Learning Gaussian Graphical Models from a Glauber Trajectory Without Mixing

Eric Shen* — eys@mit.edu — MIT

Tony Wu* — tonyyzwu@mit.edu — MIT

Mahbod Majid — mahbod@mit.edu — MIT

Ankur Moitra† — moitra@mit.edu — MIT

April 21, 2026

Abstract

We study the task of learning the structure of a d-sparse Gaussian graphical model on n variables from a single trajectory of Glauber dynamics. Beyond algorithmic considerations, many applications present temporally correlated observations rather than i.i.d. samples. Moreover, in the classical i.i.d. setting, polynomial-time structure learning from a sublinear in n number of samples is suspected to be computationally hard without additional assumptions on the precision matrix. Motivated in part by this, we design the first polynomial-time algorithm that recovers the conditional-independence graph from a single Glauber trajectory, with a trajectory-length guarantee that does not depend on the mixing time.

Technically, our algorithm has three components. First, we estimate the conditional variances and rescale the trajectory to reduce to the unit-diagonal case, without changing the underlying graph. Second, we design a local edge test that extracts adjacency information from short update windows by isolating pairwise influence. Third, we aggregate these local statistics using a robust median-based estimator, and prove accuracy despite temporal dependence arising from a single trajectory.

*Equal contribution.

†Supported in part by DARPA expMath, a Microsoft Trustworthy AI Grant, NSF-CCF 2430381, an ONR grant, and a David and Lucile Packard Fellowship.

Contents

1 Introduction

A *Gaussian Graphical Model* (GGM) on n vertices is a mean-zero Gaussian random variable $X \sim \mathcal{N}(0, \Sigma)$. The relevant object for the graph structure is the *precision matrix* $\Theta := \Sigma^{-1}$. We associate to Θ an undirected graph $G = (V, E)$ by putting an edge (i, j) whenever $\Theta_{ij} \neq 0$. The key fact is that, for Gaussians, zeros in Θ exactly encode conditional independences: for distinct $i, j \in V$,

$$X_i \perp X_j \mid X_{V \setminus \{i,j\}} \quad \Longleftrightarrow \quad \Theta_{ij} = 0\,.$$

This is known as the Markov property. An important measure of complexity of GGMs is *sparsity*: we say the GGM is d-sparse if every vertex has at most d neighbors in G, equivalently each row of Θ has at most d nonzero off-diagonal entries.

GGMs provide a natural way to represent *conditional dependence structure* among many interacting variables. The literature on GGM applications is too vast to survey here, but representative examples include neuroscience and brain connectivity [DMG+20, HLS+10], genomics [YZLM22], metabolic pathway reconstruction [KSI+11], climate science [ZFLH14], financial systemic-risk modeling [CG16], and environmental psychology [BMS+19]. A recurring regime in such applications is high dimensionality, where the number of variables n can be comparable to or larger than the number of available observations, motivating our focus on sparse GGMs, in which the conditional-independence graph has maximum degree at most d.

Algorithmically, the main challenge is already present in *structure learning*: recovering the edge set E (equivalently, the support of Θ). Indeed, once G is known, estimating the coefficients reduces to running n low-dimensional (regression) problems, each involving only the d neighbors of a node. More precisely, for each node $i \in V$, the conditional distribution of X_i given the remaining coordinates is

$$X_i = -\sum_{j \in N(i)} \frac{\Theta_{ij}}{\Theta_{ii}} X_j + \xi_i, \quad \xi_i \sim \mathcal{N}\left(0, \frac{1}{\Theta_{ii}}\right)\,.$$

Thus, given G, estimating Θ reduces to n regressions of X_i onto $\{X_j : j \in N(i)\}$, which recover the coefficients $\{-\Theta_{ij}/\Theta_{ii}\}_{j \in N(i)}$ and the noise variance $1/\Theta_{ii}$.

In the classical i.i.d. data model, Misra, Vuffray and Lokhov [MVL20] studied the information-theoretic sample complexity of learning sparse GGMs *without* assuming bounded spectrum or incoherence. The only assumption they make is the following guarantee on the minimum normalized edge strength

$$\frac{|\Theta_{ij}|}{\sqrt{\Theta_{ii}\Theta_{jj}}} \geq \alpha \qquad \forall (i, j) \in E\,, \tag{non-degeneracy}$$

which is a natural non-degeneracy condition ensuring that present edges are not arbitrarily weak. It is important to note that this constraint does *not* impose any assumptions on the minimum eigenvalue of the normalized matrix, and the spectrum may be arbitrarily ill-conditioned. For a simple demonstration of this, see appendix Section D.

They show that information-theoretically $O(d \log n/\alpha^2)$ i.i.d samples suffice for learning the graph structure. Earlier work of Wang, Wainwright, and Ramchandran [WWR10] shows that at least $\Omega(\log n/\alpha^2)$ i.i.d samples are necessary for this task, and it is currently unknown which of the upper bound or the lower bound is tight. However, the price paid is computational: the algorithm of [MVL20], uses an exhaustive search based on an ℓ_0 constrained sparse linear regression with running time $n^{\Omega(d)}$. It is further suspected and believed that for learning GGMs and the related problem of sparse linear regression, no polynomial-time algorithm exists that uses $o(n)$ many samples [Mek24]. In particular, in the fixed-design, worst-case setting, *proper* sparse linear regression—meaning the algorithm must output a k-sparse predictor when a k-sparse solution exists—is NP-hard [Nat95, ZWJ14].

In many scientific settings, observations are not i.i.d.; instead we observe a system evolving over time. A natural stylized model for such temporal dependence is a Markov chain whose stationary distribution is a GGM, for instance single-site Gibbs sampling (Glauber dynamics). If the chain mixes rapidly, then by spacing observations by at least the mixing time one can obtain approximately independent draws and reduce to the classical i.i.d. setting.

However, mixing-based reductions can be unsatisfactory even for Gaussian targets. Indeed, for a multivariate normal target, single-site Gibbs has an explicit linear-operator description, and its convergence rate is controlled by the spectrum of an associated update matrix [Ami91, RS97]. Moreover, this convergence behavior is invariant under diagonal rescaling of the coordinates, so it is expressed in terms of the normalized precision matrix $\Theta' = D^{-1/2}\Theta D^{-1/2}$, where $D = \mathrm{diag}(\Theta_{11}, \ldots, \Theta_{nn})$. For example, a standard spectral-gap bound for single-site Gibbs implies

$$t_{\mathrm{mix}}(\varepsilon) \approx n \, \frac{1}{\lambda_{\min}(\Theta')} \, \log(1/\varepsilon), \tag{1}$$

where the equality is up to absolute constants [RS97, Ami91]. Since $\lambda_{\min}(\Theta')$ can be arbitrarily small under the non-degeneracy constraint,[1] the mixing time can be arbitrarily large.

This leads to our main question:

> *Is it possible to learn the structure of a sparse GGM from a single Glauber trajectory* without *waiting for the chain to mix, and without imposing additional assumptions on the precision matrix Θ?*

We answer this question in the affirmative by giving an efficient algorithm that recovers the graph from a single trajectory, with no dependence on the mixing time and without imposing additional assumptions on the precision matrix beyond the non-degeneracy condition. In the next section we formalize the model and state our main theorem.

1.1 Results

We begin by recalling two definitions that formalize our setting: (α, d)-sparse Gaussian graphical models and single-site Glauber dynamics.

Definition 1.1 ((α, d)-sparse Gaussian graphical model)**.** Let $\Sigma \in \mathbb{R}^{n\times n}$ be positive definite and let $\Theta := \Sigma^{-1}$. For $\alpha > 0$ and an integer $d \geq 1$, we say that $X \sim \mathcal{N}(0, \Sigma)$ is an (α, d)-sparse GGM if the graph $G = (V, E)$ on $V = [n]$ defined by

$$(i, j) \in E \quad \Longleftrightarrow \quad \Theta_{ij} \neq 0$$

has maximum degree at most d, and moreover every edge has normalized strength at least α, i.e.,

$$\left| \frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} \right| \geq \alpha \qquad \forall (i, j) \in E,$$

We refer to G as the (conditional-independence) graph or the sparsity pattern of the model.

[1] See Section D for a simple demonstration of this.

Next we define the Glauber dynamics for a GGM. We work with the continuous-time dynamics, where each coordinate has an independent rate-1 Poisson clock, so the chain performs n updates per unit time in expectation.

Definition 1.2 (Continuous-time Glauber dynamics for a GGM)**.** The continuous-time Glauber dynamics for a GGM with precision matrix Θ is a random process $\{Y^{(t)}\}_{t\geq 0}$ taking values in $\mathbb{R}^n$. It is initialized at an arbitrary (possibly random or worst-case) vector $Y^{(0)}$, and is updated at random times $\{S^{(\ell)}\}_{\ell\in\mathbb{N}}$ with $S^{(0)} = 0$, where $S^{(\ell)}$ is the time of the ℓ-th update. The inter-update times $\{S^{(\ell+1)} - S^{(\ell)}\}_{\ell\geq 0}$ are i.i.d. sampled from the exponential distribution with parameter n, namely $\mathrm{Exp}(n)$, so the chain performs n updates per unit time in expectation. The process is piecewise constant between updates; define the embedded discrete-time chain $X^{(\ell)} := Y^{(t)}$ for $t \in [S^{(\ell)}, S^{(\ell+1)})$.

At each update time $S^{(\ell)}$, an index $I^{(\ell)} \in [n]$ is chosen uniformly at random. Let $i = I^{(\ell)}$. Then we resample coordinate i from its conditional distribution given the others:

$$X_i^{(\ell)} \sim \mathcal{N}\left(-\sum_{j\in N(i)} \frac{\Theta_{ij}}{\Theta_{ii}} X_j^{(\ell-1)},\ \frac{1}{\Theta_{ii}}\right),$$

and set $X_j^{(\ell)} = X_j^{(\ell-1)}$ for all $j \neq i$. Here $N(i) = \{j \neq i : \Theta_{ij} \neq 0\}$ denotes the neighborhood of i in the sparsity graph of Θ.

We are now ready to present our main result.

Theorem 1.3 (Main Theorem (Structure learning))**.** *There exists a polynomial-time algorithm which, given a Glauber trajectory of length T from an (α, d)-sparse Gaussian graphical model on n vertices, recovers the sparsity pattern (i.e., $\mathrm{supp}(\Theta)$) with probability at least $1-\delta$, provided*

$$T = \Omega\left(\frac{d^3 \log(n/\delta)}{\alpha^5}\right).$$

A few remarks are in order. First, the guarantee makes no mixing-time assumption and does not require any additional regularity conditions on Θ beyond sparsity and the edge-strength condition in Definition 1.1. Second, in the continuous-time dynamics of Definition 1.2, the chain performs n single-site updates per unit time in expectation; thus, observing the chain up to time horizon T corresponds to about nT single-site updates on average. Third (on optimality), information-theoretic lower bounds of $\Omega(\log(d)/\alpha^2)$ were previously known (see Theorem 2 in [TRD25]). Their statement is parameterized by $\beta_{\min} := \min_{(i,j)\in E} |\Theta_{ij}|/\Theta_{ii}$, whereas we work with the symmetric normalization $\min_{(i,j)\in E} |\Theta_{ij}|/\sqrt{\Theta_{ii}\Theta_{jj}}$. Under the condition that Θ has diagonal 1, these parameters become equivalent. In Section 6 (Theorem 6.4) using similar techniques, we show an improved information-theoretic lower bound that $\Omega(n\log(n)/\alpha^2)$ single-site updates are necessary for structure learning from a single trajectory (for success probability $1/2$). Note that single-site number of updates and continuous-time Glauber trajectory lengths are equivalent up to a factor of n, therefore this implies a lower bound of $\Omega(\log(n)/\alpha^2)$ on the continuous-time trajectory length.[2] Thus, while our result matches the correct logarithmic dependence, it is possible that the polynomial dependence on d and α in Theorem 1.3 is not tight and could potentially be improved.

[2] See Section B for a formal statement of this equivalence.

Theorem 1.3 follows from the more general parameter learning theorem below. Our parameter learning theorem, Theorem 1.4, learns each coordinate up to multiplicative factor ϵ. And the structure learning theorem follows from the parameter learning theorem by choosing $\epsilon = O(1)$, then outputting an edge if the estimated $\widehat{\Theta}_{ij} > \frac{\alpha}{2}$.

