^{(1-c)d}})^{\epsilon}\big)$.

*Proof.* A reduction to Theorem 6.11 by the same adversary reducing Theorem 4.17 to 4.15. □

## 6.3 Agnostic Learning AND functions

**Theorem 6.13** (refuting $\eta$-noisy $k$AND)**.** For $k \ge 2$, the planted $k$AND is refutable from any $\eta$-noisy $n^{k/2} \cdot O(\frac{(k\log n)^3}{(1-2\eta)^5})$ data in $t_{\mathtt{sdp}}(n^k) \cdot O(2^k n)$ time.

*Proof.* Changing $\beta \approx \frac{1}{2s}$ to $(1-2\eta)/2$ in Theorem 5.7's completeness analysis proves Theorem 6.13 since the target AND function $f$ is a single-term planted $k$DNF satisfying $1 - 2\eta = \mathbb{E}[(-1)^Y] + 2\mathrm{bias}_f(\mathcal{D})$ over the data distribution $(X, Y) \sim \mathcal{D}$, where $\eta := \mathsf{Pr}[Y \ne f(X)]$. □

**Theorem 6.14** (refuting $\eta$-noisy planted $k$XOR)**.** For $k \ge 2$, the planted $k$XOR is refutable from any $\eta$-noisy $n^{k/2} \cdot O(\frac{(k\log n)^3}{(1-2\eta)^5})$ data in $t_{\mathtt{sdp}}(n^{\lfloor\frac{k-1}{2}\rfloor}) \cdot O(n)$ time.

*Proof.* Adapt Theorem 5.7's bias measurement to the canonical $k$XOR function $f = \bigoplus_{i=1}^{k} \theta \circ \mathbf{x}_i$:

*Bias measurement:* $\mathrm{bias}_f(\mathcal{D}) := \frac{1}{m}\sum_{j=1}^{m}(-1)^{y(j)+1}1[f(x(j)) = 1] = \sum_{a\in[n]^k}\hat{M}(a)\prod_{i=1}^{k}(-1)^{\theta(a_i)}$,

*Fourier coefficients:* $\hat{M}(a) = \frac{1}{m}\sum_{j:\lfloor x_i(j)/2\rfloor_{i=1}^{k}=a}(-1)^{y(j)+1}\prod_{i=1}^{k}(-1)^{x_i(j)+1}$.

Theorem 5.7's computational complexity analysis brings Theorem 6.14's running time since the above bias measurement fixes $w = (k]$ rather than running over $w \subset (k]$. □

**Theorem 6.15** (Theorem 1.21 for $k$AND)**.** For $k \ge 2$, the planted $k$AND is agnostically learnable from any $\eta$-noisy $n^{\lceil k/2\rceil} \cdot O\big((\frac{k\log n}{\varepsilon(1-2\eta)})^2\big)$ data in $t_{\mathtt{sdp}}(n^k) \cdot O\big(2^k n(\frac{k\log n}{1-2\eta})^2\big)$ learning time.

*Proof.* Build Theorem 5.9's weak hypothesis from Theorem 6.13's refuter and apply Theorem 6.3's agnostic boosting. For $\beta \approx (1-2\eta)/2$, $\nu_0 = \frac{c_{6.3}}{(\beta\gamma_{\rm g}/2)^2}$, $m \gg n^{\lceil k/2\rceil} \cdot O\big((\frac{k\log n}{\varepsilon(1-2\eta)})^2\big)$, Theorem 5.9's UGEB holds, and Theorem 6.3's agnostic boosting finishes within $\nu_0 \cdot t_{\mathtt{sdp}}(n^k) \cdot O(2^k n)$ time. □

**Theorem 6.16** (Theorem 1.21 for planted $k$XOR)**.** For $k \ge 2$, the planted $k$XOR is agnostically learnable from any $\eta$-noisy $n^{\lceil k/2\rceil} \cdot O\big((\frac{k\log n}{\varepsilon(1-2\eta)})^2\big)$ data in $t_{\mathtt{sdp}}(n^k) \cdot O\big(n(\frac{k\log n}{1-2\eta})^2\big)$ learning time.

*Proof.* Apply Theorem 6.3's agnostic boosting to Theorem 6.14's refutation as Theorem 6.15's argument did to Theorem 6.13's one. □

**Theorem 6.17** (Theorem 1.21)**.** For $k \ge 2$, the planted $k$JUNTA is agnostically learnable from any $\eta$-noisy $n^{\lceil k/2\rceil} \cdot O\big((\frac{2^k k\log n}{\varepsilon(1-2\eta)})^2\big)$ data in $t_{\mathtt{sdp}}(n^k) \cdot O\big(2^k n(\frac{2^k k\log n}{1-2\eta})^2\big)$ time.

*Proof.* Adjust Theorem 6.15's one to target an exclusive OR of at most $2^k$ terms, one of which must have the completeness's threshold $\beta \approx \frac{1-2\eta}{2\cdot 2^k}$, deducing the claimed complexities. □

**Theorem 6.18** (Theorem 1.23[47])**.** For $k \geq 2$, the planted AND is agnostically learnable from any $\eta$-noisy $n^{\lceil k/2 \rceil} \cdot O\big((\frac{k \log n}{\varepsilon(1-2\eta)})^2\big)$ data in $t_{\mathtt{sdp}}(n^k) \cdot O\big(n(\frac{k \log n}{1-2\eta})^2\big)$ learning time under any $k$-wisely $O(\frac{1}{2^k \delta(1-2\eta)})$-dense uniform flipper.

*Proof.* A reduction to Theorem 6.15 as 5.12 to 5.7, since Lemma 5.11's recall guarantees $\mathrm{bias}_f(\mathcal{D}) \geq \beta - \frac{1}{2^{k+1}\rho\gamma} \approx \beta$ for $\beta \approx (1-2\eta)/2$ and $\rho \gg \frac{1}{2^k \beta\delta}$ by Markov's inequality parameter $\gamma = \delta/\epsilon$. $\square$

## 6.4 Approximate promise-MaxCSP.

This section translates Section 6.3's Theorems to those for approximating promise-MaxCSP.

**Definition 6.19** (approximation of promise-MaxCSP)**.** The $(\beta_{\mathtt{cmp}}, \beta_{\mathtt{snd}})$-gap (or $\triangle$-gap, $\triangle := \beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}}$) approximation of promise-MaxCSP assumes either $\mathrm{acc}(P) = \beta_{\mathtt{cmp}} > \beta_{\mathtt{snd}} = \mathrm{acc}(P')$ or $P = P'$ must hold of the two unknown samplers $P$ and $P'$ of MaxCSP's constraints. It asks to discern which is the case by observing the i.i.d. outcomes $\mathcal{D} \sim P^m$ and $\mathcal{D}' \sim P'^m$ as follows:

*Verifiable Completeness:* Show a witness to verify[48] $|\mathrm{acc}(\mathcal{D}) - \mathrm{acc}(\mathcal{D}')| \leq \frac{3}{4}\triangle \to P = P'$.

It attaches proof-theoretic refutation demand to the previous models. It covers Feige's (resp. Barak, Kindler, and Steurer's) hypothesis [Fei02] (resp. [BKS13]) by taking $P_{\mathtt{cmp}}$ and $P_{\mathtt{snd}}$ over $k$CNF's (resp. $k$JUNTA's) satisfiable versus random constraints and Alekhnovich's hypothesis [Ale11] by LPN's random ones of Hamming-distance noise $k$ versus $k+1$. Moreover, it involves distinguishment problems between "malicious" $\omega(m)$ constraints $\tilde{\mathcal{D}}$ and $\tilde{\mathcal{D}}'$ with a slight difference $\mathrm{acc}(\tilde{\mathcal{D}}) - \mathrm{acc}(\tilde{\mathcal{D}}') = \epsilon$ by taking empirical distributions to draw $m$ i.i.d. constraints $\mathcal{D}$ and $\mathcal{D}'$ from $\tilde{\mathcal{D}}$ and $\tilde{\mathcal{D}}'$, respectively.

**Theorem 6.20** (Theorem 1.24)**.** Any gap approximation of the promise-Max$k$SAT under a marginally uniform distribution requires either $\Omega(n^{\frac{1-\epsilon}{2}k})$ constraints or $\Omega(n^\epsilon)$ SoS-degree.

*Proof.* A reduction to Theorem 6.9 by filtering $m \gg n^{\frac{1-\epsilon}{2}k}$ positive data $\mathcal{G}_\kappa$ for $\kappa \in \{\mathtt{cmp}, \mathtt{snd}\}$:

$$\underset{\textit{(discard negative examples)}}{\textit{Positive constraint sampler}}: \quad \eta_\kappa P(x)|x,1\rangle + (1-\eta_\kappa)P(x, f(x))|x, f(x)\rangle$$

$$\text{of } f = \bigvee_{i=1}^{k} \theta \circ x_i \text{ and } \frac{1-1/2^k}{\eta_\kappa + (1-1/2^k)(1-\eta_\kappa)} = \beta_{\mathtt{cmp}} \cdot 1[\kappa = \mathtt{cmp}] + \beta_{\mathtt{snd}} \cdot 1[\kappa = \mathtt{snd}].$$

This mixture joint-distribution has the claimed accuracies $\beta_\kappa$ since $P(f(x) = 0) = 1/2^k$. The $2m$ outcomes emitted from a mixture source $\mathcal{G} := \mathcal{G}_{\mathtt{cmp}} \otimes |+1\rangle \sqcup \mathcal{G}_{\mathtt{snd}} \otimes |-1\rangle$ must take the weighted accuracy gap $\mathrm{acc}(\mathcal{G}_{\mathtt{cmp}}) - \mathrm{acc}(\mathcal{G}_{\mathtt{snd}}) \leq (1+\epsilon)(\beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}})$ with significance $e^{-\frac{\gamma^2}{2+\gamma} \cdot (\beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}})m} < o(\delta)$ by Chernoff bound of $\gamma = \frac{\epsilon\triangle}{\beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}}}$. Definition 6.19's verifiable completeness obliges to prove $\mathbf{deg}_{\mathrm{SoS}}\big[|\mathrm{acc}(\mathcal{G}_{\mathtt{cmp}}) - \mathrm{acc}(\mathcal{G}_{\mathtt{snd}})| = \mathrm{err}_\theta(\mathcal{G}) \geq \frac{3\triangle}{4}\big] \leq \zeta_{\mathrm{D}}/3 := n^\epsilon$, a contradiction against Theorem 6.9's CSP instances $\mathcal{G}_\kappa := (\mathcal{I} \sqcup \mathcal{J}_\kappa, \mathcal{E}, \mathcal{S})$. Here $\mathcal{J}_\kappa = \{(x_\kappa(j), 1)\}_j$ collects the only $m$ positive examples emitted from the positive constraint sampler, and $\mathcal{S}_{\kappa,j} = \{0,1\}^{\mathcal{I}[j]} - \{(x_{\kappa,i}(j) \bmod 2)_{i=1}^k\}$. Take $(k-1)$-uniform random variables $X_{\kappa,j} \in \mathcal{S}_{\kappa,j}$.

$$\begin{aligned}
\zeta_{\mathrm{D}}/3 &\leq \mathbf{deg}_{\mathrm{SoS}}\Big[\mathrm{err}_\theta(\mathcal{G}) \geq \textstyle\sum_{\kappa \in \{\mathtt{cmp},\mathtt{snd}\}} s_\kappa \frac{|\mathcal{J}_\kappa|}{2m}(1 - \mathrm{unsat}_{\mathcal{G}_\kappa}(\mathbf{x}))\Big] \quad (\text{where } s_\kappa := (-1)^{1[\kappa = \mathtt{cmp}]}) \\
&\leq \mathbf{deg}_{\mathrm{SoS}}\Big[\mathrm{err}_\theta(\mathcal{G}) \geq \tfrac{1+\epsilon}{2}(\beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}}) + \textstyle\sum_\kappa \Big(\frac{s_\kappa \sum_{j \in \mathcal{J}'_\kappa} \Pr[X_{\kappa,j} \notin \mathcal{S}_{\kappa,j}]}{2m} + \frac{|\mathcal{J}_\kappa| - |\mathcal{J}'_\kappa|}{2m}\Big)\Big] \\
&\leq \mathbf{deg}_{\mathrm{SoS}}\Big[\mathrm{err}_\theta(\mathcal{G}) \geq \tfrac{1+\epsilon}{2}\triangle + n^{-\frac{1}{2}+\zeta}\Big] \leq \mathbf{deg}_{\mathrm{SoS}}\Big[\mathrm{err}_\theta(\mathcal{G}) \geq \tfrac{3\triangle}{4}\Big] < \tfrac{\zeta_{\mathrm{D}}}{3}.
\end{aligned}$$

$\square$

[47] Set $k = \log \frac{\epsilon}{(1-2\eta)\delta}$. Take Theorem 2.9's $\frac{1}{2}$-dense $dn$-bit flipper of cardinality $O(k 2^k \log(dn))$.

[48] The verifiable-completeness threshold $\frac{3}{4}\triangle$ could be any $c\triangle$ between $1/2 < c < 1$.

**Theorem 6.21** (Theorem 1.25)**.** Any gap approximation of the promise-Max$k$XOR under a marginally uniform distribution demands either $\Omega\big(n^{(1-\epsilon)k/2}\big)$ constsrints or $\Omega(n^{\epsilon})$ SoS-degree.

*Proof.* A reduction to Theorem 6.10 by letting Max$k$XOR approximate $\mathcal{P} = \mathcal{P}_{\mathtt{cmp}} \otimes |+1\rangle \sqcup \mathcal{P}_{\mathtt{snd}} \otimes |-1\rangle$ drawn from the following mixture and deriving a contradiction as in Theorem 6.20:

$$\underset{\textit{(discard negative examples)}}{\textit{Positive-constraint sampler}}: \quad \eta_\kappa P(x) \otimes |x, 0\rangle + (1-\eta_\kappa) P(x, f(x)) |x, f(x)\rangle,$$

$$\text{where } \frac{1/2}{\eta_\kappa + (1/2)(1-\eta_\kappa)} = \beta_{\mathtt{c}} \cdot 1[\kappa = \mathtt{cmp}] + \beta_{\mathtt{snd}} \cdot 1[\kappa = \mathtt{snd}].$$ ☐

**Theorem 6.22** (Theorem 1.27[49])**.** At $e(G) = (1-c)k$ for $0 < c < 1$, any gap approximation of the promise-Max$k$SAT under a marginal uniform distribution shifted by any flipper $G$ requires either $\Omega\big(n^{(1-\epsilon)t_{4.13}/2}\big)$ constraints or $\Omega\big(n^{\epsilon}\big)$ SoS-degree .

*Proof.* A reduction to Theorem 6.20, like Theorem 6.11 to 6.9. ☐

**Theorem 6.23** (Theorem 1.26 for promise-Max$k$SAT)**.** The promise-Max$k$SAT is $\triangle$-gap approximable by any $n^{k/2} \cdot O\big((k \log n)^3/\triangle^5\big)$ constraints in $t_{\mathtt{sdp}}\ (n^k) \cdot O(2^k n)$ time.

*Proof.* Feed the difference of the i.i.d. random outcomes to Theorem 5.7's completeness proof in the following manner, instead of Definition 3.1's random-label dataset $\mathcal{U}$. Draw $\mathcal{D} \sim P^m(x, y)$ and $\mathcal{D}' \sim P'^m(x, y)$ and measure their bias difference $|\mathrm{bias}_f(P, P')| = |\frac{1}{m} \sum_{(x,y)\in\mathcal{D}'} 1[f(x) = 1] - \frac{1}{m} \sum_{(x',y')\in\mathcal{D}'} 1[f(x') = 1]|$. Theorem 5.7's bias measurement on the random outcomes from a mixture $X \otimes |Y\rangle \sim \frac{1}{2} P(x) \otimes |+1\rangle + \frac{1}{2} P'(x) \otimes |-1\rangle$ distinguishes between $|\mathrm{bias}_f(P_{\mathtt{cmp}}, P_{\mathtt{snd}})| \approx\triangle$ (or larger) versus $|\mathrm{bias}_f(P, P)| \approx 0$. The former produces Theorem 6.13's completeness proof by replacing $1-2\eta$ therein with $\triangle= \beta_{\mathtt{cmp}} - \beta_{\mathtt{snd}}$, and the latter Theorem 5.7's soundness one. ☐

**Theorem 6.24** (Theorem 1.26 for promise-Max$k$XOR)**.** The promise-Max$k$XOR is $\triangle$-gap approximable by any $n^{k/2} \cdot O\big((k \log n)^3/\triangle^5\big)$ constraints in $t_{\mathtt{sdp}}(n^k) \cdot O(n)$ time.

*Proof.* Adjust Theorem 6.23's argument to Theorem 6.14's bias measurement. ☐

**Theorem 6.25** (Theorem 1.26)**.** The promise-Max$k$CSP is $\triangle$-gap approximable by any $n^{k/2} \cdot O\big((k \log n)^3\ 2^{5k}/\triangle^5\big)$ constraints in $t_{\mathtt{sdp}}(n^k) \cdot O(2^k n)$ time.

*Proof.* Replace Theorem 6.23's completeness's threshold to $\beta \approx\triangle\ /2^k$ instead of $\triangle$ as we did in Theorem 6.17's one since the target predicate is an exclusive OR of (at most) $2^k$ terms. ☐

**Theorem 6.26** (approximating promise-MaxSAT)**.** The promise-MaxSAT is $\triangle$-gap approximable by any $n^{k/2} \cdot O\big((k \log n)^3/\triangle^5\big)$ constraints in $t_{\mathtt{sdp}}(n^k) \cdot O(2^k n)$ time under any $k$-wisely $O(\frac{1}{2^k \delta \triangle})$-dense uniform flipper.

*Proof.* A reduction to Theorem 6.23 as 6.18 to 6.15 via 5.11. ☐

[49] $k = (c+1) \log \frac{1}{2\varepsilon}$ implies $\beta_{\mathtt{snd}} \geq 1 - 1/2^k \Leftrightarrow \beta_{\mathtt{snd}} \geq 1 - (2\varepsilon)^{c+1}$.

# 7 Inverting Planted Functions in Smoothed Analysis

We have so far confirmed that efficiently PAC learning the planted $k$DNF took $\Omega(n^{(1-\epsilon)k/2})$ data necessary for the uniform distribution, and $\tilde{\Omega}(n^{\lceil k/2\rceil})$ data sufficient for any distribution. It was so for agnostic learning the planted $k$JUNTA, approximating Max$k$SAT, and refuting $k$SAT, too. However, previous works have already broken this $n^{k/2}$ barrier under the uniform distribution [CM01, Vio05, MST06, BQ12, ABR16, LV17], e.g., inverting $k$CSP in $O(n^{k/3})$ time by analyzing the *correlation* $\mathbb{E}[(-1)^{\sum_{i\in w} X_i+Y} \mid \forall i \in w, \lfloor X_i/2\rfloor = a_i]$ on a *location* (or *place*) $(w,a) \in \binom{d}{k} \times [n]^k$ under the uniform random $X \sim [2n)^k$ [App16]. Our smoothed analysis will work under any distribution to make the correlation analysis invert the monotone DNF in only $\tilde{O}(n)$ time. Moreover, the correlation analysis on larger min-entropy can invert even non-monotone functions approximated by low-degree polynomials over $\mathbb{F}_p$.

## 7.1 Inverting Monotone DNF

The correlation analysis of the uniform random data can learn monotone DNF [KLV94, SM00, Ser04, Fel12], monotone Boolean functions [BT96, OS07], monotone JUNTA [MOS04], halfspaces [TTV09, OS11, DDFS14], and LPN [Val15]. We will extend them to any pairwise dense data distribution to learn monotone DNF via approximate inclusion-exclusion [LN90, KLS96, TT99].

**Definition 7.1** (approximating inclusion-exclusion of monotone DNF)**.** For a monotone DNF expression $f = \bigvee_{\kappa\in f} f_\kappa$ of $f_\kappa := \bigwedge_{i\in f_\kappa} \mathbf{x}_i$, write $f_{\vee w} := \bigvee_{\kappa\in w} f_\kappa$ and $f_{\wedge w} := \bigwedge_{\kappa\in w} f_\kappa$. Inclusion-exclusion expands logical expressions $f \equiv b, f' \equiv b'$ of DNF $f, f'$, and $b, b_f, b' \in \{0,1\}$ as

*Inclusion-Exclusion (IE)*: $\mathbf{ie}_c(f \equiv b) := \sum_{|w|=b}^{c-1} \sum_{w\subset f} (-1)^{|w|+b} f_{\wedge w}$.

*Doubled Inclusion-Exclusion*: $\mathbf{ie}_c(f \equiv b, f' \equiv b') := \sum_{\substack{|w\cup w'|\le c-1,\\ b\le|w|,b'\le|y'|}} \sum_{\substack{w\subset f,\\ w'\subset f'}} (-1)^{|w|+|w'|+b+b'} (f_{\wedge w} \wedge f'_{\wedge w'})$.

*(Inclusion-Exclusion on average*: $\mu_c(f \equiv b) := \sum_{|w|=b}^{c-1} \sum_{w\subset f} (-1)^{|w|+b} 2^{-|f_{\wedge w}|}$.

*Doubled IE on average*: $\mu_c(f \equiv b, f' \equiv b') := \sum_{\substack{|w\cup w'|\le c-1,\\ b\le|w|,b'\le|w'|}} \sum_{\substack{w\subset f,\\ w'\subset f'}} (-1)^{|w|+|w'|+b+b'} 2^{-|f_{\wedge w}\cup f'_{\wedge w'}|}$.

The IE of tripled DNF formulas $f \equiv b, f' \equiv b', f'' \equiv b''$ develops in the same manner. Observe that if $x \in \{0,1\}^d$ satisfies $c' - 1 \ge c$ terms of $f$, its contribution to $\mathbf{ie}_{c'}(f \equiv b) - \mathbf{ie}_c(f \equiv b)$ is $|\sum_{\kappa=c}^{c'-1}(-1)^\kappa\binom{c'-1}{\kappa}| = |\sum_{\kappa=0}^{c-1}(-1)^\kappa\binom{c'-1}{\kappa}| = \binom{c'-2}{c-1}$, which we call the *truncated coefficient* of $x$. Its contribution to $\mathbf{ie}_{c'}(f \equiv b, f' \equiv b') - \mathbf{ie}_c(f \equiv b, f' \equiv b')$ is the same amount $\binom{c'-2}{c-1}$, once the $x$ satisfies $c' - 1$ terms of $f \cup f'$.

**Definition 7.2** ($\rho$-spread)**.** A random vector $X \sim \prod_{i=1}^d \mathcal{S}_i$ is $\rho$-spread with significance $\delta$ if

$$\text{$\rho$-spread: } \forall \mathcal{S} \subset \prod_{i=1}^d \mathcal{S}_i, \forall i, |\{x_i \in \mathcal{S}_i \mid x \in \mathcal{S}\}|/|\mathcal{S}_i| \le \rho \Rightarrow \mathsf{Pr}[\forall i, X_i \notin \mathcal{S} \cap \mathcal{S}_i] \ge 1-\delta.$$

**Lemma 7.3.** Any 1-wisely $\rho$-dense random vector is $\frac{\delta\rho}{d(1-\delta+\delta^2/2)}$-spread with significance $\delta$.

*Proof.* Suppose $\forall i, \frac{|\mathcal{S}\cap\mathcal{S}_i|}{|\mathcal{S}_i|} \le \frac{\delta\rho}{d(1-\delta+\delta^2/2)}$. Lemma 2.5's LLL at $\alpha_i = \delta/d$ applies to

$$\mathsf{Pr}[X_i \in \mathcal{S} \cap \mathcal{S}_i] \le \frac{|\{x_i \in \mathcal{S}_i \mid x \in \mathcal{S}\}|}{|\mathcal{S}_i|\rho} \le \frac{\delta}{d(1-\delta+\delta^2/2)} < p_i(1-p_i)^{d-1}$$

of dependent $n$ events $X_i \in \mathcal{S} \cap \mathcal{S}_i$, deriving $\mathsf{Pr}[\forall i, X_i \notin \mathcal{S} \cap \mathcal{S}_i] \ge (1-p_i)^d > 1-\delta$. □

**Definition 7.4** ($(\alpha,\beta)$-inversion)**.** We say that a randomized algorithm $\mathcal{A}$ $(\alpha,\beta)$-inverts $\{f\}$ planting $\theta \in \{0,1\}^{dn}$ on data $(X,Y) \sim \mathcal{D}$ if it can retrieve the hidden parameter $\theta_i(a)$ of any $\alpha$-heavy $\beta$-correlated place $(i,a) \in (d] \times [n)$ as follows, where $\delta_{\mathtt{inv}} := \delta/d$.

$$\begin{aligned}
\textit{Correlation:}\quad & \mathbf{corr}_i(\mathcal{D}) = \mathbf{corr}_i(X,Y) := \mathbb{E}[(-1)^{X_i+Y}] - \mathbb{E}[(-1)^{X_i}]\mathbb{E}[(-1)^{Y}]. \\
\textit{Invariance:}\quad & 0 < \exists \mu_i < 1, \forall (i,a), \big||\mathbf{corr}_i(\mathcal{D})| - \mu_i\big| \ll \beta. \\
(\alpha,\beta)\textit{-inversion:}\quad & \mathsf{Pr}_{\mathcal{D},\mathcal{A}}\begin{bmatrix}\mathsf{Pr}[\lfloor X_i/2\rfloor = a] \geq \alpha/n \wedge |\mathbf{corr}_i(\mathcal{D})| \geq \beta \\ \Rightarrow \mathcal{A}(\mathcal{D},i,a) = \theta_i(a)\end{bmatrix} \geq 1 - O(\delta_{\mathtt{inv}}).
\end{aligned}$$

---

**Algorithm 1** $(\alpha,\beta)$-inversion of monotone DNF

---

Given data $(X,Y) \sim \mathcal{D}$ and a query $(i,a)$ (an index-attribute pair to invert).

1: Filter $\mathcal{D}$ to $(X_{i,a}, Y_{i,a}) \sim \mathcal{D}_{i,a} := \{(x,y) \in \mathcal{D} \mid \lfloor \frac{x_i}{2} \rfloor = a\}$. If $\frac{|D_{i,a}|}{|\mathcal{D}|} < \frac{\alpha}{n}$, then return `?`.

2: Compute the data's output bias $\mathbb{E}[(-1)^{Y_{i,a}}]$ (or use the already computed value).

3: Compute $\mathbf{corr}_i(X_{i,a}, Y_{i,a})$ and return zero if it is $\geq \beta$, one if $\leq -\beta$, and `?` otherwise.

---

**Theorem 7.5** (inverting canonical DNF)**.** Let $\beta_{7.5} := \max\big(\frac{\binom{s}{c}}{2^{ck-1}\delta_{\mathtt{inv}}}, (\frac{ks}{\alpha\delta_{\mathtt{inv}}^4 \rho n})^{1/2}\big)\binom{s-2}{c-1}$. Suppose $\beta_{7.5} \leq \beta \ll 1$. Algorithm 1 can $(\alpha,\beta)$-invert the canonical planted DNF $\{\bigvee_{\kappa=1}^{s}\bigwedge_{i=1}^{k} \theta_{i+\kappa k} \circ x_{i+\kappa k} \mid \theta \in \{0,1\}^{ks}\}$ from any noise-free $n \cdot O\big(\binom{s-2}{c-1}^2 / (\alpha\beta^2\delta_{\mathtt{inv}}^3)\big)$ data with pairwisely $\rho$-dense attributes under any $\epsilon\beta$-away $2ck$-independent flipper over $\{0,1\}^{ksn}$.

*Proof.* Definition 7.1's IE calculates Definition 7.4's $\mathbf{corr}_i(\mathcal{D}_{i,a})$ and exhibits Algorithm 1's inversion performance. For the target canonical DNF expression $f = \bigvee_{\kappa=1}^{s}\bigwedge_{i\in f_\kappa} \mathbf{x}_i$, write $f_{-\kappa} := \bigvee_{\kappa' \in (s] \setminus \{\kappa\}} f_{\kappa'}$ and $f_{\kappa - i} = \bigwedge_{i' \in f_\kappa - \{i\}} \mathbf{x}_{i'}$. They express the relevance and irrelevance of $\mathbf{x}_i$ to $f$ by $f_{\mathtt{rel},i} := f_{\kappa-i} \equiv 1 \wedge f_{-\kappa} \equiv 0$, $f_{\mathtt{ir0}} := f_{\kappa-i} \equiv f_{-\kappa} \equiv 0$, and $f_{\mathtt{ir1}} := f_{-\kappa} \equiv 1$ as follows. Let $\mu(f \equiv 1) := 2^{-|f|}$, $\mu(f \equiv 0) := 1 - 2^{-|f|}$, $\mu_c := \mu_c(f_{\mathtt{ir0}}) - \mu_c(f_{\mathtt{ir1}})$, and $\mu_i := \mu_c(f_{\mathtt{rel},i})$.

$$\begin{aligned}
\textit{Relevance, irrelevance}:\quad & f_{\mathtt{rel},i} \equiv 0 \Rightarrow f \equiv \mathbf{x}_i, \quad f_{\mathtt{ir0}} \equiv 0 \Rightarrow f \equiv 0, \ \text{ and } f_{\mathtt{ir1}} \equiv 1 \Rightarrow f \equiv 1. \\
\textit{rel+ir0+ir1 decomposition}:\quad & 1[f_{\mathtt{rel},i}] + 1[f_{\mathtt{ir0}}] + 1[f_{\mathtt{ir1}}] = 1. \\
\textit{Averages:}\quad & \mu_c(f_{\mathtt{rel},i}) = \mu(f_{\kappa-i} \equiv 1)\mu_{c-1}(f_{-\kappa} \equiv 0),\ \mu_c(f_{\mathtt{ir0}}) = \mu(f_{\kappa-i} \equiv 0)\mu_c(f_{-\kappa} \equiv 0), \\
& \mu_c(f_{\mathtt{ir1}}) = \mu_{c-1}(f_{-\kappa} \equiv 1), \text{ and } \mathbb{E}[(-1)^{Y_{i,a}}] \approx \mu_c.
\end{aligned}$$

**Claim:** If $G$ is perfectly $2ck$-independent and $\mathcal{D}_{i,a} = \{\big(G(x(j)), y(j)\big)\}_j$ satisfies the disjointness, the other four assertions must hold with high confidence.

$$\begin{aligned}
\textit{Disjointness}:\quad & \forall (j \neq j'), \forall (i' \neq i), \lfloor x_i(j)/2 \rfloor = a \wedge \lfloor x_{i'}(j')/2 \rfloor \neq \lfloor x_{i'}(j')/2 \rfloor. \\
\textit{Low degree:}\quad & (\forall w, |w| \geq c), f_{-\kappa,\wedge w}\big(\theta \circ G(x(j))\big) \approx 0. \\
\textit{Relevance:}\quad & \mathsf{Pr}_G[f_{-\kappa}\big(\theta \circ G(x(j))\big) = b] \approx \mu_c(f_{-\kappa} \equiv b). \\
\textit{Correlation on shift}:\quad & \mathbb{E}_G[\mathbf{corr}_i\big(G(x(j)), y(j)\big)] \approx (-1)^{\theta_i(a)}\mu_i. \\
\textit{Correlation on data}:\quad & \mathbb{E}_{J,G}[\mathbf{corr}_i(G(x(J)), y(J))] \approx (-1)^{\theta_i(a)}\mu_i.
\end{aligned}$$

**Low-degree:** Since every term contains $k$ (or $k-1$ in $f_{\kappa-i}$) variables in a disjoint manner, the $ck$-independence of $G$ over the first $ck-1$ variables $\mathbf{x}_{i'}$ of $(f_{\kappa-i} \vee f_{-\kappa})_{\wedge w}$ evaluates

$$\mathsf{Pr}[\neg\text{low-deg}(\theta \circ G(x(j)))] \leq \mathsf{Pr}\big[(\exists w, |w| \geq c), \forall i' \in f_{-\kappa,\wedge w}, \theta_{i'} \circ G(x_{i'}(j)) = 1\big] \leq \binom{s}{c}/2^{ck-1}.$$

**Relevance:** The inclusion-exclusion formula of $f_{-\kappa} \equiv b$ under low-deg approximates

$$\begin{aligned}
&\Big|\mathsf{Pr}[f_{-\kappa}\big(\theta \circ G(x(j))\big) = b] - \mu_c(f_{-\kappa} \equiv b)\Big| \\
&= \Big|\mathsf{Pr}[f_{-\kappa}\big(\theta \circ G(x(j))\big) = b] - \mathbb{E}[\mathbf{ie}_c(f_{-\kappa} \equiv b)(\theta \circ G(x(j)))]\Big| \le \binom{s-2}{c-1}\mathsf{Pr}[\neg\text{low-deg}\big(G(x(j))\big)].
\end{aligned}$$

The first equality stands on the $ck$-independence of $G$. The second one bounds the truncation error of $\mathbf{ie}_c$ at $x' = \theta \circ G(x(j))$ by $\mathsf{Pr}[\neg\text{low-deg}(x')]$ times the truncated coefficient $\binom{s-2}{c-1}$ of $x'$.

