Optimal structure and parameter learning of Ising models

Andrey Lokhov

Theoretical Division and Center for Nonlinear Studies
Los Alamos National Laboratory

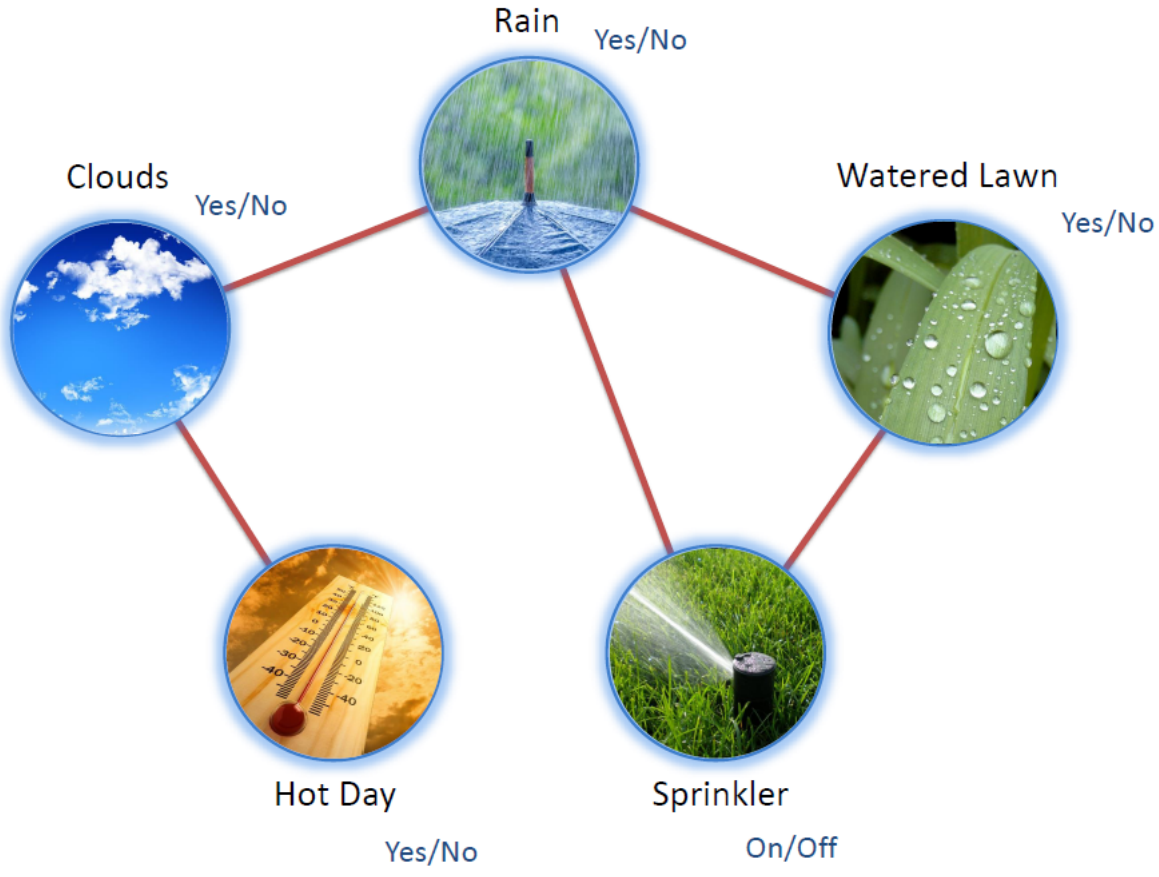
Vuffray, Misra, Lokhov, Chertkov, *NIPS*, 2016

Lokhov, Vuffray, Misra, Chertkov, *arXiv:1612.05024*, 2016

Information Sciences Institute

Graphical models: a quick example

Graphical model = graph of conditional dependencies
Node = random variable
Edge = direct correlation



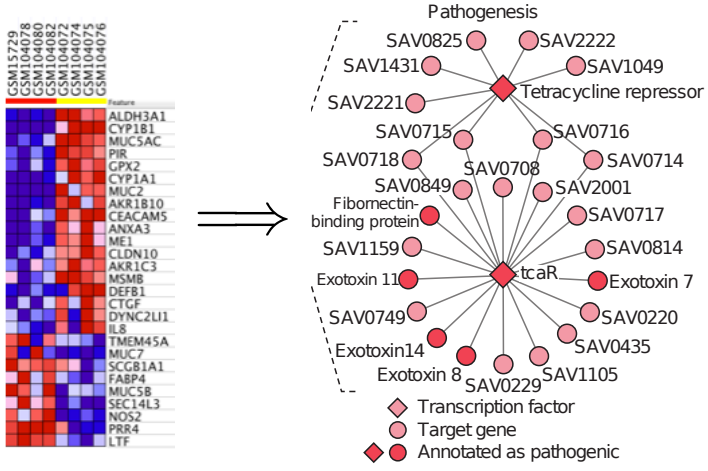
"Watered Lawn" conditioned on "Rain" is independent on "Clouds" and "Hot Day", but still depends on "Sprinkler"

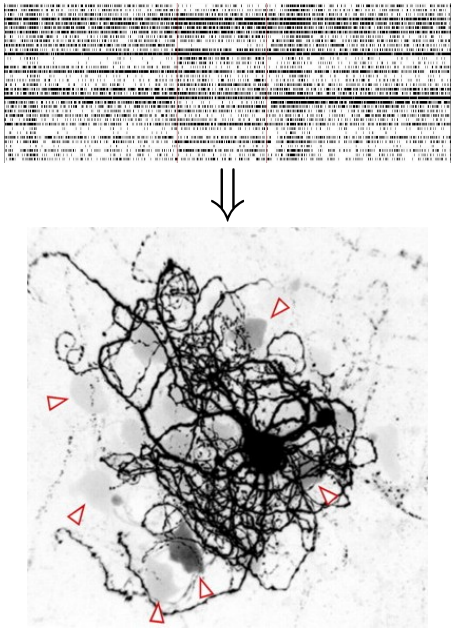
What is the network reconstruction problem?

- ✓ Can we reconstruct the graph from daily observations?
- ✓ Can we estimate the direct correlations strengths?
- ✓ How many days of observations do we need?