Theorem 1.4 (Main Theorem (Parameter learning))**.** *Let* $0 < \alpha, \delta, \varepsilon < 1$ *and let* $n, d \in \mathbb{N}$*. Given a Glauber trajectory of length*

$$T = \Omega\left(\frac{d^3 \log(n/\delta)}{\alpha^5 \varepsilon^5}\right)$$

evolving according to an (α, d)*-sparse GGM with precision matrix* Θ*, there is a polynomial-time algorithm that outputs an estimate* $\widehat{\Theta}$ *such that*

$$\left|\widehat{\Theta}_{ij} - \Theta_{ij}\right| \leq \varepsilon |\Theta_{ij}| \quad \forall i, j,$$

with success probability $1 - \delta$*.*

Finally we discuss the practicality of our algorithms.

Remark 1.5 (On Practicality). Despite our asymptotic improvements in trajectory length, constants that appear in our analysis preclude the practicality of our algorithm. We believe that these constants are not significantly improvable, and indeed, our experiments suggest that our algorithms do not have practical trajectory length requirements. We briefly explain the theoretical necessity for these large constants at the end of Section 5. Obtaining practical algorithms for this problem is an interesting direction for future work.

To see a full proof of Theorems 1.3 and 1.4 see Section 5. For a technical overview of the algorithm see Section 2.

1.2 Related work

Gaussian GGMs from Glauber dynamics. Most closely related to our work is the recent work of Tirukkonda, Rayas, and Dasarathy [TRD25], which gives the first algorithmic guarantees for structure learning in Gaussian graphical models from a single Glauber trajectory, along with information-theoretic lower bounds. In Section C we discuss a technical issue in the proof of one of their lemmas; for now, we summarize their assumptions and guarantees as stated. They impose several regularity conditions beyond a minimum edge-strength assumption. Concretely, in addition to requiring $|\Theta_{ij}|/\Theta_{ii} \geq \beta_{\min}$ for all $(i, j) \in E$, they assume (i) an upper bound $|\Theta_{ij}|/\Theta_{ii} \leq \beta_{\max}$ for all $(i, j) \in E$ (Assumption A1), (ii) bounded marginals $\Sigma_{ii} \leq \sigma^2_{\max}$ and non-degenerate conditionals $1/\Theta_{ii} \geq \sigma^2_{\min}$ (Assumption A2), and (iii) a sample decay condition $d\beta_{\max} < 1 - \Omega(1)$ (Assumption A3).[3] Meanwhile, we make no additional assumptions beyond the minimum edge-strength assumption. Most importantly, in the diagonal-1 and $d = O(1)$ setting, the last assumption implies very fast mixing via the Gershgorin circle theorem: $\lambda_{\min}(\Theta) \geq 1 - d\beta_{\max} = \Omega(1)$ and $\lambda_{\max}(\Theta) \leq 1 + d\beta_{\max} = O(1)$ (see Equation (1)). Therefore, under their assumptions one can essentially simulate approximate i.i.d. samples by observing the Glauber trajectory at constant time intervals. However, in our setting, the mixing time may be arbitrarily slow.

[3] As stated it is written as $d\beta_{\max} < 1$; however, the proof of Lemma 4 explicitly uses that this gap is at least a constant.

Learning graphical models from Glauber dynamics. A growing line of work studies structure learning when observations arise from local Markov dynamics rather than i.i.d. sampling. Bresler, Gamarnik, and Shah [BGS17] initiated this direction for discrete graphical models, showing that observing a single-site Glauber trajectory can make structure learning computationally tractable for Ising models. More recently, Gaitonde and Mossel [GM24] gave a unified analysis for learning Ising models from Glauber trajectories, and Gaitonde, Moitra, and Mossel [GMM25a] gave the first efficient algorithms in the observation model that reveals only actual configuration changes rather than all update attempts, with extensions to reversible single-site chains such as Metropolis. In a different direction, Gaitonde, Moitra, and Mossel [GMM25b] show that for higher-order Markov random fields, access to Glauber dynamics trajectories can bypass i.i.d. hardness barriers (e.g., noisy parity), yielding efficient learning.

Technically, our algorithm is close in spirit to several methods in this line of work: we examine short windows of the trajectory to probe the interaction between a candidate pair (i, j), and we exploit sparsity to ensure that with sufficiently large probability no neighbor of i or j updates within the window, allowing the local effect of (i, j) to be isolated from confounding updates. That said, structure learning for GGMs differs in important ways from the Ising setting. First, the variables are continuous, so one cannot rely on discrete indicator statistics present in previous work whose empirical averages directly estimate event probabilities. Second, in the Gaussian case a key difficulty is *anti-concentration*: to detect an edge (i, j) one needs the neighbor's influence on the conditional mean to be typically non-negligible. Unlike the Ising setting—where boundedness and discrete concentration can often be leveraged—Gaussian anti-concentration depends on the scale of the conditional variance (equivalently, on Θ_{ii}) and can degrade without additional regularity beyond (α, d). As a result, while the high-level philosophy is shared, the Gaussian case requires new technical ideas.

2 Technical overview

At a high level, our algorithm has three ingredients. First, we estimate the diagonal entries Θ_{ii} from short windows of the trajectory and use them to normalize the model so that the precision matrix has (approximately) unit diagonal. Second, for each candidate edge (i, j), we look for short update patterns that isolate the influence of j on a later update of i. Third, because we cannot directly tell whether hidden neighbor updates occurred inside a window, we treat such windows as contaminated and aggregate many windows using robust median-based estimators.

We begin with diagonal estimation because it already contains the main ideas of the full algorithm: short informative windows, unobserved contamination, and robust aggregation. We then explain how to estimate off-diagonal entries, and finally describe a simpler variant that is available when the chain is allowed to mix between samples.

Two nearby i-updates reveal Θ_{ii}. Suppose that at some point in the trajectory the ith coordinate updates twice, with no update to any neighbor of i in between. Write the corresponding states as

$$X^{(0)} \xrightarrow{i} X^{(1)} \xrightarrow{i} X^{(2)}.$$

Since no neighbor of i changes between the two i-updates, both updates use the same conditional mean. Hence

$$X_i^{(1)} - X_i^{(2)} \sim \mathcal{N}\left(0, \frac{2}{\Theta_{ii}}\right).$$

So every such window gives a sample whose variance is exactly $2/\Theta_{ii}$. This turns diagonal estimation into a one-dimensional robust variance-estimation problem.

The "ii" pattern and hidden contamination. Exact consecutive "ii" updates are too rare to be useful on the trajectory lengths we target. Instead, we divide the trajectory into short windows of length T and keep every window that contains at least two updates to coordinate i. A window is *clean* if, in addition, no neighbor of i updates inside that window. On a clean window, the same calculation as above shows that the difference between the two updated values of coordinate i is distributed as $\mathcal{N}(0, 2/\Theta_{ii})$.

The difficulty is that cleanliness depends on the unknown neighborhood of i, so we cannot test it directly. We therefore keep all windows with two i-updates and view the non-clean ones as contaminated samples. Because the window is short and the graph is d-sparse, the contamination rate is small. Taking the median absolute deviation of the resulting samples gives a robust estimate of Θ_{ii}.

Normalization. Using a small initial portion of the trajectory, we estimate every diagonal entry Θ_{ii}. We then rescale coordinate i by $\sqrt{\Theta_{ii}}$ (or, in the actual algorithm, by its estimate), replacing each state $X \in \mathbb{R}^n$ with X' defined by

$$X'_i = \sqrt{\Theta_{ii}}\, X_i.$$

Under exact normalization, X' is again a Glauber trajectory, now for a precision matrix with diagonal entries equal to 1. With estimated diagonals, the normalized trajectory has diagonal entries within $1 \pm \varepsilon$, and the later analysis is stable to this approximation. For the overview, it is therefore convenient to assume from this point onward that the diagonal is exactly 1.

A naive "iji" test for Θ_{ij}. To estimate an off-diagonal entry Θ_{ij}, the most natural idea is to look for a short window of the form

$$X^{(1)} \xrightarrow{i} X^{(2)} \xrightarrow{j} X^{(3)} \xrightarrow{i} X^{(4)},$$

again with no neighbor of i or j updating inside the window. On such a clean window,

$$\begin{aligned} X_i^{(2)} &= \sum_{k \neq i} -\Theta_{ik} X_k^{(1)} + \varepsilon^{(2)}, \\ X_i^{(4)} &= \sum_{k \neq i} -\Theta_{ik} X_k^{(3)} + \varepsilon^{(4)}, \end{aligned}$$

with $\varepsilon^{(2)}, \varepsilon^{(4)} \sim \mathcal{N}(0, 1)$ i.i.d. The only relevant change between the two conditional means is the value of coordinate j, so

$$X_i^{(4)} - X_i^{(2)} = -\Theta_{ij}\big(X_j^{(3)} - X_j^{(1)}\big) + \varepsilon^{(4)} - \varepsilon^{(2)}.$$

This suggests estimating Θ_{ij} from the ratio

$$\frac{X_i^{(4)} - X_i^{(2)}}{X_j^{(3)} - X_j^{(1)}}.$$

To keep the denominator well behaved, we further condition on $|X_j^{(3)} - X_j^{(1)}| > c$ for a fixed constant $c > 0$.

Why the "iji" test fails. The problem is a subtle dependence issue. In the "iji" pattern, the middle j-update depends on the value produced by the first i-update. As a result, the noise term from the first i-update is not independent of the denominator $X_j^{(3)} - X_j^{(1)}$. So the ratio above is not centered in the simple way suggested by the heuristic calculation. This is the main obstacle that forces us away from the naive "iji" statistic.

The "iiji" pattern breaks the dependence. To remove this dependence, we prepend one extra i-update and instead search for windows of the form

$$X^{(0)} \xrightarrow{i} X^{(1)} \xrightarrow{i} X^{(2)} \xrightarrow{j} X^{(3)} \xrightarrow{i} X^{(4)},$$

again requiring that no neighbor of i or j updates inside the window. Conditioned on $X^{(0)}$, the extra i-update acts as a refresh step: the second i-update is independent from the first one, so the information introduced at $X^{(1)}$ is washed out before the j-update occurs. So, conditioned on $X^{(0)}$, the noise in the first i-update is therefore independent of the later j-state. This yields the relation

$$X_i^{(4)} - X_i^{(1)} = -\Theta_{ij}\big(X_j^{(3)} - X_j^{(0)}\big) + \varepsilon^{(4)} - \varepsilon^{(1)},$$

where $\varepsilon^{(1)}, \varepsilon^{(4)} \sim \mathcal{N}(0,1)$ are independent. After conditioning on

$$\big|X_j^{(3)} - X_j^{(0)}\big| > c,$$

the ratio statistic

$$\frac{X_i^{(4)} - X_i^{(1)}}{X_j^{(3)} - X_j^{(0)}} = -\Theta_{ij} + \frac{\varepsilon^{(4)} - \varepsilon^{(1)}}{X_j^{(3)} - X_j^{(0)}}$$

is centered at $-\Theta_{ij}$ and has variance bounded by a constant. Thus each clean "iiji" window gives a noisy but informative estimate of Θ_{ij}.

Robust aggregation. As in the diagonal-estimation step, we cannot directly verify that a window is clean, because the unknown graph determines which coordinates count as neighbors. We therefore collect all windows exhibiting the visible pattern and treat windows with hidden neighbor updates as contaminated. Moreover, the variance of the ratio statistic can vary from one accepted window to another, and successive windows are not independent because they come from a single trajectory. Nevertheless, the dependence is structured enough that martingale concentration can be used in place of the usual independent-sample Chernoff argument. This shows that the sample median remains a robust estimator of the common location parameter $-\Theta_{ij}$. Estimating every Θ_{ij} to additive error $O(\alpha)$ is then enough to recover the graph structure. With tighter parameter settings, the same framework also yields multiplicative estimation of the full precision matrix.

Information-theoretic lower bound. Finally, we prove an information-theoretic lower bound showing that logarithmic trajectory length is unavoidable even without computational constraints. We construct a family of GGMs on $2n$ vertices whose graphs are disjoint unions of edges, with each candidate obtained by deleting one edge from a perfect matching. The resulting trajectories are hard to distinguish from one another. By upper-bounding the pairwise KL divergence and applying Fano's inequality, we obtain a lower bound on the number of Glauber updates required for structure learning.

This is different from the approach of [TRD25] and improves upon their result. Instead of studying classes of graphs, we focus specifically on only the set of perfect matching graphs with edits. But our analysis mainly differs by using the exact KL divergence between Gaussians on the expected behavior of the models rather than using KL Divergence of truncated Gaussians conditioning on the trajectory having small updates. Ultimately, we improve upon the result by a factor of $\log(n)$, which differs only from the upper bound by $\mathrm{Poly}(d, \frac{1}{\alpha})$.