**Correlation on shift:** The relevance on the rel+ir0+ir1 cover yields

$$\begin{aligned}
&\Big|\mathbb{E}[(-1)^{G(x_i(j))+f\left(\theta\circ G(x(j))\right)}] - \mathbb{E}[(-1)^{G(x_i(j))}]\mu - (-1)^{\theta_i(a)}\mu_i\Big| = \\
&\left|\begin{aligned}&(-1)^{\theta_i(a)}\big(\mathsf{Pr}[f_{\mathtt{rel},i}(\theta \circ G(x(j)))] - \mu_c(f_{\mathtt{rel},i})\big)\\ &+ \mathbb{E}[(-1)^{G(x_i(j))}]\textstyle\sum_{b=0}^{1}(-1)^b\big(\mathsf{Pr}[f_{\mathtt{ir}b}(\theta \circ G(x(j)))] - \mu_c(f_{\mathtt{ir}b})\big)\end{aligned}\right| \le 3\binom{s-2}{c-1}\mathsf{Pr}[\neg\text{low-deg}\big(G(x(j))\big)].
\end{aligned}$$

**Correlation on data:** Averaging the correlation over the shifted data $G(x(J))$ has a bound

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\mathbb{E}_J[(-1)^{G(x_i(J))+f(\theta\circ G(x(J)))}] - \mathbb{E}_J[(-1)^{G(x_i(J))}]\mu - (-1)^{\theta_i(a)}\mu_i - \bar{Z}\Big|\Big] \\
&\quad \le 3\binom{s-2}{c-1}\mathbb{E}_J\big[\mathsf{Pr}[\neg\text{low-deg}(G(x(J)))]\big] \\
&\text{for } \bar{Z} := \mathbb{E}_J\big[(-1)^{\theta_i(a)}\bar{Z}_{\mathtt{rel},i}(G(x(J))) + (-1)^{G(x_i(J))}\textstyle\sum_{b=0}^{1}(-1)^b\bar{Z}_{\mathtt{ir}b}(G(x(J)))\big], \\
&\bar{Z}_\kappa(x) = \hat{Z}_\kappa(x) - \mu_c(f_\kappa), \text{ and } \hat{Z}_\kappa(x) = \mathbf{ie}_c(f_\kappa)(\theta \circ x) \text{ of } \kappa \in \{\mathtt{rel}, \mathtt{ir0}, \mathtt{ir1}\}.
\end{aligned}$$

They have the zero-mean $\mathbb{E}[\bar{Z}_\kappa(G(x(J)))] = 0$. The $2k$-independence of $G$ under the disjointness makes them mutually perpendicular as $\mathbb{E}[\bar{Z}_\kappa(G(x))\bar{Z}_{\kappa'}(G(x'))] = 0$ between $x \neq x'$. Thus,

$$\begin{aligned}
\mathbb{E}[\bar{Z}^2] &= \tfrac{1}{m^2}\textstyle\sum_{j=1}^{m}\mathbb{E}\Big[\Big((-1)^{\theta_i(a)}\bar{Z}_{\mathtt{rel},i}\big(G(x(j))\big) + (-1)^{G(x_i(j))}\sum_b(-1)^b\bar{Z}_{\mathtt{ir}b}\big(G(x(j))\big)\Big)^2\Big] \\
&\le \tfrac{1}{m^2}\textstyle\sum_j \mathbb{E}\Big[\Big(\sum_\kappa \bar{Z}_\kappa\big(G(x(j))\big)\Big)^2\Big] \le \tfrac{1}{m^2}\sum_j\Big(\sum_\kappa\binom{s-2}{c-1}\Big)^2 \le 9\binom{s-2}{c-1}^2/m,
\end{aligned}$$

deriving $\mathsf{Pr}\big[|\bar{Z}| \ge \frac{3\binom{s-2}{c-1}}{(\delta_{\mathtt{inv}}m)^{1/2}}\big] \le \delta_{\mathtt{inv}}$ by Chebyshev's inequality parameter $\gamma = \frac{1}{\delta_{\mathtt{inv}}}$. A similar analysis gives $\mathbb{E}[(-1)^{y(J)}] \approx \mu$ by bounding $\mathbb{E}_G\big[\big|\mathbb{E}_J[(-1)^{y(J)}] - \mu - \bar{Z}\big|\big] \le$

$$\left|\begin{aligned}&\textstyle\sum_{b=0}^{1}(-1)^b\big(\mathsf{Pr}[\theta_i \circ \mathbf{x}_i(G(x)) = b, f_{\mathtt{rel},i} \circ \theta(G(x))] - \frac{1}{2}\mu_c(f_{\mathtt{rel},i})\big)\\ &+ \textstyle\sum_{b=0}^{1}(-1)^b\big(\mathsf{Pr}[f_{\mathtt{ir}b}(\theta \circ G(x))] - \mu_c(f_{\mathtt{ir}b})\big)\end{aligned}\right| \le 4\binom{s-2}{c-1}\mathsf{Pr}[\neg\text{low-deg}(G(x))].$$

**Disjoint:** The pairwisely $\rho$-dense attributes give $\mathsf{Pr}[\lfloor x_{i'}(J)/2\rfloor \mid \lfloor x_i(J)/2\rfloor = a] \le \frac{1/(\rho n^2)}{\alpha/n}$ under $\mathsf{Pr}[\lfloor x_i(J)/2\rfloor = a] \ge \alpha/n$. Lemma 7.3 distributes the attributes $\{\lfloor x_{i'J}/2\rfloor, i' \in [ks] - \{i\} \mid \lfloor x_i(J)/2\rfloor = a\}$ as $\frac{\alpha\delta_{\mathtt{inv}}\rho}{ks(1-\delta_{\mathtt{inv}}+\delta_{\mathtt{inv}}^2/2)}$-spread with significance $\delta_{\mathtt{inv}}$. It extracts a sub-data $\mathcal{D}_{i,a}$ of size $m \approx \binom{s-2}{c-1}^2/(\beta^2\delta_{\mathtt{inv}}^3) \le \frac{\alpha\delta_{\mathtt{inv}}\rho n}{ks(1-\delta_{\mathtt{inv}}+\delta_{\mathtt{inv}}^2/2)}$ $(\because (\frac{ks}{\alpha\delta_{\mathtt{inv}}^4\rho n})^{1/2}\binom{s-2}{c-1} \le \beta)$ from the given data $\mathcal{D}$ by Chebyshev's inequality parameter $\gamma = \frac{|\mathcal{D}|-|\mathcal{D}_{i,a}|}{\delta_{\mathtt{inv}}|\mathcal{D}|} \gg 1$ with significance $\boldsymbol{e}^{-\frac{\gamma^2}{2+\gamma}\delta_{\mathtt{inv}}|\mathcal{D}|} \ll o(\delta_{\mathtt{inv}})$.

**Invariance:** The corr-on-data measures on $|\mathcal{D}_{i,a}| = m \geq \frac{\binom{s-2}{c-1}^2}{\beta^2\delta_{\mathtt{inv}}^3}$ and $\binom{s-2}{c-1}\binom{s}{c}/2^{ck-1} \leq \beta\delta_{\mathtt{inv}}$:

$$\textit{Average invar-iance}: \mathbb{E}\Big[\Big|\mathbf{corr}_i(G(x(J)), y(J)) - (-1)^{\theta(a)}\mu_i\Big|\Big] < 7\binom{s-2}{c-1}\binom{s}{c}/2^{ck-1} + 7\binom{s-2}{c-1}/(\delta_{\mathtt{inv}}m)^{1/2} < 14\beta\delta_{\mathtt{inv}}.$$

It guarantees Definition 7.4's invariance $\left|\mathbf{corr}_i(G(x(J)), y(J)) - (-1)^{\theta_i(a)}\mu_i\right| \leq \epsilon\beta$ by Markov's inequality of $\gamma = O(\delta_{\mathtt{inv}})$ with significance $O(\delta_{\mathtt{inv}})$. Although the actual flipper $\tilde{G}$ is $\epsilon\beta$-away from the perfectly independent $G$, The **Claim**'s assertions (so the invariance as well) still hold for the $\tilde{G}$ by adding an extra statistical deviation. For example, the low-degree is a local argument at a location $v$ of the first $2ck-1$ variables of $(f_{\kappa-i} \vee f_{-\kappa})_{\wedge w}$ to bound

$$\mathsf{Pr}[\neg\text{low-deg}(\theta \circ \tilde{G}(x(j)))] \leq \binom{s}{c}/2^{ck-1} + d_{\mathtt{st}}\big(G(x_w(j)), \tilde{G}_w(x_w(j))\big) \leq 2\epsilon\beta.$$

$(\alpha, \beta)$**-inversion:** The invariance detects $\theta_i(a)$ for the following reasons. First, the correlation's average must be significant as $|\mu_i| \geq (1-\epsilon)\beta$. Otherwise, the invariance falsifies $|\mathbf{corr}_i(X_{i,a}, Y_{i,a})| \geq \beta$. Secondly, $\mu_i > 0$ by the relevance $\mu_i/\mu(f_{\kappa-i} \equiv 1) = \mu_{c-1}(f_{-\kappa} \equiv 0) \approx \mathsf{Pr}_G[f_{-\kappa}\big(\theta \circ G(x(j))\big) = 0] \geq 0$. Algorithm 1 must succeed in inverting $|\mathcal{D}| \geq (1+\epsilon)\frac{|\mathcal{D}_{i,a}|}{\alpha/n}$ data, since then $|\mathcal{D}_{i,a}| \geq \binom{s-2}{c-1}^2/(\beta^2\delta_{\mathtt{inv}}^3)$ with CB's significance $\boldsymbol{e}^{\epsilon^2/2\cdot\alpha|\mathcal{D}|} \leq o(\delta_{\mathtt{inv}})$ under $\mathsf{Pr}[\lfloor X_{i,a}/2\rfloor = a] \geq \alpha/n$. □

**Definition 7.6** (expanding DNF)**.** REVIEW3's DNF expression $f$ is $c$-wisely $k$-expanding if

$$\textit{c-wisely k-expanding: } \forall v \subset f, |v| \leq c \Rightarrow \left|\bigcup\nolimits_{\kappa \in v} f_\kappa\right| \geq k|v|.$$

**Theorem 7.7** (inverting monotone DNF)**.** For $\beta_{7.5} \leq \beta \ll 1$, Algorithm 1 $(\alpha, \beta)$-inverts a monotone variable of any planted $s$-term $k$DNF with $c$-wise $k$-expansion from any $n \cdot O\big(\frac{\binom{s-2}{c-1}^2}{\alpha\beta^2\delta_{\mathtt{inv}}^3}\big)$ data with pairwisely $\rho$-dense attributes under any $\epsilon\beta$-away $2ck$-independent flipper over $\{0,1\}^{dn}$.

*Proof.* It is similar to Theorem 7.5's one which has relied solely on the $c$-wise $k$-expansion and the monotonicity of a queried variable. This time, divide $f = f_{\vee(t]} \vee f_{\vee((s]-(t])}$ to those terms $j \in (t]$ containing $i$ and the others not holding it, and let $f_{\vee w-i} := \bigvee_{\kappa \in w} f_{\kappa-i}$ for $w \subset (t]$.

$$\textit{Relevance, irrelevance}: \; f_{\mathtt{rel},i} := f_{\vee(t]-i} \equiv 1 \wedge f_{\vee(t,s]} \equiv 0,\; f_{\mathtt{ir0}} := f_{[t]-i} \equiv f_{\vee(t,s]} \equiv 0 \text{ and } f_{\mathtt{ir1}} := f_{\vee(t,s]} \equiv 1.$$

$$\textit{rel+ir0+ir1 cover}: \; 1[f_{\mathtt{rel},i}] + 1[f_{\mathtt{ir0}}] + 1[f_{\mathtt{ir1}}] = 1.$$

$$\textit{Averages}: \; \mu_c(f_{\mathtt{rel},i}) = \mu_c(f_{\vee(t]-i} \equiv 1, f_{(t,s]} \equiv 0),\; \mu_c(f_{\mathtt{ir0}}) = \mu_c(f_{\vee(t]-i} \equiv 0, f_{\vee(t,s]} \equiv 0) \text{ and } \mu_c(f_{\mathtt{ir1}}) = \mu_{c-1}(f_{\vee(t,s]} \equiv 1).$$

Notice that the target DNF's terms $f_\kappa$ may be too long, mutually overlapping, and even contracting to each other, but the $\mathbf{ie}_c$ adjusts as follows to preserve Theorem 7.5's proof:

$$\textit{IE}: \; \mathbf{ie}_c(f \equiv b) := \sum\nolimits_{|w|=b}^{c-1} \sum\nolimits_{w \subset f, |f_{\wedge w}| < ck} (-1)^{|w|+b} f_{\wedge w}.$$

$$\textit{IE on average}: \; \mu_c(f \equiv b) := \sum\nolimits_{|w|=b}^{c-1} \sum\nolimits_{w \subset f, |f_{\wedge w}| < ck,} (-1)^{|w|+b} 2^{-|f_{\wedge w}|}.$$

$$\textit{Doubeled IE}: \; \mathbf{ie}_c(f \equiv b, f' \equiv b') := \sum\nolimits_{w \subset f, w' \subset f', |f_{\wedge w} \cup f'_{\wedge w'}| < ck} (-1)^{|w \cup w'|+b+b'} f_{\wedge w} \cup f'_{\wedge w'}.$$

□

**Algorithm 2** Properly PAC learning monotone DNF

Input a dataset $(X, Y) \sim \mathcal{D}$, initialize $h_0 \equiv 0$ and $\nu = 1$, and repeat 1–6.

1: *Stopping criterion.* Finish and output $h_{\nu-1}$ if $\mathsf{Pr}[h_{\nu-1}(X) = 0, Y = 1] < \varepsilon$.

2: *Variable selection.* Guess a set of variables $f_\nu \subset (d]$. Let $\theta_\nu = \mathcal{S}_\nu = \emptyset$.

3: *Correlation retrieval.* For each $(i, a) \in f_\nu \times [n) - \mathcal{S}_\nu$, feed $(\mathcal{D}, i, a)$ to Algorithm 1 for $(\alpha,\beta)$-*inversion.* If the answer is 0 or 1, set it to $\theta_{\nu,i}(a_i)$ and put $(i, a)$ into $\mathcal{S}_\nu$.

4: *Positively reliable cover selection.* $h_\nu(x) = \bigwedge_{(\lfloor x_i/2\rfloor,i)\in\mathcal{S}_\nu} \theta_{\nu,i} \circ x_i$.

5: *Consistency measurement.* Return `FAIL` unless both are true:

$$\begin{aligned} \textit{Recall:}\quad & \mathsf{Pr}\left[h_\nu(X) = 1 \mid Y = 1, h_{\nu-1}(X) = 0\right] \geq 1/s. \\ \textit{Small FPE:}\quad & \mathsf{Pr}\left[Y = 0, h_\nu(X) = 1\right] \leq (1 + \epsilon)2^k\beta. \end{aligned}$$

6: *Induction.* $h_\nu := h_{\nu-1} \vee h_\nu$, $\nu := \nu + 1$.

## 7.2 Linear Time Proper Learning Monotone DNF

**Theorem 7.8** (properly learning canonical DNF)**.** If $\binom{s-2}{c-1}\binom{s}{c}\frac{s^2k}{\varepsilon\delta} \ll 2^{(c-2)k}$, Algorithm 2 can PAC learn the canonical DNF $\{\bigvee_{j=1}^{s}\bigwedge_{i=1}^{k}\theta_{i+jk}\circ x_{i+jk} \mid \theta \in \{0,1\}^{ksn}\}$ in $n \cdot O\Big(\Big(\frac{k^2s^3}{\varepsilon\delta^2}2^k\binom{s-2}{c-1}\Big)^2\Big)$ time from $n \cdot O\Big(\Big(\frac{k^2s^3}{\varepsilon\delta^2}2^k\binom{s-2}{c-1}\Big)^2\Big)$ data with pairwisely $\frac{1}{n}\cdot\big(\frac{2^{2ck-1}(ks)^2}{\binom{s}{c}\delta^{1.5}}\big)^2$-dense attributes under any $\epsilon\beta$-away $2ck$-independent flipper over $\{0,1\}^{ksn}$. It is a proper PAC learning by a hypothesis class $\{\bigvee_{j=1}^{s}\bigwedge_{i=1}^{k}\theta'_{i+jk}\circ x_{i+jk} \mid \theta' \in \{0,1,*\}^{ksn}\}$ to set $\theta(\lfloor x_{i+jk}/2\rfloor) = * \Rightarrow \theta \circ x_{i+jk} \equiv 1$.

*Proof.* Set $\nu_0 = s$, $\delta_{\mathtt{inv}} = \alpha = \frac{\delta}{ks}$, $\beta = \frac{\varepsilon}{2^k\nu_0}$, and $\rho = \big(\frac{2^{ck-1}\delta_{\mathtt{inv}}}{\binom{s}{c}}\big)^2\cdot\frac{ks}{\alpha\delta_{\mathtt{inv}}^4 n}$, implying $\beta_{7.5} \ll \beta \ll 1$ by $\binom{s-2}{c-1}\binom{s}{c}\frac{s^2k}{\varepsilon\delta} \leq 2^{(c-2)k}$. Algorithm 2 may succeed if Step 5 never fails on $f_\nu = \{\mathbf{x}_{i+\nu k} \mid i \in (k]\}$:

$$\begin{aligned} &\textit{No FNE}\text{:}\ \forall(i,j), \theta'_{i+jk} \neq * \Rightarrow \theta'_{i+jk} = \theta_{i+jk}. \\ &\textit{Small FPE}\text{:}\ \mathsf{Pr}[Y = 0, h_{\nu_0-1}(X) = 1] \leq \textstyle\sum_{\nu=1}^{\nu_0-1}\mathsf{Pr}[Y = 0, h_\nu(X) = 1] \leq (1+\epsilon)2^k\beta\nu_0 \leq (1+\epsilon)\varepsilon. \end{aligned}$$

It converges to Definition 2.1's $2\varepsilon$-learning by Theorem 5.9's UGEB analysis

$$\textit{UGEB:}\ \mathsf{Pr}_D\left[P(h_{\nu_0-1}(x) \neq y) - \mathsf{Pr}[h_{\nu_0-1}(X) \neq Y] \geq \varepsilon + \varepsilon\right] \leq |\mathcal{H}|\boldsymbol{e}^{-\frac{1}{3}\varepsilon m} \leq (2^{kn})^{\nu_0-1}\boldsymbol{e}^{-\frac{1}{3}\varepsilon m} = o(\delta)$$

on $|\mathcal{H}| = \prod_{\nu=1}^{\nu_0-1}|\mathcal{H}_\nu|$ of $|\mathcal{H}_\nu|$ counting the number $2^{kn}$ of the assignments $\theta_\nu \in \{0,1\}^{kn}$.

**Recall:** Step 2 may choose the best $f_\nu$ among the $s$ terms of the target DNF to cover the remained positive examples by a ratio $\mathsf{Pr}\big[f_\nu(X) = 1 \mid Y = 1, h_{\nu-1}(X) = 0\big] \geq \frac{1}{s}$. Step 5 can attain the recall once Step 3 has correctly inverted $\theta_\nu$ on the locations $(i + jk, \lfloor X_{i+jk}/2\rfloor)$ of all $(i,j) \in (k] \times (s]$. Theorem 7.5 guarantees the inversion with significance $ks \cdot O(\delta_{\mathtt{inv}}) = O(\delta)$.

**FPE** holds under Step 3's correct inversions and Definition 7.4's prerequisite $\forall(i,j) \in (k] \times (s], \mathsf{Pr}[\lfloor X_{i+jk}/2\rfloor] \geq \alpha/n$. Lemma 2.3's $(0, \alpha/n)$-slice of $\mathsf{Pr}[\lfloor X_i/2\rfloor = a]$ over $a \in [n)$ guarantees the latter with significance $ks \cdot \alpha/n \cdot n = ks\alpha = \delta$. We may write $f_\nu = \bigwedge_{i=1}^{k}\mathbf{x}_i$ and $(k', k] := \{i \in (k] \mid (i, \lfloor X_i/2\rfloor) \in \mathcal{S}_\nu\}$. Divide the hypothesis Step 4' $h_\nu$ into $h_\nu = \sqcup_u h_u$ over $u \in \{0,1\}^{k'n}$ of $h_u(x) := h_\nu \wedge \bigwedge_{i=1}^{k'} u_i \circ x_i$. Yao's reduction takes the uniform random assignment $U \sim \{0,1\}^{k'n}$ and calculates the probability differentials between the target $h_0 = f_\nu$ and the hypothesis $h_\nu$

along a sequence $h_0,\dots,h_{k'}$ of AND functions $h_i := h_\nu \wedge \bigwedge_{\iota=1}^{i} U_\iota \circ x_\iota \wedge \bigwedge_{\iota=i+1}^{k'} \theta_\iota \circ x_\iota$. For $h_{i,b} := h_\nu \wedge \bigwedge_{\iota=1}^{i-1} U_\iota \circ x_\iota \wedge x_i \oplus b \wedge \bigwedge_{\iota=i+1}^{k'} \theta_\iota \circ x_\iota$ and $h_{i,*} := h_\nu \wedge \bigwedge_{\iota=1}^{i-1} U_\iota \circ x_i \wedge \bigwedge_{\iota=i+1}^{k'} \theta_\iota \circ x_\iota$,

$$\begin{aligned}
&\mathsf{Pr}[Y=b,h_U(X)=1]-\mathsf{Pr}[Y=b,f_\nu(X)=1]\\
&=\textstyle\sum_{i=0}^{k'-1}\big(\mathsf{Pr}[Y=b,h_{i+1}(X)=1]-\mathsf{Pr}[Y=b,h_i(X)=1]\big)\\
&=\textstyle\sum_{i=0}^{k'-1}\frac{1}{2}\mathbb{E}\big[(-1)^{\theta_i\circ X_i}1[Y=b,h_{i,*}(X)=1]\big] \quad (\because \text{Theorem 7.5's rel+ir0+ir1 decomposition})\\
&=\textstyle\sum_{i=0}^{k'-1}\frac{1}{2}\mathbb{E}\big[(-1)^{\theta_i\circ X_i}\big(1[Y=b,h_{i,*}(X)=1,f_{\mathtt{rel},i}(\theta\circ X)]+1[Y=b,h_{i,*}(X)=1,f_{\mathtt{ir}b}(\theta\circ X)]\big)\big]\\
&\le\textstyle\sum_{i=0}^{k'-1}\frac{1}{2}\big(\mathsf{Pr}[h_{i,*}(X)=1,f_{\mathtt{rel},i}(\theta\circ X)]+\mathbb{E}\big[(-1)^{\theta_i\circ X_i}1[h_{i,*}(X)=1,f_{\mathtt{ir}b}(\theta\circ X)]\big]\big)\\
&\le\textstyle\sum_{i=0}^{k'-1}\frac{1}{2}\mathsf{Pr}[\bigwedge_{\iota=1}^{i-1}U_\iota\circ X=1]\big(\mathsf{Pr}[f_{\mathtt{rel},i}(\theta\circ X)]+\mathbb{E}\big[(-1)^{\theta_i\circ X_i}\big]\mathbb{E}\big[f_{\mathtt{ir}b}(\theta\circ X)\big]\big)\\
&\le\textstyle\sum_{i=0}^{k'-1}\frac{1}{2}\cdot\frac{1}{2^{i-1}}(|\mu_i|+O(\beta\delta_{\mathtt{inv}})) \quad \Big(\because \substack{\text{Theorem 7.5's relevance, correlation-on-shift,}\\ \text{and correlation-on-data analyses}}\Big)
\end{aligned}$$

Its $b=0$ case on $\mathsf{Pr}[Y=0,f_\nu(X)=1]=0$ gives rise to FPE because $(i,\lfloor X_i/2\rfloor)\notin S_\nu$ implies $|\mu_i|<\beta$ (otherwise $\theta_i(\lfloor X_i/2\rfloor)$ is detectable) in a summation

$$\begin{aligned}
\mathsf{Pr}[Y=0,h_\nu(X)=1]&=\textstyle\sum_u\mathsf{Pr}[Y=0,h_u(X)=1]=2^k\mathsf{Pr}[Y=b,h_U(X)=1]\\
&\le 2^k(|\mu_i|+O(\beta\delta_{\mathtt{inv}}))<(1+\epsilon)2^k\beta.
\end{aligned}$$

**Computational complexity:** Algorithm 2 spends Theorem 7.5's $n\cdot O\big(\frac{\binom{s-2}{c-1}^2}{\alpha\beta^2\delta_{\mathtt{inv}}^3}\big)$ data to execute Step 3 in $kn\cdot\nu_0\cdot O\big(\frac{\binom{s-2}{c-1}^2}{\beta^2\delta_{\mathtt{inv}}^3}\big)$ time with $O(ks\delta_{\mathtt{inv}})+O(ks\alpha)=O(\delta)$ significance. □

**Theorem 7.9** (Theorem 1.4[50])**.** Suppose $\frac{s}{\varepsilon}\ll 2^k$ and $\binom{s-2}{c-1}\binom{s}{c}\frac{s^2k\ln(1/\varepsilon)}{\varepsilon^2\delta}\le 2^{(c-2)k}$. Algorithm 2 can PAC learn the planted monotone $s$-term DNF hiding $\theta\in\{0,1\}^{dn}$ and having $c$-wisely $k$-expanding terms in $n\cdot O\Big(\Big(\frac{k^2s^3}{\varepsilon^2\delta^2}2^k\binom{s-2}{c-1}(\log\frac{1}{\varepsilon})^{1.5}\Big)^2\Big)$ time. It works on any $\epsilon\varepsilon$-noisy $n\cdot O\Big(\Big(\frac{k^2s^3}{\varepsilon^2\delta^2}2^k\binom{s-2}{c-1}\log\frac{1}{\varepsilon}\Big)^2\Big)$ data with pairwisely $\frac{1}{n}\cdot(\frac{2^{2ck-1}(ks)^2}{\binom{s}{c}\delta^{1.5}})^2$-dense attributes under any $\epsilon\beta$-away $2ck$-independent flipper over $\{0,1\}^{ksn}$. It loads a properhypothesis class $\{\bigvee_{\nu=1}^{s\ln\frac{1}{\varepsilon}}\bigwedge_{i=1}^{k}\theta'_{\nu,i}\circ x_{i_\nu}\mid\theta'\in\{0,1,*\}^{ksn\ln(\frac{1}{\varepsilon})}\}$.

*Proof.* The same with Theorem 7.8's one but adopting 7.7 for Theorem Algorithm 2's step correlation retrieval, once Step 2 can have $|f_\nu|\le k$. Setting $\nu_0:=\frac{s}{\varepsilon}\ln\frac{1}{\varepsilon}$ (the other parameters are the same as Theorem 7.8) and applying Theorem 5.11's recall can provide it on $\frac{s}{2^{k+1}}\ll\varepsilon$:

$$\begin{aligned}
&\mathsf{Pr}[h_\nu(X)=1\mid Y=1,h_{\nu-1}(X)=0]\ge(\mathsf{Pr}[Y=1,h_{\nu-1}(X)=0]-\epsilon\varepsilon-\tfrac{s}{2^{k+1}})/s\ge\tfrac{(1-\epsilon)\varepsilon}{s}\\
\Rightarrow&\ \textit{Small FNE: }\mathsf{Pr}[h_{\nu_0-1}(X)=0,Y=1]=\mathsf{Pr}[Y=1]\mathsf{Pr}[h_{\nu_0-1}(X)=0\mid Y=1]\\
&=\mathsf{Pr}[Y=1]\textstyle\prod_{\nu=1}^{\nu_0-1}\big(1-\mathsf{Pr}[h_\nu(X)=1\mid Y=1,h_{\nu-1}(X)=0]\big)\le\mathsf{Pr}[Y=1](1-\tfrac{(1-\epsilon)\varepsilon}{s})^{\nu_0-1}<\varepsilon.
\end{aligned}$$ □

## 7.3 Inverting Planted Fourier Transforms over $\mathbb{Z}_q$

The standard Fourier analysis (Review4) is the correlation analysis of degree-$k$ polynomials under the uniformly distributed polarities $(X_i \bmod 2)_{i\in w}$ of $w\in\binom{[d]}{k}$. This section will extend it to smoothed analysis induced by $k$-wisely independent shifts flipping the polarities.

[50]Set $k=O(\log s)$ and $1/\varepsilon,1/\delta\le s^{O(1)}$.

**Definition 7.10** (planted Fourier transforms)**.** Planted (probabilistic) degree-$k$ Fourier transforms over $\mathbb{Z}_q$ of odd order $q \geq 3$ are degree-$k$ polynomial functions $f(x), f(x|r) : [2n)^d \to \mathbb{Z}_q$. It explains a given $\eta$-noisy data $(X,Y) \sim \mathcal{D}$ by the unknown secret parameters $\theta \in \mathbb{Z}_q^{dn}$ and known coefficients $\hat{f}_w \in \mathbb{Z}_q$ as follows, where $\theta_i \circ \mathbf{x}_i := \theta_i(\lfloor \mathbf{x}_i/2 \rfloor)(-1)^{\mathbf{x}_i}$:

$$\underset{FT}{\textit{Planted}}\text{: } f(\mathbf{x}) := \textstyle\sum_{|w|\leq k} \hat{f}_w \prod_{i\in w} \theta_i \circ \mathbf{x}_i \text{ such that } \Pr[Y \neq f(X)] \leq \eta,$$

$$\underset{\textit{planted FT}}{\textit{Probabilistic}}\text{: } f(\mathbf{x}|R) := \textstyle\sum_{|w|\leq k} \hat{f}_w \prod_{i\in w} \theta_{R,i} \circ \mathbf{x}_i \text{ such that } \forall (x,y) \in \mathcal{D}, \Pr_R[y \neq f(x|R)] \leq \eta.$$

They are non-degenerate if $\forall w \in \binom{d}{k}, \hat{f}_w \in \mathbb{Z}_q^*$. Fourier $(w,a)$-coefficient is $\theta_w(a) := \hat{f}_w \prod_{i \in w} \theta_i(a)$.

**Definition 7.11** ($(\alpha,\beta)$-inversion)**.** We say that a randomized algorithm $\mathcal{A}$ $(\alpha,\beta)$-inverts a $(w,a)$-coefficient and a parameter $\theta$ of a degree-$k$ planted FT from data $(X,Y) \sim \mathcal{D}$ if it can estimate them within accuracy $\beta$ as follows, where $\delta_{7.11} \leq \frac{\delta}{dn}$ (resp. $\delta_{7.11} \leq \frac{\delta}{n}$ when $\theta \in \mathbb{Z}_q^n$).

$$\underset{(\alpha,\beta)\textit{-inversion}}{\textit{Coefficinet}}\text{: } \Pr_{\mathcal{D},\mathcal{A}} \begin{bmatrix} \Pr[\lfloor X_w/2 \rfloor = a] \geq \alpha\mu^k \left(\text{resp.} \Pr[\lfloor X_w/2 \rfloor \subset a] \geq \alpha\mu^k\right) \\ \Rightarrow \left|\mathcal{A}(\mathcal{D}, w, a) - \theta_w(a)\right| \leq \beta \end{bmatrix} \geq 1 - O(\delta_{7.11}).$$

$$\underset{\textit{inversion}}{\textit{Parameter}}\text{: } \Pr_{\mathcal{D},\mathcal{A}} [\mathcal{A}(\mathcal{D}) = \theta] \geq 1 - O(\delta).$$

---

**Algorithm 3** $(\alpha,\beta)$-inverting Fourier coefficients

---

Input a dataset $\mathcal{D}$ and a query $(w,a) \in \binom{d}{k} \times [n)^k$. Let $\mathcal{D}_{w,a} := \{(x,y) \in \mathcal{D} \mid \lfloor x_w/2 \rfloor = a\}$ (resp. $\{(x,y) \in \mathcal{D} \mid x_w \subset a\}$ when $\theta \in \{0,1\}^n$).

1: Filter $\mathcal{D}$ to a sub-data $(X_{w,a}, Y_{w,a}) \sim \mathcal{D}_{w,a}$. If $|D_{w,a}|/|\mathcal{D}| < \alpha\mu^k$, then return ?.

2: Compute and output $\mathbf{corr}_w(\mathcal{D}_{w,a}) = \mathbf{corr}_w(X_{w,a}, Y_{w,a}) := \mathbb{E}\big[Y_{w,a} \cdot \prod_{i\in w}(-1)^{X_{w,a,i}}\big]$.

---

Algorithm 3 can invert the Fourier coefficient $\theta_w(a)$ of a target degree-$k$ planted FT $f$ through Definition 2.8's hash functional $h_w$. It will employ the following three kinds of functionals $h_w^\kappa(g) = h_{J,J'}^\kappa(g) := (h_J^\kappa(g), h_{J'}^\kappa(g))$ indexed by $\kappa \in \{\mathtt{dim}, \mathtt{hsh}, \mathtt{rem}\}$ and the random $J \neq J'$ to pick up $(g(x(J)), y(J)), (g(x(J')), y(J')) \sim \mathcal{D}_{w,a}$. Let $\delta_\kappa := \frac{\beta\delta_{7.11}}{r|\mathcal{Q}_\kappa|}$. Let $g \in \{0,1\}^{dmn}$ be a flipper $g(x_i(j)) := 2\lfloor x_i(j)/2 \rfloor + x_i(j) \oplus g(\lfloor x_i(j)/2 \rfloor)$. The Fourier $(w,a)$-coefficient inversion under $g$ consumes $m_{7.12} := \frac{2^k r^2}{\beta^2 \delta_{7.11}} + \frac{2^{2k}}{\delta_{7.11}}$ (resp. $m'_{7.12} := \frac{r^2}{\beta^2} \ln \frac{1}{\delta_{7.11}} + 2^k \ln \frac{2^k}{\delta_{7.11}}$) examples.

$$\underset{\textit{dimension}}{\textit{Small}}\text{: } h_j^{\mathtt{dim}}(g) := g(x_{w^c}(j)) \in [2n)^{d-k} \text{ and } |\mathcal{Q}_{\mathtt{dim}}| := |[2n)^{d-k}| = (2n)^{d-k}.$$

$$\underset{\textit{hashes}}{\textit{Small}}\text{: } h_j^{\mathtt{hsh}}(g) := f\big(x_w \bmod 2 \mid \lfloor x_w/2 \rfloor = a, x_{w^c} = g(x_{w^c}(j))\big) \in \mathbb{Z}_r^{\{0,1\}^w} \\ \text{and } |\mathcal{Q}_{\mathtt{hsh}}| := (2r+1)^{2^k}.$$

$$\underset{\textit{remainders}}{\textit{Sparse}}\text{: } h_j^{\mathtt{rem}}(g) = \big(f - \textstyle\sum_{v: v\cap w \neq \emptyset} \hat{f}_v \prod_{i\in v} \theta_i \circ x_i\big)\big(x_w \bmod 2 \mid \lfloor x_w/2 \rfloor = a, x_{w^c} = g(x_{w^c}(j))\big) \\ \text{and } |\mathcal{Q}_{\mathtt{rem}}| := \big|\{\xi \mid \Pr[h_J^{\mathtt{rem}}(G) = \xi] \geq \delta_{\mathtt{rem}}\}\big|.$$

**Theorem 7.12** (inverting Fourier coefficients)**.** Suppose $m = m_{7.12}$ (resp. $m'_{7.12}$) $\ll q/r$, $\beta \ll 1$, $2^k r\eta \leq \beta\delta_{7.11}$. Algorithm 3 can $(\alpha, \epsilon\beta)$-invert the $(w,a)$-coefficient of degree-$k$ planted FT over $\mathbb{Z}_q$ from any $\eta$-noisy data $\mathcal{D}_{w,a} = \{\big(G(x(j)), y(j)\big)\}_{j=1}^m$ within range $\forall j, |y(j)| \leq r$ under any $(h_w^\kappa, \delta_\kappa)$-hashed $\beta\delta_{7.11}/r$-away $2k$-independent (resp. $km$-independent) flipper $G$.

