Supplementary Material: Exponential Reduction in Sample Complexity with Learning of Ising Model Dynamics

Arkopal Dutt [1] **Andrey Y. Lokhov** [2] **Marc Vuffray** [2] **Sidhant Misra** [2]

We provide in Section S1 detailed proofs of Theorems related to the error bound and sample complexity of D-RISE/D-RPLE in the M-regime. Sections S2 and S3 specify the different optimization techniques that can be used for these estimators and the selection procedure for the regularization parameter of λ in our numerical experiments on different Ising models. Section S4 provides some remarks regarding scalings obtained for the random regular graphs in M-regime. Finally, Section S5 discusses details of preparation of the neural dataset that was used for tests on real data.

S1. Analysis of the estimators D-RISE and D-RPLE

In this section, we provide a rigorous analysis of the sample complexity of learning from Glauber dynamics in the M-regime on Ising models using the estimators of D-RISE and D-RPLE. Theorems 2 and 3 are established.

To simplify the analysis, we consider the case of Ising models with zero local magnetic fields (i.e., $H_i^\star = 0$). The probability measure of a particular configuration of spins $\underline{\sigma} \in \{-1, +1\}^n$ on the resulting Ising model with n nodes is given by

$$p(\underline{\sigma}) = \frac{1}{Z} \exp\left(\sum_{(i,j)\in E} J_{ij}^\star \sigma_i \sigma_j\right). \tag{S1}$$

As noted earlier in Section 4.2, it is useful to view the initial spin configuration $\underline{\sigma}^0$ as a query to the Glauber dynamics which returns the output of $(\underline{\sigma}^1, I^1)$ in the M-regime, where I^1 is the identity of the spin being updated at time $t = 1$. The conditional probability of updating node i through Glauber dynamics in the M-regime is then given by

$$p(\sigma_i^1 | \underline{\sigma}^0, \delta_{i,I^1} = 1) = \frac{\exp\left[\sigma_i^1 (\sum_{j\in\partial i} J_{ij}^\star \sigma_j^0)\right]}{2\cosh\left[\sum_{j\in\partial i} J_{ij}^\star \sigma_j^0\right]}. \tag{S2}$$

where the initial probability distribution over $\underline{\sigma}^0$ is the uniform distribution over all possible spin configurations:

$$p(\underline{\sigma}^0 = \underline{\sigma}) = \frac{1}{2^n}, \ \forall \underline{\sigma} \in \{-1, +1\}^n. \tag{S3}$$

Using m samples $\{(\underline{\sigma}^{0(t)}, \underline{\sigma}^{1(t)}, I^{1(t)})\}_{t\in[m]}$ generated through Glauber dynamics (Eq. S2) in the M-regime, the Ising model is learned using the D-RISE and D-RPLE estimators which are based on the following (simplified) objectives:

$$\text{D-ISO:} \quad \mathcal{S}_m(\underline{J}_u) = \frac{1}{m_u} \sum_{t=1}^{m} \exp\left[-\sigma_u^{1(t)} \left(\sum_{i\neq u} J_{ui} \sigma_i^{0(t)}\right)\right] \delta_{u,I^{1(t)}}. \tag{S4}$$

$$\text{D-PL:} \quad \mathcal{L}_m(\underline{J}_u) = -\frac{1}{m_u} \sum_{t=1}^{m} \ln\left[1 + \sigma_u^{1(t)} \tanh\left(\sum_{i\neq u} J_{ui} \sigma_u^{0(t)}\right)\right] \delta_{u,I^{1(t)}}. \tag{S5}$$

The above objective functions are given for recovering the neighborhood around node u. The analysis of the D-RISE and D-RPLE estimators closely follows the work of (Vuffray et al., 2016) which analyzed the case of learning from i.i.d. samples. We are now in a position to state formal theorems regarding estimation error and sample complexity of structure learning of the estimators D-RISE and D-RPLE.

Theorem 2 (M-regime: Error Bound on Estimates). *Let $\{\underline{\sigma}^{0(t)}, \underline{\sigma}^{1(t)}, I^{1(t)}\}_{t\in[m]}$ be m samples of spin configurations and corresponding node identities drawn through Glauber dynamics (Eq. 2), and define $m_i = \sum_{t=1}^{m} \delta_{i,I^{1(t)}}$ as the number of updates per spin i. Considering M-regime on an Ising model with maximum degree d, maximum coupling intensity β, and assume $H_i^\star = 0\ \forall i$. Then for any $\delta > 0$, the square error of the following estimators with the following choices of penalty parameter $\lambda \propto \sqrt{\frac{\ln(3n^3/\delta)}{m_u}}$ is bounded with probability at least $1-\delta$ for all nodes $u \in V$ if the number of samples satisfies*

i) D-RPLE: If $m_u \geq 2^{17} d^2 \exp(4\beta d) \ln \frac{3n^3}{\delta}$ *for* $\lambda = 4\sqrt{2}\sqrt{\frac{\ln(3n^3/\delta)}{m_u}}$ *then* $\left\|\hat{\underline{J}}_u - \underline{J}_u^\star\right\|_2 \leq 240\sqrt{2d}\exp(2\beta d)\sqrt{\frac{\ln\frac{3n^3}{\delta}}{m_u}}$,

ii) D-RISE: If $m_u \geq 2^{14} d^2 \exp(2\beta d) \ln \frac{3n^3}{\delta}$ *for* $\lambda = 4\sqrt{\frac{\ln(3n^3/\delta)}{m_u}}$ *then* $\left\|\hat{\underline{J}}_u - \underline{J}_u^\star\right\|_2 \leq 240\sqrt{d}\exp(\beta d)\sqrt{\frac{\ln\frac{3n^3}{\delta}}{m_u}}$.

The following theorem quantifies the sample complexity required for structure learning for the different estimators.

Theorem 3 (M-regime: Structure Learning of Ising Model Dynamics). *Let $\{\underline{\sigma}^{0(t)}, \underline{\sigma}^{1(t)}, I^{1(t)}\}_{t\in[m]}$ be m samples of spin configurations and corresponding node identities drawn through Glauber dynamics (Eq. 2), and define $m_i = \sum_{t=1}^{m} \delta_{i,I^{1(t)}}$ as the number of updates per spin i. Consider M-regime on an Ising model with maximum degree d, maximum coupling intensity β, minimum coupling intensity α, and assume $H_i^\star = 0\ \forall i$. Then for any $\delta > 0$, the following estimators with specified penalty parameter of form $\lambda \propto \sqrt{\frac{\ln(3n^3/\delta)}{m_u}}$ reconstructs the edge-set perfectly with probability $p(\hat{E}(\lambda,\alpha) = E) \geq 1-\delta$ if the number of samples satisfies*

i) D-RPLE: $m_u \geq \max(d, \alpha^{-2}) 2^{19} d \exp(4\beta d) \ln \frac{3n^3}{\delta}$ *for* $\lambda = 4\sqrt{2}\sqrt{\frac{\ln(3n^3/\delta)}{m_u}}$,

ii) D-RISE: $m_u \geq \max(d, \alpha^{-2}) 2^{18} d \exp(2\beta d) \ln \frac{3n^3}{\delta}$ *for* $\lambda = 4\sqrt{\frac{\ln(3n^3/\delta)}{m_u}}$.

Remark: Given the choice of the initial distribution $p(\underline{\sigma}_0)$ to be the uniform distribution, the total number of samples m required to get the number of samples m_u that satisfy Theorems 2 and 3 is $m = O(nm_u)$.

S1.1. Conditions for controlling estimation error

In order to control the error of the D-RISE and D-RPLE estimators, we enforce conditions shown to be sufficient in (Negahban et al., 2009) for such ℓ_1-regularized M-estimators. These conditions are similar to the ones shown in (Vuffray et al., 2016) and are restated here for completeness. We state them considering the D-RISE estimator but they hold for the D-RPLE estimator as well.

Condition 1. *The ℓ_1-penalty parameter strongly enforces regularization if it is greater than any partial derivatives of the objective function at $\underline{J}_u = \underline{J}_u^\star$*

$$\|\nabla S_m(\underline{J}_u^\star)\|_\infty \leq \frac{\lambda}{2}. \tag{S6}$$

The above condition ensures that $\underline{J}_u^\star$ has at most d non-zero components and then the difference of the estimates $\Delta_u = \hat{\underline{J}}_u - \underline{J}_u^\star$ lies within the set

$$K := \left\{\Delta_u \in \mathbb{R}^{n-1} | \|\Delta_u\|_1 \leq 4\sqrt{d}\,\|\Delta_u\|_2\right\}. \tag{S7}$$

Denoting the residual of the first order Taylor expansion of the objective function of the estimator:

$$\delta S_m(\Delta, \underline{J}_u^\star) = S_m(\underline{J}_u^\star + \Delta) - S_m(\underline{J}_u^\star) - \langle \nabla S_m(\underline{J}_u^\star), \Delta\rangle. \tag{S8}$$

Condition 2. *The objective function is restricted strongly convex with respect to K on a ball of radius R centered at $\underline{J}_u = \underline{J}_u^\star$, if for all $\Delta_u \in K$ such that $\|\Delta_u\|_2 \leq R$, there exists a constant $\kappa > 0$ such that*

$$\delta S_m(\Delta_u, \underline{J}_u^\star) \geq \kappa \|\Delta_u\|_2^2. \tag{S9}$$

The above condition ensures that the objective function is strongly convex in a restricted subset of $\mathbb{R}^{n-1}$. The following proposition shows that estimation error can be controlled if the above two conditions are satisfied.

