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 \( H_\infty(G) := -\log \min_y \Pr[G = y] \). The previous algorithms allowed \( H_\infty(G) = \text{poly}(n) \) for avoiding the worst-case intractability. We will derandomize them below \( H_\infty(G) \leq O(\log n) \) 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 \( \theta \circ x_i := \theta([x_i/2]) \oplus x_i \in \{0, 1\} \) in 1–4; \( \theta \circ x_i := \theta([x_i/2]) \cdot (-1)^{x_i} \in \mathbb{Z}_p \) of a large prime \( p \) in 5 and 8–10, and \( \theta \circ x_i := \theta_i \cdot [x_i/2] \cdot (-1)^{x_i} \) in 6–7. 11–13 will verify the unsatisfiability of small circuits in the worst case analysis \( (H_\infty(G) = 0) \). Suppose \( \log n \gg \log \frac{2}{\tau} + k \). Randomly pick an example \((X, Y)\) from the observed \( m \) data.

1. At \( H_\infty(G) = 0 \), MaxkCSP of any \( k \)-variate predicate \( f \) requires the sample size \( \Omega(n^{(k-1)/2}) \) \( \leq m \leq O(n^{k/2}) \) 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 \( n^\Omega(k) \) time by given access to both samplers \( P(x, y) \) and \( P'(x, y) \).

2. At \( H_\infty(G) = \text{log} s \), the planted \( s \)-term DNF demands \( m \geq n^\Omega(\log s) \) for \( c < 1 \), and \( m \leq n^2 \log s + o(1) \) for \( c > 1 \) to make \( n^\Omega(\log s) \)-time PAC learning (even under a slight noise).

3. At \( H_\infty(G) = \text{log} 1/\varepsilon \), planted AND needs \( m \geq n^\Omega(\log \frac{1}{\varepsilon}) \) for \( c < 1 \), and \( m \leq n \log \frac{1}{\varepsilon} + O(1) \) for \( c > 1 \), to make \( \max \theta P(y = f_\theta(G(x))) + \varepsilon \)-accurate agnostic learning in \( n^\Omega(\log 1/\varepsilon) \) time.

4. At \( H_\infty(G) = O(\log s) \), the monotone DNF with expanding \( s \)-terms is PAC learnable from \( m = n \cdot \text{poly}(s) \) data with \( \Pr[|X_i/2|, |X_r/2|] \geq 1/n^{1+\varepsilon} \) in \( n \cdot O(\text{log} d) \) time.

5. At \( H_\infty(G) = O(\log p) \), the kFT \( f_\theta(x) = \sum_{|w| \leq k} \hat{f}_w \prod_{i \in w} \theta \circ x_i \) over \( \mathbb{Z}_p \) of \( p \geq n^3 \) is invertible from \( m = O(n^{k+2}) \) data with \( \Pr[|X_{ij}/2|, \ldots, |X_{ik}/2|] \geq \Omega(1/n^k) \), \( |Y| \leq r \leq p^{1/2^{k+1}} \) and \( \Pr[Y \neq f_\theta(X)] \leq |X_{ij}/2|, \ldots, |X_{ik}/2|] \ll 1/(nr) \) in \( O(n^{k+3}) \) time.

6. LPN and LWE over \( \mathbb{Z}_p \) of \( p \geq n^{10} \) hiding small secrets \( \forall i, \theta_i = O(1) \) are breakable in polynomial time. 7. GapSVP \( O(n^2) \) is breakable within polynomial time.

8. At \( H_\infty(G) = O(\log n) \), any bilinear form \( \sum_{i,j=1}^n x_i M_{ij} x_j \) with sparsity \( |\{M_{ij} \in \{-1, 0, 1\} | M_{ij} \neq 0\}| \leq n^{2-o(1)} \) requires \( \Omega(n(\log n^{1-\varepsilon})) \) size for algebraic NC1 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 \( H_\infty(G) = O(n) \), any \( 2^2 \) by \( 2^2 \) matrix with sparsity \( 2^{n-n^\varepsilon} \) demands \( \exp(n^{O(1)}) \) size \( \text{PH}^k \) protocol unless it is learnable from \( m = \exp(n^\varepsilon) \) data in \( \exp(n^\varepsilon) \) time.

10. \( \text{PH}^k \neq \text{PSPACE}^k \) or \( \forall k, \text{NP} \not\subseteq \text{DEP}[k \log n] \). 11. \( \text{VP} \neq \text{VNP} \) or \( \forall k, \text{quasi-NP} \not\subseteq \text{NC}^k \).

12. \( \text{PTIME} \subseteq \text{DTIME}[n^{\text{poly} (\log n)}] \) or \( \forall \varepsilon, \forall k, \text{NTIME}[2^n] \not\subseteq \text{SIZE}[n^k] \). 13. \( \text{quasi-NP} \not\subseteq \text{TC}^0 \).

*School of Science and Engineering, Tokyo Denki University, Ishizaka, Hatoyama, Hiki, Saitama, 359-0394, Japan. Email: tsukiji@mail.dendai.ac.jp.
Keywords: agnostic learning, Algebraic circuits, Approximation 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 \( \mathcal{D} = \{z(1), \ldots, z(m)\} \) 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 \( \text{poly}(n) \) 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 \( i \in \{1, \ldots, n\} \).

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 \( y(j) = f_\theta(x(j)) = \text{sgn}(\sum_{i=1}^n \theta_i x_i(j) + \theta_0) \) 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

---

1A dataset may contain the same data multiple times.
$\theta \in \mathbb{R}^{n+1}$ [MP17]. Under a small additive Gaussian perturbation $(x(j) + G, f_\theta(x(j) + G))$, 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 $f : \{0, 1\}^n \rightarrow \{0, 1\}$ from a supervised dataset [Val84]. Elementary yet general, the most studied concept is a Disjunctive Normal Form (DNF) $f(x) := \bigvee_{s=1}^{n} \bigwedge_{i \in f_s} x_i \oplus f_{s_0}$ of given $f_s \subseteq \{1, \ldots, d\}$ and $f_{s_0} \subseteq \{0, 1\}$. We write it as $f \in s$-term DNF, and $f \in s$-term $k$DNF$_{d}$ when $\forall |f_s| = k$ [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 $\mu \in [c, 1-c]^{n}$ [Ver90]. Kalai, Samorodnitsky, and Teng [KST09, Fel12] proved that DNF is polynomial-time learnable under the product distribution of mean $\mu + G$ perturbed by the uniform random vector $G \in [-\frac{1}{2}, \frac{1}{2}]^{n}$.

**Review4:** Low-degree Fourier inversion is the most successful method of learning DNF over the real-number field $\mathbb{R}$: Learn $Y \approx \sum_{\|w\| \leq k} \theta_w \prod_{i \in w} (-1)^{X_i}$ by inverting the unknown $\theta_w \approx \mathbb{E}_X [Y \prod_{i \in w} (-1)^{X_i}]$ 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 $\mathcal{AC}^0$ [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 $\text{err}(D) = \min_{f} \text{err}_f(D)$ of the observed error rate $\text{err}_f(D) := \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}[f(x(j)) \neq y(j)]$ 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:** RkSAT refutation denies the existence of an assignment $\theta \in \{0, 1\}^n$ satisfying the OR predicate $f_\theta(x) := \bigvee_{i=1}^{k} \theta \circ x_i = \bigvee_{i=1}^{k} \theta([x_i/2]) \oplus x_i$ for all constraints$^2$ $x \in \mathcal{U} \subseteq [2^n]^{k} := \{0, 1, \ldots, 2^n-1\}$ drawn from $P(x) = \frac{1}{2^n \cdot \mathcal{U}}$, i.e., disprove $\text{err}(\mathcal{U}) = \min_{\theta} \text{err}_\theta(\mathcal{U}) = 0$ [CS88, Fei02]. It has noticed a constant $\alpha_k \approx 2^k \ln 2 - (1-\ln 2)/2$ to make a sharp threshold $\forall \varepsilon > 0, \lim_{n} \text{Pr}[\text{err}(\mathcal{U}) = 1 | m/n \geq \alpha_k - \varepsilon] = 1 = \lim_{n} \text{Pr}[\text{err}(\mathcal{U}) = 0 | m/n \leq \alpha_k + \varepsilon]$ [Fri99, MPZ02, DSS15, COP16]. Moreover, its data size complexity $m = \min |\mathcal{U}|$ of efficient refutation has attained the following dichotomy. RkSAT refutation is polynomial-time solvable above $m \geq O(2^{k(n)/2})$ [GK01, FGK05, COGL07, FO07, COCF10, BM16, AOW15], but demanding exp($n^{O(1)}$) proof-length or $n^{O(1)}$ proof-degree below $m \leq n^{(1-\varepsilon)/2}$ in the well-studied proof systems [BKPS98, AR01, BSW01, Gri01, Sch08, Tul09, BSI10, CLRS16, KMOV17, 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 $x_i(j) \mapsto 2[x_i(j)/2] + x_i(j) \oplus G_{ij}(\lceil x_i(j)/2 \rceil)$ by a flipper $G \in \{0, 1\}^{nm}$ with a small mean $\mathbb{E}[G_{ij}(\lceil x_i(j)/2 \rceil)] = \varepsilon$. Abiscon, Gurwushi, and Kothei [AGK21] recently generalized it to the kCSP refutation targeting the $f_\theta(x) = f(\theta \circ x_1, \ldots, \theta \circ x_k)$ of an arbitrary $k$-variable Boolean predicate $f(x_1, \ldots, x_k)$.

**Review7:** MaxkSAT approximation aims to measure the empirical accuracy rate $\text{acc}(D) := \max_{\theta} \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}[f_\theta(x(j)) = 1] = 1 - \text{err}(D)$ of the OR predicate $f_\theta(x) = \bigvee_{i=1}^{k} \theta \circ x_i$ on the

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


We may sometimes consider $f$-gap approximation to distinguish between $\text{acc}(\mathcal{D}) \leq \beta_{\text{and}}$ and $f_{\text{gap}} \leq \text{acc}(\mathcal{D})$ of $0 < \beta_{\text{and}} < f_{\text{gap}} \leq 1$. The $(1, 1 - 1/2 + \epsilon)$-gap approximation is hard on the $\mathbb{P} \neq \mathbb{NP}$ assumption [BGS98, Raz98, Hås01]. Exponential Time Hypothesis (ETH) conjectures that 3SAT must take $\exp(n)$ time to distinguish between $\text{acc}(\mathcal{D}) = 1$ and $\neq 1$ [IP01], which obliges 3SAT to take $\exp(n/\log(n))$ 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, . . . [Ale11]. RkCSP’s refutation of [AGK21] gives rise to the promise-MaxCSP’s approximation by only $\tilde{O}(n^{k/2})$ constraints in $n^{O(k)}$ time.

**Review8:** Planted CSP asks to invert the secret assignment $\theta \in \{0, 1\}^n$ 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 $P(x) = 1/|\{2n\}|$ [Gold00, AIK06, App13, OW14, BBKK18, AL18, FPV18]. It involves several well-studied inversion problems, e.g., the planted dSAT by $f = \bigvee_{i=1}^d x_i$, [BHL+02, JMS07, KMZ14], the noisy dLIN4 over $\mathbb{F}_2$ by $(-1)^f \approx \sum_{i} \prod_{j} (-1)^{x_i}$ [BFKL03, ABW10, Ale11, DMQN12, BS19], and the planted kDNF by $f = \bigvee_{j=1}^k \bigwedge_{i=1}^n x_{i+j}$ [DSS16].

**Review9:** LPN and LWE5 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 $(x_{i}(j))_{i,j} \in \mathbb{Z}_{q \times m}$ contaminated by i.i.d. errors $E(j) \in \mathbb{Z}_q$. LPN supposes the uniform random matrix with Bernoulli noise $\Pr[E(j) \neq 0] = \mu$ [AAB17, BCG+20, CM21, JLS21], while LWE treats Gaussian error $\Pr[E(j)] = 1/\sqrt{2\pi\gamma} e^{-E(j)^2/(2\gamma)}$ [AD97, Reg04, KS06b, BV14]. The current best attackers take min$(q^{O(n \log n)}, 2^{O(\sqrt{n})})$ 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., GapSVP 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 $\Pr_{I,J}[\mathcal{M}(I, J) = \sum_{\kappa=1}^n \theta(I, \kappa)\mathcal{M}(\kappa, J)] \approx 1$. 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 $\Pr[Y = \sum_{j=1}^n \theta(\langle X_j \rangle)|-\langle X \rangle|] \approx 1$ of a randomly picked example $(X, Y) \sim \{(2i + 1 - \mathcal{M}(i, j))/2 \mid i \in \{n\}, \mathcal{M}(i, j) \neq 0\}, \mathcal{M}(\kappa, j)]_{j=1}^m$. A more general problem asks to invert a secret $\theta \in \mathbb{Z}_q^n$ of a “noisy” degree-$k$ Fourier transform $\Pr[Y = \sum_{|w| \leq k} \hat{f}_w \prod_{i \in w} \theta([-X_i/2]) \langle -X \rangle \approx 1$ of the known coefficients $\hat{f}_w \in \mathbb{F}$.

**Review11:** Boolean Circuit lower bounds have brought learnability7, and vice versa

---

3We 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 $\theta \in \{0, 1\}^d$.

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


6GapSVP, poses a lattice $\mathcal{L} \in \mathbb{Z}^n$ together with an integer $d$ and asks to distinguish between $v(\mathcal{L}) \leq d$ and $v(\mathcal{L}) > d$ for the hidden shortest vector’s length $v(\mathcal{L}) := \min_{e \in \mathcal{L} \setminus \{0\}} \|e\|_2$.

7Learnable, 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 $AC^0$’s learnability from circuit lower bounds [Ajt83, Yao83, FSS84, Has86]. They inverted Review 4’s Fourier coefficients from quasi-polynomially many data of a low-degree polynomial over $\mathbb{R}$ derived from the $AC^0$’s circuit lower bound. Carmosino, Impagliazzo, Kabanets, and Kolokolova [CIKK16] did it on $AC^0[p]$ lower bounds [Raz87, Smo87] via Nisan-Wigderson’s pseudorandom generator (PRG) [NW94]. Murray and Williams [Wil13, Wil14a, MW19] established quasi-$NP \not\subseteq ACC$ by learning ACC from quasi-polynomially many data thanks to Beigel-Tarui’s low-degree $SYM^+$-computation. The polynomial method [Bei93, Wil14b, Hop18] was a consistent mechanics to make these lower bounds work.

**Review 12:** Algebraic circuit lower bounds have been interplaying with derandomization and learnability [SY10, CKW11, Sap14, KS19, GKS20]. Kabanets and Impagliazzo [KI04] derandomized PIT$^8$. It plugged (unproved) exponential$^{10}$ circuit size lower bounds of explicit multilinear polynomials to Nisan-Wigderson’s PRG. Low-rank patriarch derivatives [Nis91, NW96, KS03, KSS14, GKK14] have brought constant-depth circuit size lower bounds [SW01, RY09, KST16, KLSS17, KS17, LST21] and multilinear formulas [SS96, Raz90], derandomized the PIT of constant-depth circuits [KMSV13, SV18, LST21], non-properly learned multilinear depth-three circuits [BBB00, KS06a], and properly learned restricted depth-three circuits [Kay12, Sin16, KS19, GKS20, GMKP20].

Let us view these previous works of learning a dataset $D = \{(x(1), y(1)), \ldots, (x(m), y(m))\}$ through the lens of smoothed analysis. First, SA1 chooses an adversarial variate (marginal) distribution $P(x) = \sum_y P(x, y)$. SA2 disturbs the $P(x) \rightarrow P(G(x))$ by a random perturbation $G \in \mathcal{G}$ while preserving the covariate distribution $P_\theta(y|x)$ intact. SA3 generates a dataset $D$ from the perturbed distribution $P(G(x))P_\theta(y|G(x))$. Its density is the supremum of probability mass $\rho(G) = \sup \{Pr[G(x)|x] | P(x) > 0\}$ [BV06, RV07, BM12], and its min-entropy is $H_\infty(G) = -\log \rho(G)$. In particular, the worst-case complexity assumes $H_\infty(G) = 0$, while a smoothed one $H_\infty(G) \leq -\log |\mathcal{G}|$ with the equality $H_\infty(G) = -\log |\mathcal{G}| \iff \forall g, Pr[G = g] = 1/|\mathcal{G}|$. 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 $H_\infty(G) \approx mn$ 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 $H_\infty(G)$ to reduce the worst-case complexity in smoothed analysis. Reviews 1 and 2 are exponentially hard at $H_\infty(G) = 0$ but polynomial-time solvable under the Gaussian perturbation $H_\infty(G) = \frac{n}{2} \log \frac{1}{2e}$. Review 3’s DNF learning is intractable$^{11}$ at $H_\infty(G) = 0$ due to the hardness of learning [DSS16]’s canonical DNF in Review 8, but tractable under the perturbed product distribution $H_\infty(G) = \Theta(mn)$, and even under the random walk $H_\infty(G) = \Theta(m \log n)$. Review 6’s 3SAT refutation is coNP-complete at $H_\infty(G) = 0$ [Coo71], but efficiently solvable under the flipper $H_\infty(G) = \tilde{\Theta}(2^{\frac{1}{2} + (1 - \frac{1}{2}) \log(1 - \frac{1}{2})})$. Similarly, Review 7’s MaxSAT approximation is NP-complete at $H_\infty(G) = 0$ but tractable under the dense constraints $H_\infty(G) = \Theta(kn^k \log n)$. Exceptionally, Review 9’s LPN and LWE are still intractable even for the uniform random matrix $H_\infty(G) = mn \log n$. These worst-case intractable but average-case tractable problems separate unlearnable from learnable by $H_\infty(G) = 0$ versus $H_\infty(G) = \text{poly}(n)$. Derandomization effort might reduce this

---

$^8$SYM$^+$: Boolean functions $g(f(x))$ of a polynomial $f$ over $\mathbb{Z}$ and $g : \mathbb{Z} \rightarrow \{0, 1\}$. quasi-$NP$: 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 $H_x(G) = 0$ versus $H_x(G) \leq O(\log n)$. In that case, the known average-case algorithms might learn DNF, approximate MaxkSAT, 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 quasi-NP $\not\subseteq$ ACC. It motivates us to investigate these smoothed complexities by fixing the min-entropy somewhere\(^{12}\) between $0 \leq H_x(G) \leq O(\log n)$.

In this paper, we prove the following Theorems 1.1–1.3 in the asymptotic analysis on the problem’s increasing magnitudes $d, k, n, p, q, s, c$ and $1/\varepsilon$ under dominance\(^{13}\) $k + \log(ds/\varepsilon) \ll \log n$, $s/\varepsilon \leq d^{O(1)}$ and $1 \ll p, q \leq n^{O(1)}$. Our learnability proofs of the smoothed analysis may pick an appropriate perturbation $G$. The unlearnability ones must endure any considerable $G$.

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

**Theorem 1.1** (promise-MaxkCSP, informal). For any $k$-variable predicate $f$, distinguishing between $|\max_x P(y = f(x)) - \max_{x'} P'(y = f(x'))| = \Omega(1)$ and $P(x, y) \equiv P'(x, y)$ on $n^{1-\varepsilon}k$ data must take $\Omega(\exp(n^c))$ 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 $H_x(G) \leq (1 - \varepsilon) \log s$, PAC learning the planted $s$-term DNF on $n^{\Omega(\log s)}$ data must consume $\Omega(\exp(n^c))$ time. At $H_x(G) = \log s + O(\log \log n)$, it becomes PAC learnable from $n^{1/2 \log s + O(1)}$ data in $n^{O(\log s)}$ time.

Similarly, the agnostic learnability of the planted AND (Boolean conjunction) emerges at $H_x(G) \approx \log(1/\varepsilon)$ to achieve the prediction accuracy $\max_x P(y = f(x)) + \varepsilon$.

**Theorem 1.3** (agnostically learning the planted AND, informal). Below $H_x(G) \leq (1 - \varepsilon) \log 1/\varepsilon$, agnostic learning the planted AND on $n^{\Omega(\log 1/\varepsilon)}$ data demands $\Omega(\exp(n^c))$ time. At $H_x(G) = \log 1/\varepsilon + O(\log \log n)$, it is agnostically learnable from $n^{1/2 \log 1/\varepsilon + O(1)}$ data in $n^{O(\log 1/\varepsilon)}$ time.

When the min-entropy goes beyond $\log s$, Theorem 1.2’s data size barrier $n^{\Omega(\log s)}$ becomes breakable into a linear term of $n$ for the “monotone”\(^{15}\) DNF with “expanding”\(^{16}\) terms.

**Theorem 1.4** (PAC learning monotone DNF). At $H_x(G) = O(\log s)$, the planted monotone DNF 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 $n\tilde{O}(s^{\log d})$ time on $n \cdot \text{poly}(s)$ 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) $f(x) := \sum_{w \leq k} f_w \prod_{i \in w} \theta \circ x_i, \theta \circ x_i = \theta(\langle x_i/2 \rangle(1 - x_i))$ 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 $H_x(G) = 0$ unless mentioning on $G$ nor $H_x(G)$ 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 $\Omega(\exp(n^c))$ lengths or $\Omega(n^c)$ degrees in several well-studied weak proof systems.

\(^{15}\)In REVIEW3’s terminology, $f$ in DNF is monotone if $i \in f_s \land f_x \rightarrow f_s = f_x$.

\(^{16}\)We say that a DNF $f$ is $c$-wisely $k$-expanding if $|f_{\xi_1} \cup \cdots \cup f_{\xi_k}| \geq ck$ for every distinct $\xi_1, \ldots, \xi_k$.

\(^{17}\)A random variable $X \sim [2n]^d$ has pairwisely dense attributes if $\forall(i \neq i'), \Pr([X_i/2], [X_{i',2}]) \geq \Omega(1/\text{poly})$. 

6
Theorem 1.5 (inverting degree-$k$ planted FT). Let $1 \leq k \leq O(1)$, $n^{2+1/2^k-1} \ll q \in 2N + 1$, and $r = q^{1/2^k+1}$. At $H_n(G) = O(\log(nq))$, 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 $|Y| \leq r$. The variate must be as $k$-wisely sparse and noiseless at every location $(w,a) \in (\mathbb{Z}_q)^2 \times \{0,1\}^k$ as $\Pr[\forall i \in w, |X_i/2| = a_i] \geq \Omega(1/n^k)$ and $\Pr[|Y| \neq f(X) \mid \forall i \in w, |X_i/2| = a_i] \ll 1/(nr)$.

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

Theorem 1.7 (breaking GapSVP). GapSVP$_{O(n^2)}$ 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 $\rho(M) := |M|_{\neq 0}/N = |\{(i,j) \mid M(i,j) \neq 0\}|/N$. We say that an algorithm $A$ learns the matrix $M$ under $G$ if $A$ feeds the first $o(N^2)$ entries of the perturbed matrix $G(M)$ and predicts $\Pr_{G,I,J}[A(I,J) = G(M)(I,J)] \approx 1$. 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 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 $H_n(G) = O(\log n)$, the bilinear form of any $n \times n \{−1, 0, 1\}$-matrix having density $n^{-o(1)}$ requires $\Omega(n(\log \log n)^1 - c)$ size algebraic $NC^1$ circuits over $\mathbb{F}_p$ of any prime $p \geq n^{\Omega(1)}$, unless it is learnable in $n^{o(1)}$ time.

Theorem 1.9 (PH$^{cc}$’s sub-linear depth lower bound). At $H_n(G) = O(n)$, any $2^{n/2}$ by $2^{n/2}$ $\{−1, 0, 1\}$-matrix of density $\exp(−n^{\Omega(1)})$ forces any PH$^{cc}$ protocol to have depth $n^{\Omega(1)}$ unless it is learnable in $\exp(n^c)$ time.

Theorem 1.10 (PH $\neq$ PSPACE in the communication). PH$^{cc}$ $\neq$ PSPACE$^{cc}$ or $NP \nsubseteq DEP[k \log n]^{21}$.

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 $\neq$ NP). quasi-NP $\nsubseteq TC^0$.

Theorem 1.12 (P $\neq$ NP in algebra). VP $\neq$ VNP or $\forall k \geq 1$, quasi-NP $\nsubseteq NC^k$.

Theorem 1.13 (derandomizing PIT). Either PIT is solvable in deterministic $n^{\text{poly}(\log \log n)}$ time, or $\forall \epsilon > 0, \forall k \geq 1, \text{NTIME}[2^{n^\epsilon}] \nsubseteq \text{SIZE}[n^k]$.

---

18We 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 $1/k$.

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

20A protocol calculates $M(i,j)$ by communication between the two parties knowing only $i$ or $j$.

21DEP$[d]$ 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\(^{22}\) \( P_\theta(G(z_i)) = P_{G(\theta)}(z_i) \) at every \( \theta \)-th dimension, as the previous works used to have. Review1’s Gaussian shift causes \( \hat{G}(\mu, \sigma_i^2) := (\mu + \mu(G(z_i)), \sigma_i^2 + \sigma^2(G(z_i))) \). Review3’s mean shift \( \hat{G}(\mu) = \mu + \hat{G}_i \) 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 \( x_i = (\text{sgn}(z_i - \mu_i) + 1)/2 \in \{0, 1\} \). Review6’s polarity\(^{23}\) flipper induces \( \hat{G}(\theta)(x_i) = \theta(|x_i/2|) \oplus G(|x_i/2|), x_i \in \{2n\} := \{0, 1, \ldots, n - 1\} \).

Our smoothed analysis will focus on Review8’s planted functions. We will employ the most general shift satisfying both robustness \( x_i \mapsto \theta(\hat{G}(z_i))_{\theta} = \{x_i \mapsto \theta \circ x_i\}_{\theta} \) and symmetry \( \theta \circ (G(x_i)) = \hat{G}(\theta) \circ x_i \). These two notions are equivalent, inducing a unique decomposition \( G = (\Phi, \Psi) \) to an attribute permuter \( \Phi \in S_2^n \) and a polarity flipper \( \Psi \in \{0, 1\}^d \) such that \( G(x_i) := 2\Phi(|x_i/2|) + \Psi(|x_i/2|) \) and \( \hat{G}(\theta)(x_i) := \theta(\Phi(|x_i/2|)) \oplus \Psi(|x_i/2|) \). 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 \( (x(j))_{j=1}^n \in \mathcal{X}^m \) 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 \( \text{err}_h(\mathcal{D}) := P(h(x) \neq y) \approx 0 \) whenever \( \text{err}_f(\mathcal{D}) = 0 \). In addition, it obliges the learner to prove \( \text{err}_f(\mathcal{D}) = 0 \approx \text{err}_h(\mathcal{D}) \approx 0 \) in the following axiomatic systems. We study resolution (Res) [DP60, DLL62, Rob65, BSW01, MMZ+01, AM20], polynomial calculus (PC) [CEI96, BIK+96, IPS99, ABRW02, 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 \( \text{err}_f(\mathcal{D}) \approx 0 \) [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 \( \text{RP} \neq \text{NP} \), it is hard to properly PAC learn \( s \)-term \( k \)-DNF for various specific (and unspecific) \( s \) and \( k \) [Val84, Val85, PV88, AFB+08, KSO8, 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, KSO4]. 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 \( n(1-c)k/2 \) examples unless he can refute the \( k \)-SAT with that many constraints. This assumption is the so-called Feige’s hypothesis [Fei02, BKSS13], 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 \( \Omega(n^{1/2}k) \) data; otherwise, all of its SoS degree, PC degree, and Res size must be \( \Omega(n^c) \), \( \Omega(n^c) \), and \( \Omega(\exp(n^c)) \). Similarly, the noisy planted \( k \)-DNF

---

\( ^{22} \)We may write \( g_i(z_i) \) as \( g_i(z) \) or \( g(z) \) 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 \( x_i \mod 2 \in \{0, 1\} \) and \( |x_i/2| \in \mathbb{N} \) as the polarity and attribute of a variate \( x_i \in \{2n\} \).
Theorem 1.15 (hardness of learning DNF). PAC learning the planted $s$-term DNF under the uniform distribution must spend $\Omega(n^{1/4}\log s)$ data; otherwise, all of its SoS degree, PC degree, and Res size must be $\Omega(n^s)$, $\Omega(n^u)$, and $\Omega(\exp(n^c))$, respectively. Similarly, the noisy planted $s$-term DNF needs the same sample size, unless SoS-degree $\Omega(n^s)$ and LP-size $\Omega(\exp(n^c))$.

Theorem 1.16 (Theorem 1.2, hardness). At $H_s(G) = (1 - c)\log s$, $0 < c < 1$, PAC learning the planted $s$-term DNF hiding $\theta \in \{0, 1\}^d$ under the uniform distribution needs $\Omega(n^{1/2}\log s)$ 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 RksAT 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 $\theta \in \{0, 1\}^d$ is distribution-free PAC learnable from $\tilde{O}(n^{k/2})$ data in $n^{O(k)}$ time.

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

In summary, in the worst-case PAC learning, the known spectral threshold $\log m / \log n \approx k/2$ of the RksAT 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 $H_s(G)$ becomes comparable to the logarithm of the problem size (i.e., $\log s$):

PAC1: $H_s(G) = 0$ takes $n^{O(d^{1/3}\log s)}$ learning time by the current best algorithm [KS04, RS10a].
PAC2: $H_s(G) = 0$ requires $2^{\Omega(n^c)}$ time to learn $O(n^{1/2}\log s)$ data under the uniform distribution.
PAC3: $H_s(G) = c\log s$ with $c < 1$ still demands sub-exponential time for $n^{\Omega(\log s)}$ data.
PAC4: $H_s(G) = \log s + O(\log \log n)$ enables us to learn any $n^{1/2}\log s + O(1)$ 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 $\eta = \min_{f} \text{err}_f(\mathcal{D})$ by accuracy $\epsilon$ to achieve $\text{err}_f(\mathcal{D}) \leq \eta + \epsilon$ for any malicious noise rate $\eta \leq 1/2 - 2\epsilon$ [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 $2 \leq d \leq \log \frac{1}{\epsilon} - O(1)$, agnostically learning the planted $\text{AND}_d$ under the uniform distribution must consume $\Omega(n^{(1-c)d/2})$ data. Otherwise, its SoS degree must be $2^{\Omega(n^c)}$.

$^{24}$XOR := $\text{XOR}_d = \{\bigoplus_{i \in w} x_i \mid w \subset \{d\}\}$. $k$JUNTA := $\{f_k(x_i, i \in w) \mid w \subset \{d\}, |w| \leq k, f_k : \{0, 1\}^k \rightarrow \{0, 1\}\}$. 

9
Theorem 1.20 (hardness of agnostic learning planted XOR). For $2 \leq d$, agnostic learning the planted XOR$_d$ under the uniform distribution demands $\Omega(n^{2d/2})$ data or $2^\Omega(n^d)$ SoS degree.

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

Theorem 1.22 (Theorem 1.3, hardness). At $H_\infty(G) = c \log \frac{1}{\varepsilon}$, $c > 0$, agnostically learning the planted AND$_d$ hiding $\theta \in \{0, 1\}^{dn}$ under the uniform distribution must consume $\Omega\left(n^{\log(1/\varepsilon)}\right)$ data. Otherwise, its SoS degree must be $\Omega(\frac{1}{\varepsilon^2})$.

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

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

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

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

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

AGN4: $H_\infty(G) = \log \frac{1}{\varepsilon} + O(\log \log n)$ 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 $\Omega(n^{\frac{k+1}{2}})$ of the following problem: For a predicate $f(x_1, \ldots, x_k)$, prove $|\text{acc}(P^m(x, y)) - \text{acc}((P')^m(x, y))| < \frac{1}{4} (\beta_{\text{cmp}} - \beta_{\text{and}}) \rightarrow P(x, y) \equiv P'(x, y)$ under a promise that either $\text{max}_{\theta} P(y = f_\theta(x)) \geq \beta_{\text{cmp}} > \beta_{\text{and}} \geq \text{max}_{\theta} P'(y = f_\theta(x))$ or $P(x, y) \equiv P'(x, y)$ must hold. We call it $\beta_{\text{cmp}}, \beta_{\text{and}}$-gap (or $\beta_{\text{cmp}} - \beta_{\text{and}}$-gap) approximation of the promise-Max$k$SAT, promise-Max$k$XOR, and promise-Max$k$CSP when $f = \bigwedge_{i=1}^k x_i$, $f = \bigwedge_{i=1}^k x_i$, and $f : \{0, 1\}^k \rightarrow \{0, 1\}$, 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 $\Delta := \beta_{\text{cmp}} - \beta_{\text{and}}$.

Theorem 1.24 (Theorem 1.1, hardness). Any gap ($> 4^{-k}$) approximation of promise-Max$k$SAT under the marginally uniform distribution requires $\Omega(n^{\frac{k+1}{2}})$ 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 $\Omega(n^{\frac{k+1}{2}})$ constraints unless its SoS degree is $\Omega(\exp(n^\epsilon))$.

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

Theorem 1.27 (approximation hardness of the promise-MaxSAT in smoothed analysis). At $H_\infty(G) = c \log \frac{1}{\varepsilon}$ and $1 - (2\varepsilon)^{c+1} \leq \beta_{\text{and}} < \beta_{\text{cmp}} - 4^{-k}$, any $\beta_{\text{cmp}}, \beta_{\text{and}}$-gap approximation of the promise-MaxSAT under the marginally uniform distribution perturbed by any flipper $G$ requires $\Omega(n^{\frac{\log(1/\varepsilon)}{20+4\log(1/\varepsilon)}})$ constraints unless its SoS degree is $\Omega(n^{0.06})$.