*Proof.* Follow Theorem 7.5's correlation-on-data analysis over $\mathbb{Z}$. The large modulus $r|D_{w,a}| \ll |\mathbb{Z}_q|$ can calculate Algorithm 3's Step 2's summation $\sum_j y(j) \prod_{i\in w}(-1)^{g(x_i(j))} \ll q$ over $\mathbb{Z}$ rather than the ring $\mathbb{Z}_q$. Let us first do it in an ideal situation that $\Pr[y(J) = f\big(G(x(J))\big)] = 1$,

$\forall\xi, \Pr[h^\kappa_J(G)=\xi]>0 \Rightarrow \Pr[h^\kappa_J(G)=\xi]\geq\delta_\kappa$, and the shift $G$ is perfectly $2k$-independent.

**Small coset:** The $\mathcal{D}_{w,a}$ can identify the truth-table of $h^{\mathtt{hsh}}_\xi \in \{h^{\mathtt{hsh}}_j(g)\}_{g,j}$ under $h^\kappa_J(G)=\xi$. Since $G$ is $(h^\kappa_w,\delta_\kappa)$-hashed $2k$-independent, Chebyshev's inequality of $\gamma=\frac{2^k}{\epsilon^2 m}$ makes the $m\geq\frac{2^{2k}}{\delta_{7.11}}$ data in $\mathcal{D}_{w,a}$ to witness $(G(x_w(j)) \bmod 2, y(j))=(b,h^{\mathtt{hsh}}_\xi(b))$ for every $b\in\{0,1\}^w$:

$$\begin{aligned}
&\tfrac{1}{m^2}(\textstyle\sum_{j=1}^m \Pr[G(x_w(j))=b \bmod 2 \mid h^\kappa_j(G_{w^c})=\xi]-\tfrac{1}{2^k})^2\\
&\quad=\tfrac{1}{m^2}\textstyle\sum_{j=1}^m (\Pr[G(x_w(j))=b \bmod 2 \mid h^\kappa_j(G_{w^c})=\xi]-\tfrac{1}{2^k})^2=\tfrac{1-1/2^k}{m\cdot 2^k} \quad\Rightarrow\\
&\underset{on\ data}{k\text{-}unif}: \Pr_G\Big[\forall b, \big|\Pr_J\big[G(x_w(J))=b \bmod 2 \mid h^\kappa_J(G_{w^c})=\xi\big]-\tfrac{1}{2^k}\big|\geq\sqrt{\tfrac{1-1/2^k}{\gamma\cdot 2^k m}}\approx\tfrac{\epsilon}{2^k}\Big]\leq 2^k\cdot\gamma\leq O(\delta_{7.11}).\\
&\underset{hash}{Small}: \forall b\in\{0,1\}^w, \forall v\subset w, |h^{\mathtt{hsh}}_\xi(b)|\leq r \wedge |(\hat{h}^{\mathtt{hsh}}_\xi)_v|=\big|2^{-k}\textstyle\sum_{b\in\{0,1\}^k} h^{\mathtt{hsh}}_\xi(b)\prod_{i\in v}(-1)^{b_i}\big|\leq r.
\end{aligned}$$

**Correlation on data:** Fixing $h^\kappa_J(G)=\xi$ induces a substitution $x_{w^c}\leftarrow G(x_{w^c}(J))$ to collapse $h^{\mathtt{hsh}}_J(G_{w^c})$ to $h^{\mathtt{hsh}}_\xi$ and yield $h^{\mathtt{hsh}}_\xi(\mathbf{x})=\theta_w(a)\prod_{i\in w}(-1)^{\mathbf{x}_i}+\sum_{v\subset w,v\neq w}(\hat{h}^{\mathtt{hsh}}_\xi)_v\prod_{i\in v}(-1)^{\mathbf{x}_i}$ of $\mathbf{x}\in\{0,1\}^w$. It can invert $\theta_w(a)$ via the correlation-on-data analysis under the random flipper $G$:

$$\begin{aligned}
&\mathbb{E}_{G,J}[\mathbf{corr}_w\big(G(x(J)),y(J)\big)]\\
&=\textstyle\sum_\xi \mathbb{E}\big[f\big(G(x(J))\big)\prod_{i\in w}(-1)^{G(x_i(J))} \mid h^\kappa_J(G_{w^c})=\xi\big]\cdot\Pr\big[h^\kappa_J(G_{w^c})=\xi\big]\\
&=\textstyle\sum_\xi 2^{-k}\sum_{b\in\{0,1\}^w} h^{\mathtt{hsh}}_\xi(b)\prod_{i\in w}(-1)^{b_i}\cdot\Pr[h^\kappa_J(G_{w^c})=\xi]=\sum_\xi\theta_w(a)\cdot\Pr[h^\kappa_J(G_{w^c})=\xi]=\theta_w(a).
\end{aligned}$$

The zero-averaged correlations $\overline{\mathbf{corr}}_v(G(x(j))):=\mathbf{corr}_w\big(G(x(j)),y(j)\big)-\theta_w(a)$ are mutually perpendicular $\mathbb{E}_G[\overline{\mathbf{corr}}_w(G(x(j)))\overline{\mathbf{corr}}_w\big(G(x(j'))\big) \mid h^\kappa_w(G)=\xi]=0$ on the perfectly $2k$-independence assumption, inverting the $\theta_w(a)$ for variance as well:

$$\begin{aligned}
&\mathbb{E}_{G,J}\big[\overline{\mathbf{corr}}_w\big(G(x(J)),y(J)\big)^2\big]\\
&=\textstyle\sum_\xi\frac{1}{m^2}\sum_{j,j'}\mathbb{E}_G\big[\overline{\mathbf{corr}}_w(G(x(j)),y(j))\overline{\mathbf{corr}}_w(G(x(j')),y(j')) \mid h^\kappa_{j,j'}(G_{w^c})=\xi\big]\cdot\Pr_G[h^\kappa_{j,j'}(G_{w^c})=\xi]\\
&=\textstyle\sum_\xi\frac{1}{m^2}\sum_{j=1}^m\mathbb{E}_G\big[\big(\mathbf{corr}_w\big(G(x(j)),y(j)\big)-\theta_w(a)\big)^2 \mid h^\kappa_j(G_{w^c})=\xi\big]\cdot\Pr_G[h^\kappa_j(G_{w^c})=\xi]\\
&=\textstyle\sum_\xi\frac{1}{m^2}\sum_j\big(2^{-k}\sum_{b\in\{0,1\}^w}\sum_{v\subset w,v\neq w}(\hat{h}^{\mathtt{hsh}}_\xi)_v\prod_{i\in w\setminus v}(-1)^{b_i}\big)^2\cdot\Pr\big[h^\kappa_j(G_{w^c})=\xi\big]\\
&\leq\textstyle\sum_\xi\frac{1}{m^2}\sum_j\sum_{v\subset w,v\neq w}(\hat{h}^{\mathtt{hsh}}_\xi)^2_v\Pr\big[h^\kappa_J(G_{w^c})=\xi\big]\leq r^2(2^k-1)/m. \quad (\because \text{the small hash.})
\end{aligned}$$

Chebyshev's inequality parameter $\gamma=O(\frac{1}{\delta_{7.11}})$ and $m\geq\frac{2^k r^2}{\beta^2\delta_{7.11}}$ guarantees

$$\underset{on\ data}{Correlation}: \Pr_G\big[|\overline{\mathbf{corr}}_w(G(X_{w,a}))|\geq\sqrt{r^2(2^k-1)\gamma/m}\gg\beta\big]\leq O(1/\gamma)=O(\delta_{7.11}).$$

**$(\alpha,\beta)$-inversion:** The $\eta$-noisy label $\tilde{y}(j)$ preserves the above correlation-on-data analysis by

$$\begin{aligned}
&\mathbb{E}_{G,J}\big[\tilde{y}(J)\neq f\big(G(x(J))\big) \mid h^\kappa_J(G_{w^c}),G(x_w(J))\big]\leq\eta\leq\tfrac{\beta\delta_{7.11}}{2^k r} \quad (\because 2^k r\eta\leq\beta\delta_{7.11})\\
&\Rightarrow\forall b\in\{0,1\}^k, \Pr_{G,J}[\tilde{y}(J)\neq h^{\mathtt{hsh}}_\xi(b) \mid h^\kappa_J(G_{w^c}),G(x_w(J))=b]\leq\tfrac{\epsilon'\beta}{r} \ (\because \text{Markov-ineq of } \gamma=\tfrac{\delta_{7.11}}{\epsilon'})\\
&\Rightarrow\underset{under\ noise}{correration}: |\overline{\mathbf{corr}}_w(G(X_{w,a}),Y_{w,a})|\leq\epsilon'\beta+\max_j|\tilde{y}(j)|\cdot\epsilon'\beta/r\leq 2\epsilon'\beta. \quad (\because \underset{on\ data}{correration})
\end{aligned}$$

Although the actual shift $\tilde{G}_\kappa$ may take $0<\Pr[h^\kappa_J(\tilde{G}_\kappa)=\xi]<\delta_\kappa$ for $\xi\in\mathcal{Q}_\kappa$, Lemma 2.3's $(0,\delta_\kappa)$-slice bounds its contribution $\Pr_{h^\kappa_J(\tilde{G}_\kappa)}\big[\Pr[h^\kappa_J(\tilde{G}_\kappa)]<\delta_\kappa\big]\leq\delta_\kappa\cdot|\mathcal{Q}_\kappa|\leq\frac{\beta\delta_{7.11}}{r}$ in any three

$\delta_\kappa$ of $\kappa \in \{\texttt{dim}, \texttt{hsh}, \texttt{rem}\}$. The local shift $\tilde{G}_\kappa(x_w(J))$ may be $\frac{\beta\delta_{7.11}}{r}$-away from the perfect $2k$-independent $G(x_w(J))$ on the location $(w,a)$, bounding the correlation under $\tilde{G}_\kappa$ by Markov's inequality parameter $\gamma = \delta_{7.11}/\epsilon'$:

$$\begin{aligned}
&\mathbb{E}[|\overline{\mathbf{corr}}_w(\tilde{G}_\kappa(X_{w,a}), Y_{w,a})|] \\
&\le \begin{cases} \mathbb{E}[\big|\mathbf{corr}(\tilde{G}_\kappa(x(J)), \tilde{y}(J)) - \mathbf{corr}(G(x(J)), \tilde{y}(J))\big|] + \max_j |\tilde{y}(j)| \cdot \mathsf{Pr}\big[\mathsf{Pr}[h^\kappa_J(G_{w^c})] < \delta_\kappa\big] \\ \qquad\qquad +\mathbb{E}[\big|\overline{\mathbf{corr}}_w(G(x(J)), \tilde{y}(J))\big|] \mid \mathsf{Pr}[h^\kappa_J(G_{w^c}) = \xi] \ge \delta_\kappa\big] \end{cases} \\
&\le \max_j |\tilde{y}(j)| \cdot \beta\delta_{7.11}/r + \max_j |\tilde{y}(j)| \cdot \beta\delta_{7.11}/r + 2\epsilon'\beta \quad (\because \text{the correlation under noise}) \\
&\Rightarrow \underset{-inversion}{(\alpha,\epsilon\beta)}: \mathsf{Pr}[|\overline{\mathbf{corr}}_w(G(X_{w,a}), Y_{w,a})| \ge 2\beta\delta_{7.11}\gamma + 2\epsilon'\beta = 4\epsilon'\beta] \le 1/\gamma = O(\delta_{7.11}).
\end{aligned}$$

**Inversion under stronger independence:** The data size can reduce from $m \gg m_{7.12}$ to $m \gg m'_{7.12}$ under $mk$-wisely independent shift $G$. Theorem 7.12's $k$-uniformity and correlation-on-data are achievable by Chernoff bound (instead of Chebyshev's inequality for weaker independence):

$$\begin{aligned}
&\underset{on\ data}{k\text{-}unif}: \mathsf{Pr}_G\Big[\forall b, \big|\mathsf{Pr}_J\big[G(x_w(J)) = b \bmod 2 \mid h^\kappa_J(G_{w^c}) = \xi\big] - \tfrac{1}{2^k}\big| \ge \tfrac{\epsilon}{2^k}\Big] \le 2^k \cdot \boldsymbol{e}^{-\frac{\epsilon^2}{2+\epsilon}\frac{m}{2^k}} \le O(\delta_{7.11}). \\
&\underset{on\ data}{Corr}: \mathsf{Pr}_G\big[|\overline{\mathbf{corr}}_w(G(X_{w,a}))| \ge \epsilon\beta\big] = \mathsf{Pr}_G\big[\big|\tfrac{\mathbf{corr}_w(G(X_{w,a})) + 2r}{3r} - \mu\big| \ge \beta' := \tfrac{\epsilon\beta}{3r}\big] \\
&\qquad \le \boldsymbol{e}^{-\frac{(\beta'/\mu)^2}{2+\beta'/\mu}\mu m} \le O(\delta_{7.11}) \text{ for } \mu = \tfrac{\theta_w(a) + 2r}{3r}.
\end{aligned}$$

□

---

**Algorithm 4** Inverting planted parameters

---

Input a dataset $\mathcal{D}$, execute the following 1–4 and output $\theta \in \mathbb{Z}_q^{dn}$.

1: *Linear case.* When $k = 1$, query $(\mathcal{D}, i, a)$ to Algorithm 3 for $(\alpha,\beta)$-inverting every $\theta_i(a)$ of $(i,a) \in (d] \times [n)$, and finish. The following steps suppose $k \ge 2$.

2: *A base location selection.* Guess a base location $(w_0 \sqcup i_0, a_0) \in \binom{(d]}{k+1} \times [n)^{w_0 \sqcup i_0}$ at which $\theta_{w_0}(a_0)\theta_{i_0}(a_{0,i_0})$ is invertible. Let $w_{0,-i,+i'} := (w_0 \backslash i) \sqcup i'$ for $(i,i') \in w_0 \times w_0^c = w_0 \times ((d] \backslash w_0)$.

3: *Fourier inverting the base parameters.* Fix an arbitrary $i_1 \in w_0$. For all $i \in w_0$, query $(\mathcal{D}, w_0, a_0)$ and $(\mathcal{D}, w_{0,-i,+i_0}, a_0)$ to Algorithm 3, and retrieve $\theta_i(a_{0,i})$ in the following calculus:

$$\begin{aligned}
\frac{\theta_i(a_{0,i})}{\theta_{i_1}(a_{0,i_1})} = \frac{\theta_i(a_{0,i})}{\theta_{i_0}(a_{0,i_0})} \cdot \frac{\theta_{i_0}(a_{0,i_0})}{\theta_{i_1}(a_{0,i_1})} &= \frac{\prod_{\kappa \in w_0} \theta_\kappa(a_{0,\kappa})}{\prod_{\kappa \in w_{0,-i,+i_0}} \theta_\kappa(a_{0,\kappa})} \cdot \frac{\prod_{\kappa \in w_{0,-i_1,+i_0}} \theta_\kappa(a_{0,\kappa})}{\prod_{\kappa \in w} \theta_\kappa(a_{0,\kappa})} \\
\Rightarrow \theta_{i_1}^k(a_{0,i_1}) &= \prod_{i \in w_0} \theta_i(a_{0,i}) \cdot \prod_{i \in w_0} \frac{\theta_{i_1}(a_{0,i_1})}{\theta_i(a_{0,i})}.
\end{aligned}$$

4: *Fourier inverting all parameters.* Query $(\mathcal{D}, w_{0,-i,+i'}, a_0 \sqcup a_{i'})$ to Algorithm 3 for $(\alpha,\beta)$-inversion of $\theta_{i'}(a_{i'}) = \theta_i(a_{0,i}) \cdot \prod_{\kappa \in w_{0,-i,+i'}} \theta_\kappa((a_0 \sqcup a_{i'})_\kappa) / \prod_{\kappa \in w_0} \theta_\kappa(a_{0,\kappa})$ until retrieving $\theta$.

---

**Theorem 7.13** (Theorem 1.5[51])**.** Suppose $m_{7.12}$ (resp. $m'_{7.12}$) $\ll q/r$, $\beta \ll 1$, and $2^{2k} r\eta \le \beta\delta_{7.11}$. In Algorithm 4, suppose that $X$ is $k$-wisely $(\mu,\alpha)$-sparse (or $(\mu,\alpha)$-cover when $\theta \in \mathbb{Z}_q^n$) at every location $(w,a)$ queried in steps 3 and 4, and $\mathcal{D}_{w,a}$ contains noise at most $\eta$. Then, Algorithm 4 can $\theta$-invert degree-$k$ planted FT over $\mathbb{Z}_q$ in $O\big((\binom{d}{k} + dn)m\big)$ time from any data $\{\big(G(x(j)), y(j)\big)\}_{j=1}^m$ of size $m = O\big(\frac{m_{7.12}}{\alpha\mu^k}\big)$ (resp. $m = O\big(\frac{m'_{7.12}}{\alpha\mu^k}\big)$) and range $\forall j, |y(j)| \le r$ under any $(h_w^{\texttt{hsh}}, \delta_{\texttt{hsh}})$-hashed $\beta\delta_{7.11}/r$-away $2k$-independent (resp. $\beta\delta_{7.11}/r$-away $m'_{7.12}k$-independent) flipper $G$.

[51] Take $\frac{k}{\alpha\beta\delta} \le O(1)$, $q \gg n^{2+1/2^{k-1}}$, $r = q^{1/2^{k+1}}$, $\delta_{7.11} = 1/n$, $\eta \ll 1/(nr)$, and $\mu = 1/n$. Lemma 2.11 provides a probabilistic shift $G$ of cardinality $\tilde{O}(q^{1/2}(nr)^3)$.

*Proof.* If all $(\alpha, \beta)$-inversions of $\mathcal{Q} = \{(w_0, a_0), (w_{0,-i,+i_0}, a_0), (w_{0,-i,+i'}, a_0 \sqcup a_{i'})\}$ queried in Steps 3 and 4 succeed, Algorithm 4 could identify the correct integer coefficients $\theta_w(a)$ due to $\beta \ll 1/2$, so retrieving the secret parameter $\theta \in \mathbb{Z}_q^n$.

$(\alpha, \beta)$**-inverting Fourier coefficients:** Algorithm 4's Step 2 must choose a location $(w_0, a_0)$ such that $\theta_{w_0}(a_0)$ is invertible and $\mathsf{Pr}[\lfloor X_w/2 \rfloor = a] \geq \alpha\mu^k$. They may query to Algorithm 3 for $|\{(a,i) \in [n) \times (d]\}|$ locations. They can receive sufficiently many examples due to $k$-wisely $\mu$-sparse (resp. cover) over the given $m = O\big(m_{7.12}/(\alpha\mu^k)\big)$ (resp. $m = O\big(m'_{7.12}/(\alpha\mu^k)\big)$) data. CB parameter $\gamma = 1$ with significance level $|\{(a,i) \in [n) \times (d]\}| \cdot \boldsymbol{e}^{-1/2 \cdot \alpha\mu^k \cdot m} \ll o(\delta)$ guarantees:

$$\underset{\textit{many examples}}{\textit{Sufficiently}}: \quad \forall (w,a), m_{7.12} \text{ (resp. } m'_{7.12}) \ll |\mathcal{D}_{w,a}| \leq 2m \cdot \alpha\mu^k \ll q/r.$$

Since $\delta_{7.11} \ll \frac{\delta}{dn}$, Step 4 inverts $\theta_i(a)$ of $\forall (i,a) \in (d] \times [n)$ with significance $O(\delta_{7.11}) \cdot dn = O(\delta)$. $\square$

### 7.4 Inverting Linear Fourier Transforms and Breaking LWE

Theorems 7.12 has demanded a large modulus $|\mathcal{D}_{w,a}| \ll q/r$ for Fourier inverting $\theta_w(a)$ over $\mathbb{Z}_q$ from $\mathcal{D}_{w,a} \subset [2n)^d \times \mathbb{Z}_r$. Previously, *modulus amplification* have brought remarkable breakthroughs in computational complexity theory, e.g., Toda's $\mathsf{PP} = \bigoplus \mathsf{P}$ [Tod91], Beigel and Tarui's $\mathsf{ACC} \subset \mathrm{SYM} \circ \mathrm{AND}_{\mathrm{plog}(n)}$ [BT94], and Williams's $\mathsf{NEXP} \not\subset \mathsf{ACC}$ [Wil14a]. This section will show that the modulus amplification can solve LWE and even $\mathrm{GapSVP}_{\tilde{O}(n^2)}$ thanks to the well-known worst-case to average-case reduction [Ajt96, MR07, Pei09, Reg09, BLP+13].

**Lemma 7.14** (modulus amplification [Yao85, Tod91, BT94])**.** There is a degree-$(2\ell-1)$ and norm-$2^{3\ell}$ polynomial $\phi_\ell(\mathbf{x})$ with the leading coefficient $(-1)^{\ell+1}\binom{2(\ell-1)}{\ell-1}$ such that

$$\underset{\textit{amplification}}{\textit{Modulus}}: (x \equiv 0 \bmod m \Rightarrow \phi_\ell(x) \equiv 0 \bmod m^\ell) \wedge (x \equiv 1 \bmod m \Rightarrow \phi_\ell(x) \equiv 1 \bmod m^\ell).$$

**Theorem 7.15** (inverting linear Fourier transform)**.** Let $p$ be an odd prime number coprime with $\binom{2(\ell-1)}{\ell-1}$, $k = (2\ell-1)v$, and $|\mathcal{Q}_{\mathtt{rem}}| = p$ (so $\delta_{\mathtt{rem}} = \frac{\beta\delta_{7.11}}{rp}$). Suppose $m_{7.12}$ (resp. $m'_{7.12}$) $\ll p^\ell/r$, $\beta \ll 1$, and $2^{2k} r\eta/\beta \leq \delta_{7.11}$. Suppose that $X$ is $k$-wisely $(\mu, \alpha)$-sparse (resp. $k$-wisely $(\mu, \alpha)$-cover when $\theta \in \mathbb{Z}_q^n$) at every location $(w,a)$ queried in Algorithm 4, and the sub-data $\mathcal{D}_{w,a}$ contains noise at most $\eta$. Then, the linear planted FT over $\mathbb{Z}_p$ is invertible in $O\big((\binom{d}{k} + dn)m\big)$ time from any data $\{\big(G(x(j)), y(j)\big)\}_{j=1}^m$ of $m = O\big(m_{7.12}/(\alpha\mu^k)\big)$ (resp. $m = O\big(m'_{7.12}/(\alpha\mu^k)\big)$), $\forall j, |y(j)| \leq r$, and $|y((m])| \leq v$ under any $(h_w^{\mathtt{rem}}, \delta_{\mathtt{rem}})$-hashed $\beta\delta_{7.11}/r$-away $2k$-independent (resp. $(h_w^{\mathtt{rem}}, \delta_{\mathtt{rem}})$-hashed $\beta\delta_{7.11}/r$-away $(km'_{7.12})$-independent) flipper $G$.

*Proof.* The variation assumption $|y((m])| \leq s$ presents a modulus amplified polynomial

$$\mathbf{y} = \textstyle\sum_{y \in y((m])} y \cdot 1[\mathbf{y} = y] \bmod p, \quad 1[\mathbf{y} = y] = \prod_{y' \in y((m]) - \{y\}} \frac{\mathbf{y} - y'}{y - y'} \Rightarrow$$
$$\underset{\textit{amplification}}{\textit{Modulus}}: \mathbf{y} = \psi_\ell(\mathbf{y}) := \textstyle\sum_{y \in y((m])} y \cdot \phi_\ell\big(\prod_{y' \in y((m]) - \{y\}} \frac{\mathbf{y} - y'}{y - y'}\big) \bmod p^\ell \quad (\because \text{Lemma 7.14})$$

Algorithm 4 can $\theta$-invert from the modulus amplified covariate dataset $\{G(x(j)), \psi_\ell(y(j))\}_{j=1}^m$. Fixing $h_J^{\mathtt{rem}}(G_{w^c}) := \sum_{i \in w^c} \hat{f}_i \theta(a_i)(-1)^{G(x(J))} = \xi \in \mathbb{Z}_p$ determines the hash function $h_J^{\mathtt{rem}}(G_{w^c}) = \sum_{i \in w} \hat{f}_i \theta(a_i)(-1)^{\mathbf{x}_i} + \xi : \{0,1\}^w \to \mathbb{Z}_p$ and its modulus amplification $h_\xi^{\mathtt{rem}}(\mathbf{x}) := \psi_\ell(h_J^{\mathtt{rem}}(G_{w^c})) : \{0,1\}^w \to \mathbb{Z}_{p^\ell}$. Accordingly, the modulus amplified degree-$k$ Fourier transform $h_\xi^{\mathtt{rem}}(\mathbf{x})$ over $\mathbb{Z}_{p^\ell}$ makes Theorems 7.12 and 7.13's proofs valid on the $h_J^{\mathtt{rem}}(G_{w^c})$'s sparseness $|\mathcal{Q}_{\mathtt{rem}}| = p$. $\square$

**Definition 7.16** (LWE in smoothed analysis)**.** Let $q \geq 3$ be an odd number. LWE over $\mathbb{Z}_q$ presents a dataset $\{(g(x(j)), y(j))\}_{j=1}^{m}$ about the following linear planted FT disturbed by arbitrary i.i.d. noises $E_j \in \mathbb{Z}_q$. It asks to invert the hidden vector $\theta \in \mathbb{Z}_q^n$ with high confidence.

$$\text{LWE: } \quad y(j) = f(g(x(j))) := \sum_{i=1}^{n} \hat{f}_i \cdot \theta_i \cdot \lceil x_i(j)/2 \rceil \cdot (-1)^{g_i(x_i(j))} + E_j.$$

Let $\mathbf{1}_w = (1, \ldots, 1)$ be the all-one vector over $i \in w$. Algorithm 4 can invert LWE by choosing $a_0 = \mathbf{1}_{w_0}$ and making $\sum_{i \in w_0} \theta_i a_{0,i} (-1)^{G(a_{0,i})} = \sum_{i \in w_0} \pm\theta_i$ concentrate near zero under the i.i.d. signs of the small secrets $\theta_i$. This $(\alpha, \beta)$-inversion algorithm queries about $\mathcal{W}_{w_0,i_0,i_1} := \{(w_0, \mathbf{1}_{w_0})\} \sqcup \{(w_{0,-i,+i'}, \mathbf{1}_{w_{0,-i,+i'}})\}_{i \in [n]}$ of $w_0 \in \binom{[n]}{k}$, $i_0 \notin w_0$, $i_1 \in w_0$, and $i' = i'(i)$ such that $(i \in w_0 \Rightarrow i' = i_0) \wedge (i \notin w_0 \Rightarrow i' = i_1) \wedge (i \in w_0 \sqcup \{i_0\} \Rightarrow \theta_i \neq 0)$.

**Theorem 7.17** (Theorem 1.6[52])**.** Let $p$ be an odd prime coprime with $\binom{2(\ell-1)}{\ell-1}$, $v = 2r + 1$, $|\mathcal{Q}_{\mathtt{rem}}| \approx v$, $k = (2\ell - 1)w$, $\gamma_{\mathtt{sm}} = (2k \log \frac{r}{\beta\delta_{7.11}})^{1/2}$, $m_{7.17} := \frac{2^{2k}(s\gamma_{\mathtt{sm}})^5}{\beta^2\delta_{7.11}} + \frac{2^k r^2}{\beta^2\delta_{7.11}} + \frac{2^{2k}}{\delta_{7.11}}$ (resp. $m'_{7.17} := \frac{(s\gamma_{\mathtt{sm}})^2}{\beta} \ln(2^k s\gamma_{\mathtt{sm}}) + \frac{r^2}{\beta^2} \ln \frac{1}{\delta_{7.11}} + 2^k \ln \frac{2^k}{\delta_{7.11}}$.) Suppose $m_{7.17}$ (resp. $m'_{7.17}$) $\ll p^\ell/r$, $\beta \ll 1$, and $s\gamma_{\mathtt{sm}} \ll r$. Suppose that $X$ is $k$-wisely $(\mu, \alpha)$-sparse at every place $(w, \mathbf{1}_w) \in \mathcal{W}_{w,i_0,i_1}$. Then, LWE over $\mathbb{Z}_p$ can retrieve small secrets $\forall i, |\theta_i| \leq s$ in $O\big((\binom{d}{k} + n)m\big)$ time from any $m \gg m_{7.17} \cdot p/(\alpha\mu^k)$ data under any $(h_w^{\mathtt{rem}}, \delta_{\mathtt{rem}})$-hashed $\beta\delta_{7.11}/r$-away $2k$ (resp. $km'_{7.17}$) independent flipper $G$ if $h_J^{\mathtt{rem}}(G_{w^c}) \in \mathbb{Z}_p$ is $\beta\delta_{7.11}/r$-away from the uniform randomness.

*Proof.* A reduction to Theorems 7.12 and 7.15's noise-free case $\eta = 0$, because the i.i.d. noise $E_J$ filters the data $\mathcal{D}_{w,\mathbf{1}_w}$ to $\mathcal{D}_\xi = \mathcal{D}_{w,\xi} := \{(G(x(j)), y(j)) \in \mathcal{D}_{w,\mathbf{1}_w} \mid \Xi = \xi\}$ of $\Xi := h_J^{\mathtt{rem}}(G_{w^c})$, over which $G$ is $(h_w^{\mathtt{rem}}, \delta_{\mathtt{rem}})$-conditionally $\beta\delta_{7.11}/r$-away $2k$-independent. Suppose the ideal case discussed in Theorem 7.12. Chernoff (resp. Chebyshev) bound parameter $\gamma = \frac{\gamma_{\mathtt{sm}}}{k/2}$ (resp. $\frac{2^k(s\gamma_{\mathtt{sm}})^4}{(\epsilon\beta)^2 m_{7.17}}$) on $q_\xi(\mathbf{x}) := \sum_{i \in w} \theta_i (-1)^{\mathbf{x}_i} = h_j^{\mathtt{hsh}} - h_j^{\mathtt{rem}}$ makes the smallness (resp. $k$-uniform-on-data) sharper:

$$\textit{Smallness: } \mathsf{Pr}_G\big[|q_\xi(G(x_w(j)))| \geq s\gamma_{\mathtt{sm}}\big] < 2\boldsymbol{e}^{-\gamma^2/(2+\gamma)\cdot k/2} \ll \beta\delta_{7.11}/r.$$

$$\textit{Smoothness: } (|\xi| \leq r - s\gamma_{\mathtt{sm}} \Rightarrow |q_\xi(G(x_w(J))) + \xi| \leq r) \\ \wedge \ (|\xi| > r + s\gamma_{\mathtt{sm}} \Rightarrow |q_\xi(G(x_w(J))) + \xi| > r).$$

$$\textit{k-uniform on data}: \mathsf{Pr}\Big[\forall \xi \in (r - s\gamma_{\mathtt{sm}}, r + s\gamma_{\mathtt{sm}}], \forall b \in \{0,1\}^w, \\ \big|\mathsf{Pr}_J[G(x_w(J)) = b \bmod 2 \mid \Xi = \xi] - \tfrac{1}{2^k}\big| \geq \sqrt{\tfrac{1-1/2^k}{\gamma 2^k m_{7.17}}} \approx \tfrac{\epsilon\beta}{2^k(s\gamma_{\mathtt{sm}})^2}\Big] \leq 2s\gamma_{\mathtt{sm}} \cdot 2^k \cdot \gamma \leq O(\delta_{7.11}).$$

$$\mathsf{Pr}\Big[\forall \xi, \forall b, \big|\mathsf{Pr}_J[G(x_w(J)) = b \bmod 2 \mid \Xi = \xi] - \tfrac{1}{2^k}\big| \geq \tfrac{\gamma}{2^k}\Big] \leq 2^{k+1} s\gamma_{\mathtt{sm}} \cdot \boldsymbol{e}^{-\frac{\gamma^2}{1+\gamma}\frac{m'_{7.17}}{2^k}} \leq O(\delta_{7.11})$$

$$\text{by } \gamma = \tfrac{\epsilon\beta}{(s\gamma_{\mathtt{sm}})^2} \text{ for } km'_{7.17} \text{ independent flipper } \Big(\because \ m_{7.17} \gg \tfrac{2^{2k}(s\gamma_{\mathtt{sm}})^5}{\beta^2\delta_{7.11}} \text{ and } m'_{7.17} \gg \tfrac{(s\gamma_{\mathtt{sm}})^2}{\beta} \ln(2^k s\gamma_{\mathtt{sm}})\Big).$$

Let $\mathcal{C} := \{(x, y) \in \mathcal{D}_{w,\mathbf{1}_w} \mid y \in \mathbb{Z}_{2r+1}\}$ be those data having range $|y| \leq r$ and variation $|\mathbb{Z}_{2r+1}| = 2r+1 = v$. We will discard all data not belonging to $\mathcal{C} \cap (\bigsqcup_\xi \mathcal{D}_\xi)$ and apply Theorem 7.12 to $\mathcal{C} \cap \mathcal{D}_\Xi$. We call a dataset $\mathcal{D}_\Xi$ fully colliding when $\mathcal{D}_\Xi \subset \mathcal{C}$.

