Reconstruction on Trees and Low-Degree Polynomials

Frederic Koehler[*]
Stanford University
fkoehler@stanford.edu

Elchanan Mossel[†]
Massachusetts Institute of Technology
elmos@mit.edu

Abstract

The study of Markov processes and broadcasting on trees has deep connections to a variety of areas including statistical physics, graphical models, phylogenetic reconstruction, Markov Chain Monte Carlo, and community detection in random graphs. Notably, the celebrated Belief Propagation (BP) algorithm achieves Bayes-optimal performance for the reconstruction problem of predicting the value of the Markov process at the root of the tree from its values at the leaves.

Recently, the analysis of low-degree polynomials has emerged as a valuable tool for predicting computational-to-statistical gaps. In this work, we investigate the performance of low-degree polynomials for the reconstruction problem on trees. Perhaps surprisingly, we show that there are simple tree models with N leaves and bounded arity where (1) nontrivial reconstruction of the root value is possible with a simple polynomial time algorithm and with robustness to noise, but not with any polynomial of degree N^c for $c > 0$ a constant depending only on the arity, and (2) when the tree is unknown and given multiple samples with correlated root assignments, nontrivial reconstruction of the root value is possible with a simple Statistical Query algorithm but not with any polynomial of degree N^c. These results clarify some of the limitations of low-degree polynomials vs. polynomial time algorithms for Bayesian estimation problems. They also complement recent work of Moitra, Mossel, and Sandon who studied the circuit complexity of Belief Propagation. As a consequence of our main result, we are able to prove a result of independent interest regarding the performance of RBF kernel ridge regression for learning to predict the root coloration: for some $c' > 0$ depending only on the arity, $\exp(N^{c'})$ many samples are needed for the kernel regression to obtain nontrivial correlation with the true regression function (BP). We pose related open questions about low-degree polynomials and the Kesten-Stigum threshold.

1 Introduction

Understanding the computational complexity of random instances has been the goal of an extensive line of research spanning multiple decades and different research areas such as cryptography, high-dimensional statistics, complexity theory, and statistical physics. In particular, this includes work on satisfiability and refutation of random constraint satisfaction problems and on computational-to-statistical gaps. In much of this work, evidence for computational hardness is indirect because there are well-known barriers to proving hardness from classical worst-case assumptions such as

[*] Supported in part by E. Mossel's Vannevar Bush Faculty Fellowship ONR-N00014-20-1-2826, NSF award CCF-1704417, NSF Award IIS-1908774, and N. Anari's Sloan Research Fellowship. Part of this work was completed while at MIT and while participating in the program *Computational Complexity of Statistical Inference* at the Simons Institute for the Theory of Computing.

[†] Supported by Simons-NSF collaboration on deep learning NSF DMS-2031883, by Vannevar Bush Faculty Fellowship award ONR-N00014-20-1-2826 and by a Simons Investigator Award in Mathematics (622132)

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

NP-hardness (Feigenbaum and Fortnow 1993; Akavia et al. 2006; Bogdanov and Trevisan 2006; Applebaum et al. 2008).

Recently, low-degree polynomials have emerged as a powerful tool for predicting computational-to-statistical gaps. Computational-to-statistical gaps are situations where it is impossible for polynomial time algorithms to estimate a desired quantity of interest from the data, even though computationally inefficient ("information-theoretic") algorithms can succeed at the same task. Heuristics based on low-degree polynomials have especially been used in the context of Bayesian estimation and testing problems and partially motivated by connections with (lower bounds for) the powerful Sum-of-Squares proof system. More specifically, a recent line of work (e.g. Hopkins and Steurer 2017; Hopkins 2018; Kunisky et al. 2019; Bandeira et al. 2020; Gamarnik et al. 2020; Holmgren and Wein 2020; Wein 2020; Bresler and Huang 2021; Mao and Wein 2021) showed that a suitable "low-degree heuristic" can be used to predict computational-statistical gaps for a variety of problems such as recovery in the multicommunity stochastic block model, sparse PCA, tensor PCA, the planted clique problem, certification in the zero-temperature Sherrington-Kirkpatrick model, the planted sparse vector problem, and for finding solutions in random k-SAT problems. Furthermore, it was observed that the predictions from this method generally agree with those conjectured using other techniques (for example, statistical physics heuristics based on studying BP/AMP fixed points, see e.g. Decelle et al. 2011; Deshpande and Montanari 2015; Mohanty et al. 2021). Some of the merits of the low-degree polynomial framework include that it is relatively easy to use (e.g. compared to proving SOS lower bounds), and that low degree polynomials capture the power of the "local algorithms" framework used in e.g. (Gamarnik and Sudan 2014; Chen et al. 2019) as well as algorithms which incorporate global information, such as spectral methods or a constant number of iterations of Approximate Message Passing (Wein 2020).

In this work, we investigate the power of low-degree polynomials for the (average case) reconstruction problem on trees. We define the model and results in the next sections, but first give an informal summary. The goal for reconstruction on trees is to estimate the value of the Markov process at the root given its value at the leaves (in the limit where the depth of the tree goes to infinity), and two key parameters of the model are the arity of the tree d and the magnitude of the second eigenvalue λ_2 of the broadcast chain. Importantly, when $d|\lambda_2|^2 > 1$ it is known (Kesten and Stigum 1966) that nontrivial reconstruction of the root is possible just from knowing the counts of the leaves of different types, whereas when $d|\lambda_2|^2 < 1$ such count statistics have no mutual information with the root (but more complex statistics of the leaves may) Mossel and Peres 2003. This threshold $d|\lambda_2|^2 = 1$ is known as the *Kesten-Stigum threshold* (Kesten and Stigum 1966) and it plays a fundamental role in other problems, such as algorithmic recovery in the stochastic block model (Abbe 2017) and phylogenetic reconstruction (Daskalakis et al. 2006). Count statistics can be viewed as degree 1 polynomials of the leaves, which begs the question of what information more general polynomials can extract from the leaves.

In this paper, we answer this question in the limit case $\lambda_2 = 0$. Perhaps surprisingly, we find that the Kesten-Stigum threshold remains tight in the sense that even polynomials of degree N^c for a small $c > 0$ are not able to correlate with the root label (Theorem 6), whereas computationally efficient reconstruction is generally possible as long as d is a sufficiently large constant (Theorem 5) and even when a constant fraction of leaves are replaced by noise. Building on the polynomial lower bound, we prove superpolynomiallly (in fact, subexponentially) many samples are needed for Gaussian Kernel Ridge Regression (KRR) to (weakly) learn to regression function which predicts the root from the leaf colorations (Theorem 20). This gives a simple and natural model where KRR provably fails that is outside the reach of existing lower bounds such as (Kamath et al. 2020).

We also consider an analogous question where the tree is unknown, and the algorithm has access to m i.i.d. samples of the Markov process where the root is biased towards an unknown label Y^*. In this setting, polynomials of degree N^c again fail to correlate with the label Y^*, but we show that a simple algorithm, straightforwardly implementable in the Statistical Query (SQ) model (Kearns 1998), can recover Y^* in polynomial time (Theorem 11). Together, these results show that low-degree polynomials behave very differently in our setting than one might intuit based on previous work in related settings, such as in random constraint satisfaction problems or the block model.

1.1 Preliminaries

Notation. We use the standard notation $O_a(\cdot)$ to denote an upper bound with an implied constant which is allowed to depend on a; the notation $\mathrm{poly}_a(\cdot)$ is similar for denoting a bound which is polynomial in its parameters. We let $d_{TV}(P, Q)$ denote the total variation distance between distributions P and Q normalized to be in $[0, 1]$ and we use $I(X; Y \mid Z)$ for the conditional mutual information of random variables X, Y conditional on Z; see (Cover 1999). Given a vector x and a subset of coordinates S, we let x_S denote the $|S|$-dimensional vector corresponding to elements of S.

Markov processes on trees. We consider Markov processes with state space $\Sigma = [q] = \{1, \ldots, q\}$ where $q \geq 1$ is the size of the alphabet. Let $M : q \times q$ be the transition matrix of a time-homogeneous Markov chain on Σ, also referred to as the *broadcast channel*. For simplicity, we always assume henceforth that M is *ergodic* (irreducible and aperiodic, see (Durrett 2019)) so it has a unique stationary distribution π_M. Let $T = (V, E, \rho)$ be a rooted tree with vertex set V, root $\rho \in V$, and where E is the set of directed edges (u, v) where u is the parent of v in the corresponding tree. The *broadcast process* on tree T of depth ℓ with transition matrix M and root prior ν a probability measure on $[q]$ is given by

$$\mu_{\ell,\nu}(x) := \nu(x_\rho) \prod_{(u,v)\in E} M_{x_u,x_v}.$$

When not otherwise indicated, ν is the stationary distribution for M. The probability measure μ is a *Markov Random Field* on the tree T. This means that if A, B are subsets of the vertices of T and all paths in T from A to B pass through a third set of vertices S, then X_A and X_B are conditionally independent given X_S. This is called the *Markov property*, see e.g. (Lauritzen 1996). In this paper, we focus on the setting of complete d-ary trees (i.e. trees where every non-leaf node has d children, and all leaf nodes are at the same depth). For the d-ary tree of depth $\ell \geq 0$, we let L be the set of leaves of the tree, i.e. the set of vertices in the tree at depth ℓ.

Definition 1. We say that *reconstruction is possible* on the d-ary tree with channel M if

$$\inf_{\ell\geq 1} \max_{c,c'\in[q]} d_{TV}\Big(\mathcal{L}_{\mu_\ell}(X_L \mid X_\rho = c), \mathcal{L}_{\mu_\ell}(X_L \mid X_\rho = c')\Big) > 0$$

where the notation $\mathcal{L}_\mu(X|E)$ denotes the conditional law of X under μ given event E occurs, μ_ℓ is the corresponding broadcast process on the depth ℓ tree with root ρ and $L = L_\ell$ is the set of leaves.

When reconstruction is possible, the Bayes-optimal estimate of the root given the leaves can be computed in linear time by passing messages up the tree using the Belief Propagation algorithm (Mezard and Montanari 2009); for our purposes, we will not need the explicit formula for BP, which can be derived by applying Bayes rule, but refer the interested reader to the reference.

Given a matrix M, we let $\lambda_2(M)$ denote the second-largest eigenvalue of M in absolute value. The *Kesten-Stigum (KS) threshold* on the d-ary tree is given by the equation $d|\lambda_2(M)|^2 = 1$. Building upon the original work of (Kesten and Stigum 1966), it was shown that the KS threshold is sharp for the problem of *count reconstruction* on trees (Mossel and Peres 2003): count reconstruction is possible when $d|\lambda_2(M)|^2 > 1$ and impossible when $d|\lambda_2(M)|^2 < 1$.

Definition 2. Let $C(x) := (\#\{i : x_i = c\})_{c\in[q]}$ be the function which computes count statistics of an input vector x with entries in $[q]$. We say that *count-reconstruction is possible* on the d-ary tree with channel M if $\inf_{\ell\geq 1} \max_{c,c'\in[q]} d_{TV}(\mathcal{L}_{\mu_\ell}(C(X_L) \mid X_\rho = c), \mathcal{L}_{\mu_\ell}(C(X_L) \mid X_\rho = c')) > 0$ where the notation $\mathcal{L}(X|E)$ denotes the conditional law of X given event E, μ_ℓ is the corresponding broadcast process on the depth ℓ tree and $L = L_\ell$ is the set of leaves on this tree.

Next, we define a notion of noisy reconstruction which plays an important role in this paper:

Definition 3. For $\epsilon \in (0, 1)$, we say that *ϵ-noisy reconstruction is possible* on the d-ary tree with channel M if $\inf_{\ell\geq 1} \max_{c,c'} d_{TV}(\mathcal{L}_{\mu_\ell}(X'_L = \cdot \mid X_\rho = c), \mathcal{L}_{\mu_\ell}(X'_L = \cdot \mid X_\rho = c')) > 0$ where X' is the ϵ-noisy version of the broadcast process values X, generated by independently for each vertex v, setting $(X'_L)_v = (X_L)_v$ with probability $1 - \epsilon$ and otherwise sampling $(X'_L)_v$ from $Uni([q])$[3].

