MCMC Learning

Varun Kanade*
École normale supérieure
varun.kanade@ens.fr

Elchanan Mossel†
University of Pennsylvania and University of California, Berkeley
mossel@stat.berkeley.edu

June 15, 2015

Abstract

The theory of learning under the uniform distribution is rich and deep, with connections to cryptography, computational complexity, and the analysis of boolean functions to name a few areas. This theory however is very limited due to the fact that the uniform distribution and the corresponding Fourier basis are rarely encountered as a statistical model.

A family of distributions that vastly generalizes the uniform distribution on the Boolean cube is that of distributions represented by Markov Random Fields (MRF). Markov Random Fields are one of the main tools for modeling high dimensional data in many areas of statistics and machine learning.

In this paper we initiate the investigation of extending central ideas, methods and algorithms from the theory of learning under the uniform distribution to the setup of learning concepts given examples from MRF distributions. In particular, our results establish a novel connection between properties of MCMC sampling of MRFs and learning under the MRF distribution.

1 Introduction

The theory of learning under the uniform distribution is well developed and has rich and beautiful connections to discrete Fourier analysis, computational complexity, cryptography and combinatorics to name a few areas. However, these methods are very limited since they rely on the assumption that examples are drawn from the uniform distribution over the Boolean cube or other product distributions. In this paper we make a first step in extending ideas, techniques and algorithms from this theory to a much broader family of distributions, namely, to Markov Random Fields.

*This work was performed while the author was at the University of California, Berkeley and at the Simons Institute, Berkeley

†Supported NSF grants DMS 1106999 and CCF 1320105, ONR grant number N00014-14-1-0823 and grant 328025 from the Simons Foundation

1.1 Learning Under the Uniform Distribution

Since the seminal work of Linial et al. (1993), the study of learning under the uniform distribution has developed into a major area of research; the principal tool is the simple and explicit Fourier expansion of functions defined on the boolean cube ($\{-1,1\}^n$):

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S(x), \quad \chi_S(x) = \prod_{i \in S} x_i.$$

This connection allows a rich class of algorithms that are based on learning coefficients of f for several classes of functions. Moreover, this connection allows application of sophisticated results in the theory of Boolean functions including hyper-contractivity, number theoretic properties and invariance, *e.g.* (O'Donnell and Servedio, 2007, Shpilka and Tal, 2011, Klivans et al., 2002). On the other hand, the central role of the uniform distribution in computational complexity and cryptography relates learning under the uniform distribution to key themes in theoretical computer science including de-randomization, hardness and cryptography, *e.g.* (Kharitonov, 1993, Naor and Reingold, 2004, Dachman-Soled et al., 2008).

Given the elegant theoretical work in this area, it is a little disappointing that these results and techniques impose such stringent assumptions on the underlying distribution. The assumption of independent examples sampled from the uniform distribution is an idealization that would rarely, if ever, be applicable in practice. In *real* distributions, features are correlated and correlations deem the analysis of algorithms that assume independence useless. Thus, it is worthwhile to ask the following question:

Question 1: *Can the Fourier Learning Theory extend to correlated features?*

1.2 Markov Random Fields

Markov random fields are a standard way of representing high dimensional distributions (see *e.g.* (Kinderman and Snell, 1980)). Recall that a Markov random field on a finite graph $G = (V, E)$ and taking values in a discrete set $\mathcal{A}$, is a probability distribution on $\mathcal{A}^V$ of the form $\Pr[(\sigma_v)_{v \in V}] = Z^{-1} \prod_C \phi_C((\sigma_v)_{v \in C})$, where the product is over all cliques C in the graph, ϕ_C are some non-negative valued functions and Z is the normalization constant. Here $(\sigma_v)_{v \in V}$ is an assignment from $V \to \mathcal{A}$.

Markov Random Fields are widely used in vision, computational biology, biostatistics, spatial statistics and several other areas. The popularity of Markov Random Fields as modeling tools is coupled with extensive algorithmic theory studying sampling from these models, estimating their parameters and recovering them. However, to the best of our knowledge the following question has not been studied.

Question 2: *For an unknown function $f : \mathcal{A}^V \to \{-1, 1\}$ from a class $\mathcal{F}$ and labeled samples from the Markov Random Field, can we learn the function?*

Of course the problem stated above is a special case of learning a function class given a general distribution (Valiant, 1984, Kearns and Vazirani, 1994). Therefore, a learning algorithm that can be applied for a general distribution can be also applied to MRF distributions. However, the real question that we seek to ask above is the following: *Can we utilize the structure of the MRF to obtain better learning algorithms?*

1.3 Our Contributions

In this paper we begin to provide an answer to the questions posed above. We show how methods that have been used in the theory of learning under the uniform distribution can be also applied for learning from certain MRF distributions.

This may sound surprising as the theory of learning under the uniform distribution strongly relies on the explicit Fourier representation of functions. Given an MRF distribution, one can also imagine expanding a function in terms of a *Fourier basis* for the MRF, the eigenvectors of the transition matrix of the Gibbs Markov Chain associated with the MRF, which are orthogonal with respect to the MRF distribution. It seems however that this approach is naïve since:

(a) Each eigenvector is of size $|\mathcal{A}|^{|V|}$; how does one store them?

(b) How does one find these eigenvectors?

(c) How does one find the expansion of a function in terms of these eigenvectors?

MCMC Learning: The main effort in this paper is to provide an answer to the questions above. For this we use Gibbs sampling, which is a Markov chain Monte Carlo (MCMC) algorithm that is used to sample from an MRF. We will use this MCMC method as the main engine in our learning algorithms. The Gibbs MC is reversible and therefore its eigenvectors are orthogonal with respect to the MRF distribution. Also, the sampling algorithm is straightforward to implement given access to the underlying graph and potential functions. There is a vast literature studying the convergence rates of this sampling algorithm; our results require that the Gibbs samplers are rapidly mixing.

In Section 4, we show how the eigenvectors of the transition matrix of the Gibbs MC can be computed implicitly. We focus on the eigenvectors corresponding to the higher eigenvalues. These eigenvectors correspond to the stable part of the spectrum, *i.e.* the part that is not very sensitive to small perturbation. Perhaps surprisingly, despite the exponential size of the matrix, we show that it is possible to adapt the power iteration method to this setting.

A function from $\mathcal{A}^V \to \mathbb{R}$ can be viewed as a $|\mathcal{A}|^{|V|}$ dimensional vector and thus applying powers of the transition matrix to it results in another function from $\mathcal{A}^V \to \mathbb{R}$. Observe that the powers of a transition matrix define distributions in time over the state space of the the Gibbs MC. Thus, the value of the function obtained by applying powers of a transition matrix can be approximated by sampling using the Gibbs Markov chain. Our main technical result (see Theorem 1) shows that any function approximated by "top" eigenvectors of the transition matrix of the Gibbs MC can be expressed a linear combination of powers of the the transition matrix applied to a suitable collection of "basis" functions, whenever certain technical conditions hold.

The reason for focusing on the part of the spectrum corresponding to stable eigenvectors is twofold. First, it is technically easier to access this part of the spectrum. Furthermore, we think of eigenvectors corresponding to small eigenvalues as unstable. Consider Gibbs sampling as the true temporal evolution of the system and let ν be an eigenvector corresponding to a small eigenvalue. Then calculating $\nu(x)$ provides very little information on $\nu(y)$ where y is obtained from x after a short evolution of the Gibbs sampler. The reasoning just applied is a generalization of the classical reasoning for concentrating on the low frequency part of the Fourier expansion in traditional signal processing.

Noise Sensitivity and Learning: In the case of the uniform distribution, the noise sensitivity (with parameter ϵ) of a boolean function f, is defined as the probability that $f(x) \neq f(y)$, where

x is chosen uniformly at random and y is obtained from x by flipping each bit with probability ϵ. Klivans et al. (2002) gave an elegant characterization of learning in terms of noise sensitivity. Using this characterization, they showed that intersections and thresholds of halfspaces can be elegantly learned with respect to the uniform distribution. In Section 4.3, we show that the notion of noise sensitivity and the results regarding functions with low noise sensitivity can be generalized to MRF distributions.

Learning Juntas: We also consider the so-called junta learning problem. A junta is a function that depends only on a small subset of the variables. Learning juntas from i.i.d. examples is a notoriously difficult problem, see (Blum, 1992, Mossel et al., 2004). However, if the learning algorithm has access to *labeled* examples that are received from a Gibbs sampler, these correlated examples can be useful for learning juntas. We show that under standard technical conditions on the Gibbs MC, juntas can be learned in polynomial time by a very simple algorithm. These results are presented in Section 5.

Relation to Structure Learning: In this paper, we assume that learning algorithms have the ability to sample from the Gibbs Markov Chain corresponding to the MRF. While such data would be hard to come by in practice, we remark that there is a vast literature regarding learning the structure and parameters of MRFs using *unlabeled* data and that it has recently been established that this can be done efficiently under very general conditions (Bresler, 2014). Once the structure of the underlying MRF is known, Gibbs sampling is an extremely efficient procedure. Thus, the methods proposed in this work could be used in conjunction with the techniques for MRF structure learning. The eigenvectors of the transition matrix could be viewed as *features* for learning, thus the methods proposed in this paper can be viewed as feature learning.

1.4 Related Work

The idea of considering Markov Chains or Random Walks in the context of learning is not new. However, none of the results and models considered before give non-trivial improvements or algorithms in the context of MRFs. Work of Aldous and Vazirani (1995) studies a Markov chain based model where the main interest was in characterizing the number of new nodes visited. Gamarnik (1999) observed that after the mixing time a chain can simulate i.i.d. samples from the stationary distribution and thus obtained learning results for general Markov chains. Bartlett et al. (1994) and Bshouty et al. (2005) considered random walks on the discrete cube and showed how to utilize the random walk model to learn functions that cannot be easily learned from i.i.d. examples from the uniform distribution on the discrete cube. In this same model, Jackson and Wimmer (2014) showed that agnostic learning parities and PAC-learning thresholds of parities (TOPs) could be performed in quasi-polynomial time.

2 Preliminaries

Let X be an instance space. In this paper, we will assume that X is finite and in particular we are mostly interested in the case when $X = \mathcal{A}^n$, where $\mathcal{A}$ is some finite set. For $x, x' \in \mathcal{A}^n$, let $d_H(x, x')$ denote the Hamming distance between x and x', *i.e.* $d_H(x, x') = |\{i \mid x_i \neq x'_i\}|$.

Let $M = \langle X, P \rangle$ denote a time-reversible discrete time ergodic Markov chain with transition matrix P. When $X = \mathcal{A}^n$, we say that M has *single-site* transitions if for any *legal* transition $x \to x'$ it is the case that $d_H(x, x') \leq 1$, *i.e.* $P(x, x') = 0$ when $d_H(x, x') > 1$. Let $X^0 = x_0$ denote

the starting state of a Markov chain M. Let $P^t(x_0, \cdot)$ denote the distribution over states at time t, when starting from x_0. Let π denote the stationary distribution of M. Denote by $\tau_M(x_0)$ the quantity:

$$\tau_M(x_0) = \min\{t : \|P^t(x_0, \cdot) - \pi\|_{\text{TV}} \leq \frac{1}{4}\}$$

Then, define the mixing time of M as $\tau_M = \max_{x_0 \in X} \tau_M(x_0)$. We say that a Markov chain with state space $X = \mathcal{A}^n$ is rapidly mixing if $\tau_M \leq \text{poly}(n)$.

