Learning CNF Formulas Meets Lovász Local Lemma

Weiming Feng[1], Xiongxin Yang[2], Yixiao Yu[3], Yiyao Zhang[3]

[1]The University of Hong Kong [2]UC Santa Barbara [3]Nanjing University

Learning CNF formulas

Variable set: $V = \{v_1, \ldots, v_n\}$ **Literal:** $\ell_j \in \{v_j, \neg v_j\}$

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

k-CNF formula: $\Phi = (V, C)$, where each clause is a disjunction of *exactly* k literals.

Solutions: $\Omega_\Phi = \{x \in \{0,1\}^V : x \models \Phi\}$, and μ_Φ is the uniform distribution over Ω_Φ.

Learning CNF formulas from uniform random solutions

Input: a sample oracle of a k-CNF Φ, and error parameters $\varepsilon > 0$ and $\delta > 0$.

Output: a CNF $\widehat{\Phi}$ such that $d_{\mathrm{TV}}(\mu_{\widehat{\Phi}}, \mu_\Phi) \le \varepsilon$ with probability at least $1 - \delta$.

Approx. learning: $\varepsilon > 0$ *Exact learning:* $\varepsilon = 0$

Valiant's elimination algorithm [Val84]

Let $\widehat{\Phi}$ contain all possible clauses;
Repeat the following T times:
  query a sample $X \sim \mu_\Phi$;
  remove all clauses in $\widehat{\Phi}$ violated by X;
Output $\widehat{\Phi}$.

Sample complexity: $T = O_k(\frac{n^k + \log(1/\delta)}{\varepsilon})$

Computational complexity: $O_k(n^k \cdot T)$

Remark: Valiant's algorithm works even distribution-free, but cannot exact learn.

Can uniform random solutions be leveraged to reduce sample complexity?

Can uniform random solutions enable exact learning?

Lovász local lemma and satisfiability threshold

Bounded degree (k,d)-CNF: each variable appears in at most d clauses.

Random k-CNF $\Phi(k, n, \lfloor \alpha n \rfloor)$: $\lfloor \alpha n \rfloor$ i.i.d. random k-clauses.

Lovász local lemma [EL75]

(k,d)-CNF is satisfiable if
$k \ge \log d + \log k + \log e = \log d + o(k)$

Satisfiability threshold [DSS22]

$\Phi(k, n, \lfloor \alpha n \rfloor)$ is satisfiable w.h.p. if
$\alpha < \alpha_{\mathrm{sat}} = 2^k \ln 2 - (1 + \ln 2)/2 + o_k(1)$

A series of works gives efficient approximate counting and sampling algorithms within both regimes [Moi17, BGGGŠ19, FGYZ20, GGGY21, HWY23, CGGG+24, CLWYY25].

Do these satisfiable, samplable structures make uniform solutions more informative than arbitrary examples?

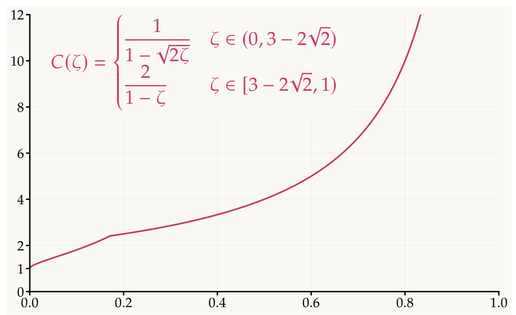
Main results at a glance

Exact learn bounded-intersection CNFs

(k,d,s)**-CNFs**: (k,d)-CNFs where each two clauses share at most s variables.

sublinear intersection $s = k^{1-\eta}$:
- sample complexity: $O_{k,\eta}(\log \frac{n}{\delta})$
- LLL regime: $k \ge \log d + o(k) + O_\eta(1)$

linear intersection $s = \zeta k$:
- sample complexity: $O_{k,\zeta}(\log \frac{n}{\delta})$
- LLL regime: $k \ge C(\zeta) \log d + o(k) + O_\zeta(1)$



Sample complexity lower bounds for (k,d,s)-CNFs

Tightness. Even disjoint clauses already force logarithmic samples:

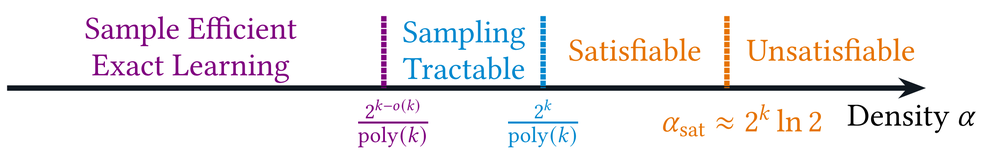
$(k,1,0)$-CNFs $\Omega_k(\log n)$ samples for exact learning.