[3] More generally, our results hold where the noise is from any full support distribution on $[q]$.

Note that in this definition, the law of $X'_L \mid X_\rho = c$ can equivalently be written as $\mathcal{L}_{\mu_\ell}(X_L \mid X_\rho = c)T_\epsilon$ where T_ϵ is the usual noise operator that independently resamples each coordinate of its input vector with probability ϵ, see e.g. O'Donnell 2014; Hopkins 2018. Finally, we recall from (Janson and Mossel 2004) the following standard definition: we say that *robust reconstruction* on the d-ary tree with channel M is possible if ϵ-noisy reconstruction is possible *for every* $\epsilon \in (0, 1)$.

Low-degree polynomials and computational-statistical gaps. As discussed in the introduction, low degree polynomials have been studied in a wide variety of contexts and settings. The recent work (Schramm and Wein 2020) showed that a version of the low-degree polynomial heuristic can predict the recovery threshold for natural Bayesian estimation problems, even when the recovery threshold is below the detection/testing threshold. In the present work, we will use the following key definition from their paper[4]

Definition 4 (Degree-D Maximum Correlation (Schramm and Wein 2020))**.** Suppose that $(X, Y) \sim P$ where X is a random vector in $\mathbb{R}^N$ and Y is a random variable valued in $\mathbb{R}$. The *degree-D maximum correlation* is defined to be

$$\mathrm{Corr}_{\leq D}(P) := \sup_{f \in \mathbb{R}[X]_{\leq D}, \mathbb{E}_P[f(X)^2] \neq 0} \frac{\mathbb{E}_P[f(X) \cdot Y]}{\sqrt{\mathbb{E}_P[f(X)^2]}}$$

where $\mathbb{R}[X]_{\leq D}$ is the space of degree at most D multivariate polynomials in variables $X_1, \ldots, X_N$ with real-valued coefficients.

As explained there, when the target label Y^* is a vector this definition can be applied with Y equal to each of the coordinates of Y^*. We note that we could rephrase our results in terms of a testing problem (as in much of the prior work on the low-degree method), but the above definition is more natural in our context (it avoids the need to introduce a "null distribution" Q). In what follows, we omit the distribution P the expectation is taken over as long as it is clear from context.

Instead of referring to polynomial degree directly, we usually use the following more convenient and equivalent definition. Suppose f is a function $[q]^n \to \mathbb{R}$. We define the (Efron-Stein) *degree* of f to be the minimal D such that there exist functions $f_S : [q]^{|S|} \to \mathbb{R}$ so that $f(x) = \sum_{S \subset [n], |S| \leq D} f_S(x_S)$. One such minimal choice of f_S is the Efron-Stein decomposition over $Uni[q]^n$, see e.g. (O'Donnell 2014); this notion is also equivalent to the minimal degree polynomial representing f where the variables are the one-hot encoding $x \mapsto (\mathbb{1}(x_i = c))_{i \in [n], c \in [q]}$.

Reconstruction below the KS threshold. In this paper, we will largely consider the problem of tree reconstruction with matrices M with $\lambda_2(M) = 0$; these exactly correspond to Markov chains which mix perfectly within a bounded number of steps. (There are many examples of such chains, for concreteness we give a very small example below in Example 13). Obviously, for such a chain M, $d|\lambda_2|^2 = 0$ for any value of d so such a model is always below the Kesten-Stigum threshold. Nevertheless, based on general results from existing work we know that near-perfect reconstruction of the root is possible (e.g. using Belief Propagation, which computes the exact posterior distribution (Mezard and Montanari 2009)). This is true as long as d is sufficiently large, and even with a constant amount of noise ϵ:

Theorem 5 (Mossel and Peres 2003, Theorem 22 below)**.** *Suppose M is a the transition matrix of a Markov chain with pairwise distinct rows*[5] *i.e. for all $i, j \in [q]$ the rows M_i and M_j are distinct vectors. Let $\delta \in (0, 1)$ be arbitrary. There exists $d_0 = d_0(M, \delta)$ and $\epsilon > 0$ such that for all $d \geq d_0$, ϵ-noisy reconstruction is possible on the d-ary tree and furthermore there exists a polynomial-time computable function $f = f_{M,\ell}$ valued in $[q]$ such that*

$$\max_{c \in [q]} \Pr(f(X'_L) \neq X_\rho \mid X_\rho = c) < \delta$$

where X'_L is the ϵ-noisy version of X_L (see Definition 3).

This exact statement does not appear in Mossel and Peres 2003 but follows from arguments presented there; for completeness, we include a proof in the appendix (Theorem 22). From the proof, we can see that a very simple recursive estimator is enough to solve this problem.

[4] In our notation X and y are swapped compared to theirs, to match the convention in the broadcast process.

[5] This condition is needed to rule out the case of e.g. a rank one matrix M where reconstruction is clearly impossible. See also (Mossel and Peres 2003) for a more complex and precise condition.

1.2 Our Results

We study the power of low-degree polynomials for the problem of reconstructing the root of a Markov process. We consider this question in the context of two very closely related versions of the model which have both been extensively studied in the literature.

Reconstruction with a known tree. In this setting, the algorithm is given access to the leaf values from a single realization of the Markov process, and the goal is to estimate the root (where we are going to be interested in estimators which are low-degree polynomials of the leaves). The tree structure is known and the estimator/polynomial is allowed to depend on this information directly.

Reconstruction with an unknown tree. In this setting, the data is still generated by a complete d-ary tree but the tree (in other words, the true ordering of the leaves) is unknown to the algorithm. This version of the model has been extensively studied due to close connections to the problem of phylogenetic reconstruction in biology, see e.g. Felsenstein 2004; Daskalakis et al. 2006; Steel 2016. Because this task is more difficult information-theoretically[6], the algorithm is given access to m i.i.d. samples from the broadcast model; we give a more precise definition of the model below.

Results for reconstruction with a known tree. We consider the problem of tree reconstruction with matrices M with $\lambda_2(M) = 0$; these exactly correspond to Markov chains which mix perfectly within a bounded number of steps. As discussed above in Preliminaries, while such models are always below the Kesten-Stigum threshold for any value of the the arity d, under fairly weak conditions on M the reconstruction problem is still solvable for d sufficiently large (Theorem 5). This is true even with noise and with a very simple reconstruction algorithm. As our main result we show that despite the algorithmic tractability of this problem, only very high degree polynomials are able to get any correlation with the root, in the same sense as Definition 4.

Theorem 6 (Corollary 15 below)**.** *Let M be the transition matrix of a Markov chain on $[q]$ and suppose that $1 \le k \le q$ is such that M^k is a rank-one matrix. For any function $f : [q]^L \to \mathbb{R}$ of Efron-Stein degree at most $2^{\lfloor \ell/(k-1) \rfloor}$ of the leaves X_L and any prior ν on the root,*

$$\mathbb{E}[f(X_L) \cdot (\mathbb{1}(X_\rho = c) - \nu(c))] = 0.$$

Remark 7 (Tightness). For fixed k, this result is tight up to the base of the exponent. When M satisfies the assumption of Theorem 5, there is a function on a constant-degree subtree to recover the root and (by Fourier expansion) this is a polynomial of degree $e^{O(\ell)}$.

To interpret this result, observe that $N = d^\ell$, so taking d a constant, the Theorem shows that polynomials of degree even N^c for an explicit constant $c = c(d, k) > 0$ fail to get any correlation with the root label. In comparison, in the previously mentioned contexts in the low-degree polynomials literature, the threshold for polynomials of degree $O(\log N)$ matches the conjectured threshold for polynomial time algorithms (see e.g. Hopkins and Steurer 2017; Hopkins 2018; Kunisky et al. 2019) and polynomials of degree N^c correspond to conjectural thresholds for subexponential time algorithms (see e.g. Ding et al. 2019; Bandeira et al. 2020).

A consequence: subexponential sample complexity lower bound for the RBF kernel. As a consequence of our main result, we can analyze the behavior of kernel regression methods in our model. Kernel ridge regression is one of the canonical methods for solving supervised learning problems, including classification problems (see e.g. Muthukumar et al. 2021). In many high-dimensional settings, it is believed that the function learned using a standard kernel (e.g. Gaussian or polynomial) is essentially a low-degree polynomial. Standard results in learning theory (Shalev-Shwartz and Ben-David 2014) imply that kernel ridge regression with a Gaussian/RBF (Radial Basis Function) kernel in N dimensions can learn a degree ℓ polynomial on the hypercube or sphere using roughly $O(N^\ell)$ samples. Establishing lower bounds on KRR is generally *much harder*. In certain particularly tractable settings (e.g. data from the uniform distribution on the hypercube) it has been recently shown explicitly that kernel regression (only) learns a low-degree polynomial (Ghorbani et al. 2021; Mei et al. 2021).

It seems plausible to guess that kernel ridge regression with a standard kernel will require subexponentially many samples of (leaf label, root label) pairs in order to learn to predict the root. We

[6] Note that in the single-sample case ($m = 1$), the information available to the algorithm would only be count statistics, which we know are insufficient for reconstruction below the KS threshold (Mossel 2004).

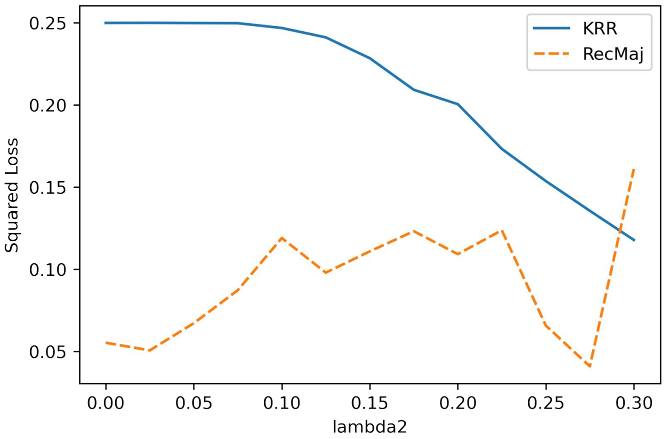

Figure 1: RBF Kernel Ridge Regression (KRR, blue line) test error in squared loss for predicting the root from leaves with data generated by broadcast operator $M_{\lambda_2} := (1-\lambda_2)M_0 + \lambda_2 I$ and varying λ_2. M_0 is from (1), and it can be directly checked that λ_2 is the second eigenvalue of M_{λ_2}. The tree is 10-ary with depth 3, and the prior is $\nu = (0.5, 0.5, 0)$. RecMaj (orange dotted line) is a baseline estimator which generalizes the one used in proof of Theorem 5. Consistent with Theorem 8, the output of KRR fails to correlate with the root coloration when $\lambda_2 = 0$ (since 0.25 is the *null risk*, the squared loss for the optimal constant predictor), even though RecMaj correlates significantly with the root for all values of λ_2. In fact, KRR fails to correlate for all values of λ_2 up to around 0.1, suggesting that the failure of KRR and low-degree polynomials should extend beyond $\lambda_2 = 0$.

are able to verify this prediction in the case of the popular RBF (Radial Basis Function) kernel. See Section C.1.1 for formal notation and background on kernel ridge regression.

Theorem 8 (Theorem 20 below)**.** *Let M be the transition matrix of a Markov chain on $[q]$ and suppose that $1 \leq k \leq q$ is such that $M^k = \pi\pi^T$ is a rank-one matrix, and suppose that π has at least two nonzero entries. Then the for any color $c \in [q]$ and prior ν for the root coloration X_ρ, the following is true. Given m i.i.d. samples $(x_1, y_1), \dots, (x_m, y_m)$ from the broadcast model on the d-ary tree with N leaves and broadcast channel M, where x_i is a one-hot encoded vector of leaf colorations and $y_i = 1(X_\rho = c) - \nu(c)$ is the centered indicator of the leaf coloration, we have that for any bandwidth $\sigma \geq 0$ and ridge parameter $\lambda \geq 0$, for w the output of ridge regression in RKHS space with those parameters and feature map φ, that with probability at least $1-\delta$*

$$\frac{\mathbb{E}_{x_0,y_0}[y_0\langle w, \varphi(x_0)\rangle]}{\sqrt{\mathbb{E}_{x_0,y_0}[y_0^2]}} = O(\sqrt{1/N})$$

provided that $m/\delta = O(e^{N^a})$ where $a = a(M,d) > 0$ is independent of the depth of the tree.