\textsuperscript{25} A joint-distribution $P(x, y)$ is marginally uniform if it does not depend on $x$ but may depend on $y$. 

10
Inverting monotone DNF, degree-\(k\) Fourier transforms, and LWE: When the min-entropy reaches \(H_{\infty}(G) = s^O(1)\), 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 \(x_i\) intact, a monotone function \(f(x_1, \ldots, x_d)\) collapses to always \(x_i\) or \(-x_i\) unless it collapses to the constants 0 or 1. Accordingly, the correlation \(\mathbb{E}[(−1)^{X_i+f(\theta_0 X)}]\) under \([X_i/2] = a\) could detect either \((-1)^{X_i+\theta_0 X_i} = (-1)^{\theta(a)}\) or \((-1)^{X_i+\theta_0 X_i} = (-1)^{\theta(a)}\) exclusively so that the statistical correlation analysis over the filtered dataset \(\{\{x, y\} \in \mathcal{D} \mid [x_i/2] = a\}\) could invert the hidden parameter \(\theta(a)\). Notice that the correlation might diminish to the statistical zero if the \(x_i\) were a non-monotone variable of \(f\). This correlation statistics gives rise to Theorem 1.4.

Similarly, suppose the target is REVIEW\(10^{\prime}\)s FT: \(f(x) := \sum_{|w| \leq k} \hat{f}_w \prod_{i \in w} \theta([x_i/2])^{(−1)^{x_i}}\). Observe the outcomes over the restricted data \(\forall i \in w, [X_i/2] = a_i\) on a “query” \((w, a) \in \binom{d}{k} \times \{0, 1\}^w\). It collapses the target function to various subfunctions \(f(x_w) : \{0, 1\}^w \to \mathbb{F}\), inducing the same Fourier coefficient \(\sum_{x_w \in \{0, 1\}^w} f(x_w) \prod_{i \in w} \theta([x_i/2]) \approx 2^w \hat{f}_w \prod_{i \in w} \theta(a_i)\) 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 \(\{\{x, y\} \in \mathcal{D} \mid \forall i \in w, [x_i/2] = a_i\}\) may retrieve the hidden \(\prod_{i \in w} \theta(a_i)\). The correlation might vanish if \(w\) were not maximal, i.e., \(\exists w \supseteq w, \hat{f}_w \neq 0\). It can invert even LWE and GapSVP due to Yao, Toda, Beigel, and Tarui’s modulus amplification [Yao90, Tod91, BT94]. LPN and LWE of REVIEW\(9\) 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 \(1 \leq c \ll k\) 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 \(E(1), \ldots, E(m) \in \mathbb{Z}_p\) is invertible with high confidence to retrieve the secret \(\theta \in \{-c, \ldots, c\}^n\) in \(\text{poly}(n)\) time under the following shift \(G \in \{0, 1\}^{nm(p−1)/2}\). It flips the matrix \(x\) by \(G(x_i(j)) \equiv [x_i(j)/2] \cdot (-1)^{x_i(j)}+G([x_i(j)/2])\) such that the random column \((G(x_i(J)))_{i=1}^n\) is \(k\)-wisely sparse and uniform at \(\forall w \in \binom{\binom{n}\{k\}}{k}\) and \(\forall b \in \mathbb{Z}_p\) as \(\text{Pr}[\forall i \in w, [\frac{x_i(j)}{2}] = 1] \geq \Omega((\frac{2}{p})^k)\) and \(\text{Pr}[\sum_{i \in w} G(x_i(J)) + E(j) = b \mid \forall i \in w, [\frac{x_i(j)}{2}] = 1] \approx \frac{1}{p}\).

We should note that Theorem 1.7’s GapSVP’s decryption [Reg04, Pei09, BLP+13] demands \(m = \text{poly}(n)\) amount of data to Theorem 1.28, while the cryptographic LWE allows no larger than \(m \leq O(n \log p)\) 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 REVIEW\(10^{\prime}\)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 emiting 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], \(\Omega(n^{2.5})\) 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 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 derivate [BS83] translates a lower bound of a matrix $M$ to a lower bound of the bilinear form $\sum_{i,j} x_i M(i,j) x_j$.

Theorems 1.9 establishes a natural sub-linear depth lower bound to learn $\text{PH}^{cc}$, 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., $\text{BPP}^{cc} \not\subseteq (\text{P}^{\text{NP}})^{cc}$ [PSS14], $(\text{P}^{\text{MA}})^{cc} \not\subseteq \text{UPP}^{cc}$ [RS10b, CM17], $\text{MA}^{cc} \not\subseteq (\text{ZPP}^{\text{NP}[1]})^{cc}$ [GPW18], $\text{AM}^{cc} \cap \text{coAM}^{cc} \not\subseteq \text{UPP}^{cc}$ [Kla11, BCH+19]. However, no explicit lower bounds are known for $\text{PH}^{cc}$ and even a much smaller $\text{AM}^{cc} \cap \text{coAM}^{cc}$ [GPW18]. Razborov [Raz98] presented super-$\text{PH}^{cc}$ 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 $\log n \ll d \ll n^{2/h}$. At $\log(G_s(G)) = O(n)$, any $\{-1,0,1\}$-matrix of density $\Omega(2^{d^2 h^2 + 4})$ demands depth $d$ for $\text{PH}_h^{cc}$ unless it is learnable in $O(2^{d^2 h^2 + 4})$ time.

Theorem 1.5’s learning algorithm derives even Theorem 1.10, separating either $\text{PSPACE}$ from $\text{PH}$ in communication complexity or quasi-$\text{NP}$ from parallel-$\text{P}$ 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$^{28}$ $\text{NC}$ 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 [Zák83] 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 [NW97, CR20] derived from Sudan, Trevisan, and Vadhalan’s PRG [STV01]. Theorem 1.5 can solve this circuit unsatisfiability problem as follows. Let $\text{CMD}$ (Connected Matrix Determinant) be an explicit language in $\text{PSPACE}^{cc}$, computing the modulo-2 determinant of the connected matrix $M$, i.e., $M(i,j) \in \{0,1\}$ and $i-j \geq 2 \Rightarrow M(i,j) = 0$.

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

**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], $\text{AC}^0$ [Ajt83, FSS84, Yao85, Hås86], and $\text{AC}^0[p]$ [Raz87, Smo87]. After 30 years of silence, Murray and Williams broke this $\text{AC}^0[p]$ lower bound barrier, establishing quasi-$\text{NP} \not\subseteq \text{ACC}$ [Wil13, Wil14a, MW19].

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

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

$^{27}$ $\text{BPP}$, $\text{ZPP}$, and $\text{UPP}$ are probabilistic polynomial-time computations with bounded, zero, and unbounded errors.

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

$^{29}$ $\text{AC}^0[\text{SYM}]$ consists of the constant-depth circuits arming all symmetric gates of unbounded fan-in.
and even deep neural networks [Dan17, Sha18, VS20, MYSSS21, VRPS21]. Previously, the constant-depth \text{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 $\text{AC}^0[m]$ by $\text{SYM}^+ = \text{SYM} \circ \text{AND}_d$ of quasi-polynomially large degree $d$ [Yao90, BT94]. Our new learning algorithm will do it even for the depth-$h$ $\text{TC}^0$.

**Lemma 1.31** ($\text{TC}^0 \subseteq \text{SYM}^+$). $\text{TC}^0_n \subseteq \text{SYM}^+ \left[ \text{deg}: (c \log n)^{2^h}, \text{norm}: \exp((c \log n)^{2^h}) \right]$.

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 $s \left( \ll \sqrt{n} \right)$ 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, inventing 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 $\text{quasi-NP} \not\subseteq \text{TC}^0$ and $\text{VP} \neq \text{VNP}$.

## 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 $\lceil a \rceil = \max \{ i \in \mathbb{Z} \mid i \leq a \}$ and floor $\lfloor a \rfloor = \min \{ i \in \mathbb{Z} \mid i \geq a \}$ of $a \in \mathbb{R}$. Let $\mathbb{Z}_q := \{(1 - q)/2, \ldots, 0, \ldots, (q - 1)/2\}$ be the integer ring modulo $q$ represented by the $q$ integers nearest to zero. Let $a \mod q := b \in \mathbb{Z}_q$ with $a - b \in q\mathbb{Z}$. For $a, b \in \mathbb{Z}$, define $a = b \mod m \iff a - b \in m\mathbb{Z}$, and $a \oplus b = (a + b \mod 2) \in \{0, 1\}$.

**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 $m < n$, $[m, n] := \{m, m + 1, \ldots, n - 1\}$, $[m, n) := \{m, m + 1, \ldots, n - 1\}$, and $[m, n) := \{m, m + 1, \ldots \}$. We sometimes abbreviate $\{a\}$ as $a$. For sets $S$ and $T$, write their disjoint union by $S \sqcup T$, a difference $S \setminus T = \{a \in S \mid a \notin T\}$, the complement $S^c = U \setminus S$ for the (predetermined) universal set $U \supset S$, a power $2^S = \{T : T \subset S\} \cong \{0, 1\}^S \cong \{\varphi : S \rightarrow \{0, 1\}\}$, a functional $T^S \cong \{\varphi : S \rightarrow T\}$, a combination $\binom{S}{k} = \{T \subset S : |T| = k\}$, and cartesian products $S \times T = \{(a, b) : a \in S, b \in T\}$, $S^n = \prod_{i=1}^n S = \{(a_1, \ldots, a_n) : a_i \in S\}$ ($S^0 = \{\text{null}\}$), and $S^* = \bigcup_{n=0}^\infty S^n$. We call $v \in S^n$ an $S$-vector (or sequence) of length $n$. Specific vectors are $a^n := (a, \ldots, a)$ and $1_i := (0, \ldots, 0, 1, 0, \ldots, 0)$ of 1 at the $i^{th}$ component.
We write \( v \subset w \) when \( v \) 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 \( \binom{\mathbb{S}}{k} = \{ v \in \mathbb{S} \mid |v| = k \} \) of some order-\( n \) set \( \mathbb{S} \). The binomial sum up to \( k < n/2 \) is close to the largest term \( \binom{n}{k} \leq \sum_{n=0}^{k} \binom{n}{k} \frac{n-k}{n-2k} \).

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

**Logics:** Propositional calculus of Boolean predicates writes the truth values 0 := FALSE, 1 := TRUE. The implication \( \phi \rightarrow \psi := \neg \phi \lor \psi \), and equivalence \( \phi \equiv \psi := (\phi \rightarrow \psi) \land (\psi \rightarrow \phi) \). 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:** \( \mathbb{S}_n = \mathbb{S}(\mathbb{S}) \) is the permutation group over a set \( \mathbb{S} \) of cardinality \( n \), say \( \mathbb{S} = [n] \). \( \mathbb{F}_q \) is the finite Galois field of order \( q \), identical with \( \mathbb{F}_q \cong \mathbb{Z} \) as rings. \( \mathbb{F}_q^* = \mathbb{F}_q \setminus \{0\} \) and \( \mathbb{Z}_q = \{ a \in \mathbb{Z}_q \mid a \) is coprime with \( q \} \) are the groups of invertible elements. The \( n \)-variate polynomial ring over \( \mathbb{F} \) is \( \mathbb{F}[x_1, \ldots, x_n] = \{ \sum_{a \in [n]} a_w \Pi_{i \in w} x_w^i \mid a_w \in \mathbb{F} \} \) having \( \mathbb{F} \)-linear summation and multiplication. The multilinear one is a quotient ring \( \{ f \in \mathbb{F}[x_0, \ldots, x_{n-1}] \mid \forall i, x_i^2 = x_i \} \cong \{ f = \sum_{w \subset [n]} a_w \Pi_{i \in w} x_i \mid a_w \in \mathbb{F} \} \). A polynomial \( f \)’s degree is \( \text{deg}(f) = \max (\sum w_i \mid a_w \neq 0) \), and the norm is \( \text{norm}(f) = \sum w_i |a_w| \). It is homogeneous of degree-\( k \) if \( \sum_{i=1}^n w_i = k \rightarrow a_w = 0 \).

**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, \( \mathbb{Z}_{\prod_{i=1}^q q_i} \cong \prod_{i=1}^n \mathbb{Z}_{q_i} \) via \( a \leftrightarrow (a \text{ mod } q_i)_{i=1}^n \).

**Matrices:** An square matrix \( \mathcal{M} \) is degenerate (non-singular, invertible) if it prohibits a non-trivial linear relation, i.e., \( a \neq 0 \Rightarrow \sum_{i,j} a_{ij} \mathcal{M}_{ij} 
eq 0 \). The \( \mathcal{M} \)’s rank measures the maximum size of a non-degenerate submatrix \( \text{rank}(\mathcal{M}) = \max \{ |I| = |J| \mid (\mathcal{M}(i,j))_{i \in I,j \in J} \) is non-degenerate). We write the \( (i,j) \)-entry \( M_{ij} = M_{i,j} = M(i,j) = M(i) \), the \( i \)th row \( M_i = (M(i))_{i \in I} \), the \( j \)th column \( M^j = M(j) = (M(j))_{j \in J} \), and \( \mathcal{M}^j = \mathcal{M}(j) = (M(j))_{j \in J} \). We measure \( \mathcal{M}_{i,j} = \{(i,j) \mid M_{ij} \neq 0 \}, |M| \neq 0 = |M_{i,j} |, \) and call \( |M|_{\neq 0} \) the density of an \( n \times m \) 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 \( \text{Pr}[X] = \text{Pr}_X[X = x] \). Write \( X \sim P(x) \) for \( \forall x, \text{Pr}[X = x] = P(x) \) and \( \text{Pr}[X|X'] \) for \( \text{Pr}_X[X = x | X' = x'] = \text{Pr}_X[x = x] = \text{Pr}[X = x, X' = x'] / \text{Pr}[X'] \). Also, \( X \sim \mathcal{X} \) is the uniform random variable \( X \sim \text{Pr}_X[X = 1] / |\mathcal{X}| \). Random variables \( X_i \) are independent if \( \text{Pr}[X_{i,j}] = \text{Pr}[X_i] \), and mutually independent if \( \text{Pr}[X_i, X_j] = \text{Pr}[X_i] \text{Pr}[X_j] \) for all \( i \neq j \), written as \( X_i \perp X_j \). Their mixture is \( X = \sum_i \rho_i X_i \) by a proportion \( \sum_i \rho_i = 1 \) obeying to \( \text{Pr}[X] = \sum_i \rho_i \text{Pr}[X_i] \). A random variable \( X \) is explicit if \( X \)’s sampler runs in \( \log |\mathcal{X}| \) time (or \( O(\log |\mathcal{X}|) \) space), i.e., there is a \( \log |\mathcal{X}| \) time (or \( O(\log |\mathcal{X}|) \) space) computable deterministic
function \( X : \mathcal{Z} \to \mathcal{X} \) to have \( \Pr_X[X] = \Pr_Z[X(Z)] \). An event (a random Boolean predicate) \( E \) occurs (becomes TRUE) with confidence \( 1 - \delta \) if \( \Pr[E] \geq 1 - \delta \), or equivalently, with significance \( \delta \) if \( \Pr[-E] \leq \delta \). We say that \( E \) happens with high confidence (low significance) if \( \mathbb{E}[E] \geq 1 - O(\delta) \).

A union bound guarantees \( \Pr[E \lor E'] \leq \Pr[E] + \Pr[E'] \), 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 \( A \) can distinguish between \( G(U_n) \) and the genuinely random \( U_m \sim \{0,1\}^m \) by accuracy \( \Pr[A(G(U_n)) \neq A(U_m)] \geq \Omega(1) \).

**Measures:** Denote by \( c, c', \tilde{c}, \ldots \) positive constants. Let \( \epsilon \) be a positive constant sufficiently close to zero, while \( (\epsilon, \delta) = (\epsilon_n, \delta_n) \) are positive variables diminishing to zero. For \( a, b \in \mathbb{R} \), write \( a \approx b \Leftrightarrow |a - b| < \epsilon \), and \( a \ll b \Leftrightarrow a/b \leq \epsilon \) taken in context. \(|S|, |X| \) and \(|v|\) are polymorphic notions denoting the number of elements in a set \( S \), the support size \( \{|x : \Pr[X = x] > 0\}\) 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 \( d_{stat}(X, X') = \frac{1}{2} \sum_{x \in \mathcal{X}} |\Pr[X = x] - \Pr[X' = x]| \). It is equal to the minimum coupling distance \( d_{stat}(X, X') = \min_{(X, X') \sim \Pr[X] \times \Pr[X']} \Pr[X \neq X'] \), so \( |\mathbb{E}[g(X)] - \mathbb{E}[g(X')]| \leq d_{stat}(X, X') \max |g(X)| \) 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 \( \text{poly}(a_n) := \{b_n \mid \exists c > 1, \lim_{n \to \infty} b_n/a_n^c \to 1\} \), \( \text{qpoly}(a_n) := \{b_n \mid \exists c > 1, \lim_{n \to \infty} b_n/a_n^{2\log^c(a_n)} \to 1\} \), and \( \text{exp}(a_n) := \{b_n \mid \exists c > 1, \lim_{n \to \infty} b_n/e^{a_n} = 1\} \). Denote \( O(a_n) = O(a_n \text{plog}(a_n)) \). Polynomial growth means \( \text{poly}(n) \), quasi-polynomial \( \text{qpoly}(n) \), exponential \( \text{exp}(n) \), sub-exponential \( \text{exp}(n') \), quasi-linear \( \tilde{O}(n) \), linear \( \Theta(n) \), sub-linear \( \Theta(n') \), poly-logarithmic \( \text{plog}(n) \), logarithmic \( \Theta(\log n) \), and constant \( O(1) \). For any sufficiently large scale \( n \), \( O(1) \ll \Theta(\log n) \ll \text{plog}(n) \ll \Theta(n') \ll \Theta(n) \ll \text{poly}(n) \ll \text{qpoly}(n) \ll \text{exp}(n) \ll \text{exp}(n') \).

**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 \( n \). \( \mathcal{P} \) 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 \( \mathcal{P} = \text{DTIME}[\text{poly}(n)] \), i.e., a computer can solve a given \( n \)-bit instance within \( \text{poly}(n) \) time. \( \mathcal{NP} \) 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 \( \in \mathcal{NP} \) asserts that a polynomial-time algorithm can ascertain whether or not a presented proof (assignment) satisfies a given instance (constraints). The class quasi-\( \mathcal{NP} \) is the same as \( \mathcal{NP} \) but allows \( \text{qpoly}(n) \) complexities for proof length and verification time. The class co\( \mathcal{NP} = \{L \subset \{0,1\}^* : \mathcal{L}^c \in \mathcal{NP}\} \) 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., \( \forall \mathcal{M} \in \mathcal{F}, \exists f \in \mathcal{P}, \forall x, \mathcal{M} \equiv f(x) \in \mathcal{L} \). 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 \( \text{DTIME}[t], \text{DSPACE}[s], \text{RTIME}[t], \text{RSPACE}[s] \) 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. SIZE[s(n)] and 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 \rightarrow \{0,1\}, respectively. 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, respectively. SIZE length from the root to that node. DEP[d] is the maximum path length, and the depth of a node is the maximum path length from the root to that node.

Communication complexity: A communication protocol is a binary \{AND, OR\}-tree to compute a function \( f(x, y) : \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\} \) by labeling to each leaf node \( w \) either \( 1[(x, y) \in \mathcal{I}_w \times \mathcal{J}_w] \) or its negation for some \( \mathcal{I}_w, \mathcal{J}_w \subset \{0,1\}^n \). DEP^{cc}[d] is the class of depth-d protocols. Its subclass PH^{cc}_h[d] \subset DEP^{cc}[d] has those protocols of all root-to-leaf paths switching at most \((h-1)\) times between AND and OR gates, and PH^{cc}_h[d] = \bigcup_{h \geq 1} \text{PH}^{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 \( \mathcal{D} \sim \left( P(G(x))P(y | G(x)) \right)^m \) of size \( m \geq \epsilon^{-2}(\ell + \log \frac{1}{\delta}) \).

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

\[ (\eta + \epsilon \xi) \text{-learning: } \exists f \in \mathcal{F}, \text{err}_f(\mathcal{D}) \leq \eta - P(y \neq h(x)) \leq \eta + \epsilon \xi. \]

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 \( h \in \text{DTIME}[t'] \) (or \( h \in \text{RTIME}[t'] \)). It defines the worst-case learning by \( H_\infty(G) = 0 \), 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 \( \text{poly}(s, k, n) \) [Str73, HY11].

\(^{31}\)The noise must be below \( \eta + \epsilon \xi \leq 1/2 - \Omega(\epsilon) \) to make the \((\eta + \epsilon \xi)\)-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 \( P(y \neq h(x)) \leq c\varepsilon \), 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: \( \{ x \mapsto \theta \circ g(x) \}_g = \{ x \mapsto \theta \circ x \}_g \).

**SHIFT2:** Symmetry: \( \forall x, \theta \circ g(x) = \hat{g}(\theta) \circ x \).

**SHIFT3:** \( \exists \phi \in \mathbb{S}_n, \exists \psi \in \{0,1\}^n, g(x) = 2\phi([x/2]) + \psi([x/2]) \circ x \).

**SHIFT4:** \( \exists \phi \in \mathbb{S}_n, \exists \psi \in \{0,1\}^n, \hat{g}(\theta) = \theta(\phi) \circ \psi \).

**Proof.** We will demonstrate the following implications. **SHIFT1 \( \Rightarrow^1 \) \( [x/2] \mapsto [g(x)/2] \) is awell-defined injective mapping \( \Rightarrow^2 \) **SHIFT3 \( \Rightarrow^3 \) \( \theta \circ g(x) = \theta(\phi([x/2]) \circ \psi([x/2]) \circ x \Rightarrow^4 \) **SHIFT4 \( \Rightarrow^5 \) **SHIFT2 \( \Rightarrow^6 \) **SHIFT1.**

\( \Rightarrow^1: \) If \( \phi \) is not well-defined, the robustness breaks down by \( [x/2] = [y/2] \wedge [g(x)/2] \neq [g(y)/2] \wedge \theta \circ g(x) \neq \theta \circ g(y) \Rightarrow \{ x \mapsto \theta \circ g(x) \}_g \neq \{ x \mapsto \theta \circ x \}_g \). Also, if \( \phi \) is not injective, then \( [x/2] \neq [y/2] \wedge [g(x)/2] = [g(y)/2] \wedge \theta \circ x \neq \theta \circ y \Rightarrow \{ x \mapsto \theta \circ g(x) \}_g \neq \{ x \mapsto \theta \circ x \}_g \).

\( \Rightarrow^2: \) The permutation \( \phi \) over \([n]\) induces \( x \mapsto (g([x/2]), g(x) \mod 2) := (\phi([x/2]), \psi([x/2])) \).

\( \Rightarrow^3: \) Review6 has defined \( \circ \) as \( \theta \circ (2\phi([x/2]) + \psi([x/2]) \circ x) = \theta(\phi([x/2]) \circ \psi([x/2]) \circ x \).

\( \Rightarrow^4: \) 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 \( \forall x, \hat{g}(\theta) \circ x = \theta \circ g(x) = \theta(\phi([x/2]) \circ \psi([x/2]) \circ x \).

\( \Rightarrow^5: \) **SHIFT4** and **SHIFT3** assert \( \hat{g}(\theta) \circ x = \theta(\phi([x/2]) \circ \psi([x/2]) \circ x = \theta \circ g(x) \).

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

### 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 \( 0 < \gamma \leq 1 \),

\[ \text{ } \text{ } (a, b) \text{-slice, } \text{(min, max)-bound}: a \leq \mathbb{E}[X \mid a \leq X \leq b] \leq b. \]  
In particular, \( \min X \leq \mathbb{E}[X] \leq \max X \).
Markov’s inequality: \( \Pr[X \geq 0] = 1 \Rightarrow \Pr[X/\mathbb{E}[X] \geq 1/\gamma] \leq \gamma. \)

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

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,

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

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

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

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

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

\[
\text{LLL: } \forall i, \Pr[-E_i] \leq \gamma_i \prod_{E_i \neq E_i}(1 - \gamma_i) \Rightarrow \Pr[\bigwedge_{i=1}^n E_i] \geq \prod_{i=1}^n (1 - \gamma_i).
\]

### 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 \( (w, x) \in \binom{\mathcal{X}}{k} \times \{0, 1\}^w \). A random bit-sequence \( X \) is:

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

- \( \varepsilon \)-away \( k \)-independent: \( \forall w, \sum_x |\Pr[\forall i \in w, X_i = x_i] - 2^{-k}| < \varepsilon. \)

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

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

- \( k \)-universal: \( \forall w, \forall x, \Pr[\forall i \in w, X_i = x_i] > 0. \)

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

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

- \( k \)-wisely \( (\mu, \alpha) \)-sparse: \( \forall w, \forall x, \Pr[X_w = x_w] > \alpha \mu^k. \)
**k-wisely \((\mu, \alpha)\)-cover:** \(\forall x \in X(k), \Pr[\exists w, X_w \subseteq x] > \alpha \mu^k\).

**\(\varepsilon\)-away k-independent:** \(\forall w, \sum_x |\Pr[\forall i \in w, X_i = x_i] - 1/|X_w|| < \varepsilon\).

**\((h_w, \delta)\)-hashed \(\varepsilon\)-away k-independent:** For a functional hash \(\{h_w : S_w^c \to \mathbb{N}\}_w\) of \(w^c = (n) - w\), \(\forall w, \forall \xi, (\Pr[h_w(X_w) = \xi] > \delta \Rightarrow \sum_x |\Pr[\forall i \in w, X_i = x_i \mid h_w(X_w) = \xi] - 1/|X_w|| < \varepsilon\).

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 \(m = 3 \cdot 2^k(\ln((\binom{n}{k})^2) + O(1))\) to gain a probabilistic existence:

\[
\Pr[\exists w \in \binom{n}{k}, \exists x \in \{0, 1\}^k, \exists (0 < \Pr[\forall i \in w, X_i(J) = x_i] < 2/2^k)] < (\binom{n}{k})^2((1 - 2^{-k})^m + e^{-\frac{1}{3} \frac{m}{2^k}}) \ll 1. \tag*{\square}
\]

**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 \(m = \frac{6}{\varepsilon^2}(\ln((\binom{n}{k})^2 \frac{n-k}{n-2k}) + O(1))\) samples (due to \(\sum_{\ell=1}^k \binom{n}{\ell} \leq (\frac{n}{k})^k \frac{n-k}{n-2k}\)) and CB parameter \(\gamma = \varepsilon\) demonstrates

\[
\Pr[1 \leq \exists \ell < k, \exists w \in \binom{n}{\ell}, |\mathbb{E}[\prod_{i \in w}(-1)^{X_i(J)}]| \geq \varepsilon] < \sum_{\ell=1}^k \binom{n}{\ell}(-e^{-\frac{n}{2\varepsilon} \frac{\ell}{2^k}} + e^{-\frac{n}{2\varepsilon} \frac{\ell}{2^k}}) \ll 1.
\]

When \(k \geq n/2\), take \(m = \frac{6}{\varepsilon^2}(\ln(2^n) + O(1))\) and apply \(\sum_{\ell=1}^k \binom{n}{\ell} \leq 2^n\) instead of \((\frac{n}{k})^k \frac{n-k}{n-2k}\). \(\square\)

**Lemma 2.11** (hashed \(k\)-independence, probabilistic). There is a \((h_w, \delta)\)-hashed \(\varepsilon\)-away \(k\)-independent random \(n\)-bit sequence of cardinality \(\frac{6\cdot2^k}{\varepsilon^2 \ln((\binom{n}{k})^2 \max_w |h_w(X_w)|)} + O(k \log n)\).

**Proof.** Lemma 10’s probabilistic method on \(m = \frac{6\cdot2^k}{\varepsilon^2}(\ln((\binom{n}{k})^2 \frac{n-k}{n-2k} \max_w |h_w(X_w)|) + O(1))\) 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:

\[
\Pr\left[\left(\exists w \in \binom{n}{k}, \exists x \in h_w(X_w), \Pr[h_w(X_w) = \xi] \geq \delta, (\emptyset \neq \exists v \subset w)\right), \left|\mathbb{E}[\prod_{i \in v}(-1)^{X_i} \mid h_w(X_w) = \xi]\right| \geq \varepsilon'\right] < \max_w |h_w(X_w)| \cdot (\binom{n}{k})^{\sum_{\ell=1}^k \binom{n}{\ell}}(e^{-\frac{\varepsilon^2}{4\varepsilon^2 \ln(2^n)}} + e^{-\frac{\varepsilon^2}{4\varepsilon^2 \ln(2^n)}}) \ll 1. \tag*{\square}
\]

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

\[
\Pr[|X - \mu(X)| \geq \gamma + \varepsilon] < e^{-|k/2|} \text{ for any } 0 \leq \gamma \leq 1 \text{ and any } k \leq \gamma^2 e^{-1/3} \mu(X),
\]

\[
\Pr[|X - \mu(X)| \geq \gamma + \varepsilon] \leq \gamma k \mathbb{E}[|X - \mu(X)|^k] \text{ of } 0 < \gamma \leq 1 \text{ for the sum } X = \sum_i X_i \text{ of perfectly } k\text{-independent } X_i. \quad \text{The claimed inequalities generalize them to an } \varepsilon\text{-away } k\text{-wise independent } X \text{ on the differential bound } |\mathbb{E}[|X - \mu(X)|^k] - \mathbb{E}[|\hat{X} - \mu(\hat{X})|^k]| \leq \varepsilon. \tag*{\square}\]
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 $\Omega(n^{1/k^2})$ 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 [LP86, NN93] and three different algebraic structures [AGHP92]. One of [AGHP92] is the inner product $X_i = \langle A^i, B \rangle := \sum_{j=0}^{\epsilon} A^j_i B_j \mod 2$ of the uniform random $A, B \sim \mathbb{F}_2^n$. The fundamental theorem of algebra assures that the vector $(X_i)_{i=0}^{n-1}$ is $(n/2^k)$-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 $2^k \frac{(k \log n)^2}{2\varepsilon}$, 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 $\varepsilon_{\text{spl}} = (2kt + 2 + \varepsilon)\varepsilon$. There is an explicit construction of $\varepsilon_{\text{spl}}$-away $k$-independent $t$-splitter $\tilde{\Psi} \in [t]^{nt}$ with $|\tilde{\Psi}| = nt^{k+1} \left(\frac{k \log t}{2\varepsilon} \log (nt \log t)\right)^2$.

**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 $(\sum_{x \in [nt]} \mathbb{E}[1[\Psi(x) = \ell] - 1/t])^2 = \sum_x (\mathbb{P}[\Psi(x) = \ell] - 1/t^2) = (n/t)(1 - 1/t)$. Although it is not precisely $t$-splitting, Chebyshev’s inequality of $\gamma = \frac{\varepsilon}{\varepsilon_{\text{spl}}}$ gives

\[
\mathbb{P}[\forall \ell \in [t], |\Psi^{-1}(\ell)| - n < \sqrt{n(1/n - (1-1/t)) < \sqrt{n/\varepsilon}} \geq 1 - \gamma t = 1 - \varepsilon.
\]

To get an exact splitter, execute $\Psi$ “sequentially” until some color gets exactly $n$ nodes, and stop there. The almost $t$-splitting may leave $t \cdot \sqrt{n/\varepsilon}$ (or less) uncolored ones, so color them appropriately to get an exact $t$-splitter $\tilde{\Psi}$. If this sequential coloring $\tilde{\Psi}$ 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 $\tilde{\Psi}(x)$ at a location $w \in \left(\frac{nt}{k}\right)$ is only:

\[
\text{Probabilistic distance: } \mathbb{P}[\exists x \in w, \Psi(x) \neq \tilde{\Psi}(x)] \leq \mathbb{P}[\neg(\text{almost } t\text{-splitting})] + \mathbb{P}[\tilde{\Psi} \text{ may leave some node in } w \text{ uncolored}]
\]

\[
\leq \varepsilon + \frac{kt\sqrt{n/\varepsilon}}{n} = (1 + kt)\varepsilon.
\]

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 $|\tilde{\Psi}| \leq t^k \left(\frac{k \log t}{2\varepsilon} \log (nt \log t)\right)^2$ having statistical distance $d_{\text{stat}}(\Psi(x), \tilde{\Psi}(x)) \leq \varepsilon/2$, yielding a $t$-splitter $\tilde{\Psi}$ 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 $\gamma = \frac{\varepsilon}{\varepsilon_{\text{spl}}}$ applies to the expected difference $\sum_{\ell \in [t]} \mathbb{E}[|\Psi^{-1}(\ell)| - |\tilde{\Psi}^{-1}(\ell)|] \leq \varepsilon^2 n$ and bounds

\[
\mathbb{P}[\exists \ell \in [t], |\Psi^{-1}(\ell)| - n > \sqrt{n/\varepsilon} + \varepsilon^2 n/\gamma] \leq \mathbb{P}[\exists \ell, |\Psi^{-1}(\ell)| - n > \sqrt{n/\varepsilon} + \mathbb{P}[\exists \ell, |\Psi^{-1}(\ell)| - |\tilde{\Psi}^{-1}(\ell)| > \varepsilon^2 n/\gamma = \varepsilon n] \leq \varepsilon + \varepsilon.
\]

$^{32}$Any random $n$-bit vector $X$ having statistical distance $d_{\text{stat}}(X, X') < 1/2$ from some perfectly $k$-independent $n$-bits $X'$ must have $|X| \geq n^{k/2}/(2k^2)$ [AGM03], although $\forall w \in \left(\frac{nt}{k}\right), d_{\text{stat}}(X_w, X'_w) \leq \varepsilon/2$ by definition.
The probabilistic distance analysis on \( \Pr[\exists x \in w, \hat{\Psi}(x) \neq \tilde{\Psi}(x)] \) derives the claimed deviation

\[
\text{Statistical distance: } 2d_{st}(\Psi_w, \hat{\Psi}_w) \leq 2d_{st}(\Psi_w, \tilde{\Psi}_w) + 2d_{st}(\tilde{\Psi}_w, \hat{\Psi}_w) \leq \varepsilon^2 + \Pr[\exists x \in w, \hat{\Psi}(x) \neq \tilde{\Psi}(x)] \\
\leq \varepsilon^2 + 2\varepsilon + \frac{kt(\sqrt{n/\varepsilon} + cn)}{n} \leq (\varepsilon + 2 + 2kt)\varepsilon. \tag*{\blacksquare}
\]

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 \( \tilde{\Psi} \) and permute the \( \ell \)-color nodes \( \{x \in [N] \mid \tilde{\Psi}(x) = \ell \} \) 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 \( (A^{i+j+k+\ell})_i \) 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 \( (i, \kappa) \in \{i, o\} \times \{x, c\} \).