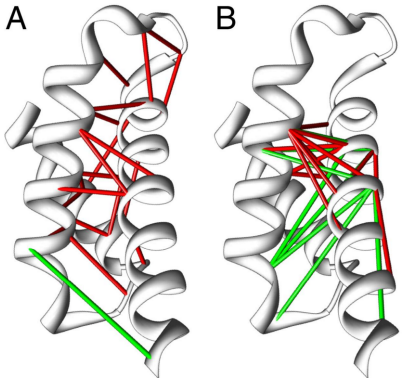
	Day 1	Day 2	Day 3	...
Clouds	Yes	No	Yes	...
Rain	No	No	Yes	...
Sprinkler	No	Yes	No	...
Hot Day	No	Yes	Yes	...
Watered Lawn	No	Yes	Yes	...

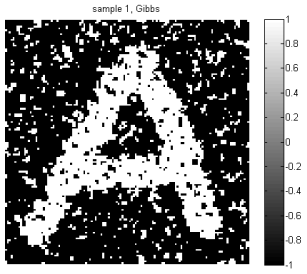
Applications of reconstruction problem: unknown models





- ✓ Gene expression
- ✓ Neuroscience
- ✓ Protein structure
- ✓ Image Segmentation
- ✓ Sociology, *etc.*





Challenges: small number of samples, noisy data

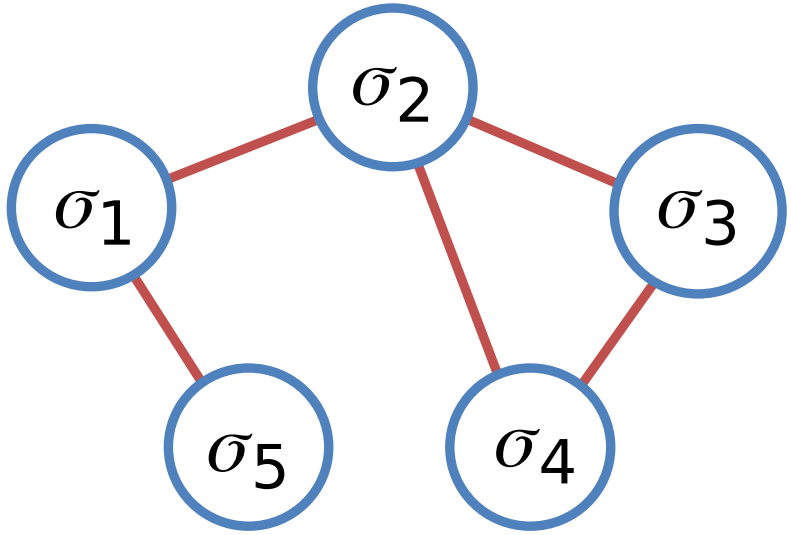
Precise formulation: Ising model

Binary variables ($\downarrow\uparrow$, on/off, 👍👎) associated with nodes:

$$\{\sigma_1, \ldots, \sigma_N\} \in \{-1, +1\}^N$$

Probability distribution on a graph $G = (V, E)$:

$$P(\sigma_1, \ldots, \sigma_N) = \frac{1}{Z} \exp\Bigg(\underbrace{\sum_{(i,j)\in E} J^*_{ij}\sigma_i\sigma_j}_{\text{direct correlations}} + \underbrace{\sum_{i\in V} H^*_i\sigma_i}_{\text{prior}} \Bigg)$$



Precise formulation: reconstruction problem

	$k = 1$	$k = 2$	...	$k = \mathrm{M}$
σ_1	+1	−1	...	+1
σ_2	−1	−1	...	−1
⋮	⋮	⋮	...	⋮
σ_N	+1	+1	...	−1

Number of variables: N

Number of samples: M

Maximum node degree: d

Couplings: $\alpha \leq |J^*_{ij}| \leq \beta$

Can we reconstruct the edge set perfectly with high probability?

How many samples M do we need for given N, d, α and β?

For simplicity, we assume for the moment that $H_i = 0$.

"Sufficient statistics"

Three facts:

1. Direct maximization of the **likelihood:** intractable (computing Z is hard!)

$$\ell = \sum_{i \in V} H_i m_i + \sum_{(i,j) \in E} J_{ij} c_{ij} - \log Z$$

$$\partial_{H_i} \log Z = m_i, \quad \partial_{J_{ij}} \log Z = c_{ij}$$

2. **Boltzmann machine:** estimation of magnetizations and pair-correlations by sampling: slow

3. However, knowledge of **first and second moments** is "enough"

Approximate methods: "sufficient statistics" $\rightarrow$ full data

Boltzmann machine Ackley *et al.* 1985	No control over sample complexity Computationally intractable
Mean field methods Tanaka 1998 Kappen and Rodriguez 1998 Roudi *et al.* 2009 Ricci-Tarsenghi 2012 Nguyen and Berg 2012	No guarantees of reconstruction No control over sample complexity Fail for strong direct correlations
Adaptive cluster expansion Cocco and Monasson 2011	No control over sample complexity Computationally expensive
Pseudo-likelihood method with "sparsistency" assumption Ravikumar *et al.* 2010	Fails for strong direct correlations Bento and Montanari 2009

Information Theory bounds

Lower bound:

$$M_{\text{opt}} \gtrsim \frac{e^{\beta d}}{\alpha^2} \log N$$

Upper bound:

$$M_{\text{opt}} \lesssim \frac{e^{4\beta d}}{\alpha^2} \log N$$

Exact methods and rigorous guarantees

	Computational complexity	**Sample complexity**
Theoretical bound for sample complexity Santhanam and Wainwright 2012		$\frac{e^{c\beta d}}{\alpha^2}\log N$, $c \in [1,4]$
Exhaustive neighborhood search Bresler, Mossel, Sly 2013	$\mathcal{O}(N^d)$	$\mathcal{O}(\log N)$
Mutual-information based greedy search Bresler 2015	$\exp\left(\frac{e^{c_1\beta d}}{\alpha^{c_2}}\right) N^2 \log N$	$\exp\left(\frac{e^{c_1\beta d}}{\alpha^{c_2}}\right) \log N$
Regularized Pseudo-Likelihood	$\frac{e^{8\beta d}}{\alpha^2} N^2 \log N$	$\frac{e^{8\beta d}}{\alpha^2} \log N$
Regularized Interaction-Screening AL, Vuffray, Misra, Chertkov 2016	$\frac{e^{6\beta d}}{\alpha^2} N^2 \log N$	$\frac{e^{6\beta d}}{\alpha^2} \log N$

