Computationally sufficient statistics for Ising models

Abhijith Jayakumar[*], **Shreya Shukla**, **Marc Vuffray**, **Andrey Y. Lokhov**, **Sidhant Misra**

Theoretical Division, Los Alamos National Laboratory, USA

Abstract

Learning Gibbs distributions using only sufficient statistics has long been recognized as a computationally hard problem. On the other hand, computationally efficient algorithms for learning Gibbs distributions rely on access to full sample configurations generated from the model. For many systems of interest that arise in physical contexts, expecting a full sample to be observed is not practical, and hence it is important to look for computationally efficient methods that solve the learning problem with access to only a limited set of statistics. We examine the trade-offs between the power of computation and observation within this scenario, employing the Ising model as a paradigmatic example. We show that the couplings of the model can be learned to a desired accuracy from statistics whose order grows linearly with the ℓ_1 width of the model and only logarithmically with the inverse of the accuracy. In particular, for a fixed accuracy, the required order is proportional to the width. This allows us to infer the model's structure, using statistics whose order grows logarithmically with the inverse of the weakest coupling, and also to learn its magnetic fields. When the width and the accuracy are fixed, the running time is polynomial in the number of variables. We also discuss a setting where prior information about structure of the model is available and show that the learning problem can be solved efficiently with even more limited observational power.

1 Introduction

Trade-offs between sample and computational complexities of learning tasks has garnered much interest in learning theory ([25, 1]). However, in certain cases, the learner's access to the data is itself constrained by limited observational power, i.e., instead of observing full samples, they may only observe some aggregated statistics or low-order moments. This raises a fundamental question: how does this observational power interact with computational tractability?

We study this question in the context of learning Gibbs distributions over discrete variables, where such a setting naturally arises as an interpolation between two sets of prior works. The first are hardness results that show that learning the parameters of such a distribution given only sufficient statistics is a computationally intractable problem [15, 4]. On the other hand, starting from the breakthrough by Bresler ([3]), provably efficient algorithms that use full independent samples from the model are known for this problem. Several other works have also looked at this learning task from other types of data models like samples drawn from a dynamic ([8]) or from a metastable distribution ([11]). All the efficient algorithms here assume that full sample configurations of the model can be observed. This motivates us to look for an intermediate regime that has been left unexplored by the aforementioned literature. We ask the question:

Is it possible to learn a discrete graphical model from a limited set of statistics, such that the computational effort scales as a polynomial in the number of variables and in the data size?

[*]Email: abhijithj@lanl.gov

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We initiate the study of this question in the Ising class of models. Here the distribution is over p binary spins, $\underline{\sigma} \in \{-1, 1\}^p$. The distribution that we are attempting to learn is a Gibbs distribution associated with a general quadratic function over these variables,

$$\mu(\underline{\sigma}) \propto \exp(E(\underline{\sigma}; \underline{\theta}^*)) \text{ where } E(\underline{\sigma}; \underline{\theta}^*) \coloneqq \sum_{1 \le u < v \le p} \theta^*_{uv} \sigma_u \sigma_v + \sum_{u=1}^{p} \theta^*_u \sigma_u, \tag{1}$$

with the convention that $\theta^*_{uv} = \theta^*_{vu}$, that will be used throughout this paper. Following previous works, our aim here will be to learn the parameters $\underline{\theta}^*$. We will assume that we know an upper bound on the ℓ_1-*width* of the model [13, 9]. That is, we know a γ such that, $\sum_{v \neq u} |\theta^*_{uv}| + |\theta^*_u| \le \gamma, \ \forall u \in [p]$.

We refer to the expectation values of monomials ($\mathbb{E}\sigma_i$, $\mathbb{E}\sigma_i\sigma_j$, $\mathbb{E}\sigma_i\sigma_j\sigma_k, \ldots$) as to the *statistics* of this model. It is well known that the parameters of the model are entirely fixed by the first ($\mathbb{E}\sigma_i$) and second order ($\mathbb{E}\sigma_i\sigma_j$) statistics. However, performing the actual inversion from these statistics to the parameter has been generally proven to be a hard problem ([15, 4]). Efficient estimators which are based on observing full sample configurations, on the other hand, can be implicitly viewed as having complete access to all the statistics of the model (up to order p). In this paper we will consider a setting where the learner is given access to only monomials up to some degree d.

For solving this learning problem, we will use a modification of the *Interaction Screening Estimator* (ISE). The ISE, first introduced in [24], is an M-estimator that works by minimizing a convex loss function for every variable $u \in [p]$ subject to constraints on the ℓ_1-width of the model. Given n samples from the true distribution,

$$\mathcal{L}_{u,n}(\underline{\theta}) \equiv \frac{1}{n} \sum_{t=1}^{n} e^{-\sigma_u^{(t)}(\theta_u + \sum_{v \neq u} \theta_{u,v} \sigma_v^{(t)})}, \tag{2}$$

$$\hat{\underline{\theta}}_u = \underset{\underline{\theta} : |\underline{\theta}| \le \gamma}{\operatorname{argmin}} \mathcal{L}_{u,n}(\theta). \tag{3}$$

This constrained optimization problem is solved for each of the p variables in the model to estimate the parameters in the energy function connected to it. It has been shown to efficiently recover the parameters and structure of Ising models and other discrete graphical models [24, 23]. As shown in these works, this estimator requires exact configurations to be evaluated. However, the exponential function, combined with the ℓ_1 constraint, suggests the possibility of approximating the exponential function with a low-degree polynomial.

This idea is explored briefly in [14] by expanding the interaction screening loss to second-order. At sufficiently low γ this second-order approximation can be expected to approximate the full IS estimator well. But this strategy is bound to fail at higher values of γ. Moreover, expanding the exponential as a polynomial can potentially lead to a loss of convexity, and in turn to the loss of learning guarantees the output of ISE. If we could find a provably effective way to approximate the exponential in the interaction screening method using, say, a degree d polynomial, then that implies that the parameters of the model can be estimated using statistics up to order $O(d)$.

In this work, we propose to approximate the gradient of the IS loss using a polynomial approximation. This allows us to analyze this problem as a convex optimization problem with a corrupted gradient oracle. From the known robustness properties of gradient descent, we show that we can get close the true optimum using this approximate gradient. Our main result (Theorem 1) shows that statistics up to order $O(\gamma + \log(1/\epsilon))$ suffice to learn every coupling to accuracy ϵ. For a fixed accuracy, the required order is therefore proportional to the width γ. Exact recovery of the structure requires accuracy $\alpha/2$, where α is the smallest magnitude of a nonzero coupling, so the structure can be recovered from statistics of order $O(\gamma + \log(1/\alpha))$. The order needed to learn the magnetic fields similarly grows logarithmically with the inverse of the target accuracy (Theorem 3). When all the statistics are estimated from the same n samples, $n = O\big(e^{8\gamma} \operatorname{poly}(\gamma, \log \frac{1}{\epsilon}) \log(p/\delta)/\epsilon^4\big)$ samples suffice (Corollary 1). This has the same scaling with p, γ and ϵ, up to factors polynomial in γ and $\log(1/\epsilon)$, as Interaction Screening with full samples [23]. The cost of acquiring the statistics in other ways, for instance from separate samples for each statistic, is discussed in Section 2.1. The running time of the algorithm is polynomial in the size of the table of statistics, $p^{O(\gamma + \log(1/\epsilon))}$. This is polynomial in p when γ and ϵ are constants independent of p, which is the standard regime in the literature on learning Ising models [13, 24].

Our results give rise to a natural question about corresponding lower bounds, i.e., what is the lowest order of statistics that will allow us to tractably learn the model. Our results point to the existence of a type of information-computation tradeoff in this learning problem. In particular, the power of statistics between information-theoretically sufficient (order 2) and that is proven in this work to be computationally sufficient (order $O(\gamma)$ for a fixed accuracy) is currently not quantified in literature.

1.1 Further Related Work

Learning from statistical queries. The Statistical Query (SQ) model, introduced by Kearns ([12]), formalizes learning scenarios in which the algorithm accesses data only through expectations of bounded functions rather than individual sample. Follow-up works have shown that many learning problems that are information-theoretically solvable become computationally intractable when restricted to SQ access alone ([7, 6]). These results show that the form in which data is observed can play a crucial role in determining computational feasibility. Our setting differs from the classical SQ model in two aspects. First, we consider access to a fixed collection of low-degree moments rather than arbitrary adaptively chosen queries. Second, we allow these moments to be reused across multiple optimization steps. Nonetheless, both settings emphasize how restrictions on observable statistics affect the computational complexity of learning.

Mean-field heuristics. An approach to the inverse Ising problem is to replace the true Gibbs measure with a tractable mean-field approximation, yielding explicit estimators for the parameters of the model purely from empirical statistics ([17]). In the simplest mean field reconstruction, inverting the correlation matrix gives a closed form estimate for the parameters of the models. More accurate variants incorporate linear-response corrections and, in particular, the Onsager reaction term ([20, 22]). These methods are attractive for their simplicity, though they are ultimately uncontrolled approximations which fail at larger γ.

Continuous variables and quantum generalizations. It is natural to consider the same problem in other types of alphabets. For continuous variable models, learning Gibbs distributions can be solved computationally efficiently using score matching, which can be viewed as morally equivalent to Interaction Screening or Pseudo-likelihood methods as it breaks down the global learning problem into single variable problems [19, 10]. However, score matching does not require access to full samples to learn. For a distribution specified by a degree-d polynomial energy function, score matching loss only requires access to statistics up to order $2d$. This shows an important distinction between continuous and discrete alphabet learning scenarios.

The quantum generalization of the Gibbs learning task has been shown to be tractable in recent works [2, 16]. These algorithms work by using a sum-of-squares hierarchy to match certain statistics of the true quantum gibbs state with those of a parametric hypothesis. This implicitly suggests, even in the classical setting, that the learning problem is solvable using a set of statistics strictly smaller than the complete set. However, in the quantum setting the structure of the model is assumed to be known and sparse. In the classical setting, this makes the learning problem manifestly tractable using some lower order moments that depend only on the sparsity. This setting is discussed in Section 2.4. These considerations show that the minimal set of statistics required to solve such a problem will also depend strongly on the prior information we have about the true model.

Notation: We use underlines (e.g., $\underline{\sigma}, \underline{\theta}$) to denote vectors. For $p \in \mathbb{N}$ we write $[p] = \{1, \dots, p\}$. The Ising configuration is $\underline{\sigma} \in \{-1, 1\}^p$ with coordinates σ_u for $u \in [p]$. Parameters are $\underline{\theta}^* = (\{\theta^*_{u,v}\}_{u \neq v}, \{\theta^*_u\}_u)$, with local parameter vector at node u denoted $\underline{\theta}_u = (\theta_u, \{\theta_{u,v}\}_{v \neq u})$. Expectations under the true model are denoted $\mathbb{E}[\cdot]$; empirical average from n i.i.d. samples are $\mathbb{E}_n[\cdot]$.

2 Interaction screening with approximate gradients

This is section we will give the main algorithm and its performance guarantees. We split the learning task into three parts:

1. Parameter learning of the second order terms ($\theta^*_{u,v}$) (Theorem 1)
2. Structure learning, i.e., finding the graph structure encoded in the $\theta^*_{u,v}$ terms (Theorem 2)

3. Parameter learning of linear terms (θ^*_u) (Theorem 3).

We show that these tasks can be solved sequentially given access to data in terms of empirical estimates of the statistics of the model up to $O(\gamma)$.

We denote by $E_u(\underline{\sigma}, \underline{\theta}_u) \coloneqq \sigma_u(\theta_u + \sum_{j \neq u} \theta_{uj}\sigma_j)$, a parametric "local" energy function. This, in effect, determines the parametrization of the single variable conditionals of the model.

2.1 Moment table and observation models

The learner never sees the model μ directly. Its only input is a table of (approximate) low-order moments. For a moment order $D \in \mathbb{N}$, define the index set $\mathcal{K}_D \coloneqq \{K \subseteq [p] : |K| \leq D\}$. For each index $K \in \mathcal{K}_D$, let $m_K \coloneqq \mathbb{E}_\mu[\prod_{i \in K} \sigma_i]$ denote the true moment (with $m_\emptyset = 1$). Note that the size of the index set is, $|\mathcal{K}_D| = \sum_{j \leq D} \binom{p}{j} \leq (ep)^D$ for $D \geq 1$.