**k-splitters:** Lemma 2.14 provides four i.i.d. \( \varepsilon_{st\text{-away}} 2k \)-independent \( k \)-splitters \( \Psi_{i,\kappa} \) of \([\sqrt{N}]\).

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

**Linear order:** Over \( \{x \in (x_r, x_z) \in [\sqrt{N}] \times [\sqrt{N}] \mid \tilde{\Psi}(x) := (\Psi_{i,\kappa}(x_r) + \Psi_{i,\kappa}(x_z)) \text{ mod } \ell \rangle \), introduce a linear order \( \#_i(x) := \left( \#_i,z(x_r), \#_i,c(x_z), \Psi_{i,\kappa}(x_z) \right) \in [\frac{\sqrt{N}}{\ell}] \times [\frac{\sqrt{N}}{\ell}] \times [k] \cong \mathbb{Z}_2^n \).

**DFT-shift:** Define \( \Phi(x) = y \Leftrightarrow \Psi_{i}(x) = \Phi(y) = \ell \land \#_\kappa(y) = \Phi_{\ell}(\#_i(x)) \).

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

\[
\Pr[\text{All } \Phi_{\ell} \text{ are non-singular linear maps}] \geq 1 - \frac{k^2n_\varepsilon^{2n+1}}{2^e}.
\]

**\( \varepsilon \)-approximate \( k \)-independence** \( \Pr[\|\tilde{\ell}, \Phi_{\ell}(z_\ell) = \Phi_{\ell}(z'_\ell)\| - 2^{-kn} \leq \varepsilon] \geq 1 - \frac{2k^2n_\varepsilon^{2n+1}}{2^e} \).

**Proof.** Let \( \varphi_{\ell}(z_\ell | w) := \sum_{i \in w} \sum_{j \in [n]} x_i^{(i+j+k+\ell)} \in \mathbb{F}_{2^e} \) on \( \mathbb{E}[-1] \langle \varphi_{\ell}(A | z), B \rangle \mid \varphi_{\ell}(a | z) \neq 0 \rangle = 0 \) promises

\[
\text{DFT-bits are unbiased: } \left| \mathbb{E}[-1] \langle \varphi_{\ell}(A | z), B \rangle \right| \leq \text{deg} \left( \varphi_{\ell}(x | w) \right)/2^e.
\]

**Permutation:** Let \( z \circ z' = \sum_{(i,j) \in z \times z'} 1_{i+j \text{ mod } 2} \) 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 \( Z \subset [n] \):

\[
\mathbb{E}[\chi_{\ell}(z_\ell, Z) \cdot \chi_{\ell}(z'_\ell, Z)] = \mathbb{E}[\prod_{j \in Z} (-1)^{\Phi_{\ell}(z_\ell)} + \Phi_{\ell}(z'_\ell)]_{j} \mid \Phi_{\ell}(z_\ell) = \Phi_{\ell}(z'_\ell)] \\
+ \mathbb{E}[\prod_{j \in Z} (-1)^{\Phi_{\ell}(z_\ell)} + \Phi_{\ell}(z'_\ell)]_{j} \mid \Phi_{\ell}(z_\ell) \neq \Phi_{\ell}(z'_\ell)]\Pr[\Phi_{\ell}(z_\ell) \neq \Phi_{\ell}(z'_\ell)] \\
= \Pr[\Phi_{\ell}(z_\ell) = \Phi_{\ell}(z'_\ell)] + 0,
\]

\[
\left| \mathbb{E}[\chi_{\ell}(z_\ell, Z) \cdot \chi_{\ell}(z'_\ell, Z)] - 2^{-n} \right| = \left| \mathbb{E}[-1] \langle \varphi_{\ell}(A | z \circ Z) + \varphi_{\ell}(A | z'_\circ Z), B \rangle \mid Z \neq \emptyset | \Pr[Z \neq \emptyset] \right| \\
\leq \text{deg} \left( \varphi_{\ell}(x | z \circ Z) + \varphi_{\ell}(x | z'_\circ Z) \right)/2^e < 2kn/2^e.
\]

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 \( \mu(A, B, Z) := \Pr[\Phi_{\ell}(Z) = 0^n] - 2^{-n} \)'s average analysis:

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

\(^33\)By definition, \( \Phi_{\ell}(\emptyset) = 0^n \) and \( \chi_{\ell}(z, \emptyset) = 1.\)
\[ \Rightarrow \Pr_{A,B}[\sum_{\ell} \mathbb{E}_{Z}[\mu(A, B, Z) \mid Z \neq \emptyset] \leq 2k^2 n/(2^e \gamma) = 2^{-2n}] \geq 1 - \gamma \]
\[ \Rightarrow \frac{1}{2^{n-1}} - \frac{1}{2^e} \leq \sum_{\ell} \mathbb{E}_{Z}[|\Pr[\Phi_{\ell}(Z) = 0]| - \frac{1}{2^n}] \mid Z \neq \emptyset] \leq 2^{-2n}. \]

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

\[
\mathbb{E}[\prod_{\ell} \chi_{\ell}(z_{\ell}, Z_{\ell}) \cdot \chi_{\ell}(z'_{\ell}, Z_{\ell})] = \mathbb{E}[\prod_{\ell} \prod_{j \in Z_{\ell}} (-1)^{\phi_{\ell}(z_{\ell}) + \phi_{\ell}(z'_{\ell})} \mid \forall \ell, \Phi_{\ell}(z_{\ell}) = \Phi_{\ell}(z'_{\ell})] \Pr[\forall \ell, \Phi_{\ell}(z_{\ell}) = \Phi_{\ell}(z'_{\ell})] + \mathbb{E}[\prod_{\ell} \prod_{j \in Z_{\ell}} (-1)^{\phi_{\ell}(z_{\ell}) + \phi_{\ell}(z'_{\ell})} \mid \exists \ell, \Phi_{\ell}(z_{\ell}) \neq \Phi_{\ell}(z'_{\ell})] \Pr[\exists \ell, \Phi_{\ell}(z_{\ell}) \neq \Phi_{\ell}(z'_{\ell})] + 0,
\]

\[
\leq \mathbb{E}[(-1)^{\sum_{\ell \in [k]} \varphi_{\ell}(A|z_{\ell} \circ Z_{\ell}) + \varphi_{\ell}(A|z_{\ell} \circ Z_{\ell})}] \mid \exists \ell, Z_{\ell} \neq \emptyset] \Pr[\exists \ell, Z_{\ell} \neq \emptyset] \leq 2kn/2^e.
\]

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

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

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

**Inversion:** \(\Phi^{-1}(y)\) is computable in \(O(n^2)\) time, once having all \(\Psi_{i, \kappa}(x_{\kappa})\) of \((i, \kappa, x_{\kappa}) \in \{1, 0\} \times \{\pi, c\} \times \sqrt{N}\) in \(O(\sqrt{N})\) time, and all linear mappings \(\Phi_{\ell}\) of \(\ell \in [k]\) in \(O(kn^2)\) time.

**Permutation:** \(\Phi\) is a permutation.

**Uniform density:** \(\mathbb{E}|(|(I, J) \cap (M \circ \Phi) \neq 0| - \mu_{\text{dss}}| \ll \mu_{\text{dss}} (1 + 4\varepsilon_{ap1})\).

**k-cover:** \(\Pr[(w, J) \subset (M \circ \Phi) \neq 0, |\Psi_{o, x}(w)| = k] - \mu_{\text{crr}} \ll \mu_{\text{crr}} (1 + 4\varepsilon_{ap1})\).

**Proof. Inversion:** The \(\Psi\)’s coloring induces Definition 2.15’s linear order \(\#_\ell(x)\) over the \(\ell\)-monotone nodes \(x \subset [N]\) \(\Psi_{i, \kappa}(x) = \ell\) \(= [2^n]\). Computing the \(n \times n\) \(F_2\)-matrices \(\Phi_i\) and inverting them for all \(\ell \in [k]\) takes only \(\tilde{O}(kn^2)\) time to execute the \(F_2\)-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 \(\Psi_{i, \kappa}\) are perfectly 2k-independent k-splitters of \([\sqrt{N}]\). Let \(\varepsilon_1 := \frac{\delta}{2^e}, \varepsilon_2 := \frac{\delta}{2k}, \varepsilon_k := \frac{\delta}{2^e},\) and \(\varepsilon_{2k} := \frac{\delta}{2^{e_{ap1}}}\). Lemma 2.16 on \(k^2 n 2^{kn} \ll 2^\delta\) has provided those \(\varepsilon\)-approx \(\ell\)-independent permutations of \((\varepsilon, t) \in \{(\varepsilon_1, 1), (\varepsilon_1, 2), (\varepsilon_2, k), (\varepsilon_{2k}, 2k)\}\) with high confidence. For \(y_{\lambda}, x_{\lambda} \in [N], v_{\lambda} \in \binom{N}{k},\) and \(j_{\lambda} \in [\sqrt{N}]\), let

\[
E(x, y) := 1[\forall \lambda, \Phi(x_{\lambda}) = y_{\lambda}] \mid \forall \lambda, \Psi_{1}(x_{\lambda}) = \Psi_{0}(y_{\lambda}) \mid \text{for } x = (x_{\lambda})_{\lambda \in \Lambda} \text{ and } y = (y_{\lambda})_{\lambda \in \Lambda},
\]
\[
E(v, j) := 1[\forall \lambda, \Phi(v_{\lambda}) = (w_{\lambda}), \forall \lambda, |\Psi_{1}(v_{\lambda})| = |\Psi_{0, x}(w)| = k] \mid \text{for } v = (v_{\lambda})_{\lambda \in \Lambda}, j = (j_{\lambda})_{\lambda \in \Lambda},
\]
\[
\overline{E}(x, y) := E(x, y) - 2^{-|\Lambda|n}, \quad \overline{E}(v, j) := E(v, j) - 2^{-|\Lambda| n}.
\]
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 $\sigma^2_{\text{dns}} \approx 3\delta^2_{\text{dns}}$ and $\sigma^2_{\text{cvr}} \approx 3\delta^2_{\text{cvr}}$. Lemma 2.16’s $k$-independent $\Phi$ calculates the first two moments of the $k$-splitting and $k$-covering claims:

\begin{align*}
\text{Uniform-density’s average:} & \quad \left| \mathbb{E}_J[[I, J] \cap (\mathcal{M} \circ \Phi) \neq 0] \right| - \mu_{\text{dns}} \\
&= \left| \left( \sum_{x \in \mathcal{M} \neq 0} \sum_{y \in \mathcal{I} \times \mathcal{V}} \mathbb{P}(\Phi(x) = y) \right) \mathbb{P}(\Phi(y) = \Phi(y)) \right| - \mu_{\text{dns}} \\
&= \frac{1}{k} \left| \mathbb{E}_J[\mathbb{E}(x, y)] \right| \leq \frac{1}{k} |\mathcal{M}| |\mathcal{I}| \mathbb{E}(x, y) = \delta_{\text{dns}}.
\end{align*}

\begin{align*}
\text{Uniform-density’s variance:} & \quad \left| \mathbb{E}_J[[I, J] \cap (\mathcal{M} \circ \Phi) \neq 0] \right|^2 - \mu_{\text{dns}}^2 \\
&= \frac{1}{k^2 N} \left| \sum_{x \in \mathcal{M} \neq 0} \sum_{y \in \mathcal{V} \times \mathcal{V}} \mathbb{E}[\mathbb{E}(x, y) \mathbb{E}(x, y)] \right| \\
&= \frac{1}{k^2 N} \left| \sum_{(x, y) \neq (x', y')} \left[ \mathbb{E}[\mathbb{E}(x, y) (1 - 2^{-n} - 2^{-n} \mathbb{E}(x, y))] + \mathbb{E}[\mathbb{E}(x, y) (1 - 2^{-n})] \right] \right| \\
&\leq \frac{1}{k^2 N} \left[ |\mathcal{M}| |\mathcal{I}| \sqrt{N} (1 + 2^{-n}) (1 - 2^{-n}) \mathbb{E}(x, y) \right] = \sigma^2_{\text{dns}}.
\end{align*}

Chebyshev’s inequality of $\gamma < \delta^{-1/2}$ applies to these moments and establishes the claimed concentrations. It must replace $\Psi_{\lambda,k}$ with Lemma 2.14’s $\varepsilon_{\text{sp1}}$-away $2k$-independent $k$-splitters $\Psi_{\lambda,k}$ so that $\mu_{\lambda}$ with $\mu_{\lambda}(1 + O(\varepsilon_{\text{sp1}}))$ and $\sigma_{\lambda}$ with $\sigma_{\lambda}(1 + O(\sigma_{\text{sp1}}))$ for $\lambda = \text{dns}, \text{cvr}$ by ratios

\begin{align*}
\frac{\mathbb{P}(\Psi_{\lambda}(x) = \Psi_{\lambda}(y))}{\mathbb{P}(\Psi_{\lambda}(x) = \Psi_{\lambda}(y))} \leq 1 + \frac{1}{\varepsilon_{\text{sp1}}} \left( d_{\text{st}}(\Psi_{\lambda,k}(x, x'), \Psi_{\lambda,k}(y, y')) + d_{\text{st}}(\Psi_{\lambda,k}(y, x), \Psi_{\lambda,k}(y, y')) \right) \leq 1 + 4\varepsilon_{\text{sp1}},
\end{align*}

\begin{align*}
\frac{\mathbb{P}(\Psi_{\lambda}(v) = \Psi_{\lambda}(w))}{\mathbb{P}(\Psi_{\lambda}(v) = \Psi_{\lambda}(w))} \leq 1 + \left( \frac{d_{\text{st}}(\Psi_{\lambda,k}(v, v'), \Psi_{\lambda,k}(w, w'))}{d_{\text{st}}(\Psi_{\lambda,k}(v, v'), \Psi_{\lambda,k}(w, w'))} \right) \leq 1 + 3\varepsilon_{\text{sp1}}.
\end{align*}

\section{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, CEL96, GK01, Fei02, App16, FPV18, BBKK18].
Definition 3.1 (refutation in smoothed analysis). A randomized algorithm \( A \) refutes \( F \) if it distinguishes between the training dataset \( D \sim (P(G(x))P(y|G(x)))^m \) with noise \( \eta \leq 1/2 - \Theta(\varepsilon) \) and the random-label \( \mathcal{U} \sim (P'(x) \cdot \frac{1}{2})^m \) drawn from an arbitrary variate distribution \( P'(x) \): 

\[
\eta\text{-noisy refutation: } \Pr_\mathcal{D,\mathcal{U}} \left[ \exists f \in F, \text{err}_f(D) \leq \eta \Rightarrow \Pr_A[A(\mathcal{U}) = \text{refute}] \approx 1 \land \Pr_A[A(D) = \text{refute}] = 0 \right] \geq 1 - O(\delta).
\]

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 \( \hat{O}(m^c) \) examples for \( c \geq 3 \). 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 \( \hat{O}(m^2) \)-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 \( \eta + c\varepsilon \leq 1/2 - c\varepsilon \) in Definition 2.1 must be Definition 3.1’s \( \eta \)-noisy refuter.

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

\[
\text{UGEB}: \quad P(\text{err}_h(\mathcal{U}) > \eta + c\varepsilon) \geq 1 - |\mathcal{H}| e^{-\gamma^2} \frac{1}{2} m \geq 1 - o(\delta).
\]

With high confidence, Definition 2.1’s \((\eta + c\varepsilon)\)-learner can get a proof of \( \text{err}(\mathcal{U}) > \eta \), but can never that of \( \text{err}(\mathcal{D}) > \eta \), 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 \( D \) and choosing \( h_\nu \in [1, 1]^{1/D} \) for \( \nu_0 \leq 2 \frac{1}{\varepsilon^2(1 - \alpha\varepsilon^2)} \) times. Finally, it outputs their majority vote \( h = (\text{sgn}(\frac{1}{\nu_0} \sum_{\nu=1}^{\nu_0} h_\nu) + 1)/2 \). It weights and performs over \( D \) as follows.

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

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

- **Boosting:** \( \forall \nu, E_{(X_\nu, Y_\nu) \sim P_\nu(x, y)}[|(-1)^{Y_\nu} - h_\nu(X_\nu)|/2] \leq 1/2 - \alpha \Rightarrow Pr_{(X, Y) \sim D}[h(X) \neq Y] \leq \varepsilon. \)

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

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

**Proof.** Yao’s reduction on binary search: Let \( A \) be Definition 3.1’s refutation algorithm. Suppose \( m = 3\log m \). Let \( \alpha \approx \frac{1}{m} \). Let \( (X, Y) \sim D \) and \( (X', Y') \sim D' \) 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 \( |b| \leq \log m \), define \( Z_{b,i} := Z_{b,i}^* = (X_i, Y_i) \) if \( 1 \leq |j| \leq |b| \); \( i = b \mod 2^{|b|-1} \land i_j = 0 \neq 1 = b_j \); \( Z_{b,i} := Z_{b,i}^* = (X_i', U_i) \) if \( 1 \leq |j| \leq |b| \), \( i = b \mod 2^{|b|-1} \land i_j = 1 \neq 0 = b_j \); \((Z_{b,i}, Z_{b,i}^*) := ((X_i, Y_i), (X_i', U_i))\) otherwise. Let \( D_b = \)
\[(Z_{b,1})_{i=0}^{m-1} \text{ and } D'_b = (Z_{b,1})_{i=0}^{m-1}. \text{ Let } A_b = 1[\text{A refutes } D'_b] - 1[\text{A refutes } D_b]. \text{ It parses the given refutation gap } E[A_{b|11}] \approx 1 \text{ into the } m \text{ pieces by } E[A_b] = E[A_b|0] + E[A_b|1], \text{ providing } E[A_{b|0}] \geq \alpha \text{ for some } b_0 \in \{0,1\}^{\log m}. \text{ Let }^3 \tilde{A}(x,y) := 1[\text{A refutes } (D_{b|0} \setminus Z_{b_0,b_0}) \sqcup (x,y)]. \text{ The binary search version of Yao’s reduction gives rise to a weak learner } \tilde{A}(x) := \tilde{A}(x,1) - \tilde{A}(x,0):\]

\[
\alpha \leq E[A_{b|0}] = E[\tilde{A}(X',U) - \tilde{A}(X,Y)] = E[(1/2)(\tilde{A}(X',Y' + 1) + \tilde{A}(X',Y')) - \tilde{A}(X,Y)]
\]

\[
\leq E[(1/2)(\tilde{A}(X',Y' + 1) - \tilde{A}(X,Y)) + \tilde{A}(X,Y)] - E[\tilde{A}(X,Y)]
\]

\[
= (1/2)E[\tilde{A}(X')(-1)^{Y'}] + E[\tilde{A}(X',Y')] - E[\tilde{A}(X,Y)]
\]

\[
\Rightarrow \text{ Advantage: } E[\tilde{A}(X')(-1)^{Y'}] \geq 2\alpha + 2[E[\tilde{A}(X,Y)] - E[\tilde{A}(X',Y')]].
\]

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

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

\[
\text{UGEGB: Pr}[P(y \neq H(x)) \geq 2\epsilon] \leq |H| e^{-\gamma/3 \cdot \tilde{m}'} < o(\delta).
\]

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 \( \nu_0 \kappa_0 \log m \cdot O(\delta,\epsilon) = O(\delta) \). For every new prediction, the learner must access \( P(x) \) and refresh \( \tilde{Z}' \) in searching \( b_0 \) of Yao’s reduction. \( \square \)

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

**Proof.** Theorem 3.3’s smoothness for \( \text{err}_f(D) \leq \epsilon \eta \) guarantees \( \text{err}_f(D_{\nu}) \leq \eta \). Definition 3.1’s \( \eta \)-noisy refutation promises \( E[A_{b|11}] \approx 1 \) in Theorem 3.4’s Yao’s reduction on binary search. It reduces Theorem 3.5 to 3.4. \( \square \)

**Theorem 3.6** (refutation to PAC learning in smoothed analysis). If noise-free \( F \) is refutable with significance \( O(\delta^2 \alpha^2 \delta) \), \( F \) 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 \( O(\delta \alpha \delta) \) 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 \( \nu_0 \kappa_0 \log m \cdot O(\delta^2 \alpha^2 \delta \gamma) = O(\delta^3 \gamma) \) with high confidence, reducing Theorem 3.6 to 3.4. \( \square \)

\[\text{UGEGB: Pr}[P(y \neq H(x)) \geq 2\epsilon] \leq |H| e^{-\gamma/3 \cdot \tilde{m}'} < o(\delta).\]
Theorem 3.7 (refutation to noisy PAC learning in smoothed analysis). If $\eta$-noisy $F$ is refutable with significance $O(\delta^2_{2d/\delta})$, $\varepsilon \eta$-noisy $F$ 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 RkSAT-refutation to planted $k$DNF-learning. This section will extend their worst-case reduction to smoothed analysis under any polarity flipper $G$ of min-entropy $H_\infty(G) = (1-c)k$, $0 < c < 1$. It reduces learning the canonical DNF to proving $\bigwedge_{i,j} \bigwedge_{(x_i,0)} \bigvee_{i=1}^{d/k} V_{i=1}^{k} (\theta \circ x_i + j + 1)$ 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 $|S_j| > 2^k - 2^{(1-c)k}$ guarantees that the local solution space $S_j := \{0,1\}^k \setminus \{f([x_{i+j,k}/2]) \circ x_{i+j,k} \oplus 1\}_{i=1}^k$ must contain a $(t-1)$-uniform subspace (Definition 4.8) for $t = \Omega(k)$. Then, any degree-$n^c$ SoS proof may “think” the shifted $k$CNF satisfiable. Consequently, PAC learning DNF requires SoS degree $\Omega(n^c)$ even under the smoothed analysis of min-entropy $H_\infty(G) = (1-c)k$.

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 $Q_D[x] = \{f(x) \in Q[x_1,\ldots,x_n]/\{\forall i, x_i^2 = x_i\} | \deg(f) \leq D\}$.

Non-negativity proof degree: $\deg_{\text{SoS}}[f(x) \geq 0] = \min \{D \mid \exists f_i \in Q_D/2[x_1,\ldots,x_n], f = \sum_i f_i^2\}$.

Positivity proof degree: $\deg_{\text{SoS}}[f(x) > 0] = \min \{D \mid \exists \varepsilon > 0, \deg_{\text{SoS}}[f \geq \varepsilon] \leq D\}$.

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 RkSAT [GK01, FO05], RkCP [COCF10, RRS17] and $t$-uniform RCSP [AOW15, AGK21] but also the matching lower bounds of RkSAT and RkXOR [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 $\psi = (x_{i+j,k})_{(i,j) \in [k] \times [m]} \in k\text{CNF}_{m} \approx [2n]^k \text{CNF}$ is unsat$_\psi(\theta) := \frac{1}{m} \sum_{j=1}^m \prod_{i=1}^k \theta \circ x_{i+j,k} \oplus 1$ at $x = \theta$ of unsat$_\psi(\theta) \in Q_k[x]$.  

---

$^{35}$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 $\Psi \in k$CNF$_n^m$ with $k \geq 3$ as follows:

SoS hardness of $k$SAT: $\Pr[\deg_{\text{SoS}}[\text{unsat}_{\Psi}(x) > 0] \geq \frac{n}{\Delta^{2/(k-2)\log \Delta}}] \geq 1 - \epsilon^k$ 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 $\{\bigwedge_{j=1}^s \bigwedge_{i=1}^k \theta \circ x_{i+jk} \mid \theta \in \{0,1\}^n\}$ under the uniform distribution requires either sample size $\Omega(n^\frac{k-4}{3})$ or SoS degree $\Omega(n^s)$.

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

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

$$P \mathcal{N} \text{ decomposition: } \text{err}_h(U) = \frac{|P|}{|P| + |\mathcal{N}|} \text{err}_h(P) + \frac{|\mathcal{N}|}{|P| + |\mathcal{N}|} \text{err}_h(\mathcal{N}).$$

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

No FPE: $\Pr[\text{err}_h(P) > 0] = \Pr[\exists(X, 1) \in P, \forall j \in [s], \exists i \in [k], \theta([X_{i+jk}/2]) \oplus X_{i+jk} = 0]$

$< |P|(1 - 1/2^k)^s < (1 + o(1))n^{(1-\epsilon)/2}n^{s/2} \leq o(\delta)$.

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

Sub-linear degree: $\Delta = (s/n)[\mathcal{N}] \approx sn^{(k-2)(1-\epsilon)/2 - \epsilon}$

$\Rightarrow \Delta \geq \frac{n}{\Delta^{2/(k-2)\log \Delta}} > (n^{(1+\frac{2}{k-2})/2}(s^{2/(k-2)}(\log s + k \log n))) \gg n^\epsilon$. \hfill \Box

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

Proof. The $\epsilon$-noisy model asks to negate $\deg_{\text{SoS}}[\text{err}_h(D) > \epsilon] < n^\epsilon$. It rewrites Theorem 4.4’s No FPE proof by Chernoff bound of $\gamma = \frac{\epsilon}{(1-2^\epsilon)^2} - 1$ on $(1 - \frac{1}{2^\epsilon})^s \leq e^{\log \epsilon} \ll \epsilon$ as follows:

Small FPE: $\Pr[\text{err}_h(P) > \epsilon] = \Pr[\epsilon < \frac{1}{|P|} \sum((X, 1) \sim P) 1[\forall j \in [s], \exists i \in [k], \theta([X_{i+jk}/2]) \oplus X_{i+jk} = 0]]$

$< e^{-\frac{\epsilon}{2}(1-1/2^k)^s} < e^{-\frac{(1-1/2^k)^s}{|P|}} < e^{-\frac{1}{2}(\omega(1))(\epsilon^{-1}(\epsilon^{-1}(\epsilon^{-1}\text{log } \epsilon))n) = o(\delta)}$.

Theorem 4.4’s sub-linear degree analysis has shown the claimed SoS degree lower bound. \hfill \Box

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 $\Psi \sim k$CNF$_n^m$ of

$$m \leq n^{10^{-4} \log \epsilon}$$

shifted by any flipper space of size $|\mathcal{G}| \leq 2^{(1-\epsilon)k}$ as follows:

SoS hardness of $k$SAT in smoothed analysis: $\Pr[\deg_{\text{SoS}}[\text{unsat}_{\Psi \in \mathcal{G}}(x) > 0] \geq n^{0.06}] \geq 1 - \epsilon^k$.

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

$^{37}$Set $k = \log \frac{s}{\log s \log n}$ 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) $S_j \subset \{0,1\}^k$ in terms of the co-cardinality $2^k - |S_j|$. Meanwhile, Theorem 4.6 allows even malicious $S_j$. 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 $(I \cup J, E)$ between a variable $i \in I$ and a constraint $j \in J$. It takes solution spaces $S_j \subset \{0,1\}^{|E[j]|}$ and presents a graphical CSP instance $G = (I \cup J, E, S)$ of density $\Delta := |J|/|I|$ to minimize $\text{unsat}_G(\theta) = \frac{1}{|I|} \sum_{j \in J} 1[(\theta(i))_{i \in E[j]} \notin S_j]$.

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

$$\text{unif}(S_j) := \max\left\{0 \leq t \leq k \mid \exists X \in S_j, \forall w \in \binom{I}{t}, \forall x \in \{0,1\}^t, \Pr[v_i \in w, X_i = x_i] = 2^{-t}\right\}.$$  

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

$$\text{d-expanding: } |u| \geq |w| - (1/2 - \zeta)|v| - (1/2)\sum_{j \in v} \text{unif}(S_j).$$

**Lemma 4.10** ($k$CSP is expanding [KMOW17]). For $3 \leq t = \Omega(k)$ and $d < \frac{|I|}{k \Delta^2/(t - 2 - \zeta)}$, any $k$CSP instance $G$ to meet $\forall v \subset J, \sum_{j \in v} \text{unif}(S_j) \geq (t - 1)|v|$ must be $k$CSP expanding: $\Pr[G \in E] \geq 1 - e^k$ for the uniform random edge set $E$.

**Proof.** The uniform random $E \sim I^{[k]}$ assures Definition 4.9’s $d$-expanding with significance

$$\Pr[G \in E] = \sum_{[v]} \Pr[G \in E | [v]] \leq \sum_{[v]} \max_{[v]} \prod_{j \in v} \Pr[E_j \in [v]] |E_j < u|$$

$$< \sum_{[v]} \left(\Pr[E_j \in [v]] |E_j < u| \right)^{|v| - 2(1/2 - \zeta)|v|}$$

$$\leq \sum_{[v]} \left(\frac{k^2 e^{2-k}(D \cdot k \Delta^2/(2 - \zeta) / |I|)^{1/2 - \zeta - 1}}{|v|^2} \right)^{|v|} \leq \sum_{[v]} \left(\frac{k^2 e^{2-k}(D \cdot k \Delta^2/(2 - \zeta) / |I|)^{1/2 - \zeta - 1}}{|v|^2} \right)^{|v|} \leq e^k, \quad (\because D \ll \frac{|I|}{k \Delta^2/(2 - \zeta)}),$$

where $\lesssim$ bounds the geometric sum by its start term $k^2 e^{2-k}(k \Delta^2/(2 - \zeta) / |I|)^{1/2 - \zeta - 1} = o(1)$. \hfill \Box

**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 $G$ of $\max_{\theta} \text{val}_{\theta}(G) < 1$ and $\forall j, |E[j]| \leq \zeta D$:

SoS hardness of expanding CSP: $\text{deg}_{\text{SoS}}[\text{unsat}_G(x) > 0] \geq \zeta D/3$.

**Lemma 4.12.** For any set $S_j \subset \{0,1\}^k$ and any integers $1 \leq t \leq r \leq k$,

$$\left(\binom{k}{t} < 2^{r-t}\right) \wedge \left(2^k - 2^{k-r} < |S_j|\right) \Rightarrow \text{unif}(S_j) \geq t.$$
Proof. Randomly generate a $k \times r$ matrix $M \sim \mathbb{F}_2^{k \times r}$. Then, all of its $t \times r$ sub-matrices happen to have the full rank $t$ with a probability of at least $1 - 2^{-t(1+2t)} > 0$. The probabilistic method provides such a matrix $M$. Divide the $k$-dimensional linear space $\mathbb{F}_2^k$ by this $M$ to make the $2^{k-r}$ (or more if $M$ 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 $|\mathbb{F}_2^k - S_j| < 2^{k-r}$ can pick a coset disjoint from $\mathbb{F}_2^k - S_j$. It gives a desired $t$-uniform random variable supported by $S_j$. \[\Box\]

**Theorem 4.13.** Let $t_{4.13} := \frac{c k}{\log e + 1.725 \log((1 + \log e) / c)} \geq 3$ and $D_{4.13} := \frac{3 n}{\Delta (t_{4.13} - 2 - 2c)} \geq k / c$ for $0 < c < 1$. Any kCSP instance $G$ with $|S_j| \geq 2^k - 2^{(1 - c) k}$ under the uniform random $E$ has

$$\Pr_{E \sim \mathcal{I}, j} [\text{deg}_{\text{SoS}}(\text{unsat}_G(x)) > 0] \geq \frac{\zeta_D}{3} \geq 1 - \epsilon^k.$$ 

Proof. Since $|S_j| \geq 2^k - 2^{(k - (r - 1))} / r = [c k]$, Theorem 4.12 of $t = t_{4.13}$ shows $\text{unif}(S_j) \geq t - 1$:

$$1 + \log e + 1.725 \log \left(\frac{1 + \log e}{c}\right) \leq c \frac{(1 + \log e)}{c} 1.725$$

$$2^{t - 1} \left(\frac{c k}{t}ight)^2 = 2^{t - 1} \left(\frac{c k}{t}ight)^2 < 2^{t - \log e + 1.725 \log \left(\frac{1 + \log e}{c} / c\right)} t = 1.$$ 

Consequently, Theorem 4.10’s RkCSP’s expansion has revealed $\Pr_{E \sim \mathcal{I}, j} [G \text{ is } D_{4.13}-\text{expanding}] \geq 1 - \epsilon^k$ for $D_{4.13} \ll \mathcal{I} / (k \Delta^2 (t_{4.13} - 2 - 2c))$, so that Theorem 4.11 with $|\mathcal{E}| = k \leq \zeta_D$ demonstrates $\text{deg}_{\text{SoS}}(\text{unsat}_G(x)) > 0 \geq \frac{\zeta_D}{3}$ with confidence $1 - \epsilon^k$. \[\Box\]

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

SoS hardness in smoothed analysis: $\Pr[\text{deg}_{\text{SoS}}(\text{unsat}_{\land g(\Psi)}(x)) > 0] \geq \frac{\zeta_D}{3} \geq 1 - \epsilon^k$. 