3 Preliminaries

3.1 Continuous-time Glauber dynamics

We consider Glauber dynamics in the continuous-time setting. The initial configuration is arbitrary, and each coordinate $i = 1, \ldots, n$ updates according to an independent Poisson process of rate 1. The following claim gives estimates on how frequently these Poisson processes update:

Lemma 3.1 (Lemma 3.1 in [GMM25b])**.** *Given an interval $I \subset \mathbb{R}_{\geq 0}$ of length T, let U_I denote the set of coordinates that update in I. For any $S \subset [n]$ with $|S| = \ell$, we have*

$$\mathbb{P}(S \cap U_I = \varnothing) = \exp(-T\ell)$$
$$\textit{and} \quad \mathbb{P}(S \subseteq U_I) \geq 1 - \ell \exp(-T).$$

Proof. For each Poisson process Π of rate 1 and interval $I \subseteq \mathbb{R}_{\geq 0}$, we have $\mathbb{P}(\Pi \cap I = \varnothing) = \exp(-|I|)$, whence the claimed bounds readily follow. □

3.2 Observing patterns

Throughout the paper, we look for intervals containing a specific "pattern," which we concretely define below:

Definition 3.2 (Pattern exhibition and strictness)**.** A *pattern* of length k is a sequence of indices $\mathcal{P} = (i_1, i_2, \ldots, i_k) \in \{1, \ldots, n\}^k$. Consider a pattern P and an interval $I = [t, t + T)$ of a continuous-time single-site update Glauber trajectory. Let $I_j = [t + (j-1)T/k, t + jT/k)$ for $j = 1, \ldots, k$. We say I *exhibits* pattern P if index i_j updates in I_j for each $j = 1, \ldots, k$. Additionally, we say I is *strict* with respect to P if, for each $j = 1, \ldots, k$, no neighbors of any index in P besides i_j updates in I_j.

We now bound the probabilities that given intervals exhibit and are strict with respect to patterns.

Lemma 3.3. *Let I be an interval of length T, and let P be a pattern containing ℓ distinct indices. Then with probability at least $\exp(-T\ell d)$, I is strict with respect to P.*

Proof. Each index in P has at most d neighbors, so we are forbidding at most Pd neighbors from updating in each I_j. By Lemma 3.1, each I_j avoids its respective indices with probability at least $\exp(-T/k \cdot \ell d)$.

The events in the k subintervals are independent, so the probability this holds for all k subintervals is at least $\exp(-T\ell d)$. □

Lemma 3.4. *Let I be an interval of length $T \leq 1/3$, and let P be a pattern of length k. Then with probability at least probability at least $T^k/2k^k$, I exhibits P.*

Proof. By Lemma 3.1,

$$\mathbb{P}(I \text{ exhibits } P) \geq (1 - \exp(-T/k))^k \geq \frac{T^k}{k^k} \exp(-T) \geq \frac{T^k}{2k^k},$$

where we use the estimates $1 - \exp(-x) \geq x \exp(-x)$ and $T \leq 1/3$. □

3.3 Corruption and robust estimation

We use several robust estimators in this paper. We work with the following corruption model. This corruption has been referred to as

Definition 3.5 (η-corruption)**.** Let $\mathcal{D}$ be a distribution, and let $X_1, \ldots, X_n \sim \mathcal{D}$ be i.i.d. samples. We say that the observed samples $\widetilde{X}_1, \ldots, \widetilde{X}_n$ are *η-corrupted* if there exist i.i.d. Bernoulli(η) random variables $\xi_1, \ldots, \xi_n$ such that for each $i = 1, \ldots, n$,

$$\widetilde{X}_i = \begin{cases} X_i, & \text{if } \xi_i = 0 \\ A_i, & \text{if } \xi_i = 1, \end{cases}$$

where each A_i is an arbitrary value. The corrupted value at index j, denoted as $\tilde{X}_j = A_i$ may be chosen adversarially, may depend on the entire trajectory up to the corrupted sample $\tilde{X}_j$. and need not be drawn from any distribution.[4]

Note that our corruption model differs from the Huber contamination model, as corrupted entries may be selected adversarially. It also differs from the strong contamination model in that each sample is independently corrupted with probability η. Further, as Glauber trajectories are sampled sequentially, corrupted samples in our corruption model may only depend on prior samples.

Under this corruption model, we use the following robust estimator for variance.

Lemma 3.6 (Robust variance estimation)**.** *Let $0 < \eta \leq 1/10$ and $0 < \delta < 1$. There is a linear-time algorithm that takes n samples from $\mathcal{N}(0, \sigma^2)$, an η-fraction of which are corrupted (as in Definition 3.5), and outputs $\widehat{\sigma}$ such that $|\widehat{\sigma} - \sigma| < 5\eta\sigma$ with probability $1 - \delta$, provided*

$$n \geq \frac{2\log(4/\delta)}{\eta^2}.$$

We also use the following mean estimator, which allows for another source of adversarial control in the choice of the variances.

Lemma 3.7 (Robust mean estimation)**.** *Let $0 < \eta \leq 1/10$ and $0 < \delta < 1$. We are given samples of the form $x^{(\ell)} = \mu + \sigma^{(\ell)}\varepsilon^{(\ell)}$, for $\ell = 1, \ldots, n$, an η-fraction of which are corrupted, where $\varepsilon^{(\ell)} \sim \mathcal{N}(0, 1)$ are i.i.d. and $\sigma^{(\ell)}$ are chosen adversarially with knowledge of $\varepsilon^{(i)}$ and $\sigma^{(i)}$ if and only if $i < \ell$, so that $|\sigma^{(\ell)}| < 1$. Then the sample median $\widehat{\mu}$ satisfies $|\widehat{\mu} - \mu| < 5\eta$ with probability $1 - \delta$, provided*

$$n \geq \frac{2\log(2/\delta)}{\eta^2}.$$

Note that Lemma 3.6 remains valid even when the corrupted samples are chosen adversarially with knowledge of the full trajectory of iterates, whereas Lemma 3.7 does not.

The proofs of these estimators are provided in Section A.

[4] This corruption model is closely related to what some recent work calls *malicious noise* or *strong malicious noise*; see, e.g., [BHMS26].

3.4 Concentration inequalities

In the proof of Lemmas 3.6 and 3.7 we will make use of the following concentration inequalities.

Fact 3.8 (Dvoretzky–Kiefer–Wolfowitz)**.** *Let $Y_1, \dots, Y_n$ be i.i.d. real-valued random variables with cdf F, and let*

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i \le t\}$$

be the empirical cdf. Then for every $\varepsilon > 0$,

$$\Pr\left(\sup_{t\in\mathbb{R}} \left|F_n(t) - F(t)\right| > \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

Fact 3.9 (Azuma–Hoeffding)**.** *Let $(M_k, \mathcal{F}_k)_{k=0}^{n}$ be a martingale, and suppose that for each $k = 1, \dots, n$,*

$$M_k - M_{k-1} \in [a_k, b_k] \qquad \textit{almost surely}.$$

Then for every $t > 0$,

$$\Pr(M_n - M_0 \ge t) \le \exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right).$$

In particular, if each increment lies in an interval of length at most 1, then

$$\Pr(M_n - M_0 \ge t) \le e^{-2t^2/n}.$$

3.5 Mixing and coupling

Definition 3.10 (Mixing and mixed chains)**.** A position X_i of the Markov chain $X_0, X_1, \dots$ is *ε-mixed* if X_i has TV-distance at most ε from its stationary distribution π. We define the *mixing time* by

$$t_{\text{mix}}(\varepsilon) = \min_t \left\{ t \ge 0 : \max_{x\in S} d_{\text{TV}}(X_t \mid X_0 = x, \pi) \le \varepsilon \right\}.$$

The stationary distribution of a Glauber trajectory with covariance matrix Σ is $\mathcal{N}(0, \Sigma)$. In fact, we show the following stronger claim.

Lemma 3.11. *Fix a coordinate i. If $x \sim \mathcal{N}(0, \Sigma)$, then after a single Glauber update to the ith coordinate, the resulting position x' still satisfies $x' \sim \mathcal{N}(0, \Sigma)$.*

Proof. Without loss of generality we flip the first coordinate; write precision and covariance in block matrix form

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{1,-1} \\ \Sigma_{-1,1} & \Sigma_{-1,-1} \end{pmatrix},$$

$$\text{and} \quad \Theta = \begin{pmatrix} \Theta_{11} & \Theta_{1,-1} \\ \Theta_{-1,1} & \Theta_{-1,-1} \end{pmatrix}.$$

Write the update as

$$x'_{-1} = x_{-1} \quad \text{and} \quad x'_1 = -\frac{\Theta_{1,-1}}{\Theta_{11}} x_{-1} + \mathcal{N}\left(0, \frac{1}{\Theta_{11}}\right).$$

Then (x_1', x_{-1}') is still Gaussian with mean 0. Further note

$$\begin{aligned}\mathrm{Cov}(x_1', x_{-1}') &= -\frac{\Theta_{1,-1}}{\Theta_{11}}\Sigma_{-1,-1} = \Sigma_{1,-1} \\ \text{and} \quad \mathrm{Var}(x_1') &= \frac{1}{\Theta_{11}} + \frac{\Theta_{1,-1}}{\Theta_{11}}\Sigma_{-1,-1}\frac{\Theta_{-1,1}}{\Theta_{11}} = \Sigma_{11}\end{aligned}$$

by Schur's complement. □

From here, we may use the coupling lemma to derive a similar statement about mixing.

Fact 3.12 (Coupling lemma)**.** *Let μ and η be distributions over $\mathbb{R}^n$. Then for any coupling ω of (μ, η) for which*

$$\mathbb{P}(X \neq Y) \geq d_{\mathrm{TV}}(\mu, \eta)$$

for $(X, Y) \sim \omega$, and equality occurs.

Corollary 3.13. *Let μ and η be distributions. Suppose for some function f, it holds that $X \in \mu$ implies $f(X) \in \eta$. Then, if Y is sampled from a distribution with TV-distance ε from μ, then $f(Y)$ follows a distribution with TV-distance at most ε from η.*

Proof. Consider the coupling achieving equality in Fact 3.12, so $\mathbb{P}(X \neq Y) = \varepsilon$. Then, under this coupling, $\mathbb{P}(f(X) \neq f(Y)) \leq \varepsilon$, which implies the desired by Fact 3.12. □

Corollary 3.14. *Fix a coordinate i. If a Glauber trajectory is ε-mixed (as in Definition 3.10), and undergoes a Glauber update to the ith coordinate, it remains ε-mixed.*

Proof. Combine Lemma 3.11 and Corollary 3.13. □

4 Reduction to normalized Gaussian graphical models

In this section we design an algorithm that reduces the problem of learning an arbitrary sparse Gaussian graphical model from a Glauber trajectory to the case where the diagonal entries of the precision matrix are 1.

To this end, we first estimate the diagonal entries up to a multiplicative factor $(1 \pm O(\varepsilon))$. We then scale the Glauber trajectory $X^{(1)}, X^{(2)}, \ldots$ by

$$X_i'^{(k)} := \sqrt{\Theta_{ii}} \cdot X_i^{(k)} \quad \forall i, k,$$

which we show is itself a Glauber trajectory. Using our estimates for Θ_{ii}, we then have a coordinate-wise $(1 \pm O(\varepsilon))$-approximation of this scaled trajectory, which also happens to be a Glauber trajectory itself.

Our main result for this section is as follows:

Theorem 4.1 (Normalization)**.** *Let $0 < \alpha, \delta < 1$ and $0 < \varepsilon \leq 1/10$, and let $n, d \in \mathbb{N}$. Given a Glauber trajectory of length*

$$T = \frac{800d\log(8n/\delta)}{\varepsilon^3} + T_{\mathrm{rest}},$$

with $T_{\mathrm{rest}} \geq 0$, evolving according to an (α, d)-sparse GGM with precision matrix Θ, there is a polynomial-time algorithm that outputs an estimate D for $\mathrm{diag}(\Theta)$ with $|D_i - \Theta_{ii}| \leq \varepsilon\Theta_{ii}$ and a trajectory of length T_{rest} evolving according to a GGM with precision matrix $\widehat{\Theta} = D^{-1/2}\Theta D^{-1/2}$ with probability $1 - \delta$.

Our main technical lemma in this section is the following lemma. We show we may estimate diagonal entries of the precision matrix via a one-dimensional robust variance estimator.

Lemma 4.2. *For fixed i, given a Glauber trajectory of length*

$$T \geq \frac{800d \log(8/\delta)}{\varepsilon^3}$$

we may retrieve in polynomial time an estimate D_i for Θ_{ii} with

$$\left| \frac{1}{\sqrt{D_i}} - \frac{1}{\sqrt{\Theta_{ii}}} \right| \leq \frac{\varepsilon}{\sqrt{\Theta_{ii}}}$$

with probability $1 - \delta$.