Definition 1 (Approximate moment table)**.** A random table $S = \{S[K]\}_{K \in \mathcal{K}_D}$ is a (D, τ, δ)-*approximate moment table* of μ if

$$\mathbb{P}\Big[\max_{K \in \mathcal{K}_D} \big|S[K] - m_K\big| \leq \tau\Big] \;\geq\; 1 - \delta.$$

The probability is over the randomness of the data-collection procedure that produces S.

The guarantee is *simultaneous* over all entries of the table, and no independence between different entries is assumed. The algorithms in this paper take such a table as their only input; once the event in Definition 1 holds, the rest of the analysis is deterministic. The accuracy and confidence parameters (τ, δ) therefore encode everything the learner needs to know about how the table was obtained. Two natural ways of producing a table, which are the two extremes of a larger family, are:

(a) **Shared dataset.** Draw n i.i.d. samples $\underline{\sigma}^{(1)}, \ldots, \underline{\sigma}^{(n)} \sim \mu$ and set $S_{D,n}[K] \coloneqq \mathbb{E}_n[\prod_{i \in K} \sigma_i] = \frac{1}{n}\sum_{t=1}^{n} \prod_{i \in K} \sigma_i^{(t)}$ for every $K \in \mathcal{K}_D$.

(b) **Separate datasets.** For every $K \in \mathcal{K}_D$ draw a fresh set of n i.i.d. samples and estimate only m_K from it.

Hybrids (different groups of moments estimated from different, possibly overlapping, datasets, or with K-dependent sample sizes) are covered in the same way. We write $\mathbb{E}_n$ for an empirical average over n samples, and $S_{D,n}$ for a table whose entries are n-sample empirical moments, in either of the two senses above.

Lemma 1 (Empirical tables are approximate moment tables)**.** *Let $S_{D,n}$ be built as in (a) or in (b). If*

$$n \;\geq\; \frac{2}{\tau^2} \log\Big(\frac{2|\mathcal{K}_D|}{\delta}\Big),$$

which holds in particular if $n \geq \frac{2}{\tau^2}\big(D\log(ep) + \log\frac{2}{\delta}\big)$, then $S_{D,n}$ is a (D, τ, δ)-approximate moment table.

Proof. Fix $K \in \mathcal{K}_D$. Each entry is an average of n i.i.d. random variables $\prod_{i \in K} \sigma_i^{(t)} \in \{-1, 1\}$ with mean m_K, so Hoeffding's inequality gives $\mathbb{P}[|S_{D,n}[K] - m_K| > \tau] \leq 2e^{-n\tau^2/2}$. A union bound over the $|\mathcal{K}_D|$ entries (no independence between entries is needed) bounds the failure probability by $2|\mathcal{K}_D|e^{-n\tau^2/2} \leq \delta$. ∎

Cost of acquiring the table. The number of samples *per moment* is the same, n, in (a) and (b), but the total measurement cost is not. In (a) each of the n samples must be revealed in enough detail to evaluate every monomial of degree $\leq D$, which in general means observing the sample in full. In (b) each observation is only needed on the $|K| \leq D$ spins of one monomial, but the total number of observations is $n\,|\mathcal{K}_D| = n\,p^{O(D)}$. Our results are stated for a generic approximate moment table so that the cost of acquiring it can be accounted for separately in each such model.

To estimate the $\theta^*_{u,v}$ parameters we use the projected gradient descent method with approximate gradients as outlined in Algorithm 1. The performance guarantees on the output of this algorithm is stated and proven in the Theorem below.

Algorithm 1: Interaction Screening with Moment-Based Approximate Gradients

Input: Degree d, approximate moment table S of order $D = d + 2 - (d \bmod 2)$ (Definition 1), ℓ_1 bound γ, Step size η, Iterations T

Output: Estimated parameters $\hat{\underline{\theta}}_u$ for every $u \in [p]$

for $u = 1$ ***to*** p **do**
  Initialize $\underline{\theta}_u^{(1)} = \{\theta_u^{(1)}, \theta_{u,1}^{(1)}, \ldots, \theta_{u,p}^{(1)}\} \leftarrow 0$
  for $t = 1$ ***to*** T **do**
    `// Gradient Approximation`
    Construct $\tilde{\nabla}\mathcal{L}_{u,S}(\underline{\theta}_u^{(t)})$ from S as follows:
    (i) Replace $e^{-E_u(\underline{\sigma}, \underline{\theta}_u^{(t)})}$ in the exact gradient by the degree-d approximation as in (6).
    (ii) Expand the resulting moments using (7).
    (iii) Substitute each monomial expectation $\mathbb{E}[\prod_{i \in K} \sigma_i]$ by the table entry $S[K]$.
    `// Gradient Step and` ℓ_1 `projection`
    $\underline{\theta}_u^{(t+1)} \leftarrow \mathcal{P}_{\|\cdot\|_1 \le \gamma}(\, \underline{\theta}_u^{(t)} - \eta \tilde{\nabla}\mathcal{L}_{u,S}(\underline{\theta}_u^{(t)})\,)$
  end
  `// Averaging`
  $\hat{\underline{\theta}}_u \leftarrow \frac{1}{T}\sum_{t=1}^{T} \underline{\theta}_u^{(t)}$
end
return $\hat{u}_\theta$

Theorem 1. *Suppose we want to achieve error $\epsilon \le 4\gamma$ on the $\theta^*_{u,v}$ parameters in the following metric,*

$$\max_{u \in [p], v \neq u} |\hat{\theta}_{u,v} - \theta^*_{u,v}| \le \epsilon.$$

Let $d = \left\lceil \frac{3\gamma + \log(\frac{16\gamma(1+\gamma)}{\epsilon^2})}{W(\frac{3}{e} + \frac{\log(16\gamma(1+\gamma)/\epsilon^2)}{\gamma e})} \right\rceil - 1$, $D = d + 2 - (d \bmod 2)$, *and*

$$\tau = \frac{\epsilon^2 e^{-4\gamma}}{16\,\gamma(1+\gamma)}.$$

Then Algorithm 1, with parameters $T = \left\lceil \frac{576\; p\gamma^2 e^{8\gamma}(1+\gamma)^2}{\epsilon^4} \right\rceil$ *and* $\eta = \frac{2\gamma e^{-\gamma}}{3\sqrt{pT}}$, *run on a* (D, τ, δ)*-approximate moment table S (Definition 1), returns with probability at least* $1 - \delta$ *an estimate satisfying the guarantee above simultaneously for every* $u \in [p]$. *The probability is over the randomness of S only.*

The W in the expression for d is the Lambert-W function in the principal branch. From the $\epsilon \le 4\gamma$, it can be seen that the argument of this function in the expression for d always remains greater than $3/e$. Since W is an increasing function in the positive real axis, the necessary moment order for this to work scales as $d = O(\gamma)$ for a fixed ϵ. More generally, since the argument of W exceeds $3/e$, the denominator is at least $W(3/e) > 0.6$, and $\log(16\gamma(1+\gamma)) \le \log 16 + 2\gamma$. Hence

$$d \;\le\; \frac{5\gamma + 2\log(1/\epsilon) + \log 16}{W(3/e)} \;=\; O\big(\gamma + \log(1/\epsilon)\big),$$

so the order grows linearly with the width and only logarithmically with the inverse of the target accuracy. Numerical experiments in Appendix H confirm this linear scaling on an example, and show that the degree needed in practice is several times smaller than the one prescribed by the theorem.

The theorem is stated for a generic approximate moment table, so the number of samples needed depends on how the table is produced. Combining it with Lemma 1 gives the following.

Corollary 1. *Let d, D, T, η be as in Theorem 1 and let $S_{D,n}$ be an empirical table built either from a shared dataset or from separate datasets (Section 2.1), with*

$$n \;\ge\; 2^9\, e^{8\gamma}\gamma^2(1+\gamma)^2\, \frac{(d+2)\log(ep) + \log(\frac{2}{\delta})}{\epsilon^4}.$$

Then, with probability at least $1-\delta$*, Algorithm 1 run on* $S_{D,n}$ *returns* $\hat{\theta}$ *with* $\max_{u,v\neq u} |\hat{\theta}_{u,v} - \theta^*_{u,v}| \leq \epsilon$.

Proof. With τ as in Theorem 1, $\frac{2}{\tau^2} = \frac{2^9 e^{8\gamma}\gamma^2(1+\gamma)^2}{\epsilon^4}$, and $D \leq d+2$. Lemma 1 then shows that $S_{D,n}$ is a (D,τ,δ)-approximate moment table, and Theorem 1 applies. ■

The number of samples n used for each moment has the same asymptotic scaling, $O(e^{8\gamma}\mathrm{poly}(\gamma)\frac{\log p}{\epsilon^4})$ for fixed ϵ, as Interaction Screening with full samples in [23]. The total measurement cost of the table, however, depends on how it is acquired (see the discussion in Section 2.1): it is n full-sample observations in the shared-dataset model and $n\,p^{O(\gamma)}$ restricted observations in the separate-datasets model. The total number of steps the algorithm takes is $pT = O(p^2\gamma^4 e^{8\gamma}/\epsilon^4)$, with the dominant cost of each step being $O(|S|)$ time taken to process the moment table given to the algorithm to construct approximate gradients, where $|S| = |\mathcal{K}_D| \leq (ep)^D$. The running time is therefore polynomial in the size of the input table. Since $D = O(\gamma)$ for a fixed target accuracy ϵ, it is polynomial in p, namely $p^{O(\gamma)}$, whenever the ℓ_1 width γ is a constant independent of p. Bounded width is a standard assumption in the literature on learning Ising models [13, 9, 24].

Proof of Theorem 1. The main idea of the proof is as follows. Consider the ideal interaction screening loss function given by

$$\mathcal{L}_u(\theta) = \mathbb{E}\left[e^{-E_u(\underline{\sigma},\underline{\theta}_u)}\right]. \tag{4}$$

For each u, the ideal IS loss function in (4) is convex and is known to have its minimum at the true model parameters $\underline{\theta}^*$, see [24]. Gradient descent for minimizing convex functions has the crucial property of *robustness*, i.e. bounded errors incurred in the gradient computation translate favorably in a controlled way to errors in the final loss function [5] (also see Lemma 4). Using this fact and by bounding the error incurred by computing the approximate gradients as in Algorithm 1 we can prove Theorem 1.

The error in the approximate gradient computation is from two sources. The first is a *polynomial approximation error* (ε_{poly}) incurred by approximating the exponential by a finite degree polynomial. The second is a *statistical approximation error* (ε_{stat}) representing the error from replacing the exact moments by the entries of the approximate moment table S.

We first describe how to obtain the approximate gradients using only moments up to order d from the data. The exact gradients of the loss function $\mathcal{L}_u$ are,

$$\frac{\partial\mathcal{L}_u(\underline{\theta}_u)}{\partial\theta_u} = -\mathbb{E}\left[\sigma_u e^{-E_u(\underline{\sigma},\underline{\theta}_u)}\right];\ \ \frac{\partial\mathcal{L}_u(\underline{\theta}_u)}{\partial\theta_{uv}} = -\mathbb{E}\left[\sigma_u\sigma_v e^{-E_u(\underline{\sigma},\underline{\theta}_u)}\right],\ \ u\neq v, \tag{5}$$

Going forward, we will focus on $\frac{\partial\mathcal{L}_u(\underline{\theta}_u)}{\partial\theta_{uv}}$ to demonstrate how the polynomial approximation works. The approximation to $\frac{\partial\mathcal{L}_u(\underline{\theta}_u)}{\partial\theta_u}$ is mechanically the same.

We begin by approximating the exponential function using a polynomial approximation

$$\mathbb{E}\left[\sigma_u\sigma_v e^{-E_u(\underline{\sigma},\underline{\theta}_u)}\right] = \sum_{k\leq d}\frac{(-1)^k\mathbb{E}[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j+\theta_u)^k]}{k!} + \varepsilon_{poly} \tag{6}$$

Here we have used ε_{poly} to denote the truncation error. Later in the proof we will set d by tuning this quantity via Lemma 3.