This establishes a new and illustrative example where KRR performs poorly in high dimensions, even though the ground truth is a relatively "simple" and the labels are closely related to the structure of the input data. Note that the conclusion implies that $\mathbb{E}[(y_0 - \langle w, \varphi(x_0)\rangle)^2] \geq (1 - O(1/\sqrt{N}))\mathbb{E}[y_0^2]$, i..e. kernel ridge regression does not significantly outperform the constant zero estimator ("null risk") unless it is given at least a subexponential number of samples. Also, as with Remark 7 this result is tight up to the power of the exponent c, since a subexponential degree polynomial exists which predicts the root well and it can provably be learned with subexponential number of samples by KRR (Shalev-Shwartz and Ben-David 2014).

In Figure 1, we test kernel ridge regression in a simulation in both the case $\lambda_2 = 0$ and $\lambda_2 > 0$: consistent with our result, KRR fails to beat the null risk when $\lambda_2 = 0$; interestingly, it also fails for moderately small values of λ_2 as well, which is related to the Open Problem we discuss later. In the figure, KRR is performed using 2000 i.i.d. samples of (x, y) pairs with x the one-hot encoded leaf colorations and y the centered indicator that the root color is 1, as in Theorem 8. Bandwidth and ridge penalty are selected via grid search on a validation set. The results for the baseline (RecMaj) are averaged over 16000 samples.

Results for reconstruction with an unknown tree. Formally, we consider the following variant of the generative model which is a variant of models in the phylogenetics literature. Briefly, in this model we generate m i.i.d. realizations of the broadcasting model, where the tree is random and the prior on the root is biased towards a random root label. We include a parameter $\epsilon \geq 0$ which can be used to add noise to the final output the of model, just as above.

Definition 9 (ϵ-Noisy Repeated Broadcast Model on Random Tree)**.** Let $\ell \geq 1, d \geq 2, m \geq 1, q \geq 1, \epsilon \geq 0$ and let M be a Markov chain on $[q]$. Define $R = R_{\ell,d,m,M,\epsilon}$ by the following process:

1. Sample $Y^* \sim Uni([q])$, and $\tau \sim Uni(S_N)$ is a random permutation. Let T be the d-ary tree on the set of leaves ordered by τ, so e.g. vertices $\tau(1)$ and $\tau(2)$ are siblings in T.
2. Sample $X^{(1)}, \dots, X^{(m)}$ i.i.d. from the ϵ-noisy broadcast process (see Definition 3) on T with prior $(2/3)\delta_{Y^*} + (1/3)Uni([q])$ and transition matrix M, where δ_{Y^*} is a delta distribution on Y^*. Let $\mathbb{X} = (X_L^{(1)}, \dots, X_L^{(m)})$.

The goal of the learning algorithm in the unknown tree model is this: given m samples of the leaves of the broadcast process, encoded in $\mathbb{X}$, reconstruct the root label Y^* which the prior is biased towards[7]. We discuss the reasons for defining the model this way: 1. The permutation τ ensures that the coordinates of $X_L^{(i)}$ behave in a symmetric way, or equivalently that the order of those coordinates is not semantically meaningful; observe that if we omitted it, then the first d coordinates would always be neighbors in the tree. This is standard in the phylogenetics literature (Steel 2016) and this kind of symmetry is also assumed in the literature on low-degree polynomial hardness, see e.g. discussion in Holmgren and Wein 2020; in sparse PCA this is analogous to how the support of the planted sparse vector is chosen uniformly at random among size-k subsets. 2. The choice that root assignments are drawn from a tilted/biased distribution is different from the previous literature motivated by phylogenetics, where the root value is generally sampled fresh each time. This does not have a significant effect on how the algorithms used to estimate the tree work. The reason for our setup is to allow for straightforward comparison between SQ and low-degree polynomial models. If the root value was sampled from an unbiased measure each time, it would not make sense for an SQ algorithm to estimate it, since SQ has no concept of individual samples.

To be formal, we define the Statistical Query VSTAT oracle analogue of R in the usual way (Feldman et al. 2017). The oracle is defined conditional on Y^* and the tree T, so the order of leaves in the tree will be consistent between different calls to the oracle. As a reminder, vector-valued queries are implemented in the SQ model by querying each coordinate of the vector individually.

Definition 10 ($VSTAT(m)$ Oracle)**.** Let Y^*, τ, T, and M be as in Definition 9. Conditional on $Y^* = y^*$ and the tree $T = t$, we define $VSTAT(m)$ to be an arbitrary oracle which given a query function $\varphi : [q]^L \to [0,1]$, returns $p + \zeta_\varphi$ where $p := \mathbb{E}_R[\varphi(X^{(1)}) \mid Y^* = y^*, T = t]$ where ζ_φ is arbitrary (can be adversarially chosen) such that $|\zeta_\varphi| \leq \max\left(\frac{1}{m}, \sqrt{\frac{p(1-p)}{m}}\right)$.

We now state our results in this model. Just as in the known tree case, there is a relatively simple algorithm which achieves nearly optimal performance in this setting when $\lambda_2(M) = 0$ and d is a large constant, and furthermore this algorithm can straightforwardly be implemented in the SQ model described above. Establishing this requires proving a new result in tree reconstruction, since (for example) the setting $\lambda_2(M) = 0$ which we care about rules out the use of Steel's evolutionary distance (see e.g. Steel 2016; Moitra 2018) commonly used in reconstruction algorithms in phylogeny, as Steel's distance is only well-defined for nonsingular phylogenies, and some kinds of tree models with singular matrices are actually computationally hard to learn (Mossel and Roch 2005).

Theorem 11 (Theorem 30 below)**.** *Suppose M is a the transition matrix of a Markov chain with pairwise distinct rows, i.e. for all $i, j \in [q]$ the rows M_i and M_j are distinct vectors, and suppose $\lambda_2(M) = 0$. There exists $d \geq 1$ and $\epsilon > 0$ so that the following result holds true for the complete d-ary tree with any depth $\ell \geq 1$. For any $\delta > 0$, there exist a polynomial time algorithm with sample complexity $m = poly_M(\log N, \log(1/\delta))$ from the ϵ-noisy repeated broadcast model (Definition 9) which with probability at least $1 - \delta$: 1) outputs the true tree T (equivalently, the true permutation τ), 2) outputs $\hat{Y}$ such that $\hat{Y} = Y^*$. Also, this algorithm can be implemented using a $VSTAT(m)$ oracle with $m = poly_M(\log(N/\delta))$ and polynomially many queries.*

[7] We could also consider the model where the root label is always Y^*. The soft bias we consider is nicer for minor technical reasons, and seems natural given we allow to add noise elsewhere in the model.

The lower bound for polynomials also applies here, just like in Theorem 6. Since it is very similar, we leave the formal statement to the appendix (Theorem 27).

1.3 Further Discussion

Related work: complexity of reconstruction on trees. Our work follows a line of previous work which identified the Kesten-Stigum threshold as a potential complexity barrier in the context of the broadcast model on trees. The work (Mossel 2016) showed that algorithms that do not use correlation between different features (named "shallow algorithms") cannot recover phylognies above the Kesten-Stigum threshold, where other ("deeper") algorithms can do so efficiently; the motivation in (Mossel 2016) was to find simple data models where depth is needed for inference. More standard complexity measures were studied in Moitra et al. 2020 who obtained a number of results on the circuit complexity of inferring the root in the broadcast process. They conjectured that below the Kesten-Stigum threshold, inferring the root is $NC1$-complete and proved it for one specific chain satisfying $\lambda_2 = 0$. Although there are some connections between low-degree polynomials and certain circuit classes (Linial et al. 1993), the results of this work and (Moitra et al. 2020) are incomparable and the techniques for establishing the lower bound are very different. Finally, we note the work (Jain et al. 2019) which studied the power of message-passing algorithms on finite alphabets: they proved such algorithms fail to recover all the way down to the Kesten-Stigum threshold, even in the simplest case of the binary symmetric channel with $q = 2$.

Message passing vs. low-degree polynomials. One of the attractive properties of the class of low-degree polynomials is that it generally captures the power of a constant (or sufficiently slowly growing) number of iterations of message-passing algorithms such as BP, AMP, and Survey Propagation (see e.g. Bresler and Huang 2021 and Appendix A of Gamarnik et al. 2020), which is interesting since a constant number of steps of these algorithms are indeed useful for many statistical tasks. On the other hand, in our models, belief propagation (which computes the exact posterior) succeeds with high probability whereas low-degree polynomials fail. This is not a contradiction: in our setting, BP requires $\Theta(\log N)$ iterations for the messages to pass from the leaves to the root and this is (as our main result shows) too large to simulate with low-degree polynomials.

SQ and Low-Degree Polynomials. The recent work (Brennan et al. 2021) established sufficient conditions for predictions to match between the Statistical Query (SQ) and low-degree polynomial heuristic, in a general setting. Nevertheless, in the unknown tree setting we consider above we saw that SQ algorithms perform significantly better than low-degree polynomials. The results of (Brennan et al. 2021) cannot be immediately applied to our setting, because we have phrased the problem as an estimation problem instead of a testing (a.k.a. distinguishing) problem; however, this is itself not the reason for the discrepancy as we could rephrase our problem in terms of testing the color of the root. Instead, the reason seems to be due to the "niceness condition" needed for their theory to apply. They show that the niceness condition will be satisfied for noise-robust problems when the "null distribution" in the testing problem is a product measure. Our setup is indeed noise robust (see Theorem 5). However, if we rephrased our problem as a testing one the null distribution will be a graphical model (with no bias at the root) and not a product measure.

Recently there has also been interest in understanding lower bounds against kernel learning algorithms (including polynomial kernels), in part motivated by connections to neural networks, and this involves connections to the SQ framework. See e.g. (Kamath et al. 2020) and references within. These methods can, for example, prove strong lower bounds against learning parities with kernel ridge regression since parities have large SQ-dimension. See also (Koehler and Risteski 2018) for another example where polynomial degree lower bounds were established for a Bayesian inference task, though only polylogarithmic in the dimension.

Noise robustness and learning parities. It was shown in (Janson and Mossel 2004) that the KS threshold is sharp for *robust reconstruction*, where we recall from above that (by definition) robust reconstruction on the d-ary tree with channel M is possible if *for all* noise levels $\epsilon \in (0, 1)$, ϵ-noisy reconstruction is possible. At first glance, this appears similar to the idea in the low-degree polynomials literature that the model should be *slightly noisy* to rule out the example of learning parities (which can be solved in the noiseless setting by Gaussian elimination, but not when there is noise). In fact the two notions are quite different: the robust reconstruction result shows that reconstruction becomes impossible for *large noise levels* $\epsilon > \epsilon_*$ above a critical threshold ϵ_*, whereas for small (but fixed) $\epsilon > 0$ the ϵ-noisy reconstruction problem often *remains solvable* — see Theorem 5. Our examples are fundamentally different to the parity example: (1) for the unknown tree version of our model, we showed that the problem is solvable with an SQ oracle whereas parities are well-known to be hard for SQ (Blum et al. 2003), (2) relatedly, the algorithms which solve our problems are not "algebraic" in nature, and (3) our results hold irrespective of adding a small amount of noise, whereas learning parities with any constant amount of noise is conjecturally hard (Valiant 2012). Altogether, we can think of these results as suggesting a new, more nuanced picture of low degree vs. robustness to noise. Whereas before the main dichotomy in the computational complexity of inference literature has been between zero noise and any noise, in our model reconstruction algorithms such as BP can tolerate a small amount of noise, but fail when the noise level crosses some critical threshold. At least in our setting with $|\lambda_2(M)| = 0$, low-degree polynomials appear to capture the latter "large noise" behavior instead of the "small noise" difficulty of the problem.