Proof. Rewrite $\text{unsat}_{\land g(\Psi)}(\theta) = \text{unsat}_G(\theta)$ by a CSP $G$ corresponding to $\Psi = (x_i \land j_k) \in k\text{CSP}^n_m$:

$$\mathcal{I} = [n], \mathcal{J} = [m], E(G) = \{[x_i \land j_k/2] : i \in [k], j \in [m]\},$$

$$S_j = \{0, 1\}^{\mathcal{J}_j} \{g([x_i \land j_k/2] \lor x_i \land j_k) \land 1\} = 1 \Pr[G = g > 0],$$

$$\text{unsat}_G(\theta) = \frac{1}{m} \sum g \Pr[G = g] \sum_{j=1}^m \theta([x_i \land j_k/2] \lor x_i \land j_k).$$

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

$$\epsilon, k, t, m = \left(0.066, D_{4.13}, t_{4.13}, n^{\frac{c k}{\log e + 1.725 \log((1 + \log e) / c)}}\right)$$

$$\Rightarrow Sublinear degree: \frac{\zeta_D}{3} = \frac{\zeta_D}{3} \geq \frac{\zeta_D}{3} \geq 1 - \epsilon^k.$$ 

**Theorem 4.15** (Theorem 1.16 for SoS degree under flipper). For $3 \leq k \leq \log \frac{s}{\log n}$ and $0 < c < 1$, PAC learning the canonical planted DNF $\{\bigwedge_{j=1}^s \bigwedge_{j=1} \theta \circ x_i \land j_k \mid \theta \in [0, 1]^n\}$ under the uniform distribution shifted by any flipper $G$ of $H_\alpha(G) = (1 - c) k$ must take each sample size $\Omega(n^{(1 - c) t_{4.13}/2})$ or SoS proof of degree $\Omega(n^s)$.

---

38This 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 \( H_s(G) = (1 - c)k \). No FPE analysis changes therein to

\[
\Pr[\exists g, \exists (g(X), 1) \in \mathcal{P}, \forall j \in (s), \exists i \in (k), \theta([X_{i+jk}/2]) \oplus X_{i+jk} \oplus g([X_{i+jk}/2]) = 0] = \frac{\zeta/n^3}{k^{(1/2(\log n)^{2 \epsilon})}} \geq \frac{\zeta/n^3}{k(t^{(1-\epsilon)/2})} (1-o(1))n^{(1-\epsilon)/2}e^{-t/2k} \leq o(\delta).
\]

Since \( \Pr[\text{deg}_{\text{SOS}}^{\text{unsat}}(\Psi_\theta(x)) > 0] \geq \Omega(\delta) \Rightarrow \exists g, \text{unsat}_{\mathcal{A}_s}^{\text{unsat}}(\Psi_\theta(x)) > 0 \), Theorem 4.4’s sub-linear degree analysis at \((t, d) = (t_{143}, d_{143})\) derives a contradiction to 4.14:

Sub-linear degree: \( \zeta d/3 = \frac{\zeta/n^3}{k^{(1/2(\log n)^{2 \epsilon})}} \geq \frac{\zeta/n^3}{k(t^{(1-\epsilon)/2})} (1-o(1))n^{(1-\epsilon)/2}e^{-t/2k} = o(\delta) \).

\[ \square \]

**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

\[
\Pr[\varepsilon < \frac{1}{|\mathcal{P}|} \sum_{j=1}^{\mathcal{X} \sim \mathcal{P}, \exists g, \exists i \in (k)} \theta([X_{i+jk}/2]) \oplus X_{i+jk} \oplus g([X_{i+jk}/2]) = 0] = \frac{\zeta/n^3}{k^{(1/2(\log n)^{2 \epsilon})}} \geq \frac{\zeta/n^3}{k(t^{(1-\epsilon)/2})} (1-o(1))n^{(1-\epsilon)/2}e^{-t/2k} = o(\delta).
\]

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

\[ \square \]

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 \( f_\theta(x_1, \ldots, x_d) \) hiding an assignment \( \theta \in \{0,1\}^d \).

**Theorem 4.17** (Theorem 1.16\textsuperscript{39} for SoS degree). For \( 3 \leq k \leq \log \frac{s}{\log(1/\varepsilon)} \) and \( 0 < c < 1 \), PAC learning the canonical planted DNF class \( \bigvee_{j=1}^{s} \bigwedge_{i=1}^{k} \theta_{i+jk} \circ x_{i+jk} \mid \theta \in \{0,1\}^{kn} \) under the uniform distribution perturbed by any shift \( G = (\mathbb{S}_n \times \{0,1\}^n)^d \) of \( H_s(G) = (1 - c)k \) requires either sample size \( \Omega((n/4^{(1-c)k})(1-c)^{t_{143}/2}) \) or SoS degree \( \Omega((n/4^{(1-c)k})^c) \).

Proof. A reduction to Theorem 4.15. Force SA1’s adversary to choose the hidden parameter \( \theta_i \in \{0,1\}^n \), \( i = i + jk \), in the following manner. Let \( O_i(a) = \{ \phi_i(a) \mid \phi_i, \psi_i \in O \} \) be the orbit permuting an attribute \( a \in [n] \), \( O_i^{-1} \circ O_i(a) = \{ \phi_i^{-1}(\phi_i(a)) \mid \phi_i, \psi_i \in O \} \), and \( A_i \subset [n] \) be a maximal attribute set of these orbits with \( O(a) \neq O(a') \Rightarrow O_i(a) \cap O_i(a') = \emptyset \). Since \( \bigcup_{O_i(a) \in (i)} O_i^{-1} \circ O_i(a) \supset [n], |A_i| \geq n/|O|^2 := n' \). Bound the adversary’s choice of the hidden parameter \( \theta \in \{0,1\}^{dn} \) to make \( \theta_i \circ x_i \) invariant modulo these orbits \( O_i(a) \), i.e., \( \forall i, [x_i/2] \in O_i(a) \Rightarrow \theta_i(x_i) = \theta_i(a) \oplus x_i \). Further, the adversary must choose \( \theta \) from \( \#_i(a) = \#_{i,t}(a') \Rightarrow \theta_i(a) = \theta_i(a') \), where \( \#_t \) is a linear order over \( A_t \). Then, learning under \( G \) reduces to learning under the induced flipper over \( \prod_{i=1}^{d} (A_i \times \{0,1\}) \). It replaces \( n \) with \( n' = n/|O|^2, 2m \) with \( 2m' \approx n'(1-c)^{t_{143}/2} = t = t_{143} \), and \( d \) with \( d' = 3n'/\alpha((s'm'/n')^{2/(2-2\epsilon)}) \). Still, Theorem 4.15’s sub-linear degree analysis derives a contradiction to Theorem 4.14:

Sub-linear degree: \( \zeta d/3 = 3n'/\alpha((s'm'/n')^{2/(2-2\epsilon)}) \gg n'^c \).

\[ \square \]

**Theorem 4.18** (Theorem 1.16 for SoS degree under noise). For \( 3 \leq k \leq \log \frac{s}{\log(1/\varepsilon)} \) and \( 0 < c < 1 \), PAC learning the \( \varepsilon \)-noisy canonical planted DNF class \( \bigvee_{j=1}^{s} \bigwedge_{i=1}^{k} \theta_{i+jk} \circ x_{i+jk} \mid \theta \in \{0,1\}^{kn} \) under the uniform distribution perturbed by any shift \( G = (\mathbb{S}_n \times \{0,1\}^n)^d \) of \( H_s(G) = (1 - c)k \) requires either sample size \( \Omega((n/4^{(1-c)k})(1-c)^{t_{143}/2}) \) or SoS degree \( \Omega((n/4^{(1-c)k})^c) \).

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

\textsuperscript{39}Set \((k, n', 2m', d', \varepsilon) = (\log \frac{s}{\log(1/\varepsilon)}, \frac{n}{4^{(1-c)k}}, n'^{(1-c)}, 3n'/\alpha((s'm'/n')^{2/(2-2\epsilon)}), 0.065) \Rightarrow n' > n'^{0.06} \).
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 \( f(\theta) : S \to \mathbb{R} \) are embeddings \( \varphi(f), \varphi(\theta) \in \mathbb{R}^s \) to a higher dimensional metric space\(^ {40} \) \( \mathbb{R}^e \). Let \( \mathcal{P} \subseteq \mathbb{R}^e \) be a polytope \( \mathcal{P} = \{ x \in \mathbb{R}^e \mid Ax \leq b \} \).

\[ \text{LP proof size: } \text{size}_{\text{LP}}[f(x) > 0] = \min \left\{ e \mid \exists \varphi, \exists \mathcal{P}, \forall \theta \in S, f(\theta) = \langle \varphi(f), \varphi(\theta) \rangle \wedge \varphi(S) \subseteq \mathcal{P} \land \min_{x \in \mathcal{P}} \langle \varphi(f), x \rangle > 0 \right\}. \]

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

Expansion: \[ \Pr[\text{size}_{\text{LP}}[\text{unsat}_G(x) > \varepsilon] \geq \exp \left( \left( \frac{|\mathcal{I}(t-2)/2}{\Delta} \right)^{2(1-\varepsilon')/k} \right) \geq 1 - o(1). \]

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

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

Sub-exp size: \( \text{size}_{\text{LP}}[\text{unsat}_\varphi(x) \geq \varepsilon/2] \geq \exp \left( \frac{2n^{k/2}}{(n^{1-\varepsilon')k/2})^{2(1-\varepsilon')/k} \right) = \exp \left( \frac{2n^{k/2}}{n^{1-\varepsilon')k/2}^{2(1-\varepsilon')/k} \right) = \exp(n^s). \)

Theorem 4.22 (Theorem 1.16 for LP size under noise). For \( 3 \leq k \leq \frac{s}{\log(1/\varepsilon)} \) and \( 0 < c < 1 \), PAC learning the \( \varepsilon \)-noisy canonical planted DNF class \( \{ V_{j=1}^s \bigwedge_{i=1}^k \theta \circ x_{i+j} \mid \theta \in \{0, 1\}^{kn} \} \) under the uniform distribution perturbed by any shift \( G \in (\mathbb{S}_n \times \{0, 1\}^d) \mathcal{H}_n(G) = (1-c)k \) requires either sample size \( \Omega((n/4^{(1-c)k}(1-c)^{kr/2}) \) or LP-size \( \Omega(\exp((n/4^{(1-c)k}(1-c)^{kr/2})) \).

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 \( \mathbb{R}^e \).
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 \(\xi_j \in F = \{ \bigvee_{i \in m} x \circ \bar{i}, \ w \in [2n]\}\) over the \(n\) Boolean indeterminates \(\{x(i)\}_{i \in [n]}\) with \(x \circ i := x([i/2]) \oplus i\) for \(i \in [2n]\),

\[
\text{Resolution proof size: } \text{size}_{\text{Res}}(\land_j,_{m+1} \xi_j \neq 1) = \\
\min \left\{ S \mid \exists \{\xi_j\}_{j=m+1}^{s}, \xi_e = 0, \forall j > m, \exists i \in [n], \exists \kappa < j, \exists \kappa' < j, \exists \xi' \in F, \right. \\
\left. \xi_k = \xi \lor x \circ (2i) \text{ and } \xi_k' = \xi \lor x \circ (2i - 1), \text{ or } \xi_j = \xi_k \lor \xi_k' \right\}.
\]

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

\[
\text{PC proof degree: } \deg_{\text{PC}}(\bigwedge_{j=1}^{m+1} \xi_j \neq 1) := \\
\min \left\{ D \mid \exists e, \exists \{\xi_j\}_{j=m+1}^{e}, \xi_e = 1 \land \forall j > m, \exists i \in [n], \exists \kappa < j, \exists \kappa' < j, \exists a \in \mathbb{Q}, \right. \\
\left. \xi_j \in \{\xi_k + a\xi_k', \xi_k \cdot x \circ (2i), \ x \circ (2i) + x \circ (2i - 1)\} \right\}.
\]

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

\[
Pr_{\psi \sim k\text{-CNF}_n^m}[\text{size}_{\text{Res}}[\text{unsat}_\psi(x) > 0] \geq \exp\left(\frac{n}{\Delta^{2/(k-2) \log \Delta}}\right)] \geq 1 - o(1).
\]

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

\[
Pr_{\psi \sim k\text{-CNF}_n^m}[\deg_{\text{PC}}[\text{unsat}_\psi(x) > 0] \geq \Omega\left(\frac{n}{\Delta^{2/(k-2) \log \Delta}}\right)] \geq 1 - o(1).
\]

**Theorem 4.27** (Theorems 1.14 and 1.15 for Res size and PC degree). For \(3 \leq k \leq \log \frac{s}{\log n}\), 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 \(\Omega(n^{(1-o)} 2^{k/2})\) unless Res-size is \(\Omega(\exp(n^s))\) and PC-degree \(\Omega(n^s)\).

**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.\(\square\)

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 \(\deg_{\text{PC}}[\text{unsat}_\psi(x) > 0] \leq D\) is rewritable to an SoS proof of \(\deg_{\text{SoS}}[\text{unsat}_\psi(x) > 0] \leq 2D\) in polynomial time.

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

\[
Pr[\deg_{\text{PC}}[\text{unsat}_\Lw^G(\psi)(x) > 0] \geq \zeta D_{13}/3] \geq 1 - \epsilon^k.
\]

\(^{41}\)The twin variable of \(x_i\) is another formal variable \(\bar{x}_i\), with the complementary axiom \(x_i + \bar{x}_i - 1 = 0\).

Theorem 4.30 (Theorem 1.16 for PC degree). For \(3 \leq k \leq \log \frac{s}{\log s} \log n\) and \(0 < c < 1\), PAC learning the canonical planted DNF \(\{V_j = 1 \wedge_{i=1}^k \theta_{i+jk} \circ x_{i+jk} \mid \theta \in \{0,1\}^{kn}\}\) under the uniform distribution perturbed by any shift of min-entropy \(H_\varepsilon(G) = (1-c)k\) requires either sample size \(\Omega\left(\left(\frac{n}{n^{1-c/k}}\right)^{(1-c)\ell}t/\varepsilon^2\right)\) or PCR degree \(\Omega\left(\left(\frac{n}{n^{1-c/k}}\right)^{c}\right)\).

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 \(d = O(d^{1/3} \log s)\) can express any planted \(s\)-term DNF \(f(x_1, \ldots, x_d)\) by

Threshold polynomial of DNF: \((-1)^f(x_1, \ldots, x_d) = \text{sgn}\left(\sum_{w} a_w(-1)\sum_{i \in w} x_i\right)\), \(a_w \in \mathbb{Q}, w \subset [n], |w| \leq d\).

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

Proof. Solve an LP instance \(\forall j \in (m), (-1)^{y(j)} = \text{sgn}\left(\sum_{w} a_w(-1)\sum_{i \in w} \theta_i \circ x_{i(j)}\right)\) of Theorem 5.2’s threshold polynomial of DNF. Inside the sign is a linear function of at most \(n' := \sum_{k=0}^{d} n^k\binom{d}{k}\) variables \(x_{w,a} = (-1)\sum_{i \in w} \theta_i(a_i) \in \{1,-1\}\) for \((w,a) \in \binom{d}{k} \times n^k, k \leq d\). Hence, Theorem 5.1’s LP algorithm can find a solution by \(O(\log(n' \varepsilon))\)-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 RkSAT 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, JKL20] to solve Grothendieck Inequality (GIE) and find refutation certificates.

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

\[
\text{Grothendieck Inequality (GIE):} \quad \max \|u_i\|, \|v_j\| \leq 1 \sum_{i,j} M_{ij} \|u_i, v_j\| \leq c_g \max \|x_i, y_j\| \leq 1 \sum_{i,j} M_{ij} x_i y_j.
\]

Symmetric GIE: \(\max \|u_i\|, \|v_j\| \leq 1 \sum_{i,j} M_{ij} \|v_i, v_j\| \leq c_g \max \|x_i, y_j\| \leq 1 \sum_{i,j} M_{ij} y_i x_j\) if \(\forall M_{ii} = 0\).

Theorem 5.5 (computational complexity of SDP [JKL20]). 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 \(t_{\text{sdp}}(n) := \tilde{O}(n^{3.5})\) when \(m = O(n)\).

\(^{42}\omega\) is the exponent in the matrix multiplication complexity. The current best is \(\omega = 2.372 \ldots\) [Str69, AW21].
Coja-Oghlan, Cooper, and Freize [COCF10, AOW15] reduced MaxkCSP’s “average-case” approximation to planted kXOR’s “strong” refutation: 

$$\text{Prove acc}(\{(x(j), y(j)): j = 1, \ldots, n\}) \leq 1/2 + \varepsilon$$

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 kDNF’s PAC learnability. Recently, Abascal, Guruswami, 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 kXOR [AGK21]). The following refutation’s proof enjoys a witness computable by SDP in $t_{\text{sdp}}(N^2) \cdot O(n)$ time with confidence $1 - \frac{1}{N}$ from any $m = \sum_{i=1}^{n} |D_i| = n \sqrt{n} \cdot O\left(\frac{\log^3 N}{\varepsilon^6}\right)$ data of $D_i \subset [2N]^2 \times \{0, 1\}$ having the i.i.d. random $m$ labels:

**Strong refutation:**

$$\max_{x \in \{0,1\}^{n}, y \in \{0,1\}^{n}} \frac{1}{m} \sum_{i=1}^{m} \sum_{(x,y) \in D_i} 1[(z \circ x_1) \oplus (z \circ x_2) = z'_1 \oplus y] \leq 1/2 + \varepsilon.$$ 

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

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

**Bias measurement:**

$$\text{bias}_f(D) := \frac{1}{m} \sum_{j=1}^{m} (-1)^{y(j)+1} f(x(j)) = 1$$

$$= \frac{1}{2^m} \sum_{j=1}^{m} (-1)^{y(j)+1} \prod_{i=1}^{k} a_i \sum_{a_i \in [n]} ((-1)^{\theta(a_i)+x_i(j)+1} + 1) \prod_{i \in w} (-1)^{\theta(a_i)},$$

**Fourier coefficients:**

$$\hat{M}_w(a) = \frac{1}{m} \sum_{j: x_w(j) / 2 = a} (-1)^{y(j)+1} \prod_{i \in w} (-1)^{x_i(j)+1}.$$ 

Let $1 \leq \kappa = ||w||/2$ or $||w||/2 \leq |k/2|$. Lift and project the bias maximization problem $\max_{\delta} \text{bias}_f(D)$ over $\theta \in \{0,1\}^{n}$ to the following QPs (Quadratic Programming) over $z = (z_a)_{a \in [n]} \in \{-1,1\}^n$ and $z' = (z'_b)_{b \in [n]}$ of $z'_b = (-1)^{\theta(b)} \in \{-1,1\}$ to bound $\text{bias}_f(D) \leq \text{val}(D)$:

**QP by lift and project:**

$$\text{val}(D) := \sum_{w \subseteq \{k\}, |w| \in 2\mathbb{Z}} \max_{a \in [n]} \hat{M}_w(z) + \sum_{v \subseteq \{k\}, |v| \in 2\mathbb{Z}+1} \max_{z,z'} \hat{M}_v(z, z'),$$

Even QP: $\hat{M}_w(z) := \sum_{a \in [n]} \sum_{a' \in [n]} \hat{M}_w(aa')z_az_{a'}$ for $|w| = 2\kappa$,

Odd QP: $\hat{M}_v(z, z') := \sum_{a \in [n]} \sum_{a' \in [n]} \hat{M}_v(aa')z_az_{a'}z'_{a}$ for $|v| = 2\kappa + 1$.

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

**Completeness:** Take a threshold $\beta \approx 1/(2s)$ as follows. We may assume $|E[(-1)^Y]| \leq \varepsilon\beta$. Otherwise, the constant function is already Definition 3.1’s refuter to distinguish between $|E[(-1)^Y]| \geq \varepsilon\beta$ and $|E[(-1)^Y]| < \varepsilon\beta$ for the random-label data. It promises the complete data err$_f(D) = 0$ to gain an advantage by choosing the heaviest $f$ from the set $s$ terms:

**Completeness:**

$$\text{bias}_f(D) = \Pr[f(X) = Y = 1] \geq \frac{1}{2s} - \frac{1}{2} |E[(-1)^Y]| := \beta.$$

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

**Soundness:**

$$\text{bias}_f(\mathcal{U}) \leq \text{val}(\mathcal{U}) \leq \frac{1}{2\kappa} \left(\sum_w \max_z \hat{M}_w(z) + \sum_v \max_{z,z'} \hat{M}_v(z, z')\right) \ll \beta.$$
Computational complexity: Theorem 5.6 solves both even and odd QPs and provides a certificate of $\text{val}(U) \ll \beta$ for the soundness data $U$. The overall confidence level is $1 - 2^k/N = 1 - o(\delta)$ to succeed in Definition 3.1’s refutation of $D' \in \{D, U\}$ only when getting a certificate of $\text{val}(D') \leq \beta$ from $m$ data in $t_{\text{sdp}}(n^k) \cdot O(n) \cdot 2^k$ time.

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 $O(n^{k/2})$ 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: $\max_{\{a\}} \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} x_i x_j \right]$ with $\forall M_{ij} = 0$ over $x \in \{-1, 1\}^N$ is approximable by ratio $44 \gamma_g := \Omega(1/\log N)$ in $t_{\text{sdp}}(N^2)$ time.

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

Proof. Theorem 5.8 with $N = n^k$ of $\kappa = [|v|/2]$ or $[|w|/2]$ approximates Theorem 5.7’s QP by lift-and-project to get even-QP’s $M_{w}$’s rounded solutions $z(w) = (z_a(w))_{a \in [n]}$. Theorem 5.8’s QP requires removing the trace $\sum_a M_{w}(aa) z_a z_a = \sum_a M_{w}(aa)$. Theorem 5.6 divides odd-QP’s $M_{o}$ into a sum over $b \in [n]$ of $M_{v,b}(z) = (-1)^{g(b)} M_{v\setminus\{j\}}(z)$ on $D_b = \{(x(j), y(j)) \in D \mid |x(j)/2| = b\}$. Theorem 5.8 provides $M_{v,b}$’s rounded solutions $z(v, b) = (z_a(v, b))_{a \in [n]}$, too. These QP’s solutions induce a hypothesis function $h : (2n)^k \rightarrow \Omega$ to bound bias$_f(D) \leq$ bias$_h(D) := \mathbb{E}_{[(-1)^{h(X)}]}$ over the empirical data $(X, Y) \in \{(x(j), y(j))\}_{j=1}^{n}$:

$$h_w(x) := \prod_{i \in [n]} (-1)^{x_i+1} [x_i/2] = a,$$

$$h_v(x) := \sum_{(a\neq a') \in [n] \times [n]} h_w(x|[a]') z_a z_a'(w), \quad g_w(x) := \sum_{a \in [n]} h_w(x|a),$$

$$h_{v,b}(x) := \prod_{i \in [n]} [x_i/2] = b \sum_{a \in [n]} h_{v\setminus\{i\}}(x|a) z_a z_a'(v, b),$$

Weak hypothesis: $h(x) := \frac{1}{2^k} \sum_{w \subset [k], |w| \in [2^k]} h_w(x) + \frac{1}{2^k} \sum_{w \subset [k], |w| \in [2^k]} \sum_{b \in [n]} h_{v,b}(x),$

Trace: $g(x) := \frac{1}{2^k} \sum_{w \subset [k], |w| \in [2^k]} g_w(x) + \frac{1}{2^k} \sum_{w \subset [k], |w| \in [2^k]} \sum_{b \in [n]} g_{v,b}(x).$

$$\text{bias}_f(D) - \text{bias}_h(D) \leq \Omega(D) - \text{bias}_h(D)$$

$$= \sum_{w \subset [k], \max_{[v] \in [2^k]} |M_w(v)|} + \sum_{v \subset [k], \max_{[v] \in [2^k]} |M_{v,b}(z)|} \leq \frac{1}{\gamma_g} \sum_{w \subset [k], |w| \in [2^k]} \text{bias}_h(D) + \frac{1}{\gamma_g} \sum_{v \subset [k], |v| \in [2^k]} \text{bias}_{h,v,b}(D)$$

$$= \frac{1}{\gamma_g} \text{bias}_h(D) \text{ by Theorem 5.8’s ratio } \gamma_g := \Omega(1/\log n^r).$$

Boosting: Theorem 5.7’s completeness proof has shown $\text{bias}_h(D) \geq \gamma_g (\text{bias}_f(D) - \text{bias}_h(D)) \geq \gamma_g (\beta - \text{bias}_h(D))$ for $\beta \approx \frac{1}{2^k}$. Theorem 3.3’s SmoothBoost turns this weak hypothesis $h(x) = h_v(x)$ feeding $D = D_v \sim (P_v \circ D)^*$ to an $\varepsilon$-accurate hypothesis in the following manner. First of all, we may assume $|\text{bias}_h(D_v)| \leq \varepsilon$. Otherwise, SmoothBoost can feed $g(x)$ or $-g(x)$ for a weak predictor. Take $\nu_0 \approx \frac{2}{\varepsilon ((1-\varepsilon) \beta \gamma_g)^2}$ and $m \gg n^{[k/2]} \cdot 2^k k (ks \log n)^2$. It is much larger than the

$\text{bias}_h(D_v) \leq \varepsilon$. Otherwise, SmoothBoost can feed $g(x)$ or $-g(x)$ for a weak predictor. Take $\nu_0 \approx \frac{2}{\varepsilon ((1-\varepsilon) \beta \gamma_g)^2}$ and $m \gg n^{[k/2]} \cdot 2^k k (ks \log n)^2$. It is much larger than the
The logarithm of the hypothesis size $|\{h_v\}_v| \leq \prod_w |\text{rng}(z(w))| \cdot \prod_{u,b} |\text{rng}(z(u,b))| \leq 2^{\sum_{n=1}^{k/2} (2^n) n^k} \cdot 2^n \sum_{n=1}^{k/2} (2^n) n^k$, so the final majority vote enjoys UGEB by Chernoff bound parameter $\gamma = 1$:

$$\text{UGEB} \cdot \prod_{v \in [n]} |\{h_v\}_v| \cdot e^{-\frac{1}{2} \epsilon m} \leq 2^{\sum_{n=1}^{k/2} (2^n) n^k} \cdot 2^{\sum_{n=1}^{k/2} (2^n) n^k} \cdot e^{-\frac{1}{2} \epsilon m} = o(\delta).$$

The overall learning time is $\nu_0(\sum_w t_{\text{adp}}(n|w|)) + \sum_{v,b} t_{\text{adp}}(n|v|)) \leq \nu_0 \cdot t_{\text{adp}}(n^k) \cdot O(2^k n)$. ☐

**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 $n^{[k/2]} \cdot O((\frac{ks \log n}{\epsilon(1-2\eta)})^2)$ data in $t_{\text{adp}}(n^k) \cdot O(\frac{2^k n}{\epsilon} (\frac{ks \log n}{1-2\eta})^2)$ time.

**Proof.** The white $\eta$-noise replaces $\beta \approx \frac{1}{2^k}$ to $\beta \approx \frac{1-2\eta}{2^k}$. 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 $k$: No FPE: $f(x) = 0 \Rightarrow \tilde{f}(x) = 0$.

Recall: $\Pr_G[f(G(x)) = 1, \tilde{f}(G(x)) = 0] \leq s/(2^{k+1} \rho)$.

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

$$\Pr_G[f(G(x)) \neq \tilde{f}(G(x))] = \Pr_{G}[f(G(x)) = 1 \land \tilde{f}(G(x)) = 0]$$

$$\leq \Pr_G[3\kappa \in (s], |f_\kappa| \geq k + 1, f_\kappa(G(x)) = 1]$$

$$= \Pr_G[3\kappa \in (s], |f_\kappa| \geq k + 1, \forall i \in f_{\kappa i}, G([x_i/2]) = \theta([x_i/2]) \oplus x_i \oplus f_{\kappa i} + 1] \leq s/(2^{k+1} \rho).$$

**Theorem 5.12** (Theorem 1.18\(^{45}\)). The planted $s$-term DNF is PAC learnable from any $n^{[k/2]} \cdot O((\frac{ks \log n}{\epsilon(1-2\eta)})^2)$ data in $t_{\text{adp}}(n^k) \cdot O(\frac{2^k n}{\epsilon} (\frac{ks \log n}{1-2\eta})^2)$ 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 bias $f(D)$. Theorem 5.11’s recall guarantees $\text{bias}_f(D) \geq (1 - \epsilon - \epsilon)/2s = (1 - \epsilon - \epsilon)/2s$ 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 $n^{[(k+1)/2]} \cdot O((\frac{ks \log n}{\epsilon(1-2\eta)})^2)$ data in $t_{\text{adp}}(n^k) \cdot O(\frac{2^k n}{\epsilon} (\frac{ks \log n}{1-2\eta})^2)$ 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 $k = \log \frac{s}{\epsilon^2}$ and $\frac{1}{\epsilon} = O(1)$. Take Lemma 2.9’s $\frac{1}{\epsilon}$-dense $dn$-bit flipper of cardinality $O(2^k \log (dn))$. 36
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 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 AND_d = \{f(x) := \bigwedge_{i \in f} x_i \oplus f_i \mid f \subset \{d\}, f_i \subset \{0, 1\}\} [Pat92, NS94, TT99, KKMS08].

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

**Proof.** Apply Theorem 6.1 to err(D) ≤ η of the target \bigwedge_{i=1}^{d} x_i function, giving a rational polynomial \ f_d of degree D = \ O(d^{1/2} \log \frac{1}{\varepsilon}) to bound \ \frac{1}{m} \sum_{j=1}^{m} |f_d(\theta \circ (j))| = (1)^{y(j)} ≤ η + \varepsilon.

Theorem 5.1 can solve this LP with \ n' = \sum_{k=0}^{d} n^{k(d)} variables in \ t = \ O(m^{1.5} n' \log(n'/\varepsilon)) time by \ O(log(n'/\varepsilon))-bit precision. The \ ell_1-norm regression chooses a hypothesis \ h = (\text{sgn}(f_d(\theta \circ x) - t)) for an appropriate threshold \ t \in [-1, 1] to become a weak empirical learner achieving \ err_h(D) ≤ \eta + \varepsilon + o(\varepsilon) [KKMS08]. Sufficiently many examples \ m = \ O(\varepsilon^2 n' \log(n'/\varepsilon)) turn this weak learner of description length \ O(n' \log(n'/\varepsilon)) to an actual one \ P(y ≠ h(x)) ≤ \eta + \varepsilon by Chernoff bound parameter \ γ = \varepsilon/\eta with significance:

\[\text{UGEB: } 2^{O(n' \log(n'/\varepsilon))} \cdot (e^{-\gamma/2+(2+\gamma)\eta|D|} \cdot 1[\eta > \varepsilon] + e^{-\gamma/3-\eta|D|} \cdot 1[0 < \eta < \varepsilon]) = o(\delta').\]

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 η'-noisy \ F is \ (1/2 - \alpha)-learnable with significance \ δ' for \ η ≤ η' ≤ 1/2 - \varepsilon, then it is \ (η + 2\varepsilon)-learnable with significance \ O(δ'/\alpha^2) under the same variate distribution \ P(x) by calling the \ (1/2 - \alpha)-learner for \ c_{\alpha^2}/\alpha^2 times. If the \ (1/2 - \alpha)-learner runs in \ t, then the \ (\eta' + 2\varepsilon)-learner in \ O(t/\alpha^2 + 1/\varepsilon^2) time.

