Learning CNF Formulas from Uniform Random Solutions in the Local Lemma Regime

Yiyao Zhang

Nanjing University

Joint work with:

Weiming Feng — The University of Hong Kong

Xiongxin Yang — UC Santa Barbara

Yixiao Yu — Nanjing University

STOC 2026

Salt Lake City, Utah, USA

CNF (Conjunctive Normal Form) Formula

Variable set: $V = \{v_1, \ldots, v_n\}$ Literal of v: $\{v, \neg v\}$

Clause set: $C = \{c_1, \ldots, c_m\}$ Clause: disjunction of literals

CNF formula $\Phi = c_1 \wedge \ldots \wedge c_m : \{\text{True}, \text{False}\}^n \to \{\text{True}, \text{False}\}$

Width k: each clause contains **exactly** k distinct literals (denoted by k-CNF)

3-CNF

clause

$$\Phi = (v_1 \vee \neg v_2 \vee v_3) \wedge (v_1 \vee v_2 \vee v_4) \wedge (v_3 \vee \neg v_4 \vee \neg v_5)$$

literal

Learning CNF Formulas from Uniform Random Solutions

Solution of a CNF formula Φ: assignment $\{\text{True}, \text{False}\}^n$ that satisfies the CNF Φ

μ_Φ **uniform** distribution over all solutions of Φ

→ Sample oracle of Φ → $X \sim \mu_\Phi$

Input: sample oracle of a k-CNF Φ, n, k, an error bound $\varepsilon \geq 0$ and $\delta > 0$

Output: a CNF formula $\hat{\Phi}$ such that $\mathrm{d}_{\mathrm{TV}}(\mu_\Phi, \mu_{\hat{\Phi}}) \leq \varepsilon$ w.p. $\geq 1 - \delta$

Approx. Learning: $\varepsilon > 0$ Exact Learning: $\varepsilon = 0$

Question: How many samples are sufficient? (Sample Complexity)

Valiant's Algorithm

Input: sample oracle of a k-CNF Φ, n, k, an error bound $\varepsilon \geq 0$ and $\delta > 0$

Output: a CNF formula $\hat{\Phi}$ such that $\mathrm{d_{TV}}(\mu_\Phi, \mu_{\hat{\Phi}}) \leq \varepsilon$ w.p. $\geq 1 - \delta$

Let $\hat{\Phi}$ contain all possible clauses;

Repeat T times: $2^k \cdot \binom{n}{k}$ in total

query a sample $X \sim \mu_\Phi$;

remove all clauses in $\hat{\Phi}$ that X violates;

Output $\hat{\Phi}$.

[Valiant 1984] **Distribution Free ✓**

Sample complexity:

$$T = O_k\left(\frac{n^k + \log(1/\delta)}{\varepsilon}\right)$$

Exact Learning ✘

Computational complexity:

$$O_k(n^k \cdot T)$$

Lovász Local Lemma

Degree d: each variable appears in **at most** d different clauses

(k, d)-CNF: width k and degree d

[Erdős, Lovász 1975]

(k, d)-CNF Φ must have a solution if

$$k \geq \log d + \log k + \log \mathrm{e} = \log d + o(k).$$

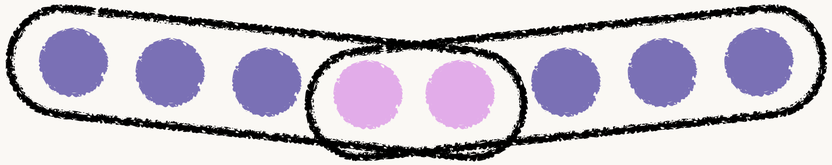
Main problem: running **Valiant's algorithm** under **local lemma condition,**

can we improve the **sample complexity** from $O(n^k)$?

is sample efficient **exact** learning possible?

Our Results: Exact Learning with Bounded Intersections

Intersection bound s: any two distinct clauses share **at most** s variables



intersection bound $s = 2$

(k, d, s)-CNF: width k, degree d and intersection bound s

Sublinear intersection $s = k^{1-\eta}$

$$k \geq \log d + o(k) + O_\eta(1)$$

Satisfying Condition!