Popular method: pseudo-likelihood with ℓ_1 penalty

Local approximation to the **likelihood**:

$$\mathcal{L}_i(J_i) = \frac{1}{M}\sum_{m=1}^{M} \ln \frac{1}{1+\exp(-2\sigma_i^{(m)}\sum_{j\neq i} J_{ij}\sigma_j^{(m)})},$$

where $J_i = \{J_{ij}\}_{j\neq i}$. **Optimization problem**:

$$\widehat{J}_i = \arg\max_{J_i}\Big[\mathcal{L}_i(J_i) - \lambda\|J_i\|_1\Big], \quad \text{where } \|J_i\|_1 = \sum_{j\neq i}|J_{ij}|$$

"Sparistency" assumption: the support of the non-zero couplings is correctly reconstructed. Valid only for the high-temperature models

Bento and Montanari, 2009

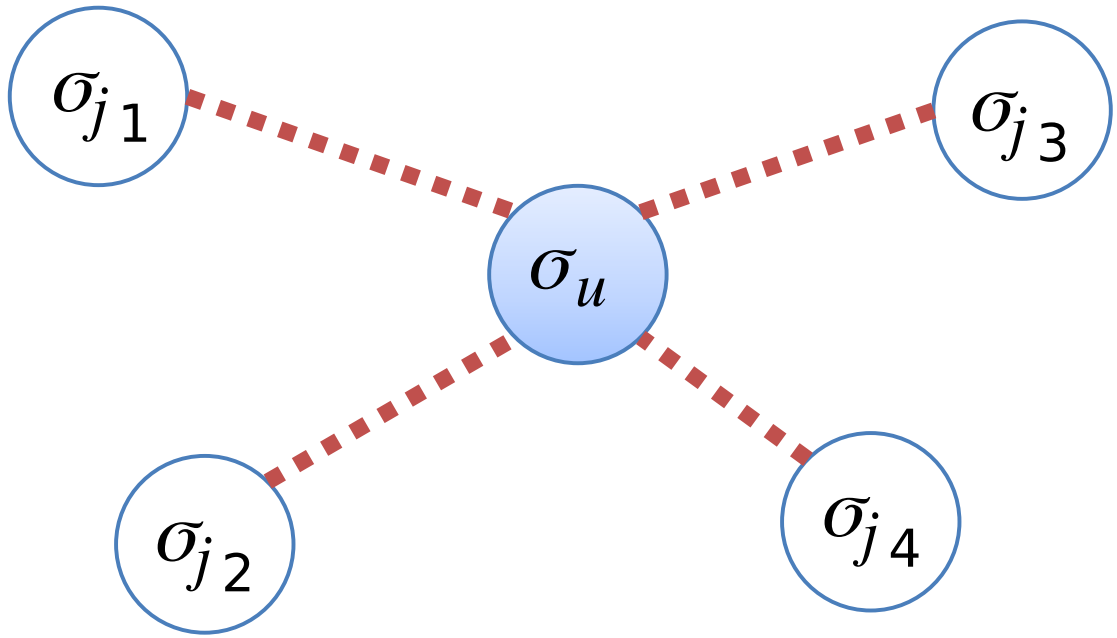
Thresholding: leads to an exact estimator

AL, Vuffray, Misra, Chertkov 2016

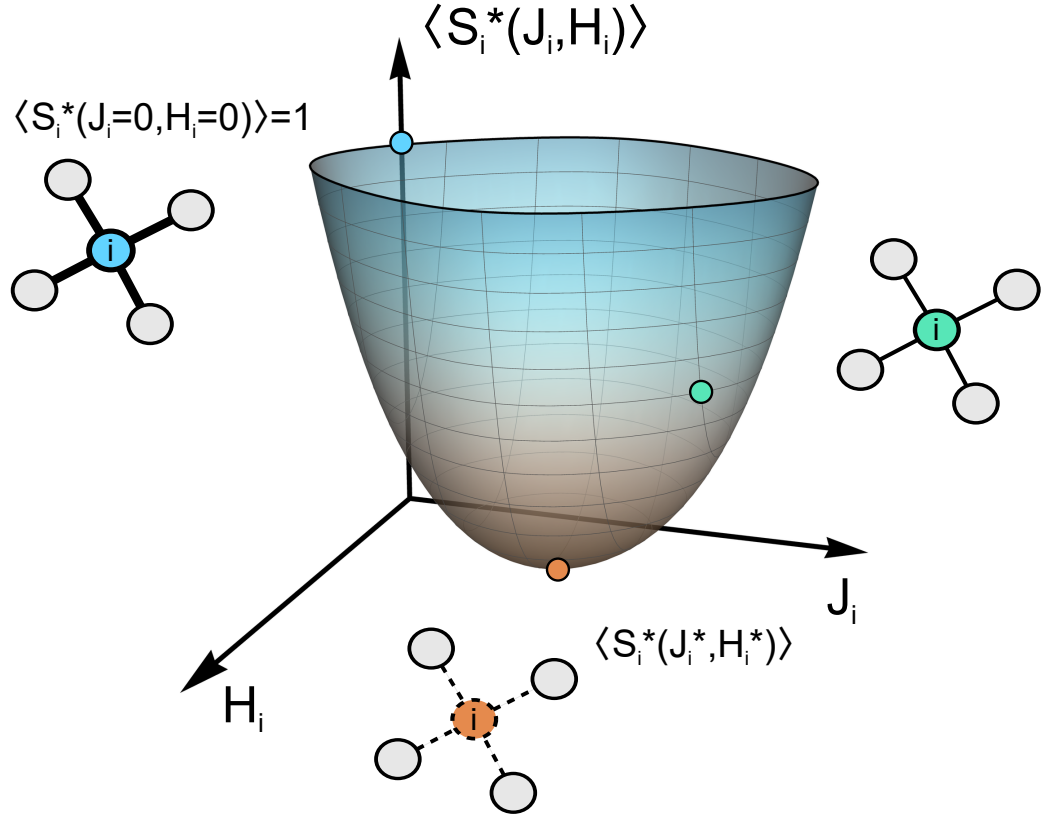
Interaction Screening Objective

$$\mathcal{S}_i(J_i) = \frac{1}{M}\sum_{m=1}^{M} \exp(-\sum_{j\neq i} J_{ij}\sigma_i^{(m)}\sigma_j^{(m)})$$

