Finding Hidden Cliques of Size $\sqrt{N/e}$ in Nearly Linear Time

Yash Deshpande* and Andrea Montanari†

April 25, 2013

Abstract

Consider an Erdös-Renyi random graph in which each edge is present independently with probability 1/2, except for a subset C_N of the vertices that form a clique (a completely connected subgraph). We consider the problem of identifying the clique, given a realization of such a random graph.

The best known algorithm provably finds the clique in linear time with high probability, provided $|\mathsf{C}_N| \geq 1.261\sqrt{N}$ [YDP11]. Spectral methods can be shown to fail on cliques smaller than $\sqrt{N}$. In this paper we describe a nearly linear time algorithm that succeeds with high probability for $|\mathsf{C}_N| \geq (1+\varepsilon)\sqrt{N/e}$ for any $\varepsilon > 0$. This is the first algorithm that provably improves over spectral methods.

We further generalize the hidden clique problem to other background graphs (the standard case corresponding to the complete graph on N vertices). For large girth regular graphs of degree $(\Delta+1)$ we prove that 'local' algorithms succeed if $|\mathsf{C}_N| \geq (1+\varepsilon)N/\sqrt{e\Delta}$ and fail if $|\mathsf{C}_N| \leq (1-\varepsilon)N/\sqrt{e\Delta}$.

1 Introduction

Numerous modern data sets have network structure, i.e. the dataset consists of observations on pairwise relationships among a set of N objects. A recurring computational problem in this context is the one of identifying a small subset of 'atypical' observations against a noisy background. This paper develops a new type of algorithm and analysis for this problem. In particular we improve over the best methods for finding a hidden clique in an otherwise random graph.

Let $G_N = ([N], E_N)$ be a graph over the vertex set $[N] \equiv \{1, 2, \dots, N\}$ and Q_0, Q_1 be two distinct probability distributions over the real line $\mathbb{R}$. Finally, let $\mathsf{C}_N \subseteq [N]$ be a subset of vertices uniformly random given its size $|\mathsf{C}_N|$. For each edge $(i,j) \in E_N$ we draw an independent random variable W_{ij} with distribution $W_{ij} \sim Q_1$ if both $i \in \mathsf{C}_N$ and $j \in \mathsf{C}_N$ and $W_{ij} \sim Q_0$ otherwise. The *hidden set problem* is to identify the set C_N given knowledge of the graph G_N and the observations $W = (W_{ij})_{(ij)\in E_N}$. We will refer to G_N as to the *background graph*. We emphasize that G_N is non-random and that it carries no information about the hidden set C_N.

In the rest of this introduction we will assume, for simplicity, $Q_1 = \delta_{+1}$ and $Q_0 = (1/2)\delta_{+1} + (1/2)\delta_{-1}$. In other words, edges $(i,j) \in E_N$ with endpoints $\{i,j\} \subseteq \mathsf{C}_N$ are labeled with $W_{ij} = +1$. Other edges $(i,j) \in E_N$ have a uniformly random label

*Y. Deshpande is with the Department of Electrical Engineering, Stanford University

†A. Montanari is with the Departments of Electrical Engineering and Statistics, Stanford University

$W_{ij} \in \{+1, -1\}$. Our general treatment in the next sections covers arbitrary subgaussian distributions Q_0 and Q_1 and does not require these distributions to be known in advance.

The special case $G_N = K_N$ (with K_N the complete graph) has attracted considerable attention over the last twenty years [Jer92] and is known as as the *hidden* or *planted clique problem*. In this case, the background graph does not play any role, and the random variables $W = (W_{ij})_{i,j\in[N]}$ can be organized in an $N \times N$ symmetric matrix (letting, by convention, $W_{ii} = 0$). The matrix W can be interpreted as the adjacency matrix of a random graph $\mathcal{R}_N$ generated as follows. Any pair of vertices $\{i, j\} \subseteq \mathsf{C}_N$ is connected by an edge. Any other pair $\{i, j\} \not\subseteq \mathsf{C}_N$ is instead connected independently with probability 1/2. (We use here $\{+1, -1\}$ instead of $\{1, 0\}$ for the entries of the adjacency matrix. This encoding is unconventional but turns out to be mathematically convenient.) Due to the symmetry of the model, the set $\mathsf{C}_N \subseteq [N]$ does not need to be random and can be chosen arbitrarily in this case.

It is easy to see that, allowing for exhaustive search, the hidden clique can be found with high probability as soon as $|\mathsf{C}_N| \geq 2(1+\varepsilon)\log_2 N$ for any $\varepsilon > 0$. This procedure has complexity $\exp[\Theta((\log N)^2)]$. Viceversa, if $|\mathsf{C}_N| \leq 2(1-\varepsilon)\log_2 N$, then the clique cannot be uniquely identified.

Despite a large body of research, the best polynomial-time algorithms to date require $|\mathsf{C}_N| \geq c\sqrt{N}$ to succeed with high probability. This was first achieved by Alon, Krivelevich and Sudakov [AKS98] through a spectral technique. It is useful to briefly discuss this class of methods and their limitations. Letting $u_{\mathsf{C}_N} \in \mathbb{R}^N$ be the indicator vector on C_N (i.e. the vector with entries $(u_{\mathsf{C}_N})_i = 1$ for $i \in \mathsf{C}_N$ and $= 0$ otherwise), we have

$$W = u_{\mathsf{C}_N} u_{\mathsf{C}_N}^{\mathsf{T}} + Z - Z_{\mathsf{C}_N,\mathsf{C}_N}\,. \tag{1.1}$$

Here $Z \in \mathbb{R}^{N\times N}$ is a symmetric matrix with i.i.d. entries $(Z_{ij})_{i<j}$ uniformly random in $\{+1, -1\}$ and $Z_{\mathsf{C}_N,\mathsf{C}_N}$ is the matrix obtained by zeroing all the entries Z_{ij} with $\{i, j\} \not\subseteq \mathsf{C}_N$. Denoting by $\|A\|_2$ the ℓ_2 operator norm of matrix A, we have $\|u_{\mathsf{C}_N} u_{\mathsf{C}_N}^{\mathsf{T}}\|_2 = \|u_{\mathsf{C}_N}\|_2^2 = |\mathsf{C}_N|$. On the other hand, a classical result by Füredi and Komlös [FK81] implies that, with high probability, $\|Z\|_2 \leq c_0\sqrt{N}$ and $\|Z_{\mathsf{C}_N,\mathsf{C}_N}\|_2 \leq c_0\sqrt{|\mathsf{C}_N|}$. Hence, if $|\mathsf{C}_N| \geq c\sqrt{N}$ with c large enough, the first term in the decomposition (1.1) dominates the others. By a standard matrix perturbation argument [DK70], letting v_1 denote the principal eigenvector of W, we have $\|v_1 - u_{\mathsf{C}_N}/\sqrt{|\mathsf{C}_N|}\|_2 \leq \varepsilon$ provided the constant $c = c(\varepsilon)$ is chosen large enough. It follows that selecting the $|\mathsf{C}_N|$ largest entries of v_1 yields an estimate $\widehat{\mathsf{C}}_N \subset [N]$ that includes at least half of the vertices of C_N: the other half can be subsequently identified through a simple procedure [AKS98].

The spectral approach does not exploit the fact that $|\mathsf{C}_N|$ is much smaller than N or –in other words– the fact that u_{C_N} is a *sparse* vector. Recent results in random matrix theory suggest that it is unlikely that the same approach can be pushed to work for $|\mathsf{C}_N| \leq (1-\varepsilon)\sqrt{N}$ (for any $\varepsilon > 0$). For instance the following is a consequence of [KY11, Theorem 2.7]. (The proof is provided in Appendix B.1)

Proposition 1.1. *Let $e_{\mathsf{C}_N} = u_{\mathsf{C}_N}/N^{1/4}$ be the normalized indicator vector on the vertex set C_N, and Z a Wigner random matrix with subgaussian entries such that $E\{Z_{ij}\} = 0$, $\mathbb{E}\{Z_{ij}^2\} = 1/N$ Denote by $v_1, v_2, v_3, \dots, v_\ell$ the eigenvectors of $W = u_{\mathsf{C}_N} u_{\mathsf{C}_N}^{\mathsf{T}} + Z$, corresponding to the ℓ largest eigenvalues.*

Assume $|\mathsf{C}_N| \geq (1+\varepsilon)\sqrt{N}$ for some $\varepsilon > 0$. Then, with high probability, $\langle v_1, e_{\mathsf{C}_N}\rangle \geq \min(\sqrt{\varepsilon}, \varepsilon)/2$. Viceversa, assume $|\mathsf{C}_N| \leq (1-\varepsilon)\sqrt{N}$. Then, with high probability for any fixed constant $\delta > 0$, $|\langle v_i, e_{\mathsf{C}_N}\rangle| \leq c\, N^{-1/2+\delta}$ for all $i \in \{1, \dots, \ell\}$ and some $c = c(\varepsilon, \ell)$.

In other words, for $|\mathsf{C}_N|$ below $\sqrt{N}$ and any fixed ℓ, the first ℓ principal eigenvectors of W are essentially no more correlated with the set C_N than a random unit vector. A natural reaction to this limitation is to try to exploit the sparsity of u_{C_N}. Ames and Vavasis [AV11] studied a convex optimization formulation wherein W is approximated by a sparse low-rank matrix. These two objectives (sparsity and rank) are convexified through the usual ℓ_1-norm and nuclear-norm relaxations. These authors prove that this convex relaxation approach is successful with high probability, provided $|\mathsf{C}_N| \geq c\sqrt{N}$ for an unspecified constant c. A similar result follows from the robust PCA analysis of Candés, Li, Ma, and Wright [CLMW11].

Dekel, Gurel-Gurevich and Peres [YDP11] developed simple iterative schemes with $O(N^2)$ complexity (see also [FR10] for similar approaches). For the best of their algorithms, these authors prove that it succeeds with high probability provided $|\mathsf{C}_N| \geq 1.261\sqrt{N}$. Finally, [AKS98] also provide a simple procedure that, given an algorithm that is successful for $|\mathsf{C}_N| \geq c\sqrt{N}$ produces an algorithm that is successful for $|\mathsf{C}_N| \geq c\sqrt{N/2}$, albeit with complexity $\sqrt{N}$ times larger.

Our first result proves that the hidden clique can be identified in nearly linear time well below the spectral threshold $\sqrt{N}$, see Proposition 1.1.

Theorem 1. *Assume $|\mathsf{C}_N| \geq (1+\varepsilon)\sqrt{N/e}$, for some $\varepsilon > 0$ independent of N. Then there exists a $O(N^2 \log N)$ time algorithm that identifies the hidden clique C_N with high probability.*

In Section 2 we will state and prove a generalization of this theorem for arbitrary –not necessarily known– distributions Q_0, Q_1.

Our algorithm is based on a quite different philosophy with respect to previous approaches to the same problem. We aim at estimating optimally the set C_N by computing the posterior probability that $i \in \mathsf{C}_N$, given edge data W. This is, in general, #P-hard and possibly infeasible if Q_0, Q_1 are unknown. We therefore consider an algorithm derived from *belief propagation*, a heuristic machine learning method for approximating posterior probabilities in graphical models. We develop a rigorous analysis of this algorithm that is asymptotically exact as $N \to \infty$, and prove that indeed the algorithm converges to the correct set of vertices C_N for $|\mathsf{C}_N| \geq (1+\varepsilon)\sqrt{N/e}$. Viceversa, the algorithm converges to an uninformative fixed point for $|\mathsf{C}_N| \leq (1-\varepsilon)\sqrt{N/e}$.

Given Theorem 1, it is natural to ask whether the threshold $\sqrt{N/e}$ has a fundamental computational meaning or is instead only relevant for our specific algorithm. Recently, [FGR$^+$12] proved complexity lower bounds for the hidden clique model, in a somewhat different framework. In the formulation of [FGR$^+$12], one can query columns of W and a new realization from the distribution of W given C_N is instantiated at each query. Assuming that each column is queried $O(1)$ times, their lower bound would require $|\mathsf{C}_N| \geq N^{1/2-\varepsilon}$. While this analysis can possibly be adapted to our setting, it is unlikely to yield a lower bound of the form $|\mathsf{C}_N| \geq c\sqrt{N}$ with a sharp constant c.

Instead, we take a different point of view, and consider the hidden set problem on a *general background graph* G_N. Let us emphasize once more that G_N is non random and that all the information about the hidden set is carried by the edge labels $W = (W_{lk})_{(l,k)\in E_N}$. In addition, we attach to the edges a collection of independent labels $U = (U_{lk})_{(l,k)\in E_N}$ i.i.d. and uniform in $[0,1]$. The U labels exist to provide for (possible) randomization in the algorithm. Given such a graph G_N with labels W, U, a vertex $i \in [N]$ and $t \geq 0$, we let $\mathsf{Ball}_{G_N}(i;t)$ denote the subgraph of G_N induced by those vertices $j \in [N]$ whose graph

distance from i is at most t. We regard $\mathsf{Ball}_{G_N}(i;t)$ as a graph rooted at i, with edge labels W_{jl}, U_{jl} inherited from G_N.

Definition 1.2. *An algorithm for the hidden set problem is said to be* t-local *if, denoting by* $\widehat{\mathsf{C}}_N$ *its output, the membership* ($i \in \widehat{\mathsf{C}}_N$) *is a function of the neighborhood* $\mathsf{Ball}_{G_N}(i;t)$. *We say that it is* local *if it is* t*-local for some* t *independent of* N.

The concept of (randomized) local algorithms was introduced in [Ang80] and formalizes the notion of an algorithm that can be run in $O(1)$ time in a distributed network. We refer to [Lin92, NS95] for earlier contributions, and to [Suo13] for a recent survey.

We say that a sequence of graphs $\{G_N\}_{N\geq 1}$ is *locally tree-like* if, for any $t \geq 0$, the fraction of vertices $i \in [N]$ such that $\mathsf{Ball}_{G_N}(i;t)$ is a tree converges to one as $N \to \infty$. As a standard example, random regular graphs are locally tree-like. The next result is proved in Section 4.

Theorem 2. *Let* $\{G_N\}_{N\geq 1}$ *be a sequence of locally tree-like graphs, with regular degree* $(\Delta+1)$, *and let* $\mathsf{C}_N \subseteq [N]$ *be a uniformly random subset of the vertices of given size* $|\mathsf{C}_N|$. *If* $|\mathsf{C}_N| \leq (1-\varepsilon)N/\sqrt{e\Delta}$ *for some* $\varepsilon > 0$, *there exists* $\xi > 0$ *independent of* Δ *and* ε *such that any local algorithm outputs a set of vertices* $\widehat{\mathsf{C}}_N$ *with* $\mathbb{E}[|\mathsf{C}_N \triangle \widehat{\mathsf{C}}_N|] \geq N\xi/\sqrt{\Delta}$ *for all* N *large enough.*

Viceversa, if $|\mathsf{C}_N| \geq (1+\varepsilon)N/\sqrt{e\Delta}$ *for some* $\varepsilon > 0$, *there exists* $\overline{\xi}(\varepsilon) > 0$ *and a local algorithm that outputs a set of vertices* $\widehat{\mathsf{C}}_N$ *satisfying* $\mathbb{E}[|\mathsf{C}_N \triangle \widehat{\mathsf{C}}_N|] \leq N\exp(-\overline{\xi}(\varepsilon)\sqrt{\Delta})$ *for all* N *large enough.*

Notice that, on a bounded degree graph, the hidden set C_N can not be identified exactly with high probability. Indeed we would not be able to assign a single vertex i with high probability of success, even if we knew exactly the status of all of its neighbors. On the other hand, purely random guessing yields $\mathbb{E}[|\mathsf{C}_N \triangle \widehat{\mathsf{C}}_N|] = N\Theta(1/\sqrt{\Delta})$. The last theorem, thus, establishes a threshold behavior: local algorithms can reconstruct the hidden set with small error if and only if $|\mathsf{C}_N|$ is larger than $N/\sqrt{e\Delta}$.

Unfortunately Theorem 2 only covers the case of sparse or locally tree-like graphs. We let $N \to \infty$ at Δ fixed and then take Δ arbitrarily large. However, if we naively apply it to the case of complete background graphs $G_N = K_N$, by setting $\Delta = N-2$, we get a threshold at $|\mathsf{C}_N| \approx \sqrt{N/e}$ which coincides with the one in Theorem 1. This suggests that $\sqrt{N/e}$ might be a fundamental limit for solving the hidden clique problem in nearly linear time. It would be of much interest to clarify whether this is indeed the case.

The contributions of this paper can be summarized as follows:

1. We develop a new algorithm based on the belief propagation heuristic in machine learning, that applies to the general hidden set problem.

2. We establish a sharp analysis of the algorithm evolution, rigorously establishing that it can be used to find hidden cliques of size $\sqrt{N/e}$ in random graphs. The analysis applies to more general noise models as well.

3. We generalize the hidden set problem to arbitrary graphs. For locally tree-like graphs of degree $(\Delta+1)$, we prove that local algorithms succeed in finding the hidden set (up to small errors) if and only if its size is larger than $N/\sqrt{e\Delta}$.

The complete graph case is treated in Section 2, with technical proofs deferred to Section 3. The locally tree-like case is instead discussed in Section 4 with proofs in Section 5.

1.1 Further related work

A rich line of research in statistics addresses the problem of identifying the non-zero entries in a sparse vector (or matrix) x from observations $W = x + Z$ where Z has typically i.i.d. standard Gaussian entries. In particular [ACDH05, ABBDL10, ACCD11, BDN12] study cases in which the sparsity pattern of x is 'structured'. For instance, we can take $x \in \mathbb{R}^{N\times N}$ a matrix with $x_{ij} = \mu$ if $\{i, j\} \subseteq \mathsf{C}_N$ and $x_{ij} = 0$ otherwise. This fits the framework studied in this paper, for G_N the complete graph and $Q_0 = \mathsf{N}(0, 1)$, $Q_1 = \mathsf{N}(\mu, 1)$. This literature however disregards computational considerations. Greedy search methods were developed in several papers, see e.g. [SN08, SWPN09].

Also, the decomposition (1.1) indicates a connection with sparse principal component analysis [ZHT06, JL09, dEGJL07, dBG08]. This is the problem of finding a sparse low-rank approximation of a given data matrix W. Remarkably, even for sparse PCA, there is a large gap between what is statistically feasible and what is achievable by practical algorithms. Berthet and Rigollet [BR13] recently investigated the implications of the assumption that hidden clique is hard to solve for $|\mathsf{C}_N = o(\sqrt{N})$ on sparse PCA.

The algorithm we introduce for the case $G_N = K_N$ is analogous to the 'linearized BP' algorithm of [MT06, GW06], and to the approximate message passing (AMP) algorithm of [DMM09, BM11, BLM12]. These ideas have been applied to low-rank approximation in [RF12]. The present setting poses however several technical challenges with respect to earlier work in this area: (*i*) The entries of the data matrix are not i.i.d.; (*ii*) They are non-Gaussian with –in general– non-zero mean; (*iii*) We seek exact recovery instead of estimation; (*iv*) The sparsity set C_N to be reconstructed scales sublinearly with N.

Finally, let us mention that a substantial literature studies the behavior of message passing algorithms on sparse random graphs [RU08, MM09]. In this paper, a large part of our technical effort is instead devoted to a similar analysis on the complete graph, in which simple local convergence arguments fail.

1.2 Notations

Throughout the paper, $[M] = \{1, 2, \dots, M\}$ denotes the set of first M integers. We employ a slight abuse of notation to write $[N]\backslash i, j$ for $[N]\backslash\{i, j\}$. The indicator function is denoted by $\mathbb{I}(\,\cdot\,)$.

We write $X \sim P$ when a random variable X has a distribution P. We will sometimes write $\mathbb{E}_P$ to denote expectation with respect to the probability distribution P. Probability and expectation will otherwise be denoted by $\mathbb{P}$ and $\mathbb{E}$. For $a \in \mathbb{R}$, $b \in \mathbb{R}_+$, $\mathsf{N}(a, b)$ denotes the Gaussian distribution with mean a and variance b. The cumulative distribution function of a standard Gaussian will be denoted by $\Phi(x) \equiv \int_{-\infty}^{x} e^{-z^2/2}\mathrm{d}z/\sqrt{2\pi}$.

Unless otherwise specified, we assume all edges in the graphs mentioned are undirected. We denote by ∂i the neighborhood of vertex i in a graph.

We will often use the phrase "for $i \in \mathsf{C}_N$" when stating certain results. More precisely, this means that for each N we are choosing an index $i_N \in \mathsf{C}_N$, which does not depend on the edge labels W.

We use c, c_0, c_1, ... and $C_1, C_2, \dots$ to denote constants independent of N and $|\mathsf{C}_N|$.

Throughout, for any random variable Z we will indicate by P_Z its law.

2 The complete graph case: Algorithm and analysis

In this section we consider the case in which the background graph is complete, i.e. $G_N = K_N$. Since G_N does not play any role in this case, we shall omit all reference to it. We will discuss the reconstruction algorithm and its analysis, and finally state a generalization of Theorem 1 to the case of general distributions Q_0, Q_1.

2.1 Definitions

In the present case the data consists of a symmetric matrix $W \in \mathbb{R}^{N\times N}$, with $(W_{ij})_{i<j}$ generated independently as follows. For an unknown set $\mathsf{C}_N \subset [N]$ we have $W_{ij} \sim Q_1$ if $\{i,j\} \subseteq \mathsf{C}_N$, and $W_{ij} \sim Q_0$ otherwise. Here Q_1 and Q_0 are distinct probability measures. We make the following assumptions:

I. Q_0 has zero mean and Q_1 has non-zero mean λ. Without loss of generality we shall further assume that Q_0 has unit variance, and that $\lambda > 0$.

II. Q_0 and Q_1 are subgaussian with common scale factor ρ.

It will be clear from the algorithm description that there is indeed no loss in generality in assuming that Q_0 has unit variance and that λ is positive. Recall that a probability distribution P is subgaussian with scale factor $\rho > 0$ if for all $y \in \mathbb{R}$ we have:

$$\mathbb{E}_P\left(e^{y(X-\mathbb{E}_P X)}\right) \le e^{\rho y^2/2}.$$

There is no loss of generality in assuming a common scale factor for Q_0 and Q_1.

The task is to identify the set C_N from a realization of the matrix W. As discussed in the introduction, the relevant scaling is $|\mathsf{C}_N| = \Theta(\sqrt{N})$ and we shall therefore define $\kappa_N \equiv |\mathsf{C}_N|/\sqrt{N}$. Further, throughout this section, we will make use of the normalized matrix

$$A \equiv \frac{1}{\sqrt{N}} W\,. \tag{2.1}$$

In several technical steps of our analysis we shall consider a sequence of instances $\{(W_{N\times N}, \mathsf{C}_N)\}_{N\ge 1}$ indexed by the dimension N, such that $\lim_{N\to\infty}\kappa_N = \kappa \in (0,\infty)$. This technical assumption will be removed in the proof of our main theorem.

2.2 Message passing and state evolution

The key innovation of our approach is the construction and analysis of a message passing algorithm that allows us to identify the hidden set C_N. As we demonstrate in Section 4, this algorithm can be derived from *belief propagation* in machine learning. However this derivation is not necessary and the treatment here will be self-contained.

The message passing algorithm is iterative and at each step $t \in \{1,2,3,\dots\}$ produces an $N \times N$ matrix θ^t whose entry (i,j) will be denoted as $\theta^t_{i\to j}$ to emphasize the fact that θ^t is not symmetric. By convention, we set $\theta^t_{i\to i} = 0$. The variables $\theta^t_{i\to j}$ will be referred to as *messages*, and their update rule is formally defined below.