**Inverting fully colliding data:** Under the fully colliding $\mathcal{D}_\Xi \subset \mathcal{C}$, Theorem 7.13 has shown

$$\textit{Sufficiently many examples}: \quad \forall (w, \mathbf{1}_w) \in \mathcal{W}_{w,i_0,i_1}, m_{7.17} \ll |\mathcal{D}_\Xi| \leq 2m \cdot \alpha\mu^k/p \ll p^\ell/r.$$

[52] Take $p \geq n^{\Omega(1)}$, $\mu = \frac{2}{p}$, $\alpha = 1$, $\max\{\ell, s\} \leq O(1)$, and $\max\{k, r\} \leq O(\log n)$ to have $2^{2k}n + 2^k r^2 n \ll p^{\ell-1}$.

It runs Algorithm 3's Step 2 over $\mathbb{Z}$ rather than $\mathbb{Z}_{p^\ell}$. The smallness bounds $|h_\xi^{\mathtt{hsh}} - \Xi| \leq s\gamma_{\mathtt{sm}}$. Also, Theorem 7.12's small-hash and correlation-on-data hold at $(w, \mathbf{1}_w) \in \mathcal{W}_{w,i_0,i_1}$. Algorithm 3 can uniquely identify $\theta_w(\mathbf{1}_w)$ of every $(w, \mathbf{1}_w) \in \mathcal{W}_{w_0,i_0,i_1}$. Algorithm 4 can invert the hidden $\theta$ from the coefficients $\theta_w(\mathbf{1}_w)$ via Theorem 7.15's modulus amplification

$$\underset{\textit{amplification}}{\textit{Modulus}}:\ \psi_\ell(q_\xi(\mathbf{x}) + \xi) = \sum\nolimits_{y \in \{y(j)\}_j} y\phi_\ell\Big(\prod\nolimits_{y' \in \{y(j)\}_j - \{y\}} \tfrac{q_\xi(\mathbf{x}) + \xi - y'}{y - y'}\Big) \equiv q_\xi(\mathbf{x}) + \xi \bmod \mathbb{Z}_{p^\ell}.$$

**Inverting partially-colliding data:** The partially-colliding $\mathcal{D}_\Xi \not\subset \mathcal{C}$ reduces to the full one by adding to Theorem 7.12's correlation accuracy an extra overhead $\beta$ as follows. It couples two symmetric partial collisions $h_{r-\ell}^{\mathtt{hsh}}(\{0,1\}^w) \cap \mathbb{Z}_{2r+1}$ and $2r+1+h_{-(r+\ell+1)}^{\mathtt{hsh}}(\{0,1\}^w) \cap \mathbb{Z}_{2r+1}$, where $2r+1 = (r-\ell) - (-(r+\ell+1))$ to make $h_{r-\ell}^{\mathtt{hsh}}(\{0,1\}^w)$ fully colliding. The following correlation analysis justifies it due to the $k$-uniform on data and smoothness in $\overset{\star}{=}$, and $s\gamma_{\mathtt{sm}} \ll r$ in $\overset{\star}{\lll}$:

$$\begin{aligned}
&\mathbb{E}\big[h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \cdot \prod\nolimits_{i \in w} (-1)^{G_i(1)} \cdot 1[h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \in \mathbb{Z}_{2r+1}, h_\Xi^{\mathtt{hsh}}(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}]\big] \\
&= \sum_{\ell=0}^{s\gamma_{\mathtt{sm}}} \sum_{\kappa=0}^{1} \left( \begin{array}{c} \mathbb{E}\big[h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \cdot \prod_{i \in w} (-1)^{G_i(1)} \cdot 1 \begin{bmatrix} h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \in \mathbb{Z}_{2r+1}, \\ h_\Xi^{\mathtt{hsh}}(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^\kappa (r-\ell) \end{bmatrix} \big] + \\ \mathbb{E}\big[h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \cdot \prod_{i \in w} (-1)^{G_i(1)} \cdot 1 \begin{bmatrix} h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \in \mathbb{Z}_{2r+1}, \\ h_\Xi^{\mathtt{hsh}}(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^{\kappa+1} (r+\ell+1) \end{bmatrix} \big] \end{array} \right) \\
&= \sum_{\ell,\kappa} \left( \begin{array}{l} \mathbb{E}\big[h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \cdot \prod_{i \in w} (-1)^{G_i(1)} \cdot 1 \begin{bmatrix} h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \in \mathbb{Z}_{2r+1}, \\ h_\Xi^{\mathtt{hsh}}(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^\kappa (r-\ell) \end{bmatrix} \big] \\ \quad + \mathbb{E}\Big[\big(h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) + (-1)^\kappa \cdot (2r+1)\big) \cdot \prod_{i \in w} (-1)^{G_i(1)} \\ \qquad \times 1 \begin{bmatrix} h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \in \mathbb{Z}_{2r+1}, \\ h_\Xi^{\mathtt{hsh}}(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^{\kappa+1} (r+\ell+1) \end{bmatrix} \Big] \\ \quad + \mathbb{E}\Big[(-1)^{\kappa+1}(2r+1) \cdot \prod_{i \in w} (-1)^{G_i(1)} \\ \qquad \times 1 \begin{bmatrix} h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \in \mathbb{Z}_{2r+1}, \\ h_\Xi^{\mathtt{hsh}}(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^{\kappa+1} (r+\ell+1) \end{bmatrix} \Big] \end{array} \right) \\
&\overset{\star}{\doteq} \sum\nolimits_{\ell,\kappa} \mathbb{E}\big[h_\Xi^{\mathtt{hsh}}(G(\mathbf{1}_w)) \cdot \prod\nolimits_{i \in w} (-1)^{G_i(1)} \cdot 1[h_\Xi^{\mathtt{hsh}}(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^\kappa (r-\ell)]\big] + \tilde{\beta}, \\
&|\tilde{\beta}| \leq \sum\nolimits_{\ell,\kappa} (2r+1) \cdot \tfrac{2^k \cdot \epsilon\beta}{2^k (s\gamma_{\mathtt{sm}})^2} \cdot \tfrac{\mathsf{Pr}\big[\Xi \in \bigsqcup_{\kappa=0}^{1} [(-1)^\kappa r - s\gamma_{\mathtt{sm}}, (-1)^\kappa r + s\gamma_{\mathtt{sm}}]\big]}{\mathsf{Pr}\big[\Xi \in [-r + s\gamma_{\mathtt{sm}}, r - s\gamma_{\mathtt{sm}}]\big]} \leq \tfrac{(2r+1)\epsilon\beta}{(s\gamma_{\mathtt{sm}})^2} \cdot \tfrac{2s\gamma_{\mathtt{sm}}}{2r - 2s\gamma_{\mathtt{sm}}} \overset{\star}{\lll} \beta.
\end{aligned}$$

**Inverting the actual data:** Since the statistical distance between the ideal shift $G$ and the actual one over $\mathcal{D}_\Xi$ is bounded by $\beta\delta_{7.11}/r$, Theorem 7.12's $(\alpha, \beta)$-inversion has demonstrated $|\mathbf{corr}_w(\mathcal{D}_\Xi) - \theta_v(\mathbf{1}_w)| \ll \beta$ under the full collision $\mathcal{D}_\Xi \subset \mathcal{C}$ with significance $O(\delta_{7.11})$. The partial one on the actural hash may add an extra accuracy cost $\epsilon\beta$ to derive $|\mathbf{corr}_w(\mathcal{C} \cap \mathcal{D}_\Xi) - \theta_w(\mathbf{1}_w)| \ll \beta$, since the actural one may deviate from the ideal $h_J^{\mathtt{hsh}}(G_{w^c})$ by statistical distance $O(\beta\delta_{7.11}/r)$.

**Inverting the almost-zero secret:** When the small secret parameter $\theta \in \mathbb{Z}_s^n$ is virtually zero as $|\{i \in (d) \mid \theta_i \neq 0\}| \leq k$, add $\mathbf{1}_w$ to $w \subset \{i : \theta_i = 0\}$, and replace each data $(x(j), y(j))$ with $(x(j), y(j) + \sum_{i \in w} \hat{f}_i \lfloor x_i(j)/2 \rfloor (-1)^{x_i(j)})$ for inverting $\theta + \mathbf{1}_w$. $\square$

**Theorem 7.18** (GapSVP to LWE [Pei09, Reg09])**.** Let $n \geq 1$ and $q \geq 2^{n/2}$ be integers, and let $0 < \alpha < 1$ be such that $\alpha q \geq 2\sqrt{n}$. The worst-case $\mathrm{GapSVP}_{\tilde{O}(n/\alpha)}$ is reducible to $\mathrm{LWE}_{n,q,\alpha}$.

**Theorem 7.19** (search-to-decision for LWE [MP12])**.** Let $q$ be a power of 2, and $\alpha$ satisfy $1/q < \alpha < 1/\omega(\sqrt{\log n})$. Then, $\mathrm{LWE}_{n,q,\alpha}$ reduces to decision $\mathrm{LWE}_{n,q,\alpha'}$ for $\alpha' = \alpha \cdot \omega(\log n)$.

**Theorem 7.20** (LWE to binary LWE [BLP+13])**.** Let $n, q, q' \geq 1$, $m \geq n' \geq 1$ be integers, where $q$ is a power of 2. Let $\alpha, \beta, \delta > 0$ and $0 < \varepsilon, \xi \ll 1$ be $n' \geq (n+1)\log q + 2\log(1/\delta)$, $\alpha \geq \sqrt{\ln(2n(1+1/\varepsilon))/\pi}/q$, $\beta = (10n'\alpha^2 + \frac{4n'}{\pi q'^2}\ln(2n'(1+1/\xi)))^{1/2}$. As decision problems, $\mathrm{LWE}_{n,m,q,\alpha}$ is reducible to $\mathrm{LWE}_{n',m,q',\beta}$ with the binary secret such that any $\zeta$-advantageous algorithm of the latter problem produces that of the former one with an advantage $\frac{\zeta-\delta}{3m} - \frac{41\varepsilon}{2} - 14\xi$.

**Theorem 7.21** (Theorem 1.7)**.** $\mathrm{GapSVP}_{\tilde{O}(n^2)}$ is solvable in probabilistic polynomial time.

*Proof.* Take $q = 2^{n/2}$, $q' = p = O(n\sqrt{\log n})$, $n' = (n+1)n/2 + 2\log(1/\delta)$, $\alpha = 1/\omega(\log n)$, $\alpha' = \epsilon/n$, and $\beta = (10n'\alpha'^2 + \frac{4n'}{\pi p^2}\ln(2n'(1+1/\xi)))^{1/2} \ll 1$. Theorem 7.18 reduces $\mathrm{GapSVP}_{\tilde{O}(n/\alpha)}$ to $\mathrm{LWE}_{n,q,\alpha}$, Theorem 7.19 reduces it to decision-$\mathrm{LWE}_{n,q,\alpha'}$, and Theorem 7.20 to decision-$\mathrm{LWE}_{n',p,\beta}$ with the binary secret. So, Theorem 7.17 inverts (search) $\mathrm{LWE}_{n',p,\beta}$ in poly-time. $\square$

# 8 Natural Lower Bounds of Matrix rigidity

This section will establish circuit lower bounds in Theorems 1.8–1.10. They apply Theorem 7.15's linear Fourier inversion to learn all sparse $\sqrt{N}$ by $\sqrt{N}$ matrices $\mathcal{M}$ having low $\mathcal{F}$-complexity of arguing circuit classes $\mathcal{F}$ in a smoothed analysis. Let $G \in \{0,1\}^N$ be any $\beta\delta_{7.11}$-away $2k_0$-independent flipper, $\Phi$ be Definition 2.15's shift, and $\tilde{\mathcal{M}}(z) := \mathcal{M}(\Phi(z))(-1)^{G(\Phi(z))}$, $z = (x, y)$.

**Definition 8.1** (learning sparse matrices in smoothed analysis)**.** Learning an $\sqrt{N}$ by $\sqrt{N}$ matrix $\mathcal{M}$ of density $|\mathcal{M}|_{\neq 0}/N$ under a shift $(G, \Phi)$ asks a learner $\mathcal{A}$ to choose rows and columns $\mathcal{I} \subset [N]$ and $\mathcal{J} \subset (N]$ to access to $\tilde{\mathcal{M}}(x, y)$ of $(x, y) \in \mathcal{I} \times (N] \sqcup [N) \times \mathcal{J}$ and predicts

$$\textit{c}\varepsilon\textit{-learning:}\ \mathsf{Pr}_G\big[\mathsf{Pr}_{X,Y}[\mathcal{A}\big(X, Y \mid \tilde{\mathcal{M}}(\mathcal{I} \times (N] \sqcup [N) \times \mathcal{J})\big) \neq \tilde{\mathcal{M}}(X,Y)] \leq c\varepsilon\big] \geq 1 - O(\delta).$$

## 8.1 Unrestricted Super-Linear Lower Bounds

An $\mathbb{F}$-linear circuit is an $n$-input $n$-output circuit computing an $\mathbb{F}$-linear form $f = \sum_i f_i g_i$ at each gate $f$ feeding the in-coming edges labeled by $f_i \in \mathbb{F}$ from the child gates $g_i$. We call it reversible [SR11, ZW17] if reversing and relabeling the edges produce a circuit computing the same linear form at every gate.

**Lemma 8.2** (reversibility)**.** Any binary $\mathbb{F}$-linear circuit computing a reversible matrix can transform to a reversible one without changing the size and depth.

*Proof.* By an induction starting from an output (fan-out 0) node. An obtained reversible circuit consists of the $n$ lines connecting the $n$ inputs to $n$ outputs having reversible $s$-input $s$-output Fredkin gates of varying $s$ per gate. Every output node is a root node of the uniquely determined maximum sub-tree having non-leaf nodes of fan-out one and leaf nodes of fan-out greater than one, excepting at most one leaf node. If the output node $o$ entails a binary tree of size $s$ computing $o = \sum_{i \in (s]} o_i g_i$ from the $s$ leaves computing $g_i$, do the following. Remove this size-$s$ subtree below $o$, take an arbitrary $g_i$ with $o_i \neq 0$, and put a new $s$-input $s$-output reversible Fredkin gate of size $s$ computing $g_i \sqcup \{g_{i'}\}_{i' \in (s]-\{i\}} \mapsto o \sqcup \{g_{i'}\}_{i' \in (s]-\{i\}}$. The $g_i$ might be an input (fan-in 0) node of fan-out one, say $\mathbf{x}_i$. There is no more than one leaf node having one fan-out due to the matrix's reversibility. This case connects $\mathbf{x}_i \to o$ by a line and proceeds to an induction step on the remainder $(n-1)$-input $(n-1)$-output circuit. In the other case, the induction step takes the $n$ output gates $g_i \sqcup \{o' \mid o' \neq o\}$ to form an $n$ by $n$ reversible matrix. $\square$

**Definition 8.3** (matrix rigidity)**.** The rigidity $\mathbf{rig}_{\mathcal{M}}(r)$ of a matrix $\mathcal{M} \in \mathbb{F}^{n\times n}$ is the minimum number of flipping entries on each row to reduce its rank to $r$:

$$\textit{Matrix rigidity: } \mathbf{rig}_{\mathcal{M}}(r) = \min\big\{\max_x |\mathcal{N}_x|_{\neq 0} \mid \mathbf{rank}(\mathcal{M}+\mathcal{N}) \leq r\big\}.$$

**Theorem 8.4** (Valiant)**.** Any matrix $\mathcal{M} \in \mathbb{F}^{n\times n}$ computable by an $\mathbb{F}$-linear circuit of fanin two, node-size $s$, and depth $d$ (a power of 2) must have rigidity $\forall t, \mathrm{rig}_{\mathcal{M}}(\frac{t}{\log d}s) \leq 2^{d_t}$ for $d_t = 2^{-t}d$. Further, for $d_{t,u} = (1-2^{-t})^u d$, truncating $\frac{tu}{\log d}s$ to their tail nodes computing the $\mathbb{F}$-linear forms forces the circuit to have depth $\max(d_t, d_{t,u})$.

*Proof.* Let $\mathcal{C} = (\mathcal{V}, \mathcal{E})$ be an arbitrary binary circuit of node-depth $\psi : \mathcal{V} \to [d]$. Cut all those nodes $v$ such that $\psi(u) < \psi(v)$ of the child nodes $u$ of $v$ differs at the $i$th bit for the most significant $i \in [\log d)$. Take those $t$ bits and fix them to bound the cut edges by at most $r \leq \frac{t}{\log d}s$. The truncated circuit has depth at most $d_t$, so every node is reachable from $2^{d_t}$ or fewer input nodes. Any input-output path passing through none of these edges must increase the accompanying node depths within $2^{-t}d$ bit patterns. As a dual, any input-output path passing some of these nodes must progress them whthin the remaining $(1-2^{-t})d$ patterns. Repeating it for $u$ times reduces it to $(1-2^{-t})^u d$ of any path through the cut edges. $\square$

**Theorem 8.5** (formulas to partial derivatives [BS83])**.** Any algebraic fanin-2 circuit of size $s$ and depth $d$ to compute a linear $\mathbf{y}$-degree polynomial $f(\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{y}_1, \ldots, \mathbf{y}_n) = \sum_{i=1}^n \frac{\partial f}{\partial \mathbf{y}_i} \cdot \mathbf{y}_i$ induces a multi-output parallel algebraic circuit of size $2s$ and depth $2d$ computing all partial derivatives $\big(\frac{\partial f}{\partial \mathbf{y}_i}(\mathbf{x})\big)_{i=1}^n$.

**Theorem 8.6** (Theorem 1.8[53])**.** Let $N = n^2 = 2^{\log k}2^{\log(N/k)}$ by even integers $\log k$ and $\log N/k$. Let $p$ be an odd prime coprime with $\binom{2(\ell-1)}{\ell-1}$, $(r_0, v, |\mathcal{Q}_{\mathtt{rem}}|) = (1, 3, p)$, $k = 3(2\ell-1)$, $\alpha = k!/k^k$, $\beta \ll 1$, $d = \Theta(\log n)$, $d_t = d/2^t$, $d_{t,u} = d(1-2^{-t})^u$, $r \leq stu/\log d$, $\delta_{7.11} = \delta/r$, and $\eta \approx 2^{2d_{t,u}+d_t}/n$. Suppose $m_{7.12} \ll \min(p^\ell, \alpha\mu^k n)$, $k \ll k_0 \approx 4^{d_{t,u}}\mu$, and $2^{2k}\eta/\beta \leq \delta_{7.11}$. Any $n$ by $n$ $\{-1,0,1\}$-matrix $\mathcal{M}$ of density $\mu$ must refute any $\mathbb{F}_p$-linear circuit of size $s$ and depth $d$ computing $\tilde{M}$ unless each row of $\tilde{M}$ is $\eta$-learnable from some $4^{d_{t,u}}$ of the first $r$ rows, and the first $m = O(m_{7.12}/(\alpha\mu^k))$ columns, in $O((\binom{k_0}{k} + k_0 r)m)$ time.

*Proof.* **Planted linear FT from matrix rigidity:** Theorem 8.4 obliges that the shifted matrix $\tilde{\mathcal{M}}(x,y) = \mathcal{M}(\Phi(x,y))(-1)^{G(\Phi(x,y))}$ realized by any $\mathbb{F}_p$-linear circuit of size $s$ and depth $d$ must have $\mathrm{rig}_{\tilde{\mathcal{M}}}(r) \leq 2^{d_t}$. In addition, a permutation matrix[54] $\mathcal{N}' \in \{-1,0,1\}^{n\times n}$ makes $\tilde{\mathcal{M}} + \mathcal{N}'$ reversible. Theorem 8.4 presents an $r$ by $n$ matrix $\mathcal{B}$ consisting of the $r$ linear forms computed by the cut edges. An $n$ by $r$ matrix $\mathcal{A}$ calculates the output matrix $\mathcal{AB} = \tilde{\mathcal{M}} + \mathcal{N}''$ with noise $\forall i, |\mathcal{N}''_x|_{\neq 0} \leq 2^{d_t} + 1$ . Lemma 8.2's reversible circuit connects each output node to at most $2^{d_{t,u}}$ edges in $B$. It deduces $\mathcal{A}^{-1}\mathcal{A} = \mathbf{1}_{r\times r}$ by $\forall i, \max(|\mathcal{A}_x|_{\neq 0}, |(\mathcal{A}^{-1})^x|_{\neq 0}) \leq 2^{d_{t,u}}$. $(\mathcal{AA}^{-1})^{\mathcal{I}}_{\mathcal{I}} = \mathcal{A}_{\mathcal{I}}(\mathcal{A}^{-1})^{\mathcal{I}} = \mathbf{1}_{r\times r}$ for any index set [55] $\mathcal{I} \in \binom{n}{r}$ of non-degenerate $A_{\mathcal{I}}$, producing $(\mathcal{AA}^{-1})^{\mathcal{I}}\mathcal{A}_{\mathcal{I}}\mathcal{B} = \tilde{\mathcal{M}} + \mathcal{N}''$ with $\mathcal{A}_{\mathcal{I}}\mathcal{B} = \tilde{\mathcal{M}}_{\mathcal{I}} + \mathcal{N}''_{\mathcal{I}}$ and $\forall x, |(\mathcal{AA}^{-1})^{\mathcal{I}}_x|_{\neq 0} \leq 4^{d_{t,u}}$. Thus, $\mathcal{N} := \mathcal{N}'' - (\mathcal{AA}^{-1})^{\mathcal{I}}\mathcal{N}''_{\mathcal{I}}$ with $\forall x, |\mathcal{N}_x| \leq 4^{d_{t,u}}(2^{d_t}+1) \approx \eta n$ brings out Definition 8.3's matrix rigidity to invert the hidden $\theta_{x_0} = (\mathcal{AA}^{-1})^{\mathcal{I}}_{x_0} \in \mathbb{F}_p^r$ of the following planted FT to invert the $x_0$th row:

---

[53] Take $k \ll k_0 \leq O(1)$, $\alpha \approx \frac{\sqrt{2\pi k}}{\mathbf{e}^k}$, $t = \frac{\epsilon}{4}\log\log\log n$, $u = (\log\log n)^{\frac{\epsilon}{2}}$, $p \gg n^{6/k}$, $r = O(\frac{n}{(\log\log n)^{\epsilon/2}})$, $s = n(\log\log n)^{1-\epsilon}$, and $\mu \approx \frac{k_0}{4^{d_{t,u}}}$. Lemma 2.13 provides an explicit $O(2^{-2d_{t,u}-d_t})$-away $2k_0$-independent flipper $|G| = O\big((2^{2d_{t,u}+d_t}\log n)^2\big)$. Lemma 2.17 gives an explicit DFT-shift $|\Phi| = n^{O(1)}$.

[54] Permutation matrics must have (at most) one non-zero entry in every row and column.

[55] $\mathbf{1}_{r\times r}$ is the $r$ by $r$ identity matrix.

*Matrix rigidity:* $(\mathcal{A}\mathcal{A}^{-1})^{\mathcal{I}}\tilde{\mathcal{M}}_{\mathcal{I}} = \tilde{\mathcal{M}} + \mathcal{N}$. Let $\mathcal{I}_{x_0} = \{x \in \mathcal{I} \mid \theta_{x_0}(x) \neq 0\} \subset \mathcal{I}$.

*Training dataset:* $\mathcal{D}_{x_0} := \{\big(\big(2\#x + \frac{1-\tilde{\mathcal{M}}(x,y)}{2} \mid x \in \mathcal{I}_{x_0}, \tilde{\mathcal{M}}(x,y) \neq 0\big), \tilde{\mathcal{M}}(x_0,y)\big)\}_{y\in\mathcal{J}}$, where $|\mathcal{J}| = m$ and $x$ is the $(1+\#x)$th smallest number in $(\mathcal{I}_{x_0}]$.

*Planted linear FT:* $f(\mathbf{x}) := \sum_{i=1}^{k_0} \theta_{x_0}(\lfloor \mathbf{x}_i/2 \rfloor)(-1)^{\mathbf{x}_i}$.

**Inverting linear planted FT:** Theorem 7.15 can invert this $f_{\theta_x}(\mathbf{x})$ on $|\mathcal{I}_{x_0}| \leq 4^{d_{t,u}}$, $k_0 \approx \mu|\mathcal{I}_{x_0}|$, $\mu_{\mathtt{hsh}} \approx \mu^k$, and $m \ll n$. Lemma 2.17's DFT-shift makes the i.i.d. $m$ samples $\mathcal{J} \subset (n]$ to have:

*Uniform density:* $\forall x_0 \in (n], \mathsf{Pr}_{Y\in\mathcal{J}}\big[\big||\mathcal{I}_{x_0} \cap \tilde{\mathcal{M}}^Y_{\neq 0}| - \mu|\mathcal{I}_{x_0}|\big| \ll \mu|\mathcal{I}_{x_0}|\big] \approx 1$.

*k-cover:* $\forall x_0 \in (n], \forall \mathcal{K} \in \binom{\mathcal{I}_{x_0}}{k}, \big|\mathsf{Pr}_{Y\in\mathcal{J}}[(\mathcal{K}, Y) \subset \tilde{\mathcal{M}}_{\neq 0}] - \mu_{\mathtt{cvr}}\big| \ll \mu_{\mathtt{cvr}}$.

*Column-wise error:* $\forall x_0 \in (n], \mathsf{Pr}_{Y\in\mathcal{J}}\big[x \in \mathcal{I}_{x_0} \Rightarrow \mathcal{N}(x,Y) = 0\big] \geq 1 - \eta\mu/\delta \cdot |\mathcal{I}_{x_0}|$ ($\because$ Markov's ineq of $\gamma = \delta$).

Chernoff bounds of $\gamma = \frac{\epsilon}{1-\epsilon}$ guarantees them with significance $n \cdot \binom{r}{k} \cdot \boldsymbol{e}^{-\gamma^2/2\cdot(1-\epsilon)m} < o(\frac{1}{n})$. $G$'s $\beta\delta_{7.11}$-away $2k_0$-independence implies its $(h^{\mathtt{rem}}_w(G), 0)$-hashed $\beta\delta_{7.11}$-away $2k$-independence. Theorem 7.15 inverts[56] the $f(\mathbf{x})$ and predicts $\tilde{\mathcal{M}}_{x_0}$ from the entries over $\mathcal{I}_{x_0} \times (n] \cup [n) \times \mathcal{J}$. $\square$

## 8.2 Lower bounds beyond $\mathsf{PH^{cc}}$

Tarui [BFS86, Tod91, Tar93] presented low-degree probabilistic polynomials to approximate $\mathsf{PH^{cc}}$ languages with a Boolean guarantee. Razborov [Raz89, Tod91, Wun12] transformed them into rigid matrices with two-sided error.

**Theorem 8.7** (probabilistic polynomials with Boolean guarantee)**.** Let $d = \sum_{\kappa=1}^h d_\kappa$, $d_{8.7} := 2de+h$ and $s_{8.7} = \prod_{\kappa=1}^h (1+2^{d_\kappa})^{2d_\kappa e+1}$. Suppose $\mathcal{L} \in \mathsf{PH}^{\mathsf{cc}}_h[d]$ has the same type of gates at depths $(d_{\kappa-1}, d_\kappa]$. It admits a low-degree linear computation $\forall (x,y) \in \{0,1\}^{n/2} \times \{0,1\}^{n/2}, \mathsf{Pr}_R[\mathcal{L}(x,y) \neq \phi_R(x,y)] \approx 0$ under the random seed $R \in \{0,1\}^{e\sum_{\kappa=1}^h 2^{d_\kappa}}$.

*Linear expression by lift and project:* $\phi_R(x,y) = \sum_{w\in\binom{n}{d_{8.7}}} \hat{\phi}_{w,R}(1[x \in \mathcal{I}_w, y \in \mathcal{J}_w])$ by $\mathcal{I}_w, \mathcal{J}_w \subset \{0,1\}^{n/2}$, and $\hat{\phi}_{w,R} \in \mathbb{Z}$ with $\sum_{w\in\binom{n}{d_{8.7}}} |\hat{\phi}_{w,R}| \leq s_{8.7}$.

*Point-wise error:* $\forall (x,y), \mathsf{Pr}_R[\mathcal{L}(x,y) \neq \phi_R(x,y)] \leq 1/2^e$.

*Boolean guarantee:* $\forall (x,y), \phi_R(x,y) \in \{0,1\} \Rightarrow \phi_R(x,y) = \mathcal{L}(x,y)$.

*Proof.* Replace binary NOR-subtrees of the $\mathsf{PH}^{\mathsf{cc}}_h[d]$ computation with Tarui's probabilistic polynomials [Tar93]. At the $\kappa$th layer of $\mathsf{PH}^{\mathsf{cc}}_h[d]$, it uses $e\sum_{\kappa=1}^h 2^{d_\kappa}$ number of the i.i.d. coin flips $R_{i,j,\ell} \in \{0,1\}$ of bias $\mathbb{E}[R_{i,j,\ell}] = 1/2^j$ and transforms a depth-$d_\kappa$ NOR-subtree to

$$\textit{Probabilistic polynomial: } \mathrm{NOR}_R(g_1, \ldots, g_{2^{d_\kappa}}) = (1 + \textstyle\sum_{i=1}^{2^{d_\kappa}} g_i)\prod_{\ell=1}^e \prod_{j=1}^{d_\kappa} (1 - \sum_{i=1}^{2^{d_\kappa}} R_{i,j,\ell} g_i)^2.$$

It satisfies the Boolean guarantee, i.e., $\mathrm{NOR}_R(g_1, \ldots, g_{2^{d_\kappa}}) = 1 \Rightarrow (1 + \sum_{i\in[2^{d_\kappa}]} g_i) = 1 \Rightarrow \forall g_i = 1$ and $\mathrm{NOR}_R(g_1, \ldots, g_{2^{d_\kappa}}) = 0 \Rightarrow \exists (1 - \sum_{i=1}^{2^{d_\kappa}} R_{i,j,\ell} g_i) = 0 \Rightarrow \exists g_i = 1$. This replacement gives a two-sided error computation. It incurs an error at most $1/2^{d+e}$ for each of the $2^d$ NOR gates, owing no more than $1/2^{d+e} \cdot 2^d = 2^{-e}$ error in total. It expands into a hierarchy of $(2d_\kappa e+1)$-degree and $(1 + 2^{d_\kappa})^{2d_\kappa e+1}$-norm polynomials at the $\kappa$th layer, yielding the claimed linear expression. $\square$

[56] Theorem 7.17's proof takes care of the almost-zero $\theta$'s case.

**Theorem 8.8** (Theorem 1.9[57]). Let $N = 2^n = 2^{\log k} 2^{\log(N/k)}$ by even integers $\log(N/k)$ and $\log k$. Let $p$ be an odd prime coprime with $\binom{2(\ell-1)}{\ell-1}$, $(r_0, v, |\mathcal{Q}_{\mathtt{rem}}|) = (1, 3, p)$, $k = 3(2\ell - 1)$, $\alpha = k!/k^k$, $\beta \ll 1$, $r \le \binom{n}{d_{8.7}}$, $\delta_{7.11} = \delta/r$, $k \ll k_0 \approx r\mu$, and $\eta \approx \frac{k_0}{\delta 2^{e_0}}$. Suppose $m_{7.12} \ll \min(p^\ell, \alpha\mu^k\sqrt{N})$, and $2^{2k}\eta/\beta \le \delta_{7.11} \ll \beta$. Any $\sqrt{N}$ by $\sqrt{N}$ $\{-1, 0, 1\}$-matrix $\mathcal{M}$ of density $\mu$ must have lower bounds $\tilde{\mathcal{M}}^{-1}(b) \notin \mathsf{PH}^{\mathsf{cc}}_h[d]$ for some $b \in \{1, -1\}$ unless $\tilde{\mathcal{M}}$ is $\eta$-learnable from $\mathcal{M}$'s $r$ rows and $m = O(\frac{pm_{7.12}}{\alpha\delta\mu^k})$ columns in $O((\binom{k_0}{k} + k_0 r)m)$ time.

*Proof.* Follow Theorem 8.6's one. Suppose $\tilde{\mathcal{M}}^{-1}(b) \in \mathsf{PH}^{\mathsf{cc}}_h[d]$ for both $b = 1, -1$. Theorem 8.7's probabilistic polynomials $\phi_{b,R}$ approximate $\tilde{\mathcal{M}}^{-1}(b) \in \mathsf{PH}^{\mathsf{cc}}_h[d]$ by point-wise noise rate no larger than $1/2^{e_0}$, providing the linear planted FT to make $\tilde{\mathcal{M}}$ learnable. Theorem 8.6's matrix rigidity argument transforms Theorem 8.7's linear expression into:

*Matrix rigidity*: $(\mathcal{A}\mathcal{A}^{-1})^{\mathcal{I}}(\tilde{\mathcal{M}}_{\mathcal{I}} + \mathcal{N}_{\mathcal{I}}) = \tilde{\mathcal{M}} + \mathcal{N}$ by $\mathcal{A} \in \mathbb{F}_p^{\sqrt{N}\times r}$ and $\mathcal{N} \in \mathbb{F}_p^{\sqrt{N}\times\sqrt{N}}$ on $\mathcal{I} \in \binom{\sqrt{N}}{r}$.