While all the results in this paper are general, we describe two basic graphical models that will aid the discussion.

2.1 Ising Model

Consider a collection of nodes, $[n] = \{1, \ldots, n\}$, and for each pair i, j, there is an associated interaction energy, β_{ij}. Suppose $([n], E)$ denotes the graph, where $\beta_{ij} = 0$ for $(i, j) \notin E$. A state σ of the system consists of an assignment of *spins*, $\sigma_i \in \{+1, -1\}$, to the nodes $[n]$. The Hamiltonian of configuration σ is defined as

$$H(\sigma) = -\sum_{(i,j)\in E} \beta_{ij}\sigma_i\sigma_j - B\sum_{i\in[n]} \sigma_i,$$

where B is the external field. The energy of a configuration σ is $\exp(-H(\sigma))$.

The Glauber dynamics on the Ising model defines the Gibbs Markov Chain $M = \langle\{-1, 1\}^n, P\rangle$, where the transitions are defined as follows:

(i) In state σ, pick a node $i \in [n]$ uniformly at random. With probability $1/2$ do nothing, otherwise

(ii) Let σ' be obtained by flipping the spin at node i. Then, with probability $\exp(-H(\sigma'))/(\exp(-H(\sigma)+\exp(-H(\sigma')))\}$, the state at the next time-step is σ'. Otherwise the state at the next time-step remains unchanged.

The stationary distribution of the above dynamics is the *Gibbs distribution*, where $\pi(\sigma) \propto \exp(-H(\sigma))$. It is known that there exists a $\beta(\Delta) > 0$ such that for all graphs of maximal degree Δ, if $\max |\beta_{i,j}| < \beta(\Delta)$ then the dynamics above is rapidly mixing (Dobrushin and Shlosman, 1985, Mossel and Sly, 2013).

2.2 Graph Coloring

Let $G = ([n], E)$ be a graph. For any $q > 0$, a *valid* q-coloring of the graph G is a function $C : V \to [q]$ such that for every $(i, j) \in E$, $C(i) \neq C(j)$. For a node i, let $N(i) = \{j \mid (i, j) \in E\}$ denote the set of neighbors of i. Consider the Markov chain defined by the following transition:

(i) In state (valid coloring) C, choose a node $i \in [n]$ uniformly at random. With probability $1/2$ do nothing, otherwise:

(ii) Let $S \subseteq [q]$ be the subset of colors defined by $S = \{C(j) \mid j \in N(i)\}$. Define C' to be the coloring obtained by choosing a random color $c \in [q] \setminus S$ and set $C'(i) = c$, $C'(j) = C(j)$ for $j \neq i$. The state at the next time-step is C'.

The stationary distribution of the above Markov chain is uniform over the valid colorings of the graph. It is known that the above chain is rapidly mixing when the condition $q \geq 3\Delta$ is satisfied, where Δ is the maximal degree of the graph (in fact much better results are known (Jerrum, 1995, Vigoda, 1999)).

3 Learning Models

Let X be a finite instance space and let $M = \langle X, P \rangle$ be an irreducible discrete-time reversible Markov chain, where P is the transition matrix. Let π_M denote the stationary distribution of M, τ_M the mixing time. We assume that the Markov chain M is *rapidly mixing*, *i.e.* $\tau_M \leq \text{poly}(\log(|X|))$ (note that if $X = \mathcal{A}^n$, $\log(|X|) = O(n)$).

We consider the problem of learning with respect to stationary distributions of rapidly mixing Markov chains (*e.g.* defined by an MRF). The two graphical models described in the previous section serve as examples of such settings. The learning algorithm has access to the *one-step* oracle, $\mathsf{OS}(\cdot)$, that when queried with a state $x \in X$, returns the state after one step. Thus, $\mathsf{OS}(x)$ is a random variable with distribution $P(x, \cdot)$ and can be used to simulate the Markov chain.

Let $\mathcal{F}$ be a class of boolean functions over X. The goal of the learning algorithm is to learn an unknown function, $f \in \mathcal{F}$, with respect to the stationary distribution π_M of the Markov chain M. As described above, the learning algorithm has the ability to simulate the Markov chain using the one-step oracle. We will consider both PAC learning and agnostic learning. Let $L : X \to \{-1, 1\}$ be a (possibly randomized) labeling function. In the case of PAC learning L is just the target function f; in the case of agnostic learning L is allowed to be completely arbitrary. Let D denote the distribution over $X \times \{-1, 1\}$, where for any $(x, y) \sim D$, $x \sim \pi_M$ and $y = L(x)$.

PAC Learning (Valiant, 1984): In PAC learning the labeling function is the target function f. The goal of the learning algorithm is to output a hypothesis, $h : X \to \{-1, 1\}$, which with probability at least $1 - \delta$ satisfies $\text{err}(h) = \Pr_{x \sim \pi_M}[h(x) \neq f(x)] \leq \epsilon$.

Agnostic Learning (Kearns et al., 1994, Haussler, 1992): In agnostic the labeling function L may be completely arbitrary. Let D be the distribution as defined above. Let $\text{opt} = \min_{f \in \mathcal{F}} \Pr_{(x,y) \sim D}[f(x) \neq y]$. The goal of the learning algorithm is to output a hypothesis, $h : X \to \{-1, 1\}$, which with probability at least $1 - \delta$ satisfies,

$$\text{err}(h) = \Pr_{(x,y) \sim D}[h(x) \neq y] \leq \text{opt} + \epsilon$$

Typically, one requires that the learning algorithm have time and sample complexity that is polynomial in n, $1/\epsilon$ and $1/\delta$. So far, we have not mentioned what access the learning algorithm has to labeled examples. We consider two possible settings.

Learning with i.i.d. examples only: In this setting, in addition to having access to the one-step oracle, $\mathsf{OS}(\cdot)$, the learning algorithm has access to the standard example oracle, which when queried returns an example $(x, L(x))$, where $x \sim \pi_M$ and L is the (possibly randomized) labeling function.

Learning with labeled examples from MC: In this setting, the learning algorithm has access to a *labeled* random walk, $(x^1, L(x^1)), (x^2, L(x^2)), \ldots$, of the Markov chain. Here x^{i+1} is the (random) state one time-step after x^i and L is the labeling function. Thus, the learning algorithm can potentially exploit *correlations* between consecutive examples.

The results in Section 4 only require access to i.i.d. examples. Note that these are sufficient to compute inner products with respect to the underlying distribution, a key requirement for *Fourier analysis*. The result in Section 5 is only applicable in the stronger setting where the learning algorithm receives examples from a labeled Markov chain. Note that since the chain is rapidly mixing, the learning algorithm by itself is able to (approximately) simulate i.i.d. random examples.

4 Harmonic Analysis using Eigenvectors

In this section, we show that the eigenvectors of the transition matrix can be (approximately) expressed as linear combinations of a suitable collection of basis functions and powers of the transition matrix applied to them.

Let $M = \langle X, P \rangle$ be a time-reversible discrete Markov chain. Let π be the stationary distribution of M. We consider the set of right-eigenvectors of the matrix P. The largest eigenvalue of P is 1 and the corresponding eigenvector has 1 in each co-ordinate. The left-eigenvector in this case is the stationary distribution. For simplicity of analysis we assume that $P(x, x) \geq 1/2$ for all x which implies that all the eigenvalues of P are non-negative. We are interested in identifying as many as possible of the remaining eigenvectors with eigenvalues less than 1.

For functions, $f, g : X \to \mathbb{R}$, define the inner-product, $\langle f, g \rangle = \mathbb{E}_{x \sim \pi}[f(x)g(x)]$, and the norm $\|f\|_2 = \sqrt{\langle f, f \rangle}$. Throughout this section, we will always consider inner products and norms with respect to the distribution π.

Since M is reversible, the right eigenvectors of P are orthogonal with respect to π. Thus, these eigenvectors can be used as a basis to represent functions from $X \to \mathbb{R}$. First, we briefly show that this approach generalizes the standard Fourier analysis on the Boolean cube, which is commonly used in uniform-distribution learning.

4.1 Fourier Analysis over the Boolean Cube

Let $\{-1, 1\}^n$ denote the boolean cube. For $S \subseteq [n]$, the parity function over S is defined as $\chi_S(x) = \prod_{i \in S} x_i$. With respect to the uniform distribution U_n over $\{-1, 1\}^n$, the set of parity functions $\{\chi_S \mid S \subseteq [n]\}$ form an orthonormal *Fourier* basis, *i.e.* for $S \neq T$, $\mathbb{E}_{x \sim U_n}[\chi_S(x)\chi_T(x)] = 0$ and $\mathbb{E}_{x \sim U_n}[\chi_S(x)^2] = 1$.

We can view the uniform distribution over $\{-1, 1\}^n$ as arising from the stationary distribution of the following simple Markov chain. For x, x', such that $x_i \neq x'_i$ and $x_j = x'_j$ for $j \neq i$, let $P(x, x') = 1/(2n)$; $P(x, x) = 1/2$. The remaining values of the matrix P are set to 0. This chain is rapidly mixing with mixing time $O(n \log(n))$ and the stationary distribution is the uniform distribution over $\{-1, 1\}^n$. It is easy to see and well known that every parity function χ_S is an eigenvector of P with eigenvalue $1 - |S|/n$. Thus, Fourier-based learning under the uniform distribution can be seen as a special case of Harmonic analysis using eigenvectors of the transition matrix.

4.2 Representing Eigenvectors Implicitly

As in the case of the uniform distribution over the boolean cube, we would like to find the eigenvectors of the transition matrix of a general Markov chain, M, and use these as an orthonormal basis for learning. Unfortunately, in most cases of interest explicit succinct representations of eigenvectors don't necessarily exist and the size of the set $|X|$ is likely to be prohibitively large, typically exponential in n, where n is the length of the vectors in X. Thus, it is not possible to use standard techniques to obtain eigenvectors of P. Here, we show how these eigenvectors may be computed implicitly.