Our overall strategy for estimating Θ_{ii} is to observe the "ii" pattern, consisting of two updates to the ith coordinate in close proximity. Under ideal conditions, we know from Definition 1.2 that these two updated ith coordinates are sampled from the same distribution with variance $1/\Theta_{ii}$. We then consider the difference between these two updated coordinates and use the robust variance estimator to estimate the variance of this difference.

4.1 Properties of the statistic

Fix i. Let $T \leq 1/3$, and let I_t denote the time interval $[t, t+T)$. Let $\mathcal{A}^{(t)}$ denote the event that I_t is strict with respect to (i, i), and $\mathcal{B}^{(t)}$ the event I_t exhibits (i, i).

If $\mathcal{A}^{(t)}$ and $\mathcal{B}^{(t)}$ both hold, let $Y^{(0)}, Y^{(1)}, Y^{(2)}$ denote the position at times t, $t + T/2$, $t + T$ respectively.

Lemma 4.3. *We have*

$$Y_i^{(1)} - Y_i^{(2)} \mid \mathcal{A}^{(t)}, \mathcal{B}^{(t)} \sim N\left(0, \frac{2}{\Theta_{ii}}\right).$$

Proof. By Definition 1.2, we have

$$\begin{aligned} Y_i^{(1)} &= -\sum_{j \neq i} -\frac{\Theta_{ij}}{\Theta_{ii}} Y_j^{(0)} + \varepsilon_1 \\ \text{and} \quad Y_i^{(2)} &= -\sum_{j \neq i} -\frac{\Theta_{ij}}{\Theta_{ii}} Y_j^{(1)} + \varepsilon_2, \end{aligned}$$

where $\varepsilon_1, \varepsilon_2 \sim N(0, \frac{1}{\Theta_{ii}})$ are i.i.d.

If $\mathcal{A}^{(t)}$ holds, $Y_j^{(1)} = Y_j^{(2)}$ for all $j \neq i$, so

$$Y_i^{(1)} - Y_i^{(2)} \mid Y^{(0)}, \mathcal{A}^{(t)}, \mathcal{B}^{(t)} = \varepsilon_1 - \varepsilon_2 \sim N\left(0, \frac{2}{\Theta_{ii}}\right),$$

and the dependence on $Y^{(0)}$ may be dropped. □

4.2 Retrieving the estimator

We divide our sample into M intervals of length T, discarding intervals that do not satisfy $\mathcal{B}^{(t)}$, sampling $Y_i^{(1)} - Y_i^{(2)}$ from the rest, and treating intervals that fail to satisfy $\mathcal{A}^{(t)}$ as corruption in order to predict $1/\Theta_{ii}$ with robust one-dimensional variance estimation.

Since the Poisson processes updating each coordinate are independent, the events $\mathcal{A}^{(t)}$ and $\mathcal{B}^{(t)}$ are independent, so our corruption rate is just $\eta = 1 - \exp(-dT)$ by Lemma 3.3.

Lemma 4.4. *In a Glauber dynamic of length MT, in at least $\frac{1}{16}MT^2$ intervals $\mathcal{B}^{(t)}$ holds with probability at least $1 - \exp(-\frac{1}{64}MT^2)$.*

Proof. From Lemma 3.4, the event $\mathcal{B}$ occurring for the ith interval is indicated by $\mathbb{1}_i = \mathrm{Ber}(T^2/8)$, so from the Chernoff bound

$$\mathbb{P}\left(\sum_{i=1}^{M} \mathbb{1}_i \leq \frac{1}{2}\frac{MT^2}{8}\right) \leq \exp\left(-\frac{1}{8}\frac{MT^2}{8}\right)$$

and the desired follows. □

Proof of Lemma 4.2. In the above, choose

$$T = \frac{\varepsilon}{d} \quad \text{and} \quad MT^2 = \frac{800\log(8/\delta)}{\varepsilon^2},$$

Note the corruption rate satisfies $\eta = 1 - \exp(-\varepsilon) \leq \varepsilon$. Now note from the weak bound $\frac{1}{64}MT^2 > \log(2/\delta)$, we have

$$\exp\left(-\frac{1}{64}MT^2\right) < \delta/2,$$

so by Lemma 4.4, with probability $1 - \delta/2$ we see $50\log(8/\delta)/\varepsilon^2$ samples.

Then, by Lemma 3.6, we can estimate $\sqrt{2/\Theta_{ii}}$ within a factor of $1 \pm \varepsilon$ with success probability $1 - \delta/2$. Taking a union bound gives the desired success rate.

Finally, the required length of the trajectory is given by

$$MT = \frac{800d\log(8/\delta)}{\varepsilon^3}.$$

□

Finally, we derive estimates for all n diagonal entries Θ_{ii}.

Corollary 4.5. *Given a Glauber dynamic of length*

$$T = \frac{800d\log(8n/\delta)}{\varepsilon^3},$$

we may retrieve in polynomial time estimates D_i for each i such that with probability at least $1 - \delta$,

$$\left|\frac{1}{\sqrt{D_i}} - \frac{1}{\sqrt{\Theta_{ii}}}\right| \leq \frac{\varepsilon}{\sqrt{\Theta_{ii}}} \quad \forall i = 1, \ldots, n.$$

Proof. Apply Lemma 4.2 with error δ/n. Then, the probability of any error among the n estimates is δ by union bound. □

Algorithm 1 $D = \text{EstimateDiagonal}(n, d, \alpha, \varepsilon, \delta)$

Let
$$T = \frac{\varepsilon}{d} \quad \text{and} \quad M = \frac{800 d^2 \log(8n/\delta)}{\varepsilon^4}.$$
Observe Glauber trajectory of length MT, split into M intervals $I_1, \ldots, I_M$ of length T.
for $i = 1, \ldots, n$ **do**
  Let $S = \{1 \le t \le M : \mathcal{B}^{(t)}\}$, where $\mathcal{B}^{(t)}$ is defined in §4.1.
  if $|S| < \frac{1}{16} MT^2$ **then**
    Return $\perp$.
  else
    By Lemma 3.6, estimate the variance of $\widehat{\sigma}^2$ of $\{Y_i^{(1)} - Y_i^{(2)} \mid t \in S\}$, where $Y_i^{(1)}$ and $Y_i^{(2)}$ are defined in §4.1
    Set $D_i = 2/\widehat{\sigma}^2$.
  end if
end for
Return D.

4.3 Normalizing the Gaussian

Let $X^{(1)}, X^{(2)}, \ldots$ denote the updates from the Glauber dynamic. As earlier, define $X'^{(1)}, X'^{(2)}, \ldots$ so that

$$X_i'^{(k)} := \sqrt{\Theta_{ii}} \cdot X_i^{(k)} \quad \forall i, k.$$

Lemma 4.6. *The above $X'^{(1)}, X'^{(2)}, \ldots$ is a Glauber trajectory with precision matrix $\Theta' = \Theta_{\text{diag}}^{-1/2} \Theta \Theta_{\text{diag}}^{-1/2}$.*

Proof. Note
$$\Theta'_{ij} = \frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}.$$

We may directly check that Definition 1.2 is preserved. We have

$$\begin{aligned} x_i' &\mapsto \sqrt{\Theta_{ii}} \left(-\sum_{j=1}^{n} \frac{\Theta_{ij}}{\Theta_{ii}} \frac{x_j'}{\sqrt{\Theta_{jj}}} + N\left(0, \frac{1}{\Theta_{ii}}\right) \right) \\ &= -\sum_{j=1}^{n} \frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} x_j' + N(0,1). \end{aligned}$$ □

To prove our main theorem, we consider the trajectory $\widehat{X}^{(1)}, \widehat{X}^{(2)}, \ldots$ defined by

$$\widehat{X}_i^{(k)} = \sqrt{D_i} \cdot X_i^{(k)} \quad \forall i, k.$$

Proof of Theorem 4.1. It suffices to show $\widehat{X}$ is a Glauber dynamic with precision matrix $\widehat{\Theta} = D^{-1/2} \Theta D^{-1/2}$, i.e.

$$\widehat{\Theta}_{ij} = \frac{\Theta_{ij}}{\sqrt{D_i D_j}}.$$

Again, we check Definition 1.2 is preserved. We have

$$\begin{aligned}\widehat{x}_i \mapsto \sqrt{D_i}\left(-\sum_{j=1}^{n} \frac{\Theta_{ij}}{\Theta_{ii}} \frac{\widehat{x}_j}{\sqrt{D_j}} + N\left(0, \frac{1}{\Theta_{ii}}\right)\right) \\ = -\sum_{j=1}^{n} \frac{\widehat{\Theta}_{ij}}{\widehat{\Theta}_{ii}} \widehat{x}_j + N\left(0, \frac{1}{\widehat{\Theta}_{ii}}\right).\end{aligned}$$

Finally, for independence reasons, we use only the first $T = \frac{800d \log(8n/\delta)}{\varepsilon^3}$ updates of the trajectory to estimate D, and truncate that half to produce the trajectory with the desired properties. □

In particular, note that by construction, $\widehat{\Theta}$ is coordinate-wise a $(1 \pm O(\varepsilon))$-approximation for Θ' (defined in Lemma 4.6).

Concretely,

Corollary 4.7. *There exist $c_1, \ldots, c_n \in 1 \pm \varepsilon$ so that for each i and k, we have $\widehat{X}_i^{(k)} = \frac{1}{c_i} X_i'^{(k)}$; moreover $\widehat{\Theta}_{ij} = c_i c_j \Theta'_{ij}$ for each i and j.*

Proof. Indeed

$$\frac{X_i'^{(k)}}{\widehat{X}_i^{(k)}} = \sqrt{\frac{\Theta_{ii}}{D_i}} =: c_i \in 1 \pm \varepsilon,$$

and further

$$\frac{\widehat{\Theta}_{ij}}{\Theta'_{ij}} = \frac{\sqrt{\Theta_{ii}\Theta_{jj}}}{\sqrt{D_i D_j}} = c_i c_j.$$ □

Thus, the trajectory $\widehat{X}$ we constructed has two properties:

- it is itself a Glauber trajectory with diagonal entries $1 \pm O(\varepsilon)$ (and precision matrix $\widehat{\Theta}$);
- it is coordinate-wise a $(1 \pm \varepsilon)$-approximation for a Glauber trajectory with diagonal entries 1 (and precision matrix Θ').

5 Main algorithm

We now move to structure-learning. Our main result is as follows:

Theorem 5.1 (Theorem 1.3)**.** *Let $0 < \alpha, \delta < 1$ and $n, d \in \mathbb{N}$. Given a Glauber trajectory of length*

$$T \geq \frac{4347095808 \cdot d^3 \log(4n/\delta)}{\alpha^5}$$

evolving according to an (α, d)-sparse GGM with precision matrix Θ, there is a polynomial-time algorithm that correctly outputs whether $i \sim j$ for each i and j with probability $1 - \delta$.

We also derive the following parameter-learning guarantees:

Corollary 5.2 (Theorem 1.4)**.** *Let $0 < \alpha, \delta, \varepsilon < 1$, and let $n, d \in \mathbb{N}$. Given a Glauber trajectory of length*

$$T = \Omega\left(\frac{d^3 \log(n/\delta)}{\alpha^5 \varepsilon^5}\right)$$

evolving according to an (α, d)-sparse GGM with precision matrix Θ, there is a polynomial time algorithm that outputs an estimate $\widehat{\Theta}$ for Θ such that

$$\left|\widehat{\Theta}_{ij} - \Theta_{ij}\right| \leq \varepsilon \left|\Theta_{ij}\right|$$

for each i and j.

To this end, we fix i and j, and evaluate the existence of each edge $i \sim j$ individually.

Lemma 5.3. *For fixed i and j, given a Glauber trajectory of length*

$$T \geq \frac{2173547904 \cdot d^3 \log(16/\delta)}{\alpha^5},$$

we may determine whether $i \sim j$ in polynomial time with probability $1 - \delta$.

To give a general overview of our main strategy, consider observing the "iji" pattern, consisting of an update to the ith coordinate, followed by an update to the jth coordinate, followed by an update to the ith coordinate in close proximity. If Δ_i denotes the difference in the ith coordinate after the two i updates, and Δ_j denotes the difference in the jth coordinate before and after the j update, we find that under ideal conditions,

$$\Delta_i \sim -\Delta_j \Theta_{ij} + \mathcal{N}(0, 2),$$

from which we perform a "linear regression" to estimate Θ_{ij}.

For dependency reasons, we instead consider the "$iiji$" update, and due to interdependencies between the instances of the "$iiji$" patterns we observe, a standard linear regression does not work. Instead, we show that we may condition on $|\Delta_j| > c$ without significant harm to the other parameters in the regression problem, and then perform mean estimation on

$$\frac{\Delta_i}{\Delta_j} \sim -\Theta_{ij} + \mathcal{N}\left(0, \frac{2}{\Delta_j^2}\right).$$

As mentioned, there is a dependency structure between observations, and of course, due to neighbor updates, there is some ε-corruption in our samples. To this end, we employ Lemma 3.7 to robustly estimate the mean under adversarial conditions.