Proposition 1 ((Vuffray et al., 2016)). *If an ℓ_1-regularized M-estimator satisfies Condition 1 and Condition 2 with $R > 3\sqrt{d}\frac{\lambda}{\kappa}$ then the error is bounded by*

$$\left\|\hat{\underline{J}}_u - \underline{J}^\star_u\right\|_2 \leq 3\sqrt{d}\frac{\lambda}{\kappa} \tag{S10}$$

S1.2. Proof of the error bound estimation

For the presentation convenience, we first state Propositions that are useful for the proof. These proofs for these Propositions are given further below.

S1.2.1. D-RISE Estimator

Proposition 2 (Gradient concentration of D-RISE). *For some node $u \in V$, let $m_u \geq \exp(2\beta\ d) \ln \frac{2n}{\delta_1}$, then with probability at least $1-\delta_1$, the components of the gradient of the D-ISO are bounded from above as*

$$\|\nabla S_m(\underline{J}^\star_u)\|_\infty \leq \epsilon_1 \tag{S11}$$

where $\epsilon_1 = 2\sqrt{\frac{\ln \frac{2n}{\delta_1}}{m_u}}$.

Proposition 3 (Restricted Strong Convexity for D-RISE). *For some node $u \in V$, let $m_u \geq 2^{11}\ d^2\ \ln \frac{n^2}{\delta_2}$, then with probability at least $1-\delta_2$, the residual of the first order Taylor expansion of D-ISO satisfies*

$$\delta S_m(\Delta_u, \underline{J}^\star_u) \geq \exp(-\beta d)\frac{\|\Delta_u\|_2^2}{4(1+2\sqrt{d}R)}, \tag{S12}$$

for all $\|\Delta_u\|_2 \leq R$.

Proof of Theorem 2(ii). Error bound on D-RISE: Let $\delta_1 = \frac{2\delta}{3n}$ and $\delta_2 = \frac{\delta}{3n}$ and $m_u \geq 2^{14} d^2 \exp(2\beta\ d) \ln \frac{3n^3}{\delta}$. Consider any node $u \in V$ and let $\hat{\underline{J}}_u$ be an optimal point of the D-ISO and $\Delta = \hat{\underline{J}}_u - \underline{J}^\star_u$. Using the values of λ and κ found in Proposition 2 and Proposition 3, we will look for values of R that satisfy

$$R > 3\sqrt{d}\frac{\lambda}{\kappa} = 12\sqrt{d}\lambda(1+2\sqrt{d}R)\exp(\beta d). \tag{S13}$$

The above inequality is satisfied for $R = 2/\sqrt{d}$. Therefore, we can apply Proposition 1 for each node u and using the union bound, we find that the error is bounded in ℓ_2-norm with probability at least $1-\delta$ for all nodes by the following quantity,

$$\left\|\hat{\underline{J}}_u - \underline{J}^\star_u\right\|_2 \leq 240\sqrt{d}\exp(\beta d)\sqrt{\frac{\ln \frac{3n^3}{\delta}}{m_u}}. \tag{S14}$$

□

S1.2.2. D-RPLE Estimator

Proposition 4 (Gradient concentration of D-RPLE). *For some node $u \in V$, let $m_u \geq \exp(2\beta\ d) \ln \frac{2n}{\delta_1}$, then with probability at least $1-\delta_1$, the components of the gradient of the D-PL are bounded from above as*

$$\|\nabla \mathcal{L}_m(\underline{J}^\star_u)\|_\infty \leq \epsilon_1 \tag{S15}$$

where $\epsilon_1 = 2\sqrt{2}\sqrt{\frac{\ln \frac{2n}{\delta_1}}{m_u}}$.

Proposition 5 (Restricted Strong Convexity for D-RPLE). *For some node $u \in V$, let $m_u \geq 2^{11}\ d^2\ \ln \frac{n^2}{\delta_2}$, then with probability at least $1-\delta_2$, the residual of the first order Taylor expansion of D-PL satisfies*

$$\delta \mathcal{L}_m(\Delta_u, \underline{J}^\star_u) \geq \exp(-2\beta d)\frac{\|\Delta_u\|_2^2}{4(1+4\sqrt{d}R)} \tag{S16}$$

Proof of Theorem 2(i). Error bound on D-RPLE: Let $\delta_1 = \frac{2\delta}{3n}$ and $\delta_2 = \frac{\delta}{3n}$ and $m_u \geq 2^{17} d^2 \exp(4\beta\, d) \ln \frac{3n^3}{\delta}$. Consider any node $u \in V$ and let $\hat{\underline{J}}_u$ be an optimal point of the D-PL and $\Delta = \hat{\underline{J}}_u - \underline{J}^\star_u$. Using the values of λ and κ found in Proposition 4 and Proposition 5, we will look for values of R that satisfy

$$R > 3\sqrt{d}\frac{\lambda}{\kappa} = 12\sqrt{d}\lambda(1 + 4\sqrt{d}R)\exp(2\beta d). \tag{S17}$$

The above inequality is satisfied for $R = 1/\sqrt{d}$. Therefore, we can apply Proposition 1 and find that the error is bounded in ℓ_2-norm by the following quantity,

$$\left\| \hat{\underline{J}}_u - \underline{J}^\star_u \right\|_2 \leq 240\sqrt{2}\sqrt{d}\exp(2\beta d)\sqrt{\frac{\ln \frac{3n^3}{\delta}}{m_u}}. \tag{S18}$$

The theorem follows by application of the union bound over all nodes. □

S1.3. Proof of structure learning theorem

Proof of Theorem 3. It is a simple application of Theorem 2 for an error equal to $\alpha/2$. □

S1.4. Gradient Concentration

S1.4.1. D-RISE Estimator

Gradient of D-ISO (Eq. S4) is given by:

$$\frac{\partial}{\partial J_{uk}} S_m(\underline{J}_u) = \frac{1}{m_u}\sum_{t=1}^{m} -\sigma_u^{1\,(t)}\sigma_k^{0\,(t)} \exp\left[-\sum_{i\in V\setminus u} J_{ui}\sigma_u^{1\,(t)}\sigma_i^{0\,(t)}\right]\delta_{u,I^{1\,(t)}} \tag{S19}$$

where $m_u = \sum_{t=1}^{m} \delta_{u,I^{1\,(t)}}$. Let us denote the term in the above summation as the following random variable

$$X_{uk}(\underline{J}_u) = -\sigma_u^1\sigma_k^0 \exp\left[-\sum_{i\in V\setminus u} J_{ui}\sigma_u^1\sigma_i^0\right] \quad \forall k \in \partial u \tag{S20}$$

The lemma below indicates that the D-RISE estimator is consistent and unbiased regardless of the choice of $p(\underline{\sigma}^0)$. This will also be useful for the concentration inequality to come after.

Lemma 1. *For any $u \in V$ and $k \in V \setminus u$, we have*

$$\mathbb{E}[X_{uk}(\underline{J}^\star_u)] = 0 \tag{S21}$$

Proof of Lemma 1. Let us note the probability distribution with respect to which we take the expectation.

$$\mathbb{E}_{p(\sigma_u^1,\underline{\sigma}^0|\delta_{u,I^1}=1)}[X_{uk}(\underline{J}^\star_u)] \tag{S22}$$

$$=\mathbb{E}_{p(\sigma_u^1|\underline{\sigma}^0,\delta_{u,I^1}=1)p(\underline{\sigma}^0)}[X_{uk}(\underline{J}^\star_u)] \tag{S23}$$

$$=\sum_{\underline{\sigma}^0}\left[\sum_{\sigma_u^1} p(\sigma_u^1|\underline{\sigma}^0,\delta_{u,I^1}=1)p(\underline{\sigma}^0)X_{uk}(\underline{J}^\star_u)\right] \tag{S24}$$

$$=\sum_{\underline{\sigma}^0} p(\underline{\sigma}^0)\left[\sum_{\sigma_u^1} p(\sigma_u^1|\underline{\sigma}^0,\delta_{u,I^1}=1)X_{uk}(\underline{J}^\star_u)\right] \tag{S25}$$

$$=\sum_{\underline{\sigma}^0}\frac{p(\underline{\sigma}^0)}{2\cosh\left(\sum_{j\in\partial u} J^\star_{uj}\sigma_j^0\right)}\left[\sum_{\sigma_u^1} -\sigma_u^1\sigma_k^0\exp\left(\sigma_u^1\left(\sum_{j\in\partial u} J^\star_{uj}\sigma_j^0 - \sum_{l\in\partial u} J^\star_{ul}\sigma_l^0\right)\right)\right] \tag{S26}$$

$$=0 \tag{S27}$$

where in the first step, we used the law of total expectations. In the second to last step, we used the definition of X_{uk} from Eq. S20 and the conditional probability from Eq. S2. □