Definition 2.1. *Let $\theta^0 \in \mathbb{R}^{N\times N}$ be an initial condition for the messages and, for each t, let $f(\,\cdot\,;t) : \mathbb{R} \to \mathbb{R}$ be a scalar function. The* message passing orbit *corresponding to the triple (A, f, θ^0) is the sequence of $\{\theta^t\}_{t\ge 0}$, $\theta^t \in \mathbb{R}^{N\times N}$ defined by letting, for each $t \ge 0$:*

$$\theta^{t+1}_{i\to j} = \sum_{\ell\in[N]\setminus i,j} A_{\ell i} f(\theta^t_{\ell\to i}, t)\,, \quad \forall\, j \neq i \in [N]\,. \tag{2.2}$$

We also define a sequence of vectors $\{\theta^t\}_{t\ge 1}$ with $\theta^t = (\theta^t_i)_{i\in[N]} \in \mathbb{R}^N$, by letting (the entries of θ^t being indexed by $i \in [N]$) given by:

$$\theta^{t+1}_i = \sum_{\ell\in[N]\setminus i} A_{\ell i} f(\theta^t_{\ell\to i}, t). \tag{2.3}$$

The functions $f(\,\cdot\,,t)$ will be chosen so that they can be evaluated in $O(1)$ operations. Each iteration can be implemented with $O(N^2)$ operations. Indeed $(\theta^{t+1}_i)_{i\in[N]}$ can be computed in $O(N^2)$ as per Eq. (2.3). Subsequently we can compute $(\theta^{t+1}_{i\to j})_{i,j\in[N]}$ in $O(N^2)$ operations by noting that $\theta^{t+1}_{i\to j} = \theta^{t+1}_i - A_{ij} f(\theta^t_{j\to i}, t)$.

The proper choice of the functions $f(\,\cdot\,,t)$ plays a crucial role in the achieving the claimed tradeoff between $|\mathsf{C}_N|$ and N. This choice will be optimized on the basis of the general analysis developed below.

Before proceeding, it is useful to discuss briefly the intuition behind the update rule introduced in Definition 2.1. For each vertex i, the message $\theta^t_{i\to j}$ and the value θ^t_i are estimates of the likelihood that $i \in \mathsf{C}_N$: they are larger for vertices that are more likely to belong to the set C_N. In order to develop some intuition on Definition 2.1, consider a conceptually simpler iteration operating as follows on variables $\vartheta^t = (\vartheta^t_i)_{i\in[N]}$. For each $i \in [N]$ we let $\vartheta^{t+1}_i = \sum_{j\in[N]} A_{ij} f(\vartheta^t_i; t)$. In the special case $f(\vartheta; t) = \vartheta$ we obtain the iteration $\vartheta^{t+1} = A\,\vartheta^t$ which is simply the power method for computing the principal eigenvector of A. As discussed in the introduction, this does not use in any way the information that $|\mathsf{C}_N|$ is much smaller than N. We can exploit this information by taking $f(\vartheta; t)$ a rapidly increasing function of ϑ that effectively selects the vertices $i \in [N]$ with ϑ^t_i large. We will see that this is indeed what happens within our analysis.

An important feature of the message passing version (operating on messages $\theta^t_{i\to j}$) is that it admits a characterization that is asymptotically exact as $N \to \infty$. In the large N limit, the messages θ^t_i (for fixed t) converge in distribution to Gaussian random variables with certain mean and variance. In order to state this result formally, we introduce the sequence of mean and variance parameters $\{(\mu_t, \tau_t^2)\}_{t\ge 0}$ by letting $\mu_0 = 1$, $\tau_0^2 = 0$ and defining , for $t \ge 0$,

$$\mu_{t+1} = \lambda\kappa\, \mathbb{E}[f(\mu_t + \tau_t\, Z, t)] \tag{2.4}$$

$$\tau^2_{t+1} = \mathbb{E}[f(\tau_t Z, t)^2], \tag{2.5}$$

Here expectation is with respect to $Z \sim \mathsf{N}(0, 1)$. We will refer to this recursion as to *state evolution*.

Lemma 2.2. *Let $f(u, t)$ be, for each $t \in \mathbb{N}$ a finite-degree polynomial. For each N, let $W \in \mathbb{R}^{N\times N}$ be a symmetric matrix distributed as per the model introduced above with $\kappa_N \equiv |\mathsf{C}_N|/\sqrt{N} \to \kappa \in (0, \infty)$. Set $\theta^0_{i\to j} = 1$ and denote the associated message passing orbit by $\{\theta^t\}_{t\ge 0}$.*

Then, for any bounded Lipschitz function $\psi : \mathbb{R} \mapsto \mathbb{R}$, the following limits hold in probability:

$$\lim_{N\to\infty} \frac{1}{|\mathsf{C}_N|} \sum_{i\in\mathsf{C}_N} \psi(\theta_i^t) = \mathbb{E}[\psi(\mu_t + \tau_t\, Z)] \tag{2.6}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i\in[N]\setminus\mathsf{C}_N} \psi(\theta_i^t) = \mathbb{E}[\psi(\tau_t\, Z)]. \tag{2.7}$$

Here expectation is with respect to $Z \sim \mathsf{N}(0,1)$ where μ_t, τ_t^2 are given by the recursion in Eqs. (2.4),(2.5).

The proof of this Lemma is deferred to Section 3.1. Naively, one would like to use the central limit theorem to approximate the distribution on the right-hand side of Eq. (2.2) or of Eq. (2.3) by a Gaussian. This is, however, incorrect because the messages $\theta^t_{\ell\to i}$ depend on the matrix A and hence the summands are not independent. In fact, the lemma would be false if we did not use the edge messages and replaced $\theta^t_{\ell\to i}$ by θ^t_ℓ in Eq. (2.2) or in Eq. (2.3).

However, for the iteration Eq. (2.2), we prove that the distribution of θ^t is approximately the same that we would obtain by using a fresh independent copy of A (given C_N) at each iteration. The central limit theorem then can be applied to this modified iteration. In order to prove that this approximation, we use the moment method, representing $\theta^t_{i\to j}$ and θ^t_i as polynomials in the entries of A. We then show that the only terms that survive in these polynomials as $N \to \infty$ are the monomials which are of degree 0, 1 or 2 in each entry of A.

2.3 Analysis of state evolution

Lemma 2.2 implies that the distribution of θ^t_i is very different depending whether $i \in \mathsf{C}_N$ or not. If $i \in \mathsf{C}_N$ then θ^t_i is approximately $\mathsf{N}(0, \tau_t^2)$. If instead $i \in \mathsf{C}_N$ then θ^t_i is approximately $\mathsf{N}(\mu_t, \tau_t^2)$.

Assume that, for some choice of the functions $f(\cdot,\cdot)$ and some t, μ_t is positive and much larger than τ_t. We can then hope to estimate C_N by selecting the indices i such that θ^t_i is above a certain threshold[1]. This motivates the following result.

Lemma 2.3. *Assume that $\lambda\kappa > e^{-1/2}$. Inductively define:*

$$p(z,\ell) = \frac{1}{\hat{L}_\ell} \sum_{k=0}^{d^*} \frac{\hat{\mu}_\ell^k z^k}{k!}\,, \qquad \hat{\mu}_{\ell+1} = \mathbb{E}[p(\hat{\mu}_\ell + Z, \ell)], \tag{2.8}$$

where $Z \sim \mathsf{N}(0,1)$ and the recursion is initialized with $p(z,0) = 1$. Here $\hat{L}_\ell$ is a normalization defined, for all $\ell \geq 1$, by $\hat{L}_\ell^2 = \mathbb{E}\Big[\Big(\sum_{k=0}^{d^*}(\hat{\mu}_\ell Z)^k/k!\Big)^2\Big]$.

Then, for any M finite there exists d^*, t^* finite such that $\hat{\mu}_{t^*} > M$.

By setting $f(\,\cdot\,,t) = p(\,\cdot\,,t)$ in the state evolution equations (2.4) *and* (2.5) *we obtain $\mu_t = \hat{\mu}_t$ and $\tau_t = 1$ for all t.*

The proof of this lemma is deferred to Section 3. Also, the proof clarifies that setting $f(\,\cdot\,,t) = p(\,\cdot\,,t)$ is the optimal choice for our message passing algorithm.

[1]The problem is somewhat more subtle because $|\mathsf{C}_N| \ll N$, see next section.

The basic intuition is as follows. Consider the state evolution equations (2.4) and (2.5). Since we are only interested in maximizing the signal-to-noise ratio μ_t/τ_t we can always normalize $f(\,\cdot\,,t)$ as to have $\tau_{t+1} = 1$. Denoting by $g(\,\cdot\,,t)$ the un-normalized function, we thus have the recursion

$$\mu_{t+1} = \lambda\kappa \frac{\mathbb{E}[g(\mu_t + Z, t)]}{\mathbb{E}[g(Z,t)^2]^{1/2}}\,.$$

We want to choose $g(\,\cdot\, t)$ as to maximize the right-hand side. It is a simple exercise of calculus to show that this happens for $g(z,t) = e^{\mu_t z}$. For this choice we obtain the iteration $\mu_{t+1} = \lambda\kappa\, e^{\mu_t^2/2}$ that diverges to $+\infty$ if and only if $\lambda\kappa > e^{-1/2}$. Unfortunately the resulting f is not a polynomial and is therefore not covered by Lemma 2.2. Lemma 2.3 deals with this problem by approximating the function $e^{\mu_t z}$ with a polynomial.

2.4 The whole algorithm and general result

As discussed above, after t iterations of the message passing algorithm we obtain a vector $(\theta_i^t)_{i\in[N]}$ wherein for each i, θ_i^t estimates the likelihood that $i \in \mathsf{C}_N$. We can therefore select a 'candidate' subset for C_N, by letting $\widetilde{\mathsf{C}}_N \equiv \{i \in [N] :\ \theta_i^t \ge \mu_t/2\}$ (this choice is motivated by the analysis of the previous section). Since however θ_i^t is approximately $\mathsf{N}(0,\tau_t^2)$ for $i \in [N]\setminus \mathsf{C}_N$, this produces a set of size $|\widetilde{\mathsf{C}}_N| = \Theta(N)$, much larger than the target C_N.

Algorithm 1 Message Passing

1: Initialize: $A(N) = W(N)/\sqrt{N}$; $\theta_i^0 = 1$ for each $i \in [N]$; d^*, t^* positive integers, $\bar\rho$ a positive constant.
2: Define the sequence of polynomials $p(\,\cdot\,,t)$ for $t \in \{0,1,\dots\}$, the values $\hat\mu_t$ as per Lemma 2.3
3: Run t^* iterations of message passing as in Eqs. (2.2), (2.3) with $f(\,\cdot\,,t) = p(\,\cdot\,,t)$
4: Find the set $\widetilde{\mathsf{C}}_N = \{i \in [N] : \theta_i^{t^*} \ge \hat\mu_{t^*}/2\}$.
5: Let $A|_{\widetilde{\mathsf{C}}_N}$ be the restriction of A to the rows and columns with index in $\tilde{\mathsf{C}}_N$, and compute by power method its principal eigenvector u^{**}.
6: Compute $\mathsf{B}_N \subseteq [N]$ of the top $|\mathsf{C}_N|$ entries (by absolute value) of u^{**}.
7: Return $\widehat{\mathsf{C}}_N = \{i \in [N] : \zeta_{\bar\rho}^{\mathsf{B}_N}(i) \ge \lambda/2\}$.

In order to overcome this problem, we apply a *cleaning* procedure to reconstruct C_N from $\widetilde{\mathsf{C}}_N$. Let $A|_{\widetilde{\mathsf{C}}_N}$ be the restriction of A to the rows and columns with index in $\widetilde{\mathsf{C}}_N$. By power iteration (i.e. by the iteration $u^{t+1} = A|_{\widetilde{\mathsf{C}}_N} u^t/\|A|_{\widetilde{\mathsf{C}}_N} u^t\|_2$, $u^t \in \mathbb{R}^{\widetilde{\mathsf{C}}_N}$, with $u^0 = (1,1,\dots,1)^{\mathsf{T}}$) we compute a good approximation $u^{**} \equiv u^{t_{**}}$ of the principal eigenvector of $A|_{\widetilde{\mathsf{C}}_N}$. We then let $\mathsf{B}_N \subseteq [N]$, $|\mathsf{B}_N| = |\mathsf{C}_N|$ be the set of indices corresponding to the $|\mathsf{C}_N|$ largest entries of $u^{t_{**}}$ (in absolute value).

The set B_N has the right size and is approximately equal to C_N. We correct the residual 'mistakes' by defining the following score for each vertex $i \in [N]$:

$$\zeta_{\bar\rho}^{\mathsf{B}_N}(i) = \sum_{j\in\mathsf{B}_N} W_{ij}\mathbb{I}_{\{|W_{ij}|\le\bar\rho\}}, \tag{2.9}$$

and returning the set $\widehat{\mathsf{C}}_N$ of vertices with large scores, e.g. $\widehat{\mathsf{C}}_N = \{i \in [N] :\ \zeta_{\bar\rho}^{\mathsf{B}_N}(i) \ge \lambda|\mathsf{B}_N|/2\}$.

Note that the 'cleaning' procedure is similar to the algorithm of [AKS98]. The analysis is however more challenging because we need to start from a set $\widetilde{\mathsf{C}}_N$ that is correlated with the matrix A.

Lemma 2.4. *Let $A = W/\sqrt{N}$ be defined as above and $\widetilde{\mathsf{C}}_N \subseteq [N]$ be any subset of the column indices (possibly dependent on A). Assume that it satisfies, for ε small enough, $|\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N| \geq (1-\varepsilon)|\mathsf{C}_N|$ and $|\widetilde{\mathsf{C}}_N \backslash \mathsf{C}_N| \leq \varepsilon |[N] \backslash \mathsf{C}_N|$.*

Then there exists $t_{**} = O(\log N)$ (number of iterations in the power method) such that the cleaning procedure *gives $\widehat{\mathsf{C}}_N = \mathsf{C}_N$ with high probability.*

The proof of this lemma can be found in Section 3 and uses large deviation bounds on the principal eigenvalue of $A|_{\mathsf{C}_N}$.

The entire algorithm is summarized in Table 1. Notice that the power method has complexity $O(N^2)$ per iteration and since we only execute $O(\log N)$ iterations, its overall complexity is $O(N^2 \log N)$. Finally the scores (2.9) can also be computed in $O(N^2)$ operations. Our analysis of the algorithm results in the following main result that generalizes Theorem 1.

Theorem 3. *Consider the hidden set problem on the complete graph $G_N = K_N$, and assume that Q_0 and Q_1 are subgaussian probability distributions with mean, respectively, 0, and $\lambda > 0$. Further assume that Q_0 has unit variance.*

If $\lambda|\mathsf{C}_N| \geq (1+\varepsilon)\sqrt{N/e}$ then there there exists a $\overline{\rho}$, d^* and t^* finite such that Algorithm 1 returns $\widehat{\mathsf{C}}_N = \mathsf{C}_N$ with high probability on input W, with total complexity $O(N^2 \log N)$.

(More explicitly, there exists $\delta(\varepsilon, N)$ with $\lim_{N\to\infty} \delta(\varepsilon, N) = 0$ such that the algorithm succeeds with probability at least $1 - \delta(\varepsilon, N)$.)

Remark 2.5. *The above result can be improved if Q_0 and Q_1 are known by taking a suitable transformation of the entries W_{ij}. In particular, assuming[2] that Q_1 is absolutely continuous with respect to Q_0, the optimal such transformation is obtained by setting*

$$A_{ij} \equiv \frac{1}{\sqrt{N}} \left[\frac{\mathrm{d}Q_1}{\mathrm{d}Q_0}(W_{ij}) - 1 \right] .$$

Here $\mathrm{d}P/\mathrm{d}Q$ denotes the Radón-Nikodym derivative of P with respect to Q. If the resulting A_{ij} is subgaussian with scale ρ/N, then our analysis above applies. Theorem 3 remains unchanged, provided the parameter λ is replaced by the ℓ_2 distance between Q_0 and Q_1:

$$\widetilde{\lambda} \equiv \left\{ \int \left[\frac{\mathrm{d}Q_1}{\mathrm{d}Q_0}(x) - 1 \right]^2 Q_0(\mathrm{d}x) \right\}^{1/2} . \tag{2.10}$$

3 Proof of Theorem 3

In this section we present the proof of Theorem 3 and of the auxiliary Lemmas 2.2, 2.3 and 2.4.

We begin by showing how these technical lemmas imply Theorem 3. First consider a sequence of instances with $\lim_{N\to\infty} |\mathsf{C}_N|/\sqrt{N} = \lim_{N\to\infty} \kappa_N = \kappa$ such that $\kappa\lambda < 1/\sqrt{e}$. We will prove that Algorithm 1 returns $\widehat{\mathsf{C}}_N = \mathsf{C}_N$ with probability converging to one as $N \to \infty$.

[2] If Q_1 is singular with respect to Q_0 the problem is simpler but requires a bit more care.

By Lemma 2.2, we have, in probability

$$\lim_{N\to\infty} \frac{|\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N|}{|\mathsf{C}_N|} = \lim_{N\to\infty} \frac{1}{|\mathsf{C}_N|} \sum_{i\in\mathsf{C}_N} \mathbb{I}(\theta_i^{t^*} \ge \hat{\mu}_{t^*}/2) = \mathbb{E}\{\mathbb{I}(\mu_{t^*} + \tau_t Z \ge \hat{\mu}_{t^*}/2)\}\,.$$

Notice that, in the second step, we applied Lemma 2.2, to the function $\psi(z) \equiv \mathbb{I}(x \ge \hat{\mu}_{t^*}/2)$. While this is not Lipschitz continuous, it can approximated from above and below pointwise by Lipschitz continuous functions. This is sufficient to obtain the claimed convergence as in standard weak convergence arguments [Bil08].

Since we used $f(\,\cdot\,,t) = p(\,\cdot\,,t)$, we have, by Lemma 2.3, $\mu_t = \hat{\mu}_t$ and $\tau_t = 1$. Denoting by $\Phi(z) \equiv \int_{-\infty}^{z} e^{-x^2/2}\mathrm{d}x/\sqrt{2\pi}$ the Gaussian distribution function, we thus have

$$\lim_{N\to\infty} \frac{|\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N|}{|\mathsf{C}_N|} = 1 - \Phi(-\hat{\mu}_t^*/2) \ge 1 - e^{-M^2/8}\,.$$

where in the last step we used $\Phi(-a) \le e^{-a^2/2}$ for $a \ge 0$ and Lemma 2.3. By taking $M^2 \ge 8\log(2/\varepsilon)$ we can ensure that the last expression is larger than $(1-\varepsilon/2)$ and therefore $|\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N| \ge (1-\varepsilon)|\mathsf{C}_N|$ with high probability.

By a similar argument we have, in probability

$$\lim_{N\to\infty} \frac{|\widetilde{\mathsf{C}}_N|}{N} = \Phi(-\hat{\mu}_t^*/2) \le \frac{\varepsilon}{2}\,,$$

and hence $|\widetilde{\mathsf{C}}_N| \le \varepsilon|[N] \setminus \mathsf{C}_N|$ with high probability.

We can therefore apply Lemma 2.4 and conclude that Algorithm 1 succeeds with high probability for $\kappa_N = |\mathsf{C}_N|/\sqrt{N} \to \kappa > 1/\sqrt{\lambda^2 e}$.

In order to complete the proof, we need to prove that the Algorithm 1 succeeds with probability at least $1-\delta(\varepsilon,N)$ for all $|\mathsf{C}_N| \ge (1+\varepsilon)\sqrt{N/(\lambda^2 e)}$. Notice that, without loss of generality we can assume $\kappa_N \in [(1+\varepsilon)/\sqrt{\lambda^2 e}, K]$ with K a large enough constant (because for $\kappa_N > K$ the problem becomes easier and –for instance– the proof of [AKS98] already works). If the claim was false there would be a sequence of values $\{\kappa_N\}_{N\ge 1}$ indexed by N such that the success probability remains bounded away from one along the sequence. But since $[(1+\varepsilon)/\sqrt{\lambda^2 e}, K]$ is compact, this sequence has a converging subsequence along which the success probability remains bounded. This contradicts the above.

3.1 Proof of Lemma 2.2

It is convenient to collate the assumptions we make on our problem instances as follows.

Definition 3.1. *We say $\{A(N), \mathcal{F}_N, \theta_N^0\}_{\{N\ge 1\}}$ is a (C,d)-*regular *sequence if:*

1. *For each N, $A(N) = W_N/\sqrt{N}$ where W_N satisfies Assumption 2.1.*
2. *For each $t \ge 0$, $f(\cdot,t) \in \mathcal{F}_N$ is a polynomial with maximum degree d and coefficients bounded in absolute value by C.*
3. *Each entry of the initial condition θ_N^0 is 1.*

Let $A^t, t \ge 1$ be i.i.d. matrices distributed as A conditional on the set C_N, and let $A^0 \equiv A$. We now define the sequence of $N \times N$ matrices $\{\xi^t\}_{t\ge 0}$ and a sequence of vectors in $\mathbb{R}^N$, $\{\xi^t\}_{t\ge 1}$ (indexed as before) given by:

$$\xi_{i\to j}^{t+1} = \sum_{\ell\in[N]\setminus\{i,j\}} A_{i\ell}^t f(\xi_{\ell\to i}^t, t) \tag{3.1}$$
$$\xi_{i\to j}^0 = \theta_i^0 \quad \forall\, j\neq i\in[N]$$
$$\xi_{i\to i}^t = 0 \quad \forall\, t\geq 0, i\in[N]$$
$$\xi_i^{t+1} = \sum_{\ell\in[N]\setminus i} A_{\ell i}^t f(\xi_{\ell\to i}^t, t) \tag{3.2}$$

The asymptotic marginals of the iterates ξ^t are easier to compute since the matrix A^{t-1} is independent of the ξ^{t-1} by definition. We proceed, hence by proving that ξ^t and θ^t have, asymptotically in N, the same moments of all orders computing the distribution for the ξ^t.

The messages $\theta_{i\to j}^t$ and $\xi_{i\to j}^t$ can be described explicitly via a sum over a family of finite rooted labeled trees. We now describe this family in detail. All edges are assumed directed towards the root. The leaves of the tree are those vertices with no children, and the set of leaves is denoted by $L(T)$. We let $V(T)$ denote the set of vertices of T and $E(T)$ the set of (directed) edges in T. The root has a label in $[N]$ called its "type". Every non-root vertex has a label in $[N]\times\{0,1,\ldots,d\}$, the first argument the label being the "type" of the vertex, and the second being the "mark". For a vertex $v\in T$ we let $l(v)$ denote its type, $r(v)$ its mark and $|v|$ its distance from the root in T.