Now we will expand the term in the expectation value using the multinomial expansion. For this purpose define,

$$\mathcal{M}_k \equiv \{(m_1,\ldots,m_p)\in\mathbb{N}^p | \sum_i m_i = k\} \qquad \text{and} \qquad C_{\underline{\theta},\underline{m}} \equiv \frac{k!\ \theta_u^{m_u}\prod_{j\neq u}\theta_{uj}^{m_j}}{\prod_{j\in[p]} m_j!}.$$

Using these we can show that,

$$\mathbb{E}[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j+\theta_u)^k] = \sum_{\underline{m}\in\mathcal{M}_k} C_{\underline{\theta},\underline{m}}\,\mathbb{E}\left[\sigma_v\sigma_u^{k+1}\prod_{j\neq u}\sigma_j^{m_j}\right] \tag{7}$$

This expansion allows us to estimate the expectation value of a random variable that depends on the optimization parameters by using parameter independent statistics computed from data. The largest value that k can take is d. From (7), we see that an estimate for $\mathbb{E}[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j+\theta_u)^k]$ for every $k\leq d$ can be constructed entirely using entries of the table S of order $D=d+2-(d \bmod 2)$, which are precisely the entries used to compute the approximate gradient in Algorithm 1.

$$\widetilde{\nabla}_S\mathcal{L}(\underline{\theta})=\sum_{k\leq d}\frac{(-1)^k\widetilde{\mathbb{E}}_S[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j+\theta_u)^k]}{k!} \tag{8}$$

where $\widetilde{\mathbb{E}}_S[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j+\theta_u)^k]$ is the expression on the right-hand side of (7) with each monomial expectation $\mathbb{E}[\sigma_v\sigma_u^{k+1}\prod_{j\neq u}\sigma_j^{m_j}]$ replaced by the table entry $S[K]$, where $K\subseteq[p]$ is the set of indices appearing with odd total exponent in the monomial (recall $\sigma_i^2=1$). As alluded to earlier, the error in this approximate gradient is given by the sum of the polynomial and the statistical approximation errors.

$$\|\nabla\mathcal{L}(\underline{\theta})-\widetilde{\nabla}_S\mathcal{L}(\underline{\theta})\|_\infty\leq\varepsilon_{poly}+\varepsilon_{stat}, \tag{9}$$

where ε_{poly} is defined by (6) and

$$\varepsilon_{stat}=\sum_{k\leq d}\frac{(-1)^k\widetilde{\mathbb{E}}_S[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j+\theta_u)^k]}{k!}-\sum_{k\leq d}\frac{(-1)^k\mathbb{E}[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j+\theta_u)^k]}{k!}. \tag{10}$$

The statistical error is bounded using appropriate concentration equalities and is given in the following lemma. The proof can be found in Appendix A.2

Lemma 2 (Statistical approximation error)**.** *Let S be a (D,τ,δ)-approximate moment table with $D=d+2-(d \bmod 2)$ and $\tau=\varepsilon e^{-\gamma}/(8\gamma)$. Then, with probability at least $1-\delta$, the statistical approximation error satisfies $\varepsilon_{stat}\leq\varepsilon/(8\gamma)$ simultaneously for every $u\in[p]$, every coordinate of the gradient, and every $\underline{\theta}_u$ in the ℓ_1 ball of radius γ.*

The next lemma bounds the polynomial approximation error. Proof of this can be found in Appendix A.1

Lemma 3 (Polynomial approximation error)**.** *Let $p_d(x)\equiv\sum_{k=0}^d\frac{(-x)^k}{k!}$. If the degree d satisfies*

$$d\geq\left\lceil\frac{C}{W\left(\frac{C}{\gamma e}\right)}\right\rceil-1, \tag{11}$$

with $C=\log(8\gamma/\varepsilon)$, then we have $\varepsilon_{poly}<\frac{\varepsilon}{8\gamma}$.

Now we will use the known robustness property of gradient descent to bound how close we can get to the true model by using this approximate gradient. This is proven in Appendix B for completeness.

Lemma 4 (Robustness of gradient descent)**.** *Let $\mathcal{L}:C\to\mathbb{R}$ be a convex function defined over the ℓ_1 ball $C=\{x\mid\ \|x\|_1\leq\gamma\}$ with minimizer $x^*\in\arg\min_{x\in C}\mathcal{L}(x)$. Suppose we have access to an approximate gradient oracle $\widetilde{\nabla}\mathcal{L}(x)$ for $x\in C$ with error bounded as*

$$\sup_{x\in C}\|\widetilde{\nabla}\mathcal{L}(x)-\nabla\mathcal{L}(x)\|_\infty\leq\varepsilon/4\gamma.$$

Let $\sup_{x\in C}\|\widetilde{\nabla}\mathcal{L}(x)\|_2\leq L$. Now, consider the projected gradient method

$$x^{t+1}=\mathcal{P}_C\big(x^t-\eta\widetilde{\nabla}\mathcal{L}(x^t)\big)$$

starting at $x^1\in C$ and with fixed step size $\eta=2\gamma/(L\sqrt{T})$. Then after $T=\left\lceil\frac{16\gamma^2L^2}{\varepsilon^2}\right\rceil$ iterations, the average $\bar{x}^T=\frac{1}{T}\sum_{t=1}^T x^t$ satisfies

$$\mathcal{L}(\bar{x}^T)-\mathcal{L}(x^*)\leq\varepsilon.$$

To apply this lemma we need to bound the norm of the approximate gradient.

Lemma 5 (Bounded gradient)**.** *If $\varepsilon \leq 16\gamma^2 e^{3\gamma}$, then $\|\widetilde{\nabla}\mathcal{L}(\underline{\theta})\|_2 \leq 3\sqrt{p}e^{\gamma}$.*

Proof. For all $\underline{\theta}$ in the ℓ_1 ball or radius γ

$$\begin{aligned}\|\widetilde{\nabla}\mathcal{L}(\underline{\theta})\|_2 &\leq \sqrt{p}\|\widetilde{\nabla}\mathcal{L}(\underline{\theta})\|_\infty \leq \sqrt{p}(\|\nabla\mathcal{L}(\underline{\theta})\|_\infty + \frac{\varepsilon}{4\gamma}),\\ &\leq \sqrt{p}(e^{\gamma} + \frac{\varepsilon}{4\gamma}) \leq 3\sqrt{p}e^{\gamma}, \end{aligned} \tag{12}$$

where the last inequality follows from the bound on ε in the statement of the lemma. ■

Substituting this in T and the learning rate in Lemma 4 gives us for $T = \frac{144p\gamma^2 e^{2\gamma}}{\varepsilon^2}$ an estimate $\widehat{\underline{\theta}}$ such that,

$$\mathcal{L}(\widehat{\underline{\theta}}) - \mathcal{L}(\underline{\theta}^*) \leq \varepsilon. \tag{13}$$

To obtain the above bound we have used the bounds derived above on ε_{poly} and ε_{stat}. Lemma 2 provides the error guarantee on the event that the table S is τ-accurate, which has probability at least $1-\delta$. On this event the bound holds for all iterations t and all nodes u simultaneously, since the table is fixed before the optimization starts and the rest of the argument is deterministic. In particular, no union bound over the T steps or the p nodes is needed.

Now we need to estimate how the error in the loss function propagates to errors in the estimated parameters. To this end, we define a curvature function around the true optimum as follows

$$\delta\mathcal{L}^*(\hat{\underline{\theta}}) \equiv \mathcal{L}(\hat{\underline{\theta}}) - \mathcal{L}(\underline{\theta}^*) - \langle\nabla\mathcal{L}(\underline{\theta}^*), \hat{\underline{\theta}} - \underline{\theta}^*\rangle \tag{14}$$

Since the gradient vanishes at $\underline{\theta}^*$, we have that, $\delta\mathcal{L}^*(\hat{\underline{\theta}}) = \mathcal{L}(\hat{\underline{\theta}}) - \mathcal{L}(\underline{\theta}^*) \leq \varepsilon$. Hence the curvature function is also upper bounded by ε. Independently we can also lower bound the curvature function using a projected infinity norm of the error vector. We state the main result below, the proof is given in Appendix C.

Lemma 6. *The curvature term is lower-bounded as $\delta\mathcal{L}^*(\hat{\underline{\theta}}_u) \geq \frac{e^{-3\gamma}}{2+2\gamma}\max_{v\neq u}|\theta^*_{u,v} - \hat{\theta}_{u,v}|^2$*

Putting these together, we get the result,

$$\max_{v\neq u}|\theta^*_{u,v} - \hat{\theta}_{u,v}| \leq e^{3\gamma/2}\sqrt{2(1+\gamma)(\mathcal{L}(\hat{\underline{\theta}}) - \mathcal{L}(\underline{\theta}^*))} \leq e^{3\gamma/2}\sqrt{2(1+\gamma)\varepsilon}$$

Now let us define a error parameter, $\epsilon \equiv \sqrt{\varepsilon}e^{3\gamma/2}\sqrt{2(1+\gamma)}$, we can guarantee the statement in the theorem with a table accuracy $\tau = \varepsilon e^{-\gamma}/(8\gamma) = \frac{\epsilon^2 e^{-4\gamma}}{16\gamma(1+\gamma)}$, as stated in Theorem 1.

■

2.2 Structure learning

In the Ising learning task, recovering the *structure* (i.e. the underlying interaction graph) is important because it determines which pairwise couplings are nonzero and hence which statistical dependencies are present. In many standard formulations, structure recovery is itself a primary goal: once the neighborhood of each node is identified, one can restrict attention to a much smaller set of parameters and thereby reduce both statistical and computational complexity. This perspective is central to several provably efficient approaches that learn Ising models from full samples, including interaction screening and related neighborhood-selection methods [3, 24, 23].

In our setting, there is a more direct motivation, moment-based gradient approximations require combining a large number of low-degree monomial statistics, and the number of relevant monomials can be reduced significantly if we only keep those involving variables in the estimated neighborhood of u. Thus, structure learning provides a way to focus the subsequent estimation steps (including magnetic field learning) on the true local interactions, avoiding unnecessary dependence on p in intermediate quantities.

Define the edge set, $E = \{(u,v) \in [p]\times[p] \mid \theta^*_{u,v} \neq 0\}$. Additionally we denote by α a known lower bound on smallest non-zero pair-wise parameter in the model, i.e. $0 < \alpha \leq \min_{(u,v)\in E}|\theta^*_{u,v}|$.

Given an estimate $\hat{\theta}$ for the pair-wise parameters, define the thresholded edge set

$$\hat{E} \equiv \{(u,v) \in [p] \times [p] \ : \ |\hat{\theta}_{u,v}| \geq \alpha/2\}. \tag{15}$$

Theorem 2 (Structure recovery by thresholding)**.** *Assume the true model satisfies a separation condition* $\min_{(u,v)\in E} |\theta^*_{u,v}| \geq \alpha$ *for some known* $\alpha > 0$*. If an estimator* $\hat{\theta}$ *satisfies* $\max_{u\in[p],v\neq u} |\hat{\theta}_{u,v} - \theta^*_{u,v}| \leq \alpha/2$*, then the thresholded edge set* (15) *exactly recovers the true structure, i.e.* $\hat{E} = E$*.*

In particular, if Algorithm 1 is run with the parameters of Theorem 1 for $\epsilon = \alpha/2$ *on a* (D, τ, δ)*-approximate moment table, with* D *and* τ *as in Theorem 1, then thresholding its output recovers* E *with probability at least* $1 - \delta$*.*

Since the guarantee of Theorem 1 holds simultaneously for every ordered pair (u, v), the two estimates $\hat{\theta}_{u,v}$ and $\hat{\theta}_{v,u}$ of the same coupling fall on the same side of the threshold, so $\hat{E}$ is symmetric and defines an undirected graph. Note also that $\alpha \leq \gamma$, so $\epsilon = \alpha/2$ satisfies the condition $\epsilon \leq 4\gamma$ of Theorem 1. This is the same standard reduction from parameter learning to structure learning that appears in prior work on interaction screening and related methods (e.g. [24]).

2.3 Learning magnetic fields

We call the coefficients of the linear terms in (1) *magnetic fields*. Even if they are present in the true model, Theorem 1 does not give a guarantee on their error. This is fundamentally due to the linear term getting swamped by higher order terms in the curvature lower bound (for instance, this happens at (32) in the proof of Lemma 7). To get around this issue, we use a re-optimization strategy.

For this we assume that the underlying graphical model has a maximum degree of D. After running Algorithm 1, we can learn the structure of the model with high probability and fix the second order coefficients. Then we solve a single-variable optimization problem for each variable in the model to estimate the magnetic fields. This can also be done with access to only $O(\gamma)$ statistics. The resulting procedure, Algorithm 2 in Appendix F, is Algorithm 1 with three changes: for each u, the couplings $\theta_{u,v}$ are fixed to their estimates $\hat{\theta}_{u,v}$; projected gradient descent runs only over the field $\theta_u \in [-\gamma, \gamma]$; and since only the derivative with respect to the field is needed, the approximate gradient is built in the same three steps from a table of order $d + 1 - (d \bmod 2)$ instead of $d + 2 - (d \bmod 2)$.