Open Problem. What happens when $\lambda_2(M) \neq 0$? It is natural to wonder if the Kesten-Stigum threshold $d|\lambda_2|^2 = 1$ is sharp for low-degree polynomial reconstruction, analogous to how it is sharp for robust reconstruction. Our main lower bound result (Theorem 6) is consistent with this intuition. Also consistent with this intuition, our simulation result Figure 1 suggests that Kernel Ridge Regression may continue to fail for small but nonzero values of $\lambda_2(M)$. See Appendix B for a more formal statement and more discussion of this important question.

2 Technical Overview

The detailed proofs of all results are given in the Appendix. Here, we explain the high-level proof ideas, which we believe are relatively clean and conceptual.Before proceeding, we give the following concrete example of a Markov chain M_0 with $\lambda_2(M_0) = 0$ and $q = 3$:

$$M_0 = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 0.25 & 0.5 & 0.25 \\ 0 & 1 & 0 \end{bmatrix}, \qquad M_0^2 = \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.5 & 0.25 \end{bmatrix}. \tag{1}$$

Since M_0^2 is rank one, it must be the case that $\lambda_2(M_0) = 0$.

Failure of low-degree polynomials (Theorem 6). We want to show that any low-degree polynomial f of the leaves of the broadcast tree fails to correlate with the root. In general, it may be very difficult to compute the maximal correlation among all low-degree polynomials; what makes it possible in our case is that the correlation is exactly zero. If $c \in [q]$ is a color and ν is the prior at the root, we want to show $\mathbb{E}[f(X_L)(1(X_\rho = c) - \nu(c))] = 0$. (Recall L is the set of leaves and X_L the leaf colorations.) The first step is to use linear of expectation to break $f(X)$ into monomials: more formally, if $f(X) = \sum_{|S| \le D} f_S(X)$ is the Efron-Stein decomposition for a polynomial of degree D, then to show the goal it clearly suffices to show

$$\mathbb{E}[f_S(X_L)(1(X_\rho = c) - \nu(c))] = 0.$$

Crucially, the monomial f_S is a function which depends only on a set of at most D leaf colorations X_D. Therefore, the result follows if we can show those leaves by themselves are independent of the root coloration. This is shown by performing an *iterative trimming* procedure on the minimal subtree spanned by the root and the leaves in S: every time there is an isolated path of length k (where M^k is rank one: $k = 2$ in the example above) all information is lost from the start of the path to its end. Using this idea and some elementary combinatorics, we can prove that if $|S|$ is small, the trimming procedure will delete everything, and so the root is indeed independent of these leaves.

Failure of RBF Kernel Ridge Regression (Theorem 8.) This result builds on the low-degree polynomials result. First, we show that if the bandwidth parameter in the kernel is taken too small, then the output of Kernel Ridge Regression (KRR) is close to zero on a new test point and so it fails to learn anything. Otherwise, we can directly show that any function with a substantial high degree polynomial component has large RKHS norm. We can also construct an interpolator of the training data which has much smaller RKHS norm, by showing that every training sample has a small "fingerprint" which uniquely identifies it and is detectable with a low-degree polynomial. It then follows that whatever the output of KRR is, it must have a small RKHS norm and cannot correlate with the true regression function.

Success of noise robust reconstruction using "high degree" algorithms (Mossel and Peres 2003). We briefly explain why noise robust reconstruction is possible with simple and computationally efficient algorithms. The key is to consider the case of a depth 1 tree: because the rows of M are distinct, if the degree of the tree is a sufficiently large constant, then by the Law of Large Numbers the empirical distribution of its children will be close to the row of M corresponding to the state of the parent, letting us reconstruct the parent with say 99.9% probability of success. Given this, it is not too hard to argue this argument works recursively and in the presence of a small adversarial noise. Note that this algorithm recursively integrates *global* information on the tree across multiple scales — in contrast, the lower bound used the fact that low-degree polynomials can only aggregate information between small sets of variables in a limited (linear) way.

Unknown tree results (Theorem 11 and Theorem 27). The lower bound for low-degree polynomials in this setting can be reduced to the previous low-degree polynomial lower bound, which leaves proving that efficient (and SQ) algorithms can successfully solve this problem. Once we reconstruct the tree, we can run any algorithm for reconstructing the root given the leaves, e.g. the one described just above or BP. As far as reconstructing the tree, we first explain how to reconstruct the first layer. We prove that the joint distribution of any two leaves has enough information to tell us if they are immediate neighbors, which determines the location of all of their parents in the tree (such a test is easy to construct if we look at the *generalized eigenvectors* of M). Now that the bottom layer of the tree structure is determined, we use the fact that we have very good estimates of their parents colorations using the algorithm described before. Crucially, since that algorithm's accuracy guarantee for reconstructing the internal node's colors is very strong and does not decay as we go further and further up the tree, we can indeed apply this argument recursively to get the whole tree.

Implementing this algorithmic approach in the SQ framework is straightforward: at the end of the day it is based on computing the joint distributions of pairs of (estimated) vertex colorations, and those are all averages over the data. On the other hand, note that this method very strongly relies on the ability of an SQ algorithm to make *adaptive queries*, since the queries made are based on the partially reconstructed tree structure, which is unknown to the algorithm before it starts.

References

Neyman, Jerzy and Egon Sharpe Pearson (1933). "IX. On the problem of the most efficient tests of statistical hypotheses". *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706, pp. 289–337.

Kesten, Harry and Bernt P Stigum (1966). "Additional limit theorems for indecomposable multidimensional Galton-Watson processes". *The Annals of Mathematical Statistics* 37.6, pp. 1463–1481.

Feigenbaum, Joan and Lance Fortnow (1993). "Random-self-reducibility of complete sets". *SIAM Journal on Computing* 22.5, pp. 994–1005.

Linial, Nathan, Yishay Mansour, and Noam Nisan (1993). "Constant depth circuits, Fourier transform, and learnability". *Journal of the ACM (JACM)* 40.3, pp. 607–620.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Kearns, Michael (1998). "Efficient noise-tolerant learning from statistical queries". *Journal of the ACM (JACM)* 45.6, pp. 983–1006.

Cover, Thomas M (1999). *Elements of information theory*. John Wiley & Sons.

Mossel, Elchanan (2001). "Reconstruction on trees: beating the second eigenvalue". *Annals of Applied Probability*, pp. 285–300.

Blum, Avrim, Adam Kalai, and Hal Wasserman (2003). "Noise-tolerant learning, the parity problem, and the statistical query model". *Journal of the ACM (JACM)* 50.4, pp. 506–519.

Mossel, Elchanan and Yuval Peres (2003). "Information flow on trees". *The Annals of Applied Probability* 13.3, pp. 817–844.

Felsenstein, J. (2004). *Inferring Phylogenies*. New York, New York: Sinauer.

Janson, Svante and Elchanan Mossel (2004). "Robust reconstruction on trees is determined by the second eigenvalue". *The Annals of Probability* 32.3B, pp. 2630–2649.

Mossel, Elchanan (2004). "Survey: Information flow on trees". *arXiv preprint math/0406446*.

Mossel, Elchanan and Sébastien Roch (2005). "Learning nonsingular phylogenies and hidden Markov models". *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 366–375.

Akavia, Adi, Oded Goldreich, Shafi Goldwasser, and Dana Moshkovitz (2006). “On basing one-way functions on NP-hardness”. *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pp. 701–710.

Bogdanov, Andrej and Luca Trevisan (2006). “On worst-case to average-case reductions for NP problems”. *SIAM Journal on Computing* 36.4, pp. 1119–1159.

Daskalakis, Constantinos, Elchanan Mossel, and Sébastien Roch (2006). “Optimal phylogenetic reconstruction”. *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pp. 159–168.

Applebaum, Benny, Boaz Barak, and David Xiao (2008). “On basing lower-bounds for learning on worst-case assumptions”. *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, pp. 211–220.

Mezard, Marc and Andrea Montanari (2009). *Information, physics, and computation*. Oxford University Press.

Dembo, Amir and Ofer Zeitouni (2010). *Large Deviations Techniques and Applications*. Springer.

Artin, Michael (2011). *Algebra*. Pearson.

Decelle, Aurelien, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová (2011). “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications”. *Physical Review E* 84.6, p. 066106.

Valiant, Gregory (2012). “Finding correlations in subquadratic time, with applications to learning parities and juntas”. *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE, pp. 11–20.

Gamarnik, David and Madhu Sudan (2014). “Limits of local algorithms over sparse random graphs”. *Proceedings of the 5th conference on Innovations in theoretical computer science*, pp. 369–376.

Mossel, Elchanan, Joe Neeman, and Allan Sly (2014). “Belief propagation, robust reconstruction and optimal recovery of block models”. *Conference on Learning Theory*. PMLR, pp. 356–370.

O’Donnell, Ryan (2014). *Analysis of boolean functions*. Cambridge University Press.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Deshpande, Yash and Andrea Montanari (2015). “Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time”. *Foundations of Computational Mathematics* 15.4, pp. 1069–1128.

Mossel, Elchanan (2016). “Deep learning and hierarchal generative models”. *arXiv preprint arXiv:1612.09057*.

Steel, Mike (2016). *Phylogeny: discrete and random processes in evolution*. SIAM.

Abbe, Emmanuel (2017). “Community detection and stochastic block models: recent developments”. *The Journal of Machine Learning Research* 18.1, pp. 6446–6531.

Feldman, Vitaly, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao (2017). “Statistical algorithms and a lower bound for detecting planted cliques”. *Journal of the ACM (JACM)* 64.2, pp. 1–37.

Hopkins, Samuel and David Steurer (2017). “Efficient bayesian estimation from few samples: community detection and related problems”. *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 379–390.

Hopkins, Samuel (2018). “Statistical inference and the sum of squares method”. PhD thesis. Cornell University.

Koehler, Frederic and Andrej Risteski (2018). “Representational power of ReLU networks and polynomial kernels: beyond worst-case analysis”. *arXiv preprint arXiv:1805.11405*.

Moitra, Ankur (2018). *Algorithmic aspects of machine learning*. Cambridge University Press.

Vershynin, Roman (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.

Chen, Wei-Kuo, David Gamarnik, Dmitry Panchenko, and Mustazee Rahman (2019). “Suboptimality of local algorithms for a class of max-cut problems”. *The Annals of Probability* 47.3, pp. 1587–1618.

Ding, Yunzi, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira (2019). “Subexponential-time algorithms for sparse PCA”. *arXiv preprint arXiv:1907.11635*.

Durrett, Rick (2019). *Probability: theory and examples*. Vol. 49. Cambridge university press.

Jain, Vishesh, Frederic Koehler, Jingbo Liu, and Elchanan Mossel (2019). “Accuracy-memory tradeoffs and phase transitions in belief propagation”. *Conference on Learning Theory*. PMLR, pp. 1756–1771.

Kunisky, Dmitriy, Alexander S Wein, and Afonso S Bandeira (2019). “Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio”. *arXiv preprint arXiv:1907.11636*.

Bandeira, Afonso S, Dmitriy Kunisky, and Alexander S Wein (2020). “Computational hardness of certifying bounds on constrained PCA problems”. *ITCS*.

Gamarnik, David, Aukosh Jagannath, and Alexander S Wein (2020). “Low-degree hardness of random optimization problems”. *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, pp. 131–140.

Holmgren, Justin and Alexander S Wein (2020). “Counterexamples to the low-degree conjecture”. *ITCS*.

Kamath, Pritish, Omar Montasser, and Nathan Srebro (2020). “Approximate is good enough: Probabilistic variants of dimensional and margin complexity”. *Conference on Learning Theory*. PMLR, pp. 2236–2262.

Moitra, Ankur, Elchanan Mossel, and Colin Sandon (2020). “Parallels Between Phase Transitions and Circuit Complexity?” *Conference on Learning Theory*. PMLR, pp. 2910–2946.