Physical interpretation: the interaction screening loss applies counter-interactions to virtually make σ_u independent:



Consistency and interpretation in the $M \to \infty$ limit



$$\frac{\partial \mathcal{S}_i^*}{\partial J_{ik}}(J_i^*) = -\langle \sigma_i \sigma_k \exp\Big(-\sum_{j\neq i} J_{ij}^* \sigma_i \sigma_j\Big)\rangle_{P(\sigma)}$$

$$\propto \sum_{\sigma_1,\ldots,\sigma_N} \sigma_i \sigma_k \exp\Big(-\sum_{j\neq i} J_{ij}^* \sigma_i \sigma_j\Big) \exp\Big(\sum_{(r,s)\in E} J_{rs}^* \sigma_r \sigma_s\Big)$$

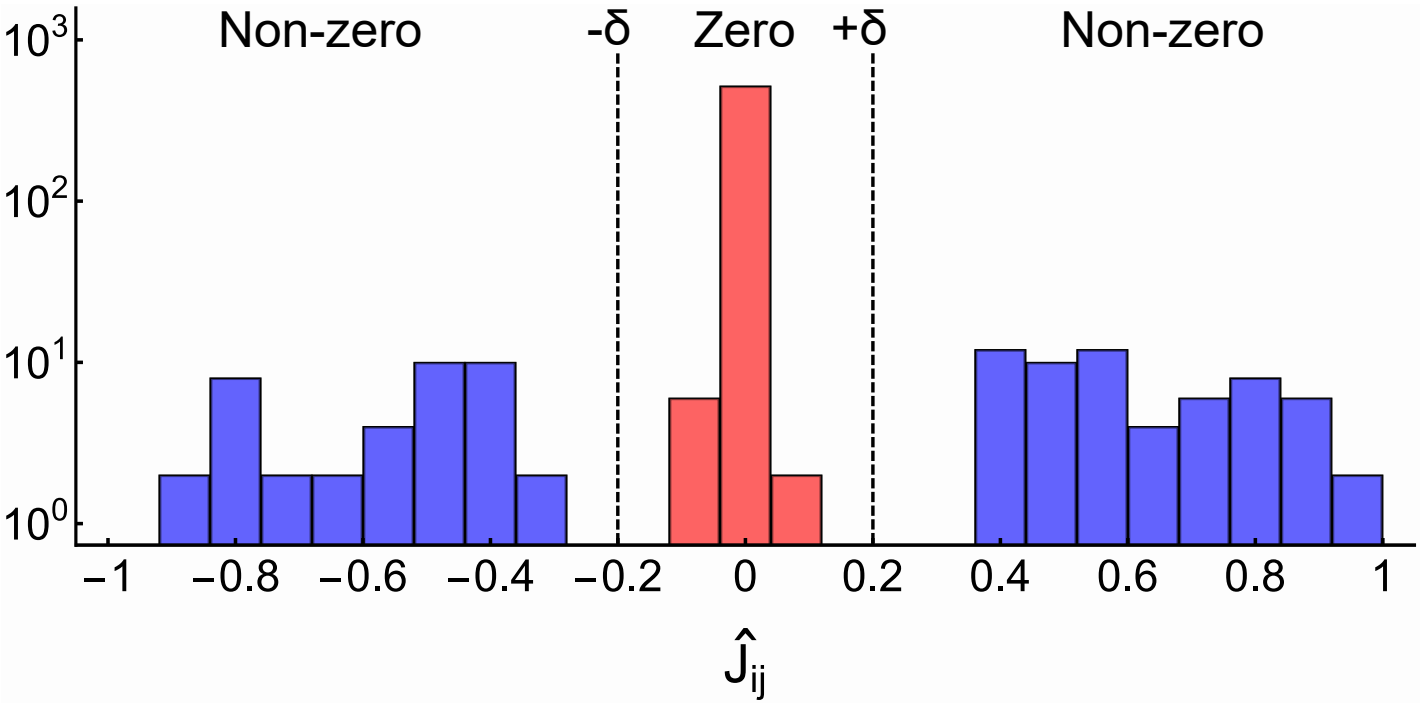
$$= \sum_{\sigma_1,\ldots,\sigma_N} \sigma_i \sigma_k \exp\Big(\sum_{(r,s)\in E\setminus i} J_{rs}^* \sigma_r \sigma_s\Big) = 0.$$

Thresholding for finite number of samples M

Regularized Interaction Screening Estimator:

$$\widehat{J}_i = \arg\min_{J_i} \left[\mathcal{S}_i(J_i) + \lambda \|J_i\|_1 \right]$$

Symmetrizing local estimations for finite M: $\widehat{J}_{ij} \leftarrow (\widehat{J}_{ij} + \widehat{J}_{ji})/2$



ER graph with $N = 25$ and $\langle d \rangle = 4$, spin glass couplings in $[-1.0, -0.4] \cup [0.4, 1.0]$ and $M = 5000$

Rigorous analysis of the RPLE and the RISE

Theorem The reconstruction error on the couplings (in the neighborhood of node i) of the RPLE with regularization parameter $\lambda = c_1\sqrt{\ln\left(N^2/\epsilon\right)/M}$ is bounded with probability $1-\epsilon/N$ as

$$\left\|\widehat{J}_i^{\mathrm{RPLE}} - J_i^*\right\|_2 \leq C_d e^{4\beta d}\sqrt{\frac{\ln\left(N^2/\epsilon\right)}{M}}.$$