Theorem 3 (Magnetic field learning)**.** *Suppose that we have coupling estimates* $\hat{\theta}_{u,v}$ *such that* $|\hat{\theta}_{u,v} - \theta^*_{u,v}| \leq \epsilon$ *for all* $(u, v) \in E$*. Fix a target accuracy* $\epsilon_h > 0$*, and let*

$$\varepsilon \coloneqq \frac{\epsilon_h^2 e^{-\gamma}}{2(1+\gamma)}, \qquad L \coloneqq (1 + D\epsilon)e^{2\gamma} + \frac{\varepsilon}{4\gamma}, \qquad T = \left\lceil \frac{16\gamma^2 L^2}{\varepsilon^2} \right\rceil, \qquad \eta = \frac{2\gamma}{L\sqrt{T}},$$

$$d = \left\lceil \frac{\gamma + \log \frac{8\gamma(1+\gamma)}{\epsilon_h^2}}{W\left(\frac{1}{2e} + \frac{1}{2\gamma e} \log \frac{8\gamma(1+\gamma)}{\epsilon_h^2}\right)} \right\rceil - 1, \text{ and}$$

$$\tau_h \coloneqq \frac{\epsilon_h^2 \, e^{-3\gamma}}{8\gamma(1+\gamma)}.$$

Then Algorithm 2, run with these parameters on a $\big(d + 1 - (d \bmod 2),\ \tau_h,\ \delta\big)$*-approximate moment table, returns with probability at least* $1 - \delta$ *estimates* $\hat{\theta}_u$ *satisfying, simultaneously for every* $u \in [p]$*,*

$$|\hat{\theta}_u - \theta^*_u| \leq \epsilon_h + \sqrt{8\gamma(1+\gamma)e^{3\gamma} D\epsilon}. \tag{16}$$

In particular, by Lemma 1, an empirical table $S_{d+1-(d \bmod 2),n}$ *built from a shared dataset or from separate datasets suffices if*

$$n \geq 2^7 \, \gamma^2 (1+\gamma)^2 e^{6\gamma} \, \frac{(d+1)\log(ep) + \log\left(\frac{2}{\delta}\right)}{\epsilon_h^4}.$$

A key difference compared to Theorem 1 is that we do not have access to the true couplings $\{\theta^*_{u,v}\}$, and instead must plug in estimates $\{\hat{\theta}_{u,v}\}$. This introduces an additional source of error in our

gradient oracle, beyond the polynomial truncation and statistical estimation errors. The sparsity of the graph controls how this errors transfers to the gradient estimate.

To quantify this effect, fix $u \in [p]$ and consider these two single-variable loss functions

$$\mathcal{R}(\theta) \equiv \mathbb{E}\, e^{-\left(\theta\sigma_u + \sigma_u \sum_{j\in N_u} \hat{\theta}_{u,j}\sigma_j\right)},\ \mathcal{R}^*(\theta) \equiv \mathbb{E}\, e^{-\left(\theta\sigma_u + \sigma_u \sum_{j\in N_u} \theta^*_{u,j}\sigma_j\right)}. \tag{17}$$

Since $\mathcal{R}^*$ is the right restriction of the complete IS loss function in 4, it is clear that $\theta^*_u \in \operatorname{argmin}_{|\theta|\le\gamma} \mathcal{R}^*(\theta)$. Their derivatives satisfy

$$\left|\frac{d\mathcal{R}^*(\theta)}{d\theta} - \frac{d\mathcal{R}(\theta)}{d\theta}\right| \le \mathbb{E}\left|e^{-\left(\theta\sigma_u + \sigma_u \sum_{j\in N_u} \theta^*_{u,j}\sigma_j\right)} - e^{-\left(\theta\sigma_u + \sigma_u \sum_{j\in N_u} \hat{\theta}_{u,j}\sigma_j\right)}\right| \tag{18}$$

$$\le e^{\gamma}\, \mathbb{E}\left|e^{-\sigma_u \sum_{j\in N_u} \theta^*_{u,j}\sigma_j} - e^{-\sigma_u \sum_{j\in N_u} \hat{\theta}_{u,j}\sigma_j}\right| \tag{19}$$

$$\le e^{2\gamma}\mathbb{E}\left|\sigma_u \sum_{j\in N_u} \sigma_j(\theta^*_{uj} - \hat{\theta}_{uj})\right| \le e^{2\gamma} D\epsilon, \tag{20}$$

where the third line uses the basic Lipschitz continuity of the exponential function, $|e^a - e^b| \le |a-b|\max(e^a, e^b)$. The rest of the proof follows that of Theorem 1, by considering an approximate optimization strategy for $\mathcal{R}^*$. The rest of the proof can be found in Appendix F.

2.4 Learning under stronger structural priors

So far we have looked at cases where the prior information on the model is only a bound on its ℓ_1 width. It is also interesting to consider cases where even stronger information about the model is known. The main motivation for this comes from studying Gibbs distributions arising in statistical physics and quantum field theory ([2, 21]). There it is often known that the energy function is defined on a $D-$degree graph representing nearest-neighbor spatial points.

Suppose that the structure of the graph (in terms of the edge set E) is known exactly and it is D-regular. Then the neighborhood, $|N_u| = D$ for every u, and the interaction screening loss at node u depends only on the spins in $\{u\} \cup N_u$:

$$\mathcal{L}_u(\underline{\theta}_u) \equiv \mathbb{E}\Big[e^{-\sigma_u\left(\theta_u + \sum_{v\in N_u} \theta_{u,v}\sigma_v\right)}\Big]. \tag{21}$$

Define the following function on the $(D+1)$ spins

$$f_u(\underline{\sigma}_{\{u\}\cup N_u}) \;\coloneqq\; e^{-\sigma_u\left(\theta_u + \sum_{v\in N_u} \theta_{u,v}\sigma_v\right)}, \qquad \underline{\sigma}_{\{u\}\cup N_u} \in \{-1,1\}^{D+1}. \tag{22}$$

Functions $\{-1,1\}^{D+1}$ admits a multilinear Fourier expansion; $f_u(\underline{\sigma}_{\{u\}\cup N_u}) = \sum_{K\subseteq\{u\}\cup N_u} \widehat{f}(\underline{\theta})_{u,K} \prod_{i\in K}\sigma_i$.

Plugging this into the definition of $\mathcal{L}_u$ yields, $\mathcal{L}_u(\underline{\theta}_u) = \sum_{K\subseteq\{u\}\cup N_u} \widehat{f}(\underline{\theta})_{u,K}\, \mathbb{E}\Big[\prod_{i\in K}\sigma_i\Big]$. Hence, evaluating (and differentiating) $\mathcal{L}_u$ only requires the moments m_K for the *local* index sets $K \subseteq \{u\}\cup N_u$, which have order $|K| \le D+1$. There are at most $p\, 2^{D+1}$ such sets, compared to $|\mathcal{K}_{D+1}| \le (ep)^{D+1}$ entries in a full moment table of the same order. Replacing each m_K by the entry $S[K]$ of an approximate moment table restricted to these local sets, and running projected gradient descent as in Algorithm 1, learns all couplings to accuracy ϵ (Theorem 4 in Appendix G). For an empirical table this requires $n = O\big(2^{2D}e^{8\gamma}\gamma^4(D + \log(p/\delta))/\epsilon^4\big)$ samples per moment.

3 Conclusion

We have shown that having access to $O(\gamma)$ statistics is both computationally and statistically efficient for learning Ising models. As it stands, we see that there is gap in our understanding on the power of statistics below this $O(\gamma)$ threshold but higher than sufficient statistics (order 2). As we have seen Section 2.4, if we assume stronger priors on the model, the learning problem can be potentially solved with even lower order statistics (when $D = o(\gamma)$). This suggests that restricting the model class in other ways (Ferromagnetic models, Sherrington-Kirkpatrick models, *etc.*) can potentially lower the order of statistics required to solve the learning problem. We leave these extensions for future work.

Acknowledgments and Disclosure of Funding

This work has been supported by the U.S. Department of Energy/Office of Science Advanced Scientific Computing Research Program.

References

[1] Alekh Agarwal. *Computational trade-offs in statistical learning*. University of California, Berkeley, 2012.

[2] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Ewin Tang. Learning quantum hamiltonians at any temperature in polynomial time. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1470–1477, 2024.

[3] Guy Bresler. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782, 2015.

[4] Guy Bresler, David Gamarnik, and Devavrat Shah. Hardness of parameter estimation in graphical models. *Advances in Neural Information Processing Systems*, 27, 2014.

[5] Sébastien Bubeck. Convex Optimization: Algorithms and Complexity, 2015. arXiv:1405.4980.

[6] Vitaly Feldman. A general characterization of the statistical query complexity. In *Conference on learning theory*, pages 785–830. PMLR, 2017.

[7] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for planted clique. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 655–664, 2013.

[8] Jason Gaitonde and Elchanan Mossel. A unified approach to learning ising models: Beyond independence and bounded width. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 503–514, 2024.

[9] Surbhi Goel, Daniel M Kane, and Adam R Klivans. Learning ising models with independent failures. In *Conference on Learning Theory*, pages 1449–1469. PMLR, 2019.

[10] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6, 2005.

[11] Abhijith Jayakumar, Andrey Y Lokhov, Sidhant Misra, and Marc Vuffray. Discrete distributions are learnable from metastable samples. *Nature Communications*, 17:8957, 2026.

[12] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[13] A. Klivans and R. Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354, Oct 2017.

[14] Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of Ising models. *Science advances*, 4(3):e1700791, 2018.

[15] Andrea Montanari. Computational Implications of Reducing Data to Sufficient Statistics, July 2015. arXiv:1409.3821.

[16] Shyam Narayanan. Improved algorithms for learning quantum hamiltonians, via flat polynomials. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4360–4385. PMLR, 30 Jun–04 Jul 2025.

[17] H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017.

[18] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[19] Chirag Pabbaraju, Dhruv Rohatgi, Anish Prasad Sevekari, Holden Lee, Ankur Moitra, and Andrej Risteski. Provable benefits of score matching. *Advances in Neural Information Processing Systems*, 36:61306–61326, 2023.

[20] Federico Ricci-Tersenghi. The bethe approximation for solving the inverse ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08015, 2012.

[21] Shreya Shukla, Abhijith Jayakumar, and Andrey Y Lokhov. Learning of statistical field theories. *arXiv preprint arXiv:2511.09859*, 2025.

[22] Toshiyuki Tanaka. Mean-field theory of Boltzmann machine learning. *Physical Review E*, 58(2):2302, 1998.

[23] Marc Vuffray, Sidhant Misra, and Andrey Lokhov. Efficient learning of discrete graphical models. *Advances in Neural Information Processing Systems*, 33:13575–13585, 2020.

[24] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2016.

[25] Alexander S Wein. Computational complexity of statistics: New insights from low-degree polynomials. *arXiv preprint arXiv:2506.10748*, 2025.

A Technical Lemmas

A.1 Proof of Lemma 3

Proof. Let

$$P_d(x) = \sum_{k=0}^{d} \frac{(-x)^k}{k!}$$

be the degree-d Taylor polynomial of e^{-x} about 0. We want a uniform bound on the error

$$E_d(\gamma) := \max_{x\in[-\gamma,\gamma]} \left|e^{-x} - P_d(x)\right|.$$

We have,

$$\left|e^{-x} - P_d(x)\right| \le \sum_{k=d+1}^{\infty} \frac{|x|^k}{k!} \le \sum_{k=d+1}^{\infty} \frac{\gamma^k}{k!}$$

Then,

$$\sum_{k=d+1}^{\infty} \frac{\gamma^k}{k!} = e^{\gamma}\Big(1 - e^{-\gamma}\sum_{k=0}^{d}\frac{\gamma^k}{k!}\Big) = e^{\gamma}\Pr\big[\mathrm{Poisson}(\gamma) \ge d+1\big].$$

Therefore,

$$E_d(\gamma) \le e^{\gamma}\Pr\big[\mathrm{Poisson}(\gamma) \ge d+1\big].$$

Use the standard Poisson (Chernoff) tail bound: for $a > \gamma$,

$$\Pr\big[\mathrm{Poisson}(\gamma) \ge a\big] \le e^{-\gamma}\Big(\frac{e\gamma}{a}\Big)^a.$$

Taking $a = d+1$ yields

$$E_d(\gamma) \le e^{\gamma}\cdot e^{-\gamma}\Big(\frac{e\gamma}{d+1}\Big)^{d+1} = \Big(\frac{e\gamma}{d+1}\Big)^{d+1} = e^{(d+1)(\log(\frac{\gamma}{d+1})+1)}$$

To finish the proof, we set $\varepsilon_{poly} = \frac{\varepsilon}{8\gamma}$. From this we get the following implicit lower bound for d that ensures that the error remains small, $e^{-(d+1)(\log(\frac{d+1}{\gamma})-1)} \le \frac{\varepsilon}{8\gamma}$. The implicit inversion can be performed by a simple yet tedious calculation and is detailed in Appendix D.