Lemma 2. *For any Ising model with n spins considering $u \in V$ and $k \in V \setminus u$, we have*

$$\mathbb{E}[X_{uk}(\underline{J}_u^\star)^2] = 1 \tag{S28}$$

Proof of Lemma 2. From direct computation, we have

$$\mathbb{E}_{p(\sigma_u^1, \underline{\sigma}^0 | \delta_{u,I^1}=1)}[X_{uk}(\underline{J}_u^\star)^2] \tag{S29}$$

$$= \sum_{\underline{\sigma}^0} p(\underline{\sigma}^0) \left[\sum_{\sigma_u^1} p(\sigma_u^1 | \underline{\sigma}^0, \delta_{u,I^1} = 1) X_{uk}(\underline{J}_u^\star)^2 \right] \tag{S30}$$

$$= \sum_{\underline{\sigma}^0} \frac{p(\underline{\sigma}^0)}{2\cosh\left(\sum\limits_{j \in \partial u} J_{uj}^\star \sigma_j^0 \right)} \left[\sum_{\sigma_u^1} \exp\left(-\sigma_u^1 \sum_{j \in \partial u} J_{uj}^\star \sigma_j^0 \right) \right] \tag{S31}$$

$$= \sum_{\underline{\sigma}^0} p(\underline{\sigma}^0) \tag{S32}$$

$$= 1 \tag{S33}$$

where in the second step we noted that $(\sigma_i^1)^2 = (\sigma_i^0)^2 = 1$ when substituting for X_{uk} from Eq. S20. □

Lemma 3. *For any Ising model with n spins with maximum degree d and maximum interaction strength β, we have for $k \neq u \in V$, we have*

$$|X_{uk}(\underline{J}_u^\star)| \leq \exp(\beta d) \tag{S34}$$

Proof of Lemma 3.

$$|X_{uk}(\underline{J}_u^\star)| = \left| -\sigma_u^1 \sigma_k^0 \exp\left(- \sum_{i \in \partial u} J_{ui}^\star \sigma_u^1 \sigma_i^0 \right) \right| \tag{S35}$$

$$= \exp\left(- \sum_{i \in \partial u} J_{ui}^\star \sigma_u^1 \sigma_i^0 \right) \tag{S36}$$

$$\leq \exp(\beta d) \tag{S37}$$

where we firstly noted that $\delta_{u,I^1} \in \{0, 1\}$. In the second step, we noted that the components of $\underline{J}_u^\star$ have a maximum value of β and at most d of them are non-zero. Further $|\sigma_u^1 \sigma_i^0| = 1$ as spins take values in $\{-1, +1\}$. □

In the M-regime, the different tuples of realizations of $(\underline{\sigma}^1, \underline{\sigma}^0, I^1)$ are independent of each other. This allows us to use the lemmas obtained above and obtain a concentration inequality in the following proof.

Proof of Proposition 2. Utilizing Lemmas 1, 2 and 3 combined with Bernstein's inequality, we have

$$p\left[\left| \frac{\partial}{\partial J_{uk}} S_m(\underline{J}_u^\star) \right| > a \right] \leq 2 \exp\left(- \frac{\frac{1}{2} a^2 m_u}{1 + \frac{1}{3} \exp(\beta d) a} \right) \tag{S38}$$

Setting

$$s = \frac{\frac{1}{2} a^2 m_u}{1 + \frac{1}{3} \exp(\beta d) a}, \tag{S39}$$

and considering $z = \frac{s}{m_u} \exp(\beta d)$, we can write

$$p\left[\left| \frac{\partial}{\partial J_{uk}} S_m(\underline{J}_u^\star) \right| > \frac{1}{3} \left(z + \sqrt{\frac{18z}{\exp(\beta d)} + z^2} \right) \right] \leq 2 \exp(-s) \tag{S40}$$

For $m_u \geq s\exp(2\beta d)$, we have $z^2 = \frac{s^2}{m_u^2}\exp(2\beta d) \leq \frac{s}{m_u}$ and that

$$\frac{1}{3}\left(z + \sqrt{\frac{18z}{\exp(\beta d)} + z^2}\right) \leq \frac{1}{3}\left(\sqrt{\frac{s}{m_u}} + \sqrt{\frac{18s}{m_u} + \frac{s}{m_u}}\right) \tag{S41}$$

$$\leq \frac{\sqrt{19}+1}{3}\sqrt{\frac{s}{m_u}} \tag{S42}$$

$$\leq 2\sqrt{\frac{s}{m_u}} \tag{S43}$$

which allows us to simplify Eq. S40:

$$p\left[\left|\frac{\partial}{\partial J_{uk}} S_m(\underline{J}_u^\star)\right| > 2\sqrt{\frac{s}{m_u}}\right] \leq 2\exp(-s) \tag{S44}$$

Setting $s = \ln\frac{2n}{\delta_1}$ and taking the union bound over every component of the gradient gives us the desired result. □

S1.4.2. D-RPLE Estimator

Gradient of D-PL (Eq. S5) is given by:

$$\frac{\partial}{\partial J_{uk}}\mathcal{L}_m(\underline{J}_u) = \frac{1}{m_u}\sum_{t=1}^{m} \sigma_k^{0\,(t)}\left[\tanh\left(\sum_{i\neq u} J_{ui}\sigma_i^{0\,(t)}\right) - \sigma_u^{1\,(t)}\right]\delta_{u,I^{1\,(t)}}. \tag{S45}$$

Let us denote the term in the above summation as the following random variable

$$Z_{uk}(\underline{J}_u) = \sigma_k^0\left[\tanh\left(\sum_{i\neq u} J_{ui}\sigma_i^0\right) - \sigma_u^1\right] \;\forall k \in \partial u \tag{S46}$$

The lemma below indicates that the D-RPLE estimator is consistent and unbiased regardless of the choice of $p(\underline{\sigma}^0)$. This will also be useful for the concentration inequality to come after.

Lemma 4. *For any $u \in V$ and $k \in V \setminus u$, we have*

$$\mathbb{E}[Z_{uk}(\underline{J}_u^\star)] = 0 \tag{S47}$$

Proof of Lemma 4. Let us note the probability distribution with respect to which we take the expectation.

$$\mathbb{E}_{p(\sigma_u^1,\underline{\sigma}^0|\delta_{u,I^1}=1)}[Z_{uk}(\underline{J}_u^\star)] \tag{S48}$$

$$= \sum_{\underline{\sigma}^0}\left[\sum_{\sigma_u^1} p(\sigma_u^1|\underline{\sigma}^0, \delta_{u,I^1}=1)p(\underline{\sigma}^0)Z_{uk}(\underline{J}_u^\star)\right] \tag{S49}$$

$$= \sum_{\underline{\sigma}^0} p(\underline{\sigma}^0)\left[\sum_{\sigma_u^1} p(\sigma_u^1|\underline{\sigma}^0, \delta_{u,I^1}=1)Z_{uk}(\underline{J}_u^\star)\right] \tag{S50}$$

$$= \sum_{\underline{\sigma}^0} p(\underline{\sigma}^0)\left[\sum_{\sigma_u^1} \frac{\exp\left[\sigma_i^1(\sum_{j\in\partial u} J_{uj}^\star\sigma_j^0)\right]}{2\cosh\left[\sum_{j\in\partial u} J_{uj}^\star\sigma_j^0\right]}\sigma_k^0\left(\tanh\left(\sum_{i\neq u} J_{ui}\sigma_i^0\right) - \sigma_u^1\right)\right] \tag{S51}$$

$$= \sum_{\underline{\sigma}^0} p(\underline{\sigma}^0)\sigma_k^0\left[\tanh\left(\sum_{i\neq u} J_{ui}\sigma_i^0\right) - \sum_{\sigma_u^1}\frac{\exp\left[\sigma_i^1(\sum_{j\in\partial u} J_{uj}^\star\sigma_j^0)\right]}{2\cosh\left[\sum_{j\in\partial u} J_{uj}^\star\sigma_j^0\right]}\sigma_u^1\right] \tag{S52}$$

$$= \sum_{\underline{\sigma}^0} p(\underline{\sigma}^0)\sigma_k^0\left[\tanh\left(\sum_{i\neq u} J_{ui}\sigma_i^0\right) - \tanh\left(\sum_{i\neq u} J_{ui}\sigma_i^0\right)\right] \tag{S53}$$

$$= 0 \tag{S54}$$

where in the first step, we used the law of total expectations. In the third to last step, we used the definition of Z_{uk} from Eq. S46 and the conditional probability from Eq. S2. □

Lemma 5. *For any Ising model with n spins with maximum degree d and maximum interaction strength β, we have for $k \neq u \in V$, we have*

$$|Z_{uk}(\underline{J}_u^\star)| \leq 2 \tag{S55}$$

Proof of Lemma 5.

$$|Z_{uk}(\underline{J}_u^\star)| = \left| \sigma_k^0 \left[\tanh\left(\sum_{i \neq u} J_{ui} \sigma_i^0 \right) - \sigma_u^1 \right] \right| \tag{S56}$$

$$\leq \left| \tanh\left(\sum_{i \neq u} J_{ui} \sigma_i^0 \right) \right| + |\sigma_u^1| \tag{S57}$$

$$\leq 2 \tag{S58}$$

where in the last step, we noted that $|\tanh(\cdot)| \leq 1$ and $|\sigma_u^1| = 1$ as spins take values in $\{-1, +1\}$. □

As noted before, the different tuples of realizations of $(\underline{\sigma}^1, \underline{\sigma}^0, I^1)$ are independent of each other in the M-regime. We use the lemmas obtained above and obtain a concentration inequality in the following proof.