For the RISE, the same error is estimated as

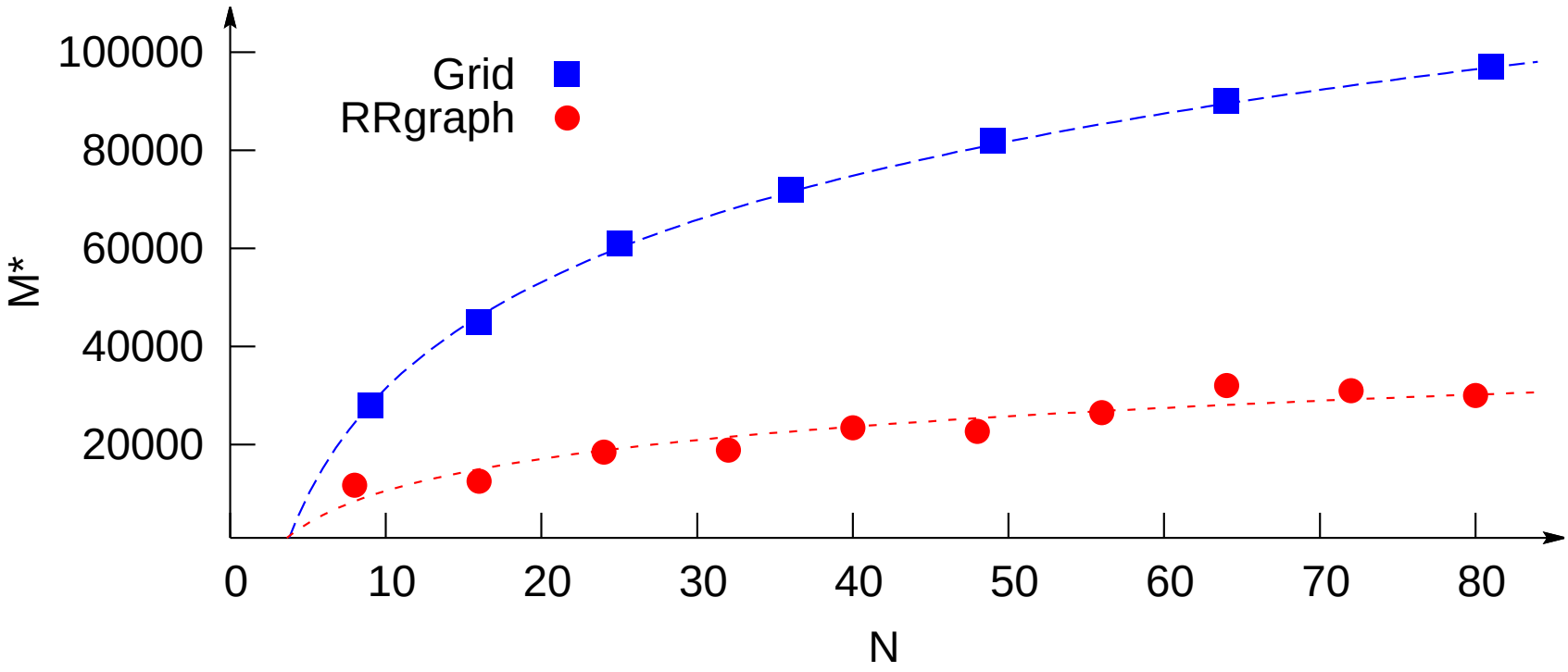
$$\left\|\widehat{J}_i^{\mathrm{RISE}} - J_i^*\right\|_2 \leq C'_d e^{3\beta d}\sqrt{\frac{\ln\left(N^2/\epsilon\right)}{M}},$$

where C_d and C'_d depend only polynomially on d, and c_1 is a constant.

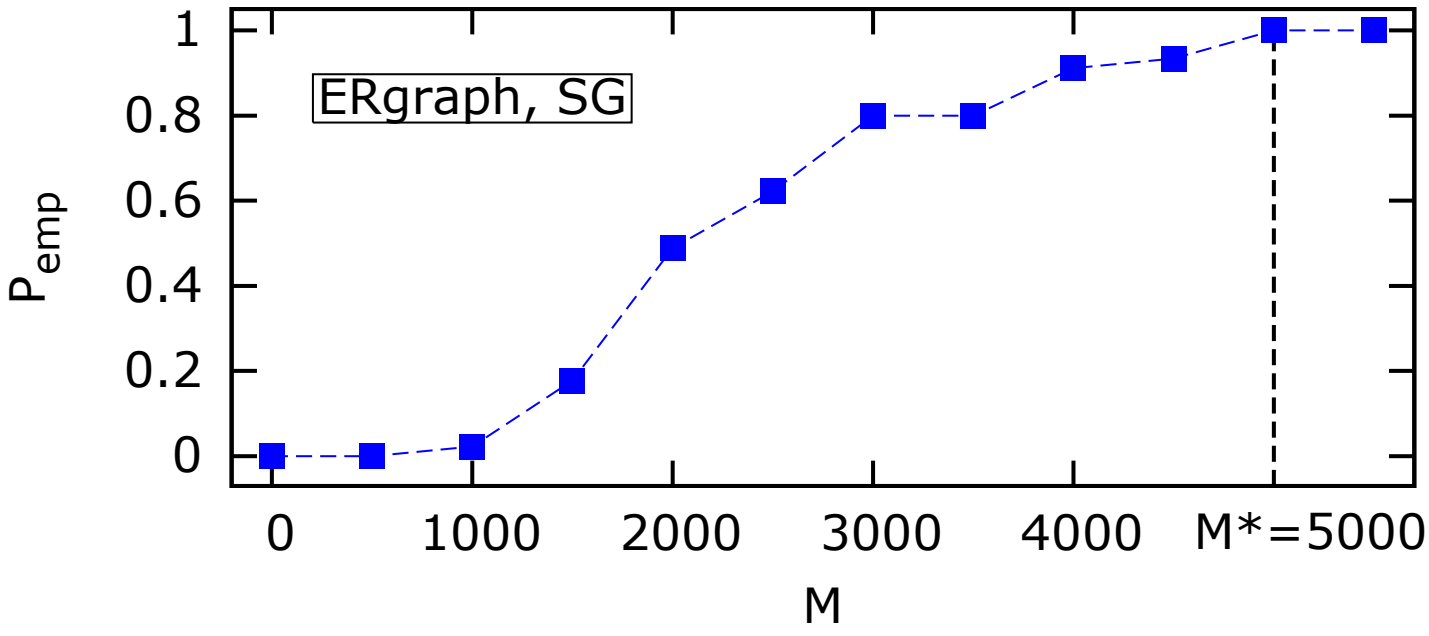
Proof using the theory of M-esimators [Negahban et al. 2009]

Check for the $\mathcal{O}(\log N)$ scaling

Ferromagnetic Ising model on a two-dimensional lattice with double periodic boundary conditions ($d = 4$) and random regular graphs ($d = 3$)