An eigenvector of the transition matrix P is a function $\nu : X \to \mathbb{R}$. Throughout this section, we will view any function $g : X \to \mathbb{R}$ as an $|X|$-dimensional vector with value $g(x)$ at position x. As such, even writing down such a vector corresponding to an eigenvector ν is not possible in

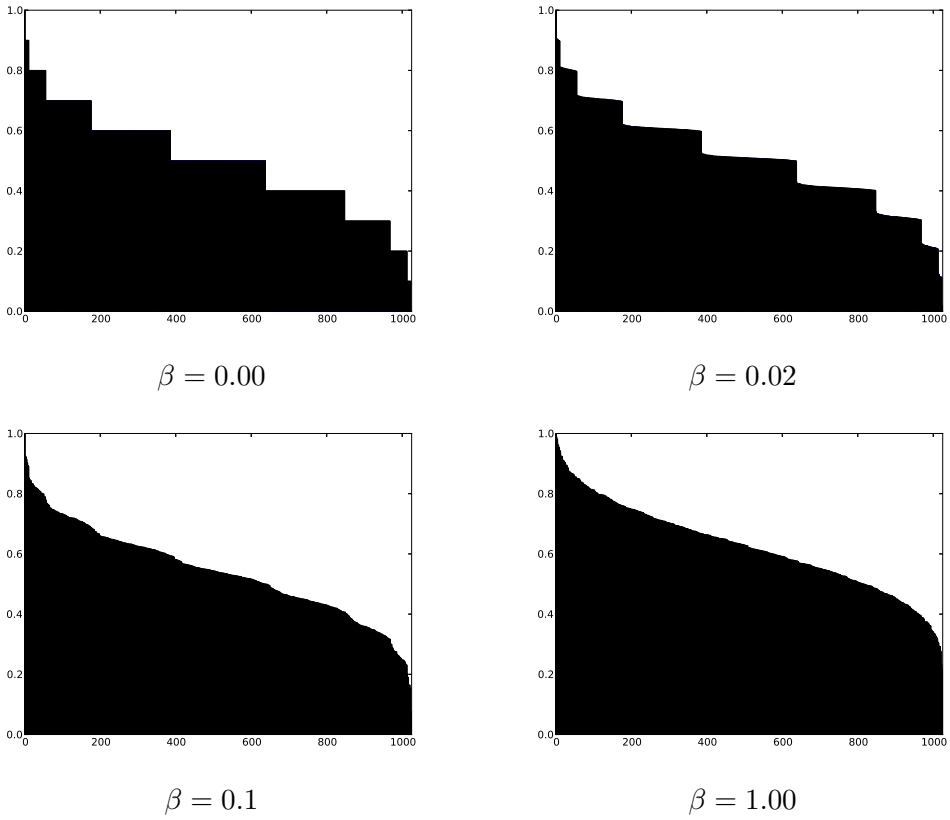

Figure 1: Spectrum of the transition matrix of the Gibbs MC for the Ising model on a cycle of length 10 for various values of β, the inverse temperature parameter.

polynomial time. Instead, our goal is to show that whenever a suitable collection of *basis functions* exists, the eigenvectors have a simple representation in terms of these *basis functions* and powers of the transition matrix applied to them, as long as the underlying Markov chain M satisfies certain conditions. The condition we require is that the *spectrum* of the transition matrix be *discrete*, *i.e.* eigenvalues show sharp drops. Between these drops, the eigenvalues may be quite close to each other, and in fact even equal. Figure 1 shows the spectrum of the transition matrix of the Ising model on a cycle of 10 nodes for various values of β, the inverse temperature parameter. The case when $\beta = 0$ corresponds to the uniform distribution on $\{-1, 1\}^{10}$. One notices that the spectrum is *discrete* for small values of β (high-temperature regime).

Next, we formally define the requirements of a discrete spectrum.

Definition 1 (Discrete Spectrum)**.** *Let P be the transition matrix of a Markov chain and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq \cdots \geq 0$ be the eigenvalues of P in non-increasing order. We say that P has an (N, k, γ, c)-discrete spectrum, if there exists a sequence $1 \leq i_1 \leq i_2 \leq \cdots \leq i_k \leq |X|$ such that the following are true*

1. *Between λ_{i_j} and λ_{i_j+1}, there is a non-trivial gap, i.e. for $j \in \{i_1, \ldots, i_k\}$, $\frac{\lambda_{j+1}}{\lambda_j} \leq \gamma < 1$*
2. *Let $i_0 = 1$, we refer to $S_j = \{i_{j-1}+1, \ldots, i_j\}$ as the j^{th} block (of eigenvalues and eigenvectors). Then the size of each block, $|S_j| \leq N$*

3. *The eigenvalue λ_{i_k} is not too small (with respect to the gap at the end of each block), $\lambda_{i_k} \geq \gamma^c$*

In general, the parameter γ will depend on n and we require that $\gamma \leq 1 - 1/\text{poly}(n)$ in order to separate eigenvectors from the various blocks. One would expect N to have dependence on both n and k and c to have some dependence on k. As an example, we note that the spectrum corresponding to the Markov chain discussed in Section 4.1 is indeed discrete with the following parameters: k can be any integer, $N = n^k$, $\gamma = 1 - (1/n)$ and $c = O(k)$.

In order to extract eigenvectors of P, we start with a collection of functions which have significant *Fourier mass* on the top eigenvectors. For an eigenvector ν, it's *Fourier coefficient* in any function $g : X \to \mathbb{R}$ is simply $\langle g, \nu \rangle$. Condition 2 in Definition 2 implicitly requires that the inner product $\langle g, \nu \rangle$ be large for ν with a large eigenvalue for some g in the set. In addition, since eigenvalues corresponding to different eigenvectors may be equal or close together, we require a set of functions where the matrix corresponding to the *Fourier coefficients* of such eigenvectors is well-conditioned. Formally, we define the notion of a *useful basis* of functions with respect to a transition matrix P which has an (N, k, γ, c)-discrete spectrum.

Definition 2 (Useful Basis)**.** *Let $\mathcal{G}$ be a collection of functions from $X \to \mathbb{R}$. We say that $\mathcal{G}$ is α-useful for an (N, k, γ, c)-discrete P if the following hold:*

1. *For every $g \in \mathcal{G}$, $\|g\|_\infty \leq 1$*
2. *Let $i_0 = 0$, then for any $1 \leq j \leq k$, if $N_j = i_j - i_{j-1}$ (the size of the j^{th} block), there exist N_j functions $g_1, \ldots, g_{N_j} \in \mathcal{G}$, such that the $N_j \times N_j$ matrix A defined by $a_{m,l} = \langle g_m, \nu_l \rangle$, where $m \in \{1, \ldots, N_j\}$ and $l \in \{i_{j-1} + 1, \ldots, i_j\}$, has smallest singular value at least $1/\alpha$. Alternatively, the operator norm of A^{-1}, $\|A^{-1}\|_{op}$ is at most α.*

The parameter α will have dependence on N—a polynomial dependence on N would result in efficient algorithms. In general, it is not known which Markov chains admit a useful basis that has a succinct representation. In the case of the uniform distribution, clearly the collection of parity functions already is such a *useful basis*. However, we observe that there are other useful bases as well. For example if for some k, one wished to extract all eigenvectors with eigenvalues at least $1 - k/n$ (parities of size at most k), one can start with the collection of functions that is disjunctions (or conjunctions) on at most k variables. Note that in this case, there is no contribution from eigenvectors with low eigenvalues (*i.e.* noise) in the basis functions. However, one would not expect to find such a *useful basis* without any contributions from eigenvectors with low eigenvalues when the stationary distribution is not product.

We now show how functions from a *useful basis* for a transition matrix with a discrete spectrum can be used to extract eigenvectors. First by applying powers of P to some function g, the contributions of eigenvectors in different blocks can be separated. However, to separate eigenvectors within a block we require an *incoherence condition* among the various g_ms (which is the second condition in Definition 2). We first show that the eigenvectors ν can be approximately represented in the following form:

$$\nu \approx \sum_{t,m} \beta_{t,m} P^t g_m,$$

where m indexes the functions in $\mathcal{G}$.

Theorem 1. *Let P be a transition matrix with an (N, k, γ, c) discrete spectrum and let $\mathcal{G}$ be an α-useful basis for P. Then for any $\epsilon > 0$, there exists τ_{max} and B such that every eigenvector ν_ℓ with $\ell \leq i_k$ can be expressed as:*

$$\nu_\ell = \sum_{t,m} \beta_{t,m}^{\ell} P^t g_m + \eta_i$$

where $\|\eta_i\|_2 \leq \epsilon$, $t \leq \tau_{max}$ and $\sum_{t,m} |\beta_{t,m}| \leq B$. Furthermore,

$$B = (2\alpha N k)^{\Theta((1+c)^{k+1})} \epsilon^{-(1+c)^k}$$

$$\tau_{max} = O\left(k(1+c)^{k-1}(\log(N) + \log(k) + \log(\alpha) + + \frac{1}{\log(1/\gamma)} + \log(\frac{1}{\epsilon}))\right)$$

The proof of the above theorem is somewhat delicate and is provided in Appendix A.1. Notice that the bounds on B and τ have a relatively mild (polynomial) dependence on most parameters except k and c. Thus, when c and k are relatively small, for example both of them constant, both B and τ are bounded by polynomials in the other parameters. Also, N may be somewhat large, in the case of the uniform distribution $N = \Theta(n^k)$—though this is still polynomial if k is constant.

We can now use the above Theorem to devise a simple learning algorithm with respect to stationary distribution of the Markov chain. In fact, the learning algorithm does not even need to explicitly estimate the values of $\beta_{t,m}^{\ell}$ in the statement of Theorem 1—the result shows that any linear combination of the eigenvectors can also be represented as a linear combination of the collection of functions $\{P^t g_m\}_{t \leq \tau_{\max}, g_m \in \mathcal{G}}$. Thus, we can treat this collection as "features" and simply perform linear regression (either L_1 or L_2) as part of the learning algorithm. The algorithm is given in Figure 2. The key idea is to show that $P^t g_m(x)$ can be approximately computed for any $x \in X$ with blackbox access to g_m and the one-step oracle $OS(\cdot)$. This is because $P^t g_m(x) = \mathbb{E}_{y \sim P^t(x,\cdot)}[g(y)]$, where $P^t(x, \cdot)$ is the distribution over X obtained by starting from x and taking t steps of the Markov chain. The functions $\phi_{t,m}$ in the algorithm are computing approximations to $P^t g_m$ and then using them as features for learning. Formally, we can prove the following theorem.

Theorem 2. *Let $M = \langle X, P \rangle$ be a Markov chain and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ denote the eigenvalues of P and ν_ℓ the eigenvector corresponding to λ_ℓ. Let π be the stationary distribution of P. Let $\mathcal{F}$ be a class of boolean functions. Suppose for some $\epsilon > 0$, there exists $\ell^*(\epsilon)$ such that for every $f \in \mathcal{F}$,*

$$\sum_{\ell > \ell^*} \langle f, \nu_\ell \rangle^2 \leq \tfrac{\epsilon^2}{4},$$

i.e. every f can be approximated (up to $\epsilon^2/4$) by the top ℓ^* eigenvectors of P. Suppose P has a (N, k, γ, c)-discrete spectrum as defined in Definition 1, with $i_k \geq \ell^*$ and that $\mathcal{G}$ is an α-useful basis for P. Then, there exists a learning algorithm that with blackbox access to functions $g \in \mathcal{G}$, the one-step oracle $\mathsf{OS}(\cdot)$ for Markov chain M, and access to random examples $(x, L(x))$ where $x \sim \pi$ and L is an arbitrary labeling function, agnostically learns $\mathcal{F}$, up to error ϵ.

Furthermore, the running time, sample complexity and the time required to evaluate the output hypothesis are bounded by a polynomial in $(Nk)^{(1+c)^{k+1}}, \epsilon^{-(1+c)^k}, |\mathcal{G}|, n$. In particular, if ϵ is a constant, c and k depend only on ϵ (and not on n), and $N \leq n^{\zeta(k)}$, where ζ may be an arbitrary function, the algorithm runs in polynomial time.