Proof of Proposition 4. Utilizing Lemmas 4 and 5 combined with Hoeffding's inequality, we have

$$p\left[\left| \frac{\partial}{\partial J_{uk}} \mathcal{L}_m(\underline{J}_u^\star) \right| > a \right] \leq 2 \exp\left(-\frac{m_u a^2}{8} \right) \tag{S59}$$

Let $a = 2\sqrt{2}\sqrt{\frac{s}{m_u}}$. We then have

$$p\left[\left| \frac{\partial}{\partial J_{uk}} \mathcal{L}_m(\underline{J}_u^\star) \right| > 2\sqrt{2}\sqrt{\frac{s}{m_u}} \right] \leq 2 \exp\left(-s\right) \tag{S60}$$

Setting $s = \ln \frac{2n}{\delta_1}$ and taking the union bound over every component of the gradient gives us the desired result. □

S1.5. Restricted Strong Convexity

The following deterministic functional inequality derived in (Vuffray et al., 2016) will be useful for latter results.

Lemma 6. *The following inequality holds for all $z \in \mathbb{R}$.*

$$e^{-z} - 1 + z \geq \frac{z^2}{2 + |z|}. \tag{S61}$$

We will also find the following deterministic functional inequality useful.

Lemma 7. *The following inequality holds for all $z \in \mathbb{R}$ and for any $x \in \mathbb{R}$.*

$$-\ln\left(1 + \tanh(x+z)\right) + \ln\left(1+x\right) + z\left(1 - \tanh(x)\right) \geq \frac{z^2}{2(1+|z|)} \operatorname{sech}^2(x). \tag{S62}$$

Proof of Lemma 7. Let us denote the function

$$f(x,z) := -\ln\left(1 + \tanh(x+z)\right) + \ln\left(1+x\right) + z\left(1 - \tanh(x)\right) \tag{S63}$$

and the auxillary function

$$g(x,z) := 2(1+|z|)f(x,z) - \operatorname{sech}^2(x) z^2. \tag{S64}$$

We show that for fixed $x \in \mathbb{R}$ and $z = 0$, $g(x, z)$ achieves its minimum at $g(x, 0) = 0$. Observe the first partial derivative of $g(x, z)$ wrt z is given by

$$\frac{\partial}{\partial z} g(x,z) = \begin{cases} 2f(x,z) + 2(1+z)\left(\tanh(x+z) - \tanh(x)\right) - 2z\,\mathrm{sech}^2(x), & z > 0 \\ -2f(x,z) + 2(1-z)\left(\tanh(x+z) - \tanh(x)\right) - 2z\,\mathrm{sech}^2(x), & z < 0 \end{cases} \tag{S65}$$

We note that the partial derivatives vanishes at zero from both the negative and positive directions of z:

$$\lim_{z \to 0_+} \frac{\partial}{\partial z} g(x,z) = \lim_{z \to 0_-} \frac{\partial}{\partial z} g(x,z) = 0 \tag{S66}$$

The second partial derivatives of $g(x, z)$ are given by

$$\frac{\partial^2}{\partial z^2} g(x,z) = \begin{cases} 4\left(\tanh(x+z) - \tanh(x)\right) + 2(1+z)\,\mathrm{sech}^2(x+z) - 2\,\mathrm{sech}^2(x), & z > 0 \\ 4\left(\tanh(x) - \tanh(x-z)\right) + 2(1-z)\,\mathrm{sech}^2(x+z) - 2\,\mathrm{sech}^2(x), & z < 0 \end{cases} \tag{S67}$$

We note that for $z > 0$, $\frac{\partial^2}{\partial z^2} g(x, z)$ is non-negative

$$\begin{aligned} \frac{\partial^2}{\partial z^2} g(x,z) &= 4\left(\tanh(x+z) - \tanh(x)\right) + 2(1+z)\,\mathrm{sech}^2(x+z) - 2\,\mathrm{sech}^2(x) && \text{(S68)} \\ &> 4\left(\tanh(x+z) - \tanh(x)\right) + 2\,\mathrm{sech}^2(x+z) - 2\,\mathrm{sech}^2(x) && \text{(S69)} \\ &= 4\left(\tanh(x+z) - \tanh(x)\right) + 2\left(\tanh^2(x) - 2\tanh^2(x+z)\right) && \text{(S70)} \\ &= 2\left(\tanh(x+z) - \tanh(x)\right)\left(2 - \tanh(x) - \tanh(x+z)\right) && \text{(S71)} \\ &> 0 && \text{(S72)} \end{aligned}$$

where in the last step, we noted that $\tanh(x)$ is a monotonically increasing function and that $\tanh(x) < 1$. A similar result can be shown for $z < 0$. Combining this with Eq. S66, we prove that for all z and any x, $g(x, z) \geq g(x, 0) = 0$ which gives us our desired result. □

Let H_{ij} denote the correlation matrix elements

$$H_{ij} = \mathbb{E}_{p(\sigma_u^1, \underline{\sigma}^0 | \delta_{u, I^1} = 1)}[\sigma_i^0 \sigma_j^0] \tag{S73}$$

and we denote the corresponding matrix as $H = [H_{ij}] \in \mathbb{R}^{|\partial u| \times |\partial u|}$. Let the empirical estimate of the correlation matrix be denoted by $\hat{H}$ and the matrix elements be given by

$$\hat{H}_{ij} = \frac{1}{m_u} \sum_{t=1}^{m} \sigma_i^{0(t)} \sigma_j^{0(t)} \delta_{u, I^{1(t)}} \tag{S74}$$

Lemma 8. *Consider some node $u \in V$. With probability at least $1 - n^2 \exp\left(-\frac{m_u \epsilon_2^2}{2}\right)$, we have*

$$|\hat{H}_{ij} - H_{ij}| \leq \epsilon_2 \tag{S75}$$

for all $i, j \in \partial u$ and $\epsilon_2 > 0$.

Proof of Lemma 8. Fix $i, j \in \partial u$. Noting that $|\sigma_i^0 \sigma_j^0 \delta_{u, I^1}| \leq 1$ and combining with Hoeffding's inequality:

$$p(|\hat{H}_{ij} - H_{ij}| > \epsilon_2) \leq 2 \exp\left(-\frac{m_u \epsilon_2^2}{2}\right) \tag{S76}$$

The proof follows by noting that the matrix H is symmetric allowing us to take the union bound over all $i < j \in V \setminus u$. □

Lemma 9. *Consider an Ising model with n spins and some node $u \in V$, then the following holds for the Hessian*

$$\Delta_u^T H \Delta_u = \|\Delta_u\|_2^2 \tag{S77}$$

Proof of Lemma 9. From direct evaluation of H, we have

$$H_{ij} = \mathbb{E}[\sigma_i^0 \sigma_j^0] \tag{S78}$$

$$= \sum_{\underline{\sigma}^0} p(\underline{\sigma}^0) \sum_{\sigma_u^1} p(\sigma_u^1 | \underline{\sigma}^0, \delta_{u,I^1} = 1) \sigma_i^0 \sigma_j^0 \tag{S79}$$

$$= \sum_{\underline{\sigma}^0} p(\underline{\sigma}^0) \sigma_i^0 \sigma_j^0 \tag{S80}$$

$$= \delta_{ij} \tag{S81}$$

where $\delta_{ij} = 1$ iff $i = j$. Thus the correlation matrix $H = I$ where I is an identity matrix of size $\mathbb{R}^{|\partial u| \times |\partial u|}$. We assumed that the initial distribution of $p(\underline{\sigma}^0)$ is given by Eq. S3. The proof follows from computation of $\Delta_u^T H \Delta_u$. □

S1.5.1. D-RISE Estimator

Lemma 10. *The residual of the first order Taylor expansion of the* D-ISO *satisfies*

$$\delta S_m(\Delta_u, \underline{J}_u^\star) \geq \exp(-\beta d) \frac{\Delta_u^T \hat{H} \Delta_u}{2 + \|\Delta_u\|_1}. \tag{S82}$$

Proof of Lemma 10. Noting the expression of the residual from Eq. S8 and that $\langle \nabla S_m(\underline{J}_u^\star), \Delta_u \rangle = \sum_{k \in \partial u} \left[\frac{\partial}{\partial J_{uk}} S_m(\underline{J}_u^\star) \right] \Delta_{uk}$, we have that

$$\delta S_m(\Delta_u, \underline{J}_u^\star) = \frac{1}{m_u} \sum_{t=1}^{m} \exp\left(- \sum_{k \in \partial u} J_{uk}^\star \sigma_u^{1\,(t)} \sigma_k^{0\,(t)} \right) \delta_{u, I^{1\,(t)}} \times \left[\exp\left(- \sum_{k \in \partial u} \Delta_{uk} \sigma_u^{1\,(t)} \sigma_k^{0\,(t)} \right) - 1 + \sum_{k \in \partial u} \Delta_{uk} \sigma_u^{1\,(t)} \sigma_k^{0\,(t)} \right] \tag{S83}$$

$$\geq \exp(-\beta d) \frac{1}{m_u} \sum_{t=1}^{m} \frac{\left(\sum_{k \in \partial u} \Delta_{uk} \sigma_u^{1\,(t)} \sigma_k^{0\,(t)} \right)^2}{2 + |\sum_{k \in \partial u} \Delta_{uk} \sigma_u^{1\,(t)} \sigma_k^{0\,(t)}|} \delta_{u, I^{1\,(t)}} \tag{S84}$$

$$\geq \exp(-\beta d) \frac{\Delta^T \hat{H} \Delta}{2 + \|\Delta_u\|_1}. \tag{S85}$$

where in the second step we used Lemma 6 considering $z = \sum_{k \in \partial u} \Delta_{uk} \sigma_u^{1\,(t)} \sigma_k^{0\,(t)}$ and noted $|\sum_{i \in \partial u} J_{ui} \sigma_i^{0\,(t)} \sigma_u^{1\,(t)}| \leq \beta d$. The proof follows from using the definition of $\hat{H}$ and observing that $|\sum_{k \in \partial u} \Delta_{uk} \sigma_u^{1\,(t)} \sigma_k^{0\,(t)}| \leq \|\Delta_u\|_1$. □