■

A.2 Proof of Lemma 2

Let S be a (D, τ, δ)-approximate moment table, and work on the event $\mathcal{E}$ on which $\max_{K \in \mathcal{K}_D} |S[K] - m_K| \leq \tau$, which has probability at least $1 - \delta$ by Definition 1.

Using the table entries in (7), we get a random approximation for the expectation value in the LHS, which we call $\widetilde{\mathbb{E}}_S[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j + \theta_u)^k]$. We use the $\widetilde{\mathbb{E}}_S$ here to emphasize that this approximation is constructed by combining table entries, and is not itself a moment of the model.

Since the σ variables square to 1, each monomial $\sigma_v\sigma_u^{k+1}\prod_{j\neq u}\sigma_j^{m_j}$ equals $\prod_{i\in K}\sigma_i$ for a set K with $|K| \leq k+2$, and for $k \leq d$ we have $|K| \leq D$, so the corresponding table entry $S[K]$ exists. From (7), we compute

$$\sum_{\underline{m}\in\mathcal{M}_k} |C_{\underline{\theta},\underline{m}}| = (\sum_{j\neq u} |\theta_{uj}| + |\theta_u|)^k \leq \gamma^k$$

for every $\underline{\theta}_u$ in the ℓ_1 ball of radius γ. On $\mathcal{E}$ this implies that,

$$\left| \mathbb{E}[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j + \theta_u)^k] - \widetilde{\mathbb{E}}_S[\sigma_u^{k+1}\sigma_v(\sum_{j\neq u}\theta_{uj}\sigma_j + \theta_u)^k] \right| \leq \gamma^k\tau.$$

Using this for each $k \leq d$ in the expression in (6), we get on $\mathcal{E}$,

$$\varepsilon_{stat} \leq \tau \sum_{k\leq d} \frac{\gamma^k}{k!} \leq \tau e^\gamma. \tag{23}$$

We set $\tau e^\gamma = \varepsilon/(8\gamma)$, i.e. $\tau = \varepsilon e^{-\gamma}/(8\gamma)$, which gives the claim. The bound holds for every coordinate of the gradient, every u and every $\underline{\theta}_u$ on the same event $\mathcal{E}$, so no further union bound is needed.

B Proof of Lemma 4

This is a modification of the proof found in [4] to the ℓ_1 ball. Define the un-projected intermediate step; $y^{t+1} = x^t - \eta\widetilde{\nabla}\mathcal{L}(x^t)$

$$\begin{aligned}
\mathcal{L}(x^t) - \mathcal{L}(x^*) &\leq \nabla\mathcal{L}(x^t)^\top(x^t - x^*) \\
&= \widetilde{\nabla}\mathcal{L}(x^t)^\top(x^t - x^*) + \big(\nabla\mathcal{L}(x^t)^\top - \widetilde{\nabla}\mathcal{L}(x^t)^\top\big)(x^t - x^*) \\
&\leq \widetilde{\nabla}\mathcal{L}(x^t)^\top(x^t - x^*) + \varepsilon\,\|x^t - x^*\|_1/4\gamma \\
&\leq \frac{1}{\eta}\,(x^t - y^{t+1})^\top(x^t - x^*) + \varepsilon/2 \\
&= \frac{1}{2\eta}\big(\|x^t - x^*\|_2^2 + \|x^t - y^{t+1}\|_2^2 - \|y^{t+1} - x^*\|_2^2\big) + \varepsilon/2 \\
&= \frac{1}{2\eta}\big(\|x^t - x^*\|_2^2 - \|y^{t+1} - x^*\|_2^2\big) + \frac{\eta}{2}\,\|\widetilde{\nabla}\mathcal{L}(x^t)\|_2^2 + \varepsilon/2\,.
\end{aligned}$$

The projection to a convex body always contracts (see Lemma 3.1 in [5]), hence, $\|y^{t+1} - x^*\|_2 \geq \|\mathcal{P}_C(y^{t+1}) - \mathcal{P}_C(x^*)\|_2 = \|x^{t+1} - x^*\|_2$. . Using this we get we get following inequality,

$$\mathcal{L}(x^t) - \mathcal{L}(x^*) \leq \frac{1}{2\eta}\big(\|x^t - x^*\|_2^2 - \|x^{t+1} - x^*\|_2^2\big) + \frac{\eta}{2}\,\|\widetilde{\nabla}\mathcal{L}(x^t)\|_2^2 + \varepsilon/2 \tag{24}$$

Now summing these from 1 to T and using the facts that $||x^1 - x^*||_2 \leq 2\gamma$ and $\|\widetilde{\nabla}\mathcal{L}(x^t)\|_2^2 \leq L$, we get,

$$\sum_{t=1}^{T}(\mathcal{L}(x^t) - \mathcal{L}(x^*)) \leq \frac{1}{\eta}(2\gamma^2) + \frac{\eta T L^2}{2} + T\varepsilon/2. \tag{25}$$

By fixing $\eta = \frac{2\gamma}{L\sqrt{T}}$ then gives us,

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(x^t) - \mathcal{L}(x^*) \leq \frac{2\gamma L}{\sqrt{T}} + \varepsilon/2. \tag{26}$$

Now we know from convexity that $\mathcal{L}(\frac{1}{T}\sum_{t=1}^{T}x^t) - \mathcal{L}(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(x^t) - \mathcal{L}(x^*)$.

Now choosing $T = \left\lceil \frac{16\gamma^2 L^2}{\varepsilon^2} \right\rceil$ gives the desired result

C Curvature bounds

Notation

1. $\underline{\theta}^*_u$, true parameters connected to the variable u
2. $\hat{\mathcal{L}}_n$, IS loss for the variable u, estimated from n samples
3. $\tilde{\underline{\theta}}_u = \operatorname{argmin} \|\underline{\theta}\|_1 \leq \gamma\hat{\mathcal{L}}(\underline{\theta})_n$
4. $\hat{\theta}_u$, an approximate minimizer of the loss $\hat{\mathcal{L}}_n$

Lemma 7. *The curvature term is lower bounded as* $\delta\mathcal{L}^*(\hat{\underline{\theta}}_u) \geq \frac{e^{-3\gamma}}{2+2\gamma}(\max_{u\neq v}|\theta^*_{u,v} - \hat{\theta}_{u,v}|^2)$

Proof. We work with the definition of the curvature function,

$$\delta\mathcal{L}^*(\underline{\theta}) \equiv \mathcal{L}(\underline{\theta}) - \mathcal{L}(\underline{\theta}^*) - \langle\nabla\mathcal{L}(\underline{\theta}^*), \underline{\theta} - \underline{\theta}^*\rangle$$

Define local energy function $E_u(\underline{\sigma};\theta) = \sum_{k\neq u}\theta_{u,k}\sigma_u\sigma_k + \theta_u\sigma_u$. Notice that this function is linear in the θ variables. So given an error parameter $\Delta \equiv \theta - \theta^*$ on the parameters we can see that,

$$\mathcal{L}(\underline{\theta}) = \mathcal{L}(\underline{\theta}^* + \underline{\Delta}) = \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu}\exp(-E_u(\underline{\sigma};\underline{\theta}^* + \underline{\Delta})) = \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu}\exp(-E_u(\underline{\sigma};\underline{\theta}^*))\ \exp(-E_u(\underline{\sigma};\underline{\Delta})). \tag{27}$$

The following is also true,

$$\langle\nabla\mathcal{L}(\underline{\theta}^*), \underline{\Delta}\rangle = \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu} - \exp(-E_u(\underline{\sigma};\underline{\theta}^*))(\sum_{k\neq u}\frac{\partial E_u(\underline{\sigma};\underline{\theta}_u)}{\partial\theta_{uk}}\Delta_{uk} + \frac{\partial E_u(\underline{\sigma};\underline{\theta}_u)}{\partial\theta_u}\Delta_u) \tag{28}$$

$$= -\mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu}\exp(-E_u(\underline{\sigma};\underline{\theta}^*))E_u(\underline{\sigma};\underline{\Delta}). \tag{29}$$

The last two expressions together imply that given a function $g(x) = e^{-x} - 1 + x$, we can write the curvature term as,

$$\delta\mathcal{L}^*(\underline{\theta}) = \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu}\exp(-E_u(\underline{\sigma},\underline{\theta}^*))g(E_u(\underline{\sigma},\underline{\Delta}_u)) \tag{30}$$

First observe that $g(x) \geq \frac{x^2}{2+|x|}$ for all $x \in \mathbb{R}$ (See Lemma 9 for the proof)

Using this and the fact that $|E_u(\underline{\sigma},\underline{\theta}^*)| \leq \gamma$ we get,

$$\delta\mathcal{L}^*(\underline{\theta}_u) \geq \frac{e^{-\gamma}}{2+2\gamma}\mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu}(E(\underline{\sigma},\underline{\Delta}))^2 \tag{31}$$

We lower bound the above in terms of the variance,

$$\mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu}(E_u(\underline{\sigma},\underline{\Delta}))^2 = \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu}(\sum_{k\neq u}\sigma_k\Delta_{u,k} + \Delta_u)^2, \tag{32}$$

$$\geq \mathop{\mathrm{Var}}_{\underline{\sigma}}\sum_{k\neq u}\sigma_k\Delta_{u,k}. \tag{33}$$

Note that the price to pay here is we lose control over the error on the magnetic field term.

Now we can use the law of variances to bound the expectation value. For any $v \neq u$ we can do the following,

$$\operatorname*{Var}_{\underline{\sigma}} \sum_{k\neq u} \sigma_k \Delta_{u,k}, \geq \mathop{\mathbb{E}}_{\underline{\sigma}_{\setminus v}} \operatorname{Var}\left[\sum_{k\neq u} \sigma_k \Delta_{u,k} \,\middle|\, \underline{\sigma}_{\setminus v}\right] \tag{34}$$

$$= \Delta_{u,v}^2 \mathop{\mathbb{E}}_{\underline{\sigma}_{\setminus v}} \operatorname{Var}[\sigma_v | \underline{\sigma}_{\setminus v}] \tag{35}$$

Now it is well known that conditional variances of a single variable Ising model can be lower bounded by a function of the ℓ_1-width ([13, 24])

$$\operatorname*{Var}_{\sigma_v \sim \mu(.|\underline{\sigma}_{\setminus v})} [\sigma_v | \underline{\sigma}_{\setminus v}] = 1 - \tanh^2(\sum_{j\neq v} \theta^*_{vj}\sigma_j + \theta^*_v) \geq 1 - \tanh^2(\gamma) \geq e^{-2\gamma}. \tag{36}$$

Using this and maximizing over all $v \neq u$ we get the following relation

$$\delta\mathcal{L}^*(\underline{\theta}) \geq \frac{e^{-\gamma}}{2+2\gamma} \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu} (E_u(\underline{\sigma}, \underline{\Delta}))^2 \geq \frac{e^{-3\gamma}}{2+2\gamma} \max_{v\neq u} |\theta_{uv} - \theta^*_{uv}|^2 \tag{37}$$

■

D Bounding d

$$e^{-(d+1)\left(\log\frac{d+1}{\gamma} - 1\right)} \leq e^{-C} \tag{38}$$

$$\iff (d+1)\left(\log\frac{d+1}{\gamma} - 1\right) \geq C. \tag{39}$$

Let $x := d+1$. Since $f(x) = x(\log(x/\gamma) - 1)$ is increasing for $x > \gamma$, the least integer x satisfying (39) is the ceiling of the solution of the equation $x(\log(x/\gamma) - 1) = C$. Let's solve this equation:

$$x\left(\log\frac{x}{\gamma} - 1\right) = C \tag{40}$$

$$\implies x\, e^{-C/x} = \gamma e. \tag{41}$$

Set $t := C/x$. Then

$$\frac{C}{t}\, e^{-t} = \gamma e \quad \implies \quad t\, e^t = \frac{C}{\gamma e}. \tag{42}$$

The W function enters here. Using the defining identity for the W function, $t\, e^t = z \iff t = W(z)$,

$$t\, e^t = \frac{C}{\gamma e} \quad \implies \quad t = W\left(\frac{C}{\gamma e}\right) \quad \implies \quad \frac{C}{x} = W\left(\frac{C}{\gamma e}\right). \tag{43}$$

Hence

$$x = \frac{C}{W\left(\frac{C}{\gamma e}\right)}, \qquad C = 3\gamma + \log\frac{16(1+\gamma)\gamma}{\epsilon^2}. \tag{44}$$

Therefore the smallest integer d satisfying (39) is

$$d = \left\lceil \frac{C}{W\left(\frac{C}{\gamma e}\right)} \right\rceil - 1 = \left\lceil \frac{3\gamma + \log\frac{16(1+\gamma)\gamma}{\epsilon^2}}{W\left(\frac{3}{e} + \frac{1}{\gamma e}\log\frac{16(1+\gamma)\gamma}{\epsilon^2}\right)} \right\rceil - 1. \tag{45}$$

(Here W denotes the principal branch when the argument is positive.)