Schramm, Tselil and Alexander S Wein (2020). “Computational barriers to estimation from low-degree polynomials”. *arXiv preprint arXiv:2008.02269*.

Wein, Alexander S (2020). “Optimal low-degree hardness of maximum independent set”. *arXiv preprint arXiv:2010.06563*.

Brennan, Matthew, Guy Bresler, Samuel Hopkins, Jerry Li, and Tselil Schramm (2021). “Statistical query algorithms and low-degree tests are almost equivalent”. *COLT*.

Bresler, Guy and Brice Huang (2021). “The Algorithmic Phase Transition of Random k-SAT for Low Degree Polynomials”. *FOCS*.

Ghorbani, Behrooz, Song Mei, Theodor Misiakiewicz, and Andrea Montanari (2021). “Linearized two-layers neural networks in high dimension”. *The Annals of Statistics* 49.2, pp. 1029–1054.

Mao, Cheng and Alexander S Wein (2021). “Optimal Spectral Recovery of a Planted Vector in a Subspace”. *arXiv preprint arXiv:2105.15081*.

Mei, Song, Theodor Misiakiewicz, and Andrea Montanari (2021). “Generalization error of random features and kernel methods: hypercontractivity and kernel matrix concentration”. *arXiv preprint arXiv:2101.10588*.

Mohanty, Sidhanth, Siqi Liu, and Prasad Raghavendra (2021). “On statistical inference when fixed points of belief propagation are unstable”. *IEEE 62st Annual Symposium on Foundations of Computer Science (FOCS)*.

Muthukumar, Vidya, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai (2021). “Classification vs regression in overparameterized regimes: Does the loss function matter?” *Journal of Machine Learning Research* 22.222, pp. 1–69.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Error bars were computed but ended up being visually negligible because of large sample size

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Only used trivial resources.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Organization of the Appendix

In the Appendices we give full proofs of all results; there is no dependence on the Technical Overview as all the information there will be repeated here in more detail. In Appendix B we state more formally the open problem described in the intro. In Appendix C we prove the results for known trees: in particular, this includes the main lower bound result, which is the failure of low degree polynomials for recovering the root; we also show how to deduce the RBF kernel lower bound using this. This is also where the RecMaj algorithm from Figure 1 is formally explained. In Appendix D we prove the results in the setting with an unknown tree; the main technical step is showing how to reconstruct the tree when $\lambda_2 = 0$ using a sample-efficient algorithm, which can be straightforwardly implemented in SQ.

B Open Problem: General Broadcast Chains

Question 12 (Kesten-Stigum is sharp for Low-Degree Polynomials?). Suppose that d and transition matrix M are such that $d|\lambda_2(M)|^2 < 1$, i.e. we are below the Kesten-Stigum threshold. Let D_N be an arbitrary function of $N = d^\ell$ such that $D_N = O(\log N)$ as $\ell \to \infty$. Is it true that the degree-D_N maximum correlation between the broadcast process at the leaves X_L and the root X_ρ in the sense of Definition 4 is asymptotically zero, i.e. $\liminf_{\ell \to \infty} \mathrm{Corr}_{\le D_N} = 0$? Equivalently, is it true that

$$\liminf_{\ell \to \infty} \max_{c \in [q]} \sup_{deg(f) \le D_N, \mathbb{E}[f(X_L)^2]=1} \mathbb{E}_{\mu_\ell}[f(X_L)(1(X_\rho = c) - \nu(c))] = 0?$$

Here we make the common choice of looking at $\log N$ degree polynomials (see e.g. Hopkins and Steurer 2017; Kunisky et al. 2019), but any degree is interesting.

With the same intuition, we ask if a similar result to Theorem 8, the lower bound for kernel ridge regression, holds below the Kestum-Stigum threshold — see Figure 1 for related simulation results, which support the failure of KRR for small values of λ_2. We note that in our experiment the threshold where KRR starts to work is much closer to $d\lambda_2 = 1$. It is quite possible that this is a finite-depth effect since the experiment was done with a relatively shallow tree. Of course, if the sharp threshold is not the Kestum-Stigum threshold it would be extremely interesting to understand what the correct threshold is as a function of the broadcast model parameters.

C Known Tree: Upper and Lower Bounds

In this section, we prove a lower bound for arbitrary markov chains M satisfying $\lambda_2(M) = 0$. From basic linear algebra (the existence of the Jordan Normal Form (Artin 2011)), we know that $\lambda_2(M) = 0$ if and only if M^k is a rank one matrix for some $1 \le k \le q$, i.e. the Markov chain mixes perfectly in a finite number of steps. For concreteness, we give an example of such a chain with $k = 2, q = 3$ below.

Example 13 (Proof of Proposition 5, Mossel 2001). The following Markov chain on $q = 3$ states is a simple example of a chain with $\lambda_2(M) = 0$: we have

$$M = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 0.25 & 0.5 & 0.25 \\ 0 & 1 & 0 \end{bmatrix}, \qquad M^2 = \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.5 & 0.25 \end{bmatrix}.$$

C.1 Failure of Low-Degree Polynomials

Theorem 14. *Let M be the transition matrix of a Markov chain on $[q]$ and suppose that $1 \le k \le q$ is such that M^k is a rank-one matrix. Let S be any subset of the leaves of the depth-ℓ complete d-ary tree $T = (V, E, \rho)$ with root ρ and let $(X_v)_{v \in V}$ denote the broadcast process on T with channel M. Let S be an arbitrary subset of the leaf nodes of this tree. If $|S| < 2^{\lfloor \ell/(k-1) \rfloor}$, then $I(X_\rho; X_S) = 0$, i.e. X_S is independent of the root value X_ρ.*

Proof. Assume for contradiction that $|S| < 2^{\lfloor \ell/(k-1) \rfloor}$ and $I(X_\rho; X_S) > 0$. Let T_S be the minimal spanning subtree of T containing the root node ρ and all of the elements of S. (Equivalently, T_S is the union of all of the root-to-leaf paths to S.)

Recall that in our convention, the edges of the tree T are directed from the parent to the child. We say that T_S contains an *isolated* length k directed path if there exists adjacent nodes $u_0, \dots, u_k$ contained in T with $(u_i, u_{i+1}) \in E$ for all $0 \le i < k$, and such that nodes $u_1, \dots, u_{k-1}$ all have degree 2 in T_S.

We show that we can reduce to the case where T_S contains no isolated length k directed paths. Otherwise, let $u_0, \dots, u_k$ be as defined above and let S_{u_k} be the subset of S consisting of descendants of u_k (note that by the definition of T_S, S_{u_k} is nonempty). Observe that

$$I(X_\rho; X_S) \le I(X_\rho; X_S, X_{u_k}) = I(X_\rho; X_{S \setminus S_{u_k}}, X_{u_k})$$

where the last equality follows by the Markov property (all nodes in S_{u_k} are descendants of u_k, so $X_{S_{u_k}}$ is independent of the root value X_ρ conditionally on $X_{S \setminus S_u}, X_{u_k}$).

Next, by the chain rule for mutual information

$$I(X_\rho; X_{S \setminus S_{u_k}}, X_{u_k}) = I(X_\rho; X_{S \setminus S_{u_k}}) + I(X_\rho; X_{u_k} \mid X_{S \setminus S_{u_k}}) = I((X_\rho; X_{S \setminus S_{u_k}})$$

where the last equality follows from the fact that

$$I(X_\rho; X_{u_k} \mid X_{S \setminus S_{u_k}}) \le I(X_{u_0}, X_\rho; X_{u_k} \mid X_{S \setminus S_{u_k}}) = I(X_{u_0}; X_{u_k} \mid X_{S \setminus S_{u_k}}) = 0$$

where in turn the first equality follows from the Markov property (X_ρ is independent of X_{u_k} conditional on X_{u_0} and $X_{S \setminus S_{u_k}}$) and the second equality follows because by the Markov property,

$$X_{S \setminus S_{u_k}} \to X_{u_0} \to X_{u_k}$$

is a Markov chain where the rightmost channel has transition matrix M^k, a rank-one matrix, so the conditional law of X_{u_k} is the stationary measure of M regardless of the value of X_{u_0}, hence X_{u_k} is conditionally independent of X_{u_0}. Combining the above claims shows that

$$I(X_\rho; X_S) \le I(X_\rho; X_{S \setminus S_{u_k}})$$

where $|S \setminus S_{u_k}| < |S|$; by monotonicity of mutual information we in fact have

$$I(X_\rho; X_S) = I(X_\rho; X_{S \setminus S_{u_k}}).$$

Repeating this argument recursively reduces to the case where T_S has no isolated length k paths.

Finally, if T_S has no isolated length k paths then every internal node of T_S is either: (a) at depth at most $k-1$, or (b) has an ancestor at graph distance at most $k-1$ away with degree at least 3. By induction, this implies that the number of nodes at depth ℓ' in T_S is at least twice as large as the number of nodes at depth $\ell' - (k-1)$. Since S is the set of nodes in T_S at depth ℓ, this implies that

$$|S| \ge 2^{\lfloor \ell/(k-1) \rfloor}$$

which completes our proof by contradiction. □

Corollary 15. *In the setting of the previous Theorem, for any function $f : [q]^L \to \mathbb{R}$ of Efron-Stein degree at most $2^{\lfloor \ell/(k-1) \rfloor}$ of the leaves X_L and any prior ν on the root,*

$$\mathbb{E}[f(X_L) \cdot (\mathbb{1}(X_\rho = c) - \nu(c))] = 0.$$

Proof. By linearity of expectation and the Efron-Stein decomposition,

$$\mathbb{E}[f(X_L) X_\rho] = \sum_{S \subset L, |S| \le D} \mathbb{E}[f_S(X_L) \cdot (\mathbb{1}(X_\rho = c) - \nu(c))]\rangle] = 0$$

where the last equality used the previous Theorem and the fact $\mathbb{E}[\mathbb{1}(X_\rho = c)] = \nu(c)$. □

C.1.1 A consequence: failure of RBF kernel regression with oracle tuning

Setting and notation. We consider the performance of RBF kernel ridge regression (with arbitrary/oracle hyperparameter selection) for predicting the color of the root given the color of the leaves. As is customary, we encode the leaf vectors using a one-hot encoding, so the input to the regression is a list of i.i.d. samples $(x_i, y_i)_{i=1}^m$ where x_i is the vector of one-hot encoded leaves, i.e. $(x_i)_{\ell,c} = \mathbb{1}(X_\ell = c)$, and for an arbitrary fixed color c, $y_i := \mathbb{1}(X_\rho = c) - \nu(c)$ is the centered indicator that the root is colored c.

Background on Kernel Ridge Regression. We remind the reader of some standard facts about kernel ridge regression and the Gaussian/RBF kernel — see (Shalev-Shwartz and Ben-David 2014) for a reference. Given a kernel $K(x, x')$, training points $x_1, \dots, x_m$, and responses $y = (y_1, \dots, y_m)$, the kernel ridge regressor with ridge parameter λ is given by solving a linear equation

$$v = (\mathsf{K} + \lambda I)^{-1} y$$

where $\mathsf{K}_{ij} = K(x_i, x_j)$ is the kernel matrix, and the predicted response for a fresh data point x_0 is given by

$$\hat{y}_0 := \sum_{i=1}^{n} v_i K(x_i, x_0).$$

As is well-known, kernel ridge regression with ridge parameter λ is equivalent to solving the ridge regression problem

$$\arg\min_{w} \sum_{i=1}^{n} (y_i - \langle w, \varphi(x_i) \rangle)^2 + \lambda \|w\|_2^2 \tag{2}$$

with feature vectors $\varphi(x)$ lying in a certain Hilbert space. Note that with this parameterization the prediction for fresh data point x_0 would just be $\hat{y}_0 = \langle w, \varphi(x_0) \rangle$ since w is in the Hilbert space. In the case of the RBF kernel $K(x, y) = e^{-\|x-y\|_2^2/2\sigma^2}$, the corresponding feature map for $x \in \mathbb{R}^d$ is

$$\varphi(x) = e^{-\|x\|^2/2\sigma^2} \left(\frac{1}{\sigma^{2(n_1 + \dots + n_d)}} \frac{x_1^{n_1} \cdots x_d^{n_d}}{\sqrt{n_1! \cdots n_d!}} \right)_{n_1, \dots, n_d \geq 0} \tag{3}$$

so that $K(x, y) = \langle \varphi(x), \varphi(y) \rangle$. Note that $\|\varphi(x)\| = 1$ since $K(x, x) = e^0 = 1$.