Definition 3.2. *Let $\mathcal{T}^t$ be the family of labeled trees T with exactly t generations satisfying the conditions:*

1. *The root of T has degree 1.*
2. *Any path $v_1, v_2 \ldots v_k$ in the tree is non-backtracking i.e. the types $l(v_i)$, $l(v_{i+1})$, $l(v_{i+2})$ are distinct.*
3. *For a vertex u that is not the a root or a leaf, the mark $r(u)$ is set to the number of children of v.*
4. *We have that $t = \max_{v\in L(T)}|v|$. All leaves $u\in L(T)$ with non-maximal depth, i.e. $|u|\leq t-1$ have mark 0.*

Let $\mathcal{T}_{i\to j}^t\subset\mathcal{T}^t$ be the subfamily satisfying, in addition, the following:

1. *The type of the root is i.*
2. *The root has a single child with type distinct from i and j.*

In a similar fashion, let $\mathcal{T}_i^t\subset\mathcal{T}^t$ be the subfamily satisfying, additionally:

1. *The type of the root is i.*
2. *The root has a single child with type distinct from i.*

Let the polynomial $f(x,t)$ be represented as:

$$f(x,t) = \sum_{i=0}^{d} q_i^t x^i$$

For a labeled tree $T \in \mathcal{T}^t$ and vector of coefficients $\mathbf{q} = (q_i^s)_{s\leq t, i\leq d}$ we now define three weights:

$$A(T) \equiv \prod_{u\to v\in E(T)} A_{l(u)l(v)} \tag{3.3}$$

$$\Gamma(T,\mathbf{q},t) \equiv \prod_{u\to v\in E(T)} q_{r(u)}^{t-|u|} \tag{3.4}$$

$$\theta(T) \equiv \prod_{u\in L(T)} (\theta^0_{l(u)})^{r(u)} \tag{3.5}$$

We now are in a position to provide an explicit expression for $\theta^t_{i\to j}$ in terms of a summation over an appropriate family of labeled trees.

Lemma 3.3. *Let $\{A(N), \mathcal{F}_N, \theta^0_N\}$ be a (C,d)-regular sequence. The orbit θ^t satisfies:*

$$\theta^t_{i\to j} = \sum_{T\in\mathcal{T}^t_{i\to j}} A(T)\Gamma(T,\mathbf{q},t)\theta(T) \tag{3.6}$$

$$\theta^t_i = \sum_{T\in\mathcal{T}^t_i} A(T)\Gamma(T,\mathbf{q},t)\theta(T) \tag{3.7}$$

Proof. We prove Eq. (3.6) using induction. The proof of Eq. (3.7) is very similar. We have, by definition, that:

$$\theta^1_{i\to j} = \sum_{\ell\in[N]\setminus i,j} \sum_{k\leq d} A_{\ell i} q^0_k (\theta^0_\ell)^k$$

This is what is given by Eq. (3.6) since $\mathcal{T}^1_{i\to j}$ is exactly the set of trees with two vertices joined by a single edge, the root having type i, the other vertex (say v) having type $l(v) \notin \{i,j\}$ and mark $r(v) \leq d$.

Now we assume Eq. (3.6) to be true up to t. For iteration $t+1$, we obtain by definition:

$$\begin{aligned}\theta^{t+1}_{i\to j} &= \sum_{\ell\in[N]\setminus\{i,j\}} A_{\ell i} \sum_{k\leq d} q^t_k (\theta^t_{\ell\to i})^k \\ &= \sum_{\ell\in[N]\setminus\{i,j\}} \sum_{k\leq d} \sum_{T_1\cdots T_k\in\mathcal{T}^t_{\ell\to i}} A_{\ell i} q^t_k \prod_{m=1}^k A(T_m)\Gamma(T_m,\mathbf{q},t)\theta(T_m)\end{aligned}$$

Notice that $\mathcal{T}^{t+1}_{i\to j}$ is in bijection with the set of pairs containing a vertex of type $\ell \notin \{i,j\}$ and a k-tuple of trees belonging to $\mathcal{T}^t_{\ell\to i}$. This is because one can form a tree in $\mathcal{T}^{t+1}_{i\to j}$ by choosing a root with type i, its child v with type $\ell \notin \{i,j\}$ and choosing a $k(\leq d)$-tuple of trees from $\mathcal{T}^t_{\ell\to i}$, identifying their roots with v and setting $r(v) = k$. With this, absorbing the factors of $A_{\ell i}$ into $\prod_{m=1}^k A(T_m)$ and q^t_k into $\prod_{m=1}^k \Gamma(T_m,\mathbf{q},t)$ yields the desired claim. □

From a very similar argument as above we obtain that:

$$\xi^t_{i\to j} = \sum_{T\in\mathcal{T}^t_{i\to j}} \bar{A}(T)\Gamma(T,\mathbf{q},t)\theta(T)$$

$$\xi^t_i = \sum_{T\in\mathcal{T}^t_i} \bar{A}(T)\Gamma(T,\mathbf{q},t)\theta(T)$$

where the weight $\bar{A}(T)$ for a labeled tree T is defined (similar to Eq. (3.3)) by:

$$\bar{A}(T) \equiv \prod_{u \to v \in E(T)} A^{t-|u|}_{l(u)l(v)} \tag{3.8}$$

We now prove that the moments of θ^t_i and ξ^t_i are asymptotically (in the large N limit) the same via the following:

Proposition 3.4. *Let $\{A(N), \mathcal{F}_N, \theta^0_N\}$ be a (C,d)-regular sequence. the conditions above. Then, for any $t \geq 1$, there exists a constant K independent of N (depending possibly on m, t, d, C) such that for any $i \in [N]$:*

$$\left|\mathbb{E}\left[(\theta^t_i)^m\right] - \mathbb{E}\left[(\xi^t_i)^m\right]\right| \leq KN^{-1/2}$$

Proof. According to our initial condition, $\theta^{0,N}$ has all entries 1. Then, using the tree representation we have that:

$$\mathbb{E}\left[(\theta^t_i)^m\right] = \sum_{T_1,\ldots,T_m \in \mathcal{T}^t_i} \left[\prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)\right] \mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] \tag{3.9}$$

$$\mathbb{E}\left[(\xi^t_i)^m\right] = \sum_{T_1,\ldots,T_m \in \mathcal{T}^t_i} \left[\prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)\right] \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] \tag{3.10}$$

Define the multiplicity $\phi(T)_{rs}$ to be the number of occurrences of an edge $u \to v$ in the tree T with types $l(u), l(v) \in \{r, s\}$. Also let $\mathbf{G}$ denote the graph obtained by identifying vertices of the same type in the tuple of trees $T_1, \ldots T_m$. We let $\mathbf{G}|_{\mathsf{C}_N}$ denote its restriction to the vertices in C_N and $\mathbf{G}|_{\mathsf{C}^c_N}$ be the graph restricted to C^c_N. Let $E(\mathbf{G}|_{\mathsf{C}_N})$ and $E(\mathbf{G}|_{\mathsf{C}^c_N})$ denote the (disjoint) edge sets of these graphs and E_J denote the edges in $\mathbf{G}$ not present in either $\mathbf{G}|_{\mathsf{C}_N}$ or $\mathbf{G}|_{\mathsf{C}^c_N}$. In other words, E_J consists of all edges in $\mathbf{G}$ with one endpoint belonging to C_N and one end point outside it. The edge sets here do not count multiplicity.

For analysis, we first split the sum over m-tuples of trees above into three terms as follows:

1. $S(A)$: the sum over all m-tuples of trees $T_1, \ldots, T_m$ such that there exists an edge rs in $E(\mathbf{G}|_{\mathsf{C}^c_N}) \cup E_J$ which is covered at least 3 times.

2. $R(A)$: the sum over all m-tuples of trees such that each edge in $E(\mathbf{G}|_{\mathsf{C}^c_N}) \cup E_J$ is covered either 0 or 2 times, and the graph $\mathbf{G}$ contains a cycle.

3. $T(A)$: the sum over all m-tuples of trees such that each edge in $E(\mathbf{G}|_{\mathsf{C}^c_N}) \cup E_J$ is covered either 0 or 2 times, and the graph $\mathbf{G}$ is a tree.

We also define analogous terms $S(\bar{A})$, $R(\bar{A})$ and $T(\bar{A})$ in the same fashion. We have that $|\prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)| \leq C^{md^{t+1}}$ since the coefficients are bounded by C and the number of edges in the tree by d^{t+1}. We thus concentrate on the portion $\mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right]$. When $\mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] = 0$, some edge in $E(\mathbf{G}|_{\mathsf{C}^c_N}) \cup E_J$ is covered exactly once. This implies $\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] = 0 = \mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right]$, since the same edge is covered only once in any generation. This guarantees that we need only consider the contributions $S(A)$, $R(A)$ and $T(A)$ as above in the sums Eq. (3.9), Eq. (3.10).

We first consider the contribution $S(A)$. We have:

$$\begin{aligned}\mathbb{E}\left[\prod_{\ell=1}^{m} A(T_\ell)\right] &= \mathbb{E}\left[\prod_{j<k}(A_{jk})^{\sum_{\ell=1}^m \phi(T_\ell)_{jk}}\right] \\ &\le \mathbb{E}\left[\prod_{j<k}|A_{jk}|^{\sum_{\ell=1}^m \phi(T_\ell)_{jk}}\right] \\ &= \prod_{j<k}\mathbb{E}\left[|A_{jk}|^{\sum_{\ell=1}^m \phi(T_\ell)_{jk}}\right] \\ &\le C_1\left(\frac{1}{\sqrt{N}}\right)^{\alpha}, \end{aligned} \tag{3.11}$$

where $\alpha = \alpha(T_1, \ldots, T_m)$ is the total number of edges (with multiplicity) in the tuple of trees $T_1, \ldots, T_m$. The last inequality follows from Lemma A.2 and observing that for any j, k, A_{jk} is subgaussian with scale parameter ρ/N. The constant $C_1 = C_1(T_1, \ldots T_m)$ absorbs the leading factors from Lemma A.2, and is independent of N.

To track the dependence on N, note that the graph $\mathbf{G}$ is connected since the roots of all the trees have type i. Let $n(\mathbf{G}|_{\mathsf{C}_N})$ [resp. $n(\mathbf{G}|_{\mathsf{C}_N^c})$] denote the number of vertices in $\mathbf{G}|_{\mathsf{C}_N}$ [resp. $\mathbf{G}|_{\mathsf{C}_N^c}$] not counting the root. Counting the edges with multiplicities, we have $3 + |E(\mathbf{G}|_{\mathsf{C}_N})| + 2(|E(\mathbf{G}|_{\mathsf{C}_N^c})| + |E_J| - 1) \le \alpha$, implying $|E(\mathbf{G}|_{\mathsf{C}_N})| + 2(|E(\mathbf{G}|_{\mathsf{C}_N^c})| + |E_J|) \le \alpha - 1$. By connectivity of $\mathbf{G}$ and the fact that each component in $\mathbf{G}|_{\mathsf{C}_N^c}$ is connected by at least one edge to a vertex of $\mathbf{G}|_{\mathsf{C}_N}$ we have that $n(\mathbf{G}|_{\mathsf{C}_N^c}) + n(\mathbf{G}|_{\mathsf{C}_N}) \le |E_J| + |E(\mathbf{G}|_{\mathsf{C}_N^c})| + |E(\mathbf{G}|_{\mathsf{C}_N})|$ and $n(\mathbf{G}|_{\mathsf{C}_N^c}) \le |E_J| + |E(\mathbf{G}|_{\mathsf{C}_N^c})|$. Combining we get:

$$\begin{aligned} n(\mathbf{G}|_{\mathsf{C}_N}) + 2n(\mathbf{G}|_{\mathsf{C}_N^c}) &\le |E(\mathbf{G}|_{\mathsf{C}_N})| + 2(|E(\mathbf{G}|_{\mathsf{C}_N^c})| + |E_J|) \\ &\le \alpha - 1 \end{aligned}$$

For a candidate graph $\mathbf{G}$, the number of possible labels of types is upper bounded by

$$\begin{aligned} \text{no. of possible labeling of } \mathbf{G} &\le 2^{n(\mathbf{G}|_{\mathsf{C}_N}) + n(\mathbf{G}|_{\mathsf{C}_N^c})} (\kappa_N \sqrt{N})^{n(\mathbf{G}|_{\mathsf{C}_N})} (N)^{n(\mathbf{G}|_{\mathsf{C}_N^c})} \\ &\le (4\kappa\sqrt{N})^{\alpha - 1}, \end{aligned}$$

for large enough N. Denote by $\mathcal{U}_i^t$ the set of trees $\mathcal{T}_i^t$ with the labels removed. We then have, using the above and Eq. (3.11)

$$\begin{aligned} |S(A)| &\le C^{md^{t+1}} \sum_{(\mathcal{U}_i^t)^m} C_1 (\sqrt{N})^{-\alpha} (4\kappa\sqrt{N})^{\alpha - 1} \\ &\le C_2 N^{-1/2}, \end{aligned} \tag{3.12}$$

where we absorbed the summation over $(\mathcal{U}_i^t)^m$ into C_2 since it is independent of N. The constant C_1 appears because the same tuple of (unlabeled) trees can yield different (candidate) graphs $\mathbf{G}$, however their total number is independent of N.

Indeed, we can do a similar calculation to obtain that $|R(A)|$ is $O(N^{-1/2})$. For such a graph, $n(\mathbf{G}|_{\mathsf{C}_N}) + n(\mathbf{G}|_{\mathsf{C}_N^c}) = |E(\mathbf{G}|_{\mathsf{C}_N})| + |E(\mathbf{G}|_{\mathsf{C}_N^c})| + |E_J| - a$ for some $a \ge 1$ when $\mathbf{G}$ has at least one cycle. We have $|E(\mathbf{G}|_{\mathsf{C}_N})| + 2(|E(\mathbf{G}|_{\mathsf{C}_N^c})| + |E_J|) \le \alpha$ by counting minimum multiplicities and $n(\mathbf{G}|_{\mathsf{C}_N^c}) \le |E(\mathbf{G}|_{\mathsf{C}_N^c})| + |E_J|$ by connectivity argument. Thus:

$$\begin{aligned} n(\mathbf{G}|_{\mathsf{C}_N}) + 2n(\mathbf{G}|_{\mathsf{C}_N^c}) &\le \alpha - a \\ &\le \alpha - 1. \end{aligned}$$

The number of possible labels for $\mathbf{G}$ is thus bounded above by $(4\kappa\sqrt{N})^{\alpha-1}$. Following the same argument as before, we get:

$$|R(A)| \leq C_3 N^{-1/2}, \tag{3.13}$$

for some constant C_3 dependent only on m, d, t, κ. We note here that the same bounds hold for $S(\bar{A})$ and $R(\bar{A})$. Indeed, let $\varphi(T)^g_{rs}$ denote the number of times an edge $u \to v$ of (distinct) types $l(u), l(v) \in \{r, s\}$ is covered with $|u| = g$. By definition, it follows that $\sum_g \varphi(T)^g_{rs} = \phi(T)_{rs}$. We then obtain:

$$\begin{aligned}
\mathbb{E}\left[\prod_{\ell=1}^{m} \bar{A}(T_\ell)\right] &= \mathbb{E}\left[\prod_{j<k}\prod_{g} (A^{g-1}_{jk})^{\sum_{\ell=1}^m \varphi(T_\ell)^g_{jk}}\right] \\
&\leq \mathbb{E}\left[\prod_{j<k}\prod_{g} |A^{g-1}_{jk}|^{\sum_{\ell=1}^m \sum_g \varphi(T_\ell)^g_{jk}}\right] \\
&= C_4 \left(\frac{1}{\sqrt{N}}\right)^{\alpha}, \qquad (3.14)
\end{aligned}$$

This can be used in place of Eq. (3.11) to obtain the required bounds on $S(\bar{A})$ and $R(\bar{A})$.

By the bounds Eq. (3.12), Eq. (3.13), to prove our result we only need to concentrate on $T(A)$. It suffices to show that $T(A) = T(\bar{A})$. We first consider the case $\mathbb{E}\left[\prod_{\ell=1}^m A(T_\ell)\right] \neq 0 = \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]$. This implies that there exists an edge rs in $E(\mathbf{G}|_{\mathsf{C}_N^c}) \cup E_J$ with multiplicity 2, but appearing in different generations in the tuple of trees. Suppose they appear on the same branch of the tree, call it T_1. Then there exists $a \to b$ and $c \to d$ with $\{l(a), l(b)\} = \{l(c), l(d)\} = \{i, j\}$ with $a \to b$ on the path from c to the root. Due to the non-backtracking property, $a \neq d$. However, then these edges form a cycle in $\mathbf{G}$ (formed of the edges from d to a) because the tree is non-backtracking and we arrive at a contradiction. Now suppose the edges $a \to b$ and $c \to d$ as above appear in different generations in distinct trees T_1 and T_2 respectively. Then as the roots of the T_ℓ's identify to the same vertex, and the trees are non back-tracking, these form a cycle in $\mathbf{G}$ and we arrive at a contradiction. Using the same argument, we see that such edges as $a \to b$ and $c \to d$ above cannot exist even on different branches of the same tree in different generations.

Now assume $\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] \neq 0$. This means that every edge in $E(\mathbf{G}|_{\mathsf{C}_N^c}) \cup E_J$ is covered exactly twice in the same generation and every edge in $E(\mathbf{G}|_{\mathsf{C}_N})$ is covered at most twice. Then, if $\mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T\ell)\right] \neq \mathbb{E}\left[\prod_{\ell=1}^m A(T\ell)\right]$, there must exist an edge $rs \in E(\mathbf{G}|_{\mathsf{C}_N})$ covered twice i.e. with multiplicity, but in two different generations. However, by the argument given previously, this is not possible. We thus obtain that $T(A) = T(\bar{A})$. Using this and the bounds on $S(A)$, $R(A)$ we obtain the required result, for an appropriately adjusted leading constant K depending on m, d, t and κ.

□

Before proceeding, we prove the following results that are useful to establish state evolution.

Lemma 3.5. *Consider the situation as assumed in Lemma 3.4. Then we have, for some constants $K_m(m, d, t, \kappa), K'_m(m, d, t, \kappa)$ independent of N that:*

$$\begin{aligned}
|\mathbb{E}[(\xi^t_{i\to j})^m]| &\leq K_m \\
|\mathbb{E}(\xi^t_i)^m| &\leq K'_m
\end{aligned}$$

Proof. We prove the claim for $\xi^t_{i\to j}$. The other claim follows by essentially the same argument. Recall from the tree representation of Lemma 3.3:

$$\mathbb{E}[(\xi^t_{i\to j})^m] = \sum_{T_1,\ldots,T_m \in \mathcal{T}^t_{i\to j}} \left[\prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)\right] \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right]$$

Using the same splitting of contributions to the above sum into $S(\bar{A})$, $R(\bar{A})$ and $T(\bar{A})$ as in Lemma 3.4, we see that it is sufficient to prove that $|T(A)|$ is bounded uniformly over N. We have:

$$\begin{aligned}|T(\bar{A})| &= \sum_{T_1,\ldots,T_m} \left[\prod_{\ell=1}^m \Gamma(T_\ell, \mathbf{q}, t)\right] \mathbb{E}\left[\prod_{\ell=1}^m \bar{A}(T_\ell)\right] \\ &\leq \sum_{T_1,\ldots T_m} C^{d^{t+1}} C_1 (\sqrt{N})^\alpha,\end{aligned}$$

where $\alpha = \alpha(T_1, \ldots, T_m)$ is the number of edges counted with multiplicity and $T_1, \ldots T_m$ ranges over m-tuples of trees such that the graph $\mathbf{G}$ (formed by identifying vertices of the same type) is a tree. Define $n(\mathbf{G}|_{\mathsf{C}_N})$, $n(\mathbf{G}|_{\mathsf{C}_N^c})$, $E(\mathbf{G}|_{\mathsf{C}_N})$, $E(\mathbf{G}|_{\mathsf{C}_N^c})$ and E_J as in Lemma 3.4. By an argument similar to that for bounding $R(A)$, we obtain that $n(\mathbf{G}|_{\mathsf{C}_N}) + 2n(\mathbf{G}|_{\mathsf{C}_N^c}) = \alpha$. Thus we get:

$$\begin{aligned}|T(\bar{A})| &\leq C_5 (\sqrt{N})^{-\alpha} (4\kappa\sqrt{N})^\alpha \\ &\leq K_m,\end{aligned}$$

where $K_m = K(m, d, t, \kappa)$ is a constant independent of N. For convenience, we make only the dependence on m explicit. The result follows, after a small change in the constant K since the other contributions $S(\bar{A})$ and $R(\bar{A})$ are $O(N^{-1/2})$. □

Lemma 3.6. *Consider the situation as in Lemma 3.4. Then we have:*

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{|\mathsf{C}_N|} \sum_{i\in\mathsf{C}_N} (\xi^t_i)^m\right) = 0$$

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{|\mathsf{C}_N|} \sum_{i\in\mathsf{C}_N\setminus j} (\xi^t_{i\to j})^m\right) = 0$$

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{N} \sum_{i\in[N]\setminus\mathsf{C}_N} (\xi^t_i)^m\right) = 0$$

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{N} \sum_{i\in[N]\setminus\mathsf{C}_N, j} (\xi^t_{i\to j})^m\right) = 0$$

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{N} \sum_{i\in[N]} (\xi^t_i)^m\right) = 0$$

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{N} \sum_{i\in[N]\setminus j} (\xi^t_{i\to j})^m\right) = 0,$$

where $\mathrm{Var}(\cdot)$ *denotes the variance of the argument. The same results hold with* θ^t *instead of* ξ^t.

Proof. We prove only the first claim in detail. The proofs for the rest of the claims follow the same analysis. To begin with:

$$\mathrm{Var}\left(\frac{1}{|\mathsf{C}_N|}\sum_{i\in\mathsf{C}_N}(\xi_i^t)^m\right)=\frac{1}{|\mathsf{C}_N|^2}\sum_{i,j\in\mathsf{C}_N}\left(\mathbb{E}\left[(\xi_i^t)^m(\xi_j^t)^m\right]-\mathbb{E}\left[(\xi_i^t)^m\right]\mathbb{E}\left[(\xi_j^t)^m\right]\right).$$

Note that the terms wherein $i=j$ are $O(|\mathsf{C}_N|)$, using Lemma 3.5. We now control each of the remaining summands, where i,j distinct, in the following fashion. Fix a pair i,j. The summand $\left(\mathbb{E}\left[(\xi_i^t)^m(\xi_j^t)^m\right]-\mathbb{E}\left[(\xi_i^t)^m\right]\mathbb{E}\left[(\xi_j^t)^m\right]\right)$ can be written as a summation over $2m$-tuples of trees $T_1,\ldots,T_m,T_1',\ldots,T_m'$ where the first m belong to $\mathcal{T}_i^t$ and the last m to $\mathcal{T}_j^t$. Let $\mathbf{G}$ denote the simple graph obtained by identifying vertices of the same type in the tuple $T_1,\ldots,T_m,T_1',\ldots,T_m'$. Let $\mathbf{G}|_{\mathsf{C}_N}$ and $\mathbf{G}|_{\mathsf{C}_N^c}$ be subgraphs defined as in Proposition 3.4. The terms in which $\mathbf{G}$ is disconnected with one component containing i and the other containing j, are identical in $\mathbb{E}\left[(\xi_i^t)^m(\xi_j^t)^m\right]$ and $\mathbb{E}\left[(\xi_i^t)^m\right]\mathbb{E}\left[(\xi_j^t)^m\right]$ and hence cancel each other. If $\mathbf{G}$ is connected, by the argument in Lemma 3.4 all terms where $\mathbf{G}$ is not a tree, or when $\mathbf{G}|_{\mathsf{C}_N^c}$ contains an edge covered thrice or more have vanishing contributions. It remains to check the contributions of terms where $\mathbf{G}$ is a connected tree, and every edge in $\mathbf{G}|_{\mathsf{C}_N^c}$ is covered at exactly twice. Defining $n(\mathbf{G}|_{\mathsf{C}_N})$, $n(\mathbf{G}|_{\mathsf{C}_N^c})$, $E(\mathbf{G}|_{\mathsf{C}_N})$, $E(\mathbf{G}|_{\mathsf{C}_N^c})$ and E_J as before, we have that $n(\mathbf{G}|_{\mathsf{C}_N})+n(\mathbf{G}|_{\mathsf{C}_N^c})\leq E(\mathbf{G}|_{\mathsf{C}_N})+E(\mathbf{G}|_{\mathsf{C}_N^c})+E_J-1$ since types i and j have been fixed. As before $n(\mathbf{G}|_{\mathsf{C}_N^c})\leq E(\mathbf{G}|_{\mathsf{C}_N^c})+E_J$ by connectivity and $E(\mathbf{G}|_{\mathsf{C}_N})+2(E(\mathbf{G}|_{\mathsf{C}_N^c})+E_J)\leq\alpha$ where α is the number of edges counted with multiplicity. This yields $n(\mathbf{G}|_{\mathsf{C}_N})+2n(\mathbf{G}|_{\mathsf{C}_N^c})\leq\alpha-1$. The total number of such terms is thus at most $O(N^{(\alpha-1)/2})$, while their weight is bounded by $O(N^{-\alpha/2})$. Their overall contribution, consequently, vanishes in limit. We thus have $\forall i\neq j\in\mathsf{C}_N$:

$$\mathbb{E}\left[(\xi_i^t)^m(\xi_j^t)^m\right]-\mathbb{E}\left[(\xi_i^t)^m\right]\mathbb{E}\left[(\xi_j^t)^m\right]\leq\varepsilon(N),$$

where $\varepsilon(N)\to 0$ as $N\to\infty$. This gives:

$$\mathrm{Var}\left(\frac{1}{|\mathsf{C}_N|}\sum_{i\in\mathsf{C}_N}(\xi_i^t)^m\right)\leq O(|\mathsf{C}_N|^{-1})+\varepsilon(N),$$

and the first claim follows.