We give the proof this theorem in Appendix A.2; the proof uses the L_1-regression technique of Kalai et al. (2005). We comment that the learning algorithm (Fig. 2) is a generalization of the

Inputs: $\tau_{\max}, W, T$, blackbox access to $g \in \mathcal{G}$ and $\mathsf{OS}(\cdot)$, labeled examples $\langle (x_i, y_i) \rangle_{i=1}^{s}$

Preprocessing: For each $t \leq \tau_{\max}$ and m such that $g_m \in \mathcal{G}$

- For each $i = 1, \ldots, s$, let

$$\phi_{t,m}(x_i) = \frac{1}{T} \sum_{j=1}^{T} g_m(\mathsf{OS}_j^t(x_i)), \tag{1}$$

  where $\mathsf{OS}_j^t(x_i)$ denotes the point obtained by an independent forward simulation of the Markov chain starting at x_i for t steps, for each j.

Linear Program: Solve the following linear program:

$$\begin{aligned} \text{minimize} & \sum_{i=1}^{s} \left| \sum_{t \leq \tau_{\max}, g_m \in \mathcal{G}} w_{t,m} \phi_{t,m}(x_i) - y_i \right| \\ \text{subject to} & \sum_{t \leq \tau_{\max}, g_m \in \mathcal{G}} |w_{t,m}| \leq W \end{aligned}$$

Output Hypothesis:

- Let $h(x) = \sum_{t \leq \tau_{\max}, g_m \in \mathcal{G}} w_{t,m} \phi_{t,m}(x)$, where $\phi_{t,m}(x)$ are defined as in step (1) above.
- Let $\theta \in [-1, 1]$ be chosen uniformly at random and output $\text{sign}(h(x) - \theta)$ as prediction

Figure 2: Agnostic Learning with respect to MRF distributions

low-degree algorithm of Linial et al. (1993). Also, when applied to the Markov chain corresponding to the uniform distribution over $\{-1,1\}^n$, this algorithm works whenever the low-degree algorithm does (albeit with slightly worse bounds). As an example, we consider the algorithm of Klivans et al. (2002) to learn arbitrary functions of halfspaces. As a main ingredient of their work, they showed that halfspaces can be approximated by the first $O(1/\epsilon^4)$ levels of the Fourier spectrum. The running time of our learning algorithm run with a useful basis consisting of parities, or conjunctions of size $O(1/\epsilon^4)$ is polynomial (for constant ϵ).

4.3 Noise Sensitivity Analysis

In light of Theorem 2, one can ask which function classes are well-approximated by top eigenvectors and for which MRFs. A generic answer is functions that are "noise-stable" with respect to the underlying Gibbs Markov chain. Below, we generalize the definition of noise sensitivity in the case of product distributions to apply under MRF distributions. In words, the noise sensitivity (with parameter t) of a boolean function f is the probability that $f(x)$ and $f(y)$ are different, where $x \sim \pi$ is drawn from the stationary distribution and y is obtained by taking t steps of the Markov chain starting at x.

Definition 3. *Let $x \sim \pi$ from the stationary distribution of P and $y \sim P^t(x,\cdot)$, the distribution obtained by taking t steps of the Gibbs MC starting at x. For a boolean function $f : X \to \{-1,1\}$, define its noise sensitivity with respect to parameter t and the transition matrix P of the Gibbs MC as*

$$\mathrm{NS}_t(f) = \Pr_{x\sim\pi, y\sim P^t(x,\cdot)}[f(x) \neq f(y)].$$

One can derive an alternative form for the noise sensitivity as follows. Let $\lambda_1 \geq \lambda_2 \geq \cdots 0$ denote the eigenvalues of P and $\nu_1, \nu_2, \ldots$ the corresponding eigenvectors. Let $\hat{f}_\ell = \langle f, \nu_\ell \rangle$. Then,

$$\begin{aligned}
\mathrm{NS}_t(f) &= \Pr_{x\sim\pi, y\sim P^t(x,\cdot)}[f(x) \neq f(y)] \\
&= \frac{1}{2}\mathbb{E}_{x\sim\pi, y\sim P^t(x,\cdot)}[1 - f(x)f(y)] \\
&= \frac{1}{2} - \frac{1}{2}\langle f, P^t f\rangle \\
&= \frac{1}{2} - \frac{1}{2}\sum_\ell \lambda_\ell^t \hat{f}_\ell^2 \qquad (2)
\end{aligned}$$

The notion of noise-sensitivity has been fruitfully used in the theory of learning under the uniform distribution (see for example Klivans et al. (2002)). The main idea is that functions that have low noise sensitivity have most of their mass concentrated on "lower order Fourier coefficients", *i.e.* eigenvectors with large eigenvalues. We show that this idea can be easily generalized in the context of MRF distributions. The proof of the following theorem is provided in Appendix A.3.

Theorem 3. *Let P be the transition matrix of the Gibbs MC of an MRF and let $f : X \to \{-1,1\}$ be a boolean function. Let ℓ^* be the largest index such that $\lambda_{\ell^*} > \rho$, then:*

$$\sum_{\ell > \ell^*} \hat{f}_\ell^2 \leq \frac{e}{e-1}\,\mathrm{NS}_{-\frac{1}{\ln\rho}}(f)$$

Thus, it is of interest to study which function classes have low noise-sensitivity with respect to certain MRFs. As an example, we consider the Ising model on graphs with bounded degrees; the Gibbs MC in this case is the Glauber dynamics. We show that the class of halfspaces have low noise sensitivity with respect to this family of MRFs. In particular, the noise sensitivity with parameter t, only depends on (t/n).

Proposition 1. *For every $\Delta \geq 0$, there exists $\beta(\Delta) > 0$ such that the following holds: For every graph G with maximum degree Δ, the Ising model with $\beta < \beta(\Delta)$ and any function of the form $f = \mathrm{sign}(\sum_{i=1}^{n} w_i x_i)$, it holds that $\mathrm{NS}_t(f) \leq \exp(-\delta(t/n))$, for some constant δ that depends only on Δ.*

The proof of the above proposition follows from Lemma 1 in Appendix A.4. As a corollary we get.

Corollary 1. *Let P be the transition matrix of the Gibbs MC of an Ising model with bounded degree Δ. Suppose that for some $\epsilon > 0$, P has an (N, k, γ, c)-discrete spectrum such that k depends only on ϵ and Δ, $\lambda_{i_k+1} < \exp(-\frac{\delta}{1} \cdot \frac{1}{\ln(4/\epsilon^2)})$ (where δ is as in Proposition 1), $\gamma = 1 - 1/\mathrm{poly}(n)$, N is $\mathrm{poly}(n)$ and c a constant, for constant ϵ, Δ. Furthermore, suppose that P admits an α-useful basis with $\alpha = \mathrm{poly}(n, 1/\epsilon)$, for the parameters (N, k, γ, c) as above. Then the class of halfspaces $\{\mathrm{sign}(\sum_i w_i x_i)\}$, is agnostically learnable with respect to the stationary distribution π of P up to error ϵ.*

Proof. Let $t = \frac{n}{\delta} \ln(4/\epsilon^2)$, where δ is from Proposition 1. Thus, $\mathrm{NS}_t(f) \leq \epsilon^2/4$. Let $\rho = \exp(-1/t)$ (as in Theorem 3); by the assumption on P, P admits an (N, k, γ, c)-distribution where k depends only on ϵ, Δ, such that $\lambda_{i_k+1} < \rho$.

Now, the algorithm in Figure 2 together with the parameter settings from Theorems 1, 2 and 3 give the desired result. □

4.4 Discussion

In this section, we proposed that approximation using eigenvectors of the transition matrix of an appropriate Markov chain may be better than just polynomial approximation, when learning with respect to distributions defined by Markov random fields (not product). We checked this for a few different Ising models to approximate the majority function. Since the computations required are fairly intensive, we could only do this for relatively small models. However, we point that the methods proposed in this paper are highly-parallelizable and not beyond the reach of large computing systems. Thus, it may be of interest to run methods proposed here on larger datasets and real-world data.

Approximation of Majority: We look at three different graphs: a cycle of length 11, the complete graph on 11 nodes and an Erdős-Rényi random graph with $n = 11$ and $p = 0.3$. We looked at the Ising model on these graphs with various different values of β. In each case, we looked at degree-k polynomial approximations for $k = 2, 4$ and also with using top n^k eigenvectors of the majority function. We see that the approximation using eigenvectors is consistently better, except possibly for very low values of β, where polynomial approximations are also quite good. The values reported in the table are squared error for the approximation.

β	Degree	Poly	Eigen
0.02	2	0.3321	0.3550
	4	0.2084	0.1645
0.05	2	0.3184	0.2322
	4	0.1937	0.1648
0.1	2	0.2238	0.1417
	4	0.1199	0.0687
0.2	2	0.1468	0.0018
	4	0.0034	0.0013

(a) K_{11}

β	Degree	Poly	Eigen
0.1	2	0.3330	0.3401
	4	0.2092	0.1606
0.2	2	0.3307	0.2229
	4	0.2052	0.1538
0.5	2	0.3113	0.1918
	4	0.1676	0.0715
1.0	2	0.1857	0.0466
	4	0.0344	0.0253

(b) C_{11}

β	Degree	Poly	Eigen
0.05	2	0.3327	0.3404
	4	0.2089	0.2172
0.1	2	0.3283	0.2240
	4	0.2034	0.1515
0.2	2	0.3017	0.1897
	4	0.1757	0.1254
0.5	2	0.0690	0.0326
	4	0.0262	0.0108

(c) $G(11, 0.3)$

Table 1: Approximation of the majority function using polynomials and eigenvectors for different Ising models

5 Learning Juntas

In this section, we consider the problem of learning the class of k-juntas. Suppose $X = \mathcal{A}^n$ is the instance space. A *k-junta* is a boolean function that depends on only k out of the n possible co-ordinates of $x \in X$. In this section, we consider the model in which we receive labeled examples from a random walk of a Markov chain (see Section 3.2).[1] In this case the learning algorithm can identify the k relevant variables by keeping track of which variables caused the function to change its value.

For a subset, $S \subseteq [n]$ of the variables and a function $b_S : S \to \mathcal{A}$, let $x_S = b_S$ denote the event, $\bigwedge_{i \in S} x_i = b_S(x_i)$, *i.e.* it fixes the assignment on the variables in S as given by the function b_S. A set S is the *junta* of function f, if the variables in S completely determine the value of f. In this case, for $b_S : S \to \mathcal{A}$, every x satisfying $x_S = b_S$ has the same value $f(x)$ and by slight abuse of notation we denote this common value by $f(b_S)$.

Figure 3 describes the simple algorithm for learning juntas. Theorem 4 gives conditions under which Algorithm 3 is guaranteed to succeed. Later, we show that the Ising model and graph coloring satisfy these conditions.

Theorem 4. *Let $X = \mathcal{A}^n$ and let $M = \langle X, P \rangle$ be a time-reversible rapidly mixing MC. Let π denote the stationary distribution of M and τ_M its mixing time. Furthermore, suppose that M has* single-site *dynamics, i.e. $P(x, x') = 0$ if $d_H(x, x') > 1$ and that the following conditions hold:*

(i) For any $S \subseteq [n]$, $b_S : S \to \mathcal{A}$ either $\pi(x_S = b_S) = 0$ or $\pi(x_S = b_S) \geq 1/(c|\mathcal{A}|)^{|S|}$, where c is a constant.