Proof of the lower bound. We now proceed to prove the subexponential RBF sample complexity lower bound in our setting. For ψ an element of the RKHS, define the orthogonal projection operator onto the space of degree J and higher polynomials $P_{\geq J}$ by

$$(P_{\geq J}\psi)_{n_1, \dots, n_d} := \begin{cases} 0 & \text{if } n_1 + \dots + n_d < J \\ \psi_{n_1, \dots, n_d} & \text{otherwise} \end{cases}.$$

From the definition, we first show that for large degree J and bandwidth σ not too tiny, $P_{\geq J}$ is very contractive when operating on feature embeddings $\varphi(x)$.

Lemma 16. *For any $x \in \mathbb{R}^d$ and $\varphi(x)$ as defined in (3) with bandwidth parameter $\sigma > 0$,*

$$\|P_{\geq J}\varphi(x)\|^2 \leq \frac{1}{\sqrt{J}} \left(\frac{e\|x\|^2}{J\sigma^2} \right)^J$$

Proof. First observe that

$$\frac{(\|x\|^2/\sigma^2)^j}{j!} = \sum_{n_1 + \dots + n_d = j} \frac{1}{n_1! \cdots n_d! (\sigma^2)^{2j}} x_1^{2n_1} \cdots x_d^{2n_d}$$

by applying the multinomial theorem. Therefore,

$$\|P_{\geq J}\varphi(x)\|^2 = e^{-\|x\|^2/\sigma^2} \sum_{j=J}^{\infty} \frac{\|x\|^{2j}}{\sigma^{2j} j!} = e^{-\|x\|^2/\sigma^2} \frac{\|x\|^{2J}}{\sigma^{2J} J!} \sum_{j=0}^{\infty} \frac{\|x\|^{2j} J!}{\sigma^{2j} (J+j)!} \leq \frac{\|x\|^{2J}}{\sigma^{2J} J!}$$

and then the stated result follows from a nonasymptotic version of Stirling's approximation. □

Next, we prove that there exists a relatively low-degree and low-RKHS norm polynomial which perfectly interpolates the training data, by showing that with high probability every sample has a small and unique "fingerprint" given by looking at a small set of well-separated leaves.

Lemma 17. *Let M be the transition matrix of a markov chain on $[q]$ and suppose that $1 \leq k \leq q$ is such that $M^k = \pi\pi^T$ is a rank-one matrix, and suppose that π has at least two nonzero entries. Then if S is a set of leaves of distance at least $2k$ from each other and $X_1, \dots, X_m$ are i.i.d. random vectors generated by the broadcast process with transition matrix M, the probability that there exists $i, j \in [m]$ such that $(X_i)_S = (X_j)_S$ is at most $\binom{m}{2}\delta^{|S|}$ where $\delta = \delta(M) \in (0, 1)$ is a constant depending only on M.*

Proof. First, let X, X' be independent samples of the leaves from the generative model and let $c = c(M) > 0$ be such that the stationary distribution π has at least two entries of size at least c. For S a set of leaves of distance at least $2k$ from each other, we have by the Markov property that the entries of X_S are independent from each other conditional on the values of the markov process X_v for all vertices v at height k above the leaves; we see then that the conditional law of the leaves X_S is $\pi^{\otimes S}$ which does not depend on X_v, so in fact the leaves X_S are unconditionally distributed according to the product measure $\pi^{\otimes S}$. Then by independence,

$$\Pr(X_S = X'_S) = \prod_{i \in S} \Pr(X_i = X'_i) \le (1-c)^{|S|}$$

where in the last step we used that regardless of the value of X_i, X'_i has a probability at least c of being different from it. □

Lemma 18. *For $x \in \{0,1\}^d$ with $\sum_i x_i = p$, and $S \subseteq [d]$ and $b_S \in \{0,1\}^d$ arbitrary, there exists $w = w(p, S)$ of (Hilbert space) norm*

$$\|w\|^2 \le 2^{|S|} e^{p/2\sigma^2} \max\left\{1, \sigma^{2|S|}\right\} \sqrt{|S|!}$$

such that

$$\langle w, \varphi(x) \rangle = 1(x_S = b_S).$$

Proof. Observe that

$$1(x_S = b_S) = \prod_{i \in S} [b_i x_i + (1-b_i)(1-x_i)]$$

which for fixed b, expands into a sum of at most $2^{|S|}$ many monomials of degree at most $|S|$ and with coefficient 1. Representing this expanded polynomial in the RKHS, using (3), then leads to the stated norm bound. □

We show that the overlap between two independent samples of the leaves from the model concentrates exponentially with a subgaussian tail:

Lemma 19. *Let M be the transition matrix of a markov chain on $[q]$ and suppose that $1 \le k \le q$ is such that $M^k = \pi\pi^T$ is a rank-one matrix. Then if X_L, X'_L are two independent random vectors of leaf colorations generated by the broadcast process on the d-ary tree with $N = |L|$ leaves and x_L, x'_L are the corresponding one-hot encodings, we have that*

$$\Pr\left(\left|\frac{1}{N}\langle x_L, x'_L\rangle - \|\pi\|_2^2\right| > t\right) \le 2e^{-cNt^2}$$

where $c = c(M, d) > 0$ is a constant not depending on N.

Proof. First, observe that if N is smaller than d^k, this bound can be proved trivially by shrinking c, so henceforth we assume N is larger than this. By the law of total probability, it is sufficient to prove the desired bound conditional on the colors X_V, X'_V where V is the set of vertices at height k above the leaves, and similar to the proof of Lemma 17 we observe by the Markov property that this makes the color of the set of children of any particular $v \in V$ independent of the colors of all non-children of v. This means that $\langle x_L, x'_L\rangle$ a sum of bounded independent random variables, and because $M^k = \pi\pi^T$ we have that its expectation is $\|\pi\|_2^2 N$, so the result follows immediately from Hoeffding's inequality (Vershynin 2018). □

Theorem 20. *Let M be the transition matrix of a markov chain on $[q]$ and suppose that $1 \le k \le q$ is such that $M^k = \pi\pi^T$ is a rank-one matrix, and suppose that π has at least two nonzero entries. Suppose that $m/\delta \le e^{cN^\epsilon}$. Then given m i.i.d. samples $(x_1, y_1), \dots, (x_m, y_m)$ from the broadcast model on the d-ary tree with N leaves and broadcast channel M, we have that for any bandwidth $\sigma \ge 0$ and ridge parameter $\lambda \ge 0$, for w the output of ridge regression in RKHS space with those parameters, that with probability at least $1 - \delta$*

$$\frac{\mathbb{E}_{x_0,y_0}[y_0\langle w, \varphi(x_0)\rangle]}{\sqrt{\mathbb{E}_{x_0,y_0}[y_0^2]}} = O(\sqrt{1/N})$$

provided that $m/\delta = O(e^{N^\epsilon})$ where $\epsilon = \epsilon(M, d) > 0$ is independent of N (equivalently, independent of the depth of the tree).

Proof. As usual, we will use that N can be assumed larger than a fixed absolute constant without loss of generality. The proof is via case analysis on the bandwidth parameter σ.

First we make an argument which covers the case of small bandwidth parameter σ. Note that for any i, $\|x_i\|^2 = N$ almost surely since there are N leaves and each leaf is one-hot encoded. By Lemma 19 and the union bound, with probability at least $1 - \delta/4$ for any $i \neq j$ in $[m]$ we have

$$\|x_i - x_j\|_2^2 = 2N - 2\langle x_i, x_j \rangle \geq 2(1 - \|\pi\|_2^2)N - O_{M,d}(\sqrt{N \log(m/\delta)})$$

so

$$\mathsf{K}_{ij} = e^{-\|x_i - x_j\|_2^2/2\sigma^2} \leq \exp\left([-(1 - \|\pi\|_2^2)N + O_{M,d}(\sqrt{N \log(m/\delta)})]/\sigma^2\right).$$

It follows that there exists $c_2 = c_2(M, d) > 0$ such that if $\sigma \leq c_2 N^{1/2 - \epsilon/4}$, then $(\mathsf{K})_{ij} \leq e^{-N^{\epsilon/3}}$ for $i \neq j$ and so by Gershgorin's disk theorem and the fact that the diagonal of K is all-ones, $\|\mathsf{K} - I\|_{OP} \leq 1/N$. Hence for the Kernel Ridge solution $v = (\mathsf{K} + \lambda I)^{-1} y$ we have $\|v\| \leq 2\|y\| \leq 2\sqrt{m}$.

Consider a fresh test set of independently sampled pairs of leaf and root colorations $(x'_1, y'_1), \ldots, (x'_{ms}, y'_{ms})$ where $s := N \log(2/\delta)$. Observe by Hoeffding's inequality that with probability at least $1 - \delta/4$,

$$\left| \frac{1}{ms} \sum_{i=1}^{ms} \left(\sum_{j=1}^{m} v_j K(x_j, x_0) \right)^2 - \mathbb{E}_{x_0} \left[\left(\sum_{j=1}^{m} v_j K(x_j, x_0) \right)^2 \right] \right| = O(\sqrt{\log(2/\delta)/s})$$

where x_0 is a fresh one-hot encoded vector of leaf colorations sampled from the same distribution and where we used the fact that $\|v\| \leq 2\sqrt{m}$ and $K(\cdot, \cdot) \leq 1$ to show that over the randomness of x_0, $\left|\sum_{j=1}^{m} v_j K(x_j, x_0)\right| \leq 2\sqrt{m}$ almost surely, which we used in order to apply Hoeffding's inequality. By repeating the argument used to show the off-diagonal entries of K are small, we have with probability at least $1 - \delta/4$

$$\frac{1}{ms} \sum_{i=1}^{ms} \left(\sum_{j=1}^{m} v_i K(x_j, x_i) \right)^2 \leq m e^{-N^{\epsilon/2}},$$

hence by the triangle inequality we have with probability at least $1 - \delta$ that

$$\mathbb{E}_{x_0} \left[\left(\sum_{i=1}^{m} v_i K(x_i, x_0) \right)^2 \right] \leq m e^{-N^{\epsilon/2}} + O(\sqrt{\log(2/\delta)/s})$$

and recalling $s = N \log(2/\delta)$ gives the result in this case.