Sample complexity:

$$T = O_k(\log \frac{n}{\delta})$$

Linear intersection $s = \zeta k$

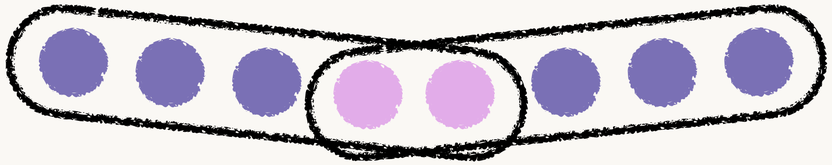
$$k \geq C(\zeta) \log d + o(k) + O_\zeta(1)$$

Sample complexity:

$$T = O_k(\log \frac{n}{\delta})$$

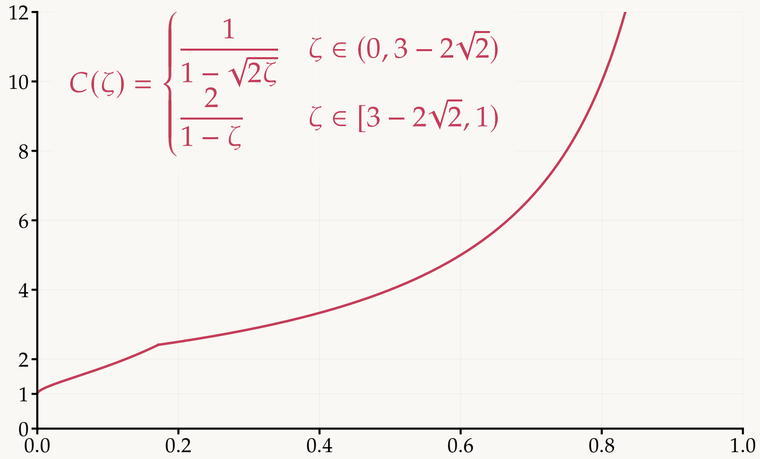
Our Results: Exact Learning with Bounded Intersections

Intersection bound s: any two distinct clauses share **at most** s variables



intersection bound $s = 2$

(k, d, s)-CNF: width k, degree d and intersection bound s



$C(\zeta) \to 1$ when $\zeta \to 0$

$C(\zeta) \to \infty$ when $\zeta \to 1$

Linear intersection $s = \zeta k$

$$k \geq C(\zeta) \log d + o(k) + O_\zeta(1)$$

$$C(\zeta) = \begin{cases} \frac{1}{1-\sqrt{2\zeta}} & \zeta \in (0, 3-2\sqrt{2}), \\ \frac{2}{1-\zeta} & \zeta \in [3-2\sqrt{2}, 1). \end{cases}$$

Sample complexity:

$$T = O_k(\log \frac{n}{\delta})$$

Our Results: Learning with Unbounded Intersections

Question: Is intersection bound necessary?

Any algorithm that exactly learns an n-variable $(k, k, k-1)$-CNF formula from uniform random solutions with prob. at least $1/3$ requires $\exp(\Omega_k(n))$ samples.

Any algorithm that approx. learns an n-variable $(k, k, k-1)$-CNF formula from uniform random solutions with TV distance ε and prob. at least $1/3$ requires $\Omega_{k,\varepsilon}\left((n/\log n)^{1-2/k}\right)$ samples.

can be further improved to $\Omega_{k,\varepsilon}(n/\log n)$!

Technique: (Distance-based) Fano's Inequality

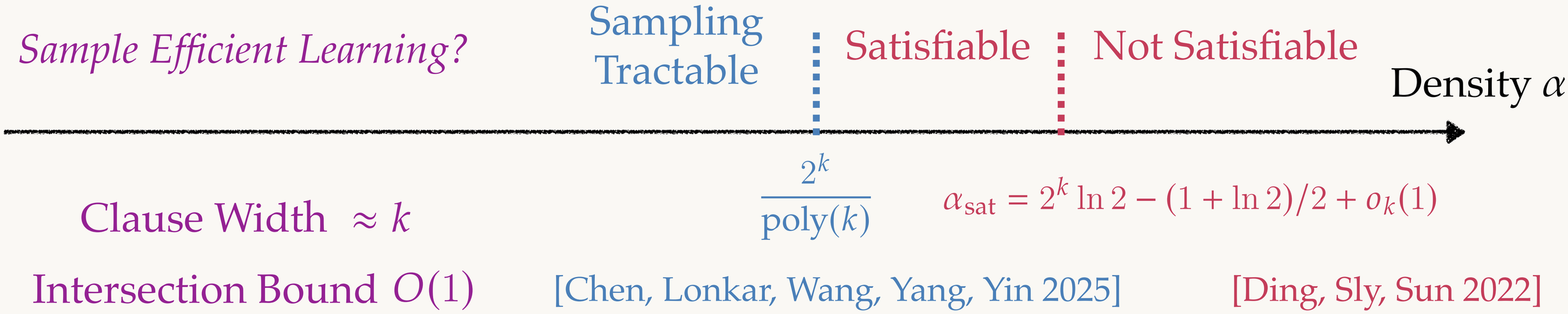
Our Results: Exact Learning Random CNF Formula