E Notes

Lemma 8 (Poisson Chernoff tail bound)**.** *For $a > b \in \mathbb{R}$,* $\Pr[Poisson(b) \geq a] \leq e^{-b}(\frac{eb}{a})^a$

Proof. Let $X \sim$ Poisson(b), then $\Pr[X = k] = \frac{e^{-b}b^k}{k!}$ .We will do a Chernoff bound to bound the tail. For any $t > 0$, from the Markov inequality we have,

$$\Pr[X \geq a] = \Pr[e^{tX} \geq e^{ta}] \leq \frac{\mathbb{E}e^{tX}}{e^{ta}} \tag{46}$$

Now $\mathbb{E}e^{tX} = e^{-b}\sum_{k=0}^{\infty}\frac{(e^t b)^k}{k!} = e^{e^t b - b}$. Hence for all $t > 0$, we have $\Pr[X \geq a] \leq e^{e^t b - b - ta}$.

Now optimium t for this falls at $\log(\frac{a}{b})$. Plugging this t into the tail bound above gives the desired answer.

■

Lemma 9. *For all $x \in \mathbb{R}$, the function $g(x) \coloneqq e^{-x} - 1 + x$ satisfies*

$$g(x) \;\geq\; \frac{x^2}{2 + |x|}.$$

Proof. Define the auxiliary function

$$M(x) \;\coloneqq\; (2 + |x|)\, g(x) - x^2 \;=\; (2 + |x|)\,(e^{-x} - 1 + x) - x^2.$$

Clearly $M(0) = 0$. We show $M(x) \geq 0$ for all x, which is equivalent to the desired inequality.

Case 1: $x \geq 0$**.** Here $|x| = x$, so

$$M(x) = (2 + x)(e^{-x} - 1 + x) - x^2.$$

Differentiating,

$$M'(x) = 1 - (1 + x)e^{-x}, \qquad M''(x) = x\, e^{-x} > 0 \quad (x > 0).$$

Hence M' is increasing on $[0, \infty)$ with $M'(0) = 0$, so $M'(x) \geq 0$ for $x \geq 0$ and M is nondecreasing on $[0, \infty)$. Since $M(0) = 0$, we get $M(x) \geq 0$ for all $x \geq 0$.

Case 2: $x \leq 0$**.** Here $|x| = -x$, so

$$M(x) = (2 - x)(e^{-x} - 1 + x) - x^2.$$

Differentiating,

$$M'(x) = 3(1 - e^{-x}) - 4x + x\, e^{-x}, \qquad M''(x) = 4(e^{-x} - 1) - x\, e^{-x}.$$

For $x < 0$ we have $e^{-x} - 1 > 0$ and $-x > 0$, hence $M''(x) > 0$. Thus M' is increasing on $(-\infty, 0]$ and $\lim_{x \to 0^-} M'(x) = 0$, which implies $M'(x) \leq 0$ for $x \leq 0$. Therefore M is nonincreasing on $(-\infty, 0]$, so for every $x \leq 0$ we have $M(x) \geq M(0) = 0$.

Combining the two cases gives $M(x) \geq 0$ for all $x \in \mathbb{R}$, i.e.

$$(2 + |x|)\big(e^{-x} - 1 + x\big) \;\geq\; x^2,$$

which is equivalent to $g(x) \geq x^2/(2 + |x|)$. Equality holds at $x = 0$. ■

F Magnetic field learning

F.1 Proof of Theorem 3

Proof. Now use the same strategy as in Theorem 1 to approximate the gradient of $\mathcal{R}(\theta)$ and expand exponential in the derivative up to order d.

Algorithm 2: Magnetic Field Learning with Moment-Based Approximate Gradients

Input: Degree d, coupling estimates $\{\hat{\theta}_{u,v}\}$, approximate moment table S of order $d+1-(d \bmod 2)$, width bound γ
, step size η, iterations T

Output: Magnetic field estimates $\{\hat{\theta}_u\}_{u\in[p]}$

for $u = 1$ ***to*** p **do**
  Initialize $\theta_u^{(1)} \leftarrow 0$
  for $t = 1$ ***to*** T **do**
    `// Gradient Approximation`
    Construct $\tilde{\nabla}\mathcal{R}_u(\theta_u^{(t)})$ from S as follows:
    (i) Replace the exponential in the exact gradient by a degree-d approximation as in (47).
    (ii) Expand the resulting moments using (48).
    (iii) Substitute each monomial expectation $\mathbb{E}[\prod_{i\in K}\sigma_i]$ by the table entry $S[K]$.
    `// Gradient Step and projection`
    $\theta_u^{(t+1)} \leftarrow \mathcal{P}_{|\cdot|\le\gamma}\big(\theta_u^{(t)} - \eta\,\tilde{\nabla}\mathcal{R}_u(\theta_u^{(t)})\big)$
  end
  `// Averaging`
  $\hat{\theta}_u \leftarrow \frac{1}{T}\sum_{t=1}^{T}\theta_u^{(t)}$
end
return $\{\hat{\theta}_u\}_{u\in[p]}$

$$\mathbb{E}\left[\sigma_u e^{-(\theta\sigma_u+\sigma_u\sum_{j\in N_u}\hat{\theta}_{uj}\sigma_j)}\right] = \sum_{k\le d}\frac{(-1)^k\mathbb{E}[\sigma_u^{k+1}(\sum_{j\in N_u}\hat{\theta}_{uj}\sigma_j+\theta)^k]}{k!} + \varepsilon_{poly} \tag{47}$$

Now the powers in this expansion can be expanded using multinomial coefficients as in (7)

$$\mathbb{E}[\sigma_u^{k+1}(\sum_{j\in N_u}\hat{\theta}_{uj}\sigma_j+\theta)^k] = \sum_{\underline{m}\in\mathcal{M}_k}\hat{C}_{\theta,\underline{m}}\,\mathbb{E}\left[\sigma_u^{k+1}\prod_{j\neq u}\sigma_j^{m_j}\right]. \tag{48}$$

Where $\mathcal{M}_k$ is defined as in (7) and $\hat{C}_{\theta,\underline{m}} \equiv \frac{k!\ \ \theta^{m_u}\prod_{j\neq u}\hat{\theta}_{uj}^{m_j}}{\prod_{j\in[p]}m_j!}$. Here we have used the convention that $\hat{\theta}_{u,j} = 0$ for $j \notin N_u$.

Since the largest value of k is d, we see that every monomial in (48) corresponds to an entry of a moment table of order $d+1-(d \bmod 2)$. Replacing each monomial expectation by the corresponding table entry gives the following approximation for the derivative of $\mathcal{R}$ w.r.t θ.

Lemma 10. *Let S be a $\big(d+1-(d \bmod 2), \tau, \delta\big)$-approximate moment table. Consider the estimate for $\sum_{k\le d}\frac{(-1)^k\mathbb{E}[\sigma_u^{k+1}(\sum_{j\in N_u}\hat{\theta}_{uj}\sigma_j+\theta)^k]}{k!}$ constructed by first expanding it in terms of monomial expectation values as in (48) and then replacing each of them by the corresponding entry of S. Then, with probability at least $1-\delta$, the estimation error is bounded by $\tau e^{2\gamma}$ simultaneously for every $u \in [p]$ and every $|\theta| \le \gamma$.*

Putting the polynomial approximation together with (20), gives us a approximation for gradient of $\mathcal{R}(\theta)$. We call this approximation $\tilde{\nabla}\mathcal{R}(\theta)$. Totally we have with probability $1-\delta$,

$$\left|\frac{d\mathcal{R}^*(\theta)}{d\theta} - \tilde{\nabla}\mathcal{R}(\theta)\right| \le e^{2\gamma}D\epsilon + \tau e^{2\gamma} + \varepsilon_{poly} \tag{49}$$

Now to keep things simpler, we define a new error term ε and choose the table accuracy τ and ε_{poly} such that $\tau e^{2\gamma} = \varepsilon_{poly} = \frac{\varepsilon}{4\gamma}$.

Now we can use this approximate gradient in Lemma 4 to bound how close we can get to $\mathcal{R}^*(\theta_u^*)$. To this end we need to bound the ℓ_1 norm (just the absolute value in this case) of the approximate gradient.

$$|\tilde{\nabla}\mathcal{R}(\theta)| \leq \left|\frac{d\mathcal{R}^*(\theta)}{d\theta}\right| + e^{2\gamma}D\epsilon + \frac{\varepsilon}{4\gamma} \leq (1 + D\epsilon)e^{2\gamma} + \frac{\varepsilon}{4\gamma}. \tag{50}$$

Choose

$$L \equiv (1 + D\epsilon)e^{2\gamma} + \frac{\varepsilon}{4\gamma}. \tag{51}$$

Then we can apply Lemma 4 with

$$T = \left\lceil \frac{16\gamma^2 L^2}{\varepsilon^2} \right\rceil, \qquad \eta = \frac{2\gamma}{L\sqrt{T}}. \tag{52}$$

Using this directly in Lemma 4 gives us an estimate $\hat{\theta}$ such that $\mathcal{R}^*(\hat{\theta}) - \mathcal{R}^*(\theta_u^*) \leq \varepsilon + 4\gamma e^{2\gamma} D\epsilon$. Now, we can estimate the error parameter estimation error by bounding the curvature of the $\mathcal{R}^*$ function at θ_u^*

$$\delta\mathcal{R}^*(\hat{\theta}) \equiv \mathcal{R}^*(\hat{\theta}) - \mathcal{R}^*(\theta^*) - \langle \nabla\mathcal{R}^*(\theta^*), \hat{\theta} - \theta^* \rangle \tag{53}$$

$$= \mathcal{R}^*(\hat{\theta}) - \mathcal{R}^*(\theta^*) \leq \varepsilon + 4\gamma e^{2\gamma} D\epsilon. \tag{54}$$

Now we can also lower bound the curvature of the function and derive the following result

Lemma 11. *The curvature term is lower bounded as* $\delta\mathcal{R}^*(\hat{\theta}) \geq \frac{e^{-\gamma}}{2+2\gamma}|\theta_u^* - \hat{\theta}|^2$

Proof. Note, that the mechanism of this proof is basically the same as 7. However, due to there being only one optimization variable, we do not have to resort to variance arguments at the end to prove a useful bound.

