STAT 291 **Spring 2024**

Extra Credit Problems

Instructions: The problems on this sheet are for extra credit. You must earn at least 60% of the points from a problem below to receive credit. Solutions should be included at the end of your solutions to an official problem set for the course. (If the last due date has already passed, you can just email solutions to us separately.)

Collaboration with your classmates is encouraged. However **you should identify everyone you worked with at the beginning of your solution PDF** (e.g. *Collaborators: Alice, Bob, and Eve*). Your solutions should be written entirely by you, even if you collaborated to solve the problems.

The first person to report each substantial mistake in this problem set (by emailing me and Yufan) will receive up to 3 points of extra credit, depending on the mistake.

Problem 1. Moment Bounds for random k-SAT (36 points)

Recall that in the random k-SAT problem, one is given N variables aims to simultaneously satisfy $M = \alpha N$ clauses of the form

$$(L_{i,1} \cdot x_{i,1}) \vee (L_{i,2} \cdot x_{i,2}) \vee \cdots \vee (L_{i,k} \cdot x_{i,k}).$$

Here the indices $(x_{i,1}, \ldots, x_{i,k}) \in \{1, 2, \ldots, N\}^k$ are a uniformly random sequence of distinct values, while the uniformly random "literals" $L_{i,j} \in \{\pm 1\}$ correspond to negations when equal to -1 and do nothing otherwise (and $\vee$ means "or").[1] The expected number of solutions is directly seen to be

$$2^{kN}(1 - 2^{-k})^{\alpha N}.$$

Thus, Markov's inequality implies the upper bound $\alpha_{SAT} \leq 2^k \log 2$ for the satisfiability threshold. This problem will investigate some improved results.

(a) Say a solution is *locally maximal* if there is no way to obtain another solution by changing 1 variable from FALSE to TRUE. Use Markov's inequality on the set $\mathcal{S}_{max}$ of locally maximal solutions to obtain an improved bound on α_{SAT}.

Next we consider the NAE-k-SAT problem, where each clause is paired with a "reversed" version where all literals are negated. For example, the reverse of $(x_4 \vee \neg x_6 \vee x_7)$ is $(\neg x_4 \vee x_6 \vee \neg x_7)$. Each "pair" of clauses is just counted as 1 constraint.

(b) Show that Markov's inequality implies $\alpha_{SAT} \leq 2^{k-1} \log 2$ for NAE-k-SAT.

(c) Show that the second moment for the number of NAE-k-SAT solutions is given by $\exp\big\{\big(f(\alpha, k) \pm o(1)\big)N\big\}$, where

$$f(\alpha, k) = \log 2 + \max_{R \in [0,1]} \left(\alpha \left(1 - \frac{4}{2^k} + \frac{2}{2^k}\big(R^k + (1-R)^k\big) \right) - R\log(R) - (1-R)\log(1-R) \right)$$

[1] For example, the clause $(\neg x_4 \vee x_6 \vee \neg x_7)$ would be written as $(-1 \cdot x_4) \vee (1 \cdot x_6) \vee (-1 \cdot x_7)$.

(d) Show that for $\alpha \leq \alpha_0$ small enough , the first and second moments of the number of NAE-k-SAT solutions match up to a factor of $e^{o(N)}$, with k fixed as $N \to \infty$. (By being more careful and using a general "sharp threshold" result, this calculation implies an almost matching lower bound for the satisfiability threshold of NAE-k-SAT.)

Similarly to p-spin models with non-zero linear term ("external field"), the second moment method fails completely for random k-SAT due to the asymmetry. A work-around is to **weight** satisfying assignments. In particular, for $0 < \lambda < 1$, let us say $\boldsymbol{\sigma} \in \{-1, 1\}^N$ has weight $w(\boldsymbol{\sigma}, C)$ for clause C, where $w(\boldsymbol{\sigma}, C) = 0$ if $\boldsymbol{\sigma}$ violates C, and otherwise $w(\boldsymbol{\sigma}, C) = \lambda^k$ for k the number of disagreements. Then define the total weight $w(\boldsymbol{\sigma})$ to be the product $\prod_{i=1}^{M} w(\boldsymbol{\sigma}, C_i)$ over the clauses C_i.

(e) Argue that to show a solution exists, it suffices to successfully apply the second moment method to $W = \sum_{\boldsymbol{\sigma}} w(\boldsymbol{\sigma})$ rather than the unweighted number of solutions.

(f) Show that λ must solve

$$(1 + \lambda)^{k-1}(1 - \lambda) = 1$$

if the second moment method is to have a chance to work for some $\alpha > 0$. In particular, this means the unweighted 2nd moment method (with $\lambda = 1$) will not work. (Hint: consider the second moment contribution from slightly non-zero overlaps.)

(g) Show that for $\alpha \leq 2^k(\log 2 - o_k(1))$, the first and second moments for W match up to a $e^{o(N)}$ factor. (Again, arguing more carefully leads to an almost sharp lower bound for α_{SAT}.)