The other claims follow using the same argument, and since $|\mathsf{C}_N|=o(N)$. □

Proposition 3.7. *Let* μ_t,τ_t *be given as in Eqs. 2.4, 2.5. Consider* $(A(N),\mathcal{F}_N,\theta_N^0)_{N\geq 1}$ *a sequence of* (C,d)*-regular MP instances. Then the following limits hold for each* $m\geq 1$ *and* $t\geq 0$*:*

$$\lim_{N\to\infty}\mathbb{E}[(\theta_i^t)^m]=\mathbb{E}[(\mu_t+Z_t)^m]\ \textit{if}\ i\in\mathsf{C}_N \tag{3.15}$$

$$\lim_{N\to\infty}\mathbb{E}[(\theta_i^t)^m]=\mathbb{E}[(Z_t)^m]\ \textit{otherwise.} \tag{3.16}$$

where $Z_t\sim\mathsf{N}(0,\tau_t^2)$.

Proof. Fix $j \neq i$. We prove by induction over t that for all $t \geq 0$ and $m \geq 1$:

$$\lim_{N\to\infty} \mathbb{E}[(\xi_{i\to j}^{t+1})^m] = \mathbb{E}[(\mu_{t+1} + Z_{t+1})^m] \quad \text{if } i \in \mathsf{C}_N \tag{3.17}$$

$$\lim_{N\to\infty} \mathbb{E}[(\xi_{i\to j}^{t+1})^m] = \mathbb{E}[(Z_{t+1})^m] \quad \text{otherwise} \tag{3.18}$$

$$\lim_{N\to\infty} \frac{1}{|\mathsf{C}_N|} \sum_{k\in\mathsf{C}_N} (\xi_{k\to j}^{t+1})^m = \mathbb{E}[(\mu_{t+1} + Z_{t+1})^m] \tag{3.19}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k\in[N]\setminus\mathsf{C}_N} (\xi_{k\to j}^{t+1})^m = \mathbb{E}[(Z_{t+1})^m], \tag{3.20}$$

where Eqs. (3.19) and (3.20) hold in probability. For $t \geq 1$, denote by $\mathfrak{F}_t$ the σ-algebra generated by $A^0, \ldots, A^{t-1}$. For convenience of notation, we write $\tilde{A}_{ij}^t$ as the centered version of A_{ij}^t. Hence $\tilde{A}_{ij} = A_{ij} - \lambda/\sqrt{N}$ if both $i, j \in \mathsf{C}_N$, else $\tilde{A}_{ij} = A_{ij}$. First consider the case where index $i \in \mathsf{C}_N$. Then we have, for any $j \neq i$:

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}\left[\xi_{i\to j}^{t+1} | \mathfrak{F}_t\right] &= \lim_{N\to\infty} \mathbb{E}\left[\sum_{\ell\in\mathsf{C}_N\setminus j} A_{\ell i}^t f(\xi_{\ell\to i}^t, t) + \sum_{\ell\in[N]\setminus\mathsf{C}_N, j} A_{\ell i}^t f(\xi_{\ell\to i}^t, t) \middle| \mathfrak{F}_t\right] \\
&= \lambda\kappa\, \mathbb{E}[f(\mu_t + Z_t, t)] \quad \text{in probability} \\
&= \mu_{t+1},
\end{aligned}$$

where $Z_t \sim \mathsf{N}(0, \tau_t^2)$. Here the second equality follows from the induction hypothesis and the third from definition. Considering the variance we have:

$$\begin{aligned}
\lim_{N\to\infty} \mathrm{Var}\left[\xi_{i\to j}^{t+1} | \mathfrak{F}_t\right] &= \lim_{N\to\infty} \mathbb{E}\left[\sum_{\ell\in[N]\setminus j} (\tilde{A}_{\ell i}^t f(\xi_{\ell\to i}^t, t))^2 \middle| \mathfrak{F}_t\right] \\
&= \lim_{N\to\infty} \frac{1}{N} \sum_{\ell\in[N]\setminus j} (f(\xi_{\ell\to i}^t, t))^2 \\
&= \mathbb{E}[f(Z_t, t)^2] \\
&= \tau_{t+1}^2,
\end{aligned}$$

where the penultimate equality holds in probability, and follows from the induction hypothesis.

Notice that $[\xi_{i\to j}^{t+1} | \mathfrak{F}_t - \mathbb{E}(\xi_{i\to j}^{t+1} | \mathfrak{F}_t)]$ is a sum of independent random variables (due to the conditioning on $\mathfrak{F}_t$). We show that, in probability, the Lindeberg condition for the central limit theorem holds. By the induction hypothesis we have, in probability:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{\ell\in[N]\setminus j} (f(\xi_{\ell\to i}^t, t))^4 = \mathbb{E}\left[(f(Z_t, t))^4\right]$$

Using this we have, for any $\varepsilon > 0$:

$$\sum_{\ell\in[N]\setminus j} \mathbb{E}\left[(\tilde{A}_{\ell i} f(\xi_{\ell\to i}^t, t))^2 \mathbb{I}_{\{|\tilde{A}_{\ell i} f(\xi_{\ell\to i}^t, t)| \geq \varepsilon\}} \middle| \mathfrak{F}_t\right] \leq C_6 \left(\frac{\rho}{\varepsilon N}\right)^2 \sum_{\ell\in[N]\setminus j} (f(\xi_{\ell\to i}^t, t))^4 \xrightarrow{p} 0,$$

using the induction hypothesis and Lemma A.2. The constant C_6 here comes from the leading factors in Lemma A.2. It follows from Lemma A.1 that for a bounded function $h : \mathbb{R} \to \mathbb{R}$ with bounded first, second and third derivatives that:

$$\lim_{N\to\infty} \mathbb{E}[h(\xi_{i\to j}^{t+1})|\mathfrak{F}_t] = \mathbb{E}[h(Z_{t+1} + \mu_{t+1})] \text{ in probability.}$$

Since the functions $h_m^+(x) = (x^m)_+$ and $h_m^-(x) = (x^m)_-$ for $m \geq 3$ can be approached pointwise by a sequence of bounded functions with bounded first, second and third derivatives and since $\mathbb{E}\left[(\xi_{i\to j}^{t+1})^m|\mathfrak{F}_t\right]$ is integrable by definition we have :

$$\lim_{N\to\infty} \mathbb{E}\left[(\xi_{i\to j}^{t+1})^m|\mathfrak{F}_t\right] = \mathbb{E}\left[(\mu_{t+1} + Z_{t+1})^m\right] \text{ in probability.}$$

By the tower property of conditional expectation and Lemma 3.5, the expectations also converge yielding the induction claim Eq. (3.17). Employing Chebyshev inequality and Lemma 3.6 on the sequence $\left\{\sum_{k\in\mathsf{C}_N}(\xi_{k\to j}^{t+1})^m/|\mathsf{C}_N|\right\}_{N\geq 1}$, we obtain the induction claim Eq. (3.19).

We now turn to the case when $i \notin \mathsf{C}_N$. By Eq. (3.1):

$$\begin{aligned}\mathbb{E}[\xi_{i\to j}^{t+1}|\mathfrak{F}_t] &= \mathbb{E}\left[\sum_{\ell\in[N]\setminus j} \tilde{A}_{\ell i}^t f(\xi_{\ell\to i}^t, t)\middle|\mathfrak{F}_t\right] \\ &= 0.\end{aligned}$$

For the variance we compute:

$$\begin{aligned}\lim_{N\to\infty} \mathrm{Var}[\xi_{i\to j}^{t+1}|\mathfrak{F}_t] &= \lim_{N\to\infty} \mathbb{E}\left[\sum_{\ell\in[N]\setminus j} (\tilde{A}_{\ell i}^t f(\xi_{\ell\to i}^t, t))^2\right] \\ &= \lim_{N\to\infty} \frac{1}{N}\sum_{i\in[N]\setminus j} (f(\xi_{\ell\to i}^t, t))^2 \\ &= \mathbb{E}[(f(Z_t, t))^2] \\ &= \tau_{t+1}^2.\end{aligned}$$

The penultimate equality holds in probability, from the induction hypothesis and the last equality by definition. Proceeding exactly as before, we obtain induction claim Eq. (3.18) and the claim Eq. (3.20).

The base case is simpler since for $t = 0$, $\mathfrak{F}_0$ is taken to be trivial. When $i \in \mathsf{C}_N$ we have:

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}\left[\xi_{i\to j}^1\right] &= \lim_{N\to\infty} \mathbb{E}\left[\sum_{\ell\in\mathsf{C}_N\setminus j} A_{\ell i}^0 f(1,0) + \sum_{\ell\in[N]\setminus\mathsf{C}_N, j} A_{\ell i}^0 f(1,0)\right] \\ &= \mu_1,\end{aligned}$$

and for the variance:

$$\begin{aligned}\lim_{N\to\infty} \mathrm{Var}\left[\xi_{i\to j}^1\right] &= \lim_{N\to\infty} \mathbb{E}\left[\sum_{\ell\in[N]\setminus j} (\tilde{A}_{\ell i}^0 f(1,0))^2\right] \\ &= \lim_{N\to\infty} \frac{1}{N}\sum_{\ell\in[N]\setminus\mathsf{C}_N, j} (f(1,0))^2 \\ &= \tau_1^2,\end{aligned}$$

by definition. It follows from the central limit theorem that $\xi^1_{i\to j} \stackrel{\mathrm{d}}{\Rightarrow} \mathsf{N}(\mu_1, \tau_1^2)$ when $i \in \mathsf{C}_N$. A very similar argument yields that $\xi^1_{i\to j} \stackrel{\mathrm{d}}{\Rightarrow} \mathsf{N}(0, \tau_1^2)$ when $i \notin \mathsf{C}_N$. Eqs. (3.17), (3.18), (3.19) and (3.20) follow for $t = 0$ using Lemma 3.5.

The proofs for ξ_i^t follow from essentially the same argument except that the required sums are modified to include the vertex j. Asymptotically in N, this has no effect on the result and we obtain the following limits:

$$\lim_{N\to\infty} \mathbb{E}[(\xi_i^t)^m] = \mathbb{E}[(\mu_t + Z_t)^m] \quad \text{if } i \in \mathsf{C}_N \tag{3.21}$$

$$\lim_{N\to\infty} \mathbb{E}[(\xi_i^t)^m] = \mathbb{E}[(Z_t)^m] \quad \text{otherwise} \tag{3.22}$$

$$\lim_{N\to\infty} \frac{1}{|\mathsf{C}_N|} \sum_{i\in\mathsf{C}_N} (\xi_i^t)^m = \mathbb{E}[(\mu_t + Z_t)^m] \text{ in probability} \tag{3.23}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i\in[N]\setminus\mathsf{C}_N} (\xi_i^t)^m = \mathbb{E}[(Z_t)^m] \text{ in probability.} \tag{3.24}$$

Using Eqs. (3.21), (3.22) and Proposition 3.4 the result follows. □

We can now prove Lemma 2.2. For brevity, we show only Eq. (2.7) as the argument for Eq. (2.6) is analogous. To show Eq. (2.7) it suffices to show that, for any subsequence $\{N_k\}$ there exists a refinement $\{N'_k\}$ such that:

$$\frac{1}{N'_k} \sum_{i=1\in[N'_k]\setminus\mathsf{C}_{N'_k}} \psi(\theta_i^t) = \mathbb{E}[\psi(Z_t)] \text{ a.s.} \tag{3.25}$$

Fix a subsequence $\{N_k\}$. By Chebyshev inequality, Lemma 3.6 and Proposition 3.7 there exists a refinement $\{N_k(1)\} \subseteq \{N_k\}$ such that:

$$\lim_{k\to\infty} \frac{1}{N_k(1)} \sum_{i\in[N_k(1)]\setminus\mathsf{C}_{N_k(1)}} \theta_i^t = \mathbb{E}[Z_t] \text{ a.s.}$$

By the same argument, for each $m \in \mathbb{N}$, there exists a refinement $\{N_k(m)\} \subseteq \{N_k(m-1)\}$ such that:

$$\lim_{k\to\infty} \frac{1}{N_k(m)} \sum_{i\in[N_k(m)]\setminus\mathsf{C}_{N_k(m)}} (\theta_i^t)^m = \mathbb{E}[Z_t] \text{ a.s.}$$

Let N'_k be the sequence $N_k(k)$. Then, for all $m \geq 1$:

$$\lim_{k\to\infty} \frac{1}{N'_k} \sum_{i\in[N'_k]\setminus\mathsf{C}_{N'_k}} (\theta_i^t)^m = \mathbb{E}[(Z_t)^m] \text{ a.s.} \tag{3.26}$$

We define the empirical measure $\mu_N(.)$ as follows:

$$\mu_N(\cdot) = \frac{1}{N} \sum_{i\in[N]\setminus\mathsf{C}_N} \delta_{\theta_i^t}(\cdot)$$

Eq. (3.26) guarantees that, almost surely, the moments of $\mu_{N'_k}$ converge to that of Z_t. By the moment method, Eq. (3.25) follows and we obtain the required result of Eq. (2.7).

3.2 Proof of Lemma 2.3

This is section is devoted to proving Lemma 2.3. In particular, our derivation will justify the construction of polynomials in the statement of the lemma, cf. Eq. (2.8).

We will first consider the state evolution recursion (2.4), (2.5) for a general sequence of functions $\{f(\,\cdot\,,t)\}_{t\geq 0}$ (not necessarily polynomials). Since we are only interested in the ratio μ_t/τ_t, there is no loss of generality in assuming that f is normalized in such a way that $\tau_t^1 = 1$ for all t, i.e. $\mathbb{E}[f(Z,t)^2] = 1$ for all t.

Lemma 3.8. *Let μ_t be defined recursively for all $t \geq 0$ by letting*

$$\mu_{t+1} = \lambda\kappa\, e^{\mu_t^2/2}\,, \qquad \mu_0 = 1\,. \tag{3.27}$$

Further, given a sequence of functions $f \equiv \{f(\,\cdot\,,t)\}_{t\geq 0}$, such that $\mathbb{E}[f(Z,t)^2] = 1$ for all t, let $\mu_t^{(f)}$ be the corresponding state evolution sequence defined by

$$\mu_{t+1}^{(f)} = \lambda\kappa\mathbb{E}[f(\mu^{(f)} + Z)]\,, \qquad \mu_0^{(f)} = 1\,.$$

Then $\mu_t^{(f)} \leq \mu_t$ for all t, with equality verifed for $t > 0$ if and only if

$$f(z,\ell) = e^{\mu_\ell z - \mu_\ell^2} \text{ for } 0 \leq \ell \leq t\,.$$

Further $\lim_{t\to\infty}\mu_t = \infty$ if and only if $\lambda\kappa > e^{-1/2}$.

Proof. For the initial condition $\mu_0^{(f)} = 1, \tau_0 = 0$, it is easy to see that the choice of normalization ensures that we need only fix $f(1,0) = 1$ which is satisfied by the choice above. We have, for $Z \sim \mathsf{N}(0,1)$, and $\ell \geq 0$:

$$\begin{aligned}
\mu_{\ell+1}^{(f)} &= \lambda\kappa\,\mathbb{E}[f(\mu_\ell^{(f)} + Z)] \\
&= \lambda\kappa\int_{\mathbb{R}} f(z) e^{-(z-\mu_\ell^{(f)})^2/2}\frac{\mathrm{d}z}{\sqrt{2\pi}} \\
&= \lambda\kappa\, e^{-(\mu_\ell^{(f)})^2/2}\mathbb{E}\left[f(Z)e^{\mu_\ell^{(f)}Z}\right] \\
&\leq \lambda\kappa\, e^{-(\mu_\ell^{(f)})^2/2}\left(\mathbb{E}\left[(f(Z,\ell)^2\right]\right)^{1/2} e^{(\mu_\ell^{(f)})^2}\,,
\end{aligned}$$

where the inequality follows from Cauchy-Schwartz. By our choice of normalization we obtain:

$$\mu_{\ell+1}^{(f)} \leq \lambda\kappa\, e^{(\mu_\ell^{(f)})^2/2}.$$

Since the inequality is satisfied as equality only by the choice $f(z,\ell) = e^{\mu_\ell z - \mu_\ell^2}$, we have proved that $\mu_t^{(f)} = \mu_t$ only for this choice.

The last statement (namely $\mu_t \to \infty$ if and only if $\lambda\kappa > e^{-1/2}$) is a simple calculus exercise. □

We are now in position to prove Lemma 2.3.

Proof of Lemma 2.3. Let $\{\mu_t\}_{t\geq 0}$ be given as per Eq. (3.27) and define $t^* \equiv \inf\{t : \mu_{t^*} > 2M\}$. The condition $\lambda\kappa > e^{-1/2}$ ensures that t^* is finite. For a fixed d, define the mappings

$g, \hat{g}_d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by letting $g(z, \mu) = e^{\mu z}$ and $\hat{g}(z, \mu) = \sum_{k=0}^{d} \mu^k z^k / k!$. Then, since the Taylor series of the exponential has infinite radius of convergence, we have, for all z, $\mu \in \mathbb{R}$,

$$\lim_{d \to \infty} \hat{g}_d(z, \mu) = g(z, \mu)\,. \tag{3.28}$$

In the rest of this proof we will –for the sake of simplicity– omit the subscript d.

For any $\mu \in \mathbb{R}$ and $Z \sim \mathsf{N}(0, 1)$, we define:

$$G(\mu) = \frac{1}{L}\mathbb{E}\left[\lambda\kappa\, g(\mu + Z, \mu)\right]$$
$$\hat{G}(\mu) = \frac{1}{\hat{L}}\mathbb{E}\left[\lambda\kappa\, \hat{g}(\mu + Z, \mu)\right],$$

where $L = (\mathbb{E}[g(Z, \mu)^2])^{1/2}$ and $\hat{L} = (\mathbb{E}[\hat{g}(Z, \mu)^2])^{1/2}$. We first obtain that:

$$|G(\mu) - \hat{G}(\mu)| \le G(\mu)\frac{|L - \hat{L}|}{\hat{L}} + \frac{\lambda\kappa}{\hat{L}}\left|\mathbb{E}\left[(g(\mu + Z, \mu) - \hat{g}(\mu + Z, \mu))e^{\mu Z}\right]\right|. \tag{3.29}$$

Note that $|g(z, \mu)|, |\hat{g}(z, \mu)| \le e^{\mu|z|}$. It follows from Eq. (3.28) and dominated convergence that $\hat{L} \to L$ and $\mathbb{E}[\hat{g}(\mu + Z, \mu)] \to \mathbb{E}[g(\mu + Z, \mu)]$ as $d \to \infty$.

By compactness, for any $\delta > 0$, we can choose $d^* < \infty$ such that $|G(\mu) - \hat{G}(\mu)| \le \delta$ for $0 \le \mu \le 2M$. Here d^* is a function of δ, M. Note that we can now rewrite the state evolution recursions as follows:

$$\mu_{\ell+1} = G(\mu_\ell),$$
$$\hat{\mu}_{\ell+1} = \hat{G}(\hat{\mu}_\ell),$$

with $\mu_0 = \hat{\mu}_0 = 1$. Define $\Delta_\ell = |\mu_\ell - \hat{\mu}_\ell|$. Then using the fact that $G(\mu)$ is convex and that $G'(\mu) = \lambda\kappa\mu\, e^{\mu^2/2}$ is bounded by $M' = G'(2M)$ we obtain:

$$\begin{aligned}
\hat{\mu}_{\ell+1} &= \hat{G}(\hat{\mu}_\ell) \\
&\ge G(\hat{\mu}_\ell) - \delta \\
&\ge G(\mu_\ell) - M'|\mu_\ell - \hat{\mu}_\ell| - \delta \\
&= \mu_{\ell+1} - M'\Delta_\ell - \delta.
\end{aligned}$$

This implies:

$$\Delta_{\ell+1} \le M'\Delta_\ell + \delta.$$

By induction, since $\Delta_0 = |\mu_0 - \hat{\mu}_0| = 0$, we obtain:

$$\begin{aligned}
\Delta_\ell &\le \left(\sum_{k=0}^{\ell-1}(M')^k\right)\delta \\
&= \frac{M'^\ell - 1}{M' - 1}\delta.
\end{aligned}$$

Now, choosing d^* such that $\delta = M(M' - 1)/2(M'^{t^*} - 1)$ we obtain that $\Delta_{t^*} \le M/2$, implying that $\hat{\mu}_{t^*} > 3M/2 > M$. □

3.3 Proof of Lemma 2.4

Let $A|_{\widetilde{\mathsf{C}}_N}$ be the matrix A, restricted to the rows (and columns) in $\widetilde{\mathsf{C}}_N$. Also let $v \in \mathbb{R}^{|\widetilde{\mathsf{C}}_N|}$ denote the unit norm indicator vector on $\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N$, i.e.

$$v_i = \begin{cases} |\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N|^{-1/2} & \text{if } i \in \widetilde{\mathsf{C}}_N \cap \mathsf{C}_N\,, \\ 0 & \text{otherwise.} \end{cases}$$

Define $\tilde{A}|_{\widetilde{\mathsf{C}}_N}$ to be a centered matrix such that:

$$A|_{\widetilde{\mathsf{C}}_N} = \frac{\lambda|\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N|}{\sqrt{N}} vv^{\mathsf{T}} + \tilde{A}|_{\widetilde{\mathsf{C}}_N}$$

Throughout this proof we assume for simplicity that $\kappa_N = \kappa$, i.e. $|\mathsf{C}_N| = \kappa\sqrt{N}$ for some constant κ independent of N. The case of κ_N dependent on N with $\lim_{N\to\infty} \kappa_N = \kappa$ can be covered by a vanishing shift in the constants presented.