(ii) For any x, x' such that $\pi(x) \neq 0$, $\pi(x') \neq 0$ and $d_H(x, x') = 1$, $P(x, x') \geq \beta$.

Then Algorithm 3 exactly *learns the class of k-junta functions with probability at least $1 - \delta$ and the running time is polynomial in $n, |\mathcal{A}|^k, \tau_M, 1/\beta, \log(1/\delta)$.*

Proof. Let f be the unknown target k-junta function. Let S be the set of variables that influence f, $|S| \leq k$. The set S is called the *junta* for f. Note that a variable i is in the junta for f, if

[1] In the model where labeled examples are received from the only from stationary distribution, it seems unlikely that any learning algorithm can benefit from access to the $\mathsf{OS}(\cdot)$ oracle. The problem of learning juntas in time $n^{o(k)}$ is a long-standing open problem even when the distribution is uniform over the Boolean cube, where the $\mathsf{OS}(\cdot)$ oracle can easily be simulated by the learner itself.

Inputs: Access to labeled examples $(x, f(x))$ from Markov Chain M

Identifying Relevant Variables

1. $\mathcal{J} = \emptyset$
2. Consider a random walk, $\langle (x^1, f(x^1)), \ldots, (x^T, f(x^T)) \rangle$.
3. For every, i, such that $f(x^i) \neq f(x^{i+1})$, if j is the variable such that $x^i_j \neq x^{i+1}_j$, add j to $\mathcal{J}$.

Learning f

1. Consider each of the $|\mathcal{A}|^{|\mathcal{J}|}$ possible assignments $b_{\mathcal{J}} \to \mathcal{A}$. We will construct a truth table for a function $h : \mathcal{A}^{\mathcal{J}} \to \mathcal{Y}$.
2. For a fixed $b_{\mathcal{J}}$, let $h(b_{\mathcal{J}})$ be the plurality label among the x^i in the random walk above for which $x^i_j = b_{\mathcal{J}}(j)$ for all $j \in \mathcal{J}$.

Output: Hypothesis h

Figure 3: Algorithm: Exact Learning k-juntas

and only if there exist $x, x' \in \mathcal{A}^n$ such that $\pi(x) \neq 0$, $\pi(x') \neq 0$, x, x' differ only at co-ordinate i and $f(x) \neq f(x')$. Otherwise, i can have no influence in determining the value of f (under the distribution π).

We claim that Algorithm 3 identifies every variable in the junta S of f. Let $b_S : S \to \mathcal{A}$, be any assignment of values to variables in S. Since S is the *junta* for f, any $x \in X$ that satisfies $x_i = b_S(i)$ for all $i \in S$, has the same value $f(x)$. By slight abuse of notation, we denote this common value by $f(b_S)$.

The fact that $i \in S$ implies that there exist assignments, b^1_S, b^2_S, such that $b^1_S(i) \neq b^2_S(i)$, $\forall j \in S$, such that $j \neq i$, $b^1_S(j) = b^2_S(j)$ and which satisfy the following: $\pi(x_S = b^1_S) \neq 0$, $\pi(x_S, b^2_S) \neq 0$. Consider the following event: x is drawn from π, x' is the state after exactly one transition, x satisfies the event $x_S = b^1_S$ and x' satisfies the event $x'_S = b^2_S$. By our assumptions, the probability of this event is at least $\beta/(c|\mathcal{A}|)^{|S|}$. Let $\alpha = \beta/(c|\mathcal{A}|)^{|S|}$. Then, if we draw x from the distribution $P^t(x_0, \cdot)$ for $t = \tau_M \ln(2/\alpha)$, instead of the *true* stationary distribution π, the probability of the above event is still at least $\alpha/2$. This is because when $t = \tau_M \ln(2/\alpha)$, the $\|P^t(x_0, \cdot) - \pi\|_{\mathrm{TV}} \leq \alpha/2$. Thus, by observing a long enough random walk, *i.e.* one with $2\tau_M \ln(1/\alpha) \log(k/\delta)/\alpha$ transitions, except with probability δ/k, the variable i will be identified as a member of the junta. Since there are at most k such variables, by a union bound all of S will be identified. Once the set S has been identified, the unknown function can be learned *exactly* by observing an example of each possible assignments to the variables in S. The above argument shows that all such assignments with non-zero measure under π already exist in the observed random walk. □

Remark 1. *We observe that the condition that the MC be* rapidly mixing *alone is sufficient to identify at least one variable of the junta. However, unlike in the case of learning from i.i.d. examples, in this learning model, identifying one variable of the junta is not equivalent to learning*

the unknown junta function. In fact, it is quite easy to construct rapidly mixing Markov chains where the influence *of some variables on the target function can be hidden, by making sure that the transitions that cause the function to change value happen only on a subset of the variables of the junta.*

We now show that the Ising model and graph coloring satisfy the conditions of Theorem 4 as long as the underlying graphs have constant degree.

Ising Model: Recall that the state space is $X = \{-1, 1\}^n$. Let $\beta(\Delta)$ be the inverse critical temperature, which is a constant independent of n as long as Δ, the maximal degree, is constant. Let $S \subseteq [n]$ and let $b_S^1 : S \to \{-1, 1\}$ and $b_S^2 : S \to \{-1, 1\}$ be two distinct assignments to variables in S. Let σ^1, σ^2 be two configurations of the Ising system such that for all $i \in S$, $\sigma_i^1 = b_S^1(i)$, $\sigma_i^2 = b_S^2(i)$ and for $i \notin S$, $\sigma_i^1 = \sigma_i^2$. Let $d^1 = \sum_{(i,j)\in E:\sigma_i^1 \neq \sigma_j^1} \beta_{ij}$ and $d^2 = \sum_{(i,j)\in E:\sigma_i^2 \neq \sigma_j^2} \beta_{ij}$. Then, since the maximum degree of the graph Δ is constant and each β_{ij} is also bounded by some constant, $|d^1 - d^2| \leq c|S|\Delta$. Then, by definition (see Section 2), $\exp(-c\beta\Delta|S|) \leq \pi(\sigma^1)/\pi(\sigma^2) \leq \exp(c\beta\Delta|S|)$. By summing over possible pairs σ^1, σ^2 that satisfy the constraints, we have $\exp(-\beta\Delta|S|) \leq \pi(x_S = b_S^1)/\pi(x_S = b_S^2) \leq \exp(\beta\Delta|S|)$. But, since there are only $2^{|S|}$ possible assignments of variables in S, the first assumption of Theorem 4 follows immediately. The second assumption follows from the definition of the transition rate matrix, *i.e.* each non-zero entry in the transition rate matrix is at least $\exp(-\beta\Delta)/2n$.

Graph Coloring: Let q be the number of colors. The state space is $[q]^n$ and *invalid* colorings have 0 mass under the stationary distribution. We assume that $q \geq 3\Delta$, where Δ is the maximum degree in the graph. This is also the assumption that ensures rapid mixing. Let $S \subseteq [n]$ be an subset of nodes. Let C_S^1 and C_S^2 be two assignments of colors to the nodes in S. Let D_1 and D_2 be the set of valid colorings such that for each $x \in D_1$, $i \in S$, $x_i = C_S^1(i)$ and for each $x \in D_2$, $i \in S$, $x_i = C_S^2(i)$. We define a map from D_1 to D_2 as follows:

1. Starting from $x \in D_1$, first for all $i \in S$, set $x_i = C_S^2(i)$. This may in fact result in an *invalid* coloring.

2. The invalid coloring is switched to a valid coloring by only modifying neighbors of nodes in S. The condition that $q \geq 3\Delta$ ensures that this can always be done.

The above map has the following properties. Let $N(S) = \{j \mid (i, j) \in E, i \in S\}$. Then, the nodes that are not in $S \cup N(S)$ do not change the color. Thus, even though the map may be a many to one map, at most $q^{|S|+|N(S)|}$ elements in D_1 may be mapped to a single element in D_2. Note that $|S| + |N(S)| \leq (\Delta + 1)|S|$. Thus, we have $\pi(D_1)/\pi(D_2) = |D_1|/|D_2| \leq q^{(\Delta+1)|S|}$. This implies the first condition of Theorem 4. The second condition follows from the definition of the transition matrix, each non-zero entry is at least $1/(2qn)$.

References

David Aldous and Umesh Vazirani. A Markovian extension of Valiant's learning model. *Inf. Comput.*, 117(2):181–186, 1995.

Peter L. Bartlett, Paul Fischer, and Klaus-Uwe Höffgen. Exploiting random walks for learning. In *Proceedings of the seventh annual conference on Computational learning theory*, COLT '94, pages 318–327, 1994.

Avrim Blum. Learning boolean functions in an infinite attribute space. *Mach. Learn.*, 9(4):373–386, 1992.

Guy Bresler. Efficiently learning Ising models on arbitrary graphs. *arXiv preprint arXiv:1411.6156*, 2014.

Nader H. Bshouty, Elchanan Mossel, Ryan O'Donnel, and Rocco A. Servedio. Learning DNF from random walks. *J. Comput. Syst. Sci.*, 71(3):250–265, Oct 2005.

Dana Dachman-Soled, Homin Lee, Tal Malkin, Rocco Servedio, Andrew Wan, and Hoeteck Wee. Optimal cryptographic hardness of learning monotone functions. In *ICALP '08: Proceedings of the 35th international colloquium on Automata, Languages and Programming, Part I*, pages 36–47, 2008.

R. L. Dobrushin and S. B. Shlosman. Constructive criterion for uniqueness of a Gibbs field. In J. Fritz, A. Jaffe, and D. Szasz, editors, *Statistical Mechanics and dynamical systems*, volume 10, pages 347–370. 1985.

David Gamarnik. Extension of the PAC framework to finite and countable Markov chains. In *Proceedings of the twelfth annual conference on Computational learning theory*, COLT '99, pages 308–317, 1999.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. ISSN 0890-5401.

Jeffrey C. Jackson and Karl Wimmer. New results for random walk learning. *Journal of Machine Learning Research (JMLR)*, 15:3635–3666, November 2014.

Mark Jerrum. A very simple algorithm for estimating the number of k-colorings of a low-degree graph. *Random Structures and Algorithms*, 7(2):157–165, 1995.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, 2008.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. In *FOCS*, pages 11–20, 2005.

Michael Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. In *Machine Learning*, pages 341–352, 1994.

Michael J. Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory*. The MIT Press, 1994.

Michael Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 372–381, 1993.

Ross Kinderman and J. Laurie Snell. *Markov Random Fields and Their Applications*. AMS, 1980.

Adam R. Klivans, Ryan O'Donnell, and Rocoo A. Servedio. Learning intersections and thresholds of halfspaces. In *FOCS*, 2002.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

Elchanan Mossel and Allan Sly. Exact thresholds for Ising-Gibbs samplers on general graphs. *The Annals of Probability*, 41(1):294–328, 2013.

Elchanan Mossel, Ryan O'Donnell, and Rocco A. Servedio. Learning functions of k relevant variables. *J. Comput. Syst. Sci.*, 69(3):421–434, 2004.

Moni Naor and Omer Reingold. Number-theoretic constructions of efficient pseudo-random functions. *Journal of the ACM (JACM)*, 51(2):231–262, 2004.

Ryan O'Donnell and Rocco A. Servedio. Learning monotone decision trees in polynomial time. *SIAM Journal on Computing*, 37(3):827–844, 2007.