Proof of Proposition 3. Using Lemma 10 and for $m_u \geq \frac{2}{\epsilon_2^2} \ln \frac{n^2}{\delta_2}$, the residual satisfies the following with probability at least $1 - \delta_2$:

$$\delta S_m(\Delta_u, \underline{J}_u^\star) \geq \exp(-\beta d) \frac{\Delta_u^T \hat{H} \Delta_u}{2 + \|\Delta_u\|_1} \tag{S86}$$

$$= \exp(-\beta d) \frac{\Delta_u^T H \Delta_u + \Delta_u^T (H - \hat{H}) \Delta_u}{2 + \|\Delta_u\|_1} \tag{S87}$$

$$\overset{(a)}{\geq} \exp(-\beta d) \frac{\Delta_u^T H \Delta_u - \epsilon_2 \|\Delta_u\|_1^2}{2 + \|\Delta_u\|_1} \tag{S88}$$

$$\overset{(b)}{\geq} \exp(-\beta d) \frac{\|\Delta_u\|_2^2 - \epsilon_2 \|\Delta_u\|_1^2}{2 + \|\Delta_u\|_1} \tag{S89}$$

where we used Lemma 8 in (a) and Lemma 9 in (b). Setting $\epsilon_2 = \frac{1}{32d}$, we note that

$$-\epsilon_2 \left\|\Delta_u\right\|_1^2 \geq -\frac{1}{2}\left\|\Delta_u\right\|_2^2 \tag{S90}$$

where we have used Condition S9 that $\left\|\Delta_u\right\|_1 \leq 4\sqrt{d}\left\|\Delta_u\right\|_2$. Combining this with $\left\|\Delta_u\right\|_2 \leq R$, we obtain the desired result. □

S1.5.2. D-RPLE ESTIMATOR

Consider the first order Taylor expansion:

$$\delta\mathcal{L}_m(\Delta, \underline{J}_u^\star) = \mathcal{L}_m(\underline{J}_u^\star + \Delta) - \mathcal{L}_m(\underline{J}_u^\star) - \langle \nabla\mathcal{L}_m(\underline{J}_u^\star), \Delta\rangle. \tag{S91}$$

Lemma 11. *The residual of the first order Taylor expansion of the* D-PL *satisfies*

$$\delta\mathcal{L}_m(\Delta_u, \underline{J}_u^\star) \geq \exp(-2\beta d)\frac{\Delta_u^T \hat{H}\Delta_u}{2(1+\left\|\Delta_u\right\|_1)}. \tag{S92}$$

Proof of Lemma 11. Noting the expression of the residual from Eq. S91 and that $\langle \nabla\mathcal{L}_m(\underline{J}_u^\star), \Delta_u\rangle = \sum_{k\in\partial u}\left[\frac{\partial}{\partial J_{uk}}\mathcal{L}_m(\underline{J}_u^\star)\right]\Delta_{uk}$, we have that

$$\begin{aligned}
\delta\mathcal{L}_m(\Delta_u, \underline{J}_u^\star) = \frac{1}{m_u}\sum_{t=1}^{m}\Bigg[-\ln\left[1+\sigma_u^{1\,(t)}\tanh\left(\sum_{i\neq u}(J_{ui}+\Delta_{ui})\sigma_i^{0\,(t)}\right)\right] + \\
\ln\left[1+\sigma_u^{1\,(t)}\tanh\left(\sum_{i\neq u}J_{ui}\sigma_i^{0\,(t)}\right)\right] + \sum_{k\in\partial u}\Delta_{uk}\sigma_k{}^{(0)}\sigma_u^{1\,(t)}\left[1-\sigma_u^{1\,(t)}\tanh\left(\sum_{i\neq u}J_{ui}\sigma_i^{0\,(t)}\right)\right]\Bigg]\delta_{u,I^{1\,(t)}}
\end{aligned} \tag{S93}$$

$$\geq \frac{1}{m_u}\sum_{t=1}^{m}\frac{\left(\sum_{k\in\partial u}\Delta_{uk}\sigma_u^{1\,(t)}\sigma_k^{0\,(t)}\right)^2}{2\left(1+\left|\sum_{k\in\partial u}\Delta_{uk}\sigma_u^{1\,(t)}\sigma_k^{0\,(t)}\right|\right)}\operatorname{sech}^2\left(\sum_{i\neq u}J_{ui}\sigma_i^{0\,(t)}\sigma_u^{1\,(t)}\right)\delta_{u,I^{1\,(t)}} \tag{S94}$$

$$\geq \exp(-2\beta d)\frac{1}{m_u}\sum_{t=1}^{m}\frac{\left(\sum_{k\in\partial u}\Delta_{uk}\sigma_u^{1\,(t)}\sigma_k^{0\,(t)}\right)^2}{2\left(1+\left|\sum_{k\in\partial u}\Delta_{uk}\sigma_u^{1\,(t)}\sigma_k^{0\,(t)}\right|\right)}\delta_{u,I^{1\,(t)}} \tag{S95}$$

$$\geq \exp(-2\beta d)\frac{\Delta_u^T\hat{H}\Delta_u}{2(1+\left\|\Delta_u\right\|_1)} \tag{S96}$$

where in the second step, we used Lemma 7 combined with the fact that $\sigma_i\tanh(x) = \tanh(\sigma_i x)$ as $\sigma_i \in \{-1,+1\}$. In the third step, we noted that $\operatorname{sech}^2(x) = 1-\tanh^2(x) \geq \exp(-2|x|)$ and $|\sum_{i\in\partial u}J_{ui}\sigma_i^{0\,(t)}\sigma_u^{1\,(t)}| \leq \beta d$. In the final step, we used the definition of $\hat{H}$ and that $|\sum_{k\in\partial u}\Delta_{uk}\sigma_u^{1\,(t)}\sigma_k^{0\,(t)}| \leq \left\|\Delta_u\right\|_1$ to complete the proof. □

Proof of Proposition 5. Using Lemma 11 and for $m_u \geq \frac{2}{\epsilon_2^2}\ln\frac{n^2}{\delta_2}$, the residual satisfies the following with probability at

least $1-\delta_2$:

$$\delta\mathcal{L}_m(\Delta_u, \underline{J}_u^\star) \geq \exp(-2\beta d)\frac{\Delta_u^T \hat{H}\Delta_u}{2(1+\|\Delta_u\|_1)} \tag{S97}$$

$$= \exp(-2\beta d)k_1\frac{\Delta_u^T H\Delta_u + \Delta_u^T(H-\hat{H})\Delta_u}{2(1+\|\Delta_u\|_1)} \tag{S98}$$

$$\overset{(a)}{\geq} \exp(-2\beta d)\frac{\Delta_u^T H\Delta_u - \epsilon_2\|\Delta_1\|_u^2}{2(1+\|\Delta_u\|_1)} \tag{S99}$$

$$\overset{(b)}{\geq} \exp(-2\beta d)\frac{\|\Delta_u\|_2^2 - \epsilon_2\|\Delta_u\|_1^2}{2(1+\|\Delta_u\|_1)} \tag{S100}$$

where we used Lemma 8 in (a) and Lemma 9 in (b). Setting $\epsilon_2 = \frac{1}{32d}$, we note that

$$-\epsilon_2\|\Delta_u\|_1^2 \geq -\frac{1}{2}\|\Delta_u\|_2^2 \tag{S101}$$

where we have used Condition S9 that $\|\Delta_u\|_1 \leq 4\sqrt{d}\|\Delta_u\|_2$. Combining this with $\|\Delta_u\|_2 \leq R$, we obtain the desired result. □

S2. Setup for numerical experiments

In this section, we describe the optimization techniques that we used in the implementation of D-RISE/D-RPLE estimators and the computer infrastructure that was used for running our numerical experiments.

S2.1. Optimization tools

The D-RISE and D-RPLE estimators involve optimization of a convex objective function. There are a variety of optimization techniques that can be used for this purpose including gradient descent type methods and interior point methods. For our numerical experiments, we used the interior point Ipopt (Biegler & Zavala, 2009) optimization package within the JuMP modeling framework for mathematical optimization in Julia.