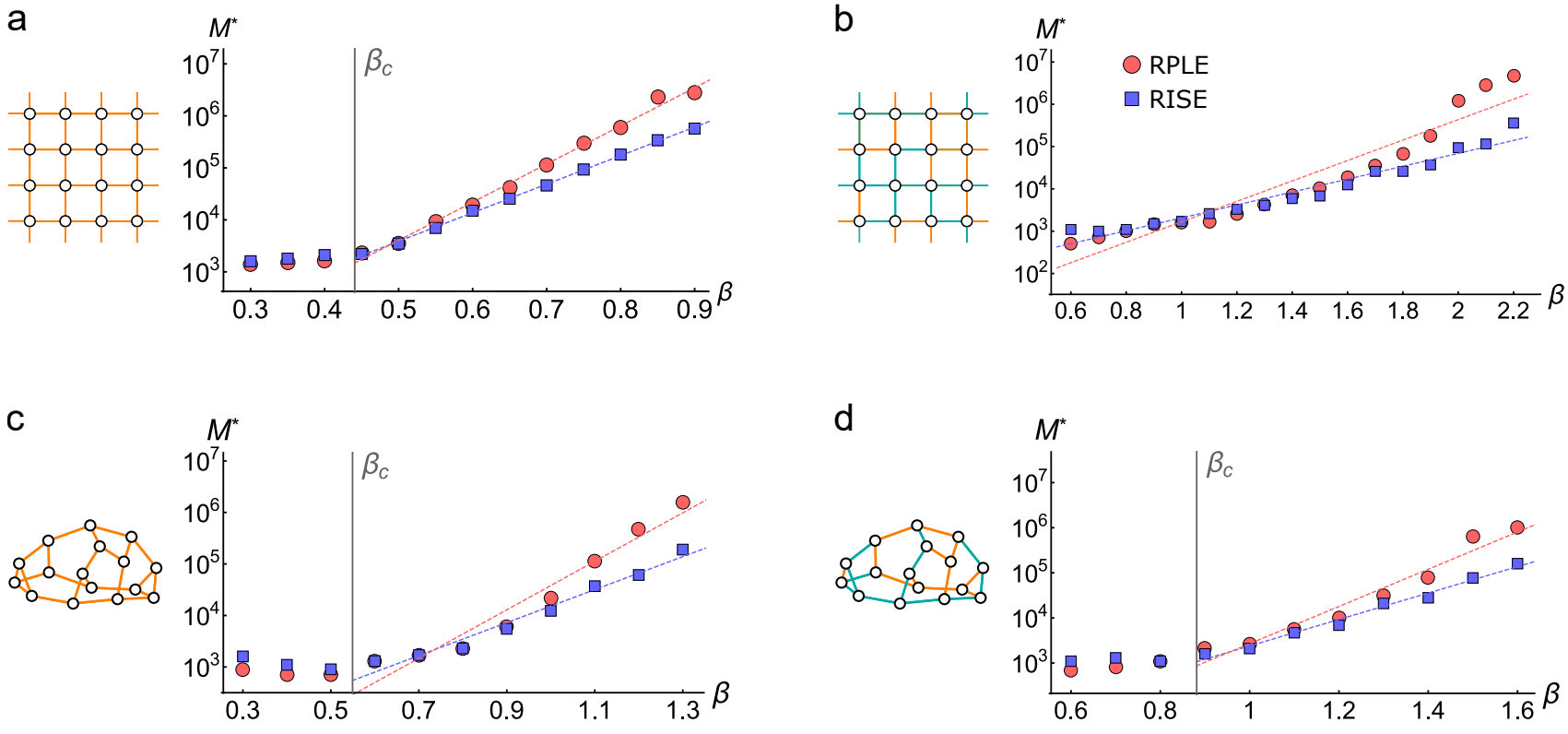


Remark: procedure for M^* selection



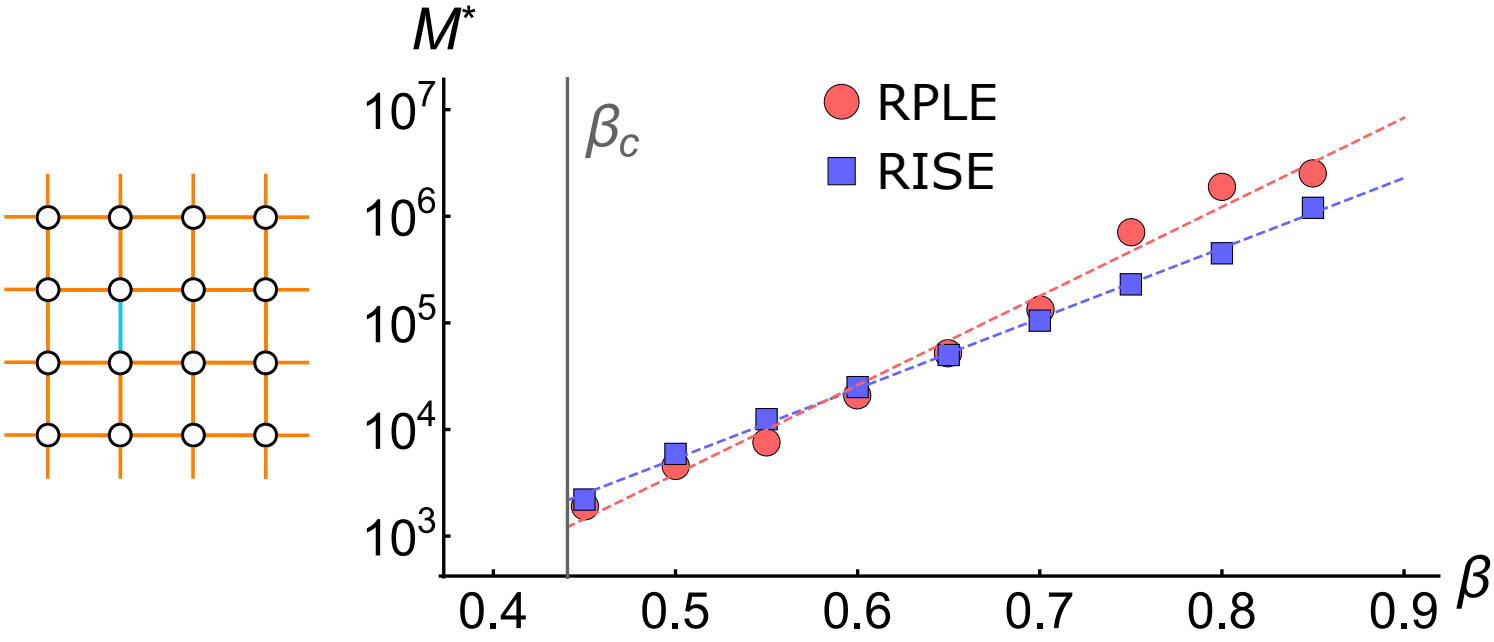
Obtained using averaging over 45 sets of samples, so that to obtain the probability of reconstruction at least 0.95 with 90% confidence