$\Phi = (k, n, m = \lfloor \alpha n \rfloor)$ n variables and $m = \lfloor \alpha n \rfloor$ random clauses of size k

Each clause is generated by independently selecting k literals uniformly at random.

Density α: average degree of the CNF formula

With high probability over the randomness of Φ,



Clause Width $\approx k$

Intersection Bound $O(1)$

[Chen, Lonkar, Wang, Yang, Yin 2025]

[Ding, Sly, Sun 2022]

Our Results: Exact Learning Random CNF Formula

$\Phi = (k, n, m = \lfloor \alpha n \rfloor)$ n variables and $m = \lfloor \alpha n \rfloor$ random clauses of size k

Each clause is generated by independently selecting k literals uniformly at random.

Density α: average degree of the CNF formula

With high probability over the randomness of Φ,

Sample Efficient Exact Learning

Sampling Tractable

Satisfiable

Not Satisfiable

Density α

$\frac{2^{k-o(k)}}{\mathrm{poly}(k)}$

$\frac{2^k}{\mathrm{poly}(k)}$

$\alpha_{\mathrm{sat}} = 2^k \ln 2 - (1 + \ln 2)/2 + o_k(1)$

Sample Complexity

$$O_k\left(n^{\exp(-\sqrt{k})} \log \frac{n}{\delta}\right)$$

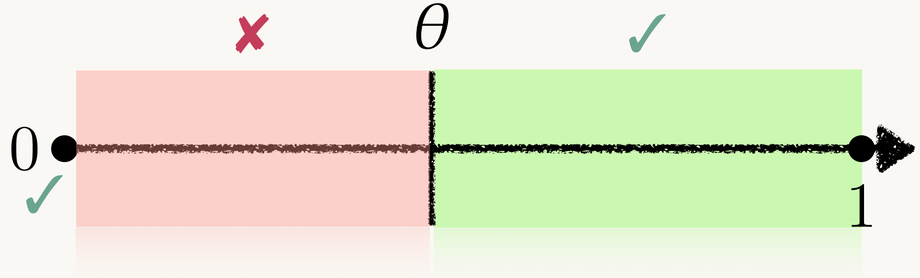
[Chen, Lonkar, Wang, Yang, Yin 2025]

[Ding, Sly, Sun 2022]

Proof Overview: Exact Learning with Bounded Intersections

θ-resilience: for any candidate clause c^*,

$$\Pr_{X \sim \mu_\Phi}[c^* \text{ is violated}] = 0 \text{ or } \Pr_{X \sim \mu_\Phi}[c^* \text{ is violated}] \geq \theta.$$



Previous applied in models with pair-wise hard constraints:

[Bresler, Gamarnik, Shah 2014]: Independent Set;

[Blanca, Chen, Štefankovič, Vigoda 2020]: Graph Coloring.

By Markov's inequality,

$O\left(\frac{k}{\theta}\log\frac{n}{\delta}\right)$ uniform random solutions suffice for exact learning.

Goal: establish $\Omega_{k,d}(1)$-resilience for (k, d, s)-CNF

Proof Overview: Exact Learning with Bounded Intersections

Goal: establish $\Omega_{k,d}(1)$-resilience for (k, d, s)-CNF

Local Uniformity [Moitra 2019] & [Feng, Guo, Yin, Zhang 2021]

Let $\Phi = (V, C)$ be a CNF formula. Each clause has width **at least** k_1 and **at most** k_2.

For any $t \geq k_2$, if $2^{k_1} \geq 2edt$, then for any $v \in V$,

$$\max\{\Pr_{X\sim\mu_\Phi}[X_v = \text{True}], \Pr_{X\sim\mu_\Phi}[X_v = \text{False}]\} \leq \frac{\exp(1/t)}{2}.$$

Main idea: control the clause width while pinning by the chain rule

Pinning reduces the clause width.

The intersection bound helps!