*Point-wise error*: $\forall(x, y), \Pr[\mathcal{N}(x, y) \neq 0] \le \Pr_R[\exists b, \phi_{b,R}(x, y) \notin \{0, 1\}] \le 2/2^{e_0}$.

*Column-wise error*: $\Pr_Y[\forall x \in \mathcal{I}, \tilde{\mathcal{M}}(x, Y) \neq 0 \Rightarrow \mathcal{N}(x, Y) = 0 \mid \sum_{x\in\mathcal{I}\setminus\tilde{\mathcal{M}}^Y_{\neq 0}}(\tilde{\mathcal{M}} + \mathcal{N})(x, Y) \bmod p] \ge 1 - \frac{2k_0}{2^{e_0}\delta}$.

Theorem 8.6 has succeeded in learning $\tilde{\mathcal{M}}$ from the first $m$-columns of $(\tilde{\mathcal{M}}_{\mathcal{I}} + \mathcal{N}_{\mathcal{I}})$ satisfying the uniform-density and $k$-cover under a fixed reminder $\sum_{x\in\mathcal{I}\setminus\tilde{\mathcal{M}}^Y_{\neq 0}}(\tilde{\mathcal{M}} + \mathcal{N})(x, Y) = \xi \in \mathbb{Z}_p$. Markov's inequality parameter $\gamma = \delta$ provides $m\gamma/p = O(\frac{m_{7.12}}{\alpha\mu^k})$ data enough to execute it. □

## 8.3 PH$^{\text{cc}}$ $\neq$ PSPACE$^{\text{cc}}$ or quasi-NP $\not\subset$ quasi-NC$^k$

As mentioned in Theorem Review11, Williams's algorithmic approach [Wil13] has established $\mathsf{NEXP} \not\subset \mathsf{ACC}$ [Wil14a] and even $\mathsf{quasi\text{-}NP} \not\subset \mathsf{ACC}$ [MW19]. This section will further extend it to prove Theorem 1.10 in virtue of Theorem 7.15's linear Fourier inversion.

**Theorem 8.9** ($\mathsf{NP}$ hierarchy [Coo72, SFM78, Žák83] ). There is a unary language $\mathcal{L} \subset \{1\}^*$ separating $\mathcal{L} \in \mathsf{NTIME}[t] - \bigcup_{t'}\{\mathsf{NTIME}[t'] \mid t'(n + 1) = o(t(n))\}$.

**Theorem 8.10** (short PCP [BSV14]). Every $t$-time verifier algorithm $\mathcal{A}(x, w)$ inputting $1^n$ and a witness $w$ can induce an $(n + \log t)^{O(1)}$-time computable generator $1^n \mapsto \mathcal{C}^{\mathcal{O}}_n$ of poly$(n + \log t)$ size circuit with an oracle $\mathcal{O}_n$ of $n = \log t + O(\log\log t)$ input bits. If $\exists w, \mathcal{A}(1^n, w) = \mathtt{accept}$ then $\exists\mathcal{O}, \mathcal{C}^{\mathcal{O}}_n$ is unsatisfiable. If $\forall w, \mathcal{A}(1^n, w) = \mathtt{reject}$, then $\forall\mathcal{O}, \mathcal{C}^{\mathcal{O}}_n$ has at least $(1 - \frac{1}{n})$ satisfying assignments. $\mathcal{C}_\nu$ is a 3CNF of the $n^{O(1)}$ inputs of the $\mathcal{O}$'s answers.

**Theorem 8.11** (easy witness lemma [MW19]). There is a universal constant $c_{8.11} > 0$ such that if $\mathsf{NTIME}[t^{c_{8.11}}] \subset \mathsf{SIZE}[n^\ell]$, then every language in $\mathsf{NTIME}[t]$ must have a witness of $\mathsf{SIZE}[n^{O(\ell^3)}]$.

**Theorem 8.12.** $\exists e_0, \forall h, \forall\ell, \forall d \ll n/e_0$, $\mathsf{P} \not\subset \mathsf{PH}^{\mathsf{cc}}_h[d]$ or $\mathsf{NTIME}[2^{\epsilon n}] \not\subset \mathsf{SIZE}[n^\ell]$ .

*Proof.* Let $N = t = O(2^{\epsilon d e_0 h/c_{8.11}})$, $n = \log t + O(\log\log t)$, $r \le \binom{n}{2de_0+h}$, $\mu \approx k_0/r$, and $p^e = N^{O(1)}$. Suppose $\mathsf{NTIME}[t^{c_{8.11}}] \subset \mathsf{SIZE}[n^\ell]$ and $\mathsf{P} \subset \mathsf{PH}^{\mathsf{cc}}_h[d]$ for a contradiction.

**Witnessing small circuits:** Theorem 8.9 presents $\mathcal{L} \in \mathsf{NTIME}[t] - \mathsf{NTIME}[\ t^{1-o(1)}]$, a unary language separating non-deterministic time hierarchy. Theorem 8.10's short PCP transfers its

[57]Take $\frac{e_0 k_0 hp}{\alpha\beta\delta} = O(1)$ and $d = n^\epsilon$. Lemma 2.13 gives an explicit $O(1)$-away $2k_0$-independent $N$-bit flipper of cardinality $O(n^2)$. Lemma 2.17 provides the DFT-shift of cardinality $N^{O(1)}$.

$t$-time verifier to an $n^{O(1)}$ non-deterministic time algorithm generating a circuit $1^n \mapsto \mathcal{C}_n^{\mathcal{O}_n}$ of size $n^{O(1)}$. Since we have assumed $\mathsf{NTIME}[t^{c_{8.11}}] \subset \mathsf{SIZE}[n^\ell]$, the easy-witness lemma (Theorem 8.11) replaces its oracle $\mathcal{O}_n$ with a witness circuit $\mathcal{W}_n$ of size $n^{O(\ell^3)}$, yielding a circuit $\mathcal{C}_n^{\mathcal{W}_n}$ of size $n^{O(1)}$. Let $1^n \mapsto \mathcal{M}$ be the 0-1 truth table of $\neg\mathcal{C}_n^{\mathcal{W}_n} = 1 - \mathcal{C}_n^{\mathcal{W}_n}$ arranged in a $\sqrt{N} \times \sqrt{N}$ $\{0,1\}$-matrix. If $1^n \in \mathcal{L}$, then all of its entries are one, while if $1^n \notin \mathcal{L}$, it has at most $\frac{1}{n}$ fraction of the one entries. Further, we reduce $\mathcal{M}$'s one-entries to make $|\mathcal{M}_U|_{\neq 0}/N = \mu$ when $|M|_{\neq 0}/N = 1$, and $\mathbb{E}[|\mathcal{M}_{\mathcal{I}}|_{\neq 0}/N] \leq \frac{\mu}{n}$ when $|\mathcal{M}|_{\neq 0}/N \leq \frac{1}{n}$, say by drawing the uniform random $U \sim [\frac{1}{\mu}]$ and forcing $\mathcal{M}_U = 0$ if $(x,y) \not\equiv U \bmod 1/\mu$. Let $\tilde{\mathcal{M}}(x,y) := \mathcal{M}_U(\varPhi(x,y)) \cdot (-1)^{G(\varPhi(x,y))}$.

**Learning small circuits:** The witness circuit $\mathcal{W}_n$ is evaluable in deterministic $n^{O(1)}$ time. Our assumption $\mathsf{P} \subset \mathsf{PH}_h^{\mathsf{cc}}[d]$ gives it a $\mathsf{PH}_h^{\mathsf{cc}}[d]$ protocol. Theorem 8.7 presents a linear expression $\phi_R(x,y)$ to approximate $\mathcal{W}_n(\varPhi(x,y))$. Guess the $\phi_R(x,y)$ by $\log(n^{O(\ell^3)})$ bits for the circuit $\mathcal{W}_n$ and $2n \cdot \sqrt{N}$ bits for its $2n$ input rectangles $1[x \in \mathcal{I}_i]$ and $1[y \in \mathcal{J}_j]$ of all $i \in (n]$ so that $1[x \in \mathcal{I}_w, y \in \mathcal{J}_w] = \prod_{i \in w \cap (0,n/2]} \prod_{j \in w \cap (n/2,n]} 1[x \in \mathcal{I}_i] 1[y \in \mathcal{J}_j]$. Since the short PCP's $\mathcal{C}_n$ is a 3CNF, Theorem 8.7's probabilistic polynomial (i.e., a NOR gate inputting $g_i = \phi_R(x,y(i))\phi_R(x'(i),y'(i))\phi_R(x''(i),y''(i))$) transforms $\tilde{\mathcal{M}}(x,y)$ to Theorem 8.8's matrix rigidity $\tilde{\mathcal{M}}' := \tilde{\mathcal{M}} + \mathcal{N}$. It induces Theorem 8.6's planted FT $f(\mathbf{x})$ that Theorem 7.15 can invert from the row data $\tilde{\mathcal{M}}'_{\mathcal{I}}$ under a guessed flipper $G$. Let $\mathcal{I} \in \binom{\sqrt{N}}{r}$ be Theorem 8.8's matrix-rigidity index set. Guess a permuter $\varPhi$ to satisfy

$$\textit{uniform density:} \quad \mathsf{Pr}_Y\left[\left||\{x \in \mathcal{I} \mid \tilde{\mathcal{M}}(x,Y) \neq 0\}| - \mu r\right| \ll \mu r\right] \approx 1.$$

$$\textit{k-cover:} \quad \forall \mathcal{K} \in \tbinom{\mathcal{I}}{k}, \left|\mathsf{Pr}_Y[(\mathcal{K},Y) \subset \tilde{M}_{\neq 0}] - \mu_{\mathtt{cvr}}\right| \ll \mu_{\mathtt{cvr}}.$$

Theorem 8.8 has shown it to meet the small column-wise error, too. Let $\{x \in \mathcal{I} \mid \tilde{\mathcal{M}}'(x,y) \neq 0\} \subset \mathcal{I}(y) \subset \mathcal{I}$ with $|\mathcal{I}(y)| = k_0$. For every $x \in [\sqrt{N})$ and $a \in \{-1,0,1\}^{k_0}$,

$$\begin{array}{l}\textit{Acceptance}\\ \textit{probability:}\\ \textit{estimation}\end{array} \left|\begin{array}{c}\frac{1}{N}\sum_x \sum_{(\mathcal{I}(y),a)} \mathbb{E}_{U,G}\left[\left|\{y \mid \tilde{M}(\mathcal{I}(y),y) = a\}\right| \cdot 1[f(x(y)) \neq 0]\right]\\ -(1 \pm O(1/2^{e_0}))\mathsf{Pr}[\tilde{\mathcal{M}}(X,Y) \neq 0]\end{array}\right| \geq 1 - O(\delta).$$

It is a consequence of Theorem 8.8 to learn $\tilde{\mathcal{M}}'$ via the probabilistic polynomial $\phi_R(x,y)$ of a guess $R$ to incur an error rate $|\mathcal{N}|_{\neq 0}/N \leq 1/2^{e_0}$. It recognizes $\mathcal{L}$ by accepting $1^n$ if the result is at least $\mu/2$, and rejecting it otherwise. It runs in the following non-deterministic time of the parameters $\frac{k_0 p}{\alpha\beta\delta} = O(1)$, $r = n^{O(\log^c n)}$ and $t = 2^n$, contradicting $\mathcal{L} \notin \mathsf{NTIME}[t^{1-o(1)}]$:

$$\begin{aligned} &|\mathcal{U}| \times (\textit{Matrix entries calculation time} + \textit{Fourier inversion time} + \textit{Acceptance probability estimation time}) \\ &= |[1/\mu]| \cdot \left(\sqrt{N}|\mathcal{I}| \cdot \tilde{O}(n) + |\mathcal{X}| \cdot (\tbinom{k_0}{k} + k_0 r) \cdot \tfrac{m_{7.12}}{\alpha\mu^k} + |\mathcal{X}| \cdot |\mathcal{I}(\mathcal{Y}) \times \{-1,0,1\}^{k_0}| \cdot \tilde{O}(p)\right) \\ &\leq \tfrac{r}{k_0} \cdot \left(\sqrt{N} \cdot r \cdot \tilde{O}(n) + \sqrt{N} \cdot O(r) \cdot \tfrac{O(r^3)}{\alpha(k_0/r)^k} + \sqrt{N} \cdot \tbinom{r}{k_0} 3^{k_0} \cdot \tilde{O}(p)\right) \ll t^{1-o(1)}. \end{aligned}$$

□

**Theorem 8.13** (*Barrington's theorem* [Bar89]). Any depth-$d$ circuit admits a permutation branching program[58] of width five and length $4^d$.

**Definition 8.14** (CMD and DCMD[59]). $\mathrm{CMD}_{n(n+1)/2}$ asks to compute the modulo-two determinant $\mathrm{CMD}_{n(n+1)/2}(\mathcal{M}) = \det(\mathcal{M}) \bmod 2$ of a Boolean connected matrix $\mathcal{M}$, i.e., $\mathcal{M}(i,j) \in \{0,1\}$

[58] A permutation branching program of width $k$ and length $\ell$ is a sequence of branching permutations $\{(x_{i_j}, f_j, g_j) \mid f_j, g_j \in \mathbb{S}_k, i \in (n], j \in (m]\}$. An input $x \in \{0,1\}^n$ guides the branches to select and compose the $\ell$ permutations $h_\ell \circ \cdots \circ h_1$ by $h_i = f_j$ if $x_{i_j} = 1$ and $h_i = g_j$ otherwise.

[59] CMD: Connected Matrix Determinant. DCMD: Decomposed CMD.

with $i \geq j+2 \Rightarrow \mathcal{M}(i,j) = 0$. $\mathrm{DCMD}_{n^3(n+1)/2}(\mathcal{M}_k, 1 \leq k \leq n^2) = \mathrm{CMD}(\sum_k \mathcal{M}_k \bmod 2)$ for connected $\mathcal{M}_k$. In particular, both CMD and DCMD belong to $\bigoplus\mathsf{L} \subset \bigoplus\mathsf{P}^{\mathsf{cc}} \subset \mathsf{PSPACE}^{\mathsf{cc}}$.

**Theorem 8.15** (CMD is $\bigoplus\mathsf{L}$-complete [IK02, CR20])**.** Any permutation branching program $C(x_1, \ldots, x_n)$ of width $k$ and length $\ell$ admits a projection mapping $p(x) : \{0,1\}^n \to \{0,1\}^{\frac{m(m+1)}{2}}$ with $m \leq k! \cdot \ell$ such that the modulo-two counting of $C(x)$'s accepting paths equals $\mathrm{CMD}(p(x))$.

**Definition 8.16** (approximate sum)**.** We say that a function $f$ admits a $\mathrm{Sum}_\varepsilon \circ \mathcal{F}$ circuit if there are functions $\mathcal{C}_i \in \mathcal{F}$ and coefficients $\alpha_i \in \mathbb{R}$ approximate $\forall x, \left|f(x) - \sum_{i=1} \alpha_i \mathcal{C}_i(x)\right| \leq \varepsilon$. Its weight is the sum of absolute coefficients $\sum_i |\alpha_i|$.

**Lemma 8.17** (boosting DCMD by CMD [CR20])**.** If a non-uniform circuit class $\mathcal{F}$ can $(1/2+\eta)$-approximate $\mathrm{DCMD}_{n^3(n+1)/2}$, $\mathrm{Sum}_\varepsilon \circ \mathcal{F}$ can compute $\mathrm{CMD}_{n(n+1)/2}$ by $O((\frac{n}{\varepsilon\eta})^2)$ circuits in $\mathcal{F}$ with the sum of absolute coefficients $O(1/\eta)$.

**Theorem 8.18** (easy witness lemma for depth [CR20])**.** If every $\mathsf{quasi}\text{-}\mathsf{NP}$ (resp. $\mathsf{NP}$) language is $(\frac{1}{2}+\frac{1}{2^{\log^k n}})$ (resp. $(\frac{1}{2}+\frac{1}{n^k})$)-approximable by circuits of $O(\log^k n)$ (resp. $k \log n$) depth for some $k \geq 1$, then every unary $\mathsf{NTIME}[\exp(n)]$ language must have a witness of $\mathsf{DEP}[n^\epsilon]$ (resp. $\mathsf{DEP}[\epsilon n]$) for any constant $\epsilon > 0$.

**Theorem 8.19** (Theorem 1.10)**.** Suppose $\mathsf{PH}^{\mathsf{cc}}$ either computes CMD or approximates DCMD by advantage $\frac{1}{2}+\frac{1}{\exp(n^{o(1)})}$. Then $\mathsf{DEP}[k \log n]$ cannot $(\frac{1}{2}+\frac{1}{n^k})$-approximate $\mathsf{NP}$ for all $k \geq 1$ , i.e., some $\mathsf{NP}$ language $\mathcal{L}$ cannot have $\mathsf{DEP}[k \log n]$ circuits $\mathcal{C}_n$ of advantage $\mathsf{Pr}_U[\mathcal{L}(U) = \mathcal{C}_n(U)] \geq \frac{1}{2} + \frac{1}{n^k}$ over the uniform random $n$-bit $U$.

*Proof.* Adapt Theorem 8.12's proof to Theorem 8.11's easy witness lemma for depth. Suppose $\mathsf{NP}$ admitted $(\frac{1}{2}+\frac{1}{n^k})$-approximation by $\mathsf{DEP}[k \log n]$ circuits. Take Theorem 8.12's parameters but $\varepsilon = o(\frac{1}{n^3 2^n})$, $d = \epsilon n$, $d' = 2(ch+2)d + 3\log n + 2\log\frac{1}{\varepsilon} + n^{o(1)}$, and $d'' = 2e_0 d' + h + 3 \ll n$.

**Witnessing shallow circuits:** The easy witness lemma for depth (Theorem 8.18) makes any $\exp(n)$-time verifier $V(1^n, y)$ to compress an $N$-bit witness $y$ of $V(1^n, y) = 1$ to a depth-$d$ circuit, i.e., $y$ must be a truth table of the circuit. Barrington's theorem (Theorem 8.13) transfers it to a permutation branching program of size $4^d$ and Theorem 8.15 to $\mathrm{CMD}_{m(m+1)/2}(p(x))$ by a projection mapping $p(x)$ of $m = 5! \cdot 4^d$. By assumption, $\mathsf{PH}_h^{\mathsf{cc}}[c \log n]$ must commute $\mathrm{CMD}_{n(n+1)/2}$ or $(\frac{1}{2}+\frac{1}{\exp(n^\zeta)})$-approximate $\mathrm{DCMD}_{n^3(n+1)/2}$, $\zeta = o(1)$, so $\mathsf{PH}_h^{\mathsf{cc}}[c \log m]$ must contain $\mathrm{CMD}_{m(m+1)/2}$ or $(\frac{1}{2}+\frac{1}{\exp(n^\zeta)})$-approximate $\mathrm{DCMD}_{m^3(m+1)/2}$. In the latter case, Theorem 8.17 writes $\mathrm{CMD}_{m(m+1)/2} \in \mathrm{Sum}_\varepsilon \circ \mathsf{PH}_h^{\mathsf{cc}}[c \log m]$ by a linear combination of $(\frac{m}{\varepsilon})^2 \exp(n^\zeta)$ $\mathsf{PH}^{\mathsf{cc}}$-circuits with the weight $\exp(n^\zeta)$. Let us derive a contradiction from the latter case since the former is easier to do it (by avoiding $\mathrm{Sum}_\varepsilon$ computation).

**Learning shallow circuits:** Let $V(x,y)$ have Theorem 8.10's short PCP's witness circuit $C_\nu^{\mathcal{W}_\nu}$. It admits a 3CNF computation, providing an $(h+3)$-layered circuit of fan-in $O(n^3)$ AND gate at the top, fan-in 3 OR gate at the second, fan-in $(\frac{m}{\varepsilon})^2 \exp(n^\zeta)$ $\mathrm{Sum}_\varepsilon$ gate at the third, and fan-in $2^{2cd}$ AND or OR gates at the remaining $h$ layers. Theorem 8.7 transfers it to a probabilistic polynomial of degree $d'' := 2e_0 d' + h + 3$ for $d_{h+3} = \log(n^3)$, $d_{h+2} = 2$, $d_{h+1} = 2\log(\frac{m}{\varepsilon}) + n^\zeta$, $d_h = \cdots = d_1 = c \log m$, and $d' = \sum_{\kappa=1}^{h+3} d_\kappa$. The probabilistic polynomial $\mathrm{NOR}(g_1, \ldots, g_{n^3})$ of the top AND gate may contain an additional error term $O(n^3 \varepsilon \exp(n^\zeta)) = o(1)$ since each $g_i$ is OR of 3 $\mathrm{Sum}_\varepsilon$ gates having an error term $\varepsilon$ and the weight $\exp(n^\zeta)$. Theorem 8.12's acceptance probability estimation recognizes $\{1^n \mid \exists y, V(1^n, y) = 1\} \in \mathsf{NTIME}[t] \backslash \mathsf{NTIME}[t^{1-o(1)}]$ in $\mathsf{NTIME}[t^{1-o(1)}]$ time, a contradiction. □

## 9 Natural Lower Bounds for NP $\not\subset$ TC$^1$ and VP $\neq$ VNP

In this section, in the worst-case analysis ($\mathrm{H}_\infty(G) = 0$), we translate number-theoretic/algebraic structures of $\mathsf{TC}^0$ and $\mathsf{VP}$ circuits into data-compressing exact learning algorithms in Lemmas 1.31 and 1.32. These learning algorihtms plug into William's program in REVIEW11 to estimate circuit's acceptance probabilities and yield the circuit lower bounds of Theorems 1.11 and 1.12.

### 9.1 quasi-NP $\not\subset$ quasi-TC$^0$

Let us briefly explain a number-theoretic mechanics to simulate $\mathsf{TC}^0$ by $\mathrm{SYM}^+$ in Lemma 1.31. It simulates every SYM gate feeding the outcomes $g(x) \in \{0,1\}$ from the previous layer by a sum of EXACT gates, and every EXACT gate by a truncated Taylor series via the Chinese remainder theorem $\sum_g g(x) = a \Leftrightarrow \sum_i \mathrm{MOD}_{p_i}(\sum_g g(x) - a) = 0 \Leftrightarrow \sum_{k=0}^{k_0} \forall t, \binom{\sum_i \mathrm{MOD}_{p_i}(\sum_g g(x)-a)}{k}(1/q_t) = a_t$ by $k = O(1)$, distinct primes $p_i \leq \ln a$, and distinct base points $q_t = t + O(1)$. Vandermode algebra in Lemmas 9.2 and 9.3 makes it a collision-free hash function. It promises the existence of $(a_t)_t$, and the modulus lifting of Lemma 9.1 turns it into a $\mathrm{SYM}^+$ computation.

**Lemma 9.1** (modulus lifting [BT94])**.** For any multi-linear polynomial $f \in \mathbb{Q}[\mathbf{x}_1, \ldots, \mathbf{x}_n]$ of $2\mathbf{norm}(f) + 1 \leq m^\ell$, and any integers $a_i \in \{0,1\} + m\mathbb{Z}$,

$$\textit{Modulus lifting:} \quad f(a_1 \bmod m, \cdots, a_n \bmod m) = f(\phi_\ell(a_1), \ldots, \phi_\ell(a_n)) \bmod m^\ell.$$

**Lemma 9.2** (Vandermonde's kernel [PR07])**.** The kernel of a generalized Vandemond matrix $\mathcal{M}_{t,n} = (a_{t'}^j)_{(t',j)\in[t]\times[n]}$ of distinct numbers $a_{t'}$ has dimension $n - t$ and admits a basis spanned by the cyclic shifts $v_0, \ldots, v_{n-t}$ of the following kinds. Let $\sigma(i) = \sum_{1\leq t_1 < \cdots < t_i \leq t} a_{t_1} a_{t_2} \cdots a_{t_i}$.

$$v_k = \big(\underbrace{0, \cdots, 0}_{k}, (-1)^t \sigma(t), \cdots, (-1)^i \sigma(i), \cdots, -\sigma(1), 1, \underbrace{0 \cdots, 0}_{n-t-k}\big).$$

**Lemma 9.3** (Vandermonde's conditional number [DSSS21])**.** For $n \in 2^{\mathbb{N}}$ and the $n$ distinct primitive $2n$th root of the unit $\zeta_i$, the conditional number $\|\mathcal{M}\|_{\mathtt{F}} \|M^{-1}\|_{\mathtt{F}}$ of the Frobenius norm $\|\mathcal{M}\|_{\mathtt{F}} := \sqrt{\mathrm{Tr}(\mathcal{M}^* \mathcal{M})}$ of the cyclic Vamdemond matrix $\mathcal{M} = (\zeta_i^j)_{i,j\in[n]}$ is $n$.

**Definition 9.4** ($\mathsf{ACC}$ circuits)**.** Let $2 = p_1 < p_2 < \cdots$ be the smallest prime numbers. An $\mathrm{SYM}_{m,q,t}$ gate is $t$-tuple set $\tilde{f} \subset \mathbb{N}^t/q$ to express $f = 1[\sum_{g\in \mathrm{in}(f)} \hat{g} \in \tilde{f}]$, i.e., each input $g$ of $f$ associates a $t$-tuple number $\hat{g} = (\hat{g}_{t'})_{t'=1}^t \in \mathbb{N}^t/q$ bounded by $\sum_g \sum_{t'} \hat{g}_{t'} \leq m/q$. A depth-$(2h+1)$ circuit $\mathrm{SYM} \circ \mathsf{ACC}_h = \mathrm{SYM}_{m_h,q_h,t_h} \circ (\mathrm{AND}_{k_d} \circ \{\mathrm{MOD}[p_1], \ldots, \mathrm{MOD}[p_{s_d}]\})_{d=1}^h$ consists of these $\mathrm{SYM}_{t_h,q_h,m_h}$ gates at the top, AND gates of fanin $k_d$ at each depth $2d$, and MOD gates of modulus $p_{\xi(\lambda_{dh})} \in \{p_1, \ldots, p_{s_d}\}$ of some $\xi \in \mathcal{Q}_{dh} := \prod_{d=1}^h (s_d]^{\Lambda_{dh}}$ of $\Lambda_{dh} := \prod_{d'=d}^h (k_{d'}]$. In this $\mathrm{SYM} \circ \mathsf{ACC}_h[\xi]$ circuit, the AND gates at depth $2d$ must take the moduli $\mathrm{AND}\big(\mathrm{MOD}[p_{\xi(1\lambda_{(d+1)h})}], \cdots, \mathrm{MOD}[p_{\xi(k_d\lambda_{(d+1)h})}]\big)$ along with a path $\lambda$ of depths from $2h$ down to $2d$.

**Lemma 9.5** (from $\mathsf{AC}^0$[SYM] via SYM$\circ\mathsf{ACC}$ to $\mathrm{SYM}^+$ (Lemma 1.31))**.** Given increasing positive integers $h \ll \triangle_h \leq \cdots \leq \triangle_1 \ll 2^{\triangle_h/h}$, and $k_d, \ell_{di}, m_d, n_d, q_{dt}, s_d$ and $t_d$ as follows[60].

$$\text{Let } k_d = O(1), m_d = \triangle_d^{\triangle_d^d(\triangle_d + O(d))}, n_d = O(\triangle_d^{d+1}), s_d = O(\triangle_d), t_d = O(\triangle_d), \tilde{p}_d = \prod_{i=1}^{s_d} p_i \approx s_d^{s_d},$$
$$\tilde{k}_d = \prod_{d'=d}^h k_{d'}, q_{dt} - t = q_{d1} = O(\triangle_d^{d+1}), \ell_{di} = \sum_{d'=d+1}^h \triangle^{2^{h-d'}} s_{d'} + i \cdot \triangle^{2^{h-d}} \text{ for } \triangle \gg s_{d+1},$$

[60]In this subsection, we often write an index $ij$ to mean $i, j$ for convenience, say $q_{dt} = q_{d,t}$.

$$u_{di} \ll \ell_{d0}, q_{dt} = 2n_{d-1} + t, \tilde{q}_{dt} = k_d! q_{dt}^{n_d} \text{ to satisfy } n_d = s_d t_d \cdot |\mathcal{Q}_{1d}|, m_d = \binom{n_d^{k_{d-1}} e^{\triangle_d} |\mathcal{Q}_{1d}| t_d}{|\mathcal{Q}_{1d}| t_d},$$

$$e^{\triangle_d} m_{d-1} \tilde{q}_{(d-1)t} t_{d-1} \ll \tilde{p}_d, e^{2\triangle_{d+1}} n_d^5 \ll (2-\epsilon)^{t_d}, \quad \prod_{d'=d}^{h} \prod_{i'=1}^{s_{d(i-1)}(d')} \left(2^{3\ell_{d'i'}} m_{d'}^{p_{i'}-1}\right)^{u_{d'i'}} \ll p_i^{\ell_{di}}.$$

Then, $\mathrm{SYM}_h \circ \cdots \circ \mathrm{SYM}_1$ circuits having $\mathrm{SYM}_d$'s fanin $e^{\triangle_d}$ transform into $\mathrm{SYM}_{r_d, u_d, m_d} \circ (\mathrm{AND}_{k_d} \circ \{\mathrm{MOD}[p_1], \ldots, \mathrm{MOD}[p_{s_d}]\})_{d=1}^{h}$ circuits, and even $\mathrm{SYM}^+[\textbf{deg}\colon \triangle^{2^h}, \textbf{norm}\colon \exp(\triangle^{2^h})]$ circuits of $\triangle = O(\triangle_1)$. These transformations are deterministic $O(\log n)$-space computation.

*Proof.* Truncated Talyor analysis transfers any $\mathrm{SYM}_d \circ \cdots \circ \mathrm{SYM}_1$-circuit $f$ to $\{(\hat{f}_{\xi t})_t \in \mathsf{ACC}_d[\xi]\}_\xi$ with $f(\mathbf{x}) = 1\left[\sum_\xi (\hat{f}_{\xi t s}(\mathbf{x}))_t \in \tilde{f}\right]$ in a recursion from the bottom-to-top layers:

*Modularize* SYM *circuit:* $f(\mathbf{x}) = \sum_{a \in f} 1\left[\sum_{g \in \mathrm{in}(f)} g(\mathbf{x}) = a\right] \quad (\because f \subset \mathbb{N} \text{ represents } f(\mathbf{x}) = 1[\sum_i \mathbf{x}_i \in f])$

$$\overset{\star}{=} \sum_{\substack{(a_{\xi' t'})_{\xi', t'} \subset \mathbb{N} \\ : (\sum_{\xi'} a_{\xi' t'})_{t'} \in \tilde{f}}} \bigwedge_{\xi' \in \mathcal{Q}_{1(d-1)}} \bigwedge_{t' \in (t_{d-1}]} 1\Big[\sum_g \sum_{\substack{\hat{g}_{\xi'} \in \mathsf{ACC}_{d-1}[\xi'] \\ : \sum_{\xi'} \hat{g}_{\xi'} \in \tilde{g}}} \tilde{q}_{(d-1)t'} \hat{g}_{\xi' t'}(\mathbf{x}) = a_{\xi' t'}\Big]$$

(by induction hypothesis for an appropriate $\tilde{f} \subset [\tilde{p}_d]^{t_{d-1}}$ taken in $\overset{\star\star}{=}$)

$$= \sum_{(a_{\xi' t'})} \bigwedge_{\xi', t'} \bigwedge_{i=1}^{s_d} \mathrm{MOD}_{p_i}\left(\sum_g \tilde{q}_{(d-1)t'} \hat{g}_{\xi' t'}(\mathbf{x}) - a_{\xi' t'}\right) \quad (\because \text{Chinese remainder by } \forall a_{\xi' t'} < \tilde{p}_d)$$

$$\overset{\star\star}{=} 1\left[\sum_\xi \left(\hat{f}_{\xi t}(\mathbf{x})\right)_t \in \tilde{f}\right] \text{ for } \hat{f}_\xi(\mathbf{x}) = (a_{\xi' t'})_{t'} \text{ and } |\{(a_{\xi' t'})_{\xi', t'} \mid \sum_{\xi'} (a_{\xi' t'})_{t'} \in \tilde{f}\}| \le m_d.$$

*Truncated Tayler series*: $\hat{f}_t(\mathbf{x}) := \sum_{k=0}^{k_d} \binom{j(\mathbf{x})}{k} (\frac{1}{q_{dt}})^k$ of $j(\mathbf{x}) := \sum_{\xi', t', i} \neg \mathrm{MOD}_{p_i}\left(\sum_g \tilde{q}_{(d-1)t'} \hat{g}_{\xi' t'}(\mathbf{x}) - a_{\xi' t'}\right) \le n_d$,

so that $\hat{f}_t(\mathbf{x}) = \sum_{\xi \in \mathcal{Q}_{1d}} \hat{f}_{\xi t}(\mathbf{x})$ of $\hat{f}_{\xi t}(\mathbf{x}) := \sum_{t'} \frac{r_{\xi t t'}}{\tilde{q}_{dt}} \prod_{k=1}^{k_d} \mathrm{MOD}_{p_{\xi(k)}}\left(\sum_g \tilde{q}_{(d-1)t'} \hat{g}_{\xi' t'}(\mathbf{x}) - a_{\xi' t'}\right)$,

$\hat{f}_{\xi t}(\mathbf{x}) \in \mathsf{ACC}_d[\xi]$ of $r_{\xi t t'} \in \mathbb{N}$, $\xi'(\lambda_{1(d-1)}) = \xi(\lambda_{1(d-1)})$, and $\xi(k) = \xi(\lambda_{1(d-1)} k)$ over $\lambda \in \Lambda_{1d}$.

We can verify $\overset{\star}{=}$ by induction on $d$ because the Vandermond algebras (Lemmas 9.2 and 9.3) guarantee $\overset{\star\star}{=}$ to incur no collision $(y_j(x))_j \ne (y_j(x'))_j \Rightarrow \sum_{f,j} y_j(x) \hat{f}(j) \ne \sum_{f,j} y_j(x') \hat{f}(j)$ of

*Hash function*: $\hat{f}(j) = (\hat{f}_t(j))_t$ for $\hat{f}_t(j) := \hat{f}_t(\mathbf{x})$ of $j := j(\mathbf{x})$.