5.1 Properties of the statistic

Fix i and j. Let $T \leq 1/3$, and let I_t denote the time interval $[t, t+T)$. Let $C^{(t)}$ denote the event that I_t is strict with respect to (i, i, j, i), and $\mathcal{D}^{(t)}$ the event I_t exhibits (i, i, j, i).

Normalize the Glauber trajectory by Theorem 4.1, and let $\widehat{X}$ consist of $(1 \pm \varepsilon)$ coordinate-wise approximations of X'. We first analyze the normalized Glauber dynamic X'.

Conditioned on $C^{(t)}$ and $\mathcal{D}^{(t)}$, let $Y'^{(k)}$ denote the value of X' at time $t + Tk/4$, for $k = 0, 1, 2, 3, 4$. Denote

$$\Delta_j'^{(t)} := Y_j'^{(3)} - Y_j'^{(0)} \quad \text{and} \quad \Delta_i'^{(t)} := Y_i'^{(4)} - Y_i'^{(1)}.$$

Lemma 5.4. *We have*

$$\Delta_i'^{(t)} \mid C^{(t)}, \mathcal{D}^{(t)}, Y'^{(0)}, Y'^{(3)} \sim -\Theta'_{ij}\Delta_j'^{(t)} + N(0,2).$$

Proof. Throughout this proof we condition on $C^{(t)}$ and $\mathcal{D}^{(t)}$. Let $\varepsilon_k \sim N(0,1)$ denote the randomness in the update for $Y_i'^{(k)}$; note the ε_k are i.i.d.

Note that $Y_i'^{(2)}$ is chosen independently of $Y_i'^{(1)}$ when conditioned on $Y'^{(0)}$, and thus $Y'^{(2)} \perp Y'^{(1)} \mid Y'^{(0)}$. It follows that $Y'^{(3)} \perp Y'^{(1)} \mid Y'^{(0)}$, so $Y'^{(1)} \mid Y'^{(0)}, Y'^{(3)}$ has the same distribution as $Y'^{(1)} \mid Y'^{(0)}$, and we may write by Definition 1.2 that

$$Y_i'^{(1)} = Y'^{(0)}, Y'^{(3)} = \sum_{k \neq i} -\Theta'_{ik} Y_k'^{(0)} + \varepsilon_1$$

and

$$\begin{aligned} Y_i'^{(4)} \mid Y'^{(0)}, Y'^{(3)} &= \sum_{k \neq i} -\Theta'_{ik} Y_k'^{(3)} + \varepsilon_4 \\ &= \sum_{k \neq i,j} -\Theta'_{ik} Y_k'^{(0)} - \Theta'_{ij} Y_j'^{(3)} + \varepsilon_4, \end{aligned}$$

and thus

$$\begin{aligned} & Y_i'^{(4)} - Y_i'^{(1)} \mid Y'^{(0)}, Y'^{(3)} \\ &= -\Theta'_{ij}\left(Y_j'^{(3)} - Y_j'^{(0)}\right) + \varepsilon_4 - \varepsilon_1 \\ &\sim -\Theta'_{ij}\left(Y_j'^{(3)} - Y_j'^{(0)}\right) + N(0,2). \end{aligned}$$

□

Now we consider actual observable data from $\widehat{X}$. Let $\widehat{Y}^{(k)}$ denote the value of $\widehat{X}$ at time $t + Tk/4$, for $k = 0, 1, 2, 3, 4$. Denote

$$\widehat{\Delta}_j^{(t)} = \widehat{Y}_j^{(3)} - \widehat{Y}_j^{(0)} \quad \text{and} \quad \widehat{\Delta}_i^{(t)} = \widehat{Y}_i^{(4)} - \widehat{Y}_i^{(1)}.$$

Lemma 5.5. *For constants $c_i, c_j \in 1 \pm \varepsilon$ as in Corollary 4.7, we have for all t with $C^{(t)}$ and $\mathcal{D}^{(t)}$ that*

$$\widehat{\Delta}_i^{(t)} \mid C^{(t)}, \mathcal{D}^{(t)}, \widehat{Y}^{(0)}, \widehat{Y}^{(3)} \sim -\frac{c_j}{c_i}\Theta'_{ij}\widehat{\Delta}_j^{(t)} + N\left(0, \frac{2}{c_i^2}\right).$$

Proof. By Corollary 4.7, we have $\widehat{\Delta}_i^{(t)} = \frac{1}{c_i}\Delta_i'^{(t)}$ and $\widehat{\Delta}_j^{(t)} = \frac{1}{c_j}\Delta_j'^{(t)}$, from which the claim follows. □

Now let $\mathcal{E}^{(t)}$ denote the condition that $|\widehat{\Delta}_j^{(t)}| \geq 0.6$. We will restrict our attention to samples where $\mathcal{E}^{(t)}$ holds.

Lemma 5.6. *We have $\mathbb{P}(\mathcal{E}^{(t)} \mid \mathcal{D}^{(t)}, \widehat{Y}^{(0)}) > 1/4$.*

Proof. As $c_j < 1.1$, we have that $|\widehat{\Delta}_j^{(t)}| \geq 0.6$ is implied by $|\Delta_j'^{(t)}| \geq 0.66$, so it suffices to show $|\Delta_j'^{(t)}| \geq 0.66$ with probability more than $1/4$.

Let Z be the value of X' right before the $Y'^{(3)}$ update. Then, by Definition 1.2, $Y_j'^{(3)}$ given Z is a Gaussian with variance 1, i.e.

$$\Delta_j'^{(t)} \mid Z \sim \mathcal{N}(m, 1)$$

for some m dependent on Z.

Then for any Z, we have

$$\begin{aligned}\mathbb{P}(|\Delta_j'^{(t)}| > 0.66 \mid Z) &= \mathop{\mathbb{P}}_{x\sim N(m,1)}(|x| \geq 0.66 \mid Z) \\ &\geq \mathop{\mathbb{P}}_{x\sim N(|m|,1)}(x \geq 0.66 \mid Z) \\ &\geq \mathop{\mathbb{P}}_{x\sim N(0,1)}(x \geq 0.66) > 1/4,\end{aligned}$$

implying that

$$\mathbb{P}(|\Delta_j'^{(t)}| > 0.66) > 1/4$$

regardless of Z. □

Lemma 5.7. *For some $\sigma < 2.62$, we have*

$$-\frac{\widehat{\Delta}_i^{(t)}}{\widehat{\Delta}_j^{(t)}} \,\Bigg|\, \widehat{Y}^{(0)}, \widehat{Y}^{(3)}, C^{(t)}, \mathcal{D}^{(t)}, \mathcal{E}^{(t)} \sim \frac{c_j}{c_i}\Theta'_{ij} + N(0, \sigma^2)$$

Proof. Note that $\mathcal{E}^{(t)}$ is determined by $\widehat{Y}^{(0)}$ and $\widehat{Y}^{(3)}$, so from Lemma 5.5 we have

$$\frac{\widehat{\Delta}_i^{(t)}}{\widehat{\Delta}_j^{(t)}} \,\Bigg|\, \widehat{Y}^{(0)}, \widehat{Y}^{(3)}, C^{(t)}, \mathcal{D}^{(t)}, \mathcal{E}^{(t)} \sim -\frac{c_j}{c_i}\Theta'_{ij} + \mathcal{N}\left(0, \frac{2}{c_i^2\left(\widehat{\Delta}_j^{(t)}\right)^2}\right),$$

Since $c_i > 0.9$ and $|\widehat{\Delta}_j^{(t)}| > 0.6$, the upper bound follows from

$$\frac{\sqrt{2}}{c_i\widehat{\Delta}_j^{(t)}} < 2.62.$$

□

5.2 Retrieving the estimator

We divide our Glauber trajectory into M intervals of length T, for some M and T we will decide later (so the trajectory has length MT). We discard intervals that do not satisfy $\mathcal{D}^{(t)}$ and $\mathcal{E}^{(t)}$, sampling $-\widehat{\Delta}_i^{(t)}/\widehat{\Delta}_j^{(t)}$ from the rest, and treating intervals that fail to satisfy $C^{(t)}$ as corruption.

Let η be the fraction of intervals we sample from in which $C^{(t)}$ does not hold, i.e. the corruption factor that will be passed into Lemma 3.7. We first bound η.

Lemma 5.8. *Our corruption rate is*

$$\eta \leq 4(1 - \exp(-2Td)).$$

Proof. Since the Poisson processes updating each coordinate are independent, the events $C^{(t)}$ and $\mathcal{D}^{(t)}$ are independent, so by Lemma 3.3 we have $\mathbb{P}(\neg C^{(t)} \mid \mathcal{D}^{(t)}) = 1 - \exp(-2Td)$. By Lemma 5.6, we have $\mathbb{P}(\mathcal{E}^{(t)} \mid \mathcal{D}^{(t)}) > 1/4$, so Bayes' theorem we have

$$\begin{aligned}\eta &= \mathbb{P}(\neg C^{(t)} \mid \mathcal{D}^{(t)}, \mathcal{E}^{(t)}) \\ &= \frac{\mathbb{P}(\mathcal{E}^{(t)} \mid \neg C^{(t)}, \mathcal{D}^{(t)})\mathbb{P}(\neg C^{(t)} \mid \mathcal{D}^{(t)})}{\mathbb{P}(\mathcal{E}^{(t)} \mid \mathcal{D}^{(t)})} \\ &< 4(1 - \exp(-2Td)),\end{aligned}$$

as claimed. □

Now, we show we are able to collect sufficiently many samples in which $\mathcal{D}^{(t)}$ and $\mathcal{E}^{(t)}$ hold.

Lemma 5.9. *For a Glauber dynamic of length MT (split into M intervals of length T as above), in at least $MT^4/4096$ intervals do $\mathcal{D}^{(t)}$ and $\mathcal{E}^{(t)}$ hold with probability at least $1 - \exp(-MT^4/16384)$.*

Proof. For the ith interval, we have $\mathbb{P}(\mathcal{D}) \geq T^4/512$ by Lemma 3.4 and $\mathbb{P}(\mathcal{E} \mid \mathcal{D}) \geq 1/4$ by Lemma 5.6. Thus, the event we keep the ith interval is dominated by $\mathbb{1}_i = \mathrm{Ber}(T^4/2048)$.

By Chernoff,

$$\mathbb{P}\left(\sum_{i=1}^{M} \mathbb{1}_i \leq \frac{1}{2}\frac{MT^4}{2048}\right) \leq \exp\left(-\frac{1}{8}\frac{MT^4}{2048}\right),$$

and the desired follows. □

Proof of Lemma 5.3. Note we may isolate the first $T_{\mathrm{diag}} := 800000d\log(16/\delta)$ of the trajectory, then apply Lemma 4.2 to retrieve approximations D_{ii} and D_{jj} of the diagonal up to multiplicative factor $(1 \pm \varepsilon)$, and thus all necessary readings of $\widehat{X}_i$ and $\widehat{X}_j$ with probability at least $1 - \delta/2$. We then focus on the remainder of the trajectory.

In Lemma 5.9, set ϕ, T, and M so that

$$\phi = \frac{\alpha}{257}, \quad T = \frac{\phi}{d}, \quad MT^4 = \frac{128\log(8/\delta)}{\phi^2}.$$

Note the corruption rate satisfies

$$\eta < 4(1 - \exp(-2\phi)) < 8\phi.$$

Henceforth we simply set $\eta := 8\phi < 1/10$.

Then, note from the weak bound $MT^4/16384 > \log(4/\delta)$, we have

$$\exp\left(-\frac{MT^4}{16384}\right) \leq \delta/4,$$

so by Lemma 5.9, with probability $1 - \delta/4$ we see

$$\frac{MT^4}{4096} = \frac{32\log(8/\delta)}{\phi^2} = \frac{2\log(8/\delta)}{\eta^2}$$

samples.

Then, by Lemma 3.7, with success probability $1-\delta/4$, we obtain an estimate $\widehat{\mu}$ for $-\frac{c_j}{c_i}\Theta'_{ij}$ within additive error $5\eta(2.62) < 9\alpha/22$. By flipping the sign we may assume we have an estimate for $\frac{c_j}{c_i}\Theta'_{ij}$ within $9\alpha/22$ accuracy.

However,

- If $i \nsim j$, then $\Theta'_{ij} = 0$, so $|\mu'| < 9\alpha/22$.
- If $i \sim j$, then $|\Theta'_{ij}| > \alpha$, so

$$\left|\frac{c_j}{c_i}\Theta'_{ij}\right| > \frac{9\alpha}{11},$$

  and $|\mu'| > 9\alpha/22$.