Now we cover the remaining set of bandwidth parameters where $\sigma > c_2 N^{1/2 - \epsilon/4}$. By the combination of Lemma 17 applied with $|S| = C_M \log(m/\delta)$ and Lemma 18, we have that there exists w such that for every x_i

$$\langle w, \varphi(x_i) \rangle = y_i$$

and

$$\|w\| \leq (m/\delta)^{C'_M} e^{N/4\sigma^2} \sigma^{C_M \log m/\delta} \sqrt{(C_M \log m/\delta)!}. \tag{4}$$

It follows that the output of KRR with any ridge parameter $\lambda \geq 0$ has norm at most the rhs of (4) (otherwise, replacing the output with w would shrink the norm without decreasing the training error in (2)). Next, by Lemma 16 we have that for any x and degree J

$$\begin{aligned} \langle P_{\geq J} w, \varphi(x) \rangle &= \langle w, P_{\geq J} \varphi(x) \rangle \\ &\leq \|w\| \|P_{\geq J} \varphi(x)\| \\ &\leq \|w\| \frac{1}{\sqrt{J}} \left(\frac{e\|x\|^2}{J\sigma^2} \right)^J = \|w\| \frac{1}{\sqrt{J}} \left(\frac{eN}{J\sigma^2} \right)^J \end{aligned}$$

so taking as in Corollary 15 $J = 2^{\lfloor \ell/(k-1) \rfloor} = N^{\epsilon}$ where this equation defines ϵ and using that

$$N/J\sigma^2 = N^{1-\epsilon}/\sigma^2 = O(\sigma^{-\epsilon/(1-\epsilon/2)}),$$

we have that for any w satisfying (4),

$$
\begin{aligned}
|\langle P_{\geq J} w, \varphi(x) \rangle| &\leq \|w\| (c_3/\sigma^{\epsilon/(1-\epsilon/2)})^{N^\epsilon} \\
&\leq (m/\delta)^{C'_M} e^{N^{\epsilon/2}/4c_2} \sigma^{C_M \log m/\delta} \sqrt{(C_M \log m/\delta)!} (c_3/\sigma^{\epsilon/(1-\epsilon/2)})^{N^\epsilon} = O((1/\sigma)^{N^\epsilon/2}).
\end{aligned}
$$

Since by Corollary 15 and Cauchy-Schwarz we have that

$$
\mathbb{E}_{x_0,y_0}[y_0 \langle w, \varphi(x_0) \rangle] = \mathbb{E}_{x_0,y_0}[y_0 \langle P_{\geq J} w, \varphi(x_0) \rangle] \leq \sqrt{\mathbb{E}_{x_0,y_0}[y_0^2]} \sqrt{\mathbb{E}_{x_0,y_0}[\langle P_{\geq J} w, \varphi(x_0) \rangle^2]}
$$

combining this with the bound on $|\langle P_{\geq J} w, \varphi(x) \rangle|$ completes the proof. □

C.2 Success of noise-robust reconstruction using non-low-degree algorithms

Above we saw that when $|\lambda_2(M)| = 0$, very high degree polynomials are needed to get any estimate correlated with the root. Nevertheless, for "most" matrices M with $|\lambda_2(M)| = 0$ and for degree d sufficiently large as a function M there exists a simple recursive and noise-robust method which witnesses the fact that reconstructing the root is possible. If one likes, this recursive function can trivially be expressed as a polynomial: then it will be a very high-degree polynomial that is nonetheless robust to noise.

The reason for the qualifier "most" in the discussion above is that there are some degenerate M for which the task is clearly impossible: e.g. if M is rank one (so it does not depend on its input). There are other similar examples, e.g. the chain on 3 states which deterministically transitions from state 1 to state 2, and such that at states 2 and 3 the chain flips a fair coin to transition to either state 2 or 3. With this clarified, we can now state the known positive result for reconstruction.

Theorem 21 (Theorem 6.1 of Mossel 2004)**.** *Suppose M is a the transition matrix of a Markov chain with pairwise distinct rows, i.e. for all $i, j \in [q]$ the rows M_i and M_j are distinct vectors. Then there exists $d_0 = d_0(M)$ such that for all $d \geq d_0$, reconstruction is possible on the d-ary tree.*

A variant of the condition in this Theorem gives a tight characterization of Markov chains where reconstruction is possible on the infinite d-ary tree for sufficiently large d, see Theorem 2.1 of Mossel and Peres 2003.

By revisiting the proof of Theorem, we get the following slightly more precise result which we will use in later sections. This result shows that for any desired accuracy δ, for sufficiently large degrees d there exists a noise-tolerant estimator f which reconstructs the root correctly with probability at least $1 - \delta$ uniformly of the color of the root.

Theorem 22 (Proof of Theorem 2.1 of Mossel and Peres 2003)**.** *Suppose M is a the transition matrix of a Markov chain with pairwise distinct rows, i.e. for all $i, j \in [q]$ the rows M_i and M_j are distinct vectors. Let $\delta \in (0, 1)$ be arbitrary. There exists $d_0 = d_0(M, \delta), \epsilon > 0$ such that for all $d \geq d_0$, ϵ-noisy reconstruction is possible on the d-ary tree and furthermore there exists a polynomial-time computable function $f = f_{M,\ell}$ valued in $[q]$ such that*

$$
\max_{c \in [q]} \Pr(f(X'_L) \neq X_\rho \mid X_\rho = c) < \delta
$$

where X'_L is the ϵ-noisy version of X_L (see Definition 3).

Proof sketch. As explained above, this result follows from examination of the proof of Theorem 2.1 in Mossel and Peres 2003. For the reader's convenience, we summarize the main idea of the proof.

In the base case of a depth 1 tree, reconstruction of the root with probability at least $1 - \delta$ is possible provided d is a suitably large constant, because by basic large deviations theory (Sanov's Theorem (Dembo and Zeitouni 2010)) the empirical distribution of the children will concentrate around the row of M corresponding to the root label (which by assumption is distinct from all of the other rows). This procedure is also robust to a small amount of noise, which handles the case where $\epsilon > 0$ and in fact even if the ϵ proportion of children assigned labels by the noise process choose their labels adversarially. When doing the induction, the result of the reconstruction process at lower levels of the tree can therefore (by conditional independence) be modeled as the true values with a small amount of adversarial noise and this allows the same argument to show that at each level each vertex is recovered correctly with probability at least $1 - \epsilon$ (where we take $\epsilon := \delta$). □

Remark 23 (RecMaj in Figure 1). The RecMaj algorithm in Figure 1 corresponds to the algorithm described in the above proof sketch: i.e. a recursive algorithm which to reconstruct the coloration of a vertex, looks at the reconstructions of its children, takes the empirical distribution, and picks the corresponding row of M which is closest in ℓ_2 norm.

C.3 Low-Degree Polynomials succeed above the KS threshold

The Kesten-Stigum threshold is the sharp threshold for *count reconstruction* defined earlier. The definition of count reconstruction informally says that there is a nontrivial amount of mutual information between count statistics at the leaves and the value of the Markov Random Field at the root. To relate count reconstruction to low-degree polynomials, we use the following more precise result:

Lemma 24 (Proof of Theorem 1.4 of Mossel and Peres 2003)**.** *Suppose that $d|\lambda_2(M)|^2 > 1$. There exist coefficients $s_c \in \mathbb{C}$ for $c \in [q]$ such that the random variable*

$$S = \sum_{c \in [q]} s_c \#\{X_\ell = c : \ell \in L\}$$

satisfies

$$\mathbb{E}[S \mid X_\rho = c] = v_c$$

where v is an unit-norm eigenvector of M in its second-largest eigenspace, i.e. achieving $\|Mv\| = |\lambda_2(M)|$, and such that

$$\mathbb{E}[|S|^2 \mid X_\rho = c] \in [A, B]$$

where $0 < A \leq B$ are constants depending only on d and M (in particular, they are independent of the depth of the tree).

As a consequence of this, we immediately obtain that low-degree polynomials (in fact, degree 1 polynomials) have nontrivial correlation with the root above the KS threshold, in the same sense as Definition 4.

C.3.1 A Question: Bayes-Optimal Reconstruction

We saw above that degree-1 polynomials of the leaves are sufficient to achieve nontrivial correlation with the root, provided that the model we consider is above the KS threshold. A natural question is whether higher degree polynomials have a significant advantage over degree-1 polynomials for estimating the value of the root. Relevant to this question, we recall the following result and conjecture from Mossel et al. 2014 which concerns noise-robust recovery with the Binary Symmetric Channel (equivalently, the Ising model on trees without external field):

Theorem 25 (Theorem 3.2 of Mossel et al. 2014)**.** *There exists an absolute constant $C \geq 1$ such that the following result is true. For $\theta \geq 0$ let*

$$M = \begin{bmatrix} (1+\theta)/2 & (1-\theta)/2 \\ (1-\theta)/2 & (1+\theta)/2 \end{bmatrix}$$

and observe that $\lambda_2(M) = \theta$. If $d\theta^2 > C$, then for all $\epsilon < 1$ and X'_L defined by the ϵ-noisy broadcast model,

$$\begin{aligned} &\lim_{\ell \to \infty} d_{TV}(\mathcal{L}_{\mu_\ell}(X'_L = \cdot \mid X_\rho = 1), \mathcal{L}_{\mu_\ell}(X'_L = \cdot \mid X_\rho = 0)) \\ &= \lim_{\ell \to \infty} d_{TV}(\mathcal{L}_{\mu_\ell}(X_L = \cdot \mid X_\rho = 1), \mathcal{L}_{\mu_\ell}(X_L = \cdot \mid X_\rho = 0)) \end{aligned}$$

in other words, if $\epsilon < 1$ is fixed then in the limit of infinite depth the probability of reconstructing the root correctly is the same as in the noiseless case $\epsilon = 0$.

(Recall that the equivalence of the statement in terms of TV and in terms of maximum probability of reconstructing the root follows from the Neyman-Pearson Lemma Neyman and Pearson 1933.) This statement is conjectured to hold with $C = 1$ Mossel et al. 2014 and as explained there, is closely related to Bayes-optimal recovery in the stochastic block model. Based on this, we ask the following question:

Question 26. Do polynomials of degree $O(\log N)$ achieve asymptotically Bayes-optimal recovery with the above channel when $d\theta^2 > 1$? More precisely, does there exist a polynomial threshold function f of degree $O(\log N)$ which asymptotically achieves

$$\Pr(f(X_L) = X_\rho) = (1 + o(1)) \Pr(\operatorname{sgn}(\mathbb{E}[X_\rho \mid X_L] - 1/2) = X_\rho)$$

where the rhs is the error of the Bayes-optimal estimator.

It seems likely the answer to this question is positive. The reason for this is the following: (1) if the conjectured strengthening of Theorem 25 is true, then it implies that the combination of a majority vote up to some depth and $\omega(1)$ number of rounds of belief propagation achieves Bayes-optimal recovery, and (2) a constant or very slowly growing number of rounds of belief propagation can be simulated with low-degree polynomials (see Appendix of Gamarnik et al. 2020), and the threshold used in the majority vote should also be approximable by polynomials. We state the conjecture with $O(\log N)$ degree polynomials since this is informally considered to correspond to "polynomial time algorithms" in the low-degree framework (Hopkins 2018; Kunisky et al. 2019), but based on the above discussion it seems likely that a smaller degree than $O(\log N)$ is sufficient, e.g. any degree going to infinity with N may be sufficient.

D Unknown Tree Setting

In this section, we show that for any channel M satisfying the conditions of Theorem 22, i.e. such that for sufficiently large d reconstructing the root is possible (in the known tree setting/in the usual sense), then in the unknown tree setting that a relatively simple algorithm succeeds at reconstructing the root with a polynomial number of samples, and this algorithm can be straightforwardly implemented in the SQ (Statistical Query) model with polynomial number of queries and error tolerance.

The key step in the algorithm for reconstructing the root is a method of reconstructing the tree, which lets us reduce to the known tree setting. This kind of problem has previously been extensively studied in the context of phylogenetic reconstruction with particular channels M coming from biology, and for example algorithms with polynomial runtime and sample complexity are known in the case that M is a nonsingular matrix (Mossel and Roch 2005). In the present context, we are very interested in the case of singular matrices (e.g. those with $\lambda_2(M) = 0$) so we cannot rely on existing results.

Model. We remind the reader that in the unknown tree setting, we are in the model of Definition 9. This means that an unknown Y^* is sampled from $Uni([q])$, and the algorithm seeks to reconstruct Y^* given access to m i.i.d. samples $X_L^{(1)}, \ldots, X_L^{(m)}$ of the leaves generated by the broadcasting process with root prior $(2/3)\delta_{Y^*} + (1/3)Uni([q])$, i.e. the root is biased/tilted towards the unknown Y^*. When we say the tree is "unknown" in this model, it means that the algorithm is not given a priori knowledge of the true order of the leaves, e.g. the algorithm does not know at the beginning whether coordinates 1 and 2 of $X_L^{(1)}$ correspond to siblings or to leaves far apart in the tree (this is completely analogous to the situation in phylogenetic reconstruction Steel 2016). In the definition of this model, this is modeled by shuffling the order of the leaves by an unknown permutation τ; note that this order is kept consistent between each sample.