Assume that $\widetilde{\mathsf{C}}_N$ is a fixed subset selected independently of A. Then the matrix $\tilde{A}|_{\widetilde{\mathsf{C}}_N}$ has independent, zero-mean entries which are subgaussian with scale factor ρ/N. Let u denote the principal eigenvector of $A|_{\widetilde{\mathsf{C}}_N}$. The set $\mathsf{B}_N \subset [N]$ consists of the indices of the $|\mathsf{C}_N|$ entries of u with largest absolute value.

We first show that the set B_N contains a large fraction of C_N. By the condition on $\widetilde{\mathsf{C}}_N$, we have that $\|A|_{\widetilde{\mathsf{C}}_N} - \tilde{A}|_{\widetilde{\mathsf{C}}_N}\|_2 \geq \lambda\kappa(1-\varepsilon)$. By Lemma A.3 in Appendix A, for a fixed δ, $\|\tilde{A}|_{\widetilde{\mathsf{C}}_N}\|_2 \leq \lambda(1-\varepsilon)\kappa\delta$ with probability at least $2(5\xi)^N e^{-N(\xi-1)}$ where $\xi = \delta^2/32\rho\varepsilon$.

Using matrix perturbation theory, we get

$$\begin{aligned} \|u - v\|_2 &\leq \sqrt{2}\sin\theta(u,v) \\ &\leq \sqrt{2}\frac{\|\tilde{A}|_{\widetilde{\mathsf{C}}_N}\|_2}{\lambda(1-\varepsilon)\kappa_N - \|\tilde{A}|_{\widetilde{\mathsf{C}}_N}\|_2} \\ &\leq 1.9\,\delta, \end{aligned}$$

where the second inequality follows by the SIN θ theorem [DK70].

We run $t_{**} = O(\log N/\delta)$ iterations of the power method, with initialization $u^0 = (1,1,\ldots,1)^{\mathsf{T}}/|\widetilde{\mathsf{C}}_N|^{1/2}$. By the same perturbation argument, there is a $\Theta(1)$ gap between the largest and second largest eigenvalue of $A|_{\widetilde{\mathsf{C}}_N}$, and $\langle u^0, u\rangle \geq N^{-c}$. It follows by a standard argument that the output u^{**} of the power method is an approximation to the leading eigenvector u with a fixed error $\|u - u^{**}\| \leq \delta/10$. This implies that $\|u^{**} - v\| \leq 2\delta$, by the triangle inequality. Let $u^{\perp}$ ($u^{\|}$) denote the projection of u^{**} orthogonal to (resp. onto) v. Thus we have $\|u^{\perp}\|_2^2 \leq 4\delta^2$. It follows that at most $36\delta^2|\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N|$ entries in $u^{\perp}$ have magnitude exceeding $(1/3)|\widetilde{\mathsf{C}}_N \cup \mathsf{C}_N|^{-1/2}$. Notice that $u^{\|} = u^{**} - u^{\perp}$ and $u^{\|}$ is a multiple of v. Consequently, we can assume B_N is selected using $u^{\|}$, instead of v. This observation along with the bound above guarantees that at most $36\delta^2|\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N|$ entries are misclassified, i.e.

$$\begin{aligned} |\mathsf{B}_N \cap \mathsf{C}_N| &\geq (1 - 36\delta^2)|\widetilde{\mathsf{C}}_N \cap \mathsf{C}_N| && (3.30) \\ &\geq (1-\delta)(1-\varepsilon)|\mathsf{C}_N|. && (3.31) \end{aligned}$$

Here we assume $\delta \leq 1/36$.

The above argument proves that the desired result for any fixed set $\widetilde{\mathsf{C}}_N$ independent of A with a probability at least $1-2(5\xi)^N e^{-N\xi/2}$ where $\xi = \delta^2/32\rho\varepsilon$, for a universal constant c and N large enough. In order to extend it to all sets $\widetilde{\mathsf{C}}_N$ (possibly dependent on the matrix A), we can take a union bound over all possible choices of $\widetilde{\mathsf{C}}_N$ and obtain the required result. For all N large enough, the number of choices satisfying the conditions of Lemma 2.4 is bounded by:

$$\#N(\varepsilon) \le 2e^{N(\varepsilon - \varepsilon\log\varepsilon)}.$$

Choosing $\delta = \varepsilon^{1/4}$, it follows from the union bound that for some ε small enough, we have that Eq. (3.31) holds with probability at least $1-4e^{-N\upsilon'}$ where $\upsilon'(\rho,\varepsilon)\to\infty$ as $\varepsilon\to 0$. Recall that the score $\zeta_{\bar\rho}^{\mathsf{B}_N}(i)$ for a vertex i is given by:

$$\begin{aligned}\zeta_{\bar\rho}^{\mathsf{B}_N}(i) &= \frac{1}{|\mathsf{C}_N|}\sum_{j\in\mathsf{B}_N} W_{ij}\mathbb{I}_{\{|W_{ij}|\le\bar\rho\}} \\ &= \frac{1}{|\mathsf{C}_N|}\sum_{j\in\mathsf{C}_N} W'_{ij} + \frac{1}{|\mathsf{C}_N|}\left(\sum_{j\in\mathsf{B}_N\setminus\mathsf{C}_N} W'_{ij} - \sum_{j\in\mathsf{C}_N\setminus\mathsf{B}_N} W'_{ij}\right),\end{aligned}$$

where $W'_{ij} = W_{ij}\mathbb{I}_{\{|W_{ij}|\le\bar\rho\}}$. The truncated variables are subgaussian with the parameters λ'_ℓ, ρ' for $\ell = 0,1$ as according $W_{ij}\sim Q_0, Q_1$. (Here λ'_ℓ denote the means after truncation) Also, for $\bar\rho$ large enough, we may take

$$\begin{aligned}\lambda'_1 &\ge \frac{7}{8}\lambda\,,\\ \lambda'_0 &\le \frac{1}{8}\lambda\,,\\ \rho' &\le 2\rho.\end{aligned}$$

It follows that since the sum $\left(\sum_{j\in\mathsf{C}_N} W'_{ij}\right)/|\mathsf{C}_N|$ is subgaussian with parameters $\lambda'_\ell, \rho/|\mathsf{C}_N|$ the following holds with high probability:

$$\begin{aligned}\zeta_{\bar\rho}^{\mathsf{B}_N}(i) &\ge \frac{3}{4}\lambda - 2\bar\rho(\delta+\varepsilon) \text{ if } i\in\mathsf{C}_N\\ \zeta_{\bar\rho}^{\mathsf{B}_N}(i) &\le \frac{1}{4}\lambda + 2\bar\rho(\delta+\varepsilon) \text{ otherwise.}\end{aligned}$$

Choosing $\varepsilon \le (\lambda/20\bar\rho)^4$ yields the desired result.

4 The sparse graph case: Algorithm and proof of Theorem 2

In this section we consider the general hidden set problem on locally tree-like graphs, as defined in the introduction. We will introduce the reconstruction algorithm and the basic idea of its analysis. A formal proof of Theorem 2 will be presented in Section 5 and builds on these ideas.

Throughout this section we consider a sequence of locally tree-like graphs $\{G_N\}_{N\ge1}$, $G_N = ([N], E_N)$, indexed by the number of vertices N. For notational simplicity, we shall assume that these graphs are $(\Delta+1)$-regular, although most of the ideas can be easily

generalized. We shall further associate to each vertex i a binary variable X_i, with $X_i = 1$ if $i \in \mathsf{C}_N$ and $X_i = 0$ otherwise. We write $X = (X_i)_{i\in[N]}$ for the vector of these variables. It is mathematically convenient to work with a slightly different model for the vertex labels X_i: we will assume that the X_i are i.i.d. such that:

$$\mathbb{P}(X_i = 1) = \frac{\kappa}{\sqrt{\Delta}}\left(1 + \frac{\kappa}{\sqrt{\Delta}}\right)^{-1}.$$

For convenience of exposition, we also define:

$$\widetilde{\kappa}(\Delta) \equiv \kappa\left(1 + \frac{\kappa}{\sqrt{\Delta}}\right)^{-1}.$$

Notice that this leads to a set $\mathsf{C}_N = \{i \in [N] : X_i = 1\}$ that has a random size which concentrates sharply around $N\widetilde{\kappa}/\sqrt{\Delta}$. This is a slightly different model from what we consider earlier: C_N is uniformly random and of a fixed size. However, if we condition on the size $|\mathsf{C}_N|$, the i.i.d. model reduces to the earlier model. We prove in Appendix B.3, that the results of the i.i.d. model still hold for the earlier model. In view of this, throughout this section we will stick to the i.i.d model.

In order to motivate the algorithm, consider the conditional distribution of W given X, and assume for notational simplicity that Q_0, Q_1 are discrete distributions. We then have $\mathbb{P}(W|X = x) = \prod_{(i,j)\in E_N} Q_{x_i x_j}(W_{ij})$. Here the subscript $x_i x_j$ means the product of x_i and x_j. The posterior distribution of x is therefore a Markov random field (pairwise graphical model) on G_N:

$$\mathbb{P}(X = x|W) = \frac{1}{Z(W)}\prod_{(i,j)\in E_N} Q_{x_i x_j}(W_{ij}) \prod_{i\in[N]} \Big(\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\Big)^{x_i}\Big(1 - \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\Big)^{1-x_i}.$$

Here $Z(W)$ is an appropriate normalization. *Belief propagation* (BP) is a heuristic method for estimating the marginal distribution of this posterior, see [WJ08, MM09, KF09] for introductions from several points of view. For the sake of simplicity, we shall describe the algorithm for the case $Q_1 = \delta_{+1}$, $Q_0 = (1/2)\delta_{+1} + (1/2)\delta_{-1}$, whence $W_{ij} \in \{+1, -1\}$. At each iteration t, the algorithm updates 'messages' $\gamma^t_{i\to j}$, $\gamma^t_{j\to i} \in \mathbb{R}_+$, for each $(i, j) \in E_N$. As formally clarified below, these messages correspond to 'odds ratios' for vertex i to be in the hidden set.

Starting from $\gamma^0_{i\to j} = 1$ for all i, j, messages are updated as follows:

$$\gamma^{t+1}_{i\to j} = \kappa \prod_{\ell\in\partial i\setminus j}\left(\frac{1 + (1 + W_{i,\ell})\gamma^t_{\ell\to i}/\sqrt{\Delta}}{1 + \gamma^t_{\ell\to i}/\sqrt{\Delta}}\right). \tag{4.1}$$

where ∂i denotes the set of neighbors of i in G_N. We further compute the vertex quantities γ^t_i as

$$\gamma^{t+1}_i = \widetilde{\kappa} \prod_{\ell\in\partial i}\left(\frac{1 + (1 + W_{i,\ell})\gamma^t_{\ell\to i}/\sqrt{\Delta}}{1 + \gamma^t_{\ell\to i}/\sqrt{\Delta}}\right). \tag{4.2}$$

Note that γ^t_i is a function of the (labeled) neighborhood $\mathsf{Ball}_{G_N}(i; t)$. The nature of this function is clarified by the next result, that is an example of a standard result in the literature on belief propagation [WJ08, MM09, KF09].

Proposition 4.1. *Let $W_{\mathsf{Ball}_{G_N}(i;t)}$ be the set of edge labels in the subgraph $\mathsf{Ball}_{G_N}(i;t)$. If $\mathsf{Ball}_{G_N}(i;t)$ is a tree, then*

$$\frac{\mathbb{P}(X_i = 1|W_{\mathsf{Ball}_{G_N}(i;t)})}{\mathbb{P}(X_i = 0|W_{\mathsf{Ball}_{G_N}(i;t)})} = \frac{\gamma_i^t}{\sqrt{\Delta}}\,.$$

Given this result, we can attempt to estimate C_N on locally tree-like graphs by running BP for t iterations and subsequently thresholding the resulting odds-ratios. In other words we let

$$\widehat{\mathsf{C}}_N \equiv \big\{i \in [N] :\ \gamma_i^t \ge \sqrt{\Delta}\big\}\,. \tag{4.3}$$

By Proposition 4.1, this corresponds to maximizing the posterior probability $\mathbb{P}(X_i = x_i|W_{\mathsf{Ball}_{G_N}(i;t)})$ for all vertices i such that $\mathsf{Ball}_{G_N}(i;t)$ is a tree. This in turn minimizes the misclassification rate $\mathbb{P}(i \in \mathsf{C}_N; i \notin \widehat{\mathsf{C}}_N) + \mathbb{P}(i \in \widehat{\mathsf{C}}_N; i \notin \mathsf{C}_N)$. The resulting error rate is

$$\Big(1 - \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\Big)\mathbb{P}\big(\gamma_i^t \ge \sqrt{\Delta}\big|X_i = 0\big) + \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\mathbb{P}\big(\gamma_i^t < \sqrt{\Delta}\big|X_i = 1\big)\,. \tag{4.4}$$

In order to characterize this misclassification rate, we let $\widetilde{\mathsf{Tree}}(t)$ denote the regular t-generations with degree $(\Delta + 1)$ at each vertex except the leaves, rooted at vertex $\circ$ and labeled as follows. Each vertex i is labeled with $X_i \in \{0, 1\}$ independently with $\mathbb{P}(X_i = 1) = \widetilde{\kappa}/\sqrt{\Delta}$. Each edge (i, j) has label an independent $W_{ij} \sim Q_1$ if $X_i = X_j = 1$ and $W_{ij} \sim Q_0$ otherwise.

Let $\widetilde{\gamma}^t(x_\circ)$ a random variable distributed as the odds ratio for $X_\circ = 1$ on $\widetilde{\mathsf{Tree}}(t)$ when the true root value is $x_\circ$

$$\widetilde{\gamma}^t(x_\circ) \equiv \sqrt{\Delta}\,\frac{\mathbb{P}(X_\circ = 1|W_{\widetilde{\mathsf{Tree}}(t)})}{\mathbb{P}(X_\circ = 0|W_{\widetilde{\mathsf{Tree}}(t)})}\,, \qquad W_{\widetilde{\mathsf{Tree}}(t)} \sim \mathbb{P}(W_{\widetilde{\mathsf{Tree}}(t)} = \cdot\,|X_\circ = x_\circ)\,. \tag{4.5}$$

The following characterization is a direct consequence of the fact Proposition 4.1 and the fact that G_N is locally tree-like. For completeness, we provide a proof in Appendix B.2

Proposition 4.2. *Let $\widehat{\mathsf{C}}_N$ be the estimated hidden set for the BP rule (4.3) after t iterations. We then have*

$$\lim_{N\to\infty} \frac{1}{N}\,\mathbb{E}[|\mathsf{C}_N \triangle \widehat{\mathsf{C}}_N|] = \Big(1 - \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\Big)\mathbb{P}\big(\widetilde{\gamma}^t(0) \ge \sqrt{\Delta}\big) + \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\mathbb{P}\big(\widetilde{\gamma}^t(1) < \sqrt{\Delta}\big)\,.$$

Further, if $\widehat{\mathsf{C}}_N$ is estimated by any *t-local algorithm, then $\liminf_{N\to\infty} N^{-1}\,\mathbb{E}[|\mathsf{C}_N \triangle \widehat{\mathsf{C}}_N|]$ is at least as large as the right-hand side.*

We have therefore reduced the proof of Theorem 2 to controlling the distribution of the random variables $\widetilde{\gamma}^t(0)$, $\widetilde{\gamma}^t(1)$. These can be characterized by a recursion over t. For κ small we have the following.

Lemma 4.3. *Assume $\kappa < 1/\sqrt{e}$. Then there exists constants $\gamma_* < \infty$, $\delta_* = \delta_*(\kappa)$ and $\Delta_* = \Delta_*(\kappa) < \infty$ such that, for all $\Delta > \Delta_*(\kappa)$ and all $t \ge 0$, we have*

$$\mathbb{P}(\widetilde{\gamma}^t(1) \le 5\gamma_*) \ge \frac{3}{4}\,.$$

For large κ, we have instead the following.

Lemma 4.4. *Assume $\kappa > 1/\sqrt{e}$. Then there exists $c_* = c_*(\kappa) > 0$ $\Delta_* = \Delta_*(\kappa) < \infty$, $t_* = t_*(\kappa, \Delta) < \infty$ such that, for all $\Delta > \Delta_*(\kappa)$ we have*

$$\Big(1 - \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\Big)\mathbb{P}\big(\widetilde{\gamma}^t(0) \geq \sqrt{\Delta}\big) + \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\mathbb{P}\big(\widetilde{\gamma}^t(1) < \sqrt{\Delta}\big) \leq e^{-c_*\sqrt{\Delta}}\,.$$

Lemma 4.3 and Proposition 4.2 together imply one part of Theorem 2. Indeed, for Δ large enough, we have that the misclassification error is $\Omega(N/\sqrt{\Delta})$, which is the same order as choosing a random subset of size $\widetilde{\kappa}N/\sqrt{\Delta}$. Similarly, Lemma 4.4 in conjunction with Proposition 4.2 yields the second half of Theorem 2.

5 Proof of Lemma 4.3 and 4.4

In this section we prove Lemma 4.3 and 4.4 that are the key technical results leading to Theorem 2. We start by establishing some facts that are useful in both cases and then pass to the proofs of the two lemmas.

5.1 Setup: Recursive construction of $\widetilde{\gamma}^t(0)$, $\widetilde{\gamma}^t(1)$

As per Proposition 4.1, the likelihood ratio $\widetilde{\gamma}^t(x_\circ)$ can be computed by applying the BP recursion on the tree $\widetilde{\mathsf{Tree}}(t)$. In order to set up this recursion, let $\mathsf{Tree}(t)$ denote the t-generation tree, with root of degree Δ, and other non-leaf vertices of degree $\Delta + 1$. The tree $\mathsf{Tree}(t)$ carries labels x_i, W_{ij} in the same fashion as $\widetilde{\mathsf{Tree}}(t)$. Thus, $\mathsf{Tree}(t)$ differs from $\widetilde{\mathsf{Tree}}(t)$ only in the root degree. We then let

$$\gamma^t(x_\circ) \equiv \sqrt{\Delta}\,\frac{\mathbb{P}(X_\circ = 1|W_{\mathsf{Tree}(t)})}{\mathbb{P}(X_\circ = 0|W_{\mathsf{Tree}(t)})}\,, \qquad W_{\mathsf{Tree}(t)} \sim \mathbb{P}(W_{\mathsf{Tree}(t)} = \,\cdot\,|X_\circ = x_\circ)\,. \tag{5.1}$$

It is then easy to obtain the distributional recursion (here and below $\stackrel{\mathrm{d}}{=}$ indicates equality in distribution)

$$\gamma^{t+1}(0) \stackrel{\mathrm{d}}{=} \kappa \prod_{\ell=1}^{\Delta}\left(\frac{1 + (1 + A_\ell^t)\gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}\right), \tag{5.2}$$

$$\gamma^{t+1}(1) \stackrel{\mathrm{d}}{=} \kappa \prod_{\ell=1}^{\Delta}\left(\frac{1 + (1 + \widetilde{A}_\ell^t)\gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}\right). \tag{5.3}$$

This recursion is initialized with $\gamma^0(0) = \gamma^0(1) = \kappa$. Here $\gamma_\ell^t(0), \gamma_\ell^t(1)$, $\ell \in [\Delta]$ are Δ i.i.d. copies of $\gamma_t(0), \gamma_t(1)$, A_i^t, $i \in [\Delta]$, are i.i.d. uniform in $\{\pm 1\}$, x_i, $i \in [\Delta]$ are i.i.d. Bernoulli with $\mathbb{P}(x_i = 1) = \widetilde{\kappa}/\sqrt{\Delta}$. Finally $\widetilde{A}_i^t = A_i^t$ if $x_i = 0$ and $\widetilde{A}_i^t = 1$ if $x_i = 1$.

The distribution of $\widetilde{\gamma}^t(0)$, $\widetilde{\gamma}^t(1)$ can then be obtained from the one of $\gamma^t(0)$, $\gamma^t(1)$ as follows:

$$\widetilde{\gamma}^{t+1}(0) \stackrel{\mathrm{d}}{=} \kappa \prod_{\ell=1}^{\Delta+1}\left(\frac{1 + (1 + A_\ell^t)\gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}\right), \tag{5.4}$$

$$\widetilde{\gamma}^{t+1}(1) \stackrel{\mathrm{d}}{=} \kappa \prod_{\ell=1}^{\Delta+1}\left(\frac{1 + (1 + \widetilde{A}_\ell^t)\gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}\right). \tag{5.5}$$

5.2 Useful estimates

A first useful fact is the following relation between the moments of $\gamma^t(0)$ and $\gamma^t(1)$.

Lemma 5.1. *Let $\gamma(0), \gamma(1), \widetilde{\gamma}(0), \widetilde{\gamma}(1)$ be defined as in Eqs. (5.2), (5.3), (5.4) and (5.5). Then, for each positive integer a we have:*

$$\mathbb{E}\left[(\gamma^t(0))^a\right] = \kappa\, \mathbb{E}\left[(\gamma^t(1))^{a-1}\right]$$
$$\mathbb{E}\left[(\widetilde{\gamma}^t(0))^a\right] = \kappa\, \mathbb{E}\left[(\widetilde{\gamma}^t(1))^{a-1}\right].$$

Proof. It suffices to show that:

$$\frac{\mathrm{d}P_{\gamma^t(1)}}{\mathrm{d}P_{\gamma^t(0)}}(\gamma) = \frac{\gamma^t}{\kappa},$$

where the left-hand side denotes the Radon-Nikodym derivative of $P_{\gamma^t(1)}$ with respect to $P_{\gamma^t(0)}$. Let ν^t denote the posterior probability of $x_\circ = 0$ given the labels on $\mathsf{Tree}(t)$. Let $\nu^t(x_\circ)$ be distributed as ν^t conditioned on the event $X_\circ = x_\circ$ for $x_\circ = 0, 1$. In other words

$$\nu^t \equiv \mathbb{P}(X_\circ = 1 | W_{\mathsf{Tree}(t)}),$$
$$\nu^t(x_\circ) \equiv \mathbb{P}(X_\circ = 1 | W_{\mathsf{Tree}(t)}), \qquad W_{\mathsf{Tree}(t)} \sim \mathbb{P}(W_{\mathsf{Tree}(t)} = \cdot\, | X_\circ = x_\circ).$$

By Bayes rule we then have:

$$\frac{\mathrm{d}P_{\nu^t(0)}}{\mathrm{d}P_{\nu^t}} = \frac{\nu^t}{1 - \widetilde{\kappa}/\sqrt{\Delta}},$$
$$\frac{\mathrm{d}P_{\nu^t(1)}}{\mathrm{d}P_{\nu^t}} = \frac{1 - \nu^t}{\widetilde{\kappa}/\sqrt{\Delta}}.$$

Using this and the fact that $\nu^t = (1 + \gamma^t/\sqrt{\Delta})^{-1}$ by Eq. (5.1), we get

$$\frac{\mathrm{d}P_{\nu^t(0)}}{\mathrm{d}P_{\nu^t}} = \frac{1}{\left(1 + \gamma^t/\sqrt{\Delta}\right)(1 - \kappa/\sqrt{\Delta})},$$
$$\frac{\mathrm{d}P_{\nu^t(1)}}{\mathrm{d}P_{\nu^t}} = \frac{\gamma^t/\sqrt{\Delta}}{\left(1 + \gamma^t/\sqrt{\Delta}\right)\widetilde{\kappa}/\sqrt{\Delta}}.$$

It follows from this and that the mapping from ν^t to the likelihood γ is bijective and Borel that:

$$\frac{\mathrm{d}P_{\gamma^t(0)}}{\mathrm{d}P_{\gamma^t(1)}} = \gamma^t \left(\frac{\widetilde{\kappa}}{1 - \widetilde{\kappa}/\sqrt{\Delta}}\right)^{-1}$$
$$= \frac{\gamma^t}{\kappa}.$$

Here the last equality follows from the definition of $\widetilde{\kappa}$. A similar argument yields the same result for $\widetilde{\gamma}^t(0)$ and $\widetilde{\gamma}(1)$. □

Our next result is a general recursive upper bound on the moments of $\gamma^t(1)$.