Proof Overview: Exact Learning Random CNF Formula

With high probability over the randomness of Φ,

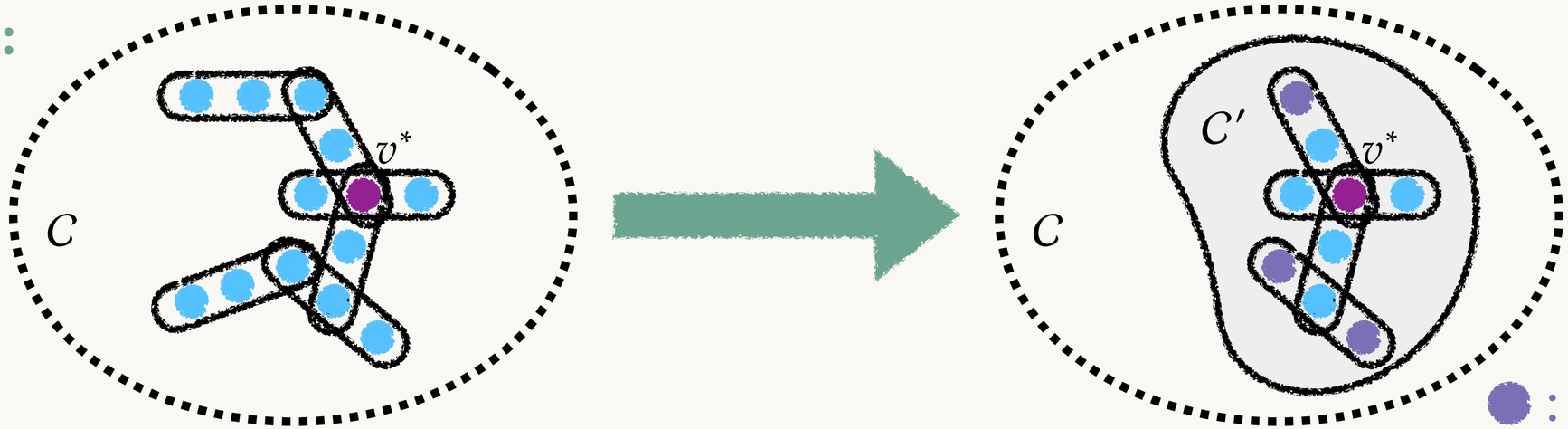
Goal: fix a variable v^* and an assignment σ_{v^*}, we want to show that

(by chain rule)

$$\Pr_{X\sim\mu_\Phi}[X_{v^*} = \sigma_{v^*}] \gtrsim n^{-\exp(-k^{4/5})}.$$

Main challenge: max-degree can be $d \approx \dfrac{\log n}{\log\log n}$ *fails to directly apply local uniformity*

Main Idea:



● : variables in S

Proof Overview: Exact Learning Random CNF Formula

Step 1: Separating high-degree variables. [GGGY21], [HWY23], [CGG+24], [CLW+25]

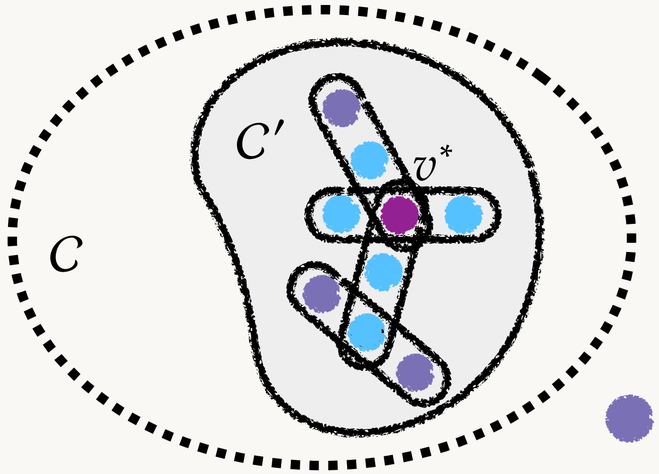
Good variable has bounded degree.

Step 2: Pre-revealing process. Inspired by frozen paradigm in [He, Wu, Yang 2023]

Draw $X \sim \mu_\Phi$;

Start from v^* and keep revealing values of good variables in a BFS manner; Subset S

Stop until X_{v^*} depends only on $C' \subseteq C$ conditioned on the revealing result.