**Theorem 6.4** (noisy refutation to agnostic learning). Let \ δ_{\varepsilon^4} := \frac{\delta}{m^4 \log^4 \frac{\delta}{m \log \frac{\delta}{c}}}. If \ η'-noisy \ F is refutable for any \ η ≤ η' ≤ 1/2 - \varepsilon with significance \ O(δ_{\varepsilon^4}) from \ m data in \ t time, \ η'-noisy \ F is agnostic learnable from \ m^2 \cdot O(\log \frac{n}{\varepsilon} \log \frac{1}{\varepsilon}) data in \ m^4 t \cdot O(\log^3 m \log \frac{m}{\varepsilon}) + 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 ≈ \frac{1}{m}, \nu_0 = c_{\alpha^2}/\alpha^2, \kappa_0 \gg \left(\frac{\log m}{\alpha}\right)^2 \log \frac{\nu_0 \log m}{\delta}, \m \gg \left(\frac{\alpha}{\varepsilon}\right)^2 \log \frac{m}{\varepsilon} and \ \tilde{m}' \gg \frac{\nu_0 \kappa_0 \log m \cdot O(t/\alpha^2)}{\varepsilon^2} + O(1/\varepsilon^2) time, and succeed with significance level \ \nu_0 \kappa_0 \log m \cdot O(\delta_{\varepsilon^4}) = O(\delta).\]

**Theorem 6.5** (noisy refutation to agnostic learning in smoothed analysis). If \ η'-noisy \ F is refutable for any \ η ≤ η' ≤ 1/2 - \varepsilon with significance \ O(δ_{\varepsilon^4}/\delta), \ η'-noisy \ 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.
6.2 Proof Theoretical 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 \( G = (I \cup J, \mathcal{E}) \) is \( r \)-bounded \((d, t)\)-expanding if the number of edge-induced \((d, t)\)-expanding subgraphs are bounded by \( r \):

\[
\text{r-rounded } \text{(d, t)-expansion: } \left| \{(u \cup v, w) \mid \emptyset \neq w \subset \mathcal{E}, u \cup v = \mathcal{E}[w], (\forall j \in u \Rightarrow |w[j]| \geq t) \} \right| \leq r.
\]

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

\[
\text{r-rounded \ (d, t)-expanding: } \Pr [G \text{ is } |\mathcal{I}|^{\frac{1}{2}+\zeta}\Delta\text{-bounded, } (d, t)\text{-expanding}] \geq 1 - \epsilon^k.
\]

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

\[
\sum_{\emptyset \neq w \subset \mathcal{E}} \Pr [v = \mathcal{J}[w], u = \mathcal{I}[w], |v| \leq D, |u| \leq k|v|, |u| + (\frac{t}{2} - \zeta)|v| - \frac{t - 1}{2} \leq |w| \leq k|v|]
\leq \sum_{|v|, |u|, |v|} (e^{2+|v|} \left( \frac{|v|}{|w|} \right)^{\frac{1}{2} - \zeta - 1} \Delta)^{|v|} \frac{|v| - 1}{2} \nu^{|w|}
\leq \sum_{|v|, |u|, |v|} (k^2 e^{2+k} \left( \frac{|v|}{|w|} \right)^{\frac{1}{2} - \zeta - 1} \Delta)^{|v| - 1} \cdot k^2 e^{2+k} \left( \frac{|v|}{|w|} \right)^{-\zeta - \frac{1}{2} \Delta}
\leq \sum_{|v|, |u|, |v|} e^{2+k} k^2 \left( |\mathcal{I}|^{\frac{1}{2}+\zeta} \Delta \right)^{|v| - 1} \cdot (k^2 e^{2+k} \left( kD\Delta \frac{|\mathcal{I}|^{\frac{1}{2}+\zeta} \Delta}{|\mathcal{I}|} \right)^{\frac{1}{2} - \zeta - 1})^{|v| - 1}
\leq 4k^6 e^{4+2k} |\mathcal{I}|^{\frac{1}{2}+\zeta} \Delta (kD\Delta \frac{|\mathcal{I}|^{\frac{1}{2}+\zeta} \Delta}{|\mathcal{I}|} \zeta)^{|v| - 1} = \epsilon^k |\mathcal{I}|^{\frac{1}{2}+\zeta} \Delta.
\]

The right-hand side of \( \leq \) 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 \((d, t)\)-expanding CSP \( G \) with \( \forall j \in J, |\mathcal{J}[j]| \leq \zeta d \), 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 \( X_j \in \{0, 1\}^{|\mathcal{J}|} \),

\[
\text{SoS hardness on bounded expansion: } \deg_{\text{SoS}}[\text{unsat}_G(x)] > \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}, \mathcal{J}'} \text{Pr}_{X_j}[X_j \notin \mathcal{S}_j] + \frac{|\mathcal{J}'| - |\mathcal{J}'|}{|\mathcal{J}|} \geq \frac{\zeta d}{3}.
\]

**Theorem 6.9** (Theorem 1.19). For \( 2 \leq d \leq \log(1/\epsilon) - O(1) \) and \( 0 \leq \eta \leq 1/2 - O(\epsilon) \), 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 \( \Omega(n(1-\epsilon) d/2) \) 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 \( \deg_{\text{SoS}}[\text{err}_D] > \eta \) \( \ll \zeta d/3 := n^\epsilon \). Let us learn a joint-distribution \( P(x, f(x)) \) having the uniform variate \( P(x) = 1/(2n)^d \) and the white-\( \bar{\eta} \)-noisy covariate:

\[
\text{White noisy constraint sampler: } \bar{\eta} P(x) \otimes |x, 0 \rangle + \bar{\eta} P(x) \otimes |x, 1 \rangle + (1 - 2\bar{\eta}) P(x, f(x)) |x, f(x)\rangle
\]

of \( f(x) = \bigwedge_{i=1}^d \theta \circ x_i \) and \( \bar{\eta} := \eta + (\epsilon + \epsilon) \epsilon \leq 1 - \Omega(\epsilon) \).

This mixture draws a data \( (X_j, Y_j) \sim D \) by first throwing the \( \bar{\eta} : \bar{\eta} : 1 - 2\bar{\eta} \)-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 \( \Pr(\text{err}_\theta(D) \leq \eta + c\varepsilon) < |\{0, 1\}^n|e^{-\gamma^2/2\bar{\eta}m} < o(\delta) \), so Definition 2.1 obliges the SoS learner to prove \( \text{err}_\theta(D) > \eta \). Similarly, the hitting sets \( \mathcal{J}_b := \{ j \mid B_j = b \} \) must have cardinality \( \forall b, |\mathcal{J}_b|/m \leq \varepsilon\bar{\eta}/\tilde{\eta} \) with significance \( 2e^{-\gamma^2/2\bar{\eta}m} = o(\delta) \) by Chernoff bounds of \( \gamma = \varepsilon\bar{\eta}/\tilde{\eta} \). The \( \mathcal{J}_b \) with \( b = 0, 1 \) induce CSP instances \( \mathcal{G}_b = (\mathcal{I} \sqcup \mathcal{J}_b, \mathcal{E}, \mathcal{S}_b) \) of the uniformity \( t = d \):

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

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

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

These CSP instances \( \mathcal{G}_b \) appeal \( \frac{|\mathcal{I}|}{\log(1/\varepsilon) - \text{ormity}} \geq \frac{1}{\log(1/\varepsilon) - \text{ormity}} \geq \frac{1}{\text{ormity}} \geq D \) to Lemma 6.7’s SoS hardness of bounded expansion, yielding a contradiction:

\[
\forall b \in \{0, 1\}, \zeta D/3 \leq \text{deg}_{\text{SoS}}[\text{unsat}_{\mathcal{G}_b} / \mathcal{S}_{b,j}] > \frac{1}{|\mathcal{J}_b|} \sum_{j \in \mathcal{J}_b} \Pr_{X_{b,j}}(X_{b,j} \notin \mathcal{S}_{b,j}) + \frac{|\mathcal{J}_b| - |\mathcal{J}_b'|}{|\mathcal{J}_b|} \Rightarrow \zeta D/3 \leq \text{deg}_{\text{SoS}}[\text{err}_\theta(D)] = \frac{1}{\log(1/\varepsilon) - \text{ormity}} \leq \text{deg}_{\text{SoS}}[\text{err}_\theta(D) > \eta] < \zeta D/3,
\]

\[
\varepsilon: \text{-ormity} \leq \frac{\log 1}{\varepsilon} - O(1) \Rightarrow \eta - \frac{2\bar{\eta}m^{-\frac{1}{2} + \zeta}}{2 + 2\varepsilon \tilde{\eta}} < \text{deg}_{\text{SoS}}[\text{err}_\theta(D) > \eta] < \zeta D/3.
\]

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

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

\[
\zeta D/3 \leq \text{deg}_{\text{SoS}}[\text{err}_\theta(D) > 2\bar{\eta}m^{-\frac{1}{2} + \zeta} + 2\varepsilon \tilde{\eta}] \leq \text{deg}_{\text{SoS}}[\text{err}_\theta(D) > \eta] < \zeta D/3,
\]

where \( \zeta \leq \frac{\log 1}{\varepsilon} - O(1) \) (Theorem 6.9).
Let $t = t_{\text{nc}}$ of Theorem 4.14. Since $|S_{i,j}| = 2^{(1-c)\eta}$ and Lemma 4.12’s cosets disjointly cover $S_{i,j}$, Lemma 4.12 presents $t$-uniform random variables $X_{b,j}$, deriving a contradiction to Theorem 6.8:

$$
\frac{c_0}{3} \leq \deg_{\text{SoS}}[err_{\mathcal{D}}(x) > \eta(1 - \frac{1}{2n}) + 2\eta n^{-\frac{1}{2} + \frac{1}{2}} + 2\epsilon \varepsilon \eta] \leq \deg_{\text{SoS}}[err_D(x) > \eta] < \frac{c_0}{3}.
$$

**Theorem 6.12** (Theorem 1.22). For $0 < c < 1$, $2 \leq d \leq \frac{1}{c} \log(1/\epsilon) - O(1)$, and $\Omega(1) \leq \eta \leq 1/2 - O(\epsilon)$, agnostic learning $\eta$-noisy, agnostic learning the $\eta$-noisy canonical planted AND

$$
\{\bigwedge_{i=1}^n \theta \land x_i \mid \theta \in \{0,1\}^{dn}\}
$$

under the uniform distribution perturbed by any shift of $H_s(G) = (1 - \sigma)^d$ demands either sample size $\Omega\left(\frac{n}{2(1-c)\eta}\right)^{(1-c)\eta n/2}$ or SoS proof of degree $\Omega\left(\frac{n}{2(1-c)\eta}\right)^{\epsilon}$.

**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 $k \geq 2$, the planted $k$AND is refutable from any $\eta$-noisy $n^{k/2} \cdot O\left(\frac{(k \log n)^3}{(1-2\eta)^5}\right)$ data in $t_{\text{adp}}(n^k) \cdot O(2^kn)$ time.

**Proof.** Changing $\beta \approx \frac{1}{2m}$ 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 kDNF satisfying $1 - 2\eta = \mathbb{E}[-1]^Y + 2\text{bias}_{f(D)}$ over the data distribution $(X, Y) \sim \mathcal{D}$, where $\eta := \Pr[Y \neq f(X)]$.

**Theorem 6.14** (refuting $\eta$-noisy planted $k$XOR). For $k \geq 2$, the planted $k$XOR is refutable from any $\eta$-noisy $n^{k/2} \cdot O\left(\frac{(k \log n)^3}{(1-2\eta)^5}\right)$ data in $t_{\text{adp}}(n^k) \cdot O(n)$ time.

**Proof.** Adapt Theorem 5.7’s bias measurement to the canonical $k$XOR function $f = \bigoplus_{i=1}^n \theta \land x_i$:

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

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

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 $k \geq 2$, the planted $k$AND is agnostically learnable from any $\eta$-noisy $n^{k/2} \cdot O\left(\frac{(k \log n)^3}{(1-2\eta)^5}\right)$ data in $t_{\text{adp}}(n^k) \cdot O(2^kn^2(\log n)^2)$ 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 $\beta \approx (1 - 2\eta)/2$, $\nu_0 = \frac{c_0}{(2m)^2}$, $m \gg n^{k/2} \cdot O\left(\frac{(k \log n)^3}{(1-2\eta)^5}\right)$, Theorem 5.9’s UGEB holds, and Theorem 6.3’s agnostic boosting finishes within $\nu_0 \cdot t_{\text{adp}}(n^k) \cdot O(2^kn)$. time.

**Theorem 6.16** (Theorem 1.21 for planted $k$XOR). For $k \geq 2$, the planted $k$XOR is agnostically learnable from any $\eta$-noisy $n^{k/2} \cdot O\left(\frac{(k \log n)^3}{(1-2\eta)^5}\right)$ data in $t_{\text{adp}}(n^k) \cdot O(n^2(\log n)^2)$ 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 $k \geq 2$, the planted $k$JUNTA is agnostically learnable from any $\eta$-noisy $n^{k/2} \cdot O\left(\frac{(2^k k \log n)^3}{(1-2\eta)^5}\right)$ data in $t_{\text{adp}}(n^k) \cdot O(2^k n(2^k k \log n)^2)$ 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 $\beta \approx \frac{1-2\eta}{2^k}$, 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 \( n^{[k/2]} \cdot O \left( \frac{k \log n}{\epsilon(1-2\eta)} \right)^2 \) data in \( t_{\text{sdp}}(n^k) \cdot O \left( n^{(k \log n)/(1-2\eta)} \right) \) learning time under any \( k \)-wisely \( O \left( \frac{1}{2^{k}(1-2\eta)} \right) \)-dense uniform flipper.

Proof. A reduction to Theorem 6.15 as 5.12 to 5.7, since Lemma 5.11’s recall guarantees bias\( \left( \frac{1}{2^k(1-2\eta)} \right) \), since \( 0 \leq \eta \approx \beta \) for \( \beta \approx (1-2\eta)/2 \) and \( \rho \geq \frac{1}{2^{k}(1-2\eta)} \) by Markov’s inequality parameter \( \gamma = \delta / \epsilon \) \( \Box \).

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 \((\beta_{\text{cmp}}, \beta_{\text{snd}})\)-gap (or \( \Delta \)-gap, \( \Delta:= \beta_{\text{cmp}} - \beta_{\text{snd}} \)) approximation of promise-MaxCSP assumes either \( \text{acc}(P) = \beta_{\text{cmp}} > \beta_{\text{snd}} = \text{acc}(P') \) 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 \( D \sim P_m \) and \( D' \sim P_m' \) as follows:

Verifiable Completeness: Show a witness to verify\(^{48} \) \( |\text{acc}(D) - \text{acc}(D')| \leq \frac{\Delta}{3} \rightarrow P = P' \).

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 \( P_{\text{cmp}} \) and \( P_{\text{snd}} \) over \( k\text{CNF} \)’s (resp. \( k\text{JUNTA} \)’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 distinguishing problems between “malicious” \( \omega(m) \) constraints \( D \) and \( D' \) with a slight difference \( \text{acc}(D) - \text{acc}(D') = \epsilon \) by taking empirical distributions to draw \( m \) i.i.d. constraints \( D \) and \( D' \) from \( D \) and \( D' \), respectively.

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

Proof. A reduction to Theorem 6.9 by filtering \( m \gg n^{\frac{1-k}{2}} \) positive data \( G_{\kappa} \) for \( \kappa \in \{\text{cmp}, \text{snd}\} \):

\[
\begin{align*}
& \text{Positive constraint sampler} \\
& \text{(discard negative examples):} \\
& \eta_{\kappa} P(x)|x,1) + (1-\eta_{\kappa})P(x,f(x))|x,f(x) \\
& \text{of } f = \sum_{i=1}^{k} \theta \circ x_i \text{ and } \eta_{\kappa} \cdot (1-\eta_{\kappa}) = \beta_{\text{cmp}} \cdot 1[\kappa = \text{cmp}] + \beta_{\text{snd}} \cdot 1[\kappa = \text{snd}].
\end{align*}
\]

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 \( G := G_{\text{cmp}} \otimes |1 \rangle \cup G_{\text{snd}} \otimes |1 \rangle \) must take the weighted accuracy gap \( \text{acc}(G_{\text{cmp}}) - \text{acc}(G_{\text{snd}}) \leq (1 + \epsilon)(\beta_{\text{cmp}} - \beta_{\text{snd}}) \) with significance \( e^{-\frac{2^k}{\epsilon}[(\beta_{\text{cmp}} - \beta_{\text{snd}})^m + o(\delta)]} \) by Chernoff bound of \( \gamma = \frac{\epsilon \delta}{\beta_{\text{cmp}} - \beta_{\text{snd}}} \). Definition 6.19’s verifiable completeness obliges to prove \( \text{deg}_{\text{SoS}} [\text{acc}(G_{\text{cmp}}) - \text{acc}(G_{\text{snd}})] = \text{err}_{\theta}(G) \geq \frac{3\delta}{4} \leq \zeta \Delta/3 := n^{\epsilon} \), a contradiction against Theorem 6.9’s CSP instances \( G_{\kappa} := (I \cup J_{\kappa}, E, S) \). Here \( J_{\kappa} := \{(x_{\kappa}(j), 1)\}_{j} \)

\[ \zeta \Delta/3 \leq \text{deg}_{\text{SoS}} [\text{err}_{\theta}(G) \geq \sum_{\kappa \in \{\text{cmp, snd}\}} s_{\kappa} \sum_{j} |x_{\kappa}(j)\rangle (1 - \text{unsat}_{G_{\kappa}}(x)) \] (where \( s_{\kappa} := \left( -1 \right)^{1[\kappa = \text{cmp}]} \))

\[ \leq \text{deg}_{\text{SoS}} [\text{err}_{\theta}(G) \geq 1 + \epsilon \cdot \left( \beta_{\text{cmp}} - \beta_{\text{snd}} \right)] + \sum_{\kappa} \left( \frac{s_{\kappa} \sum_{j} \text{Pr}[x_{\kappa}(j) \notin S_{\kappa,j}] + |x_{\kappa}(j)\rangle\langle x_{\kappa}(j)|]}{2m} \right) \]

\[ \leq \text{deg}_{\text{SoS}} [\text{err}_{\theta}(G) \geq 1 + \epsilon \cdot \Delta + n^{\frac{1-k}{2}} \cdot \zeta] \leq \text{deg}_{\text{SoS}} [\text{err}_{\theta}(G) \geq \frac{3\delta}{4}] \leq \zeta \Delta/3. \]

\(^{47}\)Set \( k = \log \left( \frac{1}{1 - 2\eta} \right) \). Take Theorem 2.9’s \( \frac{1}{2} \)-dense \( dn \)-bit flipper of cardinality \( O(k^2 \log (dn)) \).

\(^{48}\)The verifiable-completeness threshold \( \frac{\Delta}{3} \) could be any \( \epsilon \) between \( 1/2 < \epsilon < 1 \).
Theorem 6.21 (Theorem 1.25). Any gap approximation of the promise-Max$k$XOR under a marginally uniform distribution demands either $\Omega\left(\frac{n^{(1-c)k^2}}{n}\right)$ constraints or $\Omega\left(n^c\right)$ SoS-degree.

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

\[ \eta \mathcal{P}(x) \otimes |x,0\rangle + (1-\eta)\mathcal{P}(x,f(x))|x,f(x)\rangle, \]

where \( \eta = \frac{1/2}{\eta + (1/2)(1-\eta)} = \beta_c \cdot 1[\kappa = \text{cmp}] + \beta_{\text{snd}} \cdot 1[\kappa = \text{snd}] \).

\(\square\)

Theorem 6.22 (Theorem 1.27\textsuperscript{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 $\Omega\left(n^{(1-c)k/2}\right)$ constraints or $\Omega\left(n^c\right)$ SoS-degree.

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

\(\square\)

Theorem 6.23 (Theorem 1.26 for promise-Max$k$SAT). The promise-Max$k$SAT is $\Delta$-gap approximable by any $n^{k/2} \cdot O\left(\left(k \log n\right)^3/\Delta^5\right)$ constraints in $t_{\text{adp}}(n^k) \cdot O(2^k n)$ 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 $|\text{bias}_f(P,P')| = |\frac{1}{m} \sum_{(x,y) \in \mathcal{D}} 1[f(x) = 1] - \frac{1}{m} \sum_{(x',y') \in \mathcal{D}'} 1[f(x') = 1]|$. 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 $|\text{bias}_f(P_{\text{cmp}},P_{\text{snd}})| \approx \Delta$ (or larger) versus $|\text{bias}_f(P,P)| \approx O$. The former produces Theorem 6.13’s completeness proof by replacing $1 - 2\eta$ therein with $\Delta = \beta_{\text{cmp}} - \beta_{\text{snd}}$, and the latter Theorem 5.7’s soundness one.

\(\square\)

Theorem 6.24 (Theorem 1.26 for promise-Max$k$XOR). The promise-Max$k$XOR is $\Delta$-gap approximable by any $n^{k/2} \cdot O\left(\left(k \log n\right)^3/\Delta^5\right)$ constraints in $t_{\text{adp}}(n^k) \cdot O(n)$ time.

Proof. Adjust Theorem 6.23’s argument to Theorem 6.14’s bias measurement.

\(\square\)

Theorem 6.25 (Theorem 1.26). The promise-Max$k$CSP is $\Delta$-gap approximable by any $n^{k/2} \cdot O\left(\left(k \log n\right)^3/\Delta^5\right)$ constraints in $t_{\text{adp}}(n^k) \cdot O(2^k n)$ time.

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

\(\square\)

Theorem 6.26 (approximating promise-Max$k$SAT). The promise-Max$k$SAT is $\Delta$-gap approximable by any $n^{k/2} \cdot O\left(\left(k \log n\right)^3/\Delta^5\right)$ constraints in $t_{\text{adp}}(n^k) \cdot O(2^k n)$ time under any $k$-wisely $O\left(\frac{1}{2^k \Delta}\right)$-dense uniform flipper.

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

\(\square\)

\textsuperscript{49}$k = (c+1)\log \frac{1}{4c}$ implies $\beta_{\text{snd}} \geq 1 - 1/2^k \iff \beta_{\text{snd}} \geq 1 - (2c)^{c+1}$.
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 $\tilde{\Omega}(n^{k/2})$ 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 $\mathbb{E}[(\sum_{i \in w} X_i + Y) \mid \forall i \in w, [X_i/2] = a_i]$ on a location (or place) $(w, a) \in \binom{[n]}{k} \times [n]^k$ 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 $f_{\kappa} := \bigwedge_{i \in f_{\kappa}} x_i$, write $f_{\bigwedge w} := \bigwedge_{\kappa \in w} f_{\kappa}$ and $f_{\bigwedge w} := \bigwedge_{\kappa \in w} f_{\kappa}$. Inclusion-exclusion expands logical expressions $f \equiv b, f' \equiv b'$ of DNF $f, f'$, and $b, b', b' \in \{0, 1\}$ as

\[
\text{Inclusion-Exclusion (IE)}: \quad \text{ie}_c(f \equiv b) := \sum_{|w| = b} \sum_{w \subseteq f} (-1)^{|w|+b} f_{\bigwedge w}.
\]

\[
\text{Double Inclusion-Exclusion (DIE)}: \quad \text{ie}_c(f \equiv b, f' \equiv b') := \sum_{|w| \leq |w'| \leq c-1} \sum_{b \leq |w|, b' \leq |w'|} \text{ie}_c((-1)^{|w|+|w'|+b+b'} (f_{\bigwedge w} \land f'_{\bigwedge w'})).
\]

\[
\text{Inclusion-Exclusion on average}: \quad \mu_c(f \equiv b, f' \equiv b') := \frac{1}{\binom{c}{2}} \sum_{|w| \leq |w'| \leq c-1} \sum_{b \leq |w|, b' \leq |w'|} \text{ie}_c((-1)^{|w|+|w'|+b+b'} (f_{\bigwedge w} \land f'_{\bigwedge w'})).
\]

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 $\ell - 1 \geq c$ terms of $f$, its contribution to $\text{ie}_c(f \equiv b) - \text{ie}_c(f \equiv b)$ is $\sum_{\kappa=c}^{\ell-1} (\binom{\ell}{\kappa} - \binom{c}{\kappa})$. The call $\text{ie}_c(f \equiv b, f' \equiv b')$ the truncated coefficient of $x$. Its contribution to $\text{ie}_c(f \equiv b, f' \equiv b') - \text{ie}_c(f \equiv b, f' \equiv b')$ is the same amount $(\binom{\ell-2}{c-1})$, once the $x$ satisfies $\ell - 1$ terms of $f \equiv f'$.

Definition 7.2 ($\rho$-spread). A random vector $X \sim \prod_{i=1}^n S_i$ is $\rho$-spread with significance $\delta$ if

\[
\rho\text{-spread}: \forall S \subseteq \prod_{i=1}^n S_i, \forall i, \{|x_i \in S_i \mid x \in S\}/|S_i| \leq \rho \Rightarrow \Pr[\forall i, X_i \not\in S_i \cap S_i] \geq 1 - \delta.
\]

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

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

\[
\Pr[X_i \in S \cap S_i] \leq \frac{|\{x_i \in S_i \mid x \in S\}|}{|S_i|} \leq \frac{\delta}{d(1-\delta^{d+\delta^2/2})} < p_i (1 - p_i)^{d-1}
\]

of dependent $n$ events $X_i \in S \cap S_i$, deriving $\Pr[\forall i, X_i \not\in S \cap S_i] \geq (1 - p_i)^d > 1 - \delta$. \qed
**Definition 7.4 ((α, β)-inversion).** We say that a randomized algorithm \( A(\alpha, \beta) \)-inverts \( \{ f \} \) planting \( \theta \in \{ 0, 1 \}^d \) 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 \( \delta_{\text{inv}} := \delta/d \).

- **Correlation:** \( \text{corr}(\mathcal{D}) = \text{corr}(X, Y) := \mathbb{E}[(\pm 1)^{X_i} Y] - \mathbb{E}[(\pm 1)^{X_i}]\mathbb{E}[(\pm 1)^{Y}] \).
- **Invariance:** \( 0 < \exists \mu_i < 1, \forall(i, a), |\text{corr}(\mathcal{D})| - \mu_i \ll \beta. \)
- **(α, β)-inversion:** \( \Pr_{\mathcal{D}, A} \left[ \left| \Pr[X_i/2] = a \right| \geq \alpha/n \land |\text{corr}(\mathcal{D})| \geq \beta \right] \geq 1 - O(\delta_{\text{inv}}). \)

**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 \((X_{i,a}, Y_{i,a}) \sim \mathcal{D}_{i,a} := \{(x, y) \in \mathcal{D} \mid [\frac{x}{2}] = a\} \).
2. If \( |\mathcal{D}_{i,a}| < \frac{\alpha}{n} \), then return \(?\).
3. Compute \( \text{corr}_{i}(X_{i,a}, Y_{i,a}) \) and return zero if it is \( \geq \beta \), one if \( \leq -\beta \), and \(?\) otherwise.

**Theorem 7.5** (inverting canonical DNF). Let \( \beta_{\text{rel}} := \max(\frac{\epsilon}{2(\beta - 1)\delta_{\text{inv}}}, (\frac{k_s}{\alpha \delta_{\text{inv}}})^{1/2}) (\frac{s-2}{c-1}) \). Suppose \( \beta_{\text{rel}} \leq \beta \ll 1 \). Algorithm 1 can \((\alpha, \beta)\)-invert the canonical planted DNF \( \{ \bigvee_{\kappa=1}^{k_s} \bigwedge_{i=1}^{k_s} [\theta_{i+k_k} \circ x_{i+k_k}] \mid \theta \in \{ 0, 1 \}^{k_s} \} \) from any noise-free \( \mathcal{O}(\frac{1}{\epsilon \beta^2 \delta_{\text{inv}}}) \) data with pairwise \( \rho \)-dense attributes under any \( \epsilon \beta \)-away \( 2ck \)-independent flipper over \( \{ 0, 1 \}^{kn} \).

**Proof.** Definition 7.1’s IE calculates Definition 7.4’s \( \text{corr}_{i}(D_{i,a}) \) and exhibits Algorithm 1’s inversion performance. For the target canonical DNF expression \( f = \bigvee_{\kappa=1}^{k_s} \bigwedge_{i=1}^{k_s} x_i \), write \( f_{-\kappa} := \bigvee_{\kappa \not\in \{i\}}^{\kappa=1} f_{\kappa} \) and \( f_{\kappa-i} := \bigwedge_{\kappa \not\in \{i\}}^{\kappa=1} x_{\kappa} \). They express the relevance and irrelevance of \( x_i \) to \( f \) by \( f_{\text{rel},i} := f_{\kappa-i} \equiv 1 \land f_{-\kappa} \equiv 0 \), \( f_{\text{ir},0} := f_{\kappa-i} \equiv f_{-\kappa} \equiv 0 \), and \( f_{\text{ir},1} := f_{\kappa-i} \equiv f_{-\kappa} \equiv 1 \) as follows. Let \( \mu(f \equiv 1) := 2^{-|f|}, \mu(f \equiv 0) := 1 - 2^{-|f|}, \mu_c := \mu_c(f_{\text{ir},0}) + \mu_c(f_{\text{ir},1}) \), and \( \mu_i := \mu_c(f_{\text{rel},i}) \).

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

**Disjointness:** \( \forall(j : \not\equiv j'), \forall(i \not\equiv i'), [x_i(j)/2] = a \land [x_{i'}(j')/2] \neq [x_{i'}(j'/2)] \).

**Low degree:** \( \forall w, |w| \geq c, f_{-\kappa,\wedge w}(\theta \circ G(x(j))) \approx 0. \)

**Relevance:** \( \Pr[G(f_{-\kappa}(\theta \circ G(x(j))) = b)] \approx \mu_c(f_{-\kappa} \equiv b). \)

**Correlation on shift:** \( \mathbb{E}[G][\text{corr}_{i}(G(x(j)), y(j))] \approx (-1)^{\theta_i(a)} \mu_i. \)

**Correlation on data:** \( \mathbb{E}[G][\text{corr}_{i}(G(x(j)), y(j))] \approx (-1)^{\theta_i(a)} \mu_i. \)

**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 \( x_i \) of \( \{ f_{\kappa-i} \not\equiv f_{-\kappa} \}_{\wedge w} \) evaluates

\[
\Pr[\text{low-deg}(\theta \circ G(x(j)))] \leq \Pr[\exists w, |w| \geq c, \forall i' \in f_{-\kappa,\wedge w}, \theta_i \circ G(x_{i'}(j)) \equiv 1] \leq \left(\frac{s}{c}\right)/2^{ck-1}.
\]
Relevance: The inclusion-exclusion formula of $f_{-\kappa} \equiv b$ under low-deg approximates

$$|\Pr[f_{-\kappa}(\theta \circ G(x(j))) = b] - \mu_c(f_{-\kappa} \equiv b)|$$

$$= |\Pr[f_{-\kappa}(\theta \circ G(x(j))) = b] - \mathbb{E}[\mathbf{i}_e(f_{-\kappa} \equiv b)(\theta \circ G(x(j)))]| \leq \binom{s-2}{c-1} \Pr[-\text{low-deg}(G(x(j)))]$$

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

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

$$\mathbb{E}[(-1)^{\theta(a)} \mathbf{f} \mathbf{x} \mathbf{e} \mu - \mathbb{E}[(-1)^{\theta(a)} \big| \mathbf{D}]], \quad \mathbb{E}[\mathbf{Z} | \mathbf{D}], \quad \mathbb{E}[\mathbf{Z} | \mathbf{D}]$$

$$\text{for } \mathbb{Z} := \mathbb{E}_J[(-1)^{\theta(a)} \mathbf{f} \mathbf{x} \mathbf{e} \mu - \mathbb{E}_J[(-1)^{\theta(a)} \big| \mathbf{D}]]$$

$$\mathbf{Z}(x) = \mathbf{Z}_x - \mu_c(f_x), \text{ and } \mathbf{Z}_x = \mathbf{ie}_c(f_x)(\theta \circ x) \text{ of } \kappa \in \{\text{rel, ir0, ir1}\}.$$
Invariance: The corr-on-data measures on $|D_{i,a}| = m \geq \frac{(s-2)^2}{\beta^2 d_{\text{inv}}}$ and $(s-2)^2/2^{ck-1} \leq \beta d_{\text{inv}}$.

Average inequality: $E\left[|\text{corr}_i(G(x(J)), y(J)) - (-1)^{\theta(a)} \mu_i|\right] < 7(s-2)^2/(c-1)^2 + 7(s-2)^2/2^{ck-1} \leq 14\beta d_{\text{inv}}$.

It guarantees Definition 7.4’s invariance $|\text{corr}_i(G(x(J)), y(J)) - (-1)^{\theta(a)} \mu_i| \leq \epsilon \beta$ by Markov’s inequality of $\gamma = O(d_{\text{inv}})$ with significance $O(d_{\text{inv}})$. 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 $G$ 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 $(f_{\kappa-i} \lor f_{\kappa})_\wedge w$ to bound

$$\Pr[-\text{low-deg}(\theta \circ \tilde{G}(x(j)))] \leq \left(\frac{s}{c}\right)/2^{ck-1} + d_{\text{st}}(G(x_w(j)), \tilde{G}(x_w(j))) \leq 2e\beta.$$  

$(\alpha, \beta)$-inversion: The invariance detects $\theta_i(a)$ for the following reasons. First, the correlation’s average must be significant as $(1-\epsilon)\beta$. Otherwise, the invariance falsifies $|\text{corr}_i(X_{i,a}, Y_{i,a})| \geq \beta$. Secondly, $\mu_i > 0$ by the relevance $\mu_i/\mu(\kappa_{\kappa-i} \equiv 1) = \mu_{c-1}(f_{\kappa} \equiv 0) \approx \Pr_G[f_{\kappa}(\theta \circ G(x(j))) = 0] \geq 0$. Algorithm 1 must succeed in inverting $|D| \geq (1+\epsilon)|D_{i,a}|/\alpha/n$ data, since then $|D_{i,a}| > (s-2)^2/\beta^2 d_{\text{inv}}$ with CB’s significance $e^{s^2/2\alpha|D|} \leq o(d_{\text{inv}})$ under $\Pr[|X_{i,a}/2 = a|] \geq \alpha/n$. $\square$

Definition 7.6 (expanding DNF). REVIEW3’s DNF expression $f$ is $c$-wisely $k$-expanding if $\forall v \subset f, |v| \leq c \Rightarrow |\bigcup_{\kappa \in v} f_{\kappa}| \geq k|v|$.