Thus, by comparing μ' with $9\alpha/22$, we may determine whether $i \sim j$. By union bound, the failure probability is at most δ, and the required length of the trajectory is given by

$$T_{\text{diag}} + MT < \frac{2173547904 \cdot d^3 \log(16/\delta)}{\alpha^5}. \qquad \square$$

Proof of Theorem 5.1. It suffices to apply Lemma 5.3 with error $2\delta/n^2$ for each pair (i, j). Then, the probability of any error among the $\binom{n}{2}$ candidate edges is less than δ by union bound. $\square$

Proof of Corollary 5.2. We run Theorem 5.1. Instead of performing diagonal approximation up to multiplicative factor $1 \pm 1/10$ via Lemma 4.2 in Lemma 5.3, we estimate each diagonal entry up to multiplicative factor $1 \pm A\varepsilon$ for some constant $A < 1$, where $A\varepsilon < 1/10$. Report these estimates as $\widehat{\Theta}_{ii}$.

Then, we proceed as in Theorem 5.1, but we decrease the parameter α to $B\alpha\varepsilon$ for some constant B. We report $\widehat{\Theta}_{ij} = 0$ whenever Theorem 5.1 reports $i \nsim j$, so it suffices to consider i and j with $|\Theta'_{ij}| > \alpha$. Note in the proof of Lemma 5.3 that, for each i and j, we derive an estimate $\widehat{m}u_{ij} = \widehat{\mu}$ for $\frac{c_j}{c_i}\Theta'_{ij}$ with

$$\left|\widehat{\mu} - \frac{c_j}{c_i}\Theta'_{ij}\right| < \frac{9B\alpha\varepsilon}{22} = O(\varepsilon|\Theta'_{ij}|).$$

As $c_i, c_j \in 1 \pm \varepsilon$, we know

$$\left|\frac{c_j}{c_i}\Theta'_{ij} - \Theta'_{ij}\right| = O(\varepsilon|\Theta'_{ij}|),$$

whence

$$\left|\widehat{\mu} - \Theta'_{ij}\right| = O(\varepsilon|\Theta'_{ij}|),$$

and thus

$$\left|\widehat{\mu}\sqrt{\Theta_{ii}\Theta_{jj}} - \Theta_{ij}\right| = O(\varepsilon|\Theta_{ij}|).$$

We set

$$\widehat{\Theta}_{ij} := \widehat{\mu}\sqrt{\widehat{\Theta}_{ii}\widehat{\Theta}_{jj}},$$

so from Corollary 4.7 we have

$$\left|\widehat{\Theta}_{ij} - \widehat{\mu}\sqrt{\Theta_{ii}\Theta_{jj}}\right| = O\left(\varepsilon\widehat{\mu}\sqrt{\Theta_{ii}\Theta_{jj}}\right) = O(\varepsilon|\Theta_{ij}|).$$

We conclude by triangle inequality

$$\left|\widehat{\Theta}_{ij} - \Theta_{ij}\right| = O(\varepsilon|\Theta_{ij}|).$$

An appropriate choice of constants A and B gives the desired. $\square$

Algorithm 2 $\widetilde{\Theta} = \text{LearnGGM}(n, d, \alpha, \varepsilon, \delta)$

Let
$$T = \frac{\alpha}{257d} \quad \text{and} \quad M = \frac{1116792422656 d^4 \log(2n/\delta)}{\alpha^6}.$$
Let $\text{diag}(\widetilde{\Theta}) = \text{EstimateDiagonal}(n, d, \alpha, 1/10, \delta/2)$ in Algorithm 1.
Keep the unused part of the trajectory.
Let $E = \varnothing$.
Observe Glauber trajectory of length MT, split into M intervals $I_1, \ldots, I_M$ of length T.
for $i = 1, \ldots, n$ **do**
  for $j = i + 1, \ldots, n$ **do**
    Let $S = \{1 \le t \le M : \mathcal{D}^{(t)}, \mathcal{E}^{(t)}\}$, where $\mathcal{D}^{(t)}$ and $\mathcal{E}^{(t)}$ are defined in §5.1.
    if $|S| < \frac{1}{4096} MT^4$ **then**
      Return $\perp$.
    else
      By Lemma 3.7, estimate the mean of $\widehat{\mu}$ of $\{-\widehat{\Delta}_i^{(t)}/\widehat{\Delta}_j^{(t)} \mid t \in S\}$, where $\widehat{\Delta}_i^{(t)}$ and $\widehat{\Delta}_j^{(t)}$ are defined in §4.1.
      if $|\widehat{\mu}| > 9\alpha/22$ **then**
        Set $E = E \sqcup \{(i, j)\}$.
      end if
    end if
  end for
end for
Return E.

Remark 5.10 (On large constants). The constants in this section preclude the practicality of this section. We briefly explain why the constants may not be improved significantly. In the proof of Lemma 5.3, we estimate Θ'_{ij} using the robust mean estimator, with a specific corruption rate η arising from samples in which neighbors of i or j occur. For the robust mean estimator to be effective, upper-bounds on η are necessary. This corruption rate is on the order of $1 - \exp(-2dT)$, a bound of the form $T < c/d$ is needed for some small constant c specifying the corruption rate. But by Lemma 3.4, the success rate of observing $\mathcal{D}$ in an interval already grows by $T^4/512 = \Theta(T^4)$, so the number of intervals we must observe grows by $\Theta(T^{-4})$, and the length of the trajectory grows by $\Theta(T^{-5})$. In our algorithm, we make do with $\eta < 1/10$, in which case we already have $c \sim 0.05$, giving a significant constant in $\Theta(T^{-5})$ already. Together with various other constants, such as the aforementioned $1/512$, practicality is already impossible. Relaxing this requirement on η raises constants elsewhere, which ultimately does not lead to significant improvement. It is an interesting question and possible direction for further research whether the constants in our approach may be improved to give a practical algorithm.

6 Information-theoretic lower bound

In this section, we obtain information theoretic lower bounds for the number of observations. Note that in the continuous model of the Glauber dynamic, the length of the trajectory T is related to the number of updates N by $T \approx N/n$, where n is the number of vertices.

We analyze the success rate of graph learning tests, defined as follows:

Definition 6.1. A *graph-learning test* is defined as follows. The test receives a Glauber trajectory $\mathcal{T}$, generated from an (α, d)-GGM with precision matrix Θ, whose support graph is G, as well as the values α and d. The test returns an estimate $\widehat{G}$ for G as a function of the trajectory it receives. And the test is successful if and only if $G = \widehat{G}$.

From now on, we consider only well-mixed Glauber trajectories:

Definition 6.2. Let $\mathcal{T}(N, \Theta)$ denote a Glauber trajectory, with N updates, with precision matrix Θ, and with starting point sampled from $\mathcal{N}(0, \Theta^{-1})$.

Definition 6.3. Let $\mathcal{T}(i, \Theta_1)_{X^{(i)}|X^{(i-1)}=X}$ denote the distribution after the ith step of $\mathcal{T}(N, \Theta)$, conditioned on the position after the $(i-1)$th step being X. Let $\mathcal{T}(i, \Theta_1)_{X_\ell^{(i)}|X^{(i-1)}=X}$ denote the distribution of the ℓth coordinate in this setting.

Our main result this section is as follows:

Theorem 6.4. *Let $n \geq 16$ and N be positive integers, and let $0 < \alpha < 1/4$. There exists a set S of $2n$-dimensional $(\alpha, 1)$-sparse GGMs such that, if Θ is sampled uniformly from S and the trajectory $\mathcal{T}(N, \Theta)$ is generated therefrom, then the success rate of a graph-learning test is at most $1/2$, given*

$$N \leq \frac{n \log n}{8\alpha^2}.$$

Note that $(\alpha, 1)$-sparse GGMs are also (α, d)-sparse for every positive integer d, so this bound holds for all d.

Remark 6.5. Note that a continuous-time Glauber trajectory of length T contains $\Theta(nT)$ updates with high probability, so here we establish a lower-bound of $T = \Omega(\log n/\alpha^2)$.

6.1 The class of graphs

Fix n a positive integer, and consider the graph G_0 of sparsity 1 on vertices labeled $1, 2, \ldots, 2n$, with an edge $(2k-1) \sim 2k$ for each $k = 1, \ldots, n$, and no other edges. Further, let G_k be a copy of G_0, but with the edge $(2k-1) \sim 2k$ removed, for $k = 1, \ldots, n$.

For each graph G_k, define Θ_k as the following $(\alpha, 1)$-sparse GGM with underlying graph G_k:

$$(\Theta_k)_{ij} = \begin{cases} 1 & i = j \\ \alpha & i \sim j \text{ in } G_k \\ 0 & \text{else.} \end{cases}$$

Our set of GGMs is $S = \{\Theta_1, \ldots, \Theta_n\}$.

Notice that each GGM is the direct sum of 2×2 matrices, among which at least one is the identity matrix I_2, and the remaining are

$$M_\alpha = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}.$$

The eigenvalues of M_α are $1 - \alpha > 0$ and $1 + \alpha > 0$, so all Θ_k are positive semidefinite.

6.2 A bound on KL-divergence

Our main strategy is to lower-bound the KL-divergence between our candidate GGMs, and apply Fano's method.

Lemma 6.6. *For each $k = 1, \dots, n$, we have*

$$D_{\mathrm{KL}}(\mathcal{T}(N, \Theta_k) \parallel \mathcal{T}(N, \Theta_0)) \le \left(\frac{N}{n} + 1\right)\alpha^2.$$

Proof of Lemma 6.6. Without loss of generality $k = 1$. Recall both trajectories $\mathcal{T}(N, \Theta_0)$ and $\mathcal{T}(N, \Theta_k)$ begin mixed, so we first compute the KL-divergence of their initial positions, by

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{T}(0, \Theta_1)_{X^{(0)}} \parallel \mathcal{T}(0, \Theta_0)_{X^{(0)}}) &= D_{\mathrm{KL}}(\mathcal{N}(0, \Theta_1^{-1}) \parallel \mathcal{N}(0, \Theta_0^{-1})) \\ &= D_{\mathrm{KL}}(\mathcal{N}(0, M_\alpha) \parallel \mathcal{N}(0, I_2)) \\ &= -\frac{1}{2}\ln(1 - \alpha^2) < \alpha^2, \end{aligned}$$

since $\alpha < 1/4$.

Now, by coupling, we may assume the two trajectories share the same sequence of updates. Suppose both trajectories have reached position X before the ith update, and on the ith update, both trajectories update the ℓth coordinate. We explicitly write down the distribution of the ℓth coordinate after this Glauber update:

$$\begin{aligned} \mathcal{T}(i, \Theta_0)_{X_\ell^{(i)} | X^{(i-1)} = X} &= -\sum_{j \sim \ell} -(\Theta_0)_{j\ell} X_j + N(0, 1) \\ \mathcal{T}(i, \Theta_1)_{X_\ell^{(i)} | X^{(i-1)} = X} &= -\sum_{j \sim \ell} -(\Theta_1)_{j\ell} X_j + N(0, 1) \end{aligned}$$

But recall $(\Theta_0)_{j\ell} = (\Theta_1)_{j\ell}$ unless $\{j, \ell\} = \{1, 2\}$, in which case $(\Theta_0)_{12} - (\Theta_1)_{12} = \alpha$. Hence, the two distributions are normal with the same variance, and the means differ by αX_2 if $\ell = 1$, by αX_1 if $\ell = 2$, and by 0 otherwise.