Empricial study at different temperatures



(a) F lattice (b) SG lattice (c) F RR graph (d) SG RR graph

Scaling with β in the hardest case

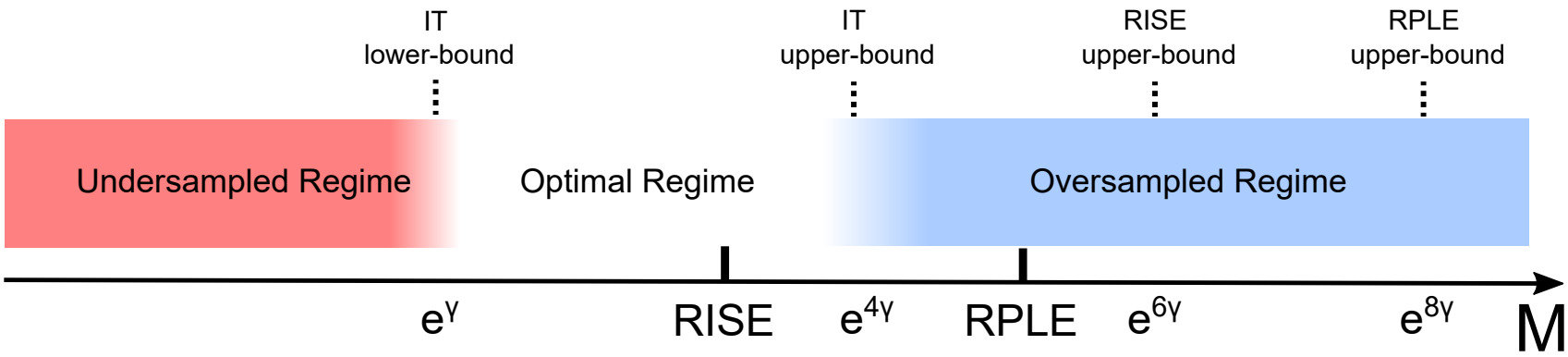


Ferromagnetic model on a two-dimensional lattice

RPLE: $M^* \propto e^{4.8\beta}$

RISE: $M^* \propto e^{3.8\beta}$

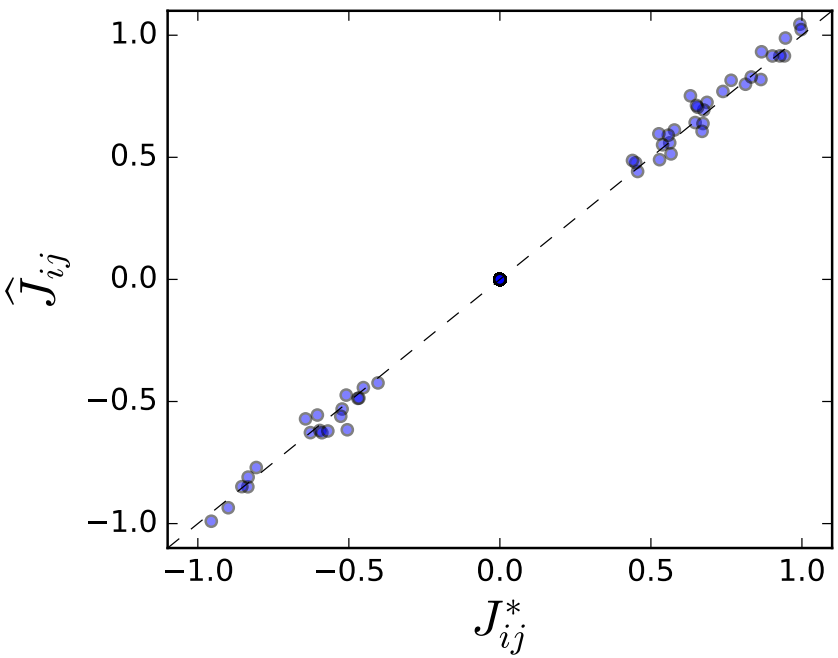
Summary of the main results

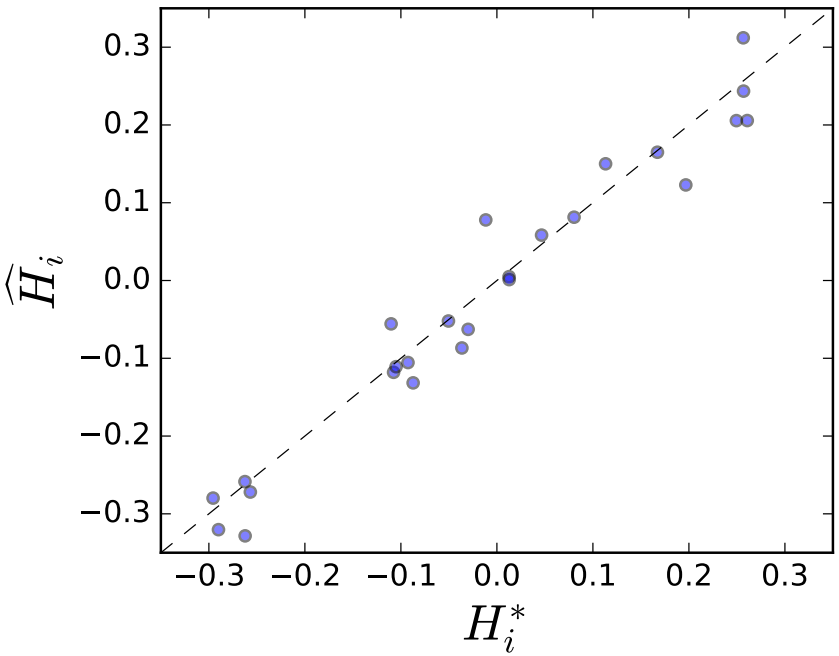


Reconstruction of model parameters

$$\mathcal{S}_i(J_i, H_i) = \frac{1}{M} \sum_{m=1}^{M} \exp(-\sum_{j \neq i} J_{ij} \sigma_i^{(m)} \sigma_j^{(m)} - H_i \sigma_i^{(m)}),$$

$$(\widehat{J}_i, \widehat{H}_i) = \arg \min_{(J_i, H_i)} \left[\ln S_i(J_i, H_i) + \lambda \|J_i\|_1 \right]$$