Theorem 7.7 (inverting monotone DNF). For $\beta r_7 \leq \beta \ll 1$, Algorithm 1 $(\alpha, \beta)$-inverts a monotone variable of any planted $s$-term $k$DNF with $c$-wise expansion from any $n \cdot O\left(\frac{(s-2)^2}{\alpha^2 d_{\text{inv}}^2}\right)$ 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 $f = f_{\forall(t)} \lor f_{\forall(s)[t]}$ to those terms $j \in (t)$ containing $i$ and the others not holding it, and let $f_{\forall-w-i} := \bigwedge_{\kappa \in w} f_{\kappa-i}$ for $w \subset (t)$.

$$\text{Relevance, irrelevance:} \ f_{r_{e1,i}} := f_{\forall(t)-i} \equiv 1 \wedge f_{\forall(s)} \equiv 0, \ f_{r_{e0}} := f_{\forall(t)-i} \equiv 1 \wedge f_{\forall(s)} \equiv 0 \ \text{and} \ f_{r_{e11}} := f_{\forall(s)} \equiv 1.$$  

$$\text{Averages:} \ \mu_c(f_{r_{e1,i}}) = \mu_c(f_{\forall(t)-i} \equiv 1, f_{\forall(s)} \equiv 0), \ \mu_c(f_{r_{e0}}) = \mu_c(f_{\forall(t)-i} \equiv 0, f_{\forall(s)} \equiv 0) \ \text{and} \ \mu_c(f_{r_{e11}}) = \mu_c-1(f_{\forall(t)} \equiv 0).$$

Notice that the target DNF’s terms $f_k$ may be too long, mutually overlapping, and even contracting to each other, but the $ie_c$ adjusts as follows to preserve Theorem 7.5’s proof:

$$IE: \ ie_c(f \equiv b) := \sum_{|w| = b} \sum_{w \subset f_i, |f_{\wedge w}| < ck} (-1)^{|w|+b} f_{\wedge w}.$$  

$$IE \text{ on average:} \mu_c(f \equiv b) := \sum_{|w| = b} \sum_{w \subset f_i, |f_{\wedge w}| < ck} (-1)^{|w|+b-2} |f_{\wedge w}|.$$  

$$\text{Doubled IE:} \ ie_c(f \equiv b, f' \equiv b') := \sum_{w \subset f_i, w' \subset f_i'} [f_{\wedge w} \cup f'_{\wedge w}] < ck (-1)^{|w \cup w'| + b + b'} f_{\wedge w} \cup f'_{\wedge w'}.$$  

$\square$
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 \(\Pr[h_{\nu-1}(X) = 0, Y = 1] < \epsilon\).
2: **Variable selection.** Guess a set of variables \(f_\nu \subset [d]\). Let \(\theta_\nu = S_\nu = \emptyset\).
3: **Correlation retrieval.** For each \((i, a) \in f_\nu \times [n] - S_\nu\), feed \((D, i, a)\) 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 \(S_\nu\).
4: **Positively reliable cover selection.** \(h_\nu(x) = \bigwedge_{\{x_i, x_j\}} \theta_{\nu,i} \circ x_i\).
5: **Consistency measurement.** Return \textsc{Fail} unless both are true:

\[\Pr[\theta_{\nu,i}(X) = 1 \mid Y = 1, h_{\nu-1}(X) = 0] \geq 1/s.\]

\[\text{Small FPE: } \Pr[Y = 0, h_\nu(X) = 1] \leq (1 + \epsilon)2^k\beta.\]

6: **Induction.** \(h_\nu := h_{\nu-1} \lor h_\nu\), \(\nu := \nu + 1\).

7.2 Linear Time Proper Learning Monotone DNF

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

**Proof.** Set \(\nu_0 = s, \delta_{inv} = \alpha = \frac{\delta}{ks}, \beta = \frac{\epsilon}{2^{2k}\nu_0},\) and \(\rho = \left(\frac{2^{2ck-\epsilon}(ks)^2}{c^2}\right)\cdot \frac{ks}{\alpha\delta_{inv}}\), implying \(\beta_{2s} \ll \beta \ll 1\) by \((s-2)/(c-1) \frac{s^2}{\epsilon^2} \ll 2^{(c-2)k}\). Algorithm 2 may succeed if Step 5 never fails on \(f_\nu = \{x_{i+jk} \mid i \in [k]\}:

\[\prod_{FNE} h_{\nu-1}(x) = \prod_{\nu \neq \nu-1} \Pr[Y = 0, h_{\nu-1}(X) = 1] \leq \prod_{\nu \neq \nu-1} \Pr[Y = 0, h_\nu(X) = 1] \leq (1 + \epsilon)2^k\beta \nu_0 \leq (1 + \epsilon)\epsilon.

It converges to Definition 2.1’s \(2\epsilon\)-learning by Theorem 5.9’s UGEB analysis

\[\text{UGEB: } \Pr(h_{\nu-1}(x) \neq y) - \Pr(h_{\nu-1}(X) \neq Y) \geq \epsilon \leq |\mathcal{H}|e^{-\frac{1}{2}\epsilon m} \leq (2kn)^{-\frac{1}{2}}e^{-\frac{1}{2}\epsilon m} = o(\delta)\]
on \(|\mathcal{H}| = \prod_{\nu=0}^{\nu-1} |\mathcal{H}|\) of \(|\mathcal{H}|\) 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 \(\Pr[f_\nu(X) = 1 \mid Y = 1, h_{\nu-1}(X) = 0] \geq \frac{1}{s}\). Step 5 can attain the recall once Step 3 has correctly inverted \(\theta_\nu\) on the locations \((i + jk, [X_{i+jk}/2])\) of all \((i, j) \in [k] \times [s]\). Theorem 7.5 guarantees the inversion with significance \(ks \cdot O(\delta_{inv}) = O(\delta)\).

**FPE** holds under Step 3’s correct inversions and Definition 7.4’s prerequisite \(\forall (i,j) \in [k] \times [s], \Pr([X_{i+jk}/2]) \geq \alpha/n\). Lemma 2.3’s \((n, \alpha/n)\)-slice of \(\Pr([X_{i+jk}/2] = a)\) over \(a \in [n]\) guarantees the latter with significance \(ks \cdot \alpha/n \cdot n = ksa = \delta\). We may write \(f_\nu = \bigwedge_{i=1}^k x_i\) and \((k', k) := \{i \in [k] \mid (i, [X_i/2]) \in S_\nu\}\). Divide the hypothesis Step 4’ \(h_\nu\) into \(h_\nu = \bigcup_u h_u\) over \(u \in \{0,1\}^{kn}\) of \(h_u(x) := h_\nu \land \bigwedge_{i=1}^{k'} u_i \circ x_i\). Yao’s reduction takes the uniform random assignment \(U \sim \{0,1\}^{kn}\) and calculates the probability differentials between the target \(h_0 = f_\nu\) and the hypothesis \(h_\nu\).
along a sequence $h_0, \ldots, h_k'$ of AND functions $h_i := h_{i'} \wedge \bigwedge_{i=1}^{i'} U_i \circ x_i$. For $h_{i, b} := h_{i'} \wedge \bigwedge_{i=1}^{i'} U_i \circ x_i \wedge b \wedge \bigwedge_{i=i+1}^{k'} \theta_i \circ x_i$ and $h_{i, s} := h_{i'} \wedge \bigwedge_{i=1}^{i'} U_i \circ x_i \wedge \bigwedge_{i=i+1}^{k'} \theta_i \circ x_i$,

$$\Pr[Y = b, h_U(X) = 1] = \Pr[Y = b, f_{v'}(X) = 1]$$

$$= \sum_{i=0}^{k'-1} \Pr[Y = b, h_{i+1}(X) = 1] - \Pr[Y = b, h_{i}(X) = 1]$$

$$= \sum_{i=0}^{k'-1} \mathbb{E} \left[ (-1)^{\theta_i \circ X_i} (1[Y = b, h_{i+1}(X) = 1] - 1[Y = b, h_{i}(X) = 1]) \right] \quad \text{(\because Theorem 7.5’s rel+ir0+ir1 decomposition)}

$$\leq \sum_{i=0}^{k'-1} \Pr[h_{i+1}(X) = 1, f_{rel,i}(\theta \circ X)] + \mathbb{E} \left[ (-1)^{\theta_i \circ X_i} (1[h_{i+1}(X) = 1] - 1[f_{rel,i}(\theta \circ X)]) \right]$$

$$\leq \sum_{i=0}^{k'-1} \Pr[h_{i+1}(X) = 1, f_{rel,i}(\theta \circ X)] + \mathbb{E} \left[ (-1)^{\theta_i \circ X_i} (1[h_{i+1}(X) = 1] - 1[f_{rel,i}(\theta \circ X)]) \right]$$

$$\leq \sum_{i=0}^{k'-1} \frac{1}{2^{i+1}} (|\mu_i| + O(\beta \delta_{inv})) \quad \text{(\because Theorem 7.5’s relevance, correlation-on-shift, and correlation-on-data analyses)}$$

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

$$\Pr[Y = 0, h_{v'}(X) = 1] = \sum_{i=0}^{k} \Pr[Y = 0, h_{i}(X) = 1] = 2^k \Pr[Y = b, h_U(X) = 1]$$

$$\leq 2^k (|\mu_i| + O(\beta \delta_{inv})) < (1 + \epsilon)2^k \beta.$$

**Computational complexity:** Algorithm 2 spends $\Theta(n \cdot O\left(\frac{(s-2)^2}{\alpha^2 \delta_{inv}}\right)$ data to execute Step 3 in $kn \cdot \nu_0 \cdot O\left(\frac{(s-2)^2}{\delta_{inv}}\right)$ time with $O(k \delta_{inv}) + O(k \alpha) = O(\delta)$ significance.

**Theorem 7.9 (Theorem 1.4\textsuperscript{50}).** Suppose $\frac{s}{\delta} \leq 2^k$ and $\left(\frac{s-2}{s-1}\right)^{\frac{s-2}{s-1}} \frac{\ln(1/\epsilon)}{\epsilon^2} \leq 2^{(c-2)k}$. Algorithm 2 can PAC learn the planted monotone s-term DNF hiding $\theta \in \{0, 1\}^d$ and having $\epsilon$-wisely $k$-expanding terms in $n \cdot O\left(\frac{(s-2)^2}{\delta_{inv}}\right)$ time. It works on any $\epsilon$-noisy $n \cdot O\left(\frac{(s-2)^2}{\delta_{inv}}\right)$ data with pairwise $\frac{1}{2} \cdot \left(\frac{(s-2)^2}{\delta_{inv}}\right)^{2}$-dense attributes under any $\epsilon$-noisy $2k$-independent flipper over $\{0, 1\}^{ks}$. It loads a proper hypothesis class $\bigvee_{i=1}^{s} \bigwedge_{i=1}^{k} \theta_{i, v_i} \circ x_{i_i}$ with $\theta' \in \{0, 1, 2\}$.

**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 $\mathbb{E}[f_{v'}] \leq k$. Setting $\nu_0 := \frac{s}{\epsilon} \ln \frac{1}{\epsilon}$ (the other parameters are the same as Theorem 7.8) and applying Theorem 5.11’s recall can provide it on $\frac{s}{2^{k+1}} \ll \epsilon$:

$$\Pr[h_{v'}(X) = 1 \mid Y = 1, h_{v'-1}(X) = 0] \geq (\Pr[Y = 1, h_{v'-1}(X) = 0] - \epsilon \mathbb{E} - \frac{s}{2^{k+1}}) / \epsilon \geq \Pr[Y = 1] \mathbb{E} [h_{v'-1}(X) = 0] \mid Y = 1]$$

$$= \Pr[Y = 1] \mathbb{E} [h_{v'-1}(X) = 1 \mid Y = 1, h_{v'-1}(X) = 0] \leq \Pr[Y = 1] ((1 - \epsilon) \nu_0 - 1) < \epsilon.$$

**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 $(X_i \mod 2)_{i \in w}$ of $w \in \binom{d}{k}$. This section will extend it to smoothed analysis induced by $k$-wisely independent shifts flipping the polarities.

\textsuperscript{50}Set $k = O(\log s)$ and $1/\epsilon, 1/\delta \leq s^{O(1)}$. 
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 $f(x), f(x|r) : [2n]^d \rightarrow \mathbb{Z}_q$. 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 $\theta_i \circ x_i := \theta_i(\lfloor|x_i/2\rfloor)\langle-1\rangle^{|x_i|}$:

\begin{align*}
\text{Planted} \ FT: & \ f(x) := \sum_{|w| \leq k} \hat{f}_w \prod_{i \in w} \theta_i \circ x_i \text{ such that } \Pr[Y \neq f(X)] \leq \eta, \\
\text{Probabilistic} \ FT: & \ f(x|\mathcal{R}) := \sum_{|w| \leq k} \hat{f}_w \prod_{i \in w} \theta_{R,i} \circ x_i \text{ such that } \forall (x,y) \in \mathcal{D}, \Pr[|f(x|\mathcal{R}) - f(x)|] \leq \eta.
\end{align*}

They are non-degenerate if $\forall w \in \binom{\emptyset}{n}^,$, $\hat{f}_w \in \mathbb{Z}_q^*$. 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 degree-$k$ planted FT from data $(X,Y) \sim \mathcal{D}$ if it can estimate them within accuracy $\beta$ as follows, where $\delta_{\text{inversion}} := \frac{\delta}{m}$ (resp. $\delta_{\text{inversion}} := \frac{\delta}{n}$ when $\theta \in \mathbb{Z}_q^*$).

\begin{align*}
\text{Coefficient} \ (\alpha, \beta)-\text{inversion}: & \ \Pr_{\mathcal{D}, \mathcal{A}} \left[ \Pr[|X_w/2| = a] \geq \alpha \mu^k \left( \text{resp. } \Pr[|X_w/2| \leq a] \geq \alpha \mu^k \right) \right] \\
& \Rightarrow |\mathcal{A}(\mathcal{D}, w,a) - \theta_{w,a}| \leq \beta \\
\text{Parameter} \ (\alpha, \beta)-\text{inversion}: & \ \Pr_{\mathcal{D}, \mathcal{A}}[\mathcal{A}(\mathcal{D}) = \theta] \geq 1 - O(\delta_{\text{inversion}}).
\end{align*}

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

1. Input a dataset $\mathcal{D}$ and a query $(w,a) \in \binom{\emptyset}{n} \times [n]^k$. Let $\mathcal{D}_{w,a} := \{(x,y) \in \mathcal{D} \mid |x_w| = a\}$ (resp. $(x,y) \in \mathcal{D} \mid x_w \subset a$ when $\theta \in \{0,1\}^n$).
2. Filter $\mathcal{D}$ to a sub-data $(X_{w,a}, Y_{w,a}) \sim \mathcal{D}_{w,a}$. If $|\mathcal{D}_{w,a}|/|\mathcal{D}| < \alpha \mu^k$, then return $\theta$.
3. Compute and output $\mathcal{C}_{w,a}(\mathcal{D}_{w,a}) = \mathcal{C}_{w,a}(X_{w,a}, Y_{w,a}) := \mathbb{E}[Y_{w,a} \cdot \prod_{i \in w} (-1)X_{w,a,i}]$.

Algorithm 3 can invert the Fourier coefficient $\theta_{w,a}(\theta)$ 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 $h^k_w(g) = h^k_{\mathbb{Z}, J}(g) := (h^k_{\mathbb{Z}, J}, h^k_{\mathbb{Z}, J})$ indexed by $\kappa \in \{\dim, \text{hash}, \text{rem}\}$ 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 $\delta_\kappa := \frac{\delta}{m_{\text{inversion}}}$. Let $g \in \{0,1\}^{m_{\text{dim}}} \ x_{i}(j)$ := $2|\{j \mid x_{i}(j)/2 | + x_{i}(j) \oplus g(|\{j \mid x_{i}(j)/2 |)$, The Fourier $(w,a)$-coefficient inversion under $g$ consumes $m_{\text{dim}} := \frac{2^{2^k \eta}}{\beta \delta_{\text{inversion}}} + \frac{2^{2k}}{\delta_{\text{inversion}}} \ (\text{resp. } m_{\text{hash}} := \frac{2^{2^k \eta}}{\beta \delta_{\text{inversion}}} + \frac{2^{2k}}{\delta_{\text{inversion}}}) \ (\text{exams})$.

\begin{align*}
\text{Small dimension: } & \ h^\text{dim}_{\beta}(g) := g(x_w(j)) \in [2n]^{d-k} \ \text{and } |Q_{\text{dim}}| := |[2n]^{d-k}| = (2n)^{d-k}.
\text{Small hashes: } & \ h^\text{hash}(g) := f(x_w \mod 2 \mid |x_w/2| = a, x_{w,c} = g(x_w(j))) \in \mathbb{Z}_r^{0,1}^{w} \\
& \text{and } |Q_{\text{hash}}| := (2^{2^k \eta} + 2k \ln 2^{2^k \eta}) \ (\text{exams})
\text{Sparse remainders: } & \ h^\text{rem}_{\beta}(g) := \left( f - \sum_{v, w \cap v = \emptyset} \hat{f}_v \prod_{i \setminus w} \theta_i \circ x_i \right) (x_w \mod 2 | |x_w/2| = a, x_{w,c} = g(x_w(j))) \ \text{and } |Q_{\text{rem}}| := \{\xi \mid \Pr[h^\text{rem}_{\beta}(G) = \xi]} \geq \delta_{\text{rem}}\}.
\end{align*}

Theorem 7.12 (inverting Fourier coefficients). Suppose $m = m_{\text{dim}} \ (\text{resp. } m_{\text{hash}} \leq \frac{q}{\beta r}, \beta \ll 1, 2^{k \eta} \beta \leq \beta_{\delta_{\text{inversion}}}$). 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 $\mathcal{D}_{w,a} = \{(G(x(j)), y(j))\}_{j=1}^{m}$ within range $\forall j, |y(j)| \leq r$ under any $(h^{\kappa}_{w,a}, \delta_\kappa)$-hashed $\beta_{\delta_{\text{inversion}}} \text{r-} \text{away} 2k$-independent (resp. $k$-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 $\sum_{j} y_{j} \prod_{i \in w, e} (-1)^{g(x_{j}(i))} \ll q$ over $\mathbb{Z}$ rather than the ring $\mathbb{Z}_q$. Let us first do it in an ideal situation that $\Pr[|y(J) - f(G(x(J)))| = 1,$
∀ξ, \Pr[h_j^\gamma(G) = ξ] > 0 \implies \Pr[h_j^\gamma(G) = ξ] ≥ δ_κ, and the shift G is perfectly 2k-independent.

**Small coset:** The \( D_{w,a} \) can identify the truth-table of \( h^\text{hash}_{\xi} \in \{h^\text{hash}_j(g)\}_{j,j} \) under \( h_j^\gamma(G) = \xi \).

Since \( G \) is \((h_w^\epsilon, δ_κ)\)-hashed 2k-independent, Chebyshev’s inequality of \( \gamma = \frac{2^k}{\epsilon/m} \) makes the \( m ≥ \frac{2^k}{\epsilon/m} \) data in \( D_{w,a} \) to witness \( (G(x_w(j)) mod 2, y(j)) = (b, h^\text{hash}_\text{w}(b)) \) for every \( b ∈ \{0, 1\}^w \):

\[
\frac{1}{m} \sum_{j=1}^{2^k} \Pr(G(x_w(j)) \equiv b \mid h^\text{hash}_\text{w}(G_w^e) = ξ) - \frac{1}{2^k} \leq \frac{1 - 1/2^k}{m - 2^k} \implies
\]

k-unif on data: \( \Pr_G[(\forall b, |\Pr_j[G(x_w(J)) \equiv b \mid h^\gamma_j(G_w^e) = ξ]) - \frac{1}{2^k} | ≥ \frac{1 - 1/2^k}{m - 2^k} \leq 2^k : γ ≤ O(δ_{11})]. \)

δ-small hash: \( \forall b ∈ \{0, 1\}^w, (\forall v ∈ w, |h^\text{hash}_\text{w}(b)| ≤ r ∧ (|h^\text{hash}_\text{w}^v|) = 2^{-k}\sum_{b\in\{0,1\}^k}h^\text{hash}_\text{w}(b) \prod_{i\in v}(-1)^b = r. \)

**Correlation on data:** Fixing \( h_j^\gamma(G) = ξ \) induces a substitution \( x_w \leftrightarrow G(x_w(J)) \) to collapse \( h^\text{hash}_\text{w}(G_w^e) \) to \( h^\text{hash}_\text{w} \) and yield \( h^\text{hash}_\text{w}(x) = θ_w(a) \prod_{i\in u}(-1)^x_i + \sum_{v⊂w, v̂≠w}h^\text{hash}_\text{w} \prod_{i∈v}(-1)^{x_i} \) of \( x ∈ \{0, 1\}^w \). It can invert \( θ_w(a) \) via the correlation-on-data analysis under the random flipper \( G \):

\[
E_{G,j}[\text{corr}_w(G(x(J)), y(J))] = \sum_{ξ} E[f(G(x(J))) \prod_{i∈w}(-1)^{G(x_i)}] h^\text{hash}_\text{w}(G_w^e) = ξ \cdot \Pr[h_j^\gamma(G_w^e) = ξ]
\]

\[
= \sum_{ξ} E[2^{-k}\sum_{b\in\{0,1\}^w}h^\text{hash}_\text{w}(b) \prod_{i∈w}(-1)^b_i \cdot \Pr[h_j^\gamma(G_w^e) = ξ] = θ_w(a) \cdot \Pr[h_j^\gamma(G_w^e) = ξ = θ_w(a)].
\]

The zero-averaged correlations \( \text{corr}_w(G(x(j)) \):= \text{corr}_w(G(x(j)), y(J)) - θ_w(a) \) are mutually perpendicular \( E_{G,j}[\text{corr}_w(G(x(j))) \text{corr}_w(G(x(j'))) | h^\text{hash}_\text{w}(G) = ξ = 0 \) on the perfectly 2k-independence assumption, inverting the \( θ_w(a) \) for variance as well:

\[
E_{G,j}[\text{corr}_w(G(x(J)), y(J))] = \frac{1}{m} \sum_{j,j'} E_G[\text{corr}_w(G(x(j)), y(j)) \cdot \text{corr}_w(G(x(j')), y(j')) | h_j^\gamma(G_w^e) = ξ \cdot \Pr_G[h_j^\gamma(G_w^e) = ξ]
\]

\[
= \frac{1}{m} \sum_{j,j'} \sum_{1} E_G[(\text{corr}_w(G(x(j)), y(j)) - θ_w(a)) \cdot \Pr_G[h_j^\gamma(G_w^e) = ξ]
\]

\[
= \frac{1}{m} \sum_{j,j'} \sum_{1} \frac{1}{m} \sum_{i∈w}(-1)^b_i \cdot \Pr[h_j^\gamma(G_w^e) = ξ]
\]

\[
≤ \frac{1}{m} \sum_{j,j'} \sum_{i∈w}(-1)^b_i \cdot \Pr[h_j^\gamma(G_w^e) = ξ] ≤ 2^{2-k}/m. \) (=: the small hash.)

Chebyshev’s inequality parameter \( \gamma = O(1/\sqrt{\epsilon/\epsilon_{11}}) \) and \( m ≥ \frac{2^k \epsilon}{\beta/\epsilon_{11}} \) guarantees

\[
Correlation on data: \Pr_G[(\text{corr}_w(G(X,ω,a)) ≥ \sqrt{2^{k-1} - 1}/m] ≥ \beta \leq O(1/\gamma) = O(\delta_{11}).
\]

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

\[
E_{G,J}[(\tilde{y}(J) ≠ f(G(x(J))) | h_j^\gamma(G_w^e), G(x_w(J))] ≤ η ≤ \frac{β\delta_{11}}{2r} \implies 2^{k-r}η ≤ \beta\delta_{11}
\]

\[
⇒ \forall b ∈ \{0, 1\}^k, \Pr_{G,J}[\tilde{y}(J) ≠ h^\text{hash}_\text{w}(b) | h_j^\gamma(G_w^e), G(x_w(J))] = b ≤ \frac{εβ}{r} \implies \text{Markov-ineq of } γ = \frac{δ_{11}}{ε'}
\]

\[
⇒ \text{correlation under noise} \| \text{corr}_w(G(X,ω,a), Y,ω,a) ≤ ε'β + max_j \tilde{y}(j) \cdot ε'/r ≤ 2ε'β. \)

Although the actual shift \( G_κ \) may take \( 0 < \Pr[h_j^\gamma(G_κ) = ξ] < δ_κ \) for \( ξ ∈ Q_κ \), Lemma 2.3’s \( (0, δ_κ) \)-slice bounds its contribution \( \Pr_{h_j^\gamma(G_κ)}[h_j^\gamma(G_κ) < δ_κ] ≤ δ_κ \cdot |Q_κ| ≤ \frac{β\δ_{11}}{r} \) in any three
\(\delta_\kappa\) of \(\kappa \in \{\text{dim.shh.ren}\}\). The local shift \(\tilde{G}_\kappa(x_w(J))\) may be \(\frac{2\beta_\mu}{r}\)-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 \(\gamma = \delta_{7,11}/\epsilon'\):

\[
\mathbb{E}[\text{corr}_w(\tilde{G}_\kappa(X_{w,a}), Y_{w,a})] \\
\leq \left\{ \begin{array}{l}
\mathbb{E}[|\text{corr}(\tilde{G}_\kappa(x(J)), \tilde{y}(J)) - \text{corr}(G(x(J)), \tilde{y}(J))| + \max_j |\tilde{y}(j)| \cdot \Pr[\text{Pr}[h_{i,j}^\kappa(G(x_w))] < \delta_\kappa] \\
+ \mathbb{E}[\text{corr}_w(G(x(J)), \tilde{y}(J)) | \text{Pr}[h_{i,j}^\kappa(G(x_w))] = \xi] \geq \delta_\kappa
\end{array} \right.
\leq \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_{\text{inversion}} \Pr[\text{corr}_w(\tilde{G}_\kappa(X_{w,a}), Y_{w,a})] \geq 2\beta \delta_{7,11} \gamma + 2\epsilon' \beta = 4\epsilon' \beta \leq 1/\gamma = O(\delta_{7,11}).

**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):

\[
\text{Corr on data: } \Pr[|\text{corr}_w(G(x_w(J))| \geq \epsilon_\beta] = \Pr[\text{corr}_w(G(x_w(J)) + 2r) - \mu] \geq \beta' : = \frac{\epsilon_\beta^2}{3r}
\leq e^{-\frac{(\beta'/\mu)^2 m}{3r}} \leq O(\delta_{7,11}) \quad \text{for } \mu = \frac{\theta_m(a) + 2r}{3r}.
\]

\[\square\]

**Algorithm 4 Inverting planted parameters**

1. **Input a dataset** \(D\), execute the following 1–4 and output \(\theta \in \mathbb{Z}_q^{dn}\).
2. **Linear case.** When \(k = 1\), query \((D, i, a)\) to Algorithm 3 for \((\alpha, \beta)\)-inverting every \(\theta_i(a)\) of \((i, a) \in (d \times [n])\), and finish. The following steps suppose \(k \geq 2\).
3. **Base location selection.** Guess a base location \((w_0 \cup i_0, a_0) \in (d, n)^{w_0 \cup i_0}\) at which \(\theta_{w_0(a_0)}(\theta_{i_0}(a_0, i_0))\) is invertible. Let \(w_0, i, a' := (w_0, i) \cup i'\) for \((i, a') \in w_0 \times w_0 = w_0 \times ((d \cup w_0)\).
4. **Fourier inverting the base parameters.** Fix an arbitrary \(i_1 \in w_0\). For all \(i \in w_0\), query \((D, w_0, a)\) and \((D, w_0, i, a)\) to Algorithm 3, and retrieve \(\theta_i(a)\) in the following calculus:

\[
\frac{\theta_i(a)}{\theta_i(a_1)} = \frac{\theta_i(a_2)}{\theta_i(a_3)} = \frac{\theta_i(a, i_0)}{\theta_i(a, i_1)} = \frac{\prod_{k \in w_0} \theta_k(a, i_0)}{\prod_{k \in w_0 \cup i_0} \theta_k(a, i_1)} = \frac{\theta_i(a, i_0)}{\theta_i(a, i_1)}.
\]

4. **Fourier inverting all parameters.** Query \((D, w_0, i, a')\) to Algorithm 3 for \((\alpha, \beta)\)-inversion of \(\theta_i(a') = \theta_i(a, i) \cdot \prod_{\kappa \in w_0 \cup i, a} \theta_{\kappa}((a_0 \cup a_1, \kappa))/\prod_{\kappa \in w_0} \theta_{\kappa}(a_0, \kappa)\) until retrieving \(\theta\).

**Theorem 7.13** (Theorem 1.5\(^{51}\)). Suppose \(m_{7,12} \ll \epsilon\gamma/r, \beta < 1, \text{ and } 2^\beta r \eta \leq \beta \delta_{7,11}\). 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 \(D_{w,a}\) contains noise at most \(\eta\). Then, Algorithm 4 can \(\theta\)-invert degree-\(k\) planted FT over \(\mathbb{Z}_q\) in \(O((d + (d + n)^m)\) time from any data \([G(x(j)) , y(j)])_{j=1}^m\) of size \(m = O(\frac{m_{7,12}^{m_{7,12}}}{\alpha \beta^2})\) (resp. \(m = O(\frac{m'_{7,12}^{m'_{7,12}}}{\alpha \beta^2})\) and range \(\forall j, |y(j)| \leq r\) under any \((h_{\text{hash}}, \delta_{\text{hash}})\)-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 \(\frac{1}{\alpha \beta^2} \leq O(1), \epsilon \gg \frac{n^{2+1/2k-1}}{r}, r = q^{1/2k+1}, \beta_{7,11} = 1/n, \eta \ll 1/(nr), \text{ and } \mu = 1/n\). Lemma 2.11 provides a probabilistic shift \(G\) of cardinality \(O(q^{1/2}(nr)^3)\).
Proof. If all \((\alpha, \beta)\)-inversions of \(Q = \{(w_0, a_0), (w_{0-i_0}, a_0), (w_{0-i_1}, a_0)\} \) 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 \(\theta \in \mathbb{Z}_q^\alpha\).

\((\alpha, \beta)\)-inverting Fourier coefficients: Algorithm 4’s Step 2 must choose a location \((w_0, a_0)\) such that \(\theta_w(a_0)\) is invertible and \(\Pr[[X/w]/2] = a \geq \alpha \mu_k\). 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. \(\mu\)-cover) over the given \(m = O(m_{12}/(\alpha \mu_k))\) (resp. \(m = O(m_{12}/(\alpha \mu_k))\)) data. CB parameter \(\gamma = 1\) with significance level \(|\{(a, i) \in [n] \times (d))\}| \cdot e^{-1/2 \cdot \alpha \mu_k} \leq q/r\) guarantees:

\[
\text{Sufficiently many examples: } \forall (w, a), m_{12} (\text{resp. } m_{12}') \leq |D_{w,a}| \leq 2m \cdot \alpha \mu_k \leq q./r.
\]

Since \(\delta_{7,11} \leq \frac{\delta}{dt}, \) Step 4 inverts \(\theta_i(a)\) of \(\forall (i, a) \in (d) \times [n]\) with significance \(O(\delta_{7,11}) \cdot dn = O(\delta). \)

7.4 Inverting Linear Fourier Transforms and Breaking LWE

Theorems 7.12 has demanded a large modulus \(|D_{w,a}| \ll q/r\) for Fourier inverting \(\theta_w(a)\) over \(Z_q\) from \(D_{w,a} \subset [2n]^d \times Z_r\). Previously, modulus amplification have brought remarkable breakthroughs in computational complexity theory, e.g., Toda’s PP = \(\bigoplus P\) [Tod91], Beigel and Tarui’s ACC \(\subset\) SYM \& AND_{plag(m)} [BT94], and Williams’s NEXP \(\not\subset\) ACC [Wil14a]. This section will show that the modulus amplification can solve LWE and even GapSVP\(_{\mathbb{Z}_q}\) 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\ell polynomial \(\phi_\ell(x)\) with the leading coefficient \((-1)^{\ell+1}(2^{\ell-1})\) such that

\[
\text{Modulus amplification: } (x \equiv 0 \mod m \Rightarrow \phi_\ell(x) \equiv 0 \mod m^\ell) \wedge (x \equiv 1 \mod m \Rightarrow \phi_\ell(x) \equiv 1 \mod m^\ell).
\]

Theorem 7.15 (inverting linear Fourier transform). Let \(p\) be an odd prime number coprime with \(2(\ell-1), k = (2\ell - 1)v\) and \(|Q_{\text{rem}}| = p\) so \(\delta_{\text{rem}} = \frac{\beta \delta_{7,11}}{tp}\). Suppose \(m_{12}\) (resp. \(m_{12}’\)) \(\ll p^\ell/r\), \(\beta \ll 1\), and \(2k^2rp\beta \leq \delta_{7,11}\). Suppose that \(X\) is \(k\)-wisely \((\mu, \alpha)\)-sparse (resp. \(k\)-wisely \((\mu, \alpha)\)-cover when \(\theta \in \mathbb{Z}_q\)) at every location \((w, a)\) queried in Algorithm 4, and the sub-data \(D_{w,a}\) contains noise at most \(\eta\). Then, the linear planted FT over \(Z_p\) is invertible in \(O((d) + dn)m\) time from any data \(\{(G(x(j)), y(j))\}_{j=1}^m\) of \(m = O(m_{12}/(\alpha \mu_k))\) (resp. \(m = O(m_{12}/(\alpha \mu_k))\)), \(\forall j, |y(j)| \leq r\), and \(|y((m))| \leq u\) under any \((h_{\text{rem}}, \delta_{\text{rem}})\)-hashed \(\beta \delta_{7,11}/r\) away \(k\)-independent (resp. \((h_{\text{rem}}, \delta_{\text{rem}})\)-hashed \(\beta \delta_{7,11}/r\)-away \(km^\ell_{12}\)-independent) flipper \(G\).