As an alternative to the Ipopt software for optimization, we also implemented the coordinate descent (CD) method for D-RISE and D-RPLE as a part of our package. The idea is to perform gradient descent only along one coordinate at a time, and then cycle through coordinates until convergence. For example, the optimization problem for each node u in D-RISE is:

$$(\widehat{\underline{J}}_u, \widehat{H}_u) = \underset{(\underline{J}_u, H_i)}{\operatorname{argmin}}\left[\frac{1}{m_u}\sum_{t=1}^{m}\exp\left(-\sigma_u^{t+1}\left(\sum_{i\neq u}J_{ui}\sigma_i^t + H_u\right)\right)\delta_{u,I^{t+1}} + \lambda\|\underline{J}_u\|_1\right], \tag{S102}$$

In coordinate descent, we optimize over one variable only at a time, e.g. over J_{uk} for some k. Once the minimum is found, we cycle through other components, and repeat until we reach the global minimum of the entire convex function. Interestingly, each iteration step is a solution of a one-dimensional optimization problem:

$$\hat{x} = \underset{x}{\operatorname{argmin}}\cosh x - \kappa\sinh x + \mu|x|, \tag{S103}$$

where the constant κ and the regularization parameter μ is given by

$$\kappa = \frac{b}{a}, \quad \mu = \begin{cases} 0, & x = H_u \\ \lambda/a, & \text{otherwise.} \end{cases} \tag{S104}$$

with

$$a = \begin{cases} \frac{1}{m_u}\sum_{t=1}^{m} \exp\left(-\sigma_u^{t+1}\sum_{i\neq u} J_{ui}\sigma_i^t\right)\delta_{u,I^{t+1}}, & x = H_u \\ \frac{1}{m_u}\sum_{t=1}^{m} \exp\left(-\sigma_u^{t+1}\left(\sum_{i\neq u,k} J_{ui}\sigma_i^t + H_u\right)\right)\delta_{u,I^{t+1}}, & x = J_{uk}\forall k \end{cases} \tag{S105}$$

$$b = \begin{cases} \frac{1}{m_u}\sum_{t=1}^{m} \sigma_u^{t+1}\exp\left(-\sigma_u^{t+1}\sum_{i\neq u} J_{ui}\sigma_i^t\right)\delta_{u,I^{t+1}}, & x = H_u \\ \frac{1}{m_u}\sum_{t=1}^{m} \sigma_u^{t+1}\sigma_k^t\exp\left(-\sigma_u^{t+1}\left(\sum_{i\neq u,k} J_{ui}\sigma_i^t + H_u\right)\right)\delta_{u,I^{t+1}}, & x = J_{uk}\forall k \end{cases} \tag{S106}$$

and the minimum can be found explicitly (after soft-thresholding):

$$\hat{x} = \begin{cases} \log\left(\frac{\sqrt{1-\kappa^2+\mu^2}-\mu\ \text{sign}(\kappa)}{1-\kappa}\right), & \mu < |\kappa| \\ 0, & \text{otherwise.} \end{cases} \tag{S107}$$

Note that there is a slight difference in the optimization problem and solution depending on if the coordinate randomly chosen is a coupling parameter J_{ui} or magnetic field H_u. As a result of having access to the above analytical solution, coordinate descent for D-RISE does not require the choice of the step in the gradient descent. This simplification however does not occur for D-RPLE. At each step of the descent, the updated coordinate is chosen at random.

Another popular optimization method that may be applied to D-RISE and D-RPLE is stochastic gradient descent (SGD) method, although SGD requires tuning hyper-parameters such as learning rate and batch size of samples.

We note that other possible choices of optimization techniques include the entropic gradient descent method (Beck & Teboulle, 2003) (see (Vuffray et al., 2019) for a description of this method used in estimator RISE for the case of i.i.d. samples) and mirror gradient descent method of (Ben-Tal et al., 2001).

S2.2. Computational resources

The numerical experiments were run on a cluster. Each node of the cluster has a Intel Xeon Gold 6248 processor with 2×20 cores with 32GB RAM. We were able to take advantage of multiple cores on each node since learning local neighborhoods of each node in our algorithm is done independently and hence can be carried out in parallel. Jobs for different graphical model instances were distributed on the cluster. Parallelization is implemented in our package and example scripts for distributing jobs are included for completeness.

S3. Empirical selection of the regularization parameter c_λ

In this section, we describe the procedure that we used for selecting the values of the coefficient of the regularization parameter c_λ for the estimators of D-RPLE and D-RISE that we used in different regimes of Glauber dynamics and on different Ising model topologies.

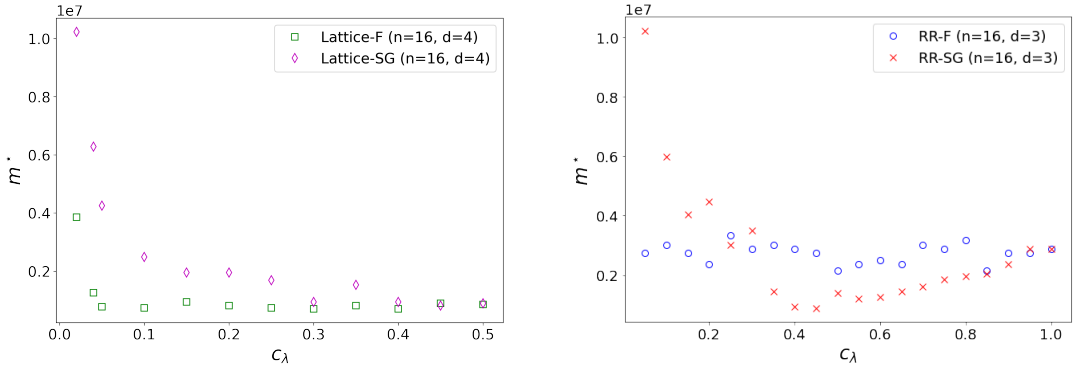
The regularization parameter λ has the following functional form

$$\lambda = c_\lambda\sqrt{\frac{\log(n^2/\delta')}{m_u}} \tag{S108}$$

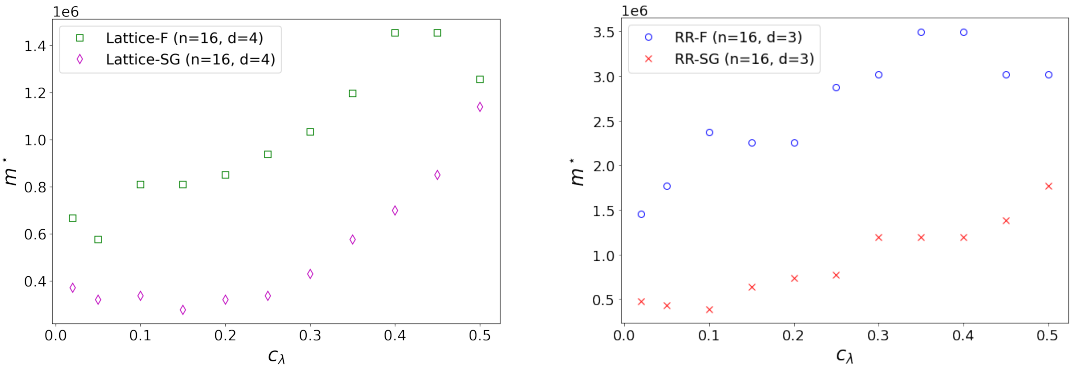
where $1-\delta'$ is the probability with which we wish to successfully reconstruct the local neighborhood of u. Note that δ' should be related to δ used in the quantifying the success of the whole graph recovery as $\delta' = \delta/n$. We follow an approach similar to that followed in Supplementary material of (Lokhov et al., 2018) to determine the optimal values of c_λ. Our experimental protocol for selecting an optimal value of c_λ on a given Ising model topology is as follows. For a fixed typical values of α and β, we determine $m^\star$ for different values of c_λ. The optimal value of c_λ is then defined as the one for which the lowest $m^\star$ was obtained. To determine consensus values across topologies, this procedure is repeated on different types

of lattices and random regular graphs. Further, as we expect different optimal values of c_λ, the studies are repeated for both the T-regime and M-regime.

The results for the T-regime and M-regime are shown in Figure S1 and Figure S2. We observe that a consensus optimal value of c_λ can be selected across lattices or random regular graphs but not across both topologies. In the T-regime, optimal choice of c_λ for D-RPLE is ≈ 0.05 on lattices and ≈ 0.1 on random regular graphs. In the T-regime, optimal choice of c_λ for D-RISE is ≈ 0.1 on lattices and ≈ 0.45 on random regular graphs. In the M-regime, optimal choice of c_λ for D-RPLE is ≈ 0.05 on lattices and ≈ 0.3 on random regular graphs. In the M-regime, optimal choice of c_λ for D-RISE is ≈ 0.1 on lattices and ≈ 0.7 on random regular graphs. We use these values for producing the scaling results in the Main Text.



(a) D-RISE: Value of $\alpha = 0.4$ for all the graphs. Value of $\beta = 0.8$ on Lattice-F and $\beta = 1.5$ on Lattice-SG. Value of $\beta = 1.2$ on RR-F and $\beta = 1.8$ on RR-SG.