*Taylor series approximation*: $\hat{f}_t(j) = \sum_{k=0}^{k_d} \binom{j}{k} (\frac{1}{q_{dt}})^{k_d} = (1 + \frac{1}{q_{dt}})^j (1 - \varepsilon_{td}(j))$,

$$\varepsilon_{td}(j) := \binom{j}{k_d} / (1 + \frac{1}{q_{dt}})^j \cdot \int_0^{\frac{1}{q_{dt}}} (1+z)^{j - k_d - 1} (\frac{1}{q_{dt}} + z)^{k_d} dz \approx 0. \quad (\because q_{dt} \gg j \text{ and } k_d \gg 1)$$

*Colliding numbers*: $y_j(\mathbf{x}) = \left|\left\{f \in \mathcal{F} \mid \sum_\xi \hat{f}_\xi(\mathbf{x}) \in \tilde{f}, j(\mathbf{x}) = j\right\}\right|$ for a given $\mathcal{F}$ of size $|\mathcal{F}| \le e^{\triangle_{d+1}}$,

where $j \in [n_d]$ and $\sum_{j=0}^{n_d} y_j(\mathbf{x}) \le e^{\triangle_{d+1}} n_d$.

*No collision*: $(y_j(x))_j \ne (y_j(x'))_j \wedge \sum_j y_j(x) \hat{f}(j) = \sum_j y_j(x') \hat{f}(j)$

$\Rightarrow$ By Lemma 9.2 of $a_t = 1 + 1/q_{dt}$, $\exists \alpha_j \in \mathbb{R}$,

$\sum_{j=0}^{n_d - t_d} \alpha_j v_j = \left((y_j(x) - y_j(x')) \cdot (1 - \varepsilon_{td}(j))\right)_j \in (\mathbb{N}^{n_d} \backslash 0^{n_d})(1 \pm \epsilon)$

$\Rightarrow$ $\beta_j := \alpha_j \sigma(t)$ has a norm $\|(\beta_j)_j\| \ge \frac{1-\epsilon}{\sqrt{n_d}}$ since Lemma 9.2's triangular matrix $(\frac{v_{kj}}{\sigma(t)})_{k,j}$ has the diagonals $\frac{v_{kk}}{\sigma(t)} \in \{1, -1\}$ and the norm $\|(\frac{v_{kj}}{\sigma(t)})^{-1}\|_{\mathtt{F}} = \sqrt{n_d}$

$\Rightarrow$ $(\sum_{j=0}^{t_d} \alpha_j \mathbf{x}^j) \cdot \prod_{t=1}^{t_d} (\mathbf{x} - a_t) = (\sum_{j=0}^{t_d} \beta_j \mathbf{x}^j) \cdot \prod_{t=1}^{t_d} (\mathbf{x}/a_t - 1)$

$= \sum_{j=0}^{n_d} (y_j(x) - y_j(x'))(1 - \varepsilon_{td}(j)) \mathbf{x}^j$ in the polynomial ring $\mathbb{Z}[\mathbf{x}]$

$\Rightarrow$ $\frac{\|(\beta_j)_j\|_2}{\|\left((y_j(x) - y_j(x') \cdot (1 - \varepsilon_{td}(j))\right)_j\|_2} \le \|(1[i=j] \prod_{i=1}^{n_d} (\frac{\zeta_i}{a_t} - 1))_{i,j}^{-1}\|_{\mathtt{F}} \cdot \|(\zeta_i^j)_{i,j}\|_{\mathtt{F}} \cdot \|(\zeta_i^j)_{i,j}^{-1}\|_{\mathtt{F}}$

$$\text{for } \zeta_i = \boldsymbol{e}^{2\boldsymbol{\pi}\sqrt{-1}\cdot(i+n_d/2)/(2n_d)} \text{ with } \left|\zeta_i/a_t - 1\right| \geq \sqrt{1+1/a_t^2}$$
$$\Rightarrow \frac{1-\epsilon}{(1+\epsilon)\sqrt{n_d}\cdot \boldsymbol{e}^{\triangle_{d+1}} n_d} \leq \frac{n_d}{\sqrt{\sum_i \prod_t (1+1/a_t^2)}}, \text{ contradicting to } \boldsymbol{e}^{2\triangle_{d+1}} n_d^5 \ll (2-\epsilon)^{t_d}.$$

The modulus lifting (Lemma 9.1) transfers the obtained $\hat{f}(\mathbf{x})(=\hat{f}_t(\mathbf{x})) \in \mathsf{ACC}_h[\xi]$ to an SYM$^+$ circuit $\check{f}(\mathbf{x})$ in a top-to-bottom recursion. Write $\hat{f}(\mathbf{x}_d)$ for the circuit $\hat{f}$ considering the MOD-gates $\mathbf{x}_{d\kappa}$ at depth $2d-1$ as the input variables $\mathbf{x}_d = (\mathbf{x}_{d\kappa})_\kappa$. So, $\mathbf{x} = \mathbf{x}_0$, $\hat{f}(\mathbf{x}) = \mathbf{x}_{h+1}$, and $\mathbf{x}_{d\kappa} \in \mathrm{MOD}[p_\kappa] \circ \mathsf{ACC}_{d-1}[\xi_\kappa]$ of $p_\kappa = p_{\xi(\lambda_d)}$ and $\xi_\kappa(\lambda_{1d}) = \xi(\lambda_{1d})$ on every path $\lambda$ passing through $\mathbf{x}_{d\kappa}$. The induction hypothesis gives $\check{f}(\mathbf{x}_d)$ of degree $u_d$ and asks to present a $\check{f}_t(\mathbf{x}_{d-1})$ of degree $u_{d-1}$ via replacing every $\mathbf{x}_{d\kappa}$ in the above $\hat{f}$'s construction of

*Truncated Tayler series*: $\tilde{\mathbf{x}}_{d\kappa} := \sum_{f\in\mathcal{F}_\kappa} \tilde{q}_{dt} \hat{f}_{\xi_\kappa t}(\mathbf{x}_{d-1}) - a_{\xi_\kappa} = \sum_f \sum_{t'} r_{\xi_\kappa t t'} \prod_{k=1}^{k_d} \mathbf{x}_{(d-1)\kappa'} - a_{\xi_\kappa}$, $\kappa' := \kappa k t'$.

*Modulus lifting*: Induce $\check{f}(\mathbf{x}_d) \mapsto \check{f}(\mathbf{x}_{d-1})$ by substitution to all $\mathbf{x}_{d\kappa} := \phi_{\ell_\kappa}(\tilde{\mathbf{x}}_{d\kappa}^{p_\kappa - 1})$ over depth $2d-1$ so that $\check{f}(\mathbf{x}_d) = \check{f}(\mathbf{x}_{d-1}) \bmod p_\kappa^{\ell_\kappa}$ on $\mathbf{norm}(\check{f}(\mathbf{x}_d)) \ll p_\kappa^{\ell_\kappa}$. Refine it to $\check{f}_{d(i-1)} \mapsto \check{f}_{di}$ by substitution $\mathbf{x}_{d\kappa(i)} := \phi_{\ell_i}(\tilde{\mathbf{x}}_{d\kappa(i)}^{p_i-1})$ to the all variable of type $\mathbf{x}_{d\kappa(i)} \in \mathrm{MOD}[p_i] \circ \mathsf{ACC}_{d-1}[\xi_{\kappa(i)}]$.

SYM$^+$ *degree:* Let $\check{f}_{d0} := \check{f}_{(d+1)s_{d+1}}$, $u_d := \mathbf{deg}(\check{f}_{d0})$, and $u_{di} := \mathbf{deg}(\mathrm{MOD}[p_i])$ = the maximum number of $\mathrm{MOD}[p_i]$-variables $\mathbf{x}_{d\kappa(i)}$ occuring in an AND-term of $\check{f}_{d0}$. They increase by $\mathbf{deg}(\check{f}_{di}) - \mathbf{deg}(\check{f}_{d(i-1)}) \leq \mathbf{deg}\big(\phi_{\ell_{di}}(\tilde{\mathbf{x}}_{d\kappa(i)}^{p_i-1})\big) \cdot u_{di} = (2\ell_{di}-1)(p_i-1)u_{di} := v_{di}$ for $i = 1, \ldots, s_d$, so that $\mathbf{deg}(\check{f}_{di}) \leq \sum_{d'=d}^{h} \sum_{i'=1}^{s_{di}(d')} v_{d'i'}$ of $s_{di}(d') := s_{d'} \cdot 1[d' > d] + i \cdot 1[d' = d]$.

SYM$^+$ *norm:* The ratios icrease by $\frac{\mathbf{norm}(\check{f}_{di})}{\mathbf{norm}(\check{f}_{d(i-1)})} \leq \mathbf{norm}\big(\phi_{\ell_{di}}(\tilde{\mathbf{x}}_{d\kappa(i)}^{p_i-1})\big)^{u_{di}} \leq \big(2^{3\ell_{di}} m_d^{p_i-1}\big)^{u_{di}}$,
so that $\mathbf{norm}(\check{f}_{d(i-1)}) \leq \prod_{d'=d}^{h} \prod_{i'=1}^{s_{d(i-1)}(d')} \big(2^{3\ell_{d'i'}} m_{d'}^{p_{i'}-1}\big)^{u_{d'i'}} \ll p_i^{\ell_{di}}$. $\square$

**Theorem 9.6** ([Wig94])**.** $\mathsf{NL}/\mathrm{poly} \subset \bigoplus \mathsf{L}/\mathrm{poly}$

**Theorem 9.7** (Theorem 1.11)**.** Suppose $\mathsf{AC}_h^0[\mathrm{SYM}]$ of size $2^{(\log n)^{O(1)}}$ either computes CMD or approximates DCMD by advantage $\frac{1}{2} + \frac{1}{2^{(\log n)^{O(1)}}}$. Then $\mathsf{DEP}[(\log n)^k]$ cannot $(\frac{1}{2} + \frac{1}{2^{\log^k n}})$-approximate $\mathsf{NTIME}[2^{(\log n)^{O(1)}}]$ for all $k \geq 1$.

*Proof.* Follow Theorem 8.19's proof. Suppose $\mathsf{AC}_h^0[\mathrm{SYM}]$ of size $2^{(\log n)^{O(1)}}$ approximates DCMD by advantage $\frac{1}{2} + \frac{1}{2^{(\log n)^{O(1)}}}$. Let $\{1\}^* \supset \mathcal{L} \in \mathsf{NTIME}[2^\nu] - \mathsf{NTIME}[2^\nu/\mathrm{poly}(\nu)]$. Theorem 8.11 has provided Theorem 8.10's short PCP a witness circuit $\mathcal{C}_\nu^{\mathcal{W}_\nu}$ of a 3CNF formula $\mathcal{C}_n$ and poly$(n)$-size circuit $\mathcal{W}_\nu$. If $1^\nu \in \mathcal{L}$, then the circuit $\mathcal{C}_\nu^{\mathcal{W}_\nu}$ always outputs zero, while if $1^\nu \notin L$, it outputs one but at most $1/n$ fraction of error. Instead of measuring the acceptance probability of this $\mathcal{C}_\nu^{\mathcal{W}_\nu}$ to distinguish these two cases, we will follow Williams's trick [Wil14a] to reduce the input bits from $n$ to $n - n' \approx n$ by measuring the induced OR-top circuit the acceptance probability of the OR-top circuit $\sum_{\mathbf{x}\in\{0,1\}^{\nu-n'}} \bigvee_{x\in\{0,1\}^{n'}} \mathcal{C}_\nu^{\mathcal{W}_\nu}(\mathbf{x}, x)$ of $\nu = n + n' = n + n^\epsilon$. The easy witness lemma for depth (Theorem 8.18) would compress a witness $y$ of $\exp(n)$-time verification $V(1^n, y) = 1$ into a depth-$d$ circuit with $d = n^{\epsilon/c}$ for constants $c, c'$, and $c''$. Consequently, Theorem 9.6 reduces $\bigvee_{x\in\{0,1\}^{n'}} \mathcal{C}_\nu^{\mathcal{W}_\nu}(\mathbf{x}, x)$ to a DCMD's approximation. Suppose that $\mathsf{quasi\text{-}AC}_h^0[\mathrm{SYM}]$ of fan-in $2^{(\log n)^c}$ admits the DCMD's approximation by advantage $\frac{1}{2} + \frac{1}{2^{(\log n)^{c'}}}$. Let $m = O(4^d)$, $\varepsilon_0 = o(\frac{1}{2^{(\log n)^{c'}}})$ and $e_0 \approx \frac{1}{\log \delta}$. Guess an $\mathrm{Sum}_\varepsilon \circ \mathsf{AC}_h^0[\mathrm{SYM}]$ circuit (inputting $m(m+2)/2$ bits) of fan-in $(m/\varepsilon_0)^2$ at the top $\mathrm{Sum}_\varepsilon$ gate and depths $\triangle_h = \cdots = \triangle_1 = 2^{(\log(m(m+2)/2))^c} = 2^{4^c n^\epsilon}$ of the $\mathsf{AC}_h^0$'s $h$ layers to realize the $\bigvee_{x\in\{0,1\}^{n'}} \mathcal{C}_\nu^{\mathcal{W}_\nu}(\mathbf{x}, x)$ circuit. Lemma 9.5 transforms it to $\check{f} \in$

SYM$^+$[**deg**:$2^{2^h 4^c n^\epsilon}$, **norm**: $\exp(2^{2^h 4^c n^\epsilon})$] to compute $\Pr[\hat{f}'(X) = \bigvee_{x\in\{0,1\}^{n'}} C_\nu^{\mathcal{W}_\nu}(X,x)] \geq 1-\delta$. Williams's dynamic program [Wil14a] can estimate it in a contradictory fast time

$$\underset{\textit{for acceptance probability estimation}}{\textit{Nondeterministic time}}: \mathrm{poly}(n)\cdot\big(2^n + 2^{n'}\cdot\mathbf{norm}(\check{f})\big) \ll 2^{\nu(1-o(1))}. \qquad \square$$

### 9.2 VP $\neq$ VNP

We take Raz's elusive function approach to prove Theorem 1.12. It requires set-multilinear polynomials, so we fix a number $q \in 2^{\mathbb{N}}$ of order $q = (\log n)^{O(\log n)}$ and identify a binary string $\tilde{x}$ with the $q$-nary vector $x$ via $[q]^n \ni x \cong \tilde{x} \in 2^{\tilde{n}}$. It algebraizes a language $\mathcal{L} \subset [q]^n \cong \{0,1\}^{\tilde{n}}$ to a set-multilinear polynomial $\hat{\mathcal{L}} := \sum_{x\in[q]^n}\mathcal{L}(x)\prod_{i=1}^n \mathbf{x}_{i,x_i}$, and $\hat{\mathcal{F}} := \{\hat{\mathcal{L}} \mid \mathcal{L}\in\mathcal{F}\}$.

**Theorem 9.8** (circuits to formulae [Hya79])**.** Any size-$s$ circuit computing a degree $d$ polynomial transforms to a formula of size $s^{O(\log d)}$ and depth $O(\log d)$.

**Definition 9.9** (multi-linear polynomial)**.** A polynomial is set-multilinear over variables $\mathcal{X}_1 \sqcup \cdots \sqcup \mathcal{X}_r$ if every term (monomial) contains one $\mathcal{X}_i$ variable. A circuit is set-multilinear if so is every gate over subsets of $\{\mathcal{X}_1, \cdots, \mathcal{X}_r\}$.

**Lemma 9.10** (multi-linearization)**.** Any algebraic circuit of size $s$ and depth $d$ computing a set-multilinear polynomial over variables $\mathcal{X}_1 \sqcup \cdots \sqcup \mathcal{X}_r$ can transfer to a set-multilinear circuit of size $(d+2)^r \cdot s$ and depth $2d$.

**Lemma 9.11** (Theorem 1.32)**.** Any sum $f = \sum_{k=1}^s \sum_{i,j=1}^n \mathbf{x}_i\lambda_i(k)\mu_j(k)\mathbf{x}_j$ of $s \ll n$ bilinear forms over $\mathcal{M}_{ij}(k)\in\mathbb{F}$ with multi-linearity $\forall i,\forall j,\forall k, i\neq j \Rightarrow \lambda_i(k)\mu_j(k)\neq 0$ is exactly learnable from $O(s^2 n)$ data and $O(s^2 n\log|\mathbb{F}|)$ guess bits in $O(s^2 n)$ time.

*Proof.* Without loss of generality, we may assume that given $s$ bilinear forms have disjoint keys (i.e., specific indices) $\mathcal{K} = \{i_k, j_k \mid k\in(s], \lambda_{i_k}(k)\mu_{j_k}(k)\neq 0\}$. Otherwise, there exist $2s'$ ($< 2s$) keys $\mathcal{K}$ to cover either $\{i \mid \lambda_i(k)\neq 0\}\subset\mathcal{K}$ or $\{j \mid \mu_j(k)\neq 0\}\subset\mathcal{K}$ over all $k > s'$ so that $f$ is learnable by only $s'n$ queries $f(\mathbf{1}_{i_k}+\mathbf{1}_j)$ and $f(\mathbf{1}_i+\mathbf{1}_{j_k})$ over $\{i_k, j_k\}\in\mathcal{K}$. Fix all these $\lambda_{i_k}(k)$ and $\mu_{j_k}(k)$ as non-zero values in $\mathbb{F}$, and all $\lambda_{i_k}(k')$ and $\mu_{j_k}(k')$ of $k'\neq k$ as well. The same argument holds for $\tilde{\lambda}_{i_k}$ and $\tilde{\mu}_{i_k}$ induced in Gaussian elimination.

**Gaussian elimination (Jacobian matrix triangularization)** can force $\forall(k'<k), \tilde{\lambda}_{i_{k'}}(k) := \lambda_{i_{k'}}(k) + \sum_{k'<k} a_{k',k}\lambda_{i_{k'}}(k') = 0$ and $\forall k' < k, \tilde{\mu}_{j_{k'}}(k) := \mu_{j_{k'}}(k) + \sum_{k'<k} b_{k',k}\mu_{j_{k'}}(k') = 0$ by taking the inductively induced coefficients $a_{k',k}$ and $b_{k',k}$ in $\mathbb{F}$. It makes

$$\begin{gathered}\underset{\textit{mapping}}{\overset{\textit{A quadratic}}{\textit{polynomial}}}: \big(\tilde{\lambda}_i(k), \tilde{\mu}_j(k) \mid i\notin\{i_{k'} \mid k'\leq k\}, j\notin\{j_{k'}\mid k'\leq k\}\big)_{k=1}^s \mapsto \big(\tilde{f}(\mathbf{1}_i+\mathbf{1}_j) \mid (i,j)\in\Lambda\big)_{k=1}^s,\\ \tilde{f}(\mathbf{1}_{i_k}+\mathbf{1}_j) := f(\mathbf{1}_{i_k}+\mathbf{1}_j) + \textstyle\sum_{k'<k} a_{k',k} f(\mathbf{1}_{i_{k'}}+\mathbf{1}_j),\ \tilde{f}(\mathbf{1}_i+\mathbf{1}_{j_k}) := f(\mathbf{1}_i+\mathbf{1}_{j_k}) + \sum_{k'<k} b_{k',k} f(\mathbf{1}_i+\mathbf{1}_{j_{k'}})\end{gathered}$$

an invertible mapping over $\mathbb{F}^{\sum_{k=1}^s 2(n-k)}$, so uniquely identify the all argued $\tilde{\lambda}_i(k)$ and $\tilde{\mu}_j(k)$, and all $\lambda_i(k)$ and $\mu_i(k)$ as well. Additional $s(s+1)/2$ queries to evaluate $\tilde{f}(\mathbf{1}_{i_k}+\mathbf{1}_{j_{k'}})$ over all $1\leq k'\leq k\leq s$ can determine the unargued coefficients $\lambda_{i_{k'}}(k)$ and $\mu_{j_{k'}}(k)$, too. $\square$

**Theorem 9.12.** $\widehat{\mathsf{NP}} \not\subset \mathsf{VSIZE}[2^{\frac{o(n)}{\log^3 n\log\log n}}]$ or $\forall\epsilon>0, \forall k\geq 1, \mathsf{NTIME}[\exp(n^\epsilon)] \not\subset \mathsf{SIZE}[n^k]$.

*Proof.* Follow Theorem 9.7's argument on Theorem 8.12's way to apply the easy witness lemma (Theorem 8.11). Suppose $\mathsf{NTIME}[t^{c_{8.11}}] \subset \mathsf{SIZE}[\nu^{k\cdot c_{8.11}/\epsilon}]$ for $t := 2^\nu$. Williams's trick [Wil14a] has reduced the recognition of $\mathcal{L} \in \mathsf{NTIME}[2^\nu]\backslash\mathsf{NTIME}[2^\nu/\mathrm{poly}(\nu)]$ to measuring the acceptance probability of an OR-top circuit $\mathcal{C}_n(\mathbf{x}) := \bigvee_{w\in\{0,1\}^{\nu-\tilde{n}}} \mathcal{C}_\nu^{\mathcal{W}_\nu}(\mathbf{x}, w) \in \mathsf{NP}$ of $\mathbf{x} \in [q)^n$ by taking $\tilde{n} = \frac{1-\epsilon}{3}\nu$. The assumption $\widehat{\mathsf{NP}} \subset \mathsf{VSIZE}[s]$ of $s = 2^{\frac{o(\tilde{n})}{\log^3 \tilde{n} \log\log \tilde{n}}} = 2^{\frac{o(n)}{\log^2 n}}$ presents an algebraic circuit $\hat{\mathcal{C}}_n := \widehat{C_\nu^{\mathcal{W}_\nu}}$ of a homogeneous polynomial $\hat{\mathcal{C}}_n(\mathbf{x}) = \sum_{x\in[q)^n} \mathcal{C}_n(x)\mathbf{x}_x$ of the terms $\mathbf{x}_x = \prod_{i=1}^n \mathbf{x}_{i,x_i}$.

**Learning elusive bilinear decompositions of algebraic circuits:** Theorems 9.8 and 9.10 transfer $\hat{\mathcal{C}}_n$ to a set-multilinear formula of size no more significant than $(d+2)^n s^{O(\log n)}$ and depth $d = O(\log n)$. Decompose it to a sum of bilinear forms $\hat{\mathcal{C}}_n = \sum_{k=1}^{s'} \sum_{x\in\mathcal{I}_k} \sum_{y\in\mathcal{J}_k} \mathbf{x}_x \lambda_x(k)\lambda_y(k)\mathbf{x}_y$ of $\mathcal{I}_k \times \mathcal{J}_k \cong [q)^n$ with balance $n/3 \le \log_q |\mathcal{I}_k| \le 2n/3$. There are $s' \le ((d+2)^n s^{O(\log n)})^d$ forms with $(d+2)^d \ll q$ and $s' \le s^{O(\log n)\cdot d} = 2^{o(n)}$. Lemma 9.11 can identify them by querying for $\sum_{k=1}^{s'}(|\mathcal{I}_k| + |\mathcal{J}_k|)$ times to evaluate $\hat{\mathcal{C}}_n$ in $s' \cdot q^{2n/3} \cdot 2^{\nu-\tilde{n}} \cdot \mathrm{poly}(\nu) \ll 2^{\nu(1-\epsilon+o(1))}$ time. Once getting all coefficients $\lambda_x(k)$ and $\lambda_y(k)$, one can estimate the acceptance probability in $s' \cdot q^n \cdot \sum_k |\mathcal{I}_k||\mathcal{J}_k| \le s' \cdot q^n \cdot q^n \ll 2^{\nu(2/3-\epsilon+o(1))}$ time, contradicting to $\mathcal{L} \notin \mathsf{NTIME}[2^\nu/\mathrm{poly}(\nu)]$. □

**Theorem 9.13** (generalized easy witness lemma for depth [CR20])**.** Given smooth functions[61] $\ell(n), d(n)$ and $\log s(n)$. Suppose $s(s(s(n)^{c_{9.13}})^{c_{9.13}})^{c_{9.13}} \le 2^{d(\ell(n))}$ and $t(n) := \exp\left(\frac{c_{9.13}\cdot \ell^2(n)}{d(\ell(n))}\right)$ is non-decreasing. If every $\mathsf{NTIME}[t(n)]$ language is $(1/2 + 1/s(n))$-approimable by circuits of depth $\log s(n)$, then every unary $\mathsf{NTIME}[\exp(n)]$ language must have a witness of $\mathsf{DEP}[d(n)]$.

**Theorem 9.14** (Theorem 1.12)**.** Suppose $\mathsf{VP}$ either computes CMD or approximates DCMD by advantage $\frac{1}{2} + \frac{1}{2^{(\log n)^{O(1)}}}$. Then $\mathsf{DEP}[(\log n)^k]$ cannot $(\frac{1}{2} + \frac{1}{2^{\log^k n}})$-approximate $\mathsf{NTIME}[2^{(\log n)^{k^3}}]$.

*Proof.* The same with Theorem 9.12's one but taking Theorem 8.19's way to apply the generalized easy witness lemma for depth (Theorem 9.13). Take the same parameters with Theorem 8.18 but $d(n) = \epsilon n/\log^2(n)$, $\ell(n) = \log^k n$, and $s(n) = 2^{(\log n)^{(1-\epsilon)k^{1/3}}}$, so $t(n) = \exp(\log^k n \log\log n)$. Apply Theorem 9.13 to the algebraic circuit class $\mathcal{C}$, $s = 2^d$ and $t = \mathrm{poly}(n)$, yielding $\mathcal{W}_\nu \in \mathsf{DEP}[d(n)]$ of $\frac{1-\epsilon}{3}\nu = \tilde{n}$, so $\hat{\mathcal{C}}_n \in \mathrm{Sum}_{\varepsilon_0} \circ \mathsf{DEP}[4^d]$ by Theorem 9.14 for Theorem 9.7's $m = O(4^{d(n)})$, $\varepsilon_0 = o(\frac{1}{2^{(\log n)^{c'}}})$ and $e_0 \approx \frac{1}{\log \delta}$. Theorem 9.12 has learned the elusive bilinear decomposition of the $\hat{\mathcal{C}}_n$ in a contradictory fast time. □

**Theorem 9.15** (combinatorial design [NW94])**.** For $k = O(m^2/\log n)$ and $n < 2^m$ there is $\mathcal{S}_1, \ldots, \mathcal{S}_n \subset [k]$ with $|\mathcal{S}_i| = m$ and $i \ne j \Rightarrow |\mathcal{S}_i \cap \mathcal{S}_j| \le \log n$. Such an $m$-set family $\mathcal{S}$ is constructible in deterministic $\mathrm{poly}(n, 2^k)$ time and called $(m, \log n)$-*combinatorial design.*

**Theorem 9.16** (hardness to derandomization [KI04])**.** Let $\mathcal{S}$ be an $(m, \log n)$-combinatorial design. Let $f(\mathbf{x})$ be an $m$-variate multi-linear polynomial which an algebraic circuit of size $s$ cannot compute. Let $\mathcal{C}(\mathbf{y})$ be an $n$-variate circuit of size $s'$ and degree $d$. If $(s'nmd)^5 < s$ then $\mathcal{C}(\mathbf{y}) \equiv 0 \Leftrightarrow \mathcal{C}(f(\mathbf{x}\restriction\mathcal{S}_1), \ldots, f(\mathbf{x}\restriction\mathcal{S}_n)) \equiv 0$.

**Theorem 9.17** (derandomizing PIT)**.** Either PIT is solvable in deterministic $n^{\mathrm{poly}(\log\log n)}$ time, or $\epsilon > 0, \forall k \ge 1, \mathsf{NTIME}[\exp(n^\epsilon)] \not\subset \mathsf{SIZE}[n^k]$.

*Proof.* Theorem 9.12's algebraic circuit hardness derandomizes PIT. Suppose $\widehat{\mathsf{NTIME}}[\mathrm{poly}(m)] \not\subset \mathsf{VSIZE}[s]$ for $s = 2^{\frac{o(m)}{\log^3 m \log\log m}}$. Let $m = \log n (\log\log n)^{3+2\epsilon}$. We have an $m$-variate $f(\mathbf{x}) \in \mathsf{NP}$

[61] A function $f(n)$ is smooth if $f(2n) \le cf(n)$ holds for a constant $c > 0$.

whose algebraic circuit size must be $s = 2^{\Omega(\frac{\log n (\log\log n)^{\epsilon}}{\log\log\log n})}$. Since $(s'nmd)^5 < s$ for $s' = \mathrm{poly}(n)$ and $d = n$, PIT is solvable in $|\{0,1\}^m| \cdot s' \cdot 2^{O(m^2/\log n)} = O(2^{\log n (\log\log n)^{6+4\epsilon}})$ time by exhausting the input space $x \in \{0,1\}^m$ to evaluate Theorem 9.16's $\mathcal{C}\big(f(x \restriction \mathcal{S}_1), \ldots, f(x \restriction \mathcal{S}_n)\big)$. $\square$

# 10 Discussions and Open Problems

Our effort to understand smoothed complexities of min-entropy below $O(\log n)$ has brought several new insights into machine learning, combinatorial optimization, cryptography, and computational complexity by relying on only the well-established results and methodologies in these fields. Can we go further from here without fundamentally new mathematical discoveries?

**From refutation to approximation:** Max$k$SAT of $O(n^{k-1})$ constraints required $2^{n^{1-\epsilon}}$ time to approximate $\max_\theta P(y = f_\theta(x))$ under ETH [FLP16]. Meanwhile, we have shown that promise-Max$k$SAT to distinguish between $|\max_\theta P(y = f_\theta(x) - \max_\theta P'(y = f_\theta(x))| \geq \epsilon$ and $P(x,y) \equiv P'(x,y)$ is possible with only $\tilde{O}(n^{k/2})$ constraints in $n^{O(k)}$ time. Is this sample complexity gap persistent for the other $f_\theta$ in combinatorial optimization, as well as $f_\theta(x) = \bigwedge_{i=1}^n \theta \circ x_i$? For example, MaxCUT requires the sample complexities (number of edges) $\Omega(n^{2-\epsilon})$ for $O(2^{n^{\epsilon'}})$-time approximation ([FLP16]), but only $\tilde{O}(n)$ edges for the $\tilde{O}(n)$-time distinguishment (Theorem 6.21 of $k = 2$). How about Max$k$CSP, Densest$k$Subgraph, MinBisection, etc.?

**PAC learning planted $k$DNF (in the worst-case):** We have shown that the planted $k$DNF is PAC learnable from any $\tilde{O}(n^{\lceil k/2 \rceil})$ data in $n^{O(k)}$ time. The best possible might be $\tilde{O}(n^{k/2})$ data since all sub-linear degree SoS, sub-linear degree PC, and sub-exponential time Res have demanded $\Omega(n^{(k-\epsilon)/2})$ data. Sub-exponential size LP might require $\Omega(n^{(k-\epsilon)/2})$ data learning since it was so for noisy PAC learning [BCR20].

**Linear time DNF learning in smoothed analysis:** Our correlation analysis has derived a linear time proper learning of planted monotone DNF with expanding terms. It has safely detected the correlation $\Pr[(-1)^{G(X_i)+Y} \mid \lfloor X_i/2 \rfloor = a]$ under an $O(\log s)$-independent flipper $G$. Unfortunately, the correlation of a non-monotone variable $X_i$ could vanish. Thus, linear time PAC learning (non-monotone) planted DNF in the smoothed analysis is wide open, even though PAC learnability of monotone DNF implies that of non-monotone DNF [KLV94].

**Inverting planted Fourier transform and LWE:** Fortunately, degree-$k$ multi-linear polynomials $f(\mathbf{x}_1, \ldots, \mathbf{x}_d) = \sum_{|w| \leq k} \prod_{i \in w} \theta(\lfloor \mathbf{x}_i/2 \rfloor)(-1)^{\mathbf{x}_i}$ over $\mathbb{Z}_q$ have the statistically non-zero correlation $\Pr[Y \cdot (-1)^{\sum_{i \in w} G(X_i)} \mid \lfloor X_w/2 \rfloor = a]$ at any $|w| = k$. Our smoothed analysis has retrieved the hidden Fourier coefficient $\prod_{i \in w} \theta_i(a_i)$ from any data of small $\max |Y|$ with noise $\Pr[Y \neq f(G(X)) \mid \lfloor X_w/2 \rfloor = a] \approx 0$. It has solved LWE with arbitrary i.i.d. noise in polynomial time due to the concentration of $\sum_{i \in w} \pm\theta_i$ over the randomly flipping signs of the small secrets $\forall i, |\theta_i| = O(1)$. However, it does not apply to non-constant $\theta_i$, nor a small $q$. Particularly, LWE with the random $\theta_i \in \mathbb{Z}_q$ and LPN with $q = 2$ are still away from polynomial-time inversion.

**Computational complexity lower bounds:** We have shown that either $\mathsf{PSPACE}^{\mathsf{cc}} \not\subset \mathsf{PH}^{\mathsf{cc}}$ or $\forall k, \mathsf{quasi\text{-}NP} \not\subset \mathsf{quasi\text{-}NC}^k$ must hold. The latter $\mathsf{quasi\text{-}NP} \not\subset \mathsf{quasi\text{-}NC}^k$ may not extend immediately to $\mathsf{quasi\text{-}NP} \not\subset \mathsf{quasi\text{-}NC}$ (so $\mathsf{NEXP} \not\subset \mathsf{PSPACE}$). For example, Theorem 8.12's nondeterministic time analysis allows a sparsity $|M|_{\neq 0}/\sqrt{N} \geq 2^{n-\epsilon n}$, but the hardness magnification demands a much sparser $|M|_{\neq 0} \leq 2^{cn}$ for $c < 1$ [CJW19]. We have established $\mathsf{quasi\text{-}NP} \not\subset \mathsf{TC}^0$

in Boolean circuit complexity. It might be far beyond our reach to demonstrate lower bounds of explicit problems beyond $O(\log n)$-depth or $O(\log n)$-space, say to prove $\mathsf{quasi\text{-}NP} \not\subset \mathsf{NC}^1$ and $\mathsf{quasi\text{-}NP} \not\subset \mathsf{L}$. As for algebraic circuit complexity, we have shown either $\mathsf{VP} \neq \mathsf{VNP}$ or $\forall k, \mathsf{quasi\text{-}NP} \not\subset \mathsf{NC}^k$. Extending Murray-Williams-Chen-Ren's easy witness lemmas and replacing the latter $\mathsf{quasi\text{-}NP} \not\subset \mathsf{NC}^k$ with $\mathsf{NP} \not\subset \mathsf{P}/\mathrm{poly}$ might establish $\mathsf{VP} \neq \mathsf{VNP}$.