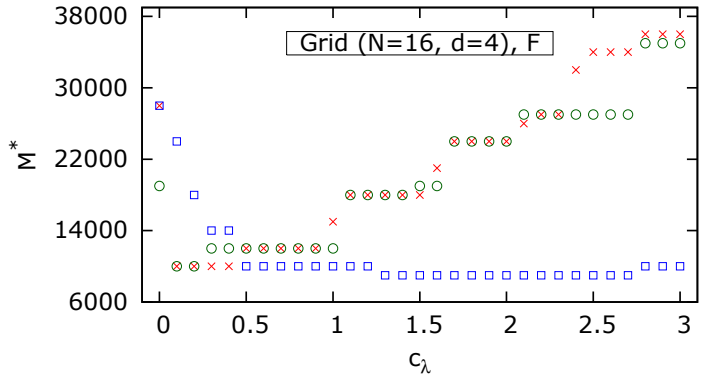
Summary

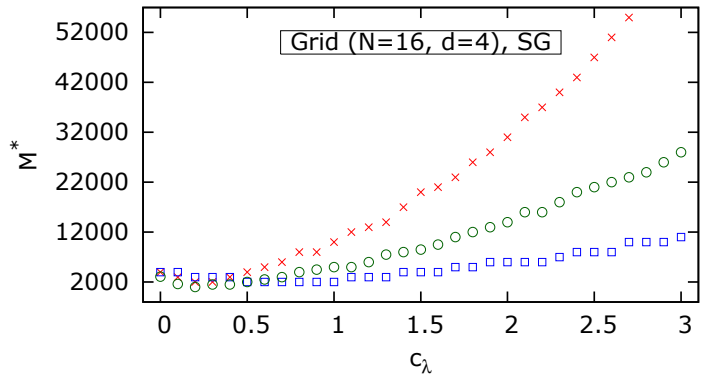
- ✓ Efficient estimator for structure and parameter learning of an arbitrary Ising model
- ✓ Provable guarantees and requirement of a near information theoretically optimal number of samples only

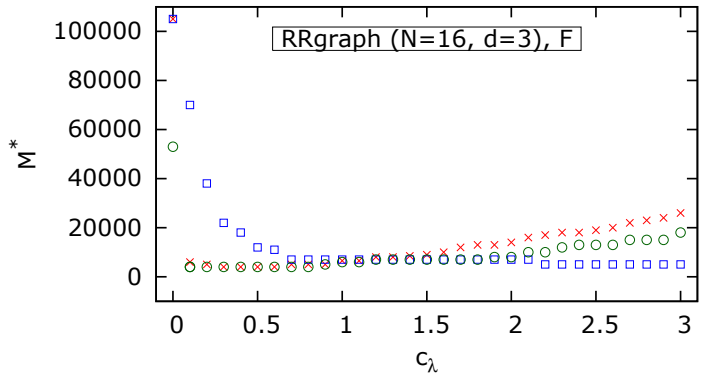
Extensions:

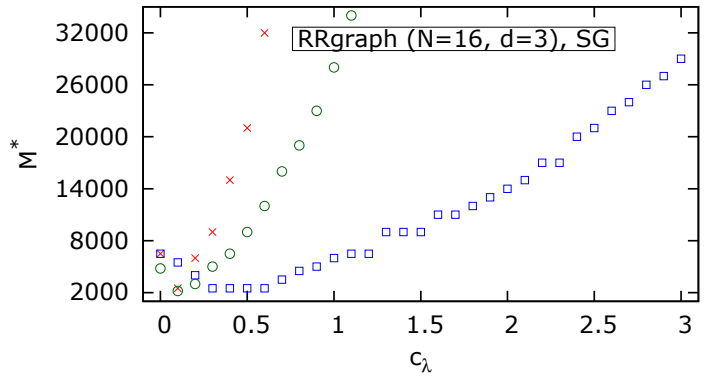
- ✓ Non-binary variables (Potts model and continuous variables)
- ✓ Non-pairwise interactions (3-body interactions and more)
- ✓ Hidden variables

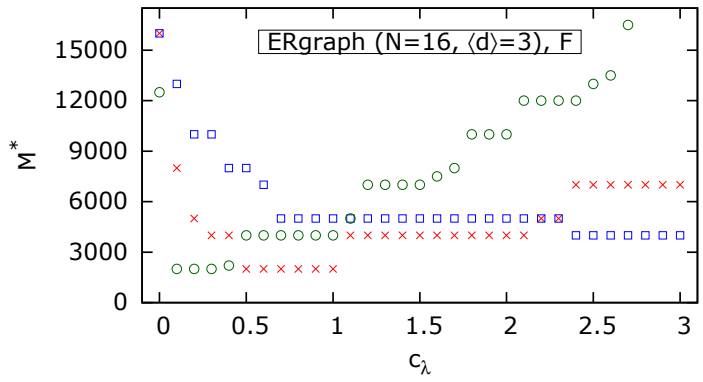
Appendix: λ selection

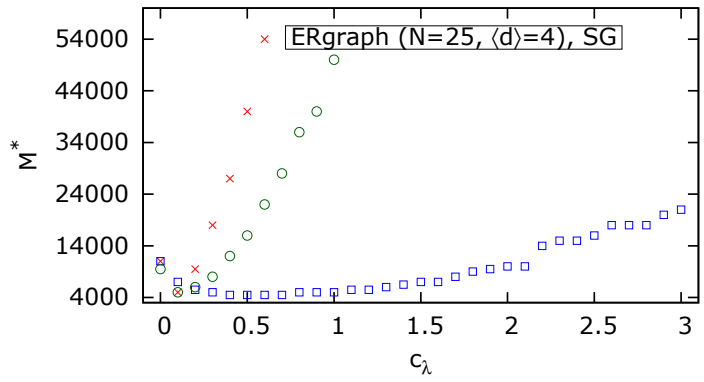