Amir Shpilka and Avishay Tal. On the minimal Fourier degree of symmetric boolean functions. In *IEEE Conference on Computational Complexity*, pages 200–209, 2011.

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, Nov 1984.

E. Vigoda. Improved bounds for sampling coloring. In *40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 51–59, 1999.

A Proofs from Section 4

A.1 Proof of Theorem 1

Proof. We divide the spectrum of P into blocks. Let k and $i_1, \ldots, i_k$ be as in Definition 1; furthermore define $i_0 = 0$ for notational convenience. For $j = 1, \ldots, k$, let $S_j = \{i_{j-1} + 1, \ldots, i_j\}$. Throughout this proof we use the letter ℓ to index eigenvectors of P—so ν_ℓ is an eigenvector with eigenvalue λ_ℓ. We want to find $\beta^\ell_{t,m}$ in order to (approximately) represent the eigenvector ν_ℓ as

$$\nu_\ell = \sum_{t,m} \beta^\ell_{t,m} P^t g_m + \eta_\ell \tag{3}$$

Also, we use the notation,

$$\bar{\nu}_\ell = \sum_{t,m} \beta^\ell_{t,m} P^t g_m, \tag{4}$$

We will show that such representations exist block by block. To begin define

$$\epsilon_1 = \left(\frac{\epsilon}{(2\alpha N)^{\frac{1+c}{c}} (Nk)^{\frac{1}{2c}}} \right)^{(1+c)^{k-1}} \tag{5}$$

and define ϵ_j according to the following recurrence,

$$\epsilon_j = 2\alpha N (Nk)^{\frac{1}{2(1+c)}} \epsilon_{j-1}^{\frac{1}{1+c}} \tag{6}$$

It is an easy calculation to verify that the solution for ϵ_j is given by

$$\epsilon_j = \left(2\alpha N (Nk)^{\frac{1}{2(1+c)}} \right)^{\frac{1+c}{c}\left(1 - \frac{1}{(1+c)^{j-1}}\right)} \epsilon_1^{\frac{1}{(1+c)^{j-1}}} \tag{7}$$

Also, define

$$B_1 = (N\alpha)^{c+1} \epsilon_1^{-c} \tag{8}$$

and let B_j be defined according the following recurrence:

$$B_j = 2\alpha N (Nk)^{\frac{1}{2(1+c)}} (\epsilon_{j-1})^{-\frac{c}{1+c}} B_{j-1} \tag{9}$$

It is an easy calculation to verify that the solution for B_j is given by

$$B_j = \left(2N\alpha (Nk)^{\frac{1}{2(1+c)}} \right)^{j-1} \cdot \left(\prod_{j'=1}^{j-1} \epsilon_{j'} \right)^{-\frac{c}{1+c}} B_1 \tag{10}$$

It can be verified that ϵ_j and B_j are increasing as a function of j as long as all ϵ_j remain smaller than 1 (which can be verified by checking that $\epsilon_k < 1$). We show by induction on j that

for any $\ell \in S_j$, $\sum_{t,m} |\beta_{t,m}^{\ell}| \leq B_j$ and $\|\eta_\ell\|_2 \leq \epsilon_j$ (recall that the norm here is with respect to the distribution π).

Consider some j and suppose that $|S_j| = N_j$. Denote by $S_{<j} = \bigcup_{j'<j} S_{j'}$, all the indices that precede those in S_j and $S_{>j} = \{\ell' \mid \ell' > i_j\}$. According to Definition 2, there exist $g_1, \ldots, g_{N_j} \in \mathcal{G}$, such that if A is the $N_j \times N_j$ matrix given by $a_{m,\ell} = \langle g_m, \nu_\ell \rangle$ for $\ell \in S_j$ and $1 \leq m \leq N_j$, then $\|A^{-1}\|_{op} \leq \alpha$. Let $\bar{a}_{\ell,m}$ denote the element in position (l, m) in A^{-1} and let $\mathcal{G}_j = \{g_1, \ldots, g_{N_j}\}$ be these specific N_j functions in $\mathcal{G}$. Also, observe that by Definition 1, $N_j \leq N$.

Let $g_m \in \mathcal{G}_j$ and for any ℓ', let $a_{m,\ell'} = \langle g_m, \nu_{\ell'} \rangle$. Then, define

$$\tilde{g}_m = g_m - \sum_{\ell' \in S_{<j}} a_{m,\ell'} \bar{\nu}_{\ell'} \tag{11}$$

Thus, $\tilde{g}_m$ is obtained from g_m by (approximately) removing contributions of eigenvectors corresponding to blocks that precede the j^{th} block. Thus, we may write $\tilde{g}_m$ as follows:

$$\begin{aligned}
\tilde{g}_m &= \sum_{\ell \in S_j} a_{m,\ell} \nu_\ell + \sum_{\ell' \in S_{<j}} a_{m,\ell'} (\nu_{\ell'} - \bar{\nu}_{\ell'}) + \sum_{\ell' \in S_{>j}} a_{m,\ell'} \nu_{\ell'} \\
&= \sum_{\ell \in S_j} a_{m,\ell} \nu_\ell + \sum_{\ell' \in S_{<j}} a_{m,\ell'} \eta_{\ell'} + \sum_{\ell' \in S_{>j}} a_{m,\ell'} \nu_{\ell'}
\end{aligned}$$

To further simplify the above equation, define $v_m^{<} = \sum_{l' \in S_{<j}} a_{m,\ell'} \eta_{\ell'}$ and $v_m^{>} = \sum_{l' \in S_{>j}} a_{m,\ell'} \nu_{\ell'}$. Then, we have

$$\tilde{g}_m = \sum_{\ell \in S_j} a_{\ell,m} \nu_\ell + v_m^{<} + v_m^{>} \tag{12}$$

In the case of $v_m^{<}$, a crude bound can be established on its norm $\|v_m^{<}\|_2$ as follows: for any $\ell' \in S_{<j}$, $\|\eta_{\ell'}\|_2 \leq \epsilon_{j-1}$ (induction hypothesis). Using the facts that $\sum_{\ell'} (a_{m,\ell'})^2 \leq 1$, and that $|S_{<j}| \leq N(j-1) \leq Nk$, by applying the Cauchy-Schwarz inequality we get $\|v_m^{<}\|_2 \leq \epsilon_{j-1}\sqrt{Nk}$.

For $v_m^{>}$, we note that $\|P v_m^{>}\|_2 \leq \lambda_{i_j+1} \|v_m^{>}\|_2$, since it only contains components corresponding to eigenvectors with eigenvalues at most λ_{i_j+1}. Also, note that $\|v_m^{>}\|_2^2 \leq \sum_{\ell' \in S_{>j}} (a_{m,\ell'})^2 \leq 1$.

We now complete the proof by induction. For, $j' = 1, \ldots, j-1$, suppose that all the eigenvectors corresponding to indices in $S_{j'}$ have representations of the form in Equation (3) with parameters $B_{j'}$ and $\epsilon_{j'}$ respectively. Recall that $\bar{a}_{\ell,m}$ is the element in position (ℓ, m) of A^{-1}, where A is the matrix defined as $a_{m,\ell} = \langle g_m, \nu_\ell \rangle$ for $g_m \in \mathcal{G}_j$ and $\ell \in S_j$. Now for any $\ell \in S_j$, we can define $\bar{\nu}_\ell$ as follows (for the value τ_j to be specified later):

$$\bar{\nu}_\ell = \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} \tilde{g}_m \tag{13}$$

Using Equation (12) in the above equation, we get

$$\begin{aligned}\bar{\nu}_\ell &= \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} \sum_{\ell' \in S_j} a_{m,\ell'} P^{\tau_j} \nu_{\ell'} + \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} P^{\tau_j} v_m^< + \lambda_\ell^{-\tau} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} v_m^> \\ &= \lambda_\ell^{-\tau_j} \sum_{\ell' \in S_j} \lambda_{\ell'}^{\tau_j} \nu_{\ell'} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} a_{m,\ell'} + \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} v_m^< + \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} v_m^>\end{aligned}$$

In the first term, we use the fact that $\sum_m \bar{a}_{\ell,m} a_{m,\ell'} = \delta_{\ell,\ell'}$ by definition. Thus, the first term reduces to ν_ℓ. We apply the triangle inequality to get

$$\begin{aligned}\|\eta_\ell\|_2 = \|\nu_\ell - \bar{\nu}_\ell\|_2 &\leq \lambda_\ell^{-\tau_j} \left\| \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} v_m^< \right\|_2 + \lambda_\ell^{-\tau_j} \left\| \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} v_m^> \right\|_2 \\ &\leq \lambda_\ell^{-\tau_j} \sqrt{\sum_{m=1}^{N_j} (\bar{a}_{\ell,m})^2} \cdot \sqrt{\sum_{m=1}^{N_j} \|P^{\tau_j} v_m^<\|_2^2} + \lambda^{-\tau_j} \sqrt{\sum_{m=1}^{N_j} (\bar{a}_{\ell,m})^2} \cdot \sqrt{\sum_{m=1}^{N_j} \|P^{\tau_j} v_m^>\|_2^2}\end{aligned} \tag{14}$$

We use the fact that $\sqrt{\sum_{i=1}^{m} (\bar{a}_{\ell,m})^2} \leq \|A^{-1}\|_F$ and that $N_j \leq N$, $\|v_m^<\|_2^2 \leq Nk(\epsilon_{j-1})^2$ to simplify the above expression. Furthermore, since P has largest eigenvalue 1, $\|P^{\tau_j} v\|_2 \leq \|v\|_2$ for any v. In the case of $v_m^>$, since the $\|v_m^>\|_2 \leq 1$ and the largest eigenvalue in it is λ_{i_j+1}, $\|P^{\tau_j} v_m^<\|_2 \leq \lambda_{i_j+1}^{\tau_j}$. Putting all these together and simplifying the above expression we get

$$\|\eta_\ell\|_2 \leq \|A^{-1}\|_F \sqrt{N} \left(\lambda_\ell^{-\tau_j} \epsilon_{j-1} \sqrt{Nk} + \left(\frac{\lambda_{i_j+1}}{\lambda_\ell} \right)^{\tau_j} \right)$$

Finally, using the fact that $\lambda_\ell \geq \lambda_{i_j}$ (since $\ell \in S_j$), we have that $\lambda_{i_j+1}/\lambda_\ell \leq \gamma$ and that $1/\lambda_\ell \leq \gamma^{-c}$. We also use the fact that $\|A^{-1}\|_F \leq \sqrt{N}\|A^{-1}\|_{op} \leq \sqrt{N}\alpha$. Thus, we get

$$\|\eta_\ell\|_2 \leq \alpha N \left(\gamma^{-c\tau_j} \epsilon_{j-1} \sqrt{Nk} + \gamma^{\tau_j} \right) \tag{15}$$

At this point we will deal with the base case $j = 1$ separately. In Equation (12) when $g_m \in \mathcal{G}_1$, $v_m^< = 0$, since the set $S_{<1}$ is empty. Thus, in Equation (14), the first term is absent if we are dealing with the case when $\ell \in S_1$, since all the $v_m^<$ in this case are 0. Thus, for $\ell \in S_1$, Equation (15) reduces to:

$$\|\eta_\ell\|_2 \leq \alpha N \gamma^{\tau_1} \tag{16}$$

Thus, by choosing $\tau_1 = -\frac{\ln(N\alpha/\epsilon_1)}{\ln(\gamma)}$, we get that for all $\ell \in S_1$, $\|\eta_\ell\|_2 \leq \epsilon_1$. Now, for $j > 1$, we can find τ_j that minimizes the RHS of Equation (15) and this is given by $\tau_j = \frac{1}{1+c} \frac{\ln(\epsilon_{j-1}\sqrt{Nk})}{\ln(\gamma)}$. It is not hard to calculate that in this case the RHS of Equation 15 exactly evaluates to ϵ_j.