# References


[AAB17] Benny Applebaum, Jonathan Avron, and Chris Brzuska. Arithmetic cryptography. *Journal of the ACM (JACM)*, 64(2):1–74, 2017.

[Aar16] Scott Aaronson. $P \stackrel{?}{=} NP$. In *Open problems in mathematics*, pages 1–122. Springer, 2016.

[AAT11] Mikhail Alekhnovich, Sanjeev Arora, and Iannis Tourlakis. Towards strong nonapproximability results in the Lovász-Schrijver hierarchy. *Computational Complexity*, 20(4):615–648, 2011.

[AB87] Noga Alon and Ravi B Boppana. The monotone circuit complexity of boolean functions. *Combinatorica*, 7(1):1–22, 1987.

[AB09] Sanjeev Arora and Boaz Barak. *Computational Complexity: a modern approach.* Cambridge University Press, 2009.

[ABE+05] Sanjeev Arora, Eli Berger, Hazan Elad, Guy Kindler, and Muli Safra. On nonapproximability for quadratic programs. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS'05)*, pages 206–215. IEEE, 2005.

[ABF+08] Misha Alekhnovich, Mark Braverman, Vitaly Feldman, Adam R Klivans, and Toniann Pitassi. The complexity of properly learning simple concept classes. *Journal of Computer and System Sciences*, 74(1):16–34, 2008.

[ABI86] Noga Alon, László Babai, and Alon Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of algorithms*, 7(4):567–583, 1986.

[ABL17] Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):50, 2017.

[ABLT06] Sanjeev Arora, Béla Bollobás, László Lovász, and Iannis Tourlakis. Proving integrality gaps without knowing the linear program. *Theory of Computing*, 2:19–51, 2006.

[ABN+92] Noga Alon, Jehoshua Bruck, Joseph Naor, Moni Naor, and Ron M. Roth. Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs. *IEEE Transactions on Information Theory*, 38(2):509–516, 1992.

[ABR16] Benny Applebaum, Andrej Bogdanov, and Alon Rosen. A dichotomy for local small-bias generators. *Journal of Cryptology*, 29(3):577–596, 2016.

[ABSRW02] Michael Alekhnovich, Eli Ben-Sasson, Alexander A Razborov, and Avi Wigderson. Space complexity in propositional calculus. *SIAM Journal on Computing*, 31(4):1184–1211, 2002.

[ABW10] Benny Applebaum, Boaz Barak, and Avi Wigderson. Public-key cryptography from different assumptions. In *42nd Annual ACM Symposium on Theory of Computing (STOC'10)*, pages 171–180. ACM, 2010.

[ACD+18] Martin R Albrecht, Benjamin R Curtis, Amit Deo, Alex Davidson, Rachel Player, Eamonn W Postlethwaite, Fernando Virdia, and Thomas Wunderer. Estimate all the {LWE, NTRU} schemes! In *11th International Conference on Security and Cryptography for Networks (SCN'18)*, pages 351–367. Springer, 2018.

[AD97] Miklós Ajtai and Cynthia Dwork. A public-key cryptosystem with worst-case/average-case equivalence. In *29th Annual ACM Symposium on Theory of Computing (STOC'97)*, pages 284–293, 1997.

[AG11] Sanjeev Arora and Rong Ge. New algorithms for learning in presence of errors. In *38th International Colloquium on Automata, Languages, and Programming (ICALP'11)*, pages 403–415. Springer, 2011.

[AGHP92] Noga Alon, Oded Goldreich, Johan Håstad, and René Peralta. Simple constructions of almost $k$-wise independent random variables. *Random Structures & Algorithms*, 3(3):289–304, 1992.

[AGK21] Jackson Abascal, Venkatesan Guruswami, and Pravesh K Kothari. Strongly refuting all semi-random boolean csps. In *32nd ACM-SIAM Symposium on Discrete Algorithms (SODA'21)*, pages 454–472. SIAM, 2021.

[AGM03] Noga Alon, Oded Goldreich, and Yishay Mansour. Almost $k$-wise independence versus $k$-wise independence. *Information Processing Letters*, 88(3):107–110, 2003.

[AGS21] Srinivasan Arunachalam, Alex Bredariol Grilo, and Aarthi Sundaram. Quantum hardness of learning shallow classical circuits. *SIAM Journal on Computing*, 50(3):972–1013, 2021.

[AH19] Albert Atserias and Tuomas Hakoniemi. Size-degree trade-offs for sums-of-squares and positivstellensatz proofs. In *34th Annual IEEE Conference on Computational Complexity (CCC'19)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[AHU74] Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman. *The design and analysis of computer algorithms*. Addison-Wesley, 1974.

[AIK06] Benny Applebaum, Yuval Ishai, and Eyal Kushilevitz. Cryptography in $\mathrm{NC}^0$. *SIAM Journal on Computing*, 36(4):845–888, 2006.

[Ajt83] Miklós Ajtai. $\Sigma_1^1$-formulae on finite structures. *Annals of pure and applied logic*, 24(1):1–48, 1983.

[Ajt96] Miklós Ajtai. Generating hard instances of lattice problems. In *28th annual ACM Symposium on Theory of Computing (STOC'96)*, pages 99–108, 1996.

[AKK99] Sanjeev Arora, David Karger, and Marek Karpinski. Polynomial time approximation schemes for dense instances of NP-hard problems. *Journal of computer and system sciences*, 58(1):193–210, 1999.

[AL88] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[AL18] Benny Applebaum and Shachar Lovett. Algebraic attacks against random local functions and their countermeasures. *SIAM Journal on Computing*, 47(1):52–79, 2018.

[Ale11] Michael Alekhnovich. More on average case vs approximation complexity. *Computational Complexity*, 20(4):755–786, 2011.

[ALN16] Albert Atserias, Massimo Lauria, and Jakob Nordström. Narrow proofs may be maximally long. *ACM Transactions on Computational Logic (TOCL)*, 17(3):19, 2016.

[Alo86] Noga Alon. Explicit construction of exponential sized families of k-independent sets. *Discrete Mathematics*, 58(2):191–193, 1986.

[AM20] Albert Atserias and Moritz Müller. Automating resolution is NP-hard. *Journal of the ACM (JACM)*, 67(5):1–17, 2020.

[AMMN06] Noga Alon, Konstantin Makarychev, Yury Makarychev, and Assaf Naor. Quadratic forms on graphs. *Inventiones mathematicae*, 163(3):499–522, 2006.

[AN06] Noga Alon and Assaf Naor. Approximating the cut-norm via Grothendieck's inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006.

[Ans00] Kurt M Anstreicher. The volumetric barrier for semidefinite programming. *Mathematics of Operations Research*, 25(3):365–380, 2000.

[AOW15] Sarah R Allen, Ryan O'Donnell, and David Witmer. How to refute a random csp. In *56th Annual IEEE Symposium on Foundations of Computer Science (FOCS'15)*, pages 689–708. IEEE, 2015.

[App13] Benny Applebaum. Pseudorandom generators with long stretch and low locality from random local one-way functions. *SIAM Journal on Computing*, 42(5):2008–2037, 2013.

[App16] Benny Applebaum. Cryptographic hardness of random local functions. *Computational Complexity*, 25(3):667–722, 2016.

[AR01] Michael Alekhnovich and Alexander A Razborov. Lower bounds for polynomial calculus: Non-binomial case. In *42nd Annual IEEE Symposium on Foundations of Computer Science (FOCS'01)*, pages 190–199. IEEE, 2001.

[AR16] Benny Applebaum and Pavel Raykov. Fast pseudorandom functions based on expander graphs. In *14th International Conference on Theory of Cryptography (TCC'16)*, pages 27–56. Springer, 2016.

[AS98] Sanjeev Arora and Shmuel Safra. Probabilistic checking of proofs: A new characterization of NP. *Journal of the ACM (JACM)*, 45(1):70–122, 1998.

[AW17] Josh Alman and Ryan Williams. Probabilistic rank and matrix rigidity. In *49th Annual ACM Symposium on Theory of Computing (STOC'17)*, pages 641–652, 2017.

[AW21] Josh Alman and Virginia Vassilevska Williams. A refined laser method and faster matrix multiplication. In *32nd ACM-SIAM Symposium on Discrete Algorithms (SODA'21)*, pages 522–539. SIAM, 2021.

[Bar89] David A Barrington. Bounded-width polynomial-size branching programs recognize exactly those languages in nc1. *Journal of Computer and System Sciences*, 38(1):150–164, 1989.

[BBB+00] Amos Beimel, Francesco Bergadano, Nader H Bshouty, Eyal Kushilevitz, and Stefano Varricchio. Learning functions represented as multiplicity automata. *Journal of the ACM (JACM)*, 47(3):506–530, 2000.

[BBCP20] Olaf Beyersdorff, Ilario Bonacina, Leroy Chew, and Jan Pich. Frege systems for quantified boolean logic. *Journal of the ACM (JACM)*, 67(2):1–36, 2020.

[BBKK18] Boaz Barak, Zvika Brakerski, Ilan Komargodski, and Pravesh K Kothari. Limits on low-degree pseudorandom generators (or: Sum-of-squares meets program obfuscation). In *37th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EuroCRYPT'18)*, pages 649–679. Springer, 2018.

[BBR94] David A Mix Barrington, Richard Beigel, and Steven Rudich. Representing boolean functions as polynomials modulo composite numbers. *Computational Complexity*, 4(4):367–382, 1994.

[BCG+20] Elette Boyle, Geoffroy Couteau, Niv Gilboa, Yuval Ishai, Lisa Kohl, and Peter Scholl. Correlated pseudorandom functions from variable-density lpn. In *61st Annual IEEE Symposium on Foundations of Computer Science (FOCS'20)*, pages 1069–1080. IEEE, 2020.

[BCH+19] Adam Bouland, Lijie Chen, Dhiraj Holden, Justin Thaler, and Prashant Nalini Vasudevan. On the power of statistical zero knowledge. *SIAM Journal on Computing*, 49(4):FOCS17–1–FOC17–58, 2019.

[BCK15] Boaz Barak, Siu On Chan, and Pravesh K Kothari. Sum of squares lower bounds from pairwise independence. In *47th annual ACM Symposium on Theory of Computing (STOC'15)*, pages 97–106. ACM, 2015.

[BCR20] Jonah Brown-Cohen and Prasad Raghavendra. Extended formulation lower bounds for refuting random csps. In *31st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'20)*, pages 305–324. SIAM, 2020.

[BD02] Avrim Blum and John Dunagan. Smoothed analysis of the perceptron algorithm for linear programming. In *13th annual ACM-SIAM Symposium on Discrete algorithms (SODA'02)*, pages 905–914. Society for Industrial and Applied Mathematics, 2002.

[BDLM01] Shai Ben-David, Philip M Long, and Yishay Mansour. Agnostic boosting. In *14th Annual Conference on Computational Learning Theory (COLT'01)*, pages 507–516. Springer, 2001.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[Bei93] Richard Beigel. The polynomial method in circuit complexity. In *8th Annual IEEE Conference on Structure in Complexity Theory (STC'08)*, pages 82–95. IEEE, 1993.

[Ber18] Christoph Berkholz. The relation between polynomial calculus, Sherali-Adams, and sum-of-squares proofs. In *35th Symposium on Theoretical Aspects of Computer Science (STACS'18)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[BFKL93] Avrim Blum, Merrick Furst, Michael Kearns, and Richard J Lipton. Cryptographic primitives based on hard learning problems. In *13th Annual International Cryptology Conference (CRYPTO'93)*, pages 278–291. Springer, 1993.

[BFKV98] Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1):35–52, 1998.

[BFS86] László Babai, Peter Frankl, and Janos Simon. Complexity classes in communication complexity theory. In *27th Annual IEEE Symposium on Foundations of Computer Science (FOCS'86)*, pages 337–347. IEEE, 1986.

[BGI+12] Boaz Barak, Oded Goldreich, Russell Impagliazzo, Steven Rudich, Amit Sahai, Salil Vadhan, and Ke Yang. On the (im) possibility of obfuscating programs. *Journal of the ACM (JACM)*, 59(2):1–48, 2012.

[BGL06] Nayantara Bhatnagar, Parikshit Gopalan, and Richard J Lipton. Symmetric polynomials over zm and simultaneous communication protocols. *Journal of Computer and System Sciences*, 72(2):252–285, 2006.

[BGMT12] Siavosh Benabbas, Konstantinos Georgiou, Avner Magen, and Madhur Tulsiani. SDP gaps from pairwise independence. *Theory of Computing*, 8(1):269–289, 2012.

[BGS98] Mihir Bellare, Oded Goldreich, and Madhu Sudan. Free bits, PCPs, and nonapproximability—towards tight results. *SIAM Journal on Computing*, 27(3):804–915, 1998.

[BHK+19] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019.

[BHL+02] Wolfgang Barthel, Alexander K Hartmann, Michele Leone, Federico Ricci-Tersenghi, Martin Weigt, and Riccardo Zecchina. Hiding solutions in random satisfiability problems: A statistical mechanics approach. *Physical review letters*, 88(18):188701, 2002.

[BIK+96] Samuel Buss, Russell Impagliazzo, Jan Krajíček, Pavel Pudlák, Alexander A. Razborov, and Jiri Sgall. Proof complexity in algebraic systems and bounded depth frege systems with modular counting. *Computational Complexity*, 6(3):256–298, 1996.

[BJS97] Roberto J Bayardo Jr and Robert Schrag. Using CSP look-back techniques to solve real-world SAT instances. In *Aaai/iaai*, pages 203–208. Providence, RI, 1997.

[BKPS98] Paul Beame, Richard Karp, Toniann Pitassi, and Michael Saks. On the complexity of unsatisfiability proofs for random $k$-CNF formulas. In *30th annual ACM Symposium on Theory of computing (STOC'98)*, pages 561–571. ACM, 1998.

[BKS13] Boaz Barak, Guy Kindler, and David Steurer. On the optimality of semidefinite relaxations for average-case and generalized constraint satisfaction. In *4th Conference on Innovations in Theoretical Computer Science (ITCS'13)*, pages 197–214. ACM, 2013.

[BKW03] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM (JACM)*, 50(4):506–519, 2003.

[BLP$^+$13] Zvika Brakerski, Adeline Langlois, Chris Peikert, Oded Regev, and Damien Stehlé. Classical hardness of learning with errors. In *45th annual ACM Symposium on Theory of Computing (STOC'13)*, pages 575–584, 2013.

[Blu83] Norbert Blum. A boolean function requiring 3n network size. *Theoretical Computer Science*, 28(3):337–345, 1983.

[BM12] Markus Bläser and Bodo Manthey. Smoothed complexity theory. In *37th International Symposium on Mathematical Foundations of Computer Science (MFCS'12)*, pages 198–209. Springer, 2012.

[BM16] Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *29th Annual Conference on Computational Learning Theory (COLT'16)*, pages 417–445. PMLR, 2016.

[BMOS05] Nader H Bshouty, Elchanan Mossel, Ryan O'Donnell, and Rocco A Servedio. Learning dnf from random walks. *Journal of Computer and System Sciences*, 71(3):250–265, 2005.

[BOCIP02] Josh Buresh-Oppenheim, Matthew Clegg, Russell Impagliazzo, and Toniann Pitassi. Homogenization and the polynomial calculus. *Computational Complexity*, 11(3-4):91–108, 2002.

[BOW10] Eric Blais, Ryan O'Donnell, and Karl Wimmer. Polynomial regression under arbitrary product distributions. *Machine learning*, 80(2-3):273–294, 2010.

[BP96] Paul Beame and Toniann Pitassi. Simplified and improved resolution lower bounds. In *37th Annual IEEE Symposium on Foundations of Computer Science (FOCS'96)*, pages 274–282. IEEE, 1996.

[BP98] Paul Beame and Toniann Pitassi. Propositional proof complexity: Past, present and future. *Bulletin of the European Association for Theoretical Computer Science*, 65:66–89, 1998.

[BPR12] Abhishek Banerjee, Chris Peikert, and Alon Rosen. Pseudorandom functions and lattices. In *31st Annual International Conference on the Theory and Applications of Cryptographic Techniques (EuroCRYPT'12)*, pages 719–737. Springer, 2012.

[BQ12] Andrej Bogdanov and Youming Qiao. On the security of Goldreich's one-way function. *Computational Complexity*, 21(1):83–127, 2012.

[BS83] Walter Baur and Volker Strassen. The complexity of partial derivatives. *Theoretical Computer Science*, 22(3):317–330, 1983.

[BS14] Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *Electronic Colloquium on Computational Complexity (ECCC)*, 21(59), 2014.

[BS15] Mark Bun and Thomas Steinke. Weighted polynomial approximations: Limits for learning and pseudorandomness. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM'15)*, pages 625–644, 2015.

[BSGH$^+$06] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of proximity, shorter PCPs, and applications to coding. *SIAM Journal on Computing*, 36(4):889–974, 2006.

[Bsh96] Nader H Bshouty. A subexponential exact learning algorithm for DNF using equivalence queries. *Information Processing Letters*, 59(1):37–39, 1996.

[BSI10] Eli Ben-Sasson and Russell Impagliazzo. Random CNF's are hard for the polynomial calculus. *Computational Complexity*, 19(4):501–519, 2010.

[BSS08] Eli Ben-Sasson and Madhu Sudan. Short pcps with polylog query complexity. *SIAM Journal on Computing*, 38(2):551–607, 2008.

[BSV14] Eli Ben-Sasson and Emanuele Viola. Short PCPs with projection queries. In *41st International Colloquium on Automata, Languages, and Programming (ICALP'14)*, pages 163–173. Springer, 2014.

[BSV19] Andrej Bogdanov, Manuel Sabin, and Prashant Nalini Vasudevan. XOR codes and sparse learning parity with noise. In *13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'19)*, pages 986–1004. SIAM, 2019.

[BSW01] Eli Ben-Sasson and Avi Wigderson. Short proofs are narrow 褰排 esolution made simple. *Journal of the ACM (JACM)*, 48(2):149–169, 2001.

[BT94] Richard Beigel and Jun Tarui. On ACC. *Computational Complexity*, 4(4):350–366, 1994.

[BT96] Nader H Bshouty and Christino Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM (JACM)*, 43(4):747–770, 1996.

[Bus91] Samuel R Buss. Propositional consistency proofs. *Annals of Pure and Applied Logic*, 52(1-2):3–29, 1991.

[Bus12] Samuel R Buss. Towards NP–P via proof complexity and search. *Annals of Pure and Applied Logic*, 163(7):906–917, 2012.

[BV06] René Beier and Berthold Vöcking. Typical properties of winners and losers in discrete optimization. *SIAM Journal on Computing*, 35(4):855–881, 2006.

[BV14] Zvika Brakerski and Vinod Vaikuntanathan. Efficient fully homomorphic encryption from (standard) LWE. *SIAM Journal on Computing*, 43(2):831–871, 2014.

[BV19] Mitali Bafna and Nikhil Vyas. Imperfect gaps in gap-ETH and PCPs. In *34th Annual IEEE Conference on Computational Complexity (CCC'19)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[Cai99] Jin-Yi Cai. Applications of a new transference theorem to Ajtai's connection factor. In *14th Annual IEEE Conference on Computational Complexity (CCC'99)*, pages 205–214. IEEE, 1999.

[CBDF+99] Nicolo Cesa-Bianchi, Eli Dichterman, Paul Fischer, Eli Shamir, and Hans Ulrich Simon. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM (JACM)*, 46(5):684–719, 1999.

[CEI96] Matthew Clegg, Jeffery Edmonds, and Russell Impagliazzo. Using the Groebner basis algorithm to find proofs of unsatisfiability. In *28th annual ACM Symposium on Theory of Computing (STOC'96)*, pages 174–183. ACM, 1996.

[CG89] Benny Chor and Oded Goldreich. On the power of two-point based sampling. *Journal of Complexity*, 5(1):96–106, 1989.

[CGH+85] Benny Chor, Oded Goldreich, Johan Håsted, Joel Freidmann, Steven Rudich, and Roman Smolensky. The bit extraction problem or t-resilient functions. In *26th Annual IEEE Symposium on Foundations of Computer Science (FOCS'85)*, pages 396–407. IEEE, 1985.

[Che52] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

[CIKK16] Marco L Carmosino, Russell Impagliazzo, Valentine Kabanets, and Antonina Kolokolova. Learning algorithms from natural proofs. In *31st Annual IEEE Conference on Computational Complexity (CCC'16)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[CKMZ83] Ashok K Chandra, Lawrence T Kou, George Markowsky, and Shmuel Zaks. On sets of boolean n-vectors with all $k$-projections surjective. *Acta Informatica*, 20(1):103–111, 1983.

[CKW11] Xi Chen, Neeraj Kayal, and Avi Wigderson. *Partial derivatives in arithmetic complexity and beyond*. Now Publishers Inc, 2011.

[CLRS16] Siu On Chan, James R Lee, Prasad Raghavendra, and David Steurer. Approximate constraint satisfaction requires large LP relaxations. *Journal of the ACM (JACM)*, 63(4):34, 2016.

[CM01] Mary Cryan and Peter Bro Miltersen. On pseudorandom generators in nc 0. In *International Symposium on Mathematical Foundations of Computer Science*, pages 272–284. Springer, 2001.

[CM17] Arkadev Chattopadhyay and Nikhil Mande. Weights at the bottom matter when the top is heavy. *Electronic Colloquium on Computational Complexity (ECCC)*, 24(83), 2017.

[CM21] Geoffroy Couteau and Pierre Meyer. Breaking the circuit size barrier for secure computation under quasi-polynomial lpn. In *50th Annual International Conference on the Theory and Applications of Cryptographic Techniques Advances in cryptology (EuroCRYPT'21)*, pages 842–870. Springer, 2021.

[CMM09] Moses Charikar, Konstantin Makarychev, and Yury Makarychev. Near-optimal algorithms for maximum constraint satisfaction problems. *ACM Transactions on Algorithms (TALG)*, 5(3):32, 2009.

[COCF10] Amin Coja-Oghlan, Colin Cooper, and Alan Frieze. An efficient sparse regularity concept. *SIAM Journal on Discrete Mathematics*, 23(4):2000–2034, 2010.

[COGL07] Amin Coja-Oghlan, Andreas Goerdt, and André Lanka. Strong refutation heuristics for random $k$-SAT. *Combinatorics, Probability and Computing*, 16(1):5–28, 2007.

[Coo72] Stephen A Cook. A hierarchy for nondeterministic time complexity. In *4th annual ACM Symposium on Theory of Computing (STOC'72)*, pages 187–192, 1972.

[COP16] Amin Coja-Oghlan and Konstantinos Panagiotou. The asymptotic $k$-SAT threshold. *Advances in Mathematics*, 288:985–1068, 2016.

[CR79] Stephen A Cook and Robert A Reckhow. The relative efficiency of propositional proof systems. *The Journal of Symbolic Logic*, 44(1):36–50, 1979.

[CR20] Lijie Chen and Hanlin Ren. Strong average-case lower bounds from non-trivial derandomization. In *52nd Annual ACM Symposium on Theory of Computing (STOC'20)*, pages 1327–1334, 2020.

[CS88] Vašek Chvátal and Endre Szemerédi. Many hard examples for resolution. *Journal of the ACM (JACM)*, 35(4):759–768, 1988.

[CW04] M Charikar and A Wirth. Maximizing quadratic programs: extending grothendieck's inequality. In *45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04)*, pages 54–60. IEEE, 2004.

[CW21] Brynmor Chapman and Ryan Williams. Smaller acc0 circuits for symmetric functions. *arXiv preprint arXiv:2107.04706*, 2021.

[Dan51] George B Dantzig. Maximization of a linear function of variables subject to linear inequalities. *Activity analysis of production and allocation*, 13:339–347, 1951.

[Dan17] Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, pages 690–696. PMLR, 2017.

[Das99] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual IEEE Symposium on Foundations of Computer Science (FOCS'99)*, pages 634–644. IEEE, 1999.

[DDFS14] Anindya De, Ilias Diakonikolas, Vitaly Feldman, and Rocco A Servedio. Nearly optimal solutions for the chow parameters problem and low-weight approximation of halfspaces. *Journal of the ACM (JACM)*, 61(2):1–36, 2014.

[Din07] Irit Dinur. The PCP theorem by gap amplification. *Journal of the ACM (JACM)*, 54(3):12–es, 2007.

[DJ19] Alexander Durgin and Brendan Juba. Hardness of improper one-sided learning of conjunctions for all uniformly falsifiable CSPs. In *30th International Conference on Algorithmic Learning Theory (ALT'19)*, pages 369–382, 2019.

[DKK+18] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'18)*, pages 2683–2702. Society for Industrial and Applied Mathematics, 2018.

[DKS18] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *50th Annual ACM Symposium on Theory of Computing (STOC'18)*, pages 1061–1073. ACM, 2018.

[DL19] Zeev Dvir and Allen Liu. Fourier and circulant matrices are not rigid. *34th Annual IEEE Conference on Computational Complexity (CCC'19)*, 2019.

[DLL62] Martin Davis, George Logemann, and Donald Loveland. A machine program for theorem-proving. *Communications of the ACM*, 5(7):394–397, 1962.

[DLSS14] Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *46th annual ACM Symposium on Theory of computing (STOC'14)*, pages 441–448, 2014.

[DMQN12] Nico Döttling, Jörn Müller-Quade, and Anderson CA Nascimento. IND-CCA secure cryptography based on a variant of the LPN problem. In *18th International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT'12)*, pages 485–503. Springer, 2012.

[DP60] Martin Davis and Hilary Putnam. A computing procedure for quantification theory. *Journal of the ACM (JACM)*, 7(3):201–215, 1960.

[DSS15] Jian Ding, Allan Sly, and Nike Sun. Proof of the satisfiability conjecture for large $k$. In *47th annual ACM Symposium on Theory of Computing (STOC'15)*, pages 59–68. ACM, 2015.

[DSS16] Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning DNF's. In *29th Annual Conference on Computational Learning Theory (COLT'16)*, pages 815–830. PMLR, 2016.

[DSSS21] Antonio J Di Scala, Carlo Sanna, and Edoardo Signorini. On the condition number of the vandermonde matrix of the nth cyclotomic polynomial. *Journal of Mathematical Cryptology*, 15(1):174–178, 2021.

[DW+00] Carlos Domingo, Osamu Watanabe, et al. Madaboost: A modification of adaboost. In *13th Annual Conference on Computational Learning Theory (COLT'00)*, pages 180–189, 2000.

[EL73] Paul Erdős and László Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. In *Colloquia Mathematica Societatis Janos Bolyai 10. Infinite and Finite Sets, Keszthely (Hungary)*. North-Holland, 1973.

[Erd59] Paul Erdös. Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38, 1959.

[Erd61] Paul Erdös. Graph theory and probability. II. *Canadian Journal of Mathematics*, 13:346–352, 1961.

[Fei02] Uriel Feige. Relations between average case complexity and approximation complexity. In *34th annual ACM Symposium on Theory of Computing (STOC'02)*, pages 534–543. ACM, 2002.

[Fei07] Uriel Feige. Refuting smoothed 3CNF formulas. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 407–417. IEEE, 2007.

[Fel06] Vitaly Feldman. Optimal hardness results for maximizing agreements with monomials. In *21st Annual IEEE Conference on Computational Complexity (CCC'06)*, pages 226–236. IEEE, 2006.

[Fel09] Vitaly Feldman. Hardness of approximate two-level logic minimization and PAC learning with membership queries. *Journal of Computer and System Sciences*, 75(1):13–26, 2009.

[Fel10] Vitaly Feldman. Distribution-specific agnostic boosting. In *First Conference on Innovations in Computer Science (ITCS'10)*, pages 241–250. ACM, 2010.

[Fel12] Vitaly Feldman. Learning DNF expressions from Fourier spectrum. In *25th Annual Conference on Learning Theory (COLT'12)*, pages 17.1–17.19. PMLR, 2012.

[FGK05] Joel Friedman, Andreas Goerdt, and Michael Krivelevich. Recognizing more unsatisfiable random $k$-SAT instances efficiently. *SIAM Journal on Computing*, 35(2):408–430, 2005.

[FGRW12] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

[FK09] Lance Fortnow and Adam R Klivans. Efficient learning algorithms yield circuit lower bounds. *Journal of Computer and System Sciences*, 75(1):27–36, 2009.

[FK15] Vitaly Feldman and Pravesh Kothari. Agnostic learning of disjunctions on symmetric distributions. *The Journal of Machine Learning Research*, 16(1):3455–3467, 2015.

[FLP16] Dimitris Fotakis, Michael Lampis, and Vangelis Th Paschos. Sub-exponential approximation schemes for CSPs: From dense to almost sparse. In *33rd Symposium on Theoretical Aspects of Computer Science (STACS'16)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[FO05] Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

[FO07] Uriel Feige and Eran Ofek. Easily refutable subformulas of large random 3CNF formulas. *Theory of Computing*, 3(1):25–43, 2007.

[FPV18] Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. *SIAM Journal on Computing*, 47(4):1294–1338, 2018.

[Fri99] Ehud Friedgut. Sharp thresholds of graph properties, and the $k$-SAT problem. *Journal of the American mathematical Society*, 12(4):1017–1054, 1999.

[FSO06] Jon Feldman, Rocco A Servedio, and Ryan O'Donnell. PAC learning axis-aligned mixtures of gaussians with no separation assumption. In *19th Annual Conference on Computational Learning Theory (COLT'06)*, pages 20–34. Springer, 2006.

[FSS84] Merrick Furst, James B Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical Systems Theory*, 17(1):13–27, 1984.

[GGM86] Oded Goldreich, Shafi Goldwasser, and Silvio Micali. How to construct random functions. *Journal of the ACM (JACM)*, 33(4):792–807, 1986.

[GHK15] Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *47th annual ACM Symposium on Theory of Computing (STOC'15)*, pages 761–770. ACM, 2015.

[GINX16] Nicolas Gama, Malika Izabachene, Phong Q Nguyen, and Xiang Xie. Structural lattice reduction: generalized worst-case to average-case reductions and homomorphic cryptosystems. In *45th Annual International Conference on the Theory and Applications of Cryptographic Techniques Advances in cryptology (EuroCRYPT'16)*, pages 528–558. Springer, 2016.

[GK01] Andreas Goerdt and Michael Krivelevich. Efficient recognition of random unsatisfiable $k$-SAT instances by spectral methods. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 294–304. Springer, 2001.

[GKK08] Parikshit Gopalan, Adam Tauman Kalai, and Adam R Klivans. Agnostically learning decision trees. In *40th annual ACM Symposium on Theory of Computing (STOC'08)*, pages 527–536. ACM, 2008.

[GKKS14] Ankit Gupta, Pritish Kamath, Neeraj Kayal, and Ramprasad Saptharishi. Approaching the chasm at depth four. *Journal of the ACM (JACM)*, 61(6):1–16, 2014.

[GKS20] Ankit Garg, Neeraj Kayal, and Chandan Saha. Learning sums of powers of low-degree polynomials in the non-degenerate case. In *61st Annual IEEE Symposium on Foundations of Computer Science (FOCS'20)*, pages 889–899. IEEE, 2020.

[GMKP20] Ignacio García-Marco, Pascal Koiran, and Timothée Pecatte. Reconstruction algorithms for sums of affine powers. *Journal of Symbolic Computation*, 98:284–318, 2020.

[Gol00] Oded Goldreich. Candidate one-way functions based on expander graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(90), 2000.

[GPV08] Craig Gentry, Chris Peikert, and Vinod Vaikuntanathan. Trapdoors for hard lattices and new cryptographic constructions. In *40th annual ACM Symposium on Theory of Computing (STOC'08)*, pages 197–206, 2008.

[GPW18] Mika Göös, Toniann Pitassi, and Thomas Watson. The landscape of communication complexity classes. *Computational Complexity*, 27(2):245–304, 2018.

[GR09] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[Gri01] Dima Grigoriev. Linear lower bound on degrees of positivstellensatz calculus proofs for the parity. *Theoretical Computer Science*, 259(1-2):613–622, 2001.

[Gro52] Alexander Grothendieck. Résumé des résultats essentiels dans la théorie des produits tensoriels topologiques et des espaces nucléaires. *Annales de l'Institut Fourier*, 4:73–112, 1952.

[GS21] Suprovat Ghoshal and Rishi Saket. Hardness of learning DNFs using halfspaces. In *53rd Annual ACM Symposium on Theory of Computing (STOC'21)*, pages 467–480, 2021.

[GST20] Nikhil Gupta, Chandan Saha, and Bhargav Thankey. A super-quadratic lower bound for depth four arithmetic circuits. *Electronic Colloquium on Computational Complexity (ECCC)*, 27(28), 2020.

[GT18] Oded Goldreich and Avishay Tal. Matrix rigidity of random toeplitz matrices. *Computational Complexity*, 27(2):305–350, 2018.

[GV01] Dima Grigoriev and Nicolai Vorobjov. Complexity of null–and positivstellensatz proofs. *Annals of Pure and Applied Logic*, 113(1-3):153–160, 2001.