We work with the definition of the curvature function,

$$\delta\mathcal{R}^*(\theta) \equiv \mathcal{R}(\theta) - \mathcal{R}(\theta^*) - \frac{d\mathcal{R}(\theta^*)}{d\theta}(\theta - \theta^*)$$

Define local energy function $F(\underline{\sigma}; \theta) = \sum_{k \neq u} \theta_{u,k}^* \sigma_u \sigma_k + \theta\sigma_u$. Notice that this function is linear in the θ variables. So given an error parameter $\Delta \equiv \theta - \theta^*$ on the parameters we can see that,

$$\mathcal{R}^*(\theta) = \mathcal{R}^*(\theta^* + \Delta) = \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu} \exp(-F(\underline{\sigma}; \theta^* + \Delta)) = \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu} \exp(-F(\underline{\sigma}; \theta^*))\ \exp(-\sigma_u\Delta). \tag{55}$$

The following is also true,

$$\frac{d\mathcal{R}(\theta^*)}{d\theta}\Delta = -\mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu} \exp(-F(\underline{\sigma}; \underline{\theta}^*))(\sigma_u\Delta). \tag{56}$$

The last two expressions together imply that given a function $g(x) = e^{-x} - 1 + x$, we can write the curvature term as,

$$\delta\mathcal{R}^*(\theta) = \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu} \exp(-F(\underline{\sigma}, \theta^*))g(\sigma_u\Delta) \tag{57}$$

First observe that $g(x) \geq \frac{x^2}{2+|x|}$ for all $x \in \mathbb{R}$ (See Lemma 9 for the proof)

Using this and the fact that $|F(\underline{\sigma}, \theta^*)| \leq \gamma$ we get,

$$\delta\mathcal{R}^*(\theta) \geq \frac{e^{-\gamma}}{2+2\gamma} \mathop{\mathbb{E}}_{\underline{\sigma}\sim\mu} (\sigma_u\Delta)^2 = \frac{e^{-\gamma}}{2+2\gamma}\ (\theta - \theta_u^*)^2 \tag{58}$$

■

Now using this we find that, $\frac{e^{-\gamma}}{2+2\gamma}|\theta_u^* - \hat{\theta}|^2 \leq \varepsilon + 4\gamma e^{2\gamma} D\epsilon$

Now define, $\epsilon_h \equiv \sqrt{\varepsilon 2(1+\gamma)}e^{\gamma/2}$. Using this we have,

$$|\theta_u^* - \hat{\theta}| \;\leq\; \sqrt{\epsilon_h^2 + 8\gamma(1+\gamma)e^{3\gamma}D\epsilon} \;\leq\; \epsilon_h + \sqrt{8\gamma(1+\gamma)e^{3\gamma}D\epsilon}$$

Choice of d and τ. Recall that we chose the table accuracy τ and the Taylor truncation error ε_{poly} so that $\tau e^{2\gamma} = \varepsilon_{poly} = \varepsilon/(4\gamma)$.

Degree d. By Lemma 3 (applied with width parameter 2γ), choosing $d+1 > 2\gamma$ ensures

$$\varepsilon_{poly} \leq \sup_{x\in[-2\gamma,2\gamma]} \left| e^{-x} - \sum_{k=0}^{d} \frac{(-x)^k}{k!} \right| \leq \exp\!\left(-(d+1)\Big(\log\frac{d+1}{2\gamma} - 1\Big)\right).$$

Therefore it suffices to choose d so that

$$\exp\!\left(-(d+1)\Big(\log\frac{d+1}{2\gamma} - 1\Big)\right) \leq \frac{\varepsilon}{4\gamma}.$$

Equivalently, substituting $\varepsilon = \epsilon_h^2 e^{-\gamma}/(2(1+\gamma))$ into the Lambert-W calculation (cf. Appendix D), one can take

$$d = \left\lceil \frac{\gamma + \log\dfrac{8\gamma(1+\gamma)}{\epsilon_h^2}}{W\!\left(\frac{1}{2e} + \frac{1}{2\gamma e}\log\dfrac{8\gamma(1+\gamma)}{\epsilon_h^2}\right)} \right\rceil - 1, \tag{59}$$

where W is the principal branch of the Lambert-W function.

Table accuracy τ. By Lemma 10, on the event that the table is τ-accurate, which has probability at least $1-\delta$, the monomial-based estimate of the truncated series in (47) has estimation error at most $\tau e^{2\gamma} = \varepsilon/(4\gamma)$ for every u. This gives

$$\tau = \frac{\varepsilon e^{-2\gamma}}{4\gamma} = \frac{\epsilon_h^2\, e^{-3\gamma}}{8\gamma(1+\gamma)} = \tau_h.$$

Since the highest monomial degree appearing in (48) is at most $d+1$, the table must have order $d+1-(d \bmod 2)$. Finally, Lemma 1 with $\frac{2}{\tau_h^2} = \frac{2^7\gamma^2(1+\gamma)^2 e^{6\gamma}}{\epsilon_h^4}$ and $|\mathcal{K}_{d+1}| \leq (ep)^{d+1}$ gives the stated sufficient sample size for empirical tables.

■

F.2 Proof of Lemma 10

The proof is the same as that of Lemma 2. Work on the event $\mathcal{E}$ on which $\max_K |S[K] - m_K| \leq \tau$ over all entries of the table, which has probability at least $1-\delta$.

Replacing the monomial expectations in (48) by the corresponding table entries gives an approximation for the expectation value in the LHS, which we denote by $\widetilde{\mathbb{E}}_S[\sigma_u^{k+1}(\sum_{j\neq u}\hat{\theta}_{uj}\sigma_j + \theta)^k]$. Since the σ variables square to 1, each monomial $\sigma_u^{k+1}\prod_{j\neq u}\sigma_j^{m_j}$ equals $\prod_{i\in K}\sigma_i$ for a set K with $|K| \leq k+1 \leq d+1$ and the parity of $k+1$, so the entry $S[K]$ exists in a table of order $d+1-(d \bmod 2)$.

From (48), we also see that,

$$\sum_{\underline{m}\in\mathcal{M}_k} |\hat{C}_{\theta,\underline{m}}| = (\sum_{j\neq u}|\hat{\theta}_{uj}| + |\theta|)^k \leq (2\gamma)^k.$$

This is because we use the ℓ_1 width as a constraint in the first round of optimization as well to find the second order parameters. On $\mathcal{E}$ this implies that,

$$\left| \mathbb{E}[\sigma_u^{k+1}(\sum_{j\neq u}\hat{\theta}_{uj}\sigma_j + \theta)^k] - \widetilde{\mathbb{E}}_S[\sigma_u^{k+1}(\sum_{j\neq u}\hat{\theta}_{uj}\sigma_j + \theta)^k] \right| \leq (2\gamma)^k\tau.$$

Using this for each $k \le d$ in the truncated series in (47), we get on $\mathcal{E}$ an estimation error of at most

$$\tau \sum_{k \le d} \frac{(2\gamma)^k}{k!} \le \tau e^{2\gamma}.$$

The bound holds for every u and every $|\theta| \le \gamma$ on the same event $\mathcal{E}$, so no further union bound is needed.

G Sample complexity of learning with structural assumptions

Theorem 4 (Interaction screening with known D-regular structure)**.** *Assume the interaction graph (edge set E) is known and D-regular, so $|N_u| = D$ for every $u \in [p]$. Let $\mathcal{K}_E := \{K : K \subseteq \{u\} \cup N_u$ for some $u \in [p]\}$ be the set of local index sets, so that $|\mathcal{K}_E| \le p\, 2^{D+1}$. Fix a target error $\epsilon \le 4\gamma$ and failure probability $\delta \in (0,1)$, and let*

$$\tau = \frac{\epsilon^2 e^{-4\gamma}}{2^{D+4}\gamma(1+\gamma)}.$$

Suppose $S = \{S[K]\}_{K \in \mathcal{K}_E}$ satisfies $\mathbb{P}\big[\max_{K \in \mathcal{K}_E} |S[K] - m_K| \le \tau\big] \ge 1 - \delta$; for example, the restriction of any $(D+1, \tau, \delta)$-approximate moment table to $\mathcal{K}_E$. For each u, let $\underline{\theta}_u = (\theta_u, (\theta_{u,v})_{v \in N_u}) \in \mathbb{R}^{D+1}$ collect the field and the couplings of u to its neighbors, the couplings to non-neighbors being fixed to zero. Then projected gradient descent on the interaction screening loss over the feasible set $\{\underline{\theta}_u \in \mathbb{R}^{D+1} : \|\underline{\theta}_u\|_1 \le \gamma\}$, using the entries of S in place of the moments, returns with probability at least $1 - \delta$ estimates satisfying, simultaneously for every $u \in [p]$,

$$\max_{v \in N_u} |\hat{\theta}_{u,v} - \theta^*_{u,v}| \le \epsilon.$$

In particular, it suffices to take

$$T = \left\lceil 2^8 (D+1)\, \gamma^2 (1+\gamma)^2 e^{8\gamma}\, \epsilon^{-4} \right\rceil, \qquad \eta = \frac{2\gamma}{L\sqrt{T}}, \qquad L = 2\sqrt{D+1}\, e^{\gamma}.$$

If S consists of empirical moments built from a shared dataset or from separate datasets, then, by the argument of Lemma 1 applied to the $|\mathcal{K}_E|$ entries, it suffices that

$$n \ge 2^{2D+9}\, \gamma^2 (1+\gamma)^2 e^{8\gamma}\, \frac{(D+1)\log 2 + \log\left(\frac{2p}{\delta}\right)}{\epsilon^4}.$$

Proof. There is no polynomial approximation error in this case unlike Theorem 1, so we only need to account for statistical error.

$$\|\nabla \mathcal{L}_u(\underline{\theta}) - \widetilde{\nabla} \mathcal{L}_u(\underline{\theta})\|_\infty \le \sum_K (|\nabla \widehat{f}_{u,K}(\underline{\theta})|) \max_{K \subseteq \{u\} \bigcup N_u} |m_K - S[K]| := \varepsilon_{stat} \tag{60}$$

Here the gradient is taken with respect to the $D+1$ coordinates of $\underline{\theta}_u$, and each coordinate of $\nabla_{\underline{\theta}} f_u(\underline{\sigma}; \underline{\theta})$ is a function on $\{-1,1\}^{D+1}$. Using the standard Fourier (Walsh–Hadamard) expansion on the boolean hypercube (see, e.g., [18, Ch. 1]), we may write

$$\nabla_{\underline{\theta}} f_u(\underline{\sigma}; \underline{\theta}) = \sum_{K \subseteq \{u\} \cup N_u} \nabla \widehat{f}_{u,K}(\underline{\theta})\, \underline{\sigma}_K, \qquad \nabla \widehat{f}_{u,K}(\underline{\theta}) = 2^{-(D+1)} \sum_{\mathbf{s} \in \{-1,1\}^{D+1}} \nabla_{\underline{\theta}} f_u(\mathbf{s}; \underline{\theta})\, s_K,$$

so that $\nabla \mathcal{L}_u(\underline{\theta}) = \sum_K \nabla \widehat{f}_{u,K}(\underline{\theta})\, m_K$ and $\widetilde{\nabla} \mathcal{L}_u(\underline{\theta}) = \sum_K \nabla \widehat{f}_{u,K}(\underline{\theta})\, S[K]$, which gives the display above. Each coordinate of $\nabla_{\underline{\theta}} f_u(\mathbf{s}; \underline{\theta})$ is of the form $-s_u s_v f_u(\mathbf{s}; \underline{\theta})$ or $-s_u f_u(\mathbf{s}; \underline{\theta})$, and $|f_u(\mathbf{s}; \underline{\theta})| \le e^{\|\underline{\theta}\|_1} \le e^{\gamma}$ on the feasible set. Since $s_K \in \{-1, 1\}$, we have the uniform bound

$$\big\|\nabla \widehat{f}_{u,K}(\underline{\theta})\big\|_\infty \le \max_{\mathbf{s} \in \{-1,1\}^{D+1}} \big\|\nabla_{\underline{\theta}} f_u(\mathbf{s}; \underline{\theta})\big\|_\infty \le e^{\gamma},$$

and therefore

$$\sum_{K \subseteq \{u\} \cup N_u} \big\|\nabla \widehat{f}_{u,K}(\underline{\theta})\big\|_\infty \le 2^{D+1} e^{\gamma}.$$

Next we bound the moment errors. On the event $\mathcal{E}$ that $\max_{K \in \mathcal{K}_E} |S[K] - m_K| \le \tau$, which has probability at least $1 - \delta$, we have for every u,

$$\varepsilon_{stat} \le 2^{D+1} e^{\gamma}\, \tau.$$

Robust gradient descent. Having established a uniform bound $\sup_{\|\underline{\theta}\|_1 \le \gamma} \|\nabla \mathcal{L}_u(\underline{\theta}) - \widetilde{\nabla} \mathcal{L}_u(\underline{\theta})\|_\infty \le \varepsilon_{stat}$, we can invoke Lemma 4 (robustness of projected gradient descent) exactly as in the proof of Theorem 1. In the notation of Lemma 4, we may take

$$\varepsilon := 4\gamma\, \varepsilon_{stat}, \qquad L := 2\sqrt{D+1}\, e^{\gamma},$$

where L bounds the ℓ_2 norm of the approximate gradient, as required by Lemma 4. Indeed, $\|\widetilde{\nabla} \mathcal{L}_u(\underline{\theta})\|_\infty \le \|\nabla \mathcal{L}_u(\underline{\theta})\|_\infty + \varepsilon_{stat} \le e^{\gamma} + \varepsilon_{stat} \le 2e^{\gamma}$, where the last step uses the choice of ε_{stat} below and $\epsilon \le 4\gamma$, which give $\varepsilon_{stat} \le 2\gamma e^{-3\gamma}/(1+\gamma) \le e^{\gamma}$; since the local parameter vector has dimension $D+1$, $\|\widetilde{\nabla} \mathcal{L}_u(\underline{\theta})\|_2 \le \sqrt{D+1}\, \|\widetilde{\nabla} \mathcal{L}_u(\underline{\theta})\|_\infty \le L$. Lemma 4 then yields the iteration complexity and step size

$$T = \left\lceil \frac{16\gamma^2 L^2}{\varepsilon^2} \right\rceil = \left\lceil \frac{L^2}{\varepsilon_{stat}^2} \right\rceil, \qquad \eta = \frac{2\gamma}{L\sqrt{T}},$$

which ensure that the averaged iterate satisfies $\mathcal{L}_u(\bar{\underline{\theta}}^T) - \mathcal{L}_u(\underline{\theta}_u^*) \le \varepsilon$.