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

\[
\text{Modulus amplification: } y = y \mod p, 1[y = y] = \prod_{y' = y((m))} \frac{y - y'}{y - y'} \Rightarrow
\]

Algorithm 4 can \(\theta\)-invert from the modulus amplified covariate dataset \(\{G(x(j)), \psi_t'(y(j))\}_{j=1}^m\). Fixing \(h_{\text{rem}}'(G_w) := \sum_{i \in w} f_i(z_j) \cdot (-1)^{\xi_i} = \xi \in \mathbb{Z}_p\) determines the hash function \(h_{\text{rem}}'(G_w) = \sum_{i \in w} f_i(z_j) \cdot (-1)^{\xi_i} = \xi \cdot \{0, 1\}^w \to \mathbb{Z}_p\) and its modulus amplification \(h_{\text{rem}}'(x) := \psi_t(h_{\text{rem}}'(G_w)) : \{0, 1\}^w \to \mathbb{Z}_p\). Accordingly, the modulus amplified degree-\(k\) Fourier transform \(h_{\text{rem}}'(x)\) over \(\mathbb{Z}_p\) makes Theorems 7.12 and 7.13’s proofs valid on the \(h_{\text{rem}}'(G_w)\)’s sparseness \(|Q_{\text{rem}}| = p\).
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.

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

Let $1_w = (1, \ldots, 1)$ be the all-one vector over $i \in w$. Algorithm 4 can invert LWE by choosing $a_0 = 1_{u_0}$ and making $\sum_{i \in u_0} \theta_i a_{0,i}(-1)^{G(a_0,i)} = \sum_{i \in u_0} \pm \theta_i$ concentrate near zero under the i.i.d. signs of the small secrets $\theta_i$. This $(\alpha, \beta)$-inversion algorithm queries about $\mathcal{W}_{u_0,i_0,i_1} := \{(w_0, 1_{u_0})\} \cup \{(w_{0,-i_1,i'}, 1_{w_{0,i_0,i'}})\}_{i \in [n]}$ of $w_0 \in \binom{[n]}{k}$, $i_0 \notin w_0$, $i_1 \in w_0$, and $i' = i'(i)$ such that $(i \in w_0 \Rightarrow i' = i_0) \land (i \notin w_0 \Rightarrow i' = i_1) \land (i \in w_0 \lor \{i_0\} \Rightarrow \theta_i \neq 0)$.

Theorem 7.17 (Theorem 1.65). Let $p$ be an odd prime co-prime with $\binom{2(\ell-1)}{\ell-1}$, $v = 2r + 1$, $|\mathcal{Q}_{\text{rem}}| \approx v$, $k = (2\ell - 1 - w)$, $\gamma_{\text{sm}} = (2k\log \frac{r}{\ell})^{1/2}$, $m_{\ell,17} := \frac{2k(s_{\gamma_{\text{sm}}})^5}{\beta^2 \gamma_{\text{sm}}^2}$, $\gamma_{\text{sm}} := \frac{(ks_{\gamma_{\text{sm}}})^5}{\beta^2 \gamma_{\text{sm}}^2}$, and $\gamma_{\text{sm}} := \frac{2k{s_{\gamma_{\text{sm}}}}^5}{\beta^2 \gamma_{\text{sm}}^2}$. Suppose $m_{\ell,17}$ (resp. $m_{\ell',17}$) is $\ll 2^{\ell/r}$ and $\gamma_{\text{sm}} \ll r$. Suppose that $X$ is $k$-wisely $(\mu, \alpha)$-sparse at every place $(w, 1_w) \in \mathcal{W}_{w,i_0,i_1}$. Then, LWE over $\mathbb{Z}_p$ can retrieve small secrets $\forall i, [\theta_i \leq s \in O((\ell + n)m)$ time from any $m \gg m_{\ell,17} \cdot p/(\alpha \beta)$ data under any $(h_{\text{rem}}^{\text{res}}, k_{\text{rem}}^{\text{res}})$-hashed $\beta_{\ell,11}/r$-away $2k$ (resp. $k_{\ell',17}$) independent flipper $G$ if $h_{\text{rem}}^{\text{res}}(G_w) \in \mathbb{Z}_p$ is $\beta_{\ell,11}/r$-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 $\mathcal{D}_{w,1_w} \to \mathcal{D}_j = \mathcal{D}_{w,i}$, $\{(G(x(j), y(j)) \in \mathcal{D}_{w,1_w} | \Xi = \xi\} \subset \Xi := h_{\text{rem}}^{\text{res}}(G_w)$, over which $G$ is $(h_{\text{rem}}^{\text{res}}, k_{\text{rem}}^{\text{res}})$-conditionally $\beta_{\ell,11}/r$-away $2k$-independent. Suppose the ideal case discussed in Theorem 7.12. Chernoff (resp. Chebyshev) bound parameter $\gamma = \frac{\gamma_{\text{sm}}}{k^{1/2}}$ (resp. $\frac{2k(s_{\gamma_{\text{sm}}})^5}{\beta^2 \gamma_{\text{sm}}^2}$) on $q^2(x) := \sum_{i \in w} \theta_i(-1)^{x_i} = h^{\text{hash}}_j - h^{\text{rem}}_j$ makes the smallness (resp. $k$-uniform-on-data) sharper:

Smallness: $Pr[G][|q^2(G(x_{w}(j)))] \geq 2e^{-\gamma^2/(2+\gamma)k/2} \ll \gamma_{\text{sm}} \gamma_{\text{sm}}$. 

Smoothness: $\lfloor |\xi| \leq r - \gamma_{\text{sm}} \Rightarrow |q^2(G(x_{w}(J)) + |\xi| \leq r \rfloor \land (|\xi| > r + \gamma_{\text{sm}} \Rightarrow |q^2(G(x_{w}(J)) + |\xi| > r \rfloor$.

$k$-uniform on data: $Pr[\forall \xi \in (r - \gamma_{\text{sm}}, r + \gamma_{\text{sm}}), \forall b \in \{0, 1\}^w, |Pr[f_{j}(G(x_{w}(J)) = b \mod 2 | \Xi = \xi] - \frac{1}{2^{|w|}} | = 2s_{\gamma_{\text{sm}}} \cdot 2k \cdot \gamma \leq O(\delta_{\ell,11})$. 

$Pr[\forall \xi, \forall b, |Pr[f_{j}(G(x_{w}(J)) = b \mod 2 | \Xi = \xi] - \frac{1}{2^{|w|}} | = 2s_{\gamma_{\text{sm}}} \cdot 2k \cdot \gamma \leq O(\delta_{\ell,11})$ by $\gamma = \frac{e\beta}{(s_{\gamma_{\text{sm}}})^2}$ for $k_{\ell,17}$ independent flipper (\*: $m_{\ell,17} \gg \frac{2k(s_{\gamma_{\text{sm}}})^5}{\beta^2 \gamma_{\text{sm}}^2}$ and $m_{\ell',17} \gg \frac{(s_{\gamma_{\text{sm}}})^2}{\beta}(2k^2)$).

Let $\mathcal{C} := \{(x, y) \in \mathcal{D}_{w,1_w} | y \in \mathbb{Z}_{2r+1}\}$ 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 $\mathcal{C} \cap (\mathcal{X} \cap \mathcal{D}_\ell)$ and apply Theorem 7.12 to $\mathcal{C} \cap \mathcal{D}_\ell$. We call a dataset $\mathcal{D}_{\ell}$ fully colliding when $\mathcal{D}_{\ell} \subset \mathcal{C}$.

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

Sufficiently many examples: $\forall(w, 1_w) \in \mathcal{W}_{w,i_0,i_1}, m_{\ell,17} \ll |\mathcal{D}_{\ell}| \ll 2m \cdot \alpha \mu/k/p \ll p^\ell/r$.

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

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

53
It runs Algorithm 3’s Step 2 over $\mathbb{Z}$ rather than $\mathbb{Z}_{2^q}$. The smallness bounds $|h^{\text{hash}} - \Xi| \leq s\gamma_w$. Also, Theorem 7.12’s small-hash and correlation-on-data hold at $(w, 1_w) \in W_{w,i_0,t_1}$. Algorithm 3 can uniquely identify $\theta_w(1_w)$ of every $(w, 1_w) \in W_{w,i_0,t_1}$. Algorithm 4 can invert the hidden $\theta$ from the coefficients $\theta_w(1_w)$ via Theorem 7.15’s modulus amplification

\[
\psi_\ell(q_\ell(x) + \xi) = \sum_{y \in \{y(j)\}_j} y\varphi_\ell\left(\prod_{y' \in \{y(j)\}_j} \frac{q_\ell(x) + \xi' - y'}{y - y'}\right) \equiv q_\ell(x) + \xi \mod \mathbb{Z}_{2^q}.
\]

**Inverting partially-colliding data:** The partially-colliding $D_\Xi \not\subset \mathcal{C}$ 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 $h^{\text{hash}}_{r-\ell}(\{0,1\}^w) \cap \mathbb{Z}_{2r+1}$ and $2r+1 + h^{\text{hash}}_{r-\ell}(\{0,1\}^w) \cap \mathbb{Z}_{2r+1}$, where $2r + 1 = (r - \ell) - (-(r + \ell + 1))$ to make $h^{\text{hash}}_{r-\ell}(\{0,1\}^w)$ fully colliding. The following correlation analysis justifies it due to the $k$-uniform on data and smoothness in $\xi$, and $s\gamma_w \ll r$ in $\ll$:

\[
\mathbb{E}[h^{\text{hash}}_w(G(1_w)) \cdot \prod_{i \in w}(-1)^{G_i(1)} \cdot 1[h^{\text{hash}}_w(G(1_w)) \in \mathbb{Z}_{2r+1}, h^{\text{hash}}_w(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}]]
\]

\[
= \sum_{\ell,\alpha=0}^{s\gamma_w} \left( \mathbb{E}[h^{\text{hash}}_w(G(1_w)) \cdot \prod_{i \in w}(-1)^{G_i(1)} \cdot 1\left[h^{\text{hash}}_w(G(1_w)) \in \mathbb{Z}_{2r+1}, h^{\text{hash}}_w(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^{\ell}(r - \ell)\right] + \right.
\]

\[
\left. \mathbb{E}[h^{\text{hash}}_w(G(1_w)) \cdot \prod_{i \in w}(-1)^{G_i(1)} \cdot 1\left[h^{\text{hash}}_w(G(1_w)) \in \mathbb{Z}_{2r+1}, h^{\text{hash}}_w(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^{\ell+1}(r + \ell + 1)\right] \right)
\]

\[
= \sum_{\ell,\alpha=0}^{s\gamma_w} \mathbb{E}[h^{\text{hash}}_w(G(1_w)) \cdot \prod_{i \in w}(-1)^{G_i(1)} \cdot 1[h^{\text{hash}}_w(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^{\ell+1}(r + \ell + 1)]
\]

\[
+ \sum_{\ell,\alpha=0}^{s\gamma_w} \left( \mathbb{E}[h^{\text{hash}}_w(G(1_w)) \cdot \prod_{i \in w}(-1)^{G_i(1)} \cdot \left[h^{\text{hash}}_w(G(1_w)) \in \mathbb{Z}_{2r+1}, h^{\text{hash}}_w(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^{\ell+1}(r + \ell + 1)\right] \right)
\]

\[
\leq \sum_{\ell,\alpha} \mathbb{E}[h^{\text{hash}}_w(G(1_w)) \cdot \prod_{i \in w}(-1)^{G_i(1)} \cdot 1[h^{\text{hash}}_w(\{0,1\}^w) \not\subset \mathbb{Z}_{2r+1}, \Xi = (-1)^{\ell+1}(r + \ell + 1)] + \beta,
\]

\[
|\beta| \leq \sum_{\ell,\alpha} (2r + 1) \cdot \frac{2^k \cdot \varepsilon \beta}{2^k(s\gamma_{w})^2} \cdot \frac{\Pr[\Xi \in [-r - s\gamma_{w}, r - s\gamma_{w}]]}{\Pr[\Xi \in [-r - s\gamma_{w}, r - s\gamma_{w}]]} \leq \frac{(2r + 1)\varepsilon \beta}{2^k(s\gamma_{w})^2} + \frac{2s\gamma_{w}}{2} \ll \beta.
\]

**Inverting the actual data:** Since the statistical distance between the ideal shift $G$ and the actual one over $D_\Xi$ is bounded by $\beta \delta_{7,11}/r$, Theorem 7.12’s $(\alpha, \beta)$-inversion has demonstrated $|\text{corr}(D_\Xi - \theta_w(1_w))| \ll \beta$ under the full collision $D_\Xi \subset \mathcal{C}$ with significance $O(\delta_{7,11})$. The partial one on the actual hash may add an extra accuracy cost $\varepsilon \beta$ to derive $|\text{corr}(G \cap D_\Xi - \theta_w(1_w))| \ll \beta$, since the actual one may deviate from the ideal $h^{\text{hash}}_w(G_w)$ by statistical distance $O(\beta \delta_{7,11}/r)$.

**Inverting the almost-zero secret:** When the small secret parameter $\theta \in \mathbb{Z}_n^*$ is virtually zero as $|\{i \in (d) | \theta_i \neq 0\}| \leq k$, add $1_w$ to $w \subset \{i : \theta_i = 0\}$, and replace each data $(x(j), y(j))$ with $x(j), y(j) + \sum_{i \in w} f_i x(i)/2 \cdot (-1)^{s(i)}$ for inverting $\theta + 1_w$.

**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 GapSVP $O(n/\alpha)$ is reducible to LWE$_{n,q,\alpha}$. 

54
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, $\text{LWE}_{n,q,\alpha}$ reduces to decision $\text{LWE}_{n,q,\alpha'}$ 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 $0 < \varepsilon, \xi < 1$ be $n' \geq (n + 1) \log q + 2 \log(1/\delta)$, $\alpha \geq \sqrt{\ln(2n(1 + 1/\varepsilon))/\pi q}$, $\beta = (10n' \alpha^2 + 4n' \pi / 2 \ln(2n'(1 + 1/\xi)))^{1/2}$. As decision problems, $\text{LWE}_{n,m,q,\alpha}$ is reducible to $\text{LWE}_{n',m,q,\beta}$ with the binary secret such that any $\zeta$-advantageous algorithm of the latter problem produces that of the former one with an advantage $\delta - \frac{4\xi}{2} - 14\xi$.

Theorem 7.21 (Theorem 1.7). 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 $\beta = (10n' \alpha^2 + 4n' \pi / 2 \ln(2n'(1 + 1/\xi)))^{1/2}$. Theorem 7.18 reduces GapSVP$\tilde{O}(n/\alpha)$ to $\text{LWE}_{n,q,\alpha}$, Theorem 7.19 reduces it to decision-$\text{LWE}_{n,q,\alpha'}$, and Theorem 7.20 to decision-$\text{LWE}_{n',p,\beta}$ with the binary secret. So, Theorem 7.17 inverts (search) $\text{LWE}_{n',p,\beta}$ in poly-time. \hfill \Box

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 $F$-complexity of arguing circuit classes $\mathcal{F}$ in a smoothed analysis. Let $G \in \{0, 1\}^N$ be any $\beta \delta_{11}$-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 $|\mathcal{M}|_{\not=0}/N$ 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 $(x, y) \in \mathcal{I} \times (N) \cup [N] \times \mathcal{J}$ and predicts

$$
\varepsilon\text{-learning: } \Pr_{G}[\Pr_{X,Y}[\mathcal{A}(X, Y) \mid \tilde{\mathcal{M}}(\mathcal{I} \times (N) \cup [N] \times \mathcal{J}) \neq \tilde{\mathcal{M}}(X, Y)] \leq \varepsilon] \geq 1 - O(\delta).
$$

8.1 Unrestricted Super-Linear Lower Bounds

An $F$-linear circuit is an $n$-input $n$-output circuit computing an $F$-linear form $f = \sum_i f_i g_i$ at each gate $f$ feeding the in-coming edges labeled by $f_i \in F$ 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 $F$-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 $g_i \cup \{g_i'\}_{i \in [s] \setminus \{i\}} \Rightarrow o \cup \{g_i'\}_{i \in [s] \setminus \{i\}}$. The $g_i$ might be an input (fan-in 0) node of fan-out one, say $x_i$. There is no more than one leaf node having one fan-out due to the matrix’s reversibility. This case connects $x_i \rightarrow o$ 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 $g_i \cup \{o' \mid o' \neq o\}$ to form an $n$ by $n$ reversible matrix. \hfill \Box
Definition 8.3 (matrix rigidity). The rigidity $\text{rig}_M(r)$ of a matrix $M \in \mathbb{F}^{n \times n}$ is the minimum number of flipping entries on each row to reduce its rank to $r$:

$$
\text{Matrix rigidity: } \text{rig}_M(r) = \min \left\{ \max_x |\mathcal{N}_x| \neq 0 \mid \text{rank}(M + \mathcal{N}) \leq r \right\}.
$$

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

Proof. Let $C = (\mathcal{V}, \mathcal{E})$ be an arbitrary binary circuit of node-depth $\psi : \mathcal{V} \rightarrow [d]$. 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 within the remaining $(1 - 2^{-t}) d$ patterns. Repeating it $u$ times reduces it to $(1 - 2^{-t}) u d$ of any path through the cut edges.

Theorem 8.5 (formulas to partial derivatives [BS83]). Any algebraic fanin-2 circuit of size $s$ and depth $d$ to compute a linear $y$-degree polynomial $f(x_1, \ldots, x_m, y_1, \ldots, y_n) = \sum_{i=1}^n \frac{\partial f}{\partial y_i} \cdot y_i$ induces a multi-output parallel algebraic circuit of size $2s$ and depth $2d$ computing all partial derivatives $\left( \frac{\partial f}{\partial y_i}(x) \right)_{i=1}^n$.

Theorem 8.6 (Theorem 1.55). Let $N = n^2 = 2^{\log k} k \log(N/k)$ by even integers $k$ and $\log N/k$. Let $p$ be an odd prime coprime with $\left(2^{(\ell-1)}, (r_0, v, |\mathcal{Q}_{\text{rem}}|) = (1, 3, p), k = 3(2^\ell - 1), \alpha = k!/k^k, \beta < 1, d = \Theta(\log n), d_t = d/2^\ell, d_{t,u} = d(1 - 2^{-t}), r \leq stu/\log d, \delta_{11} = \delta/r, \eta \approx 2^{d_{t,u}} + d_t/n$. Suppose $m_{12} \ll \min(\mu, \eta n, k) \ll k_0 \approx 4^{d_{t,u}} \mu$, and $2^{\mu k}/\beta \ll \delta_{11}$. Any $n$ by $n \{-1, 0, 1\}$-matrix $M$ of density $\mu$ must refute any $\mathbb{F}$-linear circuit of size $s$ and depth $d$ computing $\hat{M}$ unless each row of $\hat{M}$ is $\eta$-learnable from some $4^{d_{t,u}}$ of the first $r$ rows, and the first $m = O(m_{12}/(\alpha \mu k))$ columns, in $O\left(\left(\frac{k_0}{k} + k \delta r\right) m\right)$ time.

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

\[\text{Take } k \ll k_0 \leq O(1), \alpha = \frac{\sqrt{\log n}}{\sqrt{\log \log n}, t = \frac{t}{\log \log n}, u = (\log \log n)^2, p \gg n^{6/k}, r = O\left(\frac{n}{\log \log n}\right), s = n(\log \log n)^{1-\varepsilon}, \text{ and } \mu \approx \frac{k_0}{1 + u}. \]  

Lemma 2.13 provides an explicit $O(\left(2^{-2^{d_{t,u}} + d_t}\right)$-away 2$k_0$-independent flipper $|G| = O\left(\left(2^{4^{d_{t,u}} + d_t}\right) \log n\right)^2$. Lemma 2.17 gives an explicit DFT-shift $|\Phi| = n^{O(1)}$.

154\text{1}_{r \times r}$ is the $r$ by $r$ identity matrix.
Matrix rigidity: \((AA^{-1})^2\tilde{M} = \tilde{M} + \mathcal{N}\). Let \(I_{x_0} = \{x \in I \mid \theta_{x_0}(x) \neq 0\} \subseteq I\).

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

**Planted linear FT:** Theorem 7.15 can invert this \(f_{\theta_{x_0}}(x)\) on \(|I_{x_0}| \leq 4d_{\mu}, k_0 \approx \mu|I_{x_0}|, \mu_{\mathrm{hash}} \approx \mu^k\), and \(m \ll n\). Lemma 2.17’s DFT-shift makes the i.i.d. \(m\) samples \(J \subset \{n\}\) to have:

\[
\begin{align*}
\text{Uniform density:} & \quad \forall x_0 \in \{n\}, \Pr_{Y \in \mathcal{J}}[|I_{x_0} \cap \tilde{M}_Y| - \mu|I_{x_0}|] \ll \mu|I_{x_0}| \approx 1. \\
\text{k-cover:} & \quad \forall x_0 \in \{n\}, \forall K \in \binom{I_{x_0}}{k}, |\Pr_{Y \in \mathcal{J}}[(K,Y) \subset \tilde{M}_Y]| \ll \mu_{\mathrm{err}} \approx \mu_{\mathrm{err}}. \\
\text{Column-wise error:} & \quad \forall x_0 \in \{n\}, \Pr_{Y \in \mathcal{J}}[x \in I_{x_0} \Rightarrow \mathcal{N}(X,Y) = 0] \geq 1 - \eta \mu/\delta \cdot |I_{x_0}|. (\because \text{Markov’s ineq of } \gamma = \delta).
\end{align*}
\]

Chernoff bounds of \(\gamma = \frac{\epsilon}{1 - \epsilon}\) guarantees them with significance \(n \cdot \left(\frac{n}{\epsilon}\right) \cdot e^{-\gamma^2/2(1-\epsilon)m} < o\left(\frac{1}{n}\right)\).

\(G\)’s \(\beta_{d_0,1}\)-away \(2k_0\)-independence implies its (\(h_{\mathrm{hash}}^\mathbb{GF}(G), 0\))-hashed \(\beta_{d_0,1}\)-away \(2k\)-independence.

**Theorem 7.15** inverses\(^{56}\) the \(f(x)\) and predicts \(M_{x_0}\) from the entries over \(I_{x_0} \times \{n\} \cup \{n\} \times \mathcal{J}\). \(\square\)

### 8.2 Lower bounds beyond \(\phi_{\mathbb{GF}}\)

Tarui [BFS86, Tod91, Tar93] presented low-degree probabilistic polynomials to approximate \(\phi_{\mathbb{GF}}\) 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 \(d = \sum_{k=0}^{h} d_k, d_{s,7} := 2d_{e} + h\) and \(s_{\mu,7} = \prod_{k=1}^{h} (1 + 2d_k)^{2d_{e}+1}\). Suppose \(\mathcal{L} \in \phi_{\mathbb{GF}}[d]\) has the same type of gates at depths \((d_{s-1}, d_h)\). It admits a low-degree linear computation \(\forall (x,y) \in \{0,1\}^{n/2} \times \{0,1\}^{n/2}, \Pr_R[\mathcal{L}(x,y) \neq \phi_R(x,y)] \approx 0\) under the random seed \(R \in \{0,1\}^{e \sum_{k=1}^{h} 2d_k}\).

**Linear expression by lift and project:** \(\phi_R(x,y) = \sum_{w \in \binom{\mathbb{GF}}{m}} \hat{\phi}_{w,R}(1)_{x \in I_w, y \in J_w} \) by \(I_w, J_w \subset \{0,1\}^{n/2}\), and \(\hat{\phi}_{w,R} \in \mathbb{Z}\) with \(\sum_{w \in \binom{\mathbb{GF}}{m}} |\hat{\phi}_{w,R}| \leq s_{\mu,7}^2\).

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

**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 \(\phi_{\mathbb{GF}}[d]\) computation with Tarui’s probabilistic polynomials [Tar93]. At the \(k\)th layer of \(\phi_{\mathbb{GF}}[d]\), it uses \(e \sum_{k=1}^{h} 2d_k\) number of the i.i.d. coin flips \(R_{i,j,t} \in \{0,1\}\) of bias \(E[R_{i,j,t}] = 1/2^j\) and transforms a depth-\(d_e\) NOR-subtree to:

\[
\text{Probabilistic polynomial: } \phi_R(g_1, \ldots, g_{2d_e}) = \left(1 + \sum_{i=1}^{2d_e} g_i \right) \prod_{t=1}^{e} \prod_{j=1}^{d_e} (1 - \sum_{i=1}^{2d_e} R_{i,j,t} g_i)^2.
\]

It satisfies the Boolean guarantee, i.e., \(\phi_R(g_1, \ldots, g_{2d_e}) = 1 \Rightarrow (1 + \sum_{i=2d_e} g_i) = 1 \Rightarrow \forall g_i = 1\) and \(\phi_R(g_1, \ldots, g_{2d_e}) = 0 \Rightarrow (1 - \sum_{i=1}^{2d_e} R_{i,j,t} g_i) = 0 \Rightarrow \forall g_i = 1\). 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 \(2d_{e}+1\)-degree and \((1 + 2d_{e})2d_{e}+1\)-norm polynomials at the \(k\)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\textsuperscript{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 $(2^{(t-1)}, (r_0, v, |Q_{\text{real}}|) = (1, 3, p), k = 3(2\ell - 1)$, $\alpha = k! / k^k$, $\beta < 1$, $r \leq (\binom{\alpha}{r_0})$, $\delta_{11} = \delta / r$, $k / k_0 = r \mu$, and $\eta = \frac{k_0 r}{2\mu}$. Suppose $m_{\gamma_1} \leq \min(p^\ell, \alpha^\mu k \sqrt{n})$, and $\beta k^\ell / \beta \leq \delta_{11} \leq \beta$. Any $\sqrt{n}$ by $\sqrt{n} \{1, 0, 1\}$-matrix $M$ of density $\mu$ must have lower bounds $\tilde{\eta}$ than $1 - \phi$.

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

$$M_{\text{rigidity}} (AA^{-1})^T (M_T + N_T) = \tilde{M} + N \quad \text{by } A \in \mathbb{F}_{p}^\times$$

$$\text{Point-wise error: } \forall (x, y), \Pr[N(x, y) \not= 0] \leq \Pr_r[\exists b, (\phi_{b, R}(x, y) \not= \{0, 1\})] \leq 2/2^\epsilon_0.$$

$$\text{Column-wise error: } \Pr_y [\forall x \in I, (M(x, y) \not= 0 \Rightarrow N(x, y) = 0) \sum_{x \in I \setminus M_{\text{rigidity}}} (\tilde{M} + N)(x, y) \mod p] \geq 1 - 2^{k_0} / 2^{\ell^3}.$$

Theorem 8.6 has succeeded in learning $\tilde{M}$ from the first $m$-columns of $(\tilde{M}_T + N_T)$ satisfying the uniform-density and $k$-cover under a fixed reminder $\sum_{x \in I \setminus M_{\text{rigidity}}} (\tilde{M} + N)(x, y) = \xi \in \mathbb{Z}_p$.

Markov’s inequality parameter $\gamma = \delta$ provides $m\gamma / p = O(\frac{m_{\gamma_1}}{\epsilon_0 \gamma_0})$ data enough to execute it. 

\[ \Box \]

8.3 \text{PH}_{\text{cc}} \neq \text{PSPACE}_{\text{cc}} \text{ or quasi-NC}^k

As mentioned in Theorem Review\textsuperscript{11}, Williams’ algorithmic approach [Wil\textsuperscript{13}] has established $\text{NEXP} \not\subseteq \text{ACC}$ [Wil\textsuperscript{14a}] and even $\text{quasi-NC} \not\subseteq \text{ACC}$ [MW\textsuperscript{19}]. This section will further extend it to prove Theorem 1.10 in virtue of Theorem 7.15’s linear Fourier inversion.

Theorem 8.9 (NP hierarchy [Coo\textsuperscript{72}, SFM\textsuperscript{78}, Žák\textsuperscript{83}]). There is a unary language $L \subset \{1\}^*$ separating $L \in \text{NTIME}[t] - \bigcup_{\ell^s} \text{NTIME}[t^\ell(n + 1) = o(t(n))]$.

Theorem 8.10 (short PCP [BSV\textsuperscript{14}]). Every $t$-time verifier algorithm $A(x, w)$ inputting $1^n$ and a witness $w$ can induce an $(n + \log t)O(1)$-time computable generator $1^n \rightarrow C_\nu^n$ of poly$(n + \log t)$ size circuit with an oracle $O_n$ of $n = \log t + O(\log \log t)$ input bits. If $\exists w, A(1^n, w) = \text{accept}$ then $\exists O, C_\nu^n$ is unsatisfiable. If $\forall w, A(1^n, w) = \text{reject}$, then $\forall O, C_\nu^n$ has at least $(1 - \frac{1}{n})$ satisfying assignments. $C_\nu$ is a $3\text{CNF}$ of the $O(n^1)$ inputs of the $\Omega$’s answers.

Theorem 8.11 (easy witness lemma [MW\textsuperscript{19}]). There is a universal constant $c_{\text{en}} > 0$ such that if $\text{NTIME}[t_{\text{en}}] \subset \text{SIZE}[n^\ell]$, then every language in $\text{NTIME}[t]$ must have a witness of $\text{SIZE}[n^{\Omega(n^1)}]$.

Theorem 8.12. $\exists e_0, \forall h, \forall \ell, \forall d \ll n/e_0, P \not\in \text{PH}_{\text{cc}}^c[d]$ or $\text{NTIME}[2^{\text{en}}] \not\subseteq \text{SIZE}[n^\ell].$

Proof. Let $N = t = O(2^{\text{degh}/c_{\text{en}}}), n = \log t + O(\log \log t), r \leq (\frac{n}{2e_0 + h}), \mu \approx k_0 / r$, and $p^e = N^{O(1)}$. Suppose $\text{NTIME}[t_{\text{en}}] \subset \text{SIZE}[n^\ell]$ and $P \not\in \text{PH}_{\text{cc}}^c[d]$ for a contradiction.

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

\textsuperscript{57}Take $\frac{\epsilon_0 k_0}{\eta^2 \delta} = O(1)$ and $d = n^\ell$. 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 $N^{O(1)}$. 

58
Our assumption expression \( \phi \) circuit \( M \) entries. Further, we reduce \( n \) the row data \( \tilde{\phi} \) PCP’s so that \( 1 \) Theorem 8.8 has shown it to meet the small column-wise error, too. Let \( \Phi \) parameters \( k \) is at least \( \mu / \) nant CMD Definition 8.14 (CMD and DCMD 58 Theorem 8.13 59 \( \tilde{\phi} \) \( \rho \) 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 \( \tilde{\mathcal{M}}' := \mathcal{M} + \mathcal{N} \). It induces Theorem 8.6’s planted FT \( f(x) \) that Theorem 7.15 can invert from the row data \( \tilde{\mathcal{M}}_d \) under a guessed flipper \( G \). Let \( \mathcal{I} \in \binom{\mathcal{I}}{n^2} \) be Theorem 8.8’s matrix-rigidity index set. Guess a permute \( \tilde{\phi} \) to satisfy

\[
\text{uniform density: } \Pr_Y \left[ \left| \{ x \in \mathcal{I} \mid \tilde{\mathcal{M}}(x,Y) \neq 0 \} \right| - \mu s \right] \leq \mu \varepsilon \approx 1.
\]

\[
\text{k-cover: } \forall \mathcal{K} \in \left( \mathcal{I} \right), \left| \Pr_{\mathcal{K}} \left[ \{ \mathcal{K} \mid \tilde{\mathcal{M}}(x,Y) \neq 0 \} \right] - \mu_{\mathcal{K}} \right| \leq \mu_{\mathcal{K}}.
\]

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 \( x \in \mathcal{V} \) and \( a \in \{-1,0,1\}^{k_0} \),

\[
\text{Acceptance probability estimation: } \left| \frac{1}{N} \sum_{x \in \mathcal{I}(y),a} \mathbb{E}_{U,G} \left[ \{ y \mid \tilde{\mathcal{M}}(\mathcal{I}(y),y) = a \} \cdot 1[f(x,y) \neq 0] \right] \right| \geq 1 - O(\delta).
\]

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 \( |\mathcal{N}|_{\neq 0} / N \leq 1 / 2^{\alpha_0} \). It recognizes \( \mathcal{L} \) by accepting \( n \) if the result is at least \( \mu / 2 \), and rejecting it otherwise. It runs in the following non-deterministic time of the parameters \( kn / 2 \) \( O(1) \), \( n = O^{\log^c n} \) and \( t = 2^n \), contradicting \( \mathcal{L} \notin \text{NTIME}[\mu^{1-o(1)}] \):

\[
|\mathcal{U}| \times (\text{Matrix entries calculation time + Fourier inversion time + Acceptance probability estimation time})
\]

\[
= \left| \left[ 1 / \mu \right] \cdot \left( \sqrt{n} |\mathcal{I}| \cdot \tilde{O}(n) + |\mathcal{X}| \cdot \left( \begin{bmatrix} k_0 \\ k \end{bmatrix} + k_0 r \right) \cdot \frac{\mu_{\mathcal{X} \mathcal{I}}}{nr^2} + |\mathcal{X}| \cdot |\mathcal{I}(\mathcal{X})| \times \{-1,0,1\}^{k_0} \cdot \tilde{O}(p) \right) \right|
\]

\[
\leq \frac{r}{k_0} \cdot \left( \sqrt{n} \cdot r \cdot \tilde{O}(n) + \sqrt{n} \cdot \tilde{O}(r) \cdot \frac{O(r^3)}{\alpha(k_0)^2} + \sqrt{n} \cdot \left( \frac{r}{k_0} \right)^3 \tilde{O}(p) \right) \ll t^{1-o(1)}.
\]

\[\square\]

**Theorem 8.13** (Barrington’s theorem [Bar89]). Any depth-\( d \) circuit admits a permutation branching program\(^{58} \) of width \( 5 \) and length \( 4^d \).