[GW95] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[GW20] Oded Goldreich and Avi Wigderson. On the size of depth-three boolean circuits for computing multilinear functions. In *Computational Complexity and Property Testing - On the Interplay Between Randomness and Computation*, volume 12050 of *Lecture Notes in Computer Science*, pages 41–86. Springer, 2020.

[HAB02] William Hesse, Eric Allender, and David A Mix Barrington. Uniform constant-depth threshold circuits for division and iterated multiplication. *Journal of Computer and System Sciences*, 65(4):695–716, 2002.

[Hås86] Johan Håstad. Almost optimal lower bounds for small depth circuits. In *18th annual ACM Symposium on Theory of Computing (STOC'86)*, pages 6–20. ACM, 1986.

[Hås01] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM (JACM)*, 48(4):798–859, 2001.

[Hat06] Kohei Hatano. Smooth boosting using an information-based criterion. In *17th International Conference on Algorithmic Learning Theory (ALT'06)*, pages 304–318. Springer, 2006.

[Hau92] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

[HKP+17] Samuel B Hopkins, Pravesh K Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *58th Annual IEEE Symposium on Foundations of Computer Science (FOCS'17)*, pages 720–731. IEEE, 2017.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[Hop18] Samuel Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018.

[HS17] Shuichi Hirahara and Rahul Santhanam. On the average-case complexity of MCSP and its variants. In *32nd Annual IEEE Conference on Computational Complexity (CCC'17)*, pages 7:1–7:20. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[HST20] Samuel B Hopkins, Tselil Schramm, and Luca Trevisan. Subexponential LPs approximate Max-Cut. In *61st Annual IEEE Symposium on Foundations of Computer Science (FOCS'20)*, pages 943–953. IEEE, 2020.

[HY11] Pavel Hrubeš and Amir Yehudayoff. Arithmetic complexity in ring extensions. *Theory of Computing*, 7(1):119–129, 2011.

[Hya79] Laurent Hyafil. On the parallel evaluation of multivariate polynomials. *SIAM Journal on Computing*, 8(2):120–123, 1979.

[IM02] Kazuo Iwama and Hiroki Morizumi. An explicit lower bound of $5n-o(n)$ for boolean circuits. In *International Symposium on Mathematical Foundations of Computer Science*, pages 353–364. Springer, 2002.

[Imp95] Russell Impagliazzo. Hard-core distributions for somewhat hard problems. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 538–545. IEEE, 1995.

[IN96] Russell Impagliazzo and Moni Naor. Efficient cryptographic schemes provably as secure as subset sum. *Journal of cryptology*, 9(4):199–216, 1996.

[IP01] Russell Impagliazzo and Ramamohan Paturi. On the complexity of $k$-SAT. *Journal of Computer and System Sciences*, 62(2):367–375, 2001.

[IPS99] Russell Impagliazzo, Pavel Pudlák, and Jiri Sgall. Lower bounds for the polynomial calculus and the Gröbner basis algorithm. *Computational Complexity*, 8(2):127–144, 1999.

[IW97] Russell Impagliazzo and Avi Wigderson. P= BPP if E requires exponential circuits: Derandomizing the XOR lemma. In *29th annual ACM Symposium on Theory of Computing (STOC'97)*, pages 220–229, 1997.

[Jac97] Jeffrey C Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 3(55):414–440, 1997.

[JKL+20] Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song. A faster interior point method for semidefinite programming. In *61st Annual IEEE Symposium on Foundations of Computer Science (FOCS'20)*, pages 910–918, 2020.

[JLS21] Aayush Jain, Huijia Lin, and Amit Sahai. Indistinguishability obfuscation from well-founded assumptions. In *53rd Annual ACM Symposium on Theory of Computing (STOC'21)*, pages 60–73, 2021.

[JLSW20] Haotian Jiang, Yin Tat Lee, Zhao Song, and Sam Chiu-wai Wong. An improved cutting plane method for convex optimization, convex-concave games, and its applications. In *52nd Annual ACM Symposium on Theory of Computing (STOC'20)*, pages 944–953, 2020.

[JMS07] Haixia Jia, Cristopher Moore, and Doug Strain. Generating hard satisfiable formulas by hiding solutions deceptively. *Journal of Artificial Intelligence Research*, 28:107–118, 2007.

[Juk12] Stasys Jukna. *Boolean function complexity: advances and frontiers*, volume 27. Springer Science & Business Media, 2012.

[Kal85] KA Kalorkoti. A lower bound for the formula size of rational functions. *SIAM Journal on Computing*, 14(3):678–687, 1985.

[Kar84] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *16th annual ACM Symposium on Theory of Computing (STOC'84)*, pages 302–311. ACM, 1984.

[Kay12] Neeraj Kayal. Affine projections of polynomials. In *44th annual ACM symposium on Theory of computing (STOC'12)*, pages 643–662, 2012.

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

[Kha80] Khachiyan, L. G. Polynomial algorithms in linear programming. *USSR Computational Mathematics and Math. Phys.*, 20:53–72, 1980. (Russian original in Zhurnal Vychisditel'noi Matematiki i Matematicheskoi Fiziki, 20:51–68).

[Kha95] Michael Kharitonov. Cryptographic lower bounds for learnability of boolean functions on the uniform distribution. *Journal of Computer and System Sciences*, 50(3):600–610, 1995.

[KI04] Valentine Kabanets and Russell Impagliazzo. Derandomizing polynomial identity tests means proving circuit lower bounds. *computational complexity*, 13(1-2):1–46, 2004.

[KK09] Varun Kanade and Adam Kalai. Potential-based agnostic boosting. In *Advances in neural information processing systems*, pages 880–888, 2009.

[KKL88] J Kahn, G Kalai, and N Linial. The influence of variables on boolean functions. In *29th Annual IEEE Symposium on Foundations of Computer Science (FOCS'88)*, pages 68–80. IEEE Computer Society, 1988.

[KKMS08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KL93] Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[KL01] Matthias Krause and Stefan Lucks. On the minimal hardware complexity of pseudorandom function generators. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 419–430. Springer, 2001.

[KL18] Pravesh Kumar Kothari and Roi Livni. Agnostic learning by refuting. In *9th Conference on Innovations in Theoretical Computer Science (ITCS'18)*, page 55. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2018.

[Kla11] Hartmut Klauck. On arthur merlin games in communication complexity. In *26th Annual IEEE Conference on Computational Complexity (CCC'11)*, pages 189–199. IEEE, 2011.

[KLS96] Jeff Kahn, Nathan Linial, and Alex Samorodnitsky. Inclusion-exclusion: Exact and approximate. *Combinatorica*, 16(4):465–477, 1996.

[KLS09] Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10:2715–2740, 2009.

[KLSS17] Neeraj Kayal, Nutan Limaye, Chandan Saha, and Srikanth Srinivasan. An exponential lower bound for homogeneous depth four arithmetic formulas. *SIAM Journal on Computing*, 46(1):307–335, 2017.

[KLV94] Michael Kearns, Ming Li, and Leslie Valiant. Learning boolean formulas. *Journal of the ACM (JACM)*, 41(6):1298–1328, 1994.

[KM93] Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993.

[KMOW17] Pravesh K Kothari, Ryuhei Mori, Ryan O'Donnell, and David Witmer. Sum of squares lower bounds for refuting any CSP. In *49th Annual ACM Symposium on Theory of Computing (STOC'17)*, pages 132–145. ACM, 2017.

[KMR17] Pravesh K Kothari, Raghu Meka, and Prasad Raghavendra. Approximating rectangles by juntas and weakly-exponential lower bounds for LP relaxations of CSPs. In *49th Annual ACM Symposium on Theory of Computing (STOC'17)*, pages 590–603. ACM, 2017.

[KMSV13] Zohar S Karnin, Partha Mukhopadhyay, Amir Shpilka, and Ilya Volkovich. Deterministic identity testing of depth-4 multilinear circuits with bounded top fan-in. *SIAM Journal on Computing*, 42(6):2114–2131, 2013.

[KMZ14] Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Reweighted belief propagation and quiet planting for random $k$-SAT. *Journal on Satisfiability, Boolean Modeling and Computation*, 8:149–171, 2014.

[KOS04] Adam R Klivans, Ryan O'Donnell, and Rocco A Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840, 2004.

[Kra95] Jan Krajíček. On frege and extended frege proof systems. In *Feasible Mathematics II*, pages 284–319. Springer, 1995.

[KS73] Daniel J Kleitman and Joel Spencer. Families of $k$-independent sets. *Discrete mathematics*, 6(3):255–262, 1973.

[KS03] Adam R Klivans and Amir Shpilka. Learning arithmetic circuits via partial derivatives. In *Learning Theory and Kernel Machines*, pages 463–476. Springer, 2003.

[KS04] Adam R Klivans and Rocco A Servedio. Learning DNF in time $2^{\widetilde{O}(n^{1/3})}$. *Journal of Computer and System Sciences*, 68(2):303–318, 2004.

[KS05] Adam Tauman Kalaia and Rocco A Servediob. Boosting in the presence of noise. *Journal of Computer and System Sciences*, 71:266–290, 2005.

[KS06a] Adam Klivans and Amir Shpilka. Learning restricted models of arithmetic circuits. *Theory of computing*, 2(1):185–206, 2006.

[KS06b] Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 553–562. IEEE, 2006.

[KS08] Subhash Khot and Rishi Saket. Hardness of minimizing and learning DNF expressions. In *49th Annual IEEE Symposium on Foundations of Computer Science (FOCS'08)*, pages 231–240. IEEE, 2008.

[KS17] Mrinal Kumar and Shubhangi Saraf. On the power of homogeneous depth 4 arithmetic circuits. *SIAM Journal on Computing*, 46(1):336–387, 2017.

[KS19] Mrinal Kumar and Ramprasad Saptharishi. Hardness-randomness tradeoffs for algebraic computation. *Bulletin of EATCS*, 3(129), 2019.

[KSS94] Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[KSS14] Neeraj Kayal, Chandan Saha, and Ramprasad Saptharishi. A super-polynomial lower bound for regular arithmetic formulas. In *46th annual ACM symposium on Theory of Computing (STOC'14)*, pages 146–153, 2014.

[KST09] Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng. Learning and smoothed analysis. In *50th Annual IEEE Symposium on Foundations of Computer Science (FOCS'09).*, pages 395–404. IEEE, 2009.

[KST16] Neeraj Kayal, Chandan Saha, and Sébastien Tavenas. An almost cubic lower bound for depth three arithmetic circuits. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP'16)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[KV94] Michael Kearns and Leslie Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, 41(1):67–95, 1994.

[Las01] Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[Lau09] Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

[LM09] Vadim Lyubashevsky and Daniele Micciancio. On bounded distance decoding, unique shortest vectors, and the minimum distance problem. In *29th Annual International Cryptology Conference (CRYPTO'09)*, pages 577–594. Springer, 2009.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993.

[LN90] Nathan Linial and Noam Nisan. Approximate inclusion-exclusion. *Combinatorica*, 10(4):349–365, 1990.

[LNSS20] Guillaume Lagarde, Jakob Nordström, Dmitry Sokolov, and Joseph Swernofsky. Trade-offs between size and degree in polynomial calculus. In *11th Conference on Innovations in Theoretical Computer Science Conference (ITCS'20)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[Lok01] Satyanarayana V Lokam. Spectral methods for matrix rigidity with applications to size–depth trade-offs and communication complexity. *Journal of Computer and System Sciences*, 63(3):449–473, 2001.

[Lok08] Satyanarayana V Lokam. Complexity lower bounds using linear algebra. *Theoretical Computer Science*, 4(1-2):1–155, 2008.

[LPR13] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. A toolkit for ring-LWE cryptography. In *42nd Annual International Conference on the Theory and Applications of Cryptographic Techniques Advances in cryptology (EuroCRYPT'13)*, pages 35–54. Springer, 2013.

[LPS86] Alexander Lubotzky, Ralph Phillips, and Peter Sarnak. Explicit expanders and the ramanujan conjectures. In *18th annual ACM Symposium on Theory of Computing (STOC'86)*, pages 240–246, 1986.

[LRS15] James R Lee, Prasad Raghavendra, and David Steurer. Lower bounds on the size of semidefinite programming relaxations. In *47th annual ACM Symposium on Theory of Computing (STOC'15)*, pages 567–576. ACM, 2015.

[LS91] László Lovász and Alexander Schrijver. Cones of matrices and set-functions and 0–1 optimization. *SIAM journal on optimization*, 1(2):166–190, 1991.

[LS09] Nati Linial and Adi Shraibman. Learning complexity vs communication complexity. *Combinatorics, Probability and Computing*, 18(1-2):227–245, 2009.

[LST21] Nutan Limaye, Srikanth Srinivasan, and Sébastien Tavenas. Superpolynomial lower bounds against low-depth algebraic circuits. *Electronic Colloquium on Computational Complexity (ECCC)*, 28(81), 2021.

[LSW15] Yin Tat Lee, Aaron Sidford, and Sam Chiu-wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *56th Annual IEEE Symposium on Foundations of Computer Science (FOCS'15)*, pages 1049–1065. IEEE, 2015.

[Lub86] Michael Luby. A simple parallel algorithm for the maximal independent set problem. *SIAM journal on computing*, 15(4):1036–1053, 1986.

[LV17] Alex Lombardi and Vinod Vaikuntanathan. Limits on the locality of pseudorandom generators and applications to indistinguishability obfuscation. In *15th International Conference on Theory of Cryptography (TCC'17)*, pages 119–137. Springer, 2017.

[Man95] Yishay Mansour. An o (nlog log n) learning algorithm for DNF under the uniform distribution. *Journal of Computer and System Sciences*, 50(3):543–550, 1995.

[Meg01] Alexandre Megretski. Relaxations of quadratic programs in operator theory and system analysis. In *Systems, approximation, singular integral operators, and related topics*, pages 365–392. Springer, 2001.

[Mic02] Daniele Micciancio. Improved cryptographic hash functions with worst-case/average-case connection. In *34th annual ACM Symposium on Theory of computing (STOC'02)*, pages 609–618, 2002.

[Mit02] David G Mitchell. Resolution complexity of random constraints. In *International Conference on Principles and Practice of Constraint Programming*, pages 295–310. Springer, 2002.

[MMZ$^+$01] Matthew W Moskewicz, Conor F Madigan, Ying Zhao, Lintao Zhang, and Sharad Malik. Chaff: Engineering an efficient SAT solver. In *Proceedings of the 38th annual Design Automation Conference*, pages 530–535. ACM, 2001.

[MOS04] Elchanan Mossel, Ryan O'Donnell, and Rocco A Servedio. Learning functions of $k$ relevant variables. *Journal of Computer and System Sciences*, 69(3):421–434, 2004.

[MP12] Daniele Micciancio and Chris Peikert. Trapdoors for lattices: Simpler, tighter, faster, smaller. In *41st Annual International Conference on the Theory and Applications of Cryptographic Techniques Advances in cryptology (EuroCRYPT'12)*, pages 700–718. Springer, 2012.

[MP17] Marvin Minsky and Seymour A Papert. *Perceptrons: An introduction to computational geometry.* MIT press, 2017.

[MPZ02] Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

[MR07] Daniele Micciancio and Oded Regev. Worst-case to average-case reductions based on gaussian measures. *SIAM Journal on Computing*, 37(1):267–302, 2007.

[MR17] Pasin Manurangsi and Prasad Raghavendra. A birthday repetition theorem and complexity of approximating dense CSPs. In *44th International Colloquium on Automata, Languages, and Programming (ICALP'17)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[MS99] Joao Marques-Silva. The impact of branching heuristics in propositional satisfiability algorithms. In *Portuguese Conference on Artificial Intelligence*, pages 62–74. Springer, 1999.

[MS07] Michael Molloy and Mohammad R Salavatipour. The resolution complexity of random constraint satisfaction problems. *SIAM Journal on Computing*, 37(3):895–922, 2007.

[MST06] Elchanan Mossel, Amir Shpilka, and Luca Trevisan. On $\varepsilon$-biased generators in $\mathrm{NC}^0$. *Random Struct. Algorithms*, 29(1):56–81, 2006.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *51st Annual IEEE Symposium on Foundations of Computer Science (FOCS'10)*, pages 93–102. IEEE, 2010.

[MW19] Cody D Murray and R Ryan Williams. Circuit lower bounds for nondeterministic quasi-polytime from a new easy witness lemma. *SIAM Journal on Computing*, 49(5):STOC18–300, 2019.

[MYSSS21] Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. The connection between approximation, depth separation and learnability in neural networks. *arXiv preprint arXiv:2102.00434*, 2021.

[Nec66] Edward I Nechiporuk. A boolean function. *Engl. transl. in Sov. Phys. Dokl.*, 10:591–593, 1966.

[Nes00] Yurii Nesterov. Squared functional systems and optimization problems. In *High performance optimization*, pages 405–440. Springer, 2000.

[Nis91] Noam Nisan. Lower bounds for non-commutative computation. In *23rd annual ACM Symposium on Theory of computing (STOC'91)*, pages 410–418, 1991.

[NN93] Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM journal on computing*, 22(4):838–856, 1993.

[NN94] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Siam, 1994.

[Nor15] Jakob Nordström. On the interplay between proof complexity and SAT solving. *ACM SIGLOG News*, 2(3):19–44, 2015.

[NR04] Moni Naor and Omer Reingold. Number-theoretic constructions of efficient pseudo-random functions. *Journal of the ACM (JACM)*, 51(2):231–262, 2004.

[NS94] Noam Nisan and Mario Szegedy. On the degree of boolean functions as real polynomials. *Computational Complexity*, 4(4):301–313, 1994.

[NW94] Noam Nisan and Avi Wigderson. Hardness vs randomness. *Journal of computer and System Sciences*, 49(2):149–167, 1994.

[NW96] Noam Nisan and Avi Wigderson. Lower bounds on arithmetic circuits via partial derivatives. *Computational Complexity*, 6(3):217–234, 1996.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[OS07] Ryan O'Donnell and Rocco A Servedio. Learning monotone decision trees in polynomial time. *SIAM Journal on Computing*, 37(3):827–844, 2007.

[OS11] Ryan O'Donnell and Rocco A Servedio. The chow parameters problem. *SIAM Journal on Computing*, 40(1):165–199, 2011.

[OS17] Igor C Oliveira and Rahul Santhanam. Conspiracies between learning algorithms, circuit lower bounds, and pseudorandomness. In *32th Annual IEEE Conference on Computational Complexity (CCC'17)*, 2017.

[OS18] Igor Carboni Oliveira and Rahul Santhanam. Hardness magnification for natural problems. In *59th Annual IEEE Symposium on Foundations of Computer Science (FOCS'18)*, pages 65–76. IEEE, 2018.

[OSS19] Igor Carboni Oliveira, Rahul Santhanam, and Srikanth Srinivasan. Parity helps to compute majority. In *34th Computational Complexity Conference (CCC'19)*, pages 23:1–23:17. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[OW14] Ryan O'Donnell and David Witmer. Goldreich's PRG: Evidence for near-optimal polynomial stretch. In *29th Annual IEEE Conference on Computational Complexity (CCC'14)*, pages 1–12. IEEE Computer Society, 2014.

[Pap03] Christos H Papadimitriou. *Computational Complexity*. John Wiley and Sons Ltd., 2003.

[Par00] Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

[Pat92] Ramamohan Paturi. On the degree of polynomials that approximate symmetric boolean functions (preliminary version). In *24th annual ACM Symposium on Theory of computing (STOC'92)*, pages 468–474. ACM, 1992.

[Pei09] Chris Peikert. Public-key cryptosystems from the worst-case shortest vector problem. In *41st annual ACM Symposium on Theory of computing (STOC'09)*, pages 333–342, 2009.

[Pei14] Chris Peikert. Lattice cryptography for the internet. In *international workshop on post-quantum cryptography*, pages 197–219. Springer, 2014.

[Pfi76] Albrecht Pfister. Hilbert's seventeenth problem and related problems on definite forms. In *Mathematical Developments Arising from Hilbert Problems, Proceedings of Symposium in Pure Mathematics of the American Mathematical Society*, volume 28, pages 483–489, 1976.

[PP06] Ramamohan Paturi and Pavel Pudlák. Circuit lower bounds and linear codes. *Journal of Mathematical Sciences*, 134(5):2425–2434, 2006.

[PR07] Jean-Philippe Preaux and Jacques Raout. Generalized vandermonde's system and lagrange's interpolation. *arXiv preprint arXiv:0709.2153*, 2007.

[PRS88] Pavel Pudlák, Vojtěch Rödl, and Petr Savickỳ. Graph complexity. *Acta Informatica*, 25(5):515–535, 1988.

[PSS14] Periklis Papakonstantinou, Dominik Scheder, and Hao Song. Overlays and limited memory communication. In *29th Annual IEEE Conference on Computational Complexity (CCC'14)*, pages 298–308. IEEE, 2014.

[Pud94] Pavel Pudlak. Communication in bounded depth circuits. *Combinatorica*, 14(2):203–216, 1994.

[PV88] Leonard Pitt and Leslie G Valiant. Computational limitations on learning from examples. *Journal of the ACM (JACM)*, 35(4):965–984, 1988.

[Raz85] Alexander A Razborov. Lower bounds for the monotone complexity of some boolean functions. In *Soviet Math. Dokl.*, volume 31, pages 354–357, 1985.

[Raz87] Alexander A Razborov. Lower bounds on the size of bounded depth circuits over a complete basis with logical addition. *Mathematical Notes of the Academy of Sciences of the USSR*, 41(4):333–338, 1987.

[Raz89] Alexander A Razborov. On rigid matrices. Technical report, Steklov Mathematical Inssitute, 1989.

[Raz98] Ran Raz. A parallel repetition theorem. *SIAM Journal on Computing*, 27(3):763–803, 1998.

[Raz06] Ran Raz. Separation of multilinear circuit and formula size. *Theory of Computing*, 2(1):121–135, 2006.

[Raz09] Ran Raz. Multi-linear formulas for permanent and determinant are of super-polynomial size. *Journal of the ACM (JACM)*, 56(2):1–17, 2009.

[Raz10] Ran Raz. Elusive functions and lower bounds for arithmetic circuits. *Theory oF Computing*, 6:135–177, 2010.

[Raz13] Ran Raz. Tensor-rank and lower bounds for arithmetic formulas. *Journal of the ACM (JACM)*, 60(6):1–15, 2013.

[Reg04] Oded Regev. New lattice-based cryptographic constructions. *Journal of the ACM (JACM)*, 51(6):899–942, 2004.

[Reg09] Oded Regev. On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM (JACM)*, 56(6):34, 2009.

[Rob65] John Alan Robinson. A machine-oriented logic based on the resolution principle. *Journal of the ACM (JACM)*, 12(1):23–41, 1965.

[RR97] Alexander A Razborov and Steven Rudich. Natural proofs. *Journal of Computer and System Sciences*, 55(1):24–35, 1997.

[RRS17] Prasad Raghavendra, Satish Rao, and Tselil Schramm. Strongly refuting random csps below the spectral threshold. In *49th Annual ACM Symposium on Theory of Computing (STOC'17)*, pages 121–131. ACM, 2017.

[RS10a] Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *42nd annual ACM Symposium on Theory of computing (STOC'10)*, pages 755–764. ACM, 2010.

[RS10b] Alexander A Razborov and Alexander A Sherstov. The sign-rank of $\mathrm{AC}^0$. *SIAM Journal on Computing*, 39(5):1833–1855, 2010.

[RV07] Heiko Röglin and Berthold Vöcking. Smoothed analysis of integer programming. *Mathematical programming*, 110(1):21–56, 2007.

[RY09] Ran Raz and Amir Yehudayoff. Lower bounds and separations for constant depth multilinear circuits. *Computational Complexity*, 18(2):171–207, 2009.

[SA90] Hanif D Sherali and Warren P Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3(3):411–430, 1990.

[Sap14] Ramprasad Saptharishi. Recent progress on arithmetic circuit lower bounds. *Bulletin of EATCS*, 3(114), 2014.

[Sch77] A. A. Schönhage. Schnelle multiplikation von polynomen über Körpern der charakteristik 2. *Acta Informatica*, 7:395–398, 1977.

[Sch08] Grant Schoenebeck. Linear level lasserre lower bounds for certain $k$-CSPs. In *49th Annual IEEE Symposium on Foundations of Computer Science (FOCS'08)*, pages 593–602. IEEE, 2008.

[SCR+20] R Santhanam, L Chen, N Rajgopal, Ján Pich, IC Oliveira, and S Hirahara. Beyond natural proofs: Hardness magnification and locality. *Leibniz International Proceedings in Informatics*, 151, 2020.

[Ser03] Rocco A Servedio. Smooth boosting and learning with malicious noise. *Journal of Machine Learning Research*, 4(Sep):633–648, 2003.

[Ser04] Rocco A Servedio. On learning monotone DNF under product distributions. *Information and Computation*, 193(1):57–74, 2004.

[SFM78] Joel I Seiferas, Michael J Fischer, and Albert R Meyer. Separating nondeterministic time complexity classes. *Journal of the ACM (JACM)*, 25(1):146–167, 1978.

[Sha17] Abhijat Sharma. An improved lower bound for depth four arithmetic circuits. *Master's thesis, Indian Institute of Science, Bangalore, India*, 2017.

[Sha18] Ohad Shamir. Distribution-specific hardness of learning neural networks. *The Journal of Machine Learning Research*, 19(1):1135–1163, 2018.

[Sho87] Naum Z Shor. Class of global minimum bounds of polynomial functions. *Cybernetics*, 23(6):731–734, 1987.

[Sin16] Gaurav Sinha. Reconstruction of real depth-3 circuits with top fan-in 2. In *31st Conference on Computational Complexity (CCC'16)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[SM00] Yoshifumi Sakai and Akira Maruoka. Learning monotone log-term DNF formulas under the uniform distribution. *Theory of Computing Systems*, 33(1):17–33, 2000.

[Smo87] Roman Smolensky. Algebraic methods in the theory of lower bounds for boolean circuit complexity. In *19th annual ACM Symposium on Theory of computing (STOC'87)*, pages 77–82, 1987.

[SS71] Arnold Schönhage and Volker Strassen. Schnelle multiplikation grosser zahlen. *Computing*, 7(3):281–292, 1971.

[SS96] Victor Shoup and Roman Smolensky. Lower bounds for polynomial evaluation and interpolation problems. *Computational Complexity*, 6(4):301–311, 1996.

[SSS95] Jeanette P Schmidt, Alan Siegel, and Aravind Srinivasan. Chernoff–hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2):223–250, 1995.

[ST04] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

[ST09] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis: an attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, 52(10):76–84, 2009.

[Ste74] Gilbert Stengle. A nullstellensatz and a positivstellensatz in semialgebraic geometry. *Mathematische Annalen*, 207(2):87–97, 1974.

[Str69] Volker Strassen. Gaussian elimination is not optimal. *Numerische mathematik*, 13(4):354–356, 1969.

[Str73] Volker Strassen. Vermeidung von divisionen. *Journal für die reine und angewandte Mathematik*, 1973(264):184–202, 1973.

[STT07] Grant Schoenebeck, Luca Trevisan, and Madhur Tulsiani. Tight integrality gaps for Lovász-Schrijver LP relaxations of vertex cover and max cut. In *39th annual ACM Symposium on Theory of computing (STOC'07)*, pages 302–310. ACM, 2007.

[STV01] Madhu Sudan, Luca Trevisan, and Salil Vadhan. Pseudorandom generators without the XOR lemma. *Journal of Computer and System Sciences*, 62(2):236–266, 2001.

[SV18] Shubhangi Saraf and Ilya Volkovich. Black-box identity testing of depth-4 multilinear circuits. *Combinatorica*, 38(5):1205–1238, 2018.

[SW01] Amir Shpilka and Avi Wigderson. Depth-3 arithmetic circuits over fields of characteristic zero. *Computational Complexity*, 10(1):1–27, 2001.

[SY10] Amir Shpilka and Amir Yehudayoff. *Arithmetic circuits: A survey of recent results and open questions*. Now Publishers Inc, 2010.

[Tar93] Jun Tarui. Probabilistic polynomials, ac0 functions and the polynomial-time hierarchy. *Theoretical Computer Science*, 113(1):167–183, 1993.

[Tod91] Seinosuke Toda. Pp is as hard as the polynomial-time hierarchy. *SIAM Journal on Computing*, 20(5):865–877, 1991.

[Tsa96] Shi-Chun Tsai. Lower bounds on representing boolean functions as polynomials in z_m. *SIAM Journal on Discrete Mathematics*, 9(1):55–62, 1996.

[TT99] Jun Tarui and Tatsuie Tsukiji. Learning DNF by approximating inclusion-exclusion formulae. In *14th Annual IEEE Conference on Computational Complexity (CCC'99)*, pages 215–220. IEEE, 1999.

[TTV09] Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. Regularity, boosting, and efficiently simulating every high-entropy distribution. In *24th Annual IEEE Conference on Computational Complexity (CCC'09)*, pages 126–136. IEEE, 2009.

[Tul09] Madhur Tulsiani. CSP gaps and reductions in the Lasserre hierarchy. In *41st annual ACM Symposium on Theory of computing (STOC'09)*, pages 303–312. ACM, 2009.

[TW13] Madhur Tulsiani and Pratik Worah. Ls+ lower bounds from pairwise independence. In *28th Annual IEEE Conference on Computational Complexity (CCC'13)*, pages 121–132. IEEE, 2013.

[Vad17] Salil P Vadhan. On learning versus refutation. *Proceedings of Machine Learning Research vol*, 65:1–14, 2017.

[Vai90] Pravin M Vaidya. An algorithm for linear programming which requires $o(((m+n)n^2+(m+n)^{1.5}n)L)$ arithmetic operations. *Mathematical Programming*, 47(1-3):175–201, 1990.

[Val77] Leslie G Valiant. Graph-theoretic arguments in low-level complexity. In *International Symposium on Mathematical Foundations of Computer Science*, pages 162–176. Springer, 1977.

[Val79] Leslie G Valiant. Completeness classes in algebra. In *11th annual ACM Symposium on Theory of computing (STOC'79)*, pages 249–261, 1979.

[Val84] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Val85] Leslie G Valiant. Learning disjunction of conjunctions. In *9th International Joint Conference on Artificial Intelligence (IJCAI'85)*, pages 560–566, 1985.

[Val15] Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *Journal of the ACM (JACM)*, 62(2):1–45, 2015.

[Vap06] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[Vaz86] Umesh Vazirani. Randomness, adversaries and computation. *PhD thesis, University of California*, 1986.

[Ver90] Karsten A Verbeurgt. Learning DNF under the uniform distribution in quasi-polynomial time. In *3rd Annual Conference on Computational Learning Theory (COLT'90)*, pages 314–326. Springer, 1990.

[Vio05] Emanuele Viola. On constructing parallel pseudorandom generators from one-way functions. In *20th Annual IEEE Conference on Computational Complexity (CCC'05)*, pages 183–197. IEEE, 2005.

[VL91] Jan Van Leeuwen. *Handbook of theoretical computer science (vol. A) algorithms and complexity.* Mit Press, 1991.

[Vol16] Sergey Volkov. Finite bases with respect to the superposition in classes of elementary recursive functions, dissertation. *arXiv preprint arXiv:1611.04843*, 2016.

[VRPS21] Gal Vardi, Daniel Reichman, Toniann Pitassi, and Ohad Shamir. Size and depth separation in approximating natural functions with neural networks. *arXiv preprint arXiv:2102.00314*, 2021.

[VS20] Gal Vardi and Ohad Shamir. Neural networks with small weights and depth-separation barriers. *Advances in Neural Information Processing Systems*, 33, 2020.

[VW21] Nikhil Vyas and Ryan Williams. On super strong eth. *Journal of Artificial Intelligence Research*, 70:473–495, 2021.

[Weg87] Ingo Wegener. The complexity of boolean functions, 1987.

[Wig94] Avi Wigderson. NL/poly $\subseteq \oplus$L/poly. In *9th Annual IEEE Conference on Structure in Complexity Theory (SCT'09)*, pages 59–62. IEEE, 1994.

[Wig19] Avi Wigderson. *Mathematics and Computation: A Theory Revolutionizing Technology and Science*. Princeton University Press, 2019.

[Wil13] Ryan Williams. Improving exhaustive search implies superpolynomial lower bounds. *SIAM Journal on Computing*, 42(3):1218–1244, 2013.

[Wil14a] Ryan Williams. Nonuniform ACC circuit lower bounds. *Journal of the ACM (JACM)*, 61(1):1–32, 2014.

[Wil14b] Ryan Williams. The polynomial method in circuit complexity applied to algorithm design (invited talk). In *34th International Conference on Foundation of Software Technology and Theoretical Computer Science (FSTTCS'14)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

[Wil16] R Ryan Williams. Natural proofs versus derandomization. *SIAM Journal on Computing*, 45(2):497–529, 2016.

[Wit17] David Witmer. *Refutation of random constraint satisfaction problems using the sum of squares proof system.* PhD thesis, Technion–Israel Institute of Technology, 2017.

[Wun12] Henning Wunderlich. On a theorem of razborov. *Computational Complexity*, 21(3):431–477, 2012.

[Yan91] Mihalis Yannakakis. Expressing combinatorial optimization problems by linear programs. *Journal of Computer and System Sciences*, 43(3):441–466, 1991.

[Yao83] Andrew C Yao. Lower bounds by probabilistic arguments. In *24th Annual Symposium on Foundations of Computer Science (FOCS'83)*, pages 420–428. IEEE, 1983.

[Yao85] A. C.-C. Yao. Separating the polynomial-time hierarchy by oracles. In *26th Annual Symposium on Foundations of Computer Science (FOCS'85)*, pages 1–10. IEEE, IEEE, 1985.

[Yao90] AC-C Yao. On ACC and threshold circuits. In *31st Annual IEEE Symposium on Foundations of Computer Science (FOCS'90)*, pages 619–627. IEEE, 1990.

[Žák83] Stanislav Žák. A turing machine time hierarchy. *Theoretical Computer Science*, 26(3):327–333, 1983.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il