It follows that the KL-divergence for the ith update is given by

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathcal{T}(i, \Theta_1)_{X^{(i)} | X^{(i-1)} = X} \parallel \mathcal{T}(i, \Theta_0)_{X^{(i)} | X^{(i-1)} = X}\right) &= D_{\mathrm{KL}}\left(\mathcal{T}(i, \Theta_1)_{X_\ell^{(i)} | X^{(i-1)} = X} \parallel \mathcal{T}(i, \Theta_0)_{X_\ell^{(i)} | X^{(i-1)} = X}\right) \\ &= \mathop{\mathbb{E}}_{\ell}\left[\begin{cases} \frac{1}{2}\alpha^2 X_2^2 & \ell = 1 \\ \frac{1}{2}\alpha^2 X_1^2 & \ell = 2 \\ 0 & \ell \ge 3 \end{cases}\right] = \frac{\alpha^2}{4n}\left(X_1^2 + X_2^2\right). \end{aligned}$$

But by Lemma 3.11, we always have $X \sim \mathcal{N}(0, \Theta_1^{-1})$, so

$$\mathop{\mathbb{E}}_{X}\left[X_1^2\right] = (\Theta_1^{-1})_{11} = \frac{1}{1 - \alpha^2} < 2,$$

and thus

$$\mathop{\mathbb{E}}_{X^{(i-1)}}\left[D_{\mathrm{KL}}(\mathcal{T}(i, \Theta_1)_{X^{(i)} | X^{(i-1)} = X} \parallel \mathcal{T}(i, \Theta_0)_{X^{(i)} | X^{(i-1)} = X})\right] = \frac{\alpha^2}{2n(1 - \alpha^2)} < \frac{\alpha^2}{n}.$$

We conclude by linearity of expectation,

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{T}(N,\Theta_1) \parallel \mathcal{T}(N,\Theta_0)) &= D_{\mathrm{KL}}(\mathcal{T}(0,\Theta_1)_{X^{(0)}} \parallel \mathcal{T}(0,\Theta_0)_{X^{(0)}}) \\ &\quad + \sum_{i=1}^{N} \mathop{\mathbb{E}}_{X^{(i-1)}} \left[D_{\mathrm{KL}}(\mathcal{T}(i,\Theta_1)_{X^{(i)}|X^{(i-1)}=X} \parallel \mathcal{T}(i,\Theta_0)_{X^{(i)}|X^{(i-1)}=X}) \right] \\ &< \alpha^2 + \frac{N\alpha^2}{n}. \end{aligned}$$ □

6.3 Fano's method

We finish by using Fano's method to bound the success rate of a graph-learning test.

Lemma 6.7. *The success rate of a graph-learning test is bounded by*

$$\mathbb{P}\left(\Theta = \widehat{\Theta}\right) \le \frac{(N+n)\alpha^2}{n\log n} + \frac{\log 2}{\log n}.$$

Proof. Let V be sampled uniformly from $\{1,\ldots,n\}$. We may then bound the mutual information of $X \sim \Theta_V$ by

$$I(V;X) \le \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}(P_i \parallel P_0) < \alpha^2 + \frac{N\alpha^2}{n}.$$

By Fano's inequality, we have

$$\mathbb{P}\left(\Theta = \widehat{\Theta}\right) \le \frac{I(V,X)+\log 2}{\log n} \le \frac{(N+n)\alpha^2}{n\log n} + \frac{\log 2}{\log n}.$$ □

Proof of Theorem 6.4. With our given assumptions, we have

$$N + n \le \frac{n\log n}{4\alpha^2},$$

and so by Lemma 6.7, we have

$$\mathbb{P}\left(\Theta = \widehat{\Theta}\right) \le \frac{(N+n)\alpha^2}{n\log n} + \frac{\log 2}{\log n} \le \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$ □

Acknowledgements

The authors thank Jason Gaitonde for helpful discussions and for comments on an earlier version of this manuscript. The authors also thank Jonathan Bloom and Roman Bezrukavnikov for organizing the SPUR program at MIT, where this project began.

References

[Ami91] Yali Amit. On rates of convergence of stochastic relaxation for gaussian and non-gaussian distributions. *Journal of Multivariate Analysis*, 38(1):82–99, 1991.

[BGS17] Guy Bresler, David Gamarnik, and Devavrat Shah. Learning graphical models from the glauber dynamics. *IEEE Transactions on Information Theory*, 64(6):4072–4080, 2017.

[BHMS26] Guy Blanc, Yizhi Huang, Tal Malkin, and Rocco A Servedio. Is nasty noise actually harder than malicious noise? In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 6767–6787. SIAM, 2026.

[BMS+19] Nitin Bhushan, Florian Mohnert, Daniel Sloot, Lise Jans, Casper Albers, and Linda Steg. Using a gaussian graphical model to explore relationships between items and variables in environmental psychology research. *Frontiers in psychology*, 10:1050, 2019.

[CG16] Paola Cerchiello and Paolo Giudici. Conditional graphical models for systemic risk estimation. *Expert systems with applications*, 43:165–174, 2016.

[DMG+20] Martin Dyrba, Reza Mohammadi, Michel J Grothe, Thomas Kirste, and Stefan J Teipel. Gaussian graphical models reveal inter-modal and inter-regional conditional dependencies of brain alterations in alzheimer's disease. *Frontiers in aging neuroscience*, 12:99, 2020.

[GM24] 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.

[GMM25a] Jason Gaitonde, Ankur Moitra, and Elchanan Mossel. Better models and algorithms for learning ising models from dynamics. *arXiv preprint arXiv:2507.15173*, 2025.

[GMM25b] Jason Gaitonde, Ankur Moitra, and Elchanan Mossel. Bypassing the noisy parity barrier: Learning higher-order markov random fields from dynamics. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 348–359, 2025.

[HLS+10] Shuai Huang, Jing Li, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, Eric Reiman, Alzheimer's Disease NeuroImaging Initiative, et al. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.

[KSI+11] Jan Krumsiek, Karsten Suhre, Thomas Illig, Jerzy Adamski, and Fabian J Theis. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5(1):21, 2011.

[Mek24] Raghu Meka. Complexity of sparse linear regression. Invited talk at the EnCORE Workshop on Computational vs Statistical Gaps in Learning and Optimization, Institute for Pure and Applied Mathematics (IPAM), University of California, Los Angeles, March 2024.

[MVL20] Sidhant Misra, Marc Vuffray, and Andrey Y Lokhov. Information theoretic optimal learning of gaussian graphical models. In *Conference on Learning Theory*, pages 2888–2909. PMLR, 2020.

[Nat95] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[RS97] Gareth O Roberts and Sujit K Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(2):291–317, 1997.

[TRD25] Vignesh Tirukkonda, Anirudh Rayas, and Gautam Dasarathy. Structure learning in gaussian graphical models from glauber dynamics. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pages 1–6. IEEE, 2025.

[WWR10] Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic bounds on model selection for gaussian markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE, 2010.

[YZLM22] Huangdi Yi, Qingzhao Zhang, Cunjie Lin, and Shuangge Ma. Information-incorporated gaussian graphical model for gene expression data. *Biometrics*, 78(2):512–523, 2022.

[ZFLH14] Tanja Zerenner, Petra Friederichs, Klaus Lehnertz, and Andreas Hense. A gaussian graphical model approach to climate networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(2), 2014.

[ZWJ14] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948. PMLR, 2014.

A Robust estimators

In this section, we prove the accuracy of our robust estimation subroutines.

Lemma 3.6 (Robust variance estimation)**.** *Let* $0 < \eta \leq 1/10$ *and* $0 < \delta < 1$*. There is a linear-time algorithm that takes n samples from* $\mathcal{N}(0, \sigma^2)$*, an* η*-fraction of which are corrupted (as in Definition 3.5), and outputs* $\widehat{\sigma}$ *such that* $|\widehat{\sigma} - \sigma| < 5\eta\sigma$ *with probability* $1 - \delta$*, provided*

$$n \geq \frac{2\log(4/\delta)}{\eta^2} .$$

Proof. Let Φ and φ denote the cdf and pdf, respectively, of the standard Gaussian. We output $m/\Phi^{-1}(3/4)$, where m is the sample median of the absolute value of the given samples.

For the below analysis, we may assume without loss of generality $\sigma = 1$. Consider $x \sim \mathcal{N}(0, 1)$, and let its absolute value $|x|$ have cdf $G(t) = 2\Phi(t) - 1$.

Then by Dvoretzky-Keifer-Wolfowitz (Fact 3.8), the empircal distribution function $G_n(t)$ satisfies

$$\begin{aligned}\Pr\left(\sup_{x\in\mathbb{R}}\left|G_n(x) - G(x)\right| > \eta/2\right) &\leq 2\exp(-n\eta^2/2)\\ &= \delta/2.\end{aligned}$$

Further, if H_n denotes G_n under η-corruption, we know for all x that

$$\left|H_n(x) - G_n(x)\right| < \frac{k}{n},$$

where $k \sim \text{Binomial}(n, \eta)$ denotes the number of corrupted samples. By Chernoff,

$$\Pr(k > 2\eta n) \le \exp(-n\eta/3) < \delta/2,$$

so with probability $1 - \delta$, we have

$$\begin{aligned}\left|G(x) - H_n(x)\right| &\le \left|G_n(x) - G(x)\right| + \left|H_n(x) - G_n(x)\right| \\ &\le \eta/2 + 2\eta = 5\eta/2.\end{aligned}$$

In particular, if M denotes the median of H_n, then

$$\left|G(M) - \frac{1}{2}\right| < \frac{5\eta}{2}.$$

As $\eta < 1/10$, it follows that

$$\begin{aligned}\left|M - G^{-1}(1/2)\right| &< \frac{5\eta}{2} \cdot \min_{1/4 \le t \le 3/4} \frac{1}{G'(t)} \\ &\le \frac{5\eta}{2} \cdot \frac{1}{2\varphi(3/4)} < 5\eta,\end{aligned}$$

as desired. □

Lemma 3.7 (Robust mean estimation)**.** *Let $0 < \eta \le 1/10$ and $0 < \delta < 1$. We are given samples of the form $x^{(\ell)} = \mu + \sigma^{(\ell)}\varepsilon^{(\ell)}$, for $\ell = 1, \ldots, n$, an η-fraction of which are corrupted, where $\varepsilon^{(\ell)} \sim \mathcal{N}(0, 1)$ are i.i.d. and $\sigma^{(\ell)}$ are chosen adversarially with knowledge of $\varepsilon^{(i)}$ and $\sigma^{(i)}$ if and only if $i < \ell$, so that $|\sigma^{(\ell)}| < 1$. Then the sample median $\widehat{\mu}$ satisfies $|\widehat{\mu} - \mu| < 5\eta$ with probability $1 - \delta$, provided*

$$n \ge \frac{2\log(2/\delta)}{\eta^2}.$$

Proof. We output the sample median $\widehat{\mu}$. Without loss of generality $\sigma_{\max} = 1$. Below, Φ and φ are the cdf and pdf, respectively, of a standard Gaussian.

Let $\mathcal{F}_{\ell-1} = \sigma(\varepsilon^{(1)}, \ldots, \varepsilon^{(\ell-1)})$, and let $\mathbb{1}_\ell$ indicate the event that either $x^\ell > \mu + 5\eta$ or the ℓth sample is corrupted, so that

$$\mathbb{E}[\mathbb{1}_\ell \mid \mathcal{F}_{\ell-1}] \le \eta + \left[1 - \Phi\left(\frac{5\eta}{|\sigma^{(\ell)}|}\right)\right] \le 1 + \eta - \Phi(5\eta)$$

Note $M_k = \sum_{i=1}^k \mathbb{1}_\ell - k(1 + \eta - \Phi(5\eta))$ is a martingale with bounded increments in $[-1, 1]$, so by Azuma-Hoeffding (Fact 3.9),

$$\begin{aligned}\mathbb{P}(\widehat{\mu} \ge \mu + 5\eta) &\le \mathbb{P}\left(\sum_{i=1}^n \mathbb{1}_\ell \ge \frac{n}{2}\right) \\ &= \mathbb{P}\left(M_n \ge n\left(\Phi(5\eta) - \eta - \frac{1}{2}\right)\right) \\ &\le \exp\left(-2n\left(\Phi(5\eta) - \eta - \frac{1}{2}\right)^2\right)\end{aligned}$$

But for $\eta \le 1/10$, we may guarantee $\varphi(5\eta) > 3/10$, so we have

$$\begin{aligned}\Phi(5\eta) - \eta - \frac{1}{2} &= -\eta + \int_0^{5\eta} \varphi(x)\,dx \\ &> -\eta + 5\eta\varphi(5\eta) > \eta/2.\end{aligned}$$

Thus for constant c we have

$$\mathbb{P}(\widehat{\mu} > \mu + 5\eta) \le \exp(-n\eta^2/2).$$

Analogously we have

$$\mathbb{P}(\widehat{\mu} \le \mu - 5\eta) \le \exp(-n\eta^2/2),$$

so by union bound,

$$\mathbb{P}(|\widehat{\mu} - \mu| \ge 5\eta) \le 2\exp(-n\eta^2/2) \le \delta.$$

□

B Continuous time versus number of updates

The upper bounds in Theorems 1.3 and 1.4 are stated in terms of the continuous-time horizon T, whereas the lower bound in Theorem 6.4 is stated in terms of the number of single-site updates. This appendix records the standard reduction showing that these two formulations are equivalent up to the factor n.

Recall from Definition 1.2 that the jump times satisfy $S^{(0)} = 0$ and $S^{(\ell+1)} - S^{(\ell)} \sim \mathrm{Exp}(n)$ i.i.d. Let

$$N_T := \max\{\ell \ge 0 : S^{(\ell)} \le T\}$$

denote the number of updates by time T. Then $N_T \sim \mathrm{Poisson}(nT)$. Moreover, conditional on $N_T = m$, the continuous-time trajectory up to time T is exactly the first m steps of the embedded discrete-time chain, together with the jump times $S^{(1)}, \dots, S^{(m)}$. Since these jump times are independent of Θ and of the embedded chain, the only substantive difference between the two models is that in continuous time one observes a *random* number N_T of discrete-time updates.