Problem 2. The 2-Core of Random XOR-SAT (72 points)

Recall that a random k-XOR-SAT instance consists of $M = \alpha N$ clauses defined on the mod-2 variables $x_1, \ldots, x_N \in \mathbb{Z}_2$:[2]

$$x_{i_{1,j}} + \cdots + x_{i_{k,j}} \equiv a_j \pmod 2, \qquad \forall 1 \leq j \leq M.$$

During class, it was claimed that the behavior of random k-XOR-SAT is related to the 2-core of the associated random k-uniform hypergraph[3] G with M hyperedges $\{x_{i_{1,j}}, \ldots, x_{i_{k,j}}\}$. In this problem, you will investigate this claim in more detail.

(a) Let $G' \subseteq G$ be a sub-hypergraph, consisting of subsets $V_2 \subseteq [N]$ of vertices and $E_2 \subseteq [M]$ of clauses, such that for every clause $j \in E_2$, the vertices $x_{i_{1,j}}, \ldots, x_{i_{k,j}}$ are all in V_2. We say G' is 2-*stable* if every $v \in V_2$ is contained in at least 2 clauses $j \in E_2$, i.e. G' has minimum degree at least 2. Prove that:

- Recursively "pruning" degree 1 vertices yields a (possibly empty) 2-stable sub-hypergraph G'.
- The resulting G' contains any other 2-stable sub-hypergraph of G.

[2] More precisely, we assume each set $\{x_{i_{1,j}} + \cdots + x_{i_{k,j}}\}$ is a uniformly random subset of k distinct variables, and these subsets are independent for different j.

[3] A k-uniform hypergraph is just a graph with "hyper-edges" that are sets of k vertices. Thus an ordinary graph is a 2-uniform hypergraph.

The second property justifies the name 2-*core*. Recall from class that a non-empty 2-core corresponds to shattering in the set of solutions.

In the remaining parts, you may assume that $M \sim Poisson(\alpha N)$ instead of $M = \alpha N$, except at the end where you will compare these models.

(b) Consider the variable x_1 and fix a radius r. Viewing G as a random bipartite graph, prove that as $N \to \infty$, the r-neighborhood[4] of x_1 converges in total variation to a probability distribution on rooted trees, and describe the distribution explicitly. (Hint: it may be helpful to construct a coupling step-by-step.)

(c) Explain the equivalence between pruning and the following *belief propagation* recursion: each variable-to-clause edge $e_{i\to a}$ and clause-to-variable edge $e_{a\to i}$ is initialized with a message "PRESENT". At time t, these messages are updated as follows: the message $m^t_{i\to a}$ is "PRESENT" if i received at least one "PRESENT" message $m^{t-1}_{a'\to i}$ for $a' \neq a$ in the previous time-step, else the message $m_{i\to a}$ is "ABSENT". In the other direction, the message $m^t_{a\to i}$ is "ABSENT" if at least 1 "ABSENT" message $m^{t-1}_{i'\to a}$ was sent for $i' \neq i$, otherwise "PRESENT".

(d) Define the sequence $(p_0, p_1, \dots)$ by $p_0 = 1$ and

$$p_{t+1} = 1 - e^{-k\alpha p_t^{k-1}}.$$

Using the limiting tree description from before, prove that for each fixed t, the expected number of "PRESENT" messages $m^t_{i\to a}$ at time t is $k\alpha p_t(1 + o(1))N$ as $N \to \infty$.

(e) Prove that $p_* = \lim_{t\to\infty} p_t$ exists for all (k, α), and that $p_* = 1$ if and only if $x < 1 - e^{-k\alpha x^{k-1}}$ for all $x \in [0, 1]$. In general, show p_* is the smallest solution to $x = 1 - e^{-k\alpha x^{k-1}}$ on $x \in [0, 1]$.

(f) The *dynamical threshold* $\alpha_d(k)$ from class is defined by

$$\sup\{\alpha \ : \ x < 1 - e^{-k\alpha x^{k-1}} \forall x \in [0, 1]\}.$$

Show that when $\alpha < \alpha_d(k)$, for any $\varepsilon > 0$, there exists t such that in the limit $N \to \infty$, the expected proportion of clauses remaining after t steps of pruning is at most ε.[5]

(g) Prove that if $G \subseteq \widetilde{G}$, then the 2-core of G is contained in that of $\widetilde{G}$. Conclude that the 2-core has at most εN clauses with high probability when $M = \lfloor \alpha N \rfloor$ is fixed, for $\alpha < \alpha_d(k)$. (Hint: compare the fixed M and Poissonian M cases for $\alpha < \widetilde{\alpha} < \alpha_d$.)

(h) Show that for $M = \lfloor \alpha N \rfloor$ and small enough ε depending on (k, α), the expected number of subsets of at most $\sqrt{\varepsilon} N$ clauses containing no variable exactly once tends to 0 with N. Conclude that for $\alpha < \alpha_d(k)$, the 2-core is empty with high probability, both for fixed and Poisson M.

[4] That is, the rooted subgraph of G within distance r of x_1, with distinguished root x_1.

[5] In the opposite case $\alpha > \alpha_d(k)$, the size of the 2-core is given by p_* with high probability. This is harder to prove since one needs to handle pruning for a number of steps diverging with N.

Problem ∞. Survey

If you submit an extra credit problem for grading, you are encouraged to rate it in the same way as the official problems.