Lemma 5.2. *Consider random variables $\gamma^t(0), \gamma^t(1), \widetilde{\gamma}^t(0), \widetilde{\gamma}^t(1)$ that satisfy the distributional recursions in Eqs.* (5.2)*,* (5.3)*,* (5.4) *and* (5.5)*. Then we have that, for each $t \geq 0$:*

$$\begin{aligned}
\mathbb{E}\left[\gamma^{t+1}(1)\right] &\leq \kappa \exp\left(\kappa\, \mathbb{E}[\gamma^t(1)]\right) , \\
\mathbb{E}\left[\gamma^{t+1}(1)^2\right] &\leq \kappa^2 \exp\left(3\kappa\, \mathbb{E}[\gamma^t(1)]\right) , \\
\mathbb{E}\left[\gamma^{t+1}(1)^3\right] &\leq \kappa^3 \exp\left(10\kappa\, \mathbb{E}[\gamma^t(1)]\right) .
\end{aligned}$$

Moreover, we also have:

$$\begin{aligned}
\mathbb{E}\left[\widetilde{\gamma}^{t+1}(1)\right] &\leq \kappa \exp\left(\kappa\, \mathbb{E}[\gamma^t(1)]\right) \left(1 + \frac{\mathbb{E}[\gamma^t(1)}{\Delta}\right) , \\
\mathbb{E}\left[\widetilde{\gamma}^{t+1}(1)^2\right] &\leq \kappa^2 \exp\left(3\kappa\, \mathbb{E}[\gamma^t(1)]\right) \left(1 + \frac{3\mathbb{E}[\gamma^t(1)}{\Delta}\right) , \\
\mathbb{E}\left[\widetilde{\gamma}^{t+1}(1)^3\right] &\leq \kappa^3 \exp\left(10\kappa\, \mathbb{E}[\gamma^t(1)]\right) \left(1 + \frac{10\mathbb{E}[\gamma^t(1)}{\Delta}\right) .
\end{aligned}$$

Proof. Consider the first moment $\mathbb{E}[\gamma^{t+1}(1)]$. By taking expectation of Eq. (5.3) over $\{A_\ell, x_\ell\}_{1 \leq \ell \leq \Delta}$, we get

$$\begin{aligned}
\mathbb{E}\left[\gamma^{t+1}(1) | \{\gamma_\ell^t\}_{1 \leq \ell \leq \Delta}\right] &= \kappa \prod_{\ell=1}^{\Delta} \left[\left(1 - \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\right) + \frac{\widetilde{\kappa}}{\sqrt{\Delta}} \frac{1 + 2\gamma_\ell^t(1)/\sqrt{\Delta}}{1 + \gamma_\ell^t(1)/\sqrt{\Delta}}\right] \\
&= \kappa \prod_{\ell=1}^{\Delta} \left[1 + \frac{\widetilde{\kappa}}{\Delta} \frac{\gamma_\ell^t(1)}{1 + \gamma_\ell^t(1)/\sqrt{\Delta}}\right] \\
&\leq \kappa \prod_{\ell=1}^{\Delta} \left[1 + \frac{\kappa}{\Delta}\, \gamma_\ell^t(1)\right] .
\end{aligned}$$

The last inequality uses the non-negativity of $\gamma^t(1)$ and that $\kappa > \widetilde{\kappa}$. Taking expectation over $\{\gamma_\ell^t\}_{1 \leq \ell \leq \Delta}$, and using the inequality $(1 + x) \leq e^x$, we get

$$\mathbb{E}[\gamma^{t+1}(1)] \leq \kappa \exp\left(\kappa\, \mathbb{E}[\gamma^t(1)]\right) ,$$

The claim for $\mathbb{E}[\widetilde{\gamma}^{t+1}(1)]$ follows from the same argument, except we include $\Delta + 1$ factors above and retain only the last.

Next take the second moment $\mathbb{E}[\gamma^{t+1}(1)]$. Using Eq. (5.3) and proceeding as above we get:

$$
\begin{aligned}
\mathbb{E}[\gamma^{t+1}(1)^2] &= \kappa^2\,\mathbb{E}\left[\prod_{\ell=1}^{\Delta}\left(1+\frac{\widetilde{A}_\ell\gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1+\gamma_\ell^t(x_\ell)/\sqrt{\Delta}}\right)^2\right] \\
&= \kappa^2\left[1+\Big(1-\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\Big)\frac{1}{\Delta}\mathbb{E}\left[\Big(\frac{\gamma^t(0)^2}{1+\gamma^t(0)/\sqrt{\Delta}}\Big)^2\right]\right. \\
&\qquad \left.+\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\mathbb{E}\left(\frac{2\gamma^t(1)\sqrt{\Delta}+3\gamma^t(1)^2/\Delta}{(1+\gamma^t(1)/\sqrt{\Delta})^2}\right)\right]^{\Delta} \\
&\leq \kappa^2\left[1+\frac{1}{\Delta}\mathbb{E}[(\gamma^t(0))^2]+\frac{\kappa}{\Delta}\mathbb{E}\left(\frac{2\gamma^t(1)+4\gamma^t(1)^2/\sqrt{\Delta}+2\gamma^t(1)^3/\Delta^{3/2}}{(1+\gamma^t(1)/\sqrt{\Delta})^2}\right)\right]^{\Delta} \\
&\leq \kappa^2\left(1+\frac{1}{\Delta}\,\mathbb{E}[\gamma^t(0)^2]+\frac{2\kappa}{\Delta}\mathbb{E}[\gamma^t(1)]\right)^{\Delta} \\
&\leq \kappa^2\,\exp\Big[\mathbb{E}(\gamma^t(0)^2)+2\kappa\,\mathbb{E}(\gamma^t(1))\Big] \\
&\leq \kappa^2\,\exp\Big[3\kappa\,\mathbb{E}(\gamma^t(1))\Big]\,.
\end{aligned}
$$

Consider, now, the third moment of $\gamma^{t+1}(1)$. Proceeding in the same fashion as above we obtain that:

$$
\begin{aligned}
\mathbb{E}\left[\gamma^{t+1}(1)^3\right] &= \kappa^3\,\mathbb{E}\prod_{\ell=1}^{\Delta}\left(1+\frac{\widetilde{A}_\ell\gamma^t(x_\ell)/\sqrt{\Delta}}{1+\gamma^t(x_\ell)/\sqrt{\Delta}}\right)^3 \\
&\leq \kappa^3\Bigg(1+\frac{3\widetilde{\kappa}}{\Delta}\mathbb{E}\left[\gamma^t(1)\right]+\frac{3}{\Delta}\mathbb{E}\left[(\gamma^t(0))^2\right] \\
&\qquad +\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\mathbb{E}\left(\frac{3\gamma^t(1)^2/\Delta+4\gamma^t(1)^3/\Delta^{3/2}}{(1+\gamma^t(1)/\sqrt{\Delta})^3}\right)\Bigg)^{\Delta}.
\end{aligned}
$$

Since $(3z^2+4z^3)/(1+z)^3 \leq 4z$ when $z \geq 0$, and that $\kappa > \widetilde{\kappa}$, we can bound the last term above to get:

$$
\begin{aligned}
\mathbb{E}\left[\gamma^{t+1}(1)^3\right] &\leq \kappa^3\left(1+\frac{3}{\Delta}\mathbb{E}\left[\gamma^t(0)^2\right]+7\frac{\kappa}{\Delta}\mathbb{E}\left[\gamma^t(1)\right]\right)^{\Delta} \\
&\leq \kappa^3\,\exp\left(3\mathbb{E}[\gamma^t(0)^2]+7\kappa\,\mathbb{E}[\gamma^t(1)]\right) \\
&\leq \kappa^3\,\exp\left(10\kappa\,\mathbb{E}[\gamma^t(1)]\right).
\end{aligned}
$$

The bound for the third moment of $\widetilde{\gamma}^{t+1}(1)$ follows from the same argument with the inclusion of the $(\Delta+1)^{\text{th}}$ factor. □

Lemma 5.3. *Consider $\gamma^t(0), \gamma^t(1), \widetilde{\gamma}^t(0), \widetilde{\gamma}^t(1)$ satisfying the recursions Eqs.* (5.2), (5.3), (5.4) *and* (5.5). *Also let x be Bernoulli with parameter $\widetilde{\kappa}/\sqrt{\Delta}$ and A and $\widetilde{A}$ be defined like*

A_ℓ^t and $\widetilde{A}_\ell^t$ as in Eqs. (5.2), (5.3). *Then we have that, for each $t \geq 0$ and $m \in \{1,2,3\}$:*

$$\mathbb{E}\left[\left|\log\left(1+\frac{A\gamma^t(x)/\sqrt{\Delta}}{1+\gamma^t(x)\sqrt{\Delta}}\right)\right|^m\right] \leq \frac{2\,\mathbb{E}[\gamma^t(x)^m]}{\Delta^{m/2}},$$
$$\mathbb{E}\left[\left|\log\left(1+\frac{\widetilde{A}\gamma^t(x)/\sqrt{\Delta}}{1+\gamma^t(x)\sqrt{\Delta}}\right)\right|^m\right] \leq \frac{2\,\mathbb{E}[\gamma^t(x)^m]}{\Delta^{m/2}}.$$

Proof. Consider the first claim. We have:

$$\mathbb{E}\left[\left|\log\left(1+\frac{A\gamma^t(x)/\sqrt{\Delta}}{1+\gamma^t(x)\sqrt{\Delta}}\right)\right|^3\right] = \mathbb{E}\left[\frac{1}{2}\left|\log\left(1-\frac{\gamma^t(x)/\sqrt{\Delta}}{1+\gamma^t(x)/\sqrt{\Delta}}\right)\right|^m + \frac{1}{2}\left|\log\left(1+\frac{\gamma^t(x)/\sqrt{\Delta}}{1+\gamma^t(x)/\sqrt{\Delta}}\right)\right|^m\right].$$

Bounding the first term we get:

$$\begin{aligned}
\mathbb{E}\left[\left|\log\left(1-\frac{\gamma^t(x)/\sqrt{\Delta}}{1+\gamma^t(x)\Delta}\right)\right|^m\right] &= \int_0^\infty x\,\mathbb{P}\left(\log\left(1-\frac{\gamma^t(x)\sqrt{\Delta}}{1+\gamma^t(x)/\sqrt{\Delta}}\right) \leq -x^{1/m}\right)\mathrm{d}x \\
&\leq \int_0^\infty x\,\mathbb{P}\left(\gamma^t(x)/\sqrt{\Delta} \geq e^{x^{1/m}}-1\right)\mathrm{d}x \\
&\leq \frac{\mathbb{E}\left[\gamma^t(x)^m\right]}{\Delta^{m/2}}\int_0^\infty \frac{x}{(e^{x^{1/m}}-1)^m}\mathrm{d}x \\
&\leq \frac{3\,\mathbb{E}\left[\gamma^t(x)^m\right]}{\Delta^{m/2}}.
\end{aligned}$$

For the second term, using $\log(1+z) \leq z$ for $z \geq 0$ and the positivity of $\gamma^t(x)$:

$$\mathbb{E}\left[\left|\log\left(1+\frac{\gamma^t(x)/\sqrt{\Delta}}{1+\gamma^t(x)/\sqrt{\Delta}}\right)\right|^m\right] \leq \frac{\mathbb{E}\left[\gamma^t(x)^m\right]}{\Delta^{m/2}}.$$

The combination of these yields the first claim. For the second claim, we write:

$$\begin{aligned}
\mathbb{E}\left[\left|\log\left(1+\frac{\widetilde{A}\gamma^t(x)/\sqrt{\Delta}}{1+\gamma^t(x)\sqrt{\Delta}}\right)\right|^m\right] &\leq \left(1-\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\right)\mathbb{E}\left[\left|\log\left(1+\frac{A\gamma^t(0)/\sqrt{\Delta}}{1+\gamma^t(0)\sqrt{\Delta}}\right)\right|^m\right] \\
&\quad + \frac{\kappa}{\sqrt{\Delta}}\mathbb{E}\left[\left|\log\left(1+\frac{\gamma^t(1)/\sqrt{\Delta}}{1+\gamma^t(1)\sqrt{\Delta}}\right)\right|^m\right] \\
&\leq \left(1-\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\right)\frac{2\,\mathbb{E}[\gamma^t(0)^m]}{\Delta^{m/2}} + \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\frac{\mathbb{E}[\gamma^t(1)^m]}{\Delta^{m/2}} \\
&\leq \frac{2\,\mathbb{E}[\gamma^t(x)^m]}{\Delta^{m/2}}.
\end{aligned}$$

Here the penultimate inequality follows in the same fashion as for the first claim. □

5.3 Proof of Lemma 4.3

For $\kappa < 1/\sqrt{e}$, the recursive bounds in Eq. (5.2) yield bounds on the first three moments of $\gamma^t(1)$ that are uniform in t. Precisely, we have the following:

Lemma 5.4. *For $\kappa < 1/\sqrt{e}$, let $\gamma_* = \gamma_\star(\kappa)$ be the smallest positive solution of the equation*

$$\gamma = \kappa\, e^{\kappa\gamma}\,.$$

Then we have, for all $t \ge 0$:

$$\mathbb{E}\gamma^t(1) \le \gamma_*\,, \tag{5.6}$$

$$\mathbb{E}(\gamma^t(1)^2) \le \frac{\gamma_*^3}{\kappa}\,, \tag{5.7}$$

$$\mathbb{E}(\gamma^t(1)^3) \le \frac{\gamma_*^{10}}{\kappa^7}\,. \tag{5.8}$$

Moreover, we have for all $t \ge 0$:

$$\mathbb{E}\widetilde{\gamma}^t(1) \le \gamma_*\left(1 + \frac{\kappa\gamma_*}{\Delta}\right)\,, \tag{5.9}$$

$$\mathbb{E}(\widetilde{\gamma}^t(1)^2) \le \frac{\gamma_*^3}{\kappa}\left(1 + \frac{3\kappa\gamma_*}{\Delta}\right)\,, \tag{5.10}$$

$$\mathbb{E}(\widetilde{\gamma}^t(1)^3) \le \frac{\gamma_*^{10}}{\kappa^7}\left(1 + \frac{10\kappa\gamma_*}{\Delta}\right)\,.. \tag{5.11}$$

Proof. We need only prove Eq. (5.6) since the rest follow trivially from it and Lemma 5.2. The claim of Eq. (5.6) follows from induction and Lemma 5.2 since $E[\gamma^0(1)] = \kappa < \widetilde{\kappa} \le \gamma_*$, and noting that, for $\gamma < \gamma_*$, $\widetilde{\kappa}\exp(\widetilde{\kappa}\,\gamma) < \gamma_*$. □

The following is a simple consequence of the central limit theorem.

Lemma 5.5. *For any $a < b \in \mathbb{R}$, $\sigma^2 > 0$, $\rho < \infty$, there exists $n_0 = n_0(a, b, \sigma^2, \rho)$ such that the following holds for all $n \ge n_0$. Let $\{W_i\}_{1\le i\le n}$ be i.i.d. random variables, with $\mathbb{E}\{W_1\} \ge a/n$, $\mathrm{Var}(W_1) \ge \sigma^2/n$ and $\mathbb{E}\{|W_1|^3\} \le \rho/n^{3/2}$. Then*

$$\mathbb{P}\Big\{\sum_{i=1}^n W_i \ge b\Big\} \ge \frac{1}{2}\,\Phi\Big(-\frac{b-a}{\sigma}\Big)\,.$$

Proof. Let $a_0 = n\mathbb{E}\{W_2\}$ and $\sigma_0^2 \equiv n\mathrm{Var}(W_1)$. By the Berry-Esseen central limit theorem, we have

$$\begin{aligned}\mathbb{P}\Big\{\sum_{i=1}^n W_i \ge b\Big\} &\ge \Phi\Big(-\frac{b-a_0}{\sigma_0}\Big) - \frac{\rho}{\sigma_0^3\sqrt{n}}\\ &\ge \Phi\Big(-\frac{b-a}{\sigma}\Big) - \frac{\rho}{\sigma^3\sqrt{n}}\end{aligned}$$

The claim follows by taking $n_0 \ge \rho^2/[\sigma^3\Phi(-u)]^2$ with $u \equiv (b-a)/\sigma$. □

We finally prove a statement that is stronger than Lemma 4.3, since it also controls the distribution of $\widetilde{\gamma}^t(0)$.

Proposition 5.6. *Assume $\kappa < 1/\sqrt{e}$ and let γ_* be defined as per Lemma 5.4. Then there exists $\Delta_* = \Delta_*(\kappa)$, $\delta_* = \delta_*(\kappa) > 0$ such that, for all $\Delta > \Delta_*(\kappa)$ and all $t \geq 0$, we have:*

$$\mathbb{P}(\widetilde{\gamma}^t(0) \geq 5\gamma_*) \geq \delta_*,$$
$$\mathbb{P}(\widetilde{\gamma}^t(1) \leq 5\gamma_*) \geq \frac{3}{4}.$$

where δ_* is given by:

$$\delta_* = \frac{1}{2}\Phi\left(8\frac{2\kappa - \log(5\gamma_*/\kappa)}{(\kappa^3/\gamma_*)^{1/2}}\right).$$

where $\Phi(\,\cdot\,)$ is the cumulative distribution function of the standard normal.

Proof. The second bound follows from Markov inequality and Lemma 5.4 for large enough Δ. As for the first one, we have using $\Gamma^t(0) = \log \widetilde{\gamma}^t(0)$:

$$\Gamma^{t+1}(0) = \log\kappa + \sum_{\ell=1}^{\Delta+1} \log\left(1 + \frac{A_\ell \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma^t(x_\ell)/\sqrt{\Delta}}\right),$$

It follows from Lemma 5.3 and Lemma 5.4 that:

$$\mathbb{E}\left[\log\left(1 + \frac{A_\ell \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma^t(x_\ell)/\sqrt{\Delta}}\right)\right] \geq -\frac{2\kappa}{\sqrt{\Delta}}.$$

We now lower bound the variance of each summand using the conditional variance given $\gamma^t(x)$. Since $A \in \{\pm 1\}$ are independent of $\gamma^t(x)$ and uniform, we have:

$$\begin{aligned}
\mathrm{Var}\left[\log\left(1 + \frac{A_\ell \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma^t(x_\ell)/\sqrt{\Delta}}\right)\right] &\geq \mathbb{E}\left[\mathrm{Var}\left(\log\left(1 + \frac{A_\ell \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma^t(x_\ell)/\sqrt{\Delta}}\right)\middle|\gamma^t(x)\right)\right] \\
&= \frac{1}{4}\mathbb{E}\left[\left(\log\left(1 + \frac{2\gamma^t(x)}{\sqrt{\Delta}}\right)\right)^2\right] \\
&\geq \frac{1}{4}\left(\log\left(1 + \frac{\kappa}{\sqrt{\Delta}}\right)\right)^2 \mathbb{P}(\gamma^t(x) \geq \kappa/2).
\end{aligned}$$

We have, using Lemma 5.4 that:

$$\mathbb{E}[\gamma^t(x)] \geq \kappa$$
$$\mathbb{E}[\gamma^t(x)^2] \leq 2\gamma_*\kappa.$$

Using the above and the Paley-Zygmund inequality, we get:

$$\begin{aligned}
\mathrm{Var}\left[\log\left(1 + \frac{A_\ell \gamma_\ell^t(x_\ell)/\sqrt{\Delta}}{1 + \gamma^t(x_\ell)/\sqrt{\Delta}}\right)\right] &\geq \frac{1}{32\gamma_*}\left[\log\left(1 + \frac{\kappa}{\sqrt{\Delta}}\right)\right]^2 \\
&\geq \frac{\kappa^3}{64\gamma_*\Delta},
\end{aligned}$$

for Δ large enough. Now, employing Lemma 5.5 we get:

$$\mathbb{P}(\widetilde{\gamma}^t(0) \geq 5\gamma_*) \geq \frac{1}{2}\Phi\left(-8\frac{\log(5\gamma_*/\kappa) - 2\kappa}{(\kappa^3/\gamma_*)^{1/2}}\right).$$

□

5.4 Proof of Lemma 4.4

As discussed in Section 4, BP minimizes the misclassification error at vertex i among all t-local algorithms provided $\mathsf{Ball}_{G_N}(i;t)$ is a tree. Equivalently it minimizes the misclassification rate at the root of the regular tree $\widetilde{\mathsf{T}}\mathsf{ree}(t)$. We will prove Lemma 4.4 by proving that there exists a local algorithm to estimate the root value on the tree $\mathsf{Tree}(t)$, for a suitable choice of t with error rate $\exp(-\Theta(\sqrt{\Delta}))$. Note that since the labeled tree $\mathsf{Tree}(t)$ is a subtree of $\widetilde{\mathsf{T}}\mathsf{ree}(t)$, the same algorithm is local on $\mathsf{Tree}(t)$. Formally we have the following.

Proposition 5.7. *Assume $\kappa > 1/\sqrt{e}$. Then there exists $c_* = c_*(\kappa) > 0$ $\Delta_* = \Delta_*(\kappa) < \infty$, $t_* = t_*(\kappa, \Delta) < \infty$ such that, for all $\Delta > \Delta_*(\kappa)$ we can construct a t_*-local decision rule $\mathsf{F} : W_{\mathsf{Tree}(t_*)} \to \{0,1\}$ satisfying*

$$\mathbb{P}(\mathsf{F}(W_{\mathsf{Tree}(t_*)}) \neq x_\circ | X_\circ = x_\circ) \le e^{-c_*\sqrt{\Delta}}\,,$$

for $x_\circ \in \{0,1\}$.

The rest of this section is devoted to the proof of this proposition.

The decision rule F is constructed as follows. We write $t_* = t_{*,1} + t_{*,2}$ with $t_{*,1}$ and $t_{*,2}$ to be chosen below, and decompose the tree $\mathsf{Tree}(t_*)$ into its first $t_{*,1}$ generations (that is a copy of $\mathsf{Tree}(t_{*,1})$) and its last $t_{*,2}$ generations (which consist of $\Delta^{t_{*,1}}$ independent copies of $\mathsf{Tree}(t_{*,2})$). We then run a first decision rule based on BP for the copies of $\mathsf{Tree}(t_{*,2})$ that correspond to the last $t_{*,2}$ generations. This yields decisions that have a small, but independent of Δ, error probability on the nodes at generation $t_{*,1}$. We then refine these decisions by running a different algorithm on the first $t_{*,1}$ generations.

Formally, Proposition 5.7 follows from the following two lemmas, that are proved next. The first lemma provides the decision rule for the nodes at generation $t_{*,1}$, based on the last $t_{*,2}$ generations.

Lemma 5.8. *Assume $\kappa > 1/\sqrt{e}$ and let $\varepsilon > 0$ be arbitrary. Then there exists $\Delta_* = \Delta_*(\kappa, \varepsilon) < \infty$, $t_{*,2} = t_{*,2}(\kappa, \Delta) < \infty$ such that, for all $\Delta > \Delta_*(\kappa, \varepsilon)$ there exists a $t_{*,2}$-local decision rule $\mathsf{F}_2 : W_{\mathsf{Tree}(t_{*,2})} \mapsto \mathsf{F}_2(W_{\mathsf{Tree}(t_{*,2})}) \in \{0,1\}$ such that*

$$\mathbb{P}(\mathsf{F}_2(W_{\mathsf{Tree}(t_{*,2})}) \neq x_\circ | X_\circ = x_\circ) \le \varepsilon\,,$$

for $x_\circ \in \{0,1\}$.