D.1 Failure of low-degree polynomials

Theorem 27. *Let M be the transition matrix of a Markov chain on $[q]$ and suppose that $1 \le k \le q$ is such that M^k is a rank-one matrix. If $c \in [q]$ is arbitrary and f is a polynomial with Efron-Stein degree strictly less than $2^{\lfloor \ell/(k-1) \rfloor}$, then*

$$\mathbb{E}_R[f(\mathbb{X})(\mathbb{1}(Y^* = c) - 1/q)] = 0$$

where R is as defined in Definition 9.

Proof. Let $\nu(c) = 1/q$ for $c \in [q]$ denote the prior on Y^*.

By linearity of expectation and the definition of Efron-Stein degree, it suffices to show the result for functions f of the form $f_{S_1}(X_L^{(1)}) \cdots f_{S_m}(X_L^{(m)})$ where $\sum_i |S_i| < 2^{\lfloor \ell/(k-1) \rfloor}$, where each

$f_{S_i}(X_L^{(i)})$ is a function only of the coordinates of its input in S_i. Since the samples $X^{(1)}, \dots, X^{(m)}$ are conditionally independent given the value of Y^*, we have

$$\begin{aligned}
&\mathbb{E}_R\left[\left(\prod_{i=1}^m f_{S_i}(X_L^{(i)})\right)(\mathbb{1}(Y^*=c)-\nu(c))\right] \\
&\qquad = \mathbb{E}_R\left[\mathbb{E}\left[\left(\prod_{i=1}^m f_{S_i}(X_L^{(i)})\right)(\mathbb{1}(Y^*=c)-\nu(c)) \mid Y^*\right]\right] \\
&\qquad = \mathbb{E}_R\left[\left(\prod_{i=1}^m \mathbb{E}[f_{S_i}(X_L^{(i)}) \mid Y^*]\right)(\mathbb{1}(Y^*=c)-\nu(c))\right] \\
&\qquad = \mathbb{E}_R\left[\left(\prod_{i=1}^m \mathbb{E}[f_{S_i}(X_L^{(i)})]\right)(\mathbb{1}(Y^*=c)-\nu(c))\right] = 0
\end{aligned}$$

where in the first equality we used the law of total expectation, in the second equality we used the aforementioned conditional independence, in the third equality we crucially used that by Theorem 14 the low-degree polynomial $f_{S_i}(X_L^{(i)})$ is independent of the root value and thus Y^*, and in the last step we used that $Y^* \sim \nu$ by definition. □

D.2 Reconstruction Algorithm

For $c \in [q]$, let $e(c)$ or e_c denote the qth standard basis vector in $\mathbb{R}^q$. In both cases, the vector is a column vector.

Lemma 28. *Suppose that ν is a probability measure on $[q]$ and $\nu(c) > 0$ for all $c \in [q]$, then there exists a constant $\alpha = \alpha(M, \nu) > 0$ such that the following is true. Let $(X_u)_u \sim \mu$ for $u \in V$ be defined by the broadcasting process on $T = (V, E, \rho)$ with prior ν at the root and channels corresponding to $M : q \times q$ the transition matrix of an ergodic Markov chain. Then $\mu(X_u = c) > \alpha$ for all $u \in V$.*

Proof. Under the assumptions, there exists some $\beta > 0$ such that $\nu = \beta\pi_M + (1-\beta)\nu'$ for ν' a probability measure. Because π_M is the stationary distribution and the marginal law at any vertex u is νM^k for some $k \geq 0$, it follows that $\mu(X_u = c) > \beta\pi_M(c) \geq \min_c \beta\pi_M(c) =: \alpha > 0$. □

Lemma 29. *Suppose that u, v are two descendants of node w at graph distance k from w and random variables X_u, X_v, X_w follow the Markov process on trees μ with transition matrix $M : q \times q$. Then*

$$\mathbb{E}[e(X_u)e(X_v)^T] = (M^k)^T \Pi_w M^k$$

where $\Pi_w : q \times q$ is a diagonal matrix with entries the marginal law of X_w, i.e. $(\Pi_w)_{cc} = \mu(X_w = c)$ for $c \in [q]$.

Proof. Using the law of total expectation and using by the Markov property that X_u and X_v are conditionally independent given X_w, we have

$$\mathbb{E}[e(X_u)e(X_v)^T] = \mathbb{E}[\mathbb{E}[e(X_u) \mid X_w]\mathbb{E}[e(X_v)^T \mid X_w]] = \mathbb{E}[(e_{X_w}^T M^k)^T(e_{X_w}^T M^k)] = (M^k)^T \Pi_w M^k$$

where in the last equality we used the definition of Π_w and the definition of the broadcast process in terms of the transition matrix M. □

Based on this, we can recursively reconstruct the tree when the degree is sufficiently large. We note that for other channels like the BSC channel, tree reconstruction methods often handle internal nodes u by computing majorities of the nodes under them, which gives an unbiased estimate of the spin X_u, but this technique is not applicable in our setting (it's unclear that unbiased estimators exist). Nevertheless, we show that applying the estimator from Theorem 22 can be used in a similar way, provided the degree d is sufficiently large.

Theorem 30. *Suppose M is a the transition matrix of a Markov chain with pairwise distinct rows, i.e. for all $i, j \in [q]$ the rows M_i and M_j are distinct vectors. If $|\lambda_2(M)| > 0$, additionally suppose that the prior on the root of the tree is the stationary distribution of M. There exists*

$d \geq 1$ and $\epsilon > 0$ so that the following result holds true for the complete d-ary tree with any depth $\ell \geq 1$. For any $\delta > 0$, there exist a polynomial time algorithm with sample complexity $m = poly_M(\log N, \log(1/\delta))$ from the ϵ-noisy repeated broadcast model (Definition 9) which with probability at least $1 - \delta$:

1. *outputs the true tree T (equivalently, the true permutation τ)*
2. *outputs $\hat{Y}$ such that $\hat{Y} = Y^*$.*

Also, this algorithm can be implemented in the Statistical Query (SQ) model using a $VSTAT(m)$ oracle with $m = poly_M(\log(N/\delta))$ and polynomial number of queries.

Proof. Given that the algorithm can correctly output the true tree T, the fact that it outputs the correct root label follows straightforwardly from Theorem 22 by using the algorithm specified in that result to estimate the root in each sample, and then taking the majority vote over those samples (which will succeed with high probability provided we take $\Omega(\log(2/\delta))$ samples due to Hoeffding's inequality), and this can approach can also clearly be implemented in the SQ model (the SQ query is the robust reconstruction function of the leaves which outputs a vector, so we take the expectation of this and look at the largest entry of this vector). In the remainder of the proof, we show how to correctly output the true tree T with high probability.

We first prove the result in the case that $\lambda_2(M) \neq 0$ and afterwards describe how to modify the argument straightforwardly when $\lambda_2(M) = 0$. Let φ be a right eigenvector such that $M\varphi = \lambda_2\varphi$. We start by describing the algorithm which computes the estimated tree $\hat{T}$ from the bottom up: let $\alpha = \alpha(M, \delta) > 0$ be a parameter to be set later. Let $\hat{\mathbb{E}}[\cdot]$ denote the expectation over the empirical distribution of m samples, so for any function f we have $\hat{E}[f(X)] = \frac{1}{m}\sum_i f(X^{(i)})$.

1. Base case: for all leaves $u \neq v$ define $g(u, v) := |\langle \varphi, \hat{E}[e(X_u)e(X_v)^T]\varphi\rangle|$. Let $g_{max} = \max_{u \neq v} g(u, v)$ and set u, v to be neighbors in $\hat{T}$ iff $g(u, v) \geq g_{max} - \alpha$. This constructs the first layer of the tree $\hat{T}$.
2. Recursive case: suppose that we have reconstructed the first $s \geq 1$ layers of the tree (from the bottom), and the current layer of the tree has more than one element. For each pair of internal nodes u, v at the current level of the tree, let S_u, S_v be the set of leaves under these nodes and let $g_s(u, v) := |\langle \varphi, \hat{E}[e(f_{M,\ell-s}(X_{S_u}))e(f_{M,\ell-s}(X_{S_v}))^T]\varphi\rangle|$ where $f_{M,\ell-s}$ is as defined in Theorem 22. Let $g_{max} = \max_{u \neq v} g(u, v)$ and set u, v to be neighbors in $\hat{T}$ iff $g(u, v) \geq g_{max} - \alpha$. This constructs the next layer of the tree $\hat{T}$.

We now need to show that with total probability at least $1 - \delta$, $\hat{T} = T$. First we consider the behavior of the base case; for simplicity, we first describe the argument when $\epsilon = 0$. Observe that if u and v are siblings in T at depth ℓ then by Lemma 29

$$\langle \varphi, \mathbb{E}[e(X_u)e(X_v)^T]\varphi\rangle = |\lambda_2|^2 \langle \varphi, \Pi_{\ell-1}\varphi\rangle$$

where $\Pi_{\ell-1}$ is a diagonal matrix encoding the marginal law of X_w for any w at depth $\ell - 1$, and similarly, if u and v are not siblings then they are at graph distance at least 4 in T so

$$\langle \varphi, \mathbb{E}[e(X_u)e(X_v)^T]\varphi\rangle \leq |\lambda_2|^4 \langle \varphi, \Pi_{\ell-1}\varphi\rangle$$

which is smaller by a factor of $|\lambda_2|^2$. (Note, here we are using the fact that in the case $|\lambda_2| > 0$, we additionally assumed the prior at the root is stationary and so the marginal law at every depth in the tree is the stationary distribution.) Observe that by Hoeffding's inequality and the union bound we have that with probability at least $1 - \delta/n$ that in the base case step, every entry of the matrix $\hat{\mathbb{E}}[e(X_u)e(X_v)^T]$ for every pair of leaves $u \neq v$ is within additive error $O(\sqrt{\log(n/\delta)/m})$ of its expectation. It follows from this and Lemma 28 that if $\alpha = (1/C_M)(|\lambda_2|^2 - |\lambda_2|^4)$ for C_M a sufficiently large constant depending only on M, ϵ is sufficiently small with respect to α, and $m = \Omega_M(\log(n/\delta))$ then in the base case the algorithm computes neighbors correctly. Observe that at each layer, if the algorithm has correctly reconstructed T in all previous layers then the sets S_u for all nodes u in this layer are deterministic functions of T, and hence so are the queries the algorithm makes to $\hat{E}$. By a similar application of the union bound and Hoeffding's inequality as well as Theorem 22 and the assumption that d is sufficiently large with respect to M it follows that the algorithm succeeds at all subsequent layers as well.

Note that provided we take $\epsilon > 0$ is sufficiently small, we can show the base case of the argument will still succeed by using the triangle inequality, and the inductive step in the argument will succeed because of Theorem 22.

Finally, in the case that $\lambda_2(M) = 0$, we let φ be a generalized eigenvector such that $M\varphi \neq 0$ but $M^2\varphi = 0$. Note that such a vector must exist because, 0 is an eigenvalue of algebraic multiplicity $q - 1$ as M is ergodic and $\lambda_2 = 0$, and because our assumption on M rules out the case that M is rank one, so it's Jordan normal form must have at least one Jordan block with size at least 2 and this corresponds to the existence of such a generalized eigenvector φ. Now observe for such a φ that if u, v are siblings in T at depth ℓ then

$$\langle \varphi, \mathbb{E}[e(X_u)e(X_v)^T]\varphi \rangle = \langle M\varphi, \Pi_{\ell-1} M\varphi \rangle$$

which by Lemma 28 is lower bounded by a constant $C'_M > 0$, while if u, v are not siblings,

$$\langle \varphi, \mathbb{E}[e(X_u)e(X_v)^T]\varphi \rangle = 0.$$

Setting $\alpha = C'_M/2$ and defining the remaining constants similarly to above ensures the algorithm succeeds, by the same argument.

Note that in both the case $\lambda_2(M) \neq 0$ and $\lambda_2(M) = 0$, the algorithm is implemented by taking the expectation of certain functions over the samples, so it is straightforwardly implementable with SQ queries by replacing the empirical expectation with the VSTAT oracle. □