We will use the following standard Chernoff bounds for a Poisson random variable: for every $\lambda > 0$,

$$\mathbb{P}\big(\mathrm{Poisson}(2\lambda) < \lambda\big) \le e^{-\lambda/4} \qquad \text{and} \qquad \mathbb{P}\big(\mathrm{Poisson}(\lambda) > 2\lambda\big) \le e^{-\lambda/3}.$$

Proposition B.1. *The continuous-time and discrete-time formulations are interchangeable up to constant factors.*

(i) If there is an estimator that, from the first N updates of the embedded discrete-time chain, succeeds with probability at least $1-\delta$, then there is an estimator that, from a continuous-time trajectory of length $2N/n$, succeeds with probability at least $1 - \delta - e^{-N/4}$.

(ii) If there is an estimator that, from a continuous-time trajectory of length T, succeeds with probability at least $1-\delta$, then there is an estimator that, from the first $\lceil 2nT \rceil$ updates of the embedded discrete-time chain, succeeds with probability at least $1 - \delta - e^{-nT/3}$.

Proof. For *(i)*, observe the continuous-time trajectory up to time $2N/n$. If $N_{2N/n} \ge N$, run the discrete-time estimator on the first N updates of the embedded chain and ignore the rest. Since $N_{2N/n} \sim \mathrm{Poisson}(2N)$, the bad event is $\{N_{2N/n} < N\}$, which has probability at most $e^{-N/4}$.

For *(ii)*, suppose we are given the first $M = \lceil 2nT \rceil$ updates of the embedded discrete-time chain. Sample i.i.d. waiting times $W_1, \dots, W_M \sim \mathrm{Exp}(n)$, set $S^{(\ell)} = \sum_{r=1}^{\ell} W_r$, and reconstruct the corresponding continuous-time path up to time T. If $S^{(M)} > T$, this reconstruction is exact up to time T, so we may run the continuous-time estimator. The bad event is $S^{(M)} \le T$, equivalently $N_T \ge M$ for a rate-n Poisson process, and since $M \ge 2nT$ this has probability at most $e^{-nT/3}$. □

Corollary B.2. *Up to absolute constant factors, the deterministic conversion between the two models is $N \asymp nT$. In particular, the lower bound $N = \Omega(n \log n/\alpha^2)$ of Theorem 6.4 is equivalent to a continuous-time lower bound $T = \Omega(\log n/\alpha^2)$.*

C A technical gap in the analysis of prior work

We identify a gap in the analysis presented in [TRD25]. We have notified the authors, who are currently working to address it. Concretely, in the proof of Lemma 1 (step "(c)"), and again in the proof of Lemma 7 (Appendix B.5), the argument asserts that a certain conditional expectation of a ratio vanishes by invoking mean-zero and (marginal) independence of Gaussian noise terms. As we explain below, this cancellation is not justified when $(i, j) \in E$, because the denominator involves a future increment of coordinate j which depends on earlier noise injected at coordinate i.

From a technical viewpoint, this is one major reason that our mixing-free analysis relies on update patterns of the form "$iiji$" rather than on a direct "iji" ratio argument. The additional update of i is used to avoid exactly the type of dependence described above. In the separate mixing-based result, we are able to work with "iji" patterns because the analysis there is different and does not rely on this cancellation.

In the rest of this section we will discuss this gap in detail.

In Lemma 1, they consider an update pattern $n_1 < n_2 < n_3$ in which node i is updated at times n_1 and n_3 and node j is updated at time n_2, and they define a conditional expectation $\mathbb{E}_{\bar{x},c}[\cdot]$ that fixes the values of $X_{N(i)\setminus\{j\}}$ at the beginning of the window and conditions on the event $|X_j^{(n_2)} - X_j^{(n_1)}| \ge c$. In the proof, the step labeled "(c)" asserts that

$$\mathbb{E}_{\bar{x},c}\left[\frac{\varepsilon_i^{(n_3)} - \varepsilon_i^{(n_1)}}{X_j^{(n_2)} - X_j^{(n_1)}}\right] = 0,$$

citing the mean-zero and (marginal) independence of the Gaussian noise variables $\{\varepsilon_i^{(t)}\}$. A formally identical cancellation is used later in Appendix B.5 (in the proof of Lemma 7) when rewriting their test statistic as a signal term plus a ratio involving $\Delta\varepsilon_i/\Delta X_j$ and dropping the latter in conditional expectation.

Why the denominator depends on earlier noise when $(i, j) \in E$. Fix the "iji" update pattern $n_1 < n_2 < n_3$ from Lemma 1 of [TRD25], where i is updated at times n_1 and n_3 and j is updated at time n_2, and enforce their accompanying event that no neighbor of i or j (other than possibly each other) updates inside the window. Assume $(i, j) \in E$, so $\Theta_{ji} \neq 0$. Recall the Gaussian single-site update rule: when u is updated at time t,

$$X_u^{(t)} = -\sum_{k \in N(u)} \frac{\Theta_{uk}}{\Theta_{uu}} X_k^{(t-1)} + \varepsilon_u^{(t)},$$

$$\varepsilon_u^{(t)} \sim \mathcal{N}\left(0, \frac{1}{\Theta_{uu}}\right).$$

Consider the randomness only through the lens of the single noise term $\varepsilon_i^{(n_1)}$ by fixing the pre-n_1 state, the update indices, and all other Gaussian noises $\{\varepsilon_u^{(t)} : (u,t) \neq (i,n_1)\}$. Under this fixing, the update at time n_1 gives an affine representation

$$X_i^{(n_1)} = m_i + \varepsilon_i^{(n_1)}$$

for a constant m_i determined by what has been fixed. Since i is not updated between n_1 and n_2, the update of j at time n_2 uses $X_i^{(n_1)}$, and therefore

$$\begin{aligned} X_j^{(n_2)} &= -\frac{\Theta_{ji}}{\Theta_{jj}} X_i^{(n_1)} - \sum_{k \in N(j)\setminus\{i\}} \frac{\Theta_{jk}}{\Theta_{jj}} X_k^{(n_2-1)} + \varepsilon_j^{(n_2)} \\ &= m'_j - \frac{\Theta_{ji}}{\Theta_{jj}} \varepsilon_i^{(n_1)} + \varepsilon_j^{(n_2)}, \end{aligned}$$

for a constant m'_j determined by the same fixing. Consequently, the increment appearing in their denominator satisfies

$$\begin{aligned} X_j^{(n_2)} - X_j^{(n_1)} &= a + b\,\varepsilon_i^{(n_1)} + \varepsilon_j^{(n_2)}, \\ b &= -\frac{\Theta_{ji}}{\Theta_{jj}} \neq 0, \end{aligned}$$

for a constant a determined by the fixed past. In particular, when $(i,j) \in E$, the denominator is a function of $\varepsilon_i^{(n_1)}$, so it is not independent of the numerator noise term $\varepsilon_i^{(n_1)}$.

The conditioning $|X_j^{(n_2)} - X_j^{(n_1)}| \geq c$ does not repair the issue. In Lemma 1 (and again in Appendix B.5), [TRD25] conditions on the event

$$\left|X_j^{(n_2)} - X_j^{(n_1)}\right| \geq c.$$

Under the affine form above, this event is exactly

$$\left|a + b\,\varepsilon_i^{(n_1)} + \varepsilon_j^{(n_2)}\right| \geq c,$$

which depends on $\varepsilon_i^{(n_1)}$. Thus, even after imposing the "$\geq c$" condition, the ratio term in step "(c)" involves a numerator noise component that is statistically coupled to the denominator.

The ratio noise term need not have zero (conditional) expectation. As a result, the cancellation invoked in step "(c)" would require showing that an expression of the form

$$\mathbb{E}\left[\frac{\varepsilon_i^{(n_1)}}{a + b\,\varepsilon_i^{(n_1)} + \varepsilon_j^{(n_2)}} \;\middle|\; \left|a + b\,\varepsilon_i^{(n_1)} + \varepsilon_j^{(n_2)}\right| \geq c\right] = 0$$

holds under the relevant conditioning. There is no general reason for this to be true.

In fact, for $a = 0$ and $b = 1$, we have

$$\mathbb{E}\left[\frac{\varepsilon_i^{(n_1)}}{\varepsilon_i^{(n_1)} + \varepsilon_j^{(n_2)}} \;\middle|\; \left|\varepsilon_i^{(n_1)} + \varepsilon_j^{(n_2)}\right| \geq c\right] = \frac{1}{2}$$

by symmetry in $\varepsilon_i^{(n_1)}$ and $\varepsilon_j^{(n_2)}$, as they are i.i.d.

Therefore, the vanishing of the ratio noise term cannot be justified solely from mean-zero and marginal independence of Gaussian noises.

Implications and connection to our approach. The discussion above shows that the specific cancellation used in step (c) of Lemma 1 of [TRD25] is not justified as stated. The denominator contains the future increment $X_j^{(n_2)} - X_j^{(n_1)}$, which, when $(i, j) \in E$, can depend on the earlier noise $\varepsilon_i^{(n_1)}$. Consequently, the relevant ratio term need not have zero conditional expectation. Since the same cancellation is used again in Appendix B.5, in the proof of Lemma 7, the same issue propagates to the later separation argument built on that identity.

For our purposes, this explains why the mixing-free part of our analysis is based on $iiji$ patterns rather than on this direct iji ratio argument. The additional update of i removes the dependence created by the first i-update before the later update of j enters the statistic, so the conditional-independence step needed in our proofs becomes valid. By contrast, when mixing is available, we also give a separate algorithm based on iji patterns; that argument uses the mixing assumption and does not rely on the cancellation above.

D Non-degeneracy does not control eigenvalues

In this appendix we demonstrate that the reciprocal of the minimum eigenvalue of the normalized precision matrix Θ' could be ill conditioned even though the non-zero entries are bounded above 0. Further, we show that this is independent of the sparsity constraint. That if the graph of Θ' is d-sparse for $d > 2$, the edge strength parameter α, which ensures that

$$\Theta'_{i,j} > \alpha$$

is independent from the reciprocal minimum eigenvalue

$$\frac{1}{\lambda_{\min}(\Theta')}$$

Our construction is similar to that of example (5) in [MVL20].

Proposition D.1. *Fix $n \geq 2$, $\alpha < 1$, and $d \geq 2$, for any large number $N \in \mathbb{R}^+$, there exists matrices $\Theta' \in \mathbb{R}^{n\times n}$ satisfying that Θ' has minimum edge strength α and sparsity d, while $\lambda_{min}^{-1} > N$.*

Proof. Consider the matrix

$$B_\varepsilon := \begin{bmatrix} 1 & 1-\varepsilon \\ 1-\varepsilon & 1 \end{bmatrix}$$

this matrix has minimal edge strength $(1 - \varepsilon)$.

The matrix B_ε has the eigenvectors $v_1 = [1, 1]$ and $v_2 = [1, -1]$, satisfying that $B_\varepsilon v_1 = (2 - \varepsilon)v_1$ and $B_\varepsilon v_2 = (\epsilon)v_2$. So $\lambda_{\min}^{-1} = \frac{1}{\epsilon}$. This means that for all $\alpha < 1$ and any large enough number $M > N$ and $M > \frac{1}{1-\alpha}$, $B_{(\frac{1}{M})}$ has entries lower bounded by $1 - \frac{1}{N} > \alpha$, while the reciprocal of the minimum eigenvalue is M.

We can further extend the example of B_ε to higher dimension to demonstrate that edge the edge strength condition, in conjunction with sparsity constraints, could still allow for large λ_{min}^{-1}. For any $n > 2$, the block matrix

$$B_{n,\varepsilon} := \begin{bmatrix} B_\varepsilon & 0 \\ 0 & I_{n-2} \end{bmatrix},$$

where I_{n-2} denotes the identity matrix in $n-2$ dimensions, has minimum entry $1-\varepsilon$ and sparsity $d = 2$. The eigenvalues of $B_{n,\varepsilon}$ are determined by its block components, so its minimum eigenvalue is ε and maximum eigenvalue $2-\varepsilon$, thus it has $\lambda_{\min}^{-1} = \frac{1}{\varepsilon}$. By the same argument above, for any $\alpha < 1$, $d > 2$, we may choose $M > \frac{1}{1-\alpha}$ and $M > N$, then $B_{n,\frac{1}{M}}$ has $\lambda_{\min}^{-1}$ at least M while having α edge strength and d-sparsity. □