The second lemma yields a decision rule for the root, given information on the first $t_{*,1}$ generations, as well as decisions on the nodes at generation $t_{*,1}$. In order to state the theorem, we denote by $\mathsf{Le}(t)$ the set of nodes at generation t. For $\varepsilon = (\varepsilon(0), \varepsilon(1))$, we also let $Y_{\mathsf{Le}(t_{*,1})}(\varepsilon) = (Y_i(\varepsilon))_{i \in \mathsf{Le}(t_{*,1})}$ denote a collection of random variables with values in $\{0,1\}$ that are independent given the node labels $X_{\mathsf{Le}(t_{*,1})} = (X_i)_{i \in \mathsf{Le}(t_{*,1})}$ and such that, for all i,

$$\mathbb{P}\{Y_i(\varepsilon) \neq X_i | X_i = x\} \le \varepsilon(x)\,.$$

Lemma 5.9. *Fix $\kappa \in (0, \infty)$. There exists $\Delta_*(\kappa) = \Delta_*(\kappa) < \infty$, $t_{*,1} < \infty$, $c_*(\kappa) > 0$ and $\varepsilon_*(\kappa) > 0$ such that. for all $\Delta > \Delta_*$ there exists a $t_{*,1}$-local decision rule*

$$\mathsf{F}_1 : (W_{\mathsf{Tree}(t_{*,1})}, Y_{\mathsf{Le}(t_{*,1})}) \mapsto \mathsf{F}_1(W_{\mathsf{Tree}(t_{*,1})} Y_{\mathsf{Le}(t_{*,1})}) \in \{0,1\}$$

satisfying, for any $\varepsilon \le \varepsilon_*$:

$$\mathbb{P}(\mathsf{F}_1(W_{\mathsf{Tree}(t_{*,1})} Y_{\mathsf{Le}(t_{*,1})}) \neq x_\circ | X_\circ = x_\circ) \le e^{-c_*\sqrt{\Delta}}\,,$$

5.4.1 Proof of Lemma 5.8

The decision rule is constructed as follows. We run BP on the tree $\mathsf{Tree}(t_{*,2})$ under consideration, thus computing the likelihood ratio $\gamma^{t_{*,2}}$ at the root. We then set $\mathsf{F}_2(W_{\mathsf{Tree}(t_{*,2})}) = \mathbb{I}(\gamma^{t_{*,2}} \geq e^{\overline{\mu}})$. Here $\overline{\mu}$ is a threshold to be chosen below.

In order to analyze this rule, recall that $\gamma^t(x_\circ)$ denotes a random variable whose distribution is the same as the conditional distribution of γ^t, given the true value of the root $X_\circ = x_\circ$. It is convenient to define

$$\Gamma^t(x) \equiv \log \gamma^t(x)$$
$$\widehat{\kappa} \equiv \log \kappa.$$

As stated formally below, in the limit of large Δ, $\Gamma^t(0)$ ($\Gamma^t(1)$) is asymptotically Gaussian with mean $\mu_t(0)$ (resp. $\mu_t(1)$) and variance σ_t^2. The parameters $\mu_t(0), \mu_t(1), \sigma_t$ are defined by the recursion:

$$\mu_{t+1}(0) = \widehat{\kappa} - \frac{1}{2} e^{2\mu_t(0)+2\sigma_t^2}\,, \tag{5.12}$$

$$\mu_{t+1}(1) = \widehat{\kappa} + \kappa\, e^{\mu_t(1)+\sigma_t^2/2} - \frac{1}{2} e^{2\mu_t(0)+2\sigma_t^2}\,, \tag{5.13}$$

$$\sigma_{t+1}^2 = e^{2\mu_t(0)+2\sigma_t^2}\,,. \tag{5.14}$$

with initial conditions $\mu_0(0) = \mu_0(1) = \widehat{\kappa}$ and $\sigma_0^2 = 0$. Formally, we have the following:

Proposition 5.10. *Fix a time $t > 0$. Then the following limits hold as $\Delta \to \infty$:*

$$\Gamma^t(0) \stackrel{\mathrm{d}}{\Rightarrow} \mu_t(0) + \sigma_t Z$$
$$\Gamma^t(1) \stackrel{\mathrm{d}}{\Rightarrow} \mu_t(1) + \sigma_t Z$$

where $Z \sim \mathsf{N}(0,1)$ and $\mu_t(0), \mu_t(1), \sigma_t$ are defined by the state evolution recursions (5.12) to (5.14).

Proof. We prove the claims by induction. We have for $1 \leq i \leq t-1$:

$$\Gamma^{i+1}(0) = \widehat{\kappa} + \sum_{\ell=1}^{\Delta} \log\left(1 + \frac{A_\ell^i \gamma_\ell^i(x_\ell)/\sqrt{\Delta}}{1+\gamma_\ell^i(x_\ell)/\sqrt{\Delta}}\right).$$

Considering the first moment, we have:

$$\begin{aligned}
\mathbb{E}\left[\Gamma^{i+1}(0)\right] &= \widehat{\kappa} + \mathbb{E}\left[\Delta \log\left(1 + \frac{A_\ell^i \gamma^i(x)/\sqrt{\Delta}}{1+\gamma^i(x)/\sqrt{\Delta}}\right)\right] \\
&= \widehat{\kappa} + \mathbb{E}\left[\frac{\Delta}{2}\log\left(1 + \frac{\gamma^i(x)/\sqrt{\Delta}}{1+\gamma^i(x)/\sqrt{\Delta}}\right) + \frac{\Delta}{2}\log\left(1 - \frac{\gamma^i(x)/\sqrt{\Delta}}{1+\gamma^i(x)/\sqrt{\Delta}}\right)\right].
\end{aligned}$$

where we drop the subscript ℓ. Expanding similarly using the distribution of x, we find that the quantity inside the expectation converges pointwise as $\Delta \to \infty$ to:

$$-\gamma^i(0)^2 = -e^{2\Gamma^i(0)}.$$

Since $\mathbb{E}[\gamma^i(0)^2]$ is bounded by Lemma 5.2 for fixed t, we have by dominated convergence and the induction hypothesis that:

$$\begin{aligned}\lim_{\Delta\to\infty} \mathbb{E}\left[\Gamma^{i+1}(0)\right] &= \widehat{\kappa} - \frac{1}{2}e^{2\mu_i(0)+2\sigma_i^2} \\ &= \mu_{i+1}(0).\end{aligned}$$

Similarly, for the variance we have:

$$\mathrm{Var}(\Gamma^{i+1}(0)) = \mathrm{Var}\left(\sqrt{\Delta}\log\left(1+\frac{A^i\gamma^i(x)/\sqrt{\Delta}}{1+\gamma^i(x)/\sqrt{\Delta}}\right)\right).$$

Using $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$, we find that the right hand side converges pointwise to $\gamma^i(0)^2 = e^{2\Gamma^i(0)}$, yielding as before:

$$\begin{aligned}\lim_{\Delta\to\infty} \mathrm{Var}(\Gamma^{i+1}(0)) &= e^{2\mu_i(0)+2\sigma_i^2} \\ &= \sigma_{i+1}^2.\end{aligned}$$

It follows from the Lindeberg central limit theorem that Γ^{i+1} converges in distribution to $\mu_{i+1}(0) + \sigma_{i+1}Z$ where $Z \sim \mathsf{N}(0,1)$. For the base case, $\Gamma^0(0)$ is initialized with the value $\log\kappa = \widehat{\kappa}$. This is trivially the (degenerate) Gaussian given by $\mu_0(0) + \sigma_0 Z$ since $\mu_0(0) = \widehat{\kappa}$ and $\sigma_0 = 0$.

Now consider the case of $\Gamma^{i+1}(1)$. We have as before:

$$\Gamma^{i+1}(1) = \widehat{\kappa} + \sum_{\ell=1}^{\Delta}\log\left(1+\frac{\widetilde{A}^i_\ell\gamma^i_\ell(x_\ell)/\sqrt{\Delta}}{1+\gamma^i_\ell(x_\ell)/\sqrt{\Delta}}\right).$$

Computing the first moment:

$$\begin{aligned}\mathbb{E}[\Gamma^{i+1}(1)] &= \widehat{\kappa} + \Delta\,\mathbb{E}\left[\log\left(1+\frac{\widetilde{A}^i\gamma^i(x)/\sqrt{\Delta}}{1+\gamma^i(x)/\sqrt{\Delta}}\right)\right] \\ &= \widehat{\kappa} + \Delta\,\mathbb{E}\left[\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\log\left(1+\frac{\gamma^i(1)/\sqrt{\Delta}}{1+\gamma^i(1)/\sqrt{\Delta}}\right)\right. \\ &\qquad \left. + \left(1-\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\right)\log\left(1+\frac{A^i\gamma^i(0)/\sqrt{\Delta}}{1+\gamma^i(0)/\sqrt{\Delta}}\right)\right].\end{aligned}$$

The second term can be handled as before. The first term in the expectation converges pointwise to $\kappa\gamma^i(1) = \kappa e^{\Gamma^i(1)}$. Thus we get by dominated convergence as before that:

$$\begin{aligned}\lim_{\Delta\to\infty}\mathbb{E}[\Gamma^{i+1}(1)] &= \widehat{\kappa} + \kappa\, e^{\mu_i(1)+\sigma_i^2/2} - \frac{1}{2}e^{2\mu_i(0)+2\sigma_i^2} \\ &= \mu_{i+1}(0) + \kappa\, e^{\mu_i(1)+\sigma_i^2/2} \\ &= \mu_{i+1}(1).\end{aligned}$$

For the variance:

$$\mathrm{Var}\left(\Gamma^{i+1}(1)\right) = \Delta\,\mathrm{Var}\left(\log\left(1+\frac{\widetilde{A}^i\gamma^i(x)/\sqrt{\Delta}}{1+\gamma^i(x)/\sqrt{\Delta}}\right)\right).$$

Dealing with each term of the variance computation separately we get:

$$\Delta\,\mathbb{E}\left[\log^2\left(1+\frac{\widetilde{A}^i\gamma^i(x)/\sqrt{\Delta}}{1+\gamma^i(x)/\sqrt{\Delta}}\right)\right] = \Delta\,\mathbb{E}\Big[\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\log^2\left(1+\frac{\gamma^i(1)/\sqrt{\Delta}}{1+\gamma^i(1)/\sqrt{\Delta}}\right) + \left(1-\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\right)\log^2\left(1+\frac{A^i\gamma^i(0)/\sqrt{\Delta}}{1+\gamma^i(0)/\sqrt{\Delta}}\right)\Big],$$

Asymptotically in Δ, the contribution of first term vanishes, while that of the other can be computed, using dominated convergence as in the case of $\Gamma^{i+1}(0)$, as:

$$\lim_{\Delta\to\infty}\mathbb{E}\left[e^{2\Gamma^i(0)}\right] = e^{2\mu_i(0)+2\sigma_i^2} = \sigma_{i+1}^2$$

where we use the induction hypothesis. Similarly expanding:

$$\mathbb{E}\left[\sqrt{\Delta}\log\left(1+\frac{\widetilde{A}^i\gamma^i(x)/\sqrt{\Delta}}{1+\gamma^i(x)/\sqrt{\Delta}}\right)\right] = \sqrt{\Delta}\mathbb{E}\Big[\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\log\left(1+\frac{\gamma^i(1)/\sqrt{\Delta}}{1+\gamma^i(1)/\sqrt{\Delta}}\right) + \left(1-\frac{\widetilde{\kappa}}{\sqrt{\Delta}}\right)\log\left(1+\frac{A^i\gamma^i(0)/\sqrt{\Delta}}{1+\gamma^i(0)/\sqrt{\Delta}}\right)\Big].$$

In this case, the contribution of both terms goes to zero, hence asymptotically in Δ the expectation above vanishes. It follows, using these computations and the Lindeberg central limit theorem that $\Gamma^{i+1}(1)$ converges in distribution to the limit random variable $\mu_{i+1}(1)+\sigma_{i+1}Z$ where $Z\sim\mathsf{N}(0,1)$. The base case for $\Gamma^{i+1}(1)$ is the same as that for $\Gamma^{i+1}(0)$ since they are initialized at the (common) value $\widehat{\kappa}$. □

Using the last lemma we can estimate the probability of error that is achieved by thresholding the likelihood ratios γ^t.

Corollary 5.11. *Let γ^t be the likelihood ratio at the root of tree $\mathsf{Tree}(t)$. Define $\overline{\mu}_t \equiv (\mu_t(1)+\mu_t(0))/2$, and set $\mathsf{F}_{2,t}(W_{\mathsf{Tree}(t)})=1$ if $\gamma^t > e^{\overline{\mu}_t}$ and $\mathsf{F}_{2,t}(W_{\mathsf{Tree}(t)})=0$ otherwise. Then, there exists $\Delta_0(t)$ such that, for all $\Delta>\Delta_0(t)$,*

$$\mathbb{P}(\mathsf{F}_{2,t}(W_{\mathsf{Tree}(t)})\neq x_\circ|X_\circ=x_\circ) \le 2\,\Phi\Big(-\frac{\mu_t(1)-\mu_t(0)}{2\sigma_t}\Big),$$

Proof. It is easy to see, from Eqs. (5.12) and (5.13) that $\mu_t(1)\ge\mu_t(0)$. By the last lemma,

$$\lim_{\Delta\to\infty}\mathbb{P}(\mathsf{F}_{2,t}(W_{\mathsf{Tree}(t)})\neq x_\circ|X_\circ=x_\circ) = \Phi\Big(-\frac{\mu_t(1)-\mu_t(0)}{\sigma_t}\Big),$$

and this in turn yields the claim. □

Finally, we have the following simple calculus lemma, whose proof we omit.

Lemma 5.12. *Let $\mu_t(0)$, $\mu_t(1)$, σ_t be defined as per Eqs. (5.12), (5.13), (5.14). If $\kappa > 1/\sqrt{e}$, then*

$$\lim_{t\to\infty}\frac{\mu_t(1)-\mu_t(0)}{\sigma_t}=\infty\,.$$

Lemma 5.8 follows from combining this result with Corollary 5.11 and selecting $t_{*,2}$ so that $\Phi\big(-(\mu_t(1)-\mu_t(0))/\sigma_t\big)\le\varepsilon/4$ for $t=t_{*,2}$. Finally we let $\Delta_*=\Delta_0(t_{*,2})$ as per Corollary 5.11.

5.4.2 Proof of Lemma 5.9

We construct the decision rule $\mathsf{F}_1 : (W_{\mathsf{Tree}(t_{*,1})}, Y_{\mathsf{Le}(t_{*,1})}) \mapsto \mathsf{F}_1(W_{\mathsf{Tree}(t_{*,1})} Y_{\mathsf{Le}(t_{*,1})}) \in \{0,1\}$ recursively as follows. For each vertex $i \in \mathsf{Tree}(t)$ we compute a decision $m_i \in \{0,1\}$ based on the set of descendants of i, do be denoted as $D(i)$, as follows

$$m_i^{t+1} = \begin{cases} 1 \text{ if } \sum_{\ell \in D(i)} W_{i\ell}\, m_\ell^t \geq \frac{\kappa}{2}\sqrt{\Delta} \\ 0 \text{ otherwise.} \end{cases} \tag{5.15}$$

If i is a leaf, we let $m_i = Y_i$. Finally we take a decision on the basis of the value at the root:

$$F_1(W_{\widetilde{\mathsf{Tree}}(t)}, Y_{\mathsf{Le}(t)}) = m_\circ\,.$$

We recall that the Y_i's are conditionally independent given $X_{\mathsf{Le}(t)}$. We assume that $\mathbb{P}(Y_i = 1|X_i = 0) = \mathbb{P}(Y_i = 0|X_i = 1) = \varepsilon$, which does not entail any loss of generality because it can always be achieved by degrading the decision rule F_2. We define the following quantities:

$$p_t = \mathbb{P}(F_1(W_{\widetilde{\mathsf{Tree}}(t)}, Y_{\mathsf{Le}(t)}) = 1|X_\circ = 0)\,, \tag{5.16}$$

$$q_t = \mathbb{P}(F_1(W_{\widetilde{\mathsf{Tree}}(t)}, Y_{\mathsf{Le}(t)}) = 1|X_i = 1). \tag{5.17}$$

Note in particular that $p_0 = 1 - q_0 = \varepsilon$.

Lemma 5.9 follows immediately from the following.

Lemma 5.13. *Let p_t and q_t be defined as in Eqs.* (5.16), (5.17). *Then there exists $\varepsilon_* = \varepsilon_*(\kappa)$ and $t_* = t_*(\Delta, \kappa) < \infty$, $\Delta_* = \Delta_*(\kappa) < \infty$, $c_* = c_*(\kappa) > 0$ such that for $\Delta \geq \Delta_*$:*

$$p_{t_*} \leq 4\, e^{-c_*\sqrt{\Delta}} \tag{5.18}$$

$$1 - q_{t_*} \leq 4\, e^{-c_*\sqrt{\Delta}}. \tag{5.19}$$

Proof. Throughout the proof, we use $c_1, c_2, \dots$ to denote constants that can depend on κ but not on Δ or t. We first prove the following by induction:

$$p_{t+1} \leq e^{-c_1\Delta p_t} + e^{-c_2\Delta^{3/2}} + e^{-c_3/p_t}$$
$$1 - q_{t+1} \leq e^{-c_1\Delta p_t} + e^{-c_2\Delta^{3/2}} + e^{-c_3/4p_t}.$$

We let $\mathsf{Bin}(n,p)$ denote the binomial distribution with parameters n, p. First, let $D \sim \mathsf{Bin}(\Delta, \widetilde{\kappa}/\sqrt{\Delta})$ and, conditional on D:

$$N_t \sim \mathsf{Bin}(\Delta - D, p_t)$$
$$M_t \sim \mathsf{Bin}(D, q_t).$$

From the definition of p_t, q_t and Eq. (5.15) we observe that:

$$p_{t+1} = \mathbb{P}\left(\sum_{i=1}^{M_t+N_t} W_i \geq \kappa\sqrt{\Delta}/2\right)$$
$$q_{t+1} = \mathbb{P}\left(\sum_{i=1}^{N_t} W_i + M_t \geq \kappa\sqrt{\Delta}/2\right),$$

where we let $W_i \in \{\pm 1\}$ are i.i.d and uniformly distributed.

Considering the p_t recursion:

$$\begin{aligned} p_{t+1} &\leq \mathbb{P}\left(M_t + N_t \leq \frac{p_t\Delta}{2}\right) + \mathbb{P}\left(\sum_{i=1}^{p_t\Delta/2} W_i \geq \frac{\kappa\sqrt{\Delta}}{2}\right) \\ &\leq \mathbb{P}\left(N_t \leq \frac{p_t\Delta}{2}\right) + e^{-\kappa^2/4p_t} \\ &\leq \mathbb{P}\left(\tilde{N} \leq \frac{p_t\Delta}{2}\right) + \mathbb{P}\left(D \geq \frac{\Delta}{4}\right) + e^{-\kappa^2/4p_t}, \end{aligned}$$

where $\tilde{N} \sim \mathsf{Bin}(3\Delta/4, p_t)$ and the penultimate inequality follows from the Chernoff bound and positivity of M_t. Using standard Chernoff bounds, for $\Delta \geq 5$ we get:

$$p_{t+1} \leq e^{-3\Delta p_t/128} + e^{-c_1\Delta^{3/2}} + e^{-\kappa^2/4p_t},$$

where c_1 is a constant dependent only on κ. Bounding $1 - q_{t+1}$ in a similar fashion, we obtain:

$$\begin{aligned} 1 - q_{t+1} &= \mathbb{P}\left(\sum_{i=1}^{N_t} W_i + M_t \leq \kappa\sqrt{\Delta}/2\right) \\ &\leq \mathbb{P}\left(\sum_{i=1}^{N_t} W_i \leq -\kappa\sqrt{\Delta}/4\right) + \mathbb{P}\left(M_t \leq 3\kappa\sqrt{\Delta}/4\right). \end{aligned}$$

We first consider the M_t term:

$$\mathbb{P}\left(M_t \leq 3\kappa\sqrt{\Delta}/4\right) \leq \mathbb{P}\left(D \leq 7\kappa\sqrt{\Delta}/8\right) + \mathbb{P}\left(\tilde{M} \geq \kappa\sqrt{\Delta}/8\right),$$

where $\tilde{M} \sim \mathsf{Bin}(0, q_t)$. Further, using Chernoff bounds we obtain that this term is less than $2e^{-\kappa\sqrt{\Delta}/128}$. The other term is handled similar to the p_t recursion as:

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{N_t} W_i \leq -\kappa\sqrt{\Delta}/4\right) &\leq \mathbb{P}\left(N_t \leq \frac{p_t\Delta}{2}\right) + \mathbb{P}\left(\sum_{i=1}^{p_t\Delta/2} W_i \geq \kappa\sqrt{\Delta}/4\right) \\ &\leq \mathbb{P}\left(\tilde{N} \leq \frac{p_t\Delta}{2}\right) + \mathbb{P}\left(D \geq \frac{\Delta}{4}\right) + e^{-\kappa^2/16p_t} \\ &\leq e^{-3\Delta p_t/128} + e^{-c_1\Delta^{3/2}} + e^{-\kappa^2/16p_t}. \end{aligned}$$

Consequently we obtain:

$$1 - q_t \leq e^{-3\Delta p_t/128} + e^{-c_1\Delta^{3/2}} + e^{-\kappa^2/16p_t} + 2e^{-\kappa\sqrt{\Delta}/128}.$$

Simple calculus shows that, for all $\Delta > \Delta_*(\kappa)$ large enough, there exists ε_*, c_0 dependent on κ but independent of Δ such that

$$e^{-3\Delta p/128} + e^{-c_1\Delta^{3/2}} + e^{-\kappa^2/16p} < p$$

for all $p \in [c_0/\sqrt{\Delta}, \varepsilon_*]$. Since $p_0 = \varepsilon \leq \varepsilon_*$, this implies that there exists t_0 such that $p_{t_0} \leq c_0/\sqrt{\Delta}$. The claim follows by taking $t_* = t_0 + 1$ which yields that $p_{t_*} = O(e^{-\Theta(\sqrt{\Delta})})$. Further, observing that the error $1 - q_{t_*}$ has only an additional $2e^{-c_3\sqrt{\Delta}}$ component, we obtain a similar claim for $1 - q_{t_*}$. □

Acknowledgments

This work was partially supported by the NSF CAREER award CCF-0743978, the NSF grant DMS-0806211, and the grants AFOSR/DARPA FA9550-12-1-0411 and FA9550-13-1-0036.

A Some tools in probability theory

This appendix contains some useful facts in probability theory.