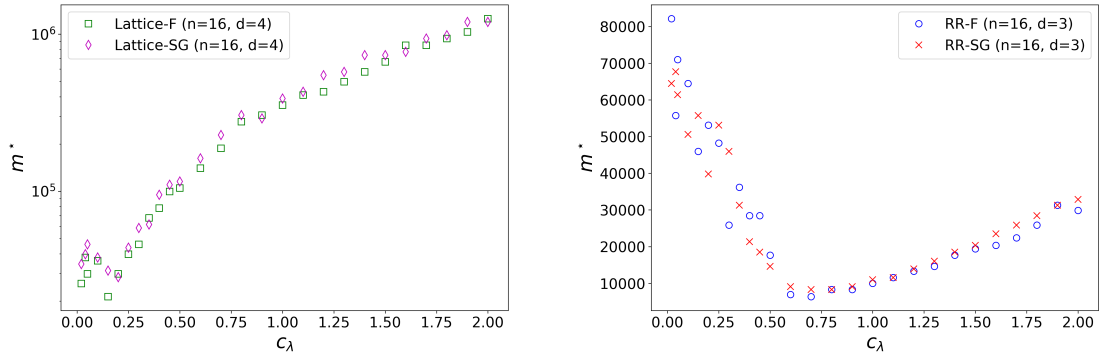
(b) D-RPLE: Value of $\alpha = 0.4$ for all the graphs. Value of $\beta = 0.8$ on Lattice-F and $\beta = 1.5$ on Lattice-SG. Value of $\beta = 1.2$ on RR-F and $\beta = 1.8$ on RR-SG.

Figure S1: **Empirical selection of c_λ in T-regime** We assess the dependence of the number of samples $m^\star$ on the regularization coefficient c_λ for the estimators of D-RISE and D-RPLE for successful structure reconstruction of Ising models of size $n = 16$. The different Ising model topologies considered are: (Lattice-F) ferromagnetic model on a periodic lattice as in Figure 1aA, (Lattice-SG) spin glass model on a periodic lattice as in as in Figure 1aC, (RR-F) ferromagnetic model on a random regular graph as in Figure 1a, and (RR-SG) spin glass model on a random regular graph as in Figure 1aD.

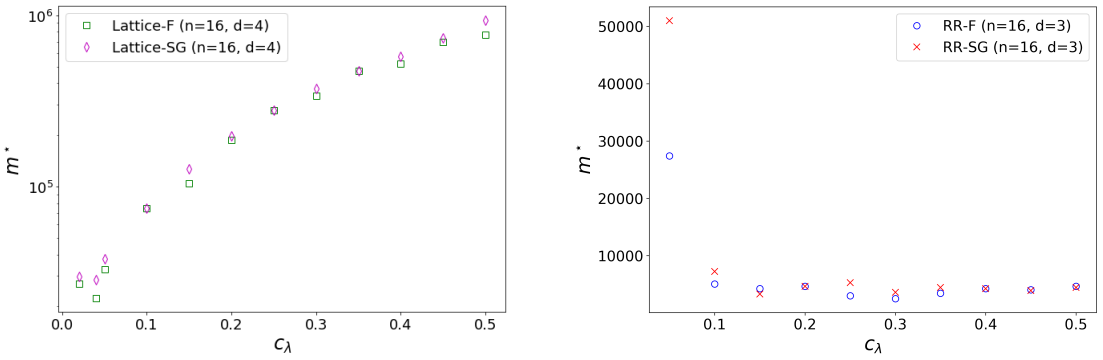
S4. Learning Random Regular Graphs in the M-regime

Here we discuss how structure learning in the M-regime can result in a sample complexity independent of $\beta = \max_{(i,j)\in E}|J_{ij}|$. The central object of our study is the conditional probability distribution in Eq. (2). For simplicity we consider the situation where the magnetic field is zero and we can rewrite this conditional distribution as

$$p(\sigma_i^{t+1}|\underline{\sigma}^t) = \frac{1 + \sigma_i^{t+1} \tanh \beta \left(\sum_{j\neq i} x_j \sigma_j^t\right)}{2}, \tag{S109}$$



(a) D-RISE: Value of $\alpha = 0.4$ for all the graphs. Value of $\beta = 1.5$ on lattices and random regular graphs.



(b) D-RPLE: Value of $\alpha = 0.4$ for all the graphs. Value of $\beta = 1.5$ on lattices and $\beta = 2.6$ on random regular graphs.

Figure S2: **Empirical selection of c_λ in M-regime** We assess the dependence of the number of samples $m^\star$ on the regularization coefficient c_λ for the estimators of D-RISE and D-RPLE for successful structure reconstruction of Ising models of size $n = 16$. The different Ising model topologies considered are: (Lattice-F) ferromagnetic model on a periodic lattice as in Figure 1bA, (Lattice-SG) spin glass model on a periodic lattice as in as in Figure 1bC, (RR-F) ferromagnetic model on a random regular graph as in Figure 1b, and (RR-SG) spin glass model on a random regular graph as in Figure 1bD.

where $x_j = J_{ij}/\beta$ if $j \in \partial i$ and $x_j = 0$ otherwise. In order to analyse the learning problem when β is large, we look at the distribution in Eq. (S109) in the limit where β goes to infinity and we obtain the following expression,

$$\lim_{\beta\to\infty} p(\sigma_i^{t+1}|\underline{\sigma}^t) = \frac{1+\sigma_i^{t+1}\,\mathrm{sign}\left(\sum_{j\neq i} x_j\sigma_j^t\right)}{2}. \tag{S110}$$

The form of the conditional distribution in Eq. (S110) implies that the update of σ_i^{t+1} is with probability one equal to $\mathrm{sign}\left(\sum_{j\neq i} x_j\sigma_j^t\right)$ whenever $\sum_{j\neq i} x_j\sigma_j^t \neq 0$, otherwise σ_i^{t+1} is updated to -1 or 1 with equal probabilities. We see in the limit of large β that the structure learning problem transforms essentially into the so-called noiseless one-bit compressive sensing problem (Boufounos & Baraniuk, 2008). In noiseless one-bit compressive sensing, we receive $t \in [1, m]$ observations of signs $y \in \mathbb{R}^m$ of the components of an unknown d-sparse vector $\underline{x} \in \mathbb{R}^n$ transformed by a known sensing matrix $A \in \mathbb{R}^{m\times n}$ i.e. $\underline{y} = \mathrm{sign}(A\underline{x})$. The objective in one-bit compressive sensing is to recover the support of $\underline{x}$ just like in our structure learning problem. In order to recover the support of $\underline{x}$, the number of observations (and the rank of A) has to be at least $m = \Omega(d^2 \ln n/\ln d)$, see (Acharya et al., 2017). The difference between compressive sensing and our structure learning problem lies in the choice of the sensing matrix. While in compressive sensing, the design of the sensing matrix is left to operators, in our case it is imposed by the distribution $p(\underline{\sigma}^t)$ as the rows of our sensing matrix correspond to i.i.d. samples of spin configurations $A_{tj} = \sigma_j^t$.

In the T-regime we expect the Glauber dynamics to mix rapidly and generate samples from the equilibrium distribution of the graphical model. In the limit $\beta \to \infty$, the equilibrium distribution is supported only by spin configurations whose energies are minimal. The number of such configurations is typically constant with respect to the system size. For example,

ferromagnetic models have no more than two states of minimal energy regardless of the number of spins. Therefore, our sensing matrix only contains a fixed number of independent rows and the one-bit compressive sensing problem cannot be solved perfectly as rank$(A) = O(1)$. It implies that for large β, structure learning with a fixed number of i.i.d. samples from the equilibrium distributions cannot be accomplished.

In the M-regime, however, we carry out one step of Glauber dynamics and the distribution $p(\underline{\sigma}^0)$ is uniform. Our sensing matrix turns out to be generated from a Bernoulli ensemble with entries equal to -1 and $+1$ at random. Such matrices are known to have a rank m with high probability (Kahn et al., 1995). This renders possible the inversion of the one-bit compressed sensing problem and consequently the possibility to solve the structure learning problem with a number of observations independent of β for β large. However, matrices with random signs do not necessarily lead to an invertible one-bit compressive sensing problem and the invertibility of the problem depends on the hidden vector $\underline{x}$. For instance, consider the simple four dimensional vectors with different support $u = (1, 1, 1, \epsilon)$ where $|\epsilon| \in (0, 1)$ and $v = (1, 1, 1, 0)$. It is easy to see that for any configurations of $\sigma_i \in \{-1, 1\}$ we have that $\text{sign}\left(\sum_i \sigma_i u_i\right) = \text{sign}\left(\sum_i \sigma_i v_i\right)$ for the quantity $\sigma_4 u_4 = \epsilon\sigma_4$ has always a smaller magnitude than $|\sigma_1 + \sigma_2 + \sigma_3| \geq 1$. Therefore, structure learning with a fixed number of samples for large β cannot be done for neighborhood of size four and couplings equal to $(\beta, \beta, \beta, \alpha)$, which explains the exponential scaling seen for the lattice instance in Fig.1b. We have a different story when we consider the three dimensional vector $w = (1, 1, \epsilon)$ with $|\epsilon| \in (0, 1)$. If we take a sensing matrix equal to $A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & -1 \end{pmatrix}$ when $\epsilon > 0$ or $A = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix}$ when $\epsilon < 0$, we see that the only three dimensional vectors x that satisfy the equation $\text{sign}(Ax) = \text{sign}(Aw)$ are those for which $x_1 > 0$, $x_2 > 0$ and $x_3 \cdot \text{sign}(\epsilon) > 0$. This means that the (signed) support of w is recoverable with a sensing matrix from the Bernoulli ensemble. It implies that structure learning with a fixed number of samples for large and even infinite β is possible for neighborhood of size three and couplings equal to (β, β, α) which explains the flat curves seen for the three-regular graphs in Fig.1b. Based on these considerations, we can extrapolate this behavior in a straightforward manner to graphs with odd and even degree d having $d - 1$ identical couplings.