**Definition 8.14** (CMD and DCMD\(^{59} \)). \( \text{CMD}_{n(n+1)/2}(\mathcal{M}) = \text{det}(\mathcal{M}) \mod 2 \) 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 \( \{(x_1, f_1, g_1), \ldots, (x_\ell, f_\ell, g_\ell)\} \). An input \( x \in \{0,1\}^n \) guides the branches to select and compose the \( \ell \) permutations \( h_0 \circ \cdots \circ h_\ell \) by \( h_i = f_i \) if \( x_{i+1} = 1 \) and \( h_i = g_i \) otherwise.

\(^{59} \) CMD: Connected Matrix Determinant. DCMD: Decomposed CMD.
with $i \geq j + 2 \Rightarrow M(i, j) = 0$. DCMD$_{n^3(n+1)/2}(M_k, 1 \leq k \leq n^2) = \text{CMD}(\bigoplus_k M_k \mod 2)$ for connected $M_k$. In particular, both CMD and DCMD belong to $\bigoplus L \subset \bigoplus \text{P}^{cc} \subset \text{PSPACE}^{cc}$.

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

**Definition 8.16** (approximate sum). We say that a function $f$ admits a Sum$_x \circ F$ circuit if there are functions $C_i \in F$ and coefficients $\alpha_i \in \mathbb{R}$ approximate $\forall x, |f(x) - \sum_i \alpha_i C_i(x)| \leq \varepsilon$. 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 $F$ can $(1/2 + \eta)$-approximate DCMD$_{n^3(n+1)/2}$, Sum$_c \circ F$ can compute CMD$_n(n+1)/2$ by $O\left(\frac{n}{\exp(\eta)}\right)$ circuits in $F$ with the sum of absolute coefficients $O(1/\eta)$.

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

**Theorem 8.19** (Theorem 1.10). Suppose PH$^{cc}$ either computes CMD or approximates DCMD by advantage $\frac{1}{2} + \frac{1}{\exp(n^\eta)}$. Then DEP$[k \log n]$ cannot $(\frac{1}{2} + \frac{1}{\exp(n^\eta)})$-approximate NP for all $k \geq 1$, i.e., some NP language $L$ cannot have DEP$[k \log n]$ circuits $C_n$ of advantage $\Pr_U[L(U) = C_n(U)] \geq \frac{1}{2} + \frac{1}{\exp(n^\eta)}$ 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 NP admitted $(\frac{1}{2} + \frac{1}{\exp(n^\eta)})$-approximation by DEP$[k \log n]$ circuits. Take Theorem 8.12’s parameters but $\varepsilon = o\left(\frac{1}{\exp(n^\eta)}\right)$, $d = cn$, $d' = 2(ch + 2)d + 3 \log n + 2 \log \frac{1}{\varepsilon} + n^{o(1)}$, and $d'' = 2ed + h + 3 \ll n$.

**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 CMD$_{n(n+1)/2}(p(x))$ by a projection mapping $p(x)$ of $m = 5! \cdot 4^d$. By assumption, PH$^{cc}[c \log m]$ must compute CMD$_{n(n+1)/2}$ or $(\frac{1}{2} + \frac{1}{\exp(m^\eta)})$-approximate DCMD$_{n^3(n+1)/2}$, $\zeta = o(1)$, so PH$^{cc}[c \log m]$ must contain CMD$_{m(m+1)/2}$ or $(\frac{1}{2} + \frac{1}{\exp(m^\eta)})$-approximate DCMD$_{m^3(m+1)/2}$. In the latter case, Theorem 8.17 writes CMD$_{m(m+1)/2} \in$ Sum$_c \circ$ PH$^{cc}[c \log m]$ by a linear combination of $(\frac{m}{\eta})^2 \exp(n^\varepsilon)$ PH$^{cc}$-circuits with the weight $\exp(n^\varepsilon)$. Let us derive a contradiction from the latter case since the former is easier to do it (by avoiding Sum$_c$ computation).

**Learning shallow circuits:** Let $V(x, y)$ have Theorem 8.10’s short PCP’s witness circuit $C^{NW}_y$. 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 $(\frac{m}{\eta})^2 \exp(n^\varepsilon)$ Sum$_c$ gate at the third, and fan-in $2^kd$ AND or OR gates at the remaining $h$ layers. Theorem 8.7 transfers it to a probabilistic polynomial of degree $d'' := 2cd + h + 3$ for $d_h = \ldots = d_1 = c \log m$, and $d'' = \sum_{r=1}^{h+3} d_r$. The probabilistic polynomial NOR($g_1, \ldots, g_{m^3}$) of the top AND gate may contain an additional error term $O(n^2 \exp(n^\varepsilon)) = o(1)$ since each $g_i$ is OR of 3 Sum$_c$ gates having an error term $\varepsilon$ and the weight $\exp(n^\varepsilon)$. Theorem 8.12’s acceptance probability estimation recognizes $\{1^n \mid \exists y, V(1^n, y) = 1\} \in$ NTIME$[\tilde{t}] \setminus$ NTIME$[t^{1-\alpha(1)}]$ in NTIME$[t^{1-\alpha(1)}]$ time, a contradiction.
9 Natural Lower Bounds for NP $\not\subseteq$ TC$^1$ and VP $\not= VNP$

In this section, in the worst-case analysis (H$_w(G) = 0$), we translate number-theoretic/algebraic structures of TC$^0$ and VP circuits into data-compressing exact learning algorithms in Lemmas 1.31 and 1.32. These learning algorithms 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 $\not\subseteq$ quasi-TC$^0$

Let us briefly explain a number-theoretic mechanics to simulate TC$^0$ by 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 $\sum g(x) = a \iff \sum \text{MOD}_{p_i}(\sum g(x) - a) = 0 \iff \sum_{k=0}^{k_0} (\sum \text{MOD}_{p_i}(\sum g(x) - a))(1/q_t) = a_t$ by $k = O(1)$, distinct primes $p_i \leq \ln a$, and distinct base points $q_t = t + O(1)$. Vandermonde 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 SYM$^+$ computation.

**Lemma 9.1** (modulus lifting [BT94]). For any multi-linear polynomial $f \in \mathbb{Q}[x_1, \ldots, x_n]$ of $2\text{norm}(f) + 1 \leq m^\ell$, and any integers $a_t \in \{0, 1\} + m\mathbb{Z}$,

$$f(a_1 \text{ mod } m, \ldots, a_n \text{ mod } m) = f(\phi(\phi(a_1), \ldots, \phi(a_n)) \text{ mod } m^\ell.$$

**Lemma 9.2** (Vandermonde’s kernel [PR07]). The kernel of a generalized Vandemond matrix $M_{t,n} = (a_{i,j}^t)(t,j)\in[t]\times[n]$ of distinct numbers $a_{i,j}$ has dimension $n - t$ and admits a basis spanned by the cyclic shifts $v_0, \ldots, v_{n-t}$ of the following kinds. Let $\sigma(i) = \sum_{1 \leq i_1 < \cdots < i_t \leq n} a_{i_1}a_{i_2}\cdots a_{i_t}$.

$$v_k = \frac{(0, \ldots, 0, (-1)^t \sigma(t), \ldots, (-1)^t \sigma(i), \ldots, -\sigma(1), 1, 0, \ldots, 0)}{n-t-k}.$$

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

**Definition 9.4** (ACC circuits). Let $2 = p_1 < p_2 < \cdots$ be the smallest prime numbers. An SYM$_{m,q,t}$ gate is $t$-tuple set $f \subset \mathbb{N}/q$ to express $f = 1[\sum_{g\in \text{in}(f)} \hat{g} \in \hat{f}]$, i.e., each input $g$ of $f$ associates a $t$-tuple number $\hat{g} = (\hat{g}_t)_{t=1}^\ell \in \mathbb{N}/q$ bounded by $\sum_g \sum_{t \leq \lambda} |\hat{g}_t| \leq m/q$. A depth-(2$h+1$) circuit SYM$\circ$ACC$_h = \text{SYM}_{m,q,t} \circ (\text{AND}_{k_d} \circ \{\text{MOD}[p_1], \ldots, \text{MOD}[p_{s_d}]\})$ consists of these SYM$_{k_d,q,t}$ gates at the top, AND gates of fanin $k_d$ at each depth $2d$, and MOD gates of modulus $p_{\xi(\lambda_d)} \in \{p_1, \ldots, p_{s_d}\}$ of some $\xi \in Q_{dh} := \prod_{d=1}^h (s_d)^{\lambda_d}$ of $\Lambda_{dh} := \prod_{d=0}^h k_d$. In this SYM$\circ$ACC$_h[\xi]$ circuit, the AND gates at depth $2d$ must take the moduli $\text{AND}(\text{MOD}[p_{\xi(1)(d+1)d}], \ldots, \text{MOD}[p_{\xi(\lambda_d, \lambda_d+1)(h)}])$ along with a path $\lambda$ of depths from $2h$ down to $2d$.

**Lemma 9.5** (from AC$^0[\text{SYM}]$ via SYM$\circ$ACC to SYM$^+$ (Lemma 1.31)). Given increasing positive integers $h \ll \lambda_h \leq \cdots \leq \lambda_1 \ll 2^{h/h}$, and $k_d, \ell_d, m_d, n_d, q_d, s_d$ and $t_d$ as follows$^{60}$.

Let $k_d = O(1)$, $m_d = \Delta_{d+O(d)}^s \Delta_{d+O(d)}^t$, $n_d = O(\Delta_{d+1}^s)$, $s_d = O(\Delta_{d+1})$, $t_d = O(\Delta_{d+1})$, $\tilde{p_d} = \prod_{i=1}^{s_d} (1_i \approx s_d^{s_d})$, $\tilde{k_d} = \prod_{d'=d}^h k_{d'}q_{d'} - t = q_{d+1} = O(\Delta_{d+1}^s), \ell_d = \Delta_{d+1}^s \Delta_{d+1}^s s_d + i \Delta_{d+1}^s$, for $\Delta \gg s_{d+1}$.

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

61
with
Truncated Taylor analysis transfers any SYM

We can verify

guarantee

= \sum_i \sum_{x_i \in X} f(x_i) = e_x \star \star \star

\begin{align*}
\text{Truncated Taylor series} & : \hat{f}_t(x) := \sum_{k=0}^{k_d} \binom{j(x)}{k} \frac{1}{q_a^k} k_d (1 + \varepsilon_{td}(j)) \\
\text{Hash function} & : \hat{f}(j) := \hat{f}_t(x) : j \Rightarrow j(x).
\end{align*}

\begin{align*}
\text{Taylor series approximation} & : \hat{f}_t(j) := \sum_{k=0}^{k_d} \binom{j}{k} \frac{1}{q_a^k} k_d (1 + \varepsilon_{td}(j)) \\
\text{Colliding numbers} & : y_j(x) := \{ f \in F : \sum_{x_i \in X} \hat{f}_t(x) \equiv f(x) \} \text{ for a given } F \text{ of size } |F| \leq e^{\Delta d + 1},
\end{align*}

No

= 1 + 1/q_d, \exists \alpha_j \in \mathbb{R},

\begin{align*}
\text{By Lemma 9.2 of } a_t & = 1 + 1/q_d, \exists \alpha_j \in \mathbb{R}, \\
\sum_{j=0}^{n_d-t} a_j y_j & = (y_j(x) - y_j(x')) \cdot (1 - \varepsilon_{td}(j)) j \equiv (N^{n_d} \setminus N^a)(1 + \varepsilon) \\
\Rightarrow \beta_j & := \alpha_j \sigma(t) \text{ has a norm } \|\beta_j\| \geq \frac{1}{\sqrt{n_d}} \text{ since Lemma 9.2's triangular matrix } \\
\text{has the diagonals } & \frac{1}{\sigma(t)} \in \{1, -1\} \text{ and the norm } \|\frac{1}{\sigma(t)}\| = \sqrt{n_d} \\
\Rightarrow (y_j(x) - y_j(x')) \cdot \prod_{t=1}^{q_d} (x - a_t) & = (y_j(x) - y_j(x')) \cdot \prod_{t=1}^{q_d} (x/a_t - 1) \\
& = \sum_{j=0}^{n_d} (y_j(x) - y_j(x')) (1 - \varepsilon_{td}(j)) x^j \text{ in the polynomial ring } \mathbb{Z}[x] \\
\Rightarrow \|\beta_j\|_2 & \leq \|1[i] = j\| \prod_{i=1}^{n_d} (n_d - 1) i_{t,j}^{-1} \|_F \cdot \|\sigma_t\|_F \cdot \|\alpha_j\|_F \cdot \|\beta_j\|_F
for \( \zeta_i = e^{2\pi \sqrt{-1} \cdot (i + n_d/2)/(2n_d)} \) with \( \left| \zeta_i/a_t - 1 \right| \geq \sqrt{1 + 1/a_t^2} \implies \frac{1 - e}{(1 + e) \sqrt{n_d}} e^{\sum_{i=1}^{n_d} |\zeta_i|} \leq \frac{n_d}{\sqrt{\sum_{i=1}^{n_d} (1 + 1/a_t^2)}}, \) contradicting to \( e^{2\delta_{d+1} n_d^5} \ll (2 - \epsilon)^{4d}. \)

The modulus lifting (Lemma 9.1) transfers the obtained \( \hat{f}(x)(\hat{f}_t(x)) \in \text{ACC}_h[\xi] \) to an \( \text{SYM}^+ \) circuit \( \hat{f}(x) \) in a top-to-bottom recursion. Write \( \hat{f}(x_d) \) for the circuit \( \hat{f} \) considering the \( \text{MOD}-gates \) \( x_{d_r} \) at depth \( 2d - 1 \) as the input variables \( x_d = (x_{d_r})_n \). So, \( x = x_0, \hat{f}(x) = x_{h+1}, \) and \( x_{d_r} \in \text{MOD}[p] \circ \text{ACC}_{2d-1}[\xi] \) of \( p = p_x(\lambda_1) \) and \( \xi(\lambda_1) = \xi(\lambda_1) \) on every path \( \lambda \) passing through \( x_{d_r}. \) The induction hypothesis gives \( \hat{f}(x_d) \) of degree \( u_d \) and asks to present a \( \hat{f}_t(x_d) \) of degree \( u_{d-1} \) via replacing every \( x_d \) in the above \( \hat{f}'s \) construction of

\[
\begin{align*}
\text{Truncated Taylor series:} & \quad \hat{x}_{d_r} := \sum_{i=1}^{2d} \hat{a}_i \hat{f}_{i+1}(x_{d-1}) - a_i = \sum_{i=1}^{2d} \hat{a}_i \sum_{\ell=1}^{k_i} x_{d-1}^{\ell}, \quad a_i := \sum_{\ell=1}^{k_i} x_{d-1}^{\ell}, \quad a := k t'.
\end{align*}
\]

\[
\text{Modulus lifting:} \quad \hat{f}_{i+1}(x_{d-1}) = \hat{f}_{i+1}(x_{d-1}) \mod p_{k_{i+1}}, \quad \text{where} \quad \hat{f}_{i+1}(x_{d-1}) \cdot \text{norm}(\hat{f}_{i+1}(x_{d-1})) = p_{k_{i+1}}.
\]

\[
\begin{align*}
\text{SYM+ degree:} & \quad \text{Let} \quad \hat{f}_d := \hat{f}_{d+1}(x_{d+1}), \quad u_d := \text{deg}(\hat{f}_d), \quad \text{and} \quad u_{d-1} := \text{deg}(\hat{f}_{d-1}).
\end{align*}
\]

\[
\text{SYM+ norm:} \quad \text{The ratios increase by} \quad \frac{\text{norm}(\hat{f}_{d-1})}{\text{norm}(\hat{f}_d)} \leq \text{norm}(\hat{f}_{d-1})^u_{d-1} \leq \text{norm}(\hat{f}_d)^u_d \leq \text{norm}(\hat{f}_{d-1})^u_{d-1} \leq \text{norm}(\hat{f}_d)^u_d.
\]

\[
\text{Theorem 9.6 ([Wig94])}. \quad \text{NL/poly} \subset \Theta(L/poly)
\]

\[
\text{Theorem 9.7 (Theorem 1.11)}. \quad \text{Suppose} \quad \text{ACC}_h[\text{SYM}] \text{ of size} \quad 2^{(\log n)^O(1)} \quad \text{either computes CMD or approximates DCMD by advantage} \quad \frac{1}{2} + \frac{1}{2(\log n)^O(1)}.
\]

\[
\text{Then} \quad \text{DEP}[\text{poly}(n)] \quad \text{cannot} \quad (\frac{1}{2} + \frac{1}{2(\log n)^O(1)}).
\]

\[
\text{Theorem 9.8 (Theorem 1.11)}. \quad \text{Suppose} \quad \text{ACC}_h[\text{SYM}] \text{ of size} \quad 2^{(\log n)^O(1)} \quad \text{approximates DCMD by advantage} \quad \frac{1}{2} + \frac{1}{2(\log n)^O(1)}.
\]

\[
\text{Let} \quad \{1\}^* \subset L \subset \text{NTIME}[2^\nu] - \text{NTIME}[2^\nu/poly(\nu)].
\]

\[
\text{Theorem 9.9 (Theorem 1.11)}. \quad \text{Suppose} \quad \text{ACC}_h[\text{SYM}] \text{ of size} \quad 2^{(\log n)^O(1)} \quad \text{approximates DCMD by advantage} \quad \frac{1}{2} + \frac{1}{2(\log n)^O(1)}.
\]

\[
\text{Theorem 9.10 (Theorem 1.11)}. \quad \text{Suppose} \quad \text{ACC}_h[\text{SYM}] \text{ of size} \quad 2^{(\log n)^O(1)} \quad \text{approximates DCMD by advantage} \quad \frac{1}{2} + \frac{1}{2(\log n)^O(1)}.
\]

\[
\text{Theorem 9.11 (Theorem 1.11)}. \quad \text{Suppose} \quad \text{ACC}_h[\text{SYM}] \text{ of size} \quad 2^{(\log n)^O(1)} \quad \text{approximates DCMD by advantage} \quad \frac{1}{2} + \frac{1}{2(\log n)^O(1)}.
\]
SYM\textsuperscript{+}[\text{deg}:2^{2^{2^{n'}}}, \text{norm} \exp(2^{2^{2^{n'}}})] to compute \( \Pr[\hat{f}'(X) = \bigvee_{x \in \{0,1\}^n} C_{\mu^{\text{W}_{\text{IE}}}}^{\text{WS}}(X, x)] \geq 1 - \delta \).

Williams’s dynamic program [Wil14a] can estimate it in a contradictory fast time

\[
\text{Nondeterministic time for acceptance probability estimation: } \text{poly}(n) \cdot (2^n + 2^{n'} \cdot \text{norm}(\hat{f})) \ll 2^{\nu(1-o(1))}.
\]

\section{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 \( q \in 2^N \) of order \( q = (\log n)^{O(\log n)} \) and identify a binary string \( \tilde{x} \) with the \( q \)-nary vector \( x \) via \([q]^n \ni x \cong \tilde{x} \in 2^n\). It algebraizes a language \( L \subseteq [q]^n \cong \{0,1\}^n \) to a set-multilinear polynomial \( \hat{L} := \sum_{x \in [q]^n} \hat{L}(x)[x_i]_{i=1}^n \), and \( \hat{F} := \{ \hat{L} \mid L \in F \} \).

\textbf{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 \( O(\log d) \).

\textbf{Definition 9.9} (multi-linear polynomial). A polynomial is set-multilinear over variables \( X_1 \sqcup \cdots \sqcup X_r \) if every term (monomial) contains one \( X_i \) variable. A circuit is set-multilinear if so is every gate over subsets of \( \{ X_1, \cdots, X_r \} \).

\textbf{Lemma 9.10} (multi-linearization). Any algebraic circuit of size \( s \) and depth \( d \) computing a set-multilinear polynomial over variables \( X_1 \sqcup \cdots \sqcup X_r \) can transfer to a set-multilinear circuit of size \( (d + 2)^r \cdot s \) and depth \( 2d \).

\textbf{Lemma 9.11} (Theorem 1.32). Any sum \( f = \sum_{k=1}^s \sum_{i,j=1}^n x_i \lambda_i(k) \mu_j(k) x_j \) of \( s \ll n \) bilinear forms over \( M_{ij}(k) \in \mathbb{F} \) 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 \( O(s^2n) \) data and \( O(s^2n \log |\mathbb{F}|) \) guess bits in \( O(s^2n) \) time.

\textbf{Proof.} Without loss of generality, we may assume that given \( s \) bilinear forms have disjoint keys (i.e., specific indices) \( K = \{i_k, j_k \mid k \in [s] \}, \lambda_{i_k}(k) \mu_{j_k}(k) \neq 0 \). Otherwise, there exist \( 2s' \leq 2s \) keys \( K \) to cover either \( \{i \mid \lambda_i(k) \neq 0\} \subseteq K \) or \( \{j \mid \mu_j(k) \neq 0\} \subseteq K \) over all \( k > s' \) so that \( f \) is learnable by only \( s'n \) queries \( f(1_{i_k} + 1_j) \) and \( f(1_i + 1_{j_k}) \) over \( \{i_k, j_k\} \in K \). 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.

\textbf{Gaussian elimination (Jacobian matrix triangulation)} 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

A quadratic polynomial mapping:

\[
\hat{f}(1_{i_k} + 1_j) := f(1_{i_k} + 1_j) + \sum_{k' < k} a_{k', k} f(1_{i_{k'}} + 1_j), \hat{f}(1_i + 1_{j_k}) := f(1_{i} + 1_{j_k}) + \sum_{k' < k} b_{k', k} f(1_i + 1_{j_{k'}})
\]

an invertible mapping over \( \mathbb{F} \sum_{k=1}^{2(n-k)} 2^{(n-k)} \), so uniquely identify the all argued \( \lambda_i(k) \) and \( \mu_j(k) \), and all \( \lambda_i(k) \) and \( \mu_j(k) \) as well. Additional \( s(s + 1)/2 \) queries to evaluate \( \hat{f}(1_{i_k} + 1_{j_{k'}}) \) over all \( 1 \leq k' \leq k \leq s \) can determine the unargued coefficients \( \lambda_{i_{k'}}(k) \) and \( \mu_{j_{k'}}(k) \), too.

\textbf{Theorem 9.12}. \( \tilde{\text{VP}} \not\subseteq \text{VSIZE}[2^{2^{2^{n'}}}] \) or \( \forall \epsilon > 0, \forall k \geq 1, \text{NTIME}[\exp(n^\epsilon)] \not\subseteq \text{SIZE}[n^k] \).
Proof. Follow Theorem 9.7’s argument on Theorem 8.12’s way to apply the easy witness lemma (Theorem 8.11). Suppose NTIME[tc,1] ⊂ SIZE[νk−cn,1/c] for ℓ := 2ℓ. Williams’ trick [Wil14a] has reduced the recognition of ℒ ∈ NTIME[2p] \ NTIME[2p/poly(ν)] to measuring the acceptance probability of an OR-top circuit Cn(x) := \bigvee_{w \in \{0, 1\}^n} C^n_{ν, w}(x, w) ∈ NP of x ∈ [q]n by taking \( \tilde{n} = \frac{1-\epsilon}{3} \nu \). The assumption \( \widehat{NTIME} \subset \text{VSIZE}[s] \) of \( s = 2^{(\log^3 n) \log \log n} = 2^{\log^3 n} \) presents an algebraic circuit \( \hat{C}_n := C^n_{ν, \tilde{n}} \) of a homogeneous polynomial \( \hat{C}_n(x) = \sum_{x \in [q]^n} C_n(x)x_{x, y} \) of the terms \( x \in \prod_{i=1}^n x_{i,x_i} \).

### Learning elusive bilinear decompositions of algebraic circuits: Theorems 9.8 and 9.10 transfer \( \hat{C}_n \) to a set-multilinear formula of size no more significant than \((d+2)^n sO(\log n)\) and depth \( d = O(\log n)\). Decompose it to a sum of bilinear forms \( \hat{C}_n = \sum_{k=1}^s \sum_{x \in I_k} \sum_{y \in J_k} \lambda_{xy}(k)\lambda_{y}(k)x_{y} \) of \( I_k \times J_k \cong [q]^n \) with balance \( n/3 \leq \log_q |I_k| \leq 2n/3 \). There are \( s' \leq ((d+2)^n sO(\log n))d \) forms with \( (d+2)^d \ll q \) and \( s' \leq sO(\log n)^d = 2^{o(n)} \). Lemma 9.11 can identify them by querying for \( \sum_{k=1}^{d^n} |I_k| \) times to evaluate \( \hat{C}_n \) in \( s' \cdot q^{2n/3} \cdot 2^{n-\epsilon} \cdot \text{poly}(\nu) \ll 2^{n(1-\epsilon+o(1))} \) time. Once getting all coefficients \( \lambda_{xy}(k) \) and \( \lambda_{y}(k) \), one can estimate the acceptance probability in \( s' \cdot q^n \cdot \sum_{k=1}^{d^n} |I_k||J_k| \leq s' \cdot q^n \cdot 2^{n(1-\epsilon+o(1))} \) time, contradicting to \( \ell \not\in \text{NTIME}[2^\ell/\text{poly}(\nu)] \). □

**Theorem 13** (generalized easy witness lemma for depth [CR20]). Given smooth functions \( \ell(n), d(n) \) and \( \log s(n) \). Suppose \( s(s(n)^{c,1})^{c,1} \leq 2^{d(\log n)} \) and \( t(n) = \exp\left(\frac{c,1}{\delta(\log n)}\right) \) is non-decreasing. If every \( \text{NTIME}[t(n)] \) language is \((1/2 + 1/s(n))-\text{approximable} \) by circuits of depth \( \log s(n) \), then every unary NTIME[exp(n)] language must have a witness of \( \text{DEP}[d(n)] \).

**Theorem 14** (Theorem 1.12). Suppose \( \text{VP} \) either computes CMD or approximates DCMD by advantage \( \frac{1}{2} + \frac{1}{2^{\log \log n}o(1)} \). Then \( \text{DEP}[\log n^k] \) cannot \( (\frac{1}{2} + \frac{1}{2^{\log \log n}o(1)})\)-approximate \( \text{NTIME}[2^{\log n^k}] \).

**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 \( d(n) = e n/\log^2(n), \ell(n) = \log^k n, \) and \( s(n) = 2^{(\log n)(1-\epsilon)k/4} \), 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 \( t = \text{poly}(n), \) yielding \( \mathcal{W}_n \in \text{DEP}[d(n)] \) of \( \frac{1-\epsilon}{3} \nu = \tilde{n} \), so \( \hat{C}_n \in \text{SUM}_{\mathcal{C}} \cap \text{DEP}[4^t] \) by Theorem 9.14 for Theorem 9.7’s \( m = O(4^d(n)), \) \( \varepsilon_0 = o\left(\frac{1}{2^{\log \log n}}\right) \) and \( \varepsilon_0 = O\left(\frac{1}{\log \delta}\right) \). Theorem 9.12 has learned the elusive bilinear decomposition of \( \hat{C}_n \) in a contradictory fast time. □

**Theorem 15** (combinatorial design [NW94]). For \( k = O(m^2/\log n) \) and \( n < 2^m \) there is \( S_1, \ldots, S_n \subset [k] \) with \( |S_i| = m \) and \( i \neq j \Rightarrow |S_i \cap S_j| \leq \log n \). Such an \( m \)-set family \( \mathcal{S} \) is constructible in deterministic \( \text{poly}(n, 2^k) \) time and called \( (m, \log n, \text{combinatorial design}). \)

**Theorem 16** (hardness to derandomization [KI04]). Let \( \mathcal{S} \) be an \( (m, \log n) \)-combinatorial design. Let \( f(x) \) be an \( m \)-variate multi-linear polynomial which an algebraic circuit of size \( s \) cannot compute. Let \( \mathcal{C}(y) \) be an \( n \)-variate circuit of size \( s' \) and degree \( d \). If \((s'nmn)^5 < s \) then \( \mathcal{C}(y) \equiv 0 \Leftrightarrow \mathcal{C}(f(\mathbf{x} \mid S_1), \ldots, f(\mathbf{x} \mid S_n)) \equiv 0. \)

**Theorem 17** (derandomizing PIT). Either PIT is solvable in deterministic \( n^{\text{poly}(\log \log n)} \) time, or \( c > 0, \forall k \geq 1, \text{NTIME}[\exp(n^k)] \not\subset \text{SIZE}[n^k] \).

**Proof.** Theorem 9.12’s algebraic circuit hardness derandomizes PIT. Suppose \( \text{NTIME}[\text{poly}(m)] \not\subset \text{VSIZE}[s] \) for \( s = 2^{\log^3 m \log \log m} \). Let \( m = \log n (\log \log n)^{3+2\epsilon} \). We have an \( m \)-variate \( f(\mathbf{x}) \in \text{NP} \).
whose algebraic circuit size must be $s = 2^{\Omega(\log n \log \log n / \log \log \log n)}$. Since $(s'nm)^5 < s$ for $s' = \text{poly}(n)$ and $d = n$, PIT is solvable in $|\{0,1\|^m \cdot s' \cdot 2^{O(m^2/\log n)} = O(2^{\log n (\log \log n)^{6+c}})$ time by exhausting the input space $x \in \{0,1\}^m$ to evaluate Theorem 9.16's $C(f(x \mid S_1), \ldots, f(x \mid S_n))$. \hfill \square

\section{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?

\textbf{From refutation to approximation:} Max$k$SAT of $O(n^{k-1})$ constraints required $2^{n^{1-c}}$ time to approximate $\max_x P(y = f_\theta(x))$ under ETH [FLP16]. Meanwhile, we have shown that promise-Max$k$SAT to distinguish between $|\max_x P(y = f_\theta(x) - \max_x P'(y = f_\theta(x))| \geq \epsilon$ 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 $f_\theta(x) =$ \text{\bigwedge}_{i=1}^n \theta \circ x_i$? For example, Max$\text{CUT}$ requires the sample complexities (number of edges) $\Omega(n^{2-c})$ for $O(2^{n^c})$-time approximation ([FLP16]), but only $\tilde{O}(n)$ edges for the $\tilde{O}(n)$-time distinguish-ment (Theorem 6.21 of $k = 2$). How about Max$k$CSP, Densest$k$Subgraph, Min$\text{Bisection}$, etc.?

\textbf{PAC learning planted kDNF (in the worst-case):} We have shown that the planted $k$DNF is PAC learnable from any $\tilde{O}(n^{k/2})$ 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 $\Omega(n^{(k-c)/2})$ data. Sub-exponential size LP might require $\Omega(n^{(k-c)/2})$ data learning since it was so for noisy PAC learning [BCR20].

\textbf{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 $\Pr((-1)^{G(X_i)+Y} \mid |X_i/2| = a)$ under an $O(\log s)$-independent flipper $G_i$. 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].

\textbf{Inverting planted Fourier transform and LWE:} Fortunately, degree-$k$ multi-linear polynomials $f(x_1, \ldots, x_d) = \sum_{|w| \leq k} \prod_{i \in w} \theta((x_i/2))(\cdot)^{X_i}$ over $\mathbb{Z}_q$ have the statistically non-zero correlation $\Pr[Y \cdot (-1)^{G(X_i)} \mid |X_i/2| = a]$ 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 $\Pr[Y \neq f(G(X)) \mid |X_i/2| = a] \approx 0$. 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.

\textbf{Computational complexity lower bounds:} We have shown that either $\text{PSPACE}^{cc} \not\subset \text{PH}^{cc}$ or $\forall k, \text{quasi-NP} \not\subset \text{quasi-NC}^k$ must hold. The latter quasi-NP \not\subset quasi-NC$^k$ may not extend immediately to quasi-NP \not\subset quasi-NC (so NEXP \not\subset PSPACE). For example, Theorem 8.12’s non-deterministic time analysis allows a sparsity $|M|_{\neq 0} \leq 2^{c \cdot n}$, but the hardness magnification demands a much sparser $|M|_{\neq 0} \leq 2^{c \cdot n}$ for $c < 1$ [CJW19]. We have established quasi-NP \not\subset \text{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 quasi-NP $\not\subset$ NC$^1$ and quasi-NP $\not\subset$ L. As for algebraic circuit complexity, we have shown either VP $\neq$ VNP or $\forall k, \text{quasi-NP} \not\subset \text{NC}^k$. Extending Murray-Williams-Chen-Ren’s easy witness lemmas and replacing the latter quasi-NP $\not\subset \text{NC}^k$ with NP $\not\subset \text{P/poly}$ might establish VP $\neq$ VNP.

References


Arkadev Chattopadhyay and Nikhil Mande. Weights at the bottom matter when the top is heavy. *Electronic Colloquium on Computational Complexity (ECCC)*, 24(83), 2017.