Lemma A.1. *Let $h : \mathbb{R} \to \mathbb{R}$ be a bounded function with first three derivatives uniformly bounded. Let $X_{n,k}$ be mutually independent random variables for $1 \leq k \leq n$ with zero mean and variance $v_{n,k}$. Define:*

$$\begin{aligned} v_n &\equiv \sum_{k=1}^{n} v_{n,k} \\ \delta_n(\varepsilon) &\equiv \sum_{k=1}^{n} \mathbb{E}[X_{n,k}^2 \mathbb{I}_{|X_{n,k}| \geq \varepsilon}] \\ S_n &\equiv \sum_{k=1}^{n} X_{n,k}. \end{aligned}$$

Also let $\mathcal{G}_n = \mathsf{N}(0, v_n)$. Then, for every n and $\varepsilon > 0$:

$$|\mathbb{E}h(S_n) - \mathbb{E}h(G_n)| \leq \left(\frac{\varepsilon}{6} + \frac{\sqrt{\varepsilon^2 + \delta_n}}{2} \right) v_n \|h'''\|_\infty + \delta_n \|h''\|_\infty$$

Proof. The lemma is proved using a standard swapping trick. The proof can be found in Amir Dembo's lecture notes [Dem13]. □

Lemma A.2. *Given a random variable X such that $\mathbb{E}(X) = \mu$. Suppose X satisfies:*

$$\mathbb{E}(e^{\lambda X}) \leq e^{\mu\lambda + \rho\lambda^2/2},$$

for all $\lambda > 0$ and some constant $\rho > 0$. Then we have for all $s > 0$:

$$\begin{aligned} &\mathbb{E}(|X|^s) \leq 2s! e^{(s+\lambda\mu)/2} \lambda^{-s}, \\ &\text{where } \lambda = \frac{1}{2\rho} \left(\sqrt{\mu^2 + 4s\rho} - \mu \right). \end{aligned}$$

Further, if $\mu = 0$, we have for $t < 1/e\rho$:

$$\mathbb{E}\left(e^{tX^2}\right) \leq \frac{1}{1 - e\rho t}$$

Proof. By an application of Markov inequality and the given condition on X:

$$\begin{aligned} \mathbb{P}(X \geq t) &\leq e^{-\lambda t} \mathbb{E}(e^{\lambda X}) \\ &\leq e^{-\lambda t + \mu\lambda + \rho\lambda^2/2}, \end{aligned}$$

for all $\lambda > 0$. By a symmetric argument:

$$\mathbb{P}(X \leq -t) \leq e^{\lambda t + \mu\lambda + \rho\lambda^2/2}$$

By the standard integration formula we have:

$$\begin{aligned}
\mathbb{E}(|X|^s) &= \int_0^\infty st^{s-1}\mathbb{P}(|X| \geq t)\,\mathrm{d}t \\
&= \int_0^\infty st^{s-1}\mathbb{P}(X \geq t)\,\mathrm{d}t + \int_0^\infty st^{s-1}\mathbb{P}(X \leq -t)\,\mathrm{d}t \\
&\leq 2e^{\mu\lambda+\rho\lambda^2/2}\int_0^\infty st^{s-1}e^{-\lambda t}\,\mathrm{d}t \\
&= 2s!\,e^{\mu\lambda+\rho\lambda^2/2}\lambda^{-s}
\end{aligned}$$

Optimizing over λ yields the desired result.

If $\mu = 0$, the optimization yields $\lambda = \sqrt{s/\rho}$. Using this, the Taylor expansion of $g(x) = e^{x^2}$ and monotone convergence we get:

$$\begin{aligned}
\mathbb{E}\left(e^{tX^2}\right) &= \sum_{k=0}^\infty \frac{t^k}{k!}\mathbb{E}(X^{2k}) \\
&\leq \sum_{k=0}^\infty (e\rho t)^k \frac{(2k)!}{k!(2k)^k} \\
&\leq \sum_{k=0}^\infty (e\rho t)^k \\
&= \frac{1}{1-e\rho t}.
\end{aligned}$$

Notice that here we remove the factor of 2 in the inequality, since this is not required for even moments of X. □

The following lemma is standard, see for instance [AKV02, Ver12].

Lemma A.3. *Let $M \in \mathbb{R}^{N\times N}$ be a symmetric matrix with entries M_{ij} (for $i \geq j$) which are centered subgaussian random variables of scale factor ρ. Then, uniformly in N:*

$$\mathbb{P}\left(\|M\|_2 \geq t\right) \leq (5\lambda)^N e^{-N(\lambda-1)},$$

where $\lambda = t^2/16N\rho e$ and $\|M\|_2$ denotes the spectral norm (or largest singular value) of M.

Proof. Divide M into its upper and lower triangular portions M^u and M^l so that $M = M^u + M_l$. We deal with each separately. Let m_i denote the i^{th} row of M^l. For a unit vector x, since M_{ij} are all independent and subgaussian with scale ρ, it is easy to see that $\langle m_j, x\rangle$ are also subgaussian with the same scale. We now bound the square exponential

moment of $\|M^l x\|$ as follows. For small enough $c \geq 0$:

$$\begin{aligned}\mathbb{E}\left(e^{c\|M^l x\|^2}\right) &= \mathbb{E}\left(\prod_{j=1}^{N} e^{c\langle m_j, x\rangle^2}\right) \\ &= \prod_{j=1}^{N} \mathbb{E}\left(e^{c\langle m_j, x\rangle^2}\right) \\ &\leq (1 - e\rho c)^N . \end{aligned} \tag{A.1}$$

Using this, we get for any unit vector x:

$$\mathbb{P}(\|M^l x\| \geq t) \leq \left(\frac{t^2}{N\rho e}\right)^N e^{-N(t^2/N\rho e - 1)},$$

where we used Markov inequality and Eq. (A.1) with an appropriate c. Let Υ be a *maximal* 1/2-net of the unit sphere. From a volume packing argument we have that $|\Upsilon| \leq 5^N$. Then from the fact that $g(x) = M^l x$ is $\|M^l\|$-Lipschitz in x:

$$\begin{aligned}\mathbb{P}\left(\|M^l\|_2 \geq t\right) &\leq \mathbb{P}\left(\max_{x\in\Upsilon} \|M^l x\| \geq t/2\right) \\ &\leq |\Upsilon| \mathbb{P}(\|M^l x\| \geq t/2).\end{aligned}$$

The same inequality holds for M^u. Now using the fact that $\|\cdot\|_2$ is a convex function and that M^u and M^l are independent we get:

$$\begin{aligned}\mathbb{P}\left(\|M\|_2 \geq t\right) &\leq \mathbb{P}\left(\|M^u\|_2 \geq t/2\right) + \mathbb{P}\left(\|M^l\|_2 \geq t/2\right) \\ &\leq 2\left(5^N \left(\frac{t^2}{16N\rho e}\right)^N e^{-N(t^2/16N\rho e - 1)}\right).\end{aligned}$$

Substituting for λ yields the result. □

B Additional Proofs

In this section we provide, for the sake of completeness, some additional proofs that are known results. We begin with Proposition 1.1.

B.1 Proof of Proposition 1.1

We assume the set C_N is generated as follows: let $X_i \in \{0, 1\}$ be the label of the index $i \in [N]$. Then X_i are i.i.d Bernoulli with parameter $\kappa/\sqrt{N}$ and the set $\mathsf{C}_N = \{i : X_i = 1\}$. The model of choosing C_N uniformly random of size $\kappa\sqrt{N}$ is similar to this model and asymptotically in N there is no difference. Notice that since $e_{\mathsf{C}_N} = u_{\mathsf{C}_N}/N^{1/4}$, we have that $\|e_{\mathsf{C}_N}\|^2$ concentrates sharply around κ and we are interested in the regime $\kappa = \Theta(1)$.

We begin with the first part of the proposition where $\kappa = 1 + \varepsilon$. Let $W_N = W/\sqrt{N}$, $Z_N = Z/\sqrt{N}$ and $e_{\mathsf{C}_N} = u_{\mathsf{C}_N}/N^{1/4}$. Since this normalization does not make a difference to the eigenvectors of W and Z we obtain from the eigenvalue equation $W_N v_1 = \lambda_1 v_1$ that:

$$e_{\mathsf{C}_N}\langle e_{\mathsf{C}_N}, v_1\rangle + Z_N v_1 = \lambda_1 v_1. \tag{B.1}$$

Multiplying by v_1 on either side:

$$\begin{aligned}\langle e_{\mathsf{C}_N}, v_1\rangle^2 &= \lambda_1 - \langle v_1, Z_N v_1\rangle \\ &\geq \lambda_1 - \|Z_N\|_2.\end{aligned}$$

Since $Z_N = Z/\sqrt{N}$ is a standard Wigner matrix with subgaussian entries, [AKV02] yields that $\|Z\|_2 \leq 2+\delta$ with probability at least $C_1 e^{-c_1 N}$ for some constants $C_1(\delta), c_1(\delta) > 0$. Further, by Theorem 2.7 of [KY11] we have that $\lambda_1 \geq 2 + \min(\varepsilon, \varepsilon^2)$ with probability at least $1 - N^{-c_2 \log\log N}$ for some constant c_2 and every N sufficiently large. It follows from this and the union bound that for N large enough, we have:

$$\langle e_{\mathsf{C}_N}, v_1\rangle^2 \geq \min(\varepsilon, \varepsilon^2)/2,$$

with probability at least $1 - N^{-c_4}$ for some constant $c_4 > 0$. The first claim then follows.

For the second claim, we start with the same eigenvalue equation (B.1). Multiplying on either side by φ_1, the eigenvector corresponding to the largest eigenvalue of Z_N we obtain:

$$\langle e_{\mathsf{C}_N}, v_1\rangle\langle e_{\mathsf{C}_N}, \varphi_1\rangle + \theta_1\langle v_1, \varphi_1\rangle = \lambda_1\langle v_1, \varphi_1\rangle\,,$$

where θ_1 is the eigenvalue of Z_N corresponding to φ_1. With this and Cauchy-Schwartz we obtain:

$$|\langle e_{\mathsf{C}_N}, v_1\rangle| \leq \frac{|\lambda_1 - \theta_1|}{|\langle \varphi_1, e_{\mathsf{C}_N}\rangle|}.$$

Let $\phi = (\log N)^{\log\log N}$. Then, using Theorem 2.7 of [KY11], for any $\delta > 0$, there exists a constant C_1 such that $|\lambda_1 - \theta_1| \leq N^{-1+\delta}$ with probability at least $1 - N^{-c_3 \log\log N}$.

Since φ_1 is independent of e_{C_N}, we observe that:

$$\begin{aligned}\mathbb{E}_e\left(\sum_{i=1}^N \varphi_1^i e^i_{\mathsf{C}_N}\right) &= N^{-3/4}(1-\varepsilon)\sum_{i=1}^N \varphi_1^i \\ \mathbb{E}_e\left(\sum_{i=1}^N (\varphi_1^i e^i_{\mathsf{C}_N})^2\right) &= \frac{1-\varepsilon}{N},\end{aligned}$$

where φ_1^i ($e^i_{\mathsf{C}_N}$) denotes the i^{th} entry of φ_1 (resp. e_{C_N}) and $\mathbb{E}_e(\cdot)$ is the expectation with respect to e_{C_N} holding Z_N (hence φ_1) constant. Using Theorem 2.5 of [KY11], it follows that there exists constants c_4, c_5, c_6, c_7 such that the following two happen with probability at least $1 - N^{-c_4 \log\log N}$. Firstly, the first expectation above is at most $(1-\varepsilon)\phi^{c_5} N^{-7/4}$. Secondly:

$$\left[\mathbb{E}_e\left(\sum_{i=1}^N (\varphi_1^i e^i_{\mathsf{C}_N})^2\right)\right]^{-1/2} \max_i |e^i_{\mathsf{C}_N}\varphi_i^1| \leq \frac{(1-\varepsilon)\phi^{c_7}}{N^{1/4}}.$$

Now, using the Berry-Esseen central limit theorem for $\langle \varphi_1, e_{\mathsf{C}_N}\rangle$ that:

$$\mathbb{P}\left(|\langle \varphi_1, e_{\mathsf{C}_N}\rangle| \leq c(N)^{1/2-\delta}\right) \leq \frac{1}{N^\delta},$$

for an appropriate constant $c = c(\varepsilon)$ and $\delta \in (0, 1/4)$. Using this and the earlier bound for $|\lambda_1 - \theta_1|$ we obtain that:

$$|\langle e_{\mathsf{C}_N}, v_1\rangle| \leq cN^{-1/2+3\delta}$$

with probability at least $1 - c'N^{-\delta}$, for some c' and sufficiently large N. The claim then follows using the union bound and the same argument for the first ℓ eigenvectors.

B.2 Proof of Proposition 4.2

For any fixed t, let $\mathcal{E}_N^t$ denote the set of vertices in G_N such that their t-neighborhoods are *not* a tree, i.e.

$$\mathcal{E}_N^t = \{i \in [N] : \mathsf{Ball}_{G_N}(i;t) \text{ is not a tree}\}.$$

For notational simplicity, we will omit the subscript G_N in the neighborhood of i. The relative size $\varepsilon_N^t = |\mathcal{E}_N^t|/N$ vanishes asymptotically in N since the sequence $\{G_N\}_{N\geq 1}$ is locally tree-like. We let $\mathsf{F}_{BP}(W_{\mathsf{Ball}(i;t)})$ denote the decision according to belief propagation at the i^{th} vertex.

From Proposition 4.1, Eqs. (4.1), (4.2), (4.5), (5.1) and induction, we observe that for any $i \in [N]\backslash\mathcal{E}_N^t$:

$$\frac{\mathbb{P}(X_i = 1|W_{\mathsf{Ball}(i;t)})}{\mathbb{P}(X_i = 0|W_{\mathsf{Ball}(i;t)})} \stackrel{\mathrm{d}}{=} \frac{\widetilde{\gamma}^t(X_i)}{\sqrt{\Delta}}.$$

We also have that:

$$|\widehat{\mathsf{C}}_N \triangle \mathsf{C}_N| = \sum_{i=1}^N \mathbb{I}(\mathsf{F}_{BP}(W_{\mathsf{Ball}(i;t)}) \neq X_i).$$

Using both of these, the fact that $\varepsilon_N^t \to 0$ and the linearity of expectation, we have the first claim:

$$\lim_{N\to\infty} \frac{\mathbb{E}|\widehat{\mathsf{C}}_N \triangle \mathsf{C}_N|}{N} = \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\mathbb{P}\left(\gamma^t(1) < \sqrt{\Delta}\right) + \left(1 - \frac{\widetilde{\kappa}}{\sqrt{\Delta}}\right)\mathbb{P}\left(\gamma^t(0) \geq \sqrt{\Delta}\right).$$

For any other decision rule $\mathsf{F}(W_{\mathsf{Ball}(i;t)})$, we have that:

$$\begin{aligned}\frac{\mathbb{E}[|\widehat{\mathsf{C}}_N \triangle \mathsf{C}_N|]}{N} &\geq (1 - \varepsilon_N^t)\mathbb{P}(\mathsf{F}(W_{\widetilde{\mathsf{Tree}}(t)}) \neq X_\circ) \\ &\geq (1 - \varepsilon_N^t)\mathbb{P}(\mathsf{F}_{BP}(W_{\widetilde{\mathsf{Tree}}(t)}) \neq X_\circ),\end{aligned}$$

since BP computes the correct posterior marginal on the root of the tree $\widetilde{\mathsf{Tree}}(t)$ and maximizing the posterior marginal minimizes the misclassification error. The second claim follows by taking the limits.

B.3 Equivalence of i.i.d and uniform set model

In Section 2 the hidden set C_N was assumed to be uniformly random given its size. However, in Section 4 we considered a slightly different model to choose C_N, wherein X_i are i.i.d Bernoulli random variables with parameter $\widetilde{\kappa}/\sqrt{\Delta}$. This leads to a set $\mathsf{C}_N = \{i : X_i = 1\}$ that has a random size, sharply concentrated around $N\widetilde{\kappa}/\sqrt{\Delta}$. The uniform set model can be obtained from the i.i.d model by simply conditioning on the size $|\mathsf{C}_N|$. In the limit of large N it is well-known that these two models are "equivalent". However, for completeness, we provide a proof that the results of Proposition 4.2 do not change when conditioned on the size $|\mathsf{C}_N| = \sum_{i=1}^N X_i$.

$$\mathbb{E}\left[|\widehat{\mathsf{C}}_N \triangle \mathsf{C}_N| \,\middle|\, |\mathsf{C}_N| = \frac{N\widetilde{\kappa}}{\sqrt{\Delta}}\right] = \sum_{i=1}^N \mathbb{P}\left(\mathsf{F}(W_{\mathsf{Ball}(i;t)}) \neq X_i \,\middle|\, \sum_{j=1}^N X_j = \frac{N\widetilde{\kappa}}{\sqrt{\Delta}}\right).$$

Let S be the event $\{\sum_{i=1}^{N} X_i = N\widetilde{\kappa}/\sqrt{\Delta}\}$. Notice that $\mathsf{F}(W_{\mathsf{Ball}(i;t)})$ is a function of $\{X_j, j \in \mathsf{Ball}(i;t)\}$ which is a discrete vector of dimension $K_t \leq (\Delta+1)^{t+1}$. A straightforward direct calculation yields that $(X_j, j \in \mathsf{Ball}(i;t))|S$ converges in distribution to $(X_j, j \in \mathsf{Ball}(i;t))$ asymptotically in N. This implies $W_{\mathsf{Ball}(i;t)}|S$ converges in distribution to $W_{\mathsf{Ball}(i;t)}$. Further, using the locally tree-like property of G_N one obtains:

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[|\widehat{\mathsf{C}}_N \triangle \mathsf{C}_N| \,\middle|\, |\mathsf{C}_N| = \frac{N\widetilde{\kappa}}{\sqrt{\Delta}}\right] = \mathbb{P}\left(\mathsf{F}(W_{\widetilde{\mathsf{T}}_{\mathsf{ree}(t)}}) \neq X_\circ\right),$$

as required.

References

[ABBDL10] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, and Gábor Lugosi, *On combinatorial testing problems*, The Annals of Statistics **38** (2010), no. 5, 3063–3092.

[ACCD11] Ery Arias-Castro, Emmanuel J Candès, and Arnaud Durand, *Detection of an anomalous cluster in a network*, The Annals of Statistics **39** (2011), no. 1, 278–304.

[ACDH05] Ery Arias-Castro, David L Donoho, and Xiaoming Huo, *Near-optimal detection of geometric objects by fast multiscale methods*, Information Theory, IEEE Transactions on **51** (2005), no. 7, 2402–2425.

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov, *Finding a large hidden clique in a random graph*, Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 1998, pp. 594–598.

[AKV02] Noga Alon, Michael Krivelevich, and Van H Vu, *On the concentration of eigenvalues of random symmetric matrices*, Israel Journal of Mathematics **131** (2002), no. 1, 259–267.

[Ang80] Dana Angluin, *Local and global properties in networks of processors*, Proceedings of the twelfth annual ACM symposium on Theory of computing, ACM, 1980, pp. 82–93.

[AV11] Brendan PW Ames and Stephen A Vavasis, *Nuclear norm minimization for the planted clique and biclique problems*, Mathematical programming **129** (2011), no. 1, 69–89.

[BDN12] Shankar Bhamidi, Partha S Dey, and Andrew B Nobel, *Energy landscape for large average submatrix detection problems in gaussian random matrices*, arXiv preprint arXiv:1211.2284 (2012).

[Bil08] Patrick Billingsley, *Probability and measure*, John Wiley & Sons, 2008.

[BLM12] M. Bayati, M. Lelarge, and A. Montanari, *Universality in message passing algorithms*, arXiv:1207.7321, 2012.

[BM11] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. on Inform. Theory **57** (2011), 764–785.

[BR13] Quentin Berthet and Philippe Rigollet, *Computational lower bounds for sparse pca*, arXiv preprint arXiv:1304.0828 (2013).

[CLMW11] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, *Robust principal component analysis?*, Journal of the ACM (JACM) **58** (2011), no. 3, 11.

[dBG08] Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui, *Optimal solutions for sparse principal component analysis*, The Journal of Machine Learning Research **9** (2008), 1269–1294.

[dEGJL07] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet, *A direct formulation for sparse pca using semidefinite programming*, SIAM review **49** (2007), no. 3, 434–448.

[Dem13] Amir Dembo, *Probability Theory*, http://www.stanford.edu/ montanar/TEACHING/Stat310A/lnotes.pdf, 2013.

[DK70] Chandler Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. iii*, SIAM Journal on Numerical Analysis **7** (1970), no. 1, pp. 1–46.

[DMM09] D. L. Donoho, A. Maleki, and A. Montanari, *Message Passing Algorithms for Compressed Sensing*, Proceedings of the National Academy of Sciences **106** (2009), 18914–18919.

[FGR+12] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao, *Statistical algorithms and a lower bound for planted clique*, arXiv preprint arXiv:1201.1214 (2012).

[FK81] Zoltán Füredi and János Komlós, *The eigenvalues of random symmetric matrices*, Combinatorica **1** (1981), no. 3, 233–241.

[FR10] Uriel Feige and Dorit Ron, *Finding hidden cliques in linear time*, DMTCS Proceedings (2010), no. 01, 189–204.

[GW06] Dongning Guo and Chih-Chun Wang, *Asymptotic mean-square optimality of belief propagation for sparse linear systems*, Information Theory Workshop, 2006. ITW'06 Chengdu. IEEE, IEEE, 2006, pp. 194–198.

[Jer92] Mark Jerrum, *Large cliques elude the metropolis process*, Random Structures & Algorithms **3** (1992), no. 4, 347–359.

[JL09] Iain M Johnstone and Arthur Yu Lu, *On consistency and sparsity for principal components analysis in high dimensions*, Journal of the American Statistical Association **104** (2009), no. 486.

[KF09] Daphne Koller and Nir Friedman, *Probabilistic graphical models: principles and techniques*, MIT press, 2009.

[KY11] Antti Knowles and Jun Yin, *The isotropic semicircle law and deformation of wigner matrices*, arXiv preprint arXiv:1110.6449 (2011).

[Lin92] Nathan Linial, *Locality in distributed graph algorithms*, SIAM Journal on Computing **21** (1992), no. 1, 193–201.

[MM09] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford, 2009.

[MT06] Andrea Montanari and David Tse, *Analysis of belief propagation for non-linear problems: The example of cdma (or: How to prove tanaka's formula)*, Information Theory Workshop, 2006. ITW'06 Punta del Este. IEEE, IEEE, 2006, pp. 160–164.

[NS95] Moni Naor and Larry Stockmeyer, *What can be computed locally?*, SIAM Journal on Computing **24** (1995), no. 6, 1259–1277.

[RF12] Sundeep Rangan and Alyson K Fletcher, *Iterative estimation of constrained rank-one matrices in noise*, Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on, IEEE, 2012, pp. 1246–1250.

[RU08] T.J. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, Cambridge, 2008.

[SN08] Xing Sun and Andrew B Nobel, *On the size and recovery of submatrices of ones in a random binary matrix*, J. Mach. Learn. Res **9** (2008), 2431–2453.

[Suo13] Jukka Suomela, *Survey of local algorithms*, ACM Computing Surveys (CSUR) **45** (2013), no. 2, 24.

[SWPN09] Andrey A Shabalin, Victor J Weigman, Charles M Perou, and Andrew B Nobel, *Finding large average submatrices in high dimensional data*, The Annals of Applied Statistics (2009), 985–1012.

[Ver12] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Compressed Sensing: Theory and Applications (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012, pp. 210–268.

[WJ08] Martin J Wainwright and Michael I Jordan, *Graphical models, exponential families, and variational inference*, Foundations and Trends in Machine Learning **1** (2008), no. 1-2, 1–305.

[YDP11] O. Gurel-Gurevich Y. Dekel and Y. Peres, *Finding Hidden Cliques in Linear Time with High Probability*, Workshop on Analytic Algorithmics and Combinatorics (San Francisco), January 2011.

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani, *Sparse principal component analysis*, Journal of computational and graphical statistics **15** (2006), no. 2, 265–286.