Necessity. Learning becomes much harder without intersection-size bounds:

$(k,k,k-1)$-CNFs $\exp(\Omega_k(n))$ samples for exact learning,
$\Omega_{k,\varepsilon}(n/\log n)$ samples for approx. learning.

Exact learn random k-CNFs near the threshold

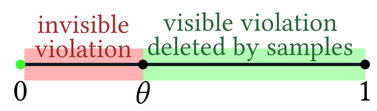
Exact learning when $\alpha \le 2^{k - \widetilde{O}(k^{4/5})}$ with sample complexity $O_k\left(n^{\exp(-\sqrt{k})} \log \frac{n}{\delta}\right)$



Resilience property

Violation probability. For a clause $c^* \notin C$ with forbidden assignment σ^*, define

$$p(c^*) = \Pr_{X \sim \mu_\Phi}[X \text{ violates } c^*] = \Pr_{X \sim \mu_\Phi}[X_{\mathrm{vbl}(c^*)} = \sigma^*].$$



$p(c^*) = 0 \Rightarrow c^*$ implied by Φ

$p(c^*) \ge \theta \Rightarrow \Pr[c^* \text{ survives } T] \le e^{-\theta T}$

Algorithmic side. If every candidate has $p(c^*) = 0$ or $p(c^*) \ge \theta$ (*resilience*), then a union bound gives that Valiant's algorithm achieves exact learning after

$$T = O_k\left(\theta^{-1} \log \frac{n}{\delta}\right).$$

Lower-bound side. Lack of resilience means false clauses can be almost invisible to samples. Constructing hard instances by gadgets together with a distance-based Fano inequality gives a proof of the lower bound.

Techniques: bounded-intersection CNFs

Local uniformity [Moi17, FGYZ20]

For CNF $\Phi = (V, C)$ with clause widths in $[k_1, k_2]$, if $t \ge k_2$ and $2^{k_1} \ge 2edt$, then

$$\forall v \in V, \qquad \max\left\{\Pr_{X \sim \mu_\Phi}[X_v = \mathrm{True}], \Pr_{X \sim \mu_\Phi}[X_v = \mathrm{False}]\right\} \le \frac{1}{2}\exp(1/t).$$

First Idea. Chain-rule pin $\mathrm{vbl}(c^*)$ to σ^* and lower-bound each step.

Issue. Residual clauses may not stay wide enough for local uniformity.

Intersection helps.

$$\widetilde{C} = \{c : |\mathrm{vbl}(c) \cap \mathrm{vbl}(c^*)| \ge t_1\}, \qquad |\widetilde{C}| \le t_2.$$

- **Filter:** only clauses in $\widetilde{C}$ can lose many literals under this pinning process.
- **Pre-satisfy:** satisfy C in advance using at most t_2 variables outside $\mathrm{vbl}(c^*)$.
- **Pin c^*:** every remaining clause has width at least $k - t_1 - t_2$; apply local uniformity.

Techniques: random CNFs

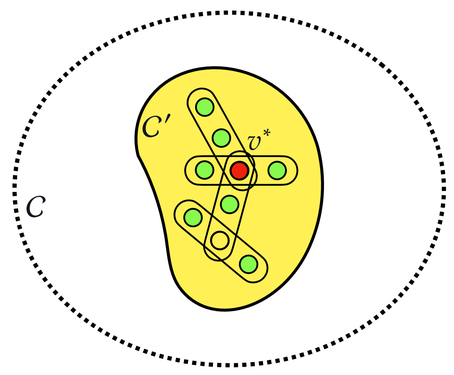
Challenge. Maximum degree can be $\frac{\log n}{\log \log n}$ $\Rightarrow$ Local uniformity directly fails.

New ideas for random CNFs

Step I: Separate high-degree variables.
- Good variables have bounded degree; good clauses have enough good variables;
- The growth of bad-clause components can be bounded.

Step II: Pre-revealing process.
- Draw $X \sim \mu_\Phi$
- Starting from v^*, BFS-reveal good variables
- Stop X_{v^*} depends only a $C' \subseteq C$ conditioned on the revealing result
- A *modified local uniformity* gives

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



Step III: Conditional-revealing process.

A new structure property: many degree-one variables.

W.h.p., every clause set $\widehat{C}$ with $2 \le |\widehat{C}| < \log n$ contains a clause c such that

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

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

Open problems

- Tight trade-off between $C(\zeta)$ and sample complexity for linear intersection $s = \zeta k$.
- Better density regime and sample complexity for random CNFs.
- Approximate learning without intersection bound with $\mathrm{poly}(n)$ samples.