We now prove a bound on B_j. Again, we look at the base case separately, when $j = 1$, $S_{<j} = \emptyset$ and so for the functions $g_m \in \mathcal{G}_1$ as in Equation (11), $\tilde{g}_m = g_m$. Thus, for $\ell \in S_1$, by looking at Equation (13), we can define: $\beta^{\ell}_{\tau_1,m} = \lambda_\ell^{-\tau_1} \bar{a}_{\ell,m}$ for $m \in \mathcal{G}_1$ and the remaining $\beta^{\ell}_{t,m}$ values are set to 0. Thus,

$$\sum_{t,m} |\beta^{\ell}_{t,m}| \leq \lambda_\ell^{-\tau_1} \sum_{m=1}^{N_1} |\bar{a}_{\ell,m}| \leq \gamma^{-c\tau_1} N\alpha \tag{17}$$

Above we used the fact that $\lambda_\ell \geq \lambda_{i_k} \geq \gamma^c$ and that $|\bar{a}_{\ell,m}| \leq \|A^{-1}\|_{op}$. But, the RHS above is exactly the quantity B_1 we defined earlier.

Next, we consider the case of $j > 1$ and we start from Equation (13).

$$\begin{aligned}
\bar{\nu}_\ell &= \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} \tilde{g}_m \\
&= \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} \left(g_m - \sum_{\ell' \in S_{<j}} a_{m,\ell'} \bar{\nu}_{\ell'} \right) \\
&= \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} P^{\tau_j} \left(g_m - \sum_{\ell' \in S_{<j}} a_{m,\ell'} \sum_{t,m'} \beta^{\ell'}_{t,m'} P^t g_{m'} \right) \\
&= \lambda_\ell^{-\tau_j} \sum_{m=1}^{N_j} \bar{a}_{\ell,m} \left(P^{\tau_j} g_m - \sum_{\ell' \in S_{<j}} a_{m,\ell'} \sum_{t,m'} \beta^{\ell'}_{t,m'} P^{t+\tau_j} g_{m'} \right)
\end{aligned} \tag{18}$$

If the above, expression is re-written to be of the form,

$$\bar{\nu}_\ell = \sum_{t,m} \beta^{\ell}_{t,m} P^t g_m,$$

we can get a bound on $\sum_{t,m} |\beta^{\ell}_{t,m}|$ as follows:

$$\sum_{t,m} |\beta^{\ell}_{t,m}| \leq \gamma^{-c\tau_j} \left(\sum_{m=1}^{N_j} |\bar{a}_{\ell,m}| \right) \cdot \left(1 + B_{j-1} \sum_{\ell' \in S_{<j}} |a_{m,\ell'}| \right)$$

Above, we use the fact that for $\ell' \in S_{<j}$, $\sum_{t,m} |\beta^{\ell'}_{t,m}| \leq B_{j-1}$. Also, note that $\sum_{m=1}^{N_j} |\bar{a}_{\ell,m}| \leq N\alpha$ and $\sum_{\ell' \in S_{<j}} |a_{m,\ell'}| \leq \sqrt{Nk}$ (since $\sum_{\ell'} (a_{m,\ell'})^2 \leq 1$ for all m), so we have

$$\begin{aligned}
\sum_{t,m} |\beta^{\ell}_{t,m}| &\leq (\epsilon_{j-1}\sqrt{Nk})^{-\frac{c}{1+c}} N\alpha (1 + \sqrt{Nk} B_{j-1}) \\
&\leq 2\sqrt{Nk} N\alpha B_{j-1} (\epsilon_{j-1}\sqrt{Nk})^{-\frac{c}{1+c}}
\end{aligned}$$

We observe that the expression on the RHS above is exactly the value B_j given by the recurrence relation in Equation (9).

Finally, by observing the RHS of Equation (13) we notice that the maximum power t, for which $\beta^{\ell}_{t,m}$ is non-zero for any ℓ, m is $\sum_{i=1}^{k} \tau_i$. Thus, the proof is complete by setting $\tau_{\max} = \sum_{j=1}^{k} \tau_j$. □

A.2 Proof of Theorem 2

Proof. Let $f \in F$ be the target function and for any ℓ, let $\hat{f}_\ell = \langle f, \nu_\ell \rangle$ denote the *Fourier* coefficients of f. Then the condition in Theorem 2 states that $\sum_{\ell > \ell^*(\epsilon)} \hat{f}_\ell^2 \leq \epsilon^2/4$.

First, we appeal to Theorem 1. In the rest of this proof, we assume that for all $\ell \leq \ell^*$, there exist $\beta_{t,m}^\ell$ such that

$$\nu_\ell = \sum_{t,m} \beta_{t,m}^\ell P^t g_m + \eta_\ell,$$

where $g_m \in \mathcal{G}$, $\|\eta_\ell\|_2 \leq \epsilon_1$. Furthermore, let B and $\tau_{\max}$ be as given by the statement of the theorem.

We first look closely at $P^t g_m$, since P is an $|X| \times |X|$ matrix and $g_m : X \to \mathbb{R}$ a function, $P^t g_m$ is also a function from $X \to \mathbb{R}$. For $x \in X$, let $\mathbb{1}_x$ denote the indicator function of the point x (it may be viewed as a vector that is 0 everywhere, except in position x where it has value 1). Then, we have

$$(P^t g_m)(x) = \mathbb{1}_x^T P^t g_m = \mathbb{E}_{y \sim P^t(x,\cdot)}[g_m(y)] \tag{19}$$

Notice that the quantity on the RHS above can be estimated by sampling. Thus, with black-box access to the oracle $\mathsf{OS}(\cdot)$ and g_m, we can estimate $(P^t g_m)(x)$. This is exactly what is done in (1) in the algorithm in Figure 2. Also, since $\|g\|_\infty \leq 1$, it is also the case that $\|P^t g_m\|_\infty \leq 1$. Thus, by a standard Chernoff-Hoeffding bound, if we set the input parameter $T = \log(\tau_{\max} \cdot |X| \cdot |\mathcal{G}|/\delta)/\epsilon_2^2$, with probability at least $1 - \delta$, it holds for every $x \in X$, for every $t < \tau_{\max}$ and every $g_m \in \mathcal{G}$, that $|\phi_{t,m}(x) - (P^t g_m)(x)| \leq \epsilon_2$. For the rest of this proof, we will treat the functions $\phi_{t,m}(x)$ as deterministic (rather than randomized) for simplicity. (This can be easily arranged by taking a sufficiently long random string used to simulate the Markov chain and treating it as advice.)

Now, consider the following:

$$\begin{aligned}
&\mathbb{E}\left[\left(f(x) - \sum_{\ell \leq \ell^*} \hat{f}_\ell \sum_{t,m} \beta_{t,m}^\ell \phi_{t,m}(x)\right)^2\right] \\
&\quad \leq 2\mathbb{E}\left[\left(f(x) - \sum_{\ell \leq \ell^*} \hat{f}_\ell \nu_\ell(x)\right)^2\right] + 2\mathbb{E}\left[\left(\sum_\ell \hat{f}_\ell \left(\nu_\ell(x) - \sum_{t,m} \beta_{t,m}^\ell \phi_{t,m}(x)\right)\right)^2\right]
\end{aligned} \tag{20}$$

Note that the first term above is at most ϵ. We will now bound the second term. (Below $\bar{\nu}_\ell$ is as defined in Equation (13).)

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_\ell \hat{f}_\ell \left(\nu_\ell(x) - \sum_{t,m} \beta_{t,m}^\ell \phi_{t,m}(x)\right)\right)^2\right] \\
&\quad \leq 2\mathbb{E}\left[\left(\sum_{\ell \leq \ell^*} \hat{f}_\ell(\nu(x) - \bar{\nu}(x))\right)^2\right] + 2\mathbb{E}\left[\left(\sum_\ell \hat{f}_\ell \left(\sum_{t,m} \beta_{t,m}^\ell ((P^t g_m)(x) - \phi_{t,m}(x))\right)\right)^2\right] \\
&\quad \leq 2\sqrt{\sum_{\ell \leq \ell^*} (\hat{f}_\ell)^2} \cdot \sqrt{\sum_{\ell \leq \ell^*} \|\eta_\ell\|_2^2} + 2\sqrt{\sum_{\ell \leq \ell^*} (\hat{f}_\ell)^2} \cdot \sqrt{\sum_{\ell < \ell^*} \left\|\sum_{t,m} \beta_{t,m}^\ell (P^t g_m - \phi_{t,m})\right\|_2^2}
\end{aligned}$$

Next we use the following facts, $\sum_{\ell \leq \ell^*} (\hat{f}_\ell)^2 \leq 1$, $\ell^* \leq Nk$ and $\|\eta_\ell\|_2 \leq \epsilon_1$. Also for the very last term, the fact that $\sum_{t,m} |\beta^\ell_{t,m}| \leq B$ and $\forall x, |(P^t g_m)(x) - \phi_{t,m}(x)| \leq \epsilon_2$, imply that $\| \sum_{t,m} \beta^\ell_{t,m} (P^t g_m - \phi_{t,m}) \|_2 \leq B\epsilon_2$. Putting everything together we get

$$\mathbb{E}\left[\left(\sum_\ell \hat{f}_\ell \left(\nu_\ell(x) - \sum_{t,m} \beta^\ell_{t,m} \phi_{t,m}(x)\right)\right)^2\right] \leq 2(\sqrt{Nk\epsilon_1} + \sqrt{BNk\epsilon_2}) \tag{21}$$

Finally, substituting (21) back into (20), we get

$$\mathbb{E}\left[\left(f(x) - \sum_{\ell \leq \ell^*} \hat{f}_\ell \sum_{t,m} \beta^\ell_{t,m} \phi_{t,m}(x)\right)^2\right] \leq 2(\frac{\epsilon^2}{4} + \sqrt{Nk\epsilon_1} + \sqrt{BNk\epsilon_2}) \tag{22}$$

By choosing $\epsilon_1 = \epsilon^2/(64Nk)$ and $\epsilon_2 = \epsilon^2/(64BNk)$ we get that the quantity is in fact at most ϵ^2. Thus, we get that

$$\mathbb{E}\left[\left|f(x) - \sum_{\ell \leq \ell^*} \hat{f}_\ell \sum_{t,m} \beta^\ell_{t,m} \phi_{t,m}(x)\right|\right] = \epsilon \tag{23}$$