From loss error to parameter error (curvature). Finally, we convert the loss optimality gap into a coordinate-wise parameter error bound using the same curvature argument as in the proof of Theorem 1. Let $\underline{\Delta}_u := \bar{\underline{\theta}}^T - \underline{\theta}_u^*$. Define the curvature functional

$$\delta \mathcal{L}_u^*(\bar{\underline{\theta}}^T) := \mathcal{L}_u(\bar{\underline{\theta}}^T) - \mathcal{L}_u(\underline{\theta}_u^*) - \langle \nabla \mathcal{L}_u(\underline{\theta}_u^*), \underline{\Delta}_u \rangle.$$

Since $\underline{\theta}_u^*$ is a minimizer of the convex loss over the feasible set, we have $\delta \mathcal{L}_u^*(\bar{\underline{\theta}}^T) \le \mathcal{L}_u(\bar{\underline{\theta}}^T) - \mathcal{L}_u(\underline{\theta}_u^*) \le \varepsilon$. Moreover, Lemma 7 (whose proof applies unchanged when the couplings to non-neighbors are fixed to zero) yields the lower bound

$$\delta \mathcal{L}_u^*(\bar{\underline{\theta}}^T) \ge \frac{e^{-3\gamma}}{2+2\gamma} \max_{v \in N_u} |\Delta_{u,v}|^2.$$

As in Theorem 1, this bound does not control the error on the field θ_u, so the guarantee is stated for the couplings only. Combining the two displays gives

$$\max_{v \in N_u} |\bar{\theta}_{u,v}^T - \theta_{u,v}^*| \le e^{3\gamma/2} \sqrt{2(1+\gamma)\,\varepsilon} = e^{3\gamma/2} \sqrt{8\gamma(1+\gamma)\,\varepsilon_{stat}}.$$

To achieve a target accuracy $\max_{v \in N_u} |\bar{\theta}_{u,v}^T - \theta_{u,v}^*| \le \epsilon$ with overall success probability at least $1-\delta$, it suffices to set

$$\varepsilon_{stat} \le \frac{\epsilon^2 e^{-3\gamma}}{8\gamma(1+\gamma)}$$

which, by the bound on ε_{stat} above, holds on $\mathcal{E}$ if $\tau = \frac{\epsilon^2 e^{-4\gamma}}{2^{D+4}\gamma(1+\gamma)}$. Since the table S is fixed before running the optimization and reused across all T iterations and all nodes u, the single event $\mathcal{E}$ implies the gradient approximation guarantee holds uniformly for every iterate and every node, and no additional union bound over the T steps or the p nodes is needed.

Finally, for empirical tables, each entry is an average of n i.i.d. random variables in $\{-1, 1\}$, so Hoeffding's inequality and a union bound over the $|\mathcal{K}_E| \le p\, 2^{D+1}$ entries give $\mathbb{P}[\mathcal{E}^c] \le 2p\, 2^{D+1} e^{-n\tau^2/2}$, which is at most δ when $n \ge \frac{2}{\tau^2}\big((D+1)\log 2 + \log \frac{2p}{\delta}\big)$. Substituting $\frac{2}{\tau^2} = \frac{2^{2D+9}\gamma^2(1+\gamma)^2 e^{8\gamma}}{\epsilon^4}$ gives the stated sample size.

■

H Numerical experiments

Our results are worst-case upper bounds. In this section we run Algorithm 1 on a small model to see how the moment order needed in practice depends on the ℓ_1 width, and how it compares with the order prescribed by Theorem 1. We also check whether restricting the learner to low-order empirical moments costs extra samples compared with Interaction Screening on full samples.

Setup. We use the long-range ferromagnetic chain on $p = 12$ spins with couplings $\theta^*_{uv} = J/(u-v)^2$ for all $u \neq v$, with $J > 0$, and no magnetic fields. Every spin interacts with every other spin, and the ℓ_1 width γ is proportional to J, so it can be tuned by changing J. With $p = 12$, the distribution, and hence every moment, can be computed exactly by enumerating the 2^{12} configurations. We run projected gradient descent with the approximate gradients of Algorithm 1, the ℓ_1 radius set to the true width γ, step size $\eta = 0.2$, and 3000 iterations (2000 in Experiment 2). We report the last iterate rather than the average; with these settings the iterates have converged, and running four times as many iterations does not change the reported errors. Unlike in our theorems, we use the constant step size and number of iterations above rather than the values of η and T in Theorem 1, which are tuned for the worst case. The error is $\max_{u \neq v} |\hat{\theta}_{uv} - \theta^*_{uv}|$, where row u of $\hat{\theta}$ is the estimate obtained from the loss of spin u.

Computing the approximate gradient. When every entry of the table is a moment of a single distribution q (the exact table, or an empirical table built from a shared dataset, in which case q is the empirical distribution of the samples), steps (ii) and (iii) of Algorithm 1 can be carried out in closed form. Expanding into monomials and substituting $S[K] = \mathbb{E}_q[\prod_{i \in K} \sigma_i]$ gives

$$\tilde{\nabla}_{uv} \mathcal{L}_{u,S}(\underline{\theta}_u) = -\,\mathbb{E}_q\Big[\sigma_u \sigma_v \sum_{k \leq d} \frac{(-E_u(\underline{\sigma}, \underline{\theta}_u))^k}{k!}\Big],$$

that is, the Interaction Screening gradient with e^{-E_u} replaced by its degree-d Taylor polynomial and the expectation taken under q. We use this form in our experiments, and we checked numerically that it agrees with the monomial expansion to machine precision. Note that the learner still only uses moments of order at most $D = d + 2 - (d \bmod 2)$. Taking $d = \infty$ in this formula gives Interaction Screening with full samples, which we use as a baseline.

Experiment 1: moment order versus width. Using the exact moment table, we find the smallest degree d for which the error is at most $\epsilon = 0.02$, for $\gamma \in \{0.25, 0.5, \ldots, 2.5\}$. The results are shown in Figure 1, together with the degree prescribed by Theorem 1 for the same ϵ and γ.

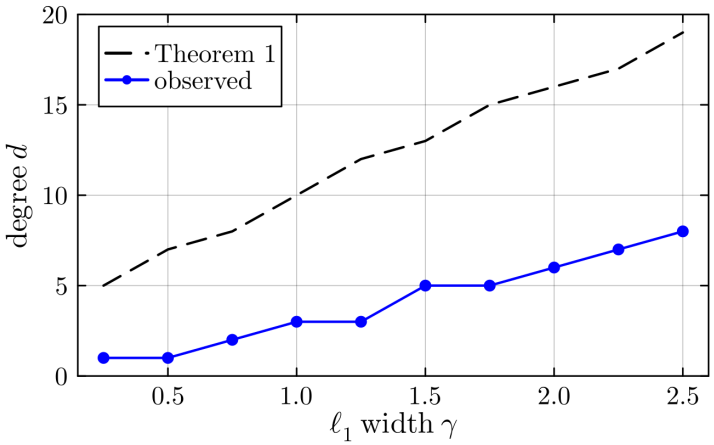

Figure 1: Smallest degree d of the Taylor approximation for which Algorithm 1, run on the exact moment table of order $d + 2 - (d \bmod 2)$, learns all couplings of the long-range chain to error 0.02 ($p = 12$). The dashed line is the degree prescribed by Theorem 1.

The required degree grows roughly linearly with γ, as predicted by Theorem 1. The prescribed degree is always larger than the observed one, by a factor between about 2.4 and 7, as expected from a worst-case bound.

Experiment 2: number of samples. We now replace the exact table by an empirical table built from a shared dataset of n i.i.d. samples, for the long-range chain with $\gamma = 1$. For this model, $d = 3$ (table order 4) is the smallest degree that reaches error 0.02 with the exact table. Figure 2 shows the error for $d \in \{1, 3, 6\}$ and for Interaction Screening with full samples ($d = \infty$), with the same

dataset used for every d in each trial. Points are medians over 10 independent datasets, and error bars show the interquartile range.

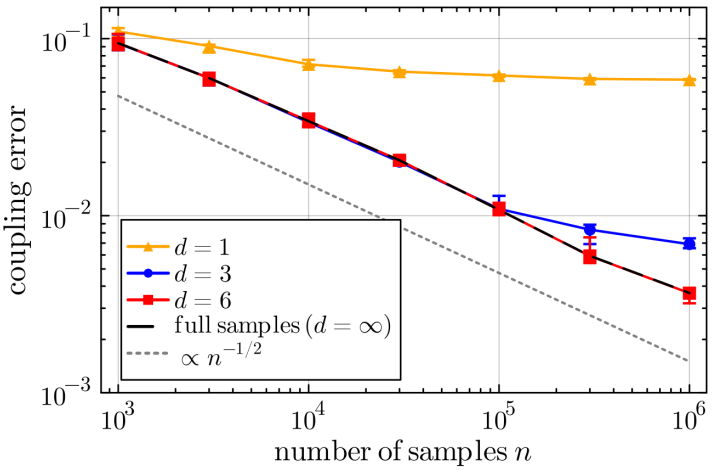

Figure 2: Coupling error of Algorithm 1 on empirical moment tables built from a shared dataset of n samples, for the long-range chain with $\gamma = 1$ ($p = 12$), compared with Interaction Screening on the full samples. The curves for $d = 6$ and for full samples coincide. Medians over 10 datasets; error bars show the interquartile range.

For $d = 6$ the error is indistinguishable from that of Interaction Screening with full samples over the whole range of n, and both decrease at the rate $n^{-1/2}$. For $d = 3$ the same holds up to $n \approx 10^5$. Beyond that point its error levels off, approaching 0.005, which is its error on the exact table. With $d = 1$ (table order 2), the error levels off much earlier, at about 0.06, which is its error on the exact table. This agrees with the main message of our results: once the moment order is large enough for the polynomial approximation error to be small, using only low-order moments costs no additional samples. The sample sizes prescribed by Corollary 1 are, as expected, very conservative. For $\gamma = 1$, $\epsilon = 0.02$ and $\delta = 0.05$ they exceed 10^{15}, whereas $n \approx 10^5$ samples suffice here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claim in the abstract is exactly supported by the main theorems in the text

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss how the corresponding lower bounds are not shown in our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Formal statements have complete set of assumptions stated

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The main results are theoretical. The only numerical experiments are in Appendix H, which specifies the model, the algorithm, all parameters, the number of trials and how the samples are generated, so that they can be reproduced.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The main results are theoretical. The only numerical experiments are in Appendix H. They are small (12 spins, computed by exact enumeration) and fully specified there; the code is not released with the paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The only numerical experiments are in Appendix H, which states the model, step size, number of iterations, ℓ_1 radius, error metric and number of trials.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The only numerical experiments are in Appendix H. Experiment 1 uses exact moments and has no randomness. In Experiment 2 we report the median over 10 independent datasets, with error bars showing the interquartile range.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The only numerical experiments are in Appendix H. Each runs on a single core of a desktop CPU in under an hour.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes it does.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no social impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No existing assets are used. Relevant prior work is cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No such study exists in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM was used for only writing assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.