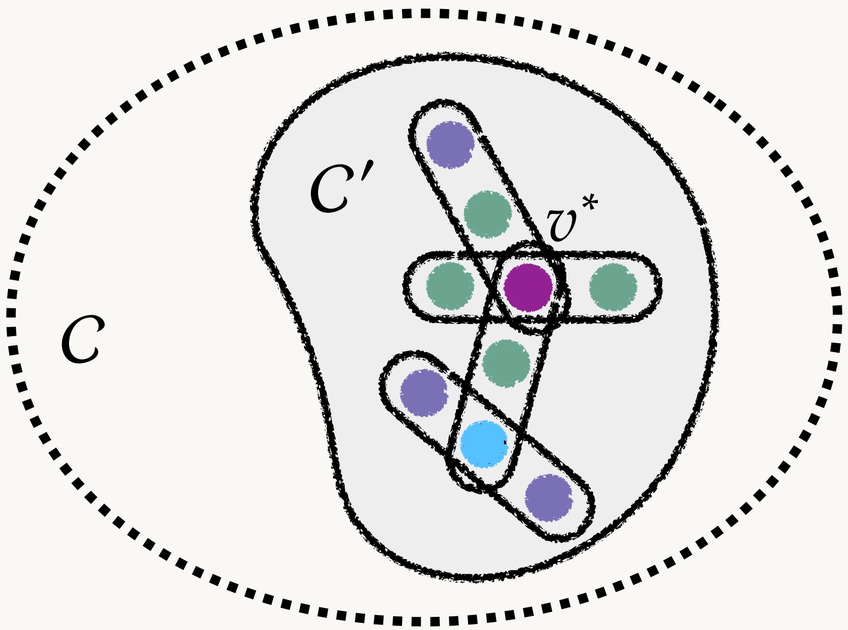
● : variables in S

By a modified local uniformity,

$$\Pr_{X \sim \mu_\Phi}[\,|C'| \le \log n] \ge 1/2.$$

Proof Overview: Exact Learning Random CNF Formula

Step 3: Conditional-revealing process. *New structure property!*



● : variables in S

● : one-degree variable

With high probability over the randomness of Φ,

for any $C' \subseteq C$ with $|C'| \leq \log n$,

there exists a clause $c \in C'$ such that

$$\left| \mathrm{vbl}(c) \setminus \bigcup_{c' \in C' \setminus \{c\}} \mathrm{vbl}(c') \right| \geq k - o(k).$$

$\mathrm{vbl}(c)$: set of variables in c

number of one-degree variables in the subformula

Pin one-degree variables and simplify the CNF formula until v^* becomes isolated.

Summary

Sample efficient **exact** learning

- **bounded** degree CNF formula with $T = O_k(\log \frac{n}{\delta})$ samples:

  sublinear intersection $s = k^{1-\eta}$ if $k \geq \log d + o(k) + O_\eta(1)$;

  linear intersection $s = \zeta k$ if $k \geq C(\zeta) \log d + o(k) + O_\zeta(1)$.

- random CNF **near the satisfiability threshold** with $O_k\left(n^{\exp(-\sqrt{k})} \log \frac{n}{\delta}\right)$ samples.

Sample complexity **lower bounds**

exact learning: $\exp\left(\Omega_k(n)\right)$

approx. learning: $\Omega_{k,\varepsilon}\left((n/\log n)^{1-2/k}\right)$ $\Omega_{k,\varepsilon}\left(n/\log n\right)$

Summary

Open Problems

Sample efficient **exact** learning

- **bounded** degree CNF formula with $T = O_k(\log \frac{n}{\delta})$ samples:

  sublinear intersection $s = k^{1-\eta}$ if $k \geq \log d + o(k) + O_\eta(1)$;

  linear intersection $s = \zeta k$ if $k \geq C(\zeta) \log d + o(k) + O_\zeta(1)$.

  Better tradeoff?

- random CNF **near the satisfiability threshold** with $O_k\left(n^{\exp(-\sqrt{k})} \log \frac{n}{\delta}\right)$ samples.

Better regime and better sample complexity?

Sample complexity **lower bounds**

Approx. learning without intersection bound with $\mathrm{poly}(n)$ *samples?*

exact learning: $\exp\left(\Omega_k(n)\right)$

approx. learning: $\Omega_{k,\varepsilon}\left((n/\log n)^{1-2/k}\right)$ $\Omega_{k,\varepsilon}\left(n/\log n\right)$

Thank you!

arXiv: 2511.02487