Thus, we have essentially shown that $\{\phi_{t,m}\}$ can be used as a suitable feature space and there is a linear form in this feature space that is a good L_1 approximation to f. This is sufficient for agnostic learning as was shown by Kalai et al. (2005). Note that the sum of coefficients on the features is bounded by $B\sqrt{Nk}$ (since B is a bound on $\sum_{t,m} |\beta^\ell_{t,m}|$ and $\sum_{\ell < \ell^*} |\hat{f}_\ell| \leq \sqrt{Nk}$). Thus, in the algorithm in Figure 2, we may set the parameters $\tau_{\max}$ (as given by Theorem 1), $W = B\sqrt{Nk}$ and $T = \log(\tau_{\max} \cdot |X| \cdot |\mathcal{G}|/\delta)/\epsilon_2^2$. The sample complexity is polynomial in W, $1/\epsilon$ as follows from standard generalization bounds (see for example Kakade et al. (2008)) and the running time of the algorithm is polynomial in $|\mathcal{G}|, T, \tau_{\max}, W, \frac{1}{\epsilon}$. The bounds given in the statement of the theorem follow from observing the values of the above quantity in the statement of Theorem 1. □

A.3 Proof of Theorem 3

Proof. The proof follows the standard proofs of these types of results. Let $t = -\frac{1}{\ln(\rho)}$

$$\begin{aligned}
\mathrm{NS}_t(f) &= \frac{1}{2} - \frac{1}{2}\langle f, P^t f \rangle \\
&= \frac{1}{2} - \frac{1}{2}\left(\sum_\ell \lambda_\ell^t \hat{f}_\ell^2\right) \\
&\geq \frac{1}{2} - \frac{1}{2}\left(\sum_{\ell \leq \ell^*} \hat{f}_\ell^2 + \sum_{\ell > \ell^*} \rho^t \hat{f}_\ell^2\right)
\end{aligned}$$

Using the fact that $\sum_{\ell \leq \ell^*} \hat{f}_\ell^2 = 1 - \sum_{\ell > \ell^*} \hat{f}_\ell^2$ (since f is boolean) and rearranging terms, we get

$$\sum_{\ell > \ell^*} \hat{f}_\ell^2 \leq \frac{1}{1 - \rho^t} \mathrm{NS}_t(f) \tag{24}$$

Then substituting the value for t completes the proofs. □

A.4 Proof of Proposition 1

Lemma 1. *For any positive integer Δ, there exists $\beta(\Delta)$, such that for all graphs G of maximum degree bounded by Δ, and for all ferromagnetic Ising models with $\beta < \beta(\Delta)$, the following holds. If $f = \mathrm{sign}(\sum_i w_i x_i)$, then for all $t \geq n$ it holds that,*

$$1 - 2\,\mathrm{NS}_t(f) \geq \delta^{t/n}$$

for some fixed $\delta > 0$ depending only on Δ and $\beta(\Delta)$.

Note that the above lemma only proves that majorities are somewhat noise stable. While one expects that if t is a very small fraction on n, majorities are very noise stable, our proof is not strong enough to prove that.

For the proof we will need to use the following well known result which goes back to Dobrushin and Shlosman (1985). The proof also follows easily from the random cluster representation of the Ising model.

Lemma 2. *For every Δ and $\eta > 0$, there exists a $\beta(\Delta, \eta) > 0$ such that for all graphs G of maximum degree bounded by Δ and for all Ising models where $\beta \leq \beta(\Delta, \eta)$, it holds that under the stationary measure for any i and any subset S of nodes,*

$$\mathbb{E}[x_i \mid x_S] \leq \eta^{d(i,S)}$$

In particular, for every i and j,

$$\mathbb{E}[x_i x_j] \leq \eta^{d(i,j)},$$

where $d(i,j)$ denotes the graph distance between i and j.

We will need a few corollaries of the above lemma.

Lemma 3. *If $\beta < \beta(\Delta, 1/(10\Delta))$, then for every set A and any weights w_i, it holds that if $f(x) = \sum_i w_i x_i$, then:*

1. $\frac{4}{5}\sum_i w_i^2 \leq \mathbb{E}[f(x)^2] \leq \frac{6}{5}\sum_i w_i^2$
2. $\mathbb{E}[(f(x))^4] \leq 10\left(\sum_i w_i^2\right)^2$

Proof. For the first claim note that

$$\begin{aligned}\mathbb{E}[f(x)^2] &= \sum_i w_i^2 + \sum_{i \neq j} w_i w_j \mathbb{E}[x_i x_j] \\ &= \sum_i w_i^2 + \sum_{d=1}^{n} \sum_{i,j:d(i,j)=d} w_i w_j \mathbb{E}[x_i x_j]\end{aligned}$$

Choose $\eta = 1/(10\Delta)$ in Lemma 2, and suppose that $\beta < \beta(\Delta, 1/(10\Delta))$, then we have that for all $i \neq j$,

$$\mathbb{E}[x_i x_j] \leq (10\Delta)^{-d(i,j)}.$$

We may thus bound,

$$\left| \sum_{i,j:d(i,j)=d} w_i w_j \mathbb{E}[x_i x_j] \right| \leq (10\Delta)^{-d} \sum_{i,j:d(i,j)=d} |w_i w_j|$$

For each i, let $v_1^i, \ldots, v_{\Delta^d}^i$ be all the nodes that are at distance d from i, where if the actual number of such nodes is less than Δ^d, we set the remaining $v_j^i = i$. Then, by applying the Cauchy Schwarz inequality, we can write:

$$\sum_{i,j:d(i,j)=d} |w_i w_j| \leq \sum_i \sum_{j=1}^{\Delta^d} |w_i w_{v_j^i}| \leq \Delta^d \sum_i w_i^2$$

So, adding up over all d, we obtain,

$$|\mathbb{E}[f(x)^2] - \sum_i w_i^2| \leq \sum_i w_i^2 \sum_{d=1}^{n} 10^{-d} \leq \frac{1}{5} \sum_i w_i^2$$

This completes the proof of the first part of the lemma.

The second part is proved analogously, however, the calculations are a bit more involved since it involves terms corresponding to four nodes at a time.

□

We can now complete the proof of Lemma 1.

Proof of Lemma 1. From the Fourier expression of noise-sensitivity (see Eq. 2) and Jensen's inequality, it is clear that if $a > 1$, then

$$1 - 2\,\mathrm{NS}_{at}(f) \geq (1 - 2\,\mathrm{NS}_t(f))^a$$

Therefore it suffices to prove the claim when $t = cn$ for some small constant c (which may depend on Δ). Our goal is therefore to show that:

$$1 - 2\,\mathrm{NS}_{cn}(f) \geq \delta > 0$$

where δ is a parameter that depends only on Δ (but not n). To prove this let $X_1, \ldots, X_n$ be the system at time 0 and let $Y_1, \ldots, Y_n$ be the system at time $t = cn$. Let $A \subset [n]$ be the random subset of spins that have not been updated from time 0 to time t. Then, the noise sensitivity is:

$$\begin{aligned}
\mathrm{NS}_\delta(f) &= \Pr\left[\mathrm{sign}(\sum_{i \in A} w_i X_i + \sum_{i \notin A} w_i X_i) \neq \mathrm{sign}(\sum_{i \in A} w_i X_i + \sum_{i \notin A} w_i Y_i) \right] \\
&\leq 2 \Pr\left[\mathrm{sign}(\sum_{i \in A} w_i X_i) \neq \mathrm{sign}(\sum_{i=1}^{n} w_i X_i) \right],
\end{aligned}$$

where the last inequality uses the fact that X_i, $i \notin A$ and Y_i, $i \notin A$ are identically distributed given A and X_i, $i \in A$ (the distribution for both is just the conditional distribution given x_i for $i \in A$).

Let $W = \sum_i w_i^2$. By Markov's inequality, it follows that for c chosen small enough with probability at least $9/10$ (over the random choice of A), we have:

$$\sum_{i \notin A} w_i^2 \leq 10^{-6} \cdot W$$

From now on, we will condition on the event that $\sum_{i \in A} w_i^2 \geq (1 - 10^{-6})W$, which we denote by $\mathcal{E}$. Under this conditioning, from Lemma 3, it follows that

$$\mathbb{E}\left[\left(\sum_{i \in A} w_i X_i\right)^2\right] \geq \frac{3}{5} W \tag{25}$$

Moreover, we claim that with probability at least $1/40$ (conditioned on the event above), it holds that:

$$\left(\sum_{i \in A} w_i X_i\right)^2 \geq \frac{W}{10}$$

Let ρ be the (conditioned on $\mathcal{E}$) probability of the above event, which we denote by $\mathcal{E}'$. Note that (25) implies that:

$$\mathbb{E}\left[\left(\sum_{i \in A} w_i X_i\right)^2 \mid \mathcal{E}'\right] \geq \frac{W}{2\rho}$$

But, then we use part two of Lemma 3 to conclude that $\rho \geq 1/40$; if not, we can derive a contradiction as follows.

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{i \in A} w_i X_i\right)^4\right] &\geq \mathbb{E}\left[\left(\sum_{i \in A} w_i X_i\right)^4 \mid \mathcal{E}'\right] \cdot \rho \\
&\geq \mathbb{E}\left[\left(\sum_{i \in A} w_i X_i\right)^2 \mid \mathcal{E}'\right]^2 \cdot \rho > 10 W^2
\end{aligned}$$

Also, conditioned on the event $\mathcal{E}$, by Markov's Inequality, we have:

$$\begin{aligned}
\Pr\left[\left(\sum_{i \notin A} w_i X_i\right)^2 \geq \tfrac{W}{100}\right] &\leq 10^{-4} \\
\Pr\left[\left(\sum_{i \notin A} w_i Y_i\right)^2 \geq \tfrac{W}{100}\right] &\leq 10^{-4}
\end{aligned}$$

Thus, conditioned on $\mathcal{E}$, by a union bound, we have that with probability at least $3/4$:

$$\operatorname{sign}\left(\sum_{i=1}^{n} w_i X_i\right) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i Y_i\right) = \operatorname{sign}\left(\sum_{i \in A} w_i X_i\right)$$

To conclude the proof, we show that when $\mathcal{E}$ does not hold, the probability that

$$\operatorname{sign}\left(\sum_{i=1}^{n} w_i X_i\right) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i Y_i\right)$$

is at least $1/2$. In fact, we show this conditioned on any A and any values of the random variables $X_i, i \in A$. Note that conditioned on A and $X_i \in A$, the random variables X_i and Y_i for $i \notin A$ are positively correlated. (Also, $(X_i)_{i \notin A}$ and $(Y_i)_{i \notin A}$ are identically distributed.) Thus, if we denote by

$$p_A = \Pr\left[\operatorname{sign}\left(\sum_{i=1}^{n} w_i X_i\right) \neq \operatorname{sign}\left(\sum_{i \in A} w_i X_i\right)\right] = \Pr\left[\operatorname{sign}\left(\sum_{i=1}^{n} w_i Y_i\right) \neq \operatorname{sign}\left(\sum_{i \in A} w_i X_i\right)\right]$$

Then, using the FKG inequality, we see that conditioned on the event $\mathcal{E}$ not occurring,

$$\Pr\left[\operatorname{sign}\left(\sum_{i=1}^{n} w_i X_i\right) \neq \operatorname{sign}\left(\sum_{i=1}^{n} w_i Y_i\right)\right] \leq 2p_A \cdot (1 - p_A) \leq \frac{1}{2}$$

This concludes the proof.

□