S5. Learning dynamics from neural spike trains

In this section, we describe how the relevant dataset from (Prentice et al., 2016) is processed to be used for learning Ising dynamics with D-RISE/D-RPLE in this study. Then, we discuss the statistics of learned model parameters, and compare results obtained with D-RISE and D-RPLE. Finally, we provide details on computation of data moments (such as correlations) from the learned Ising model to assess performance against the dataset.

S5.1. Preparation of neural dataset

The dataset contains spike trains from 152 salamander retinal ganglion cells in response to a non-repeated natural movie stimulus, of which we select spike trains for $n = 42$ neurons over 24s for our application. To obtain a time series of spin configurations over the neurons, we bin the spike trains into 20 ms time bins. The spin $\sigma_i^{(t)}$ of a neuron i in time bin t is set to 1 if it fires at least once in this time bin and -1 otherwise. We thus produce a sequence of 1.2×10^5 spin configurations (also called binary spike words). A segment of the sequence is shown as a spike raster in Figure 2. However, this can't be used directly for learning an Ising model using D-RISE or D-RPLE.

Our learning algorithms require information about the identity of nodes being updated which isn't directly available from the data recorded. The identity of the node being updated at time t is however known when there is some $l \in [n]$ for which the spin of node/neuron l flips in sign i.e., $\sigma_l^{(t+1)} = -\sigma_l^{(t)}$. There maybe more than one such node at time t. However, in Glauber dynamics, only one node is selected for update at time t. We thus only consider samples of spin configurations $\{(\underline{\sigma}^t, \underline{\sigma}^{t+1}, I^{t+1})\}$ for time bins t where $I^{(t+1)}$ can be directly inferred by searching for nodes which flipped its spin and there is only one such node. The resulting set of samples are time ordered but samples from consecutive time bins may not be chosen. Thus, it is convenient to represent the samples as $\{(\underline{\sigma}^{0(k)}, \underline{\sigma}^{1(k)}, I^{1(k)})\}_{k \in [m]}$ where k is now used to index the samples and t_k corresponds to the time t of the k-th time bin chosen. After this processing, we end up with a set of 3.2×10^4 samples corresponding to the M-regime with an unknown distribution over the initial spin configurations.

Let us briefly comment on some challenges associated with the dataset preparation procedure that we used, and point out

to directions for overcoming these limitations. As we require the identity of the updated nodes for D-RISE/D-RPLE, this requires us to only select samples where a node is observed to be flipped. Otherwise, we wouldn't know a node has been updated or not. The model fit would improve if the node identities weren't required and thus samples were no flips are observed could also be used for estimation. This however requires deriving estimators beyond D-RISE/D-RPLE that is outside the scope of this work, but would be an interesting direction for future studies. The bin size was chosen to be 20ms as this is the expected time scale of persistence of modes in the neural spike trains (Prentice et al., 2016) and was reused. Reducing the bin size decreases the probability of observing a spike or a node activation in that time bin leading to more samples with no updates. Increasing the bin size increases the probability of observing a spike or node activation in the time bin but there may also be more than one spike in the time interval for a given node which is counted only once. Choosing the bin size appropriately ensures that there is enough samples and that the Poisson rate of observation matches closely with the Poisson rate of node updates. We would expect the model fit to overfit if the bin size is too high and be more computationally expensive in the case where bin size is small.

Finally, it would be interesting to go beyond the assumption of Glauber dynamics for constructing an effective model of the data, making use of all available samples that can be costly to get. Possible extensions of our framework include: (i) considering samples with only spin history and thus not requiring identities of updated nodes, (ii) accounting for multiple nodes being updated at the same time e.g., akin to block Gibbs sampling, and (iii) considering more general Markov chain dynamics. We leave these extensions to future work.

S5.2. Statistics of learned model parameters

Ising model parameters learned using D-RISE on the set of samples prepared as discussed in the previous section are shown in Figure S3. If the task is to learn an effective Ising model for explaining the dynamics, these parameter estimates can be used directly. However, if the task is that of a model selection, i.e. learning the network structure of the model, then the following procedure can be used. In the histogram over $\hat{J}_{ij}$, we observe gaps separating a group of estimated couplings in the vicinity of zero from those with higher intensities in absolute value. We choose the threshold δ to correspond to the first symmetrical gap around zero ($\delta_- = -\delta$ and $\delta_+ = \delta$) that separates these groups. All the coupling parameters $|\hat{J}_{ij}| < \delta$ obtained after structure learning and shown in red in the histogram of Figure S3 would be set to zero. Observing these gaps and ability to choose a clear threshold indicates that the number of samples is sufficient for structure learning.

The resulting $\hat{J}_{ij}$ estimates indicate a low value of β and thus a high effective temperature of the model. Most of the couplings are weak with few strong couplings. This is in agreement with other such studies. The effective Ising model that we learn here from Glauber dynamics can be used for predicting higher order moments of the data and understanding the behavior of this population of neurons. The difference in correlation matrices computed from data assuming the samples are i.i.d. and that respecting time (presented in the Main Text) explain the difference in parameter estimates $\hat{J}_{ij}$ obtained through RISE and D-RISE as visualized in Figure S4. This once again highlights the importance of respecting dynamics and hence time correlations in the data when learning an effective Ising model.

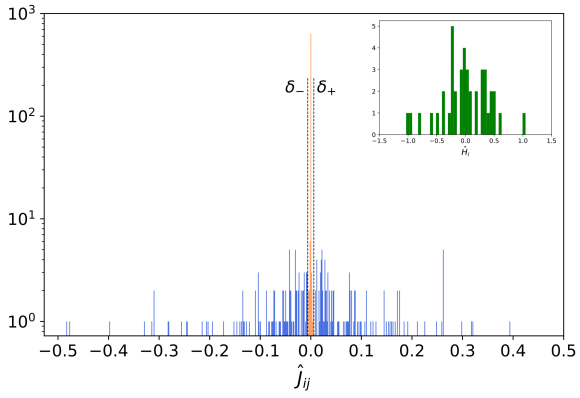


Figure S3: Ising model parameters learned from spike trains over 42 neurons using 3.2×10^4 samples. Significant couplings are in blue and thresholded couplings are in red. Reconstructed fields are in green in a separated histogram.

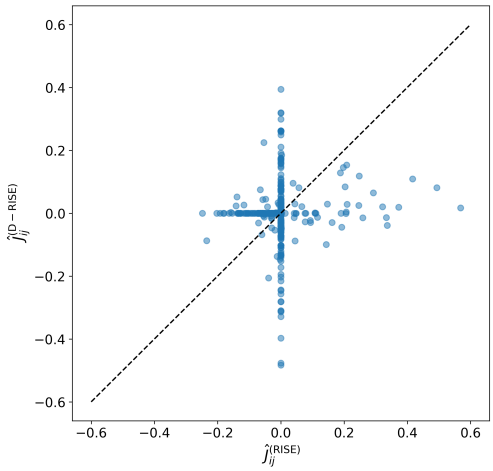


Figure S4: Comparison of Ising model parameter estimates $\hat{J}_{ij}$ obtained through RISE and D-RISE

S5.3. Computation of correlations

To assess the performance of the learned Ising model, it is common to compare predicted correlations from the model against the data. Here, we obtain predicted correlations by computing correlations on a dataset simulated using the learned Ising model under the M-regime and running Glauber dynamics. We now explain how this simulated dataset is constructed.

We note that the covariance of interest (including the corresponding probabilities) is given by

$$\mathrm{Cov}(\sigma_i^0, \sigma_j^1) = \mathbb{E}_{p(\sigma_j^1|\sigma_i^0, I^1)p(\sigma_i^0)p(I^1)}[\sigma_i^0\sigma_j^1] - \mathbb{E}_{p(\sigma_i^0)}[\sigma_i^0]\mathbb{E}_{p(\sigma_j^1|\sigma_i^0, I^1)}[\sigma_j^1] \tag{S111}$$

We note from the above expression that the resulting covariance not only depends on the Ising model dynamics which explains $p(\sigma_j^1|\sigma_i^0, I^1)$ but also the initial distribution over spin configurations $p(\sigma_i^0)$ and the probability of a node being chosen for update $p(I^1)$. In order to ensure we can compare the predicted covariance from the Ising model against that from data, we use the same $p(\sigma_i^0)$ and $p(I^1)$ as in the experimental dataset. Additionally, we note that the spin configurations in the experimental dataset only contains flipped spin configurations which also needs to be respsected.

Thus, we construct the simulated dataset by running Glauber dynamics on the M-regime using the same $p(\sigma^0)$ and $p(I^1)$ as from the experimental dataset, and only including those σ^1 where there is a flip in the sign of node given by I^1. Moments are then computed on this simulated dataset using the usual estimators for population means, covariances, correlations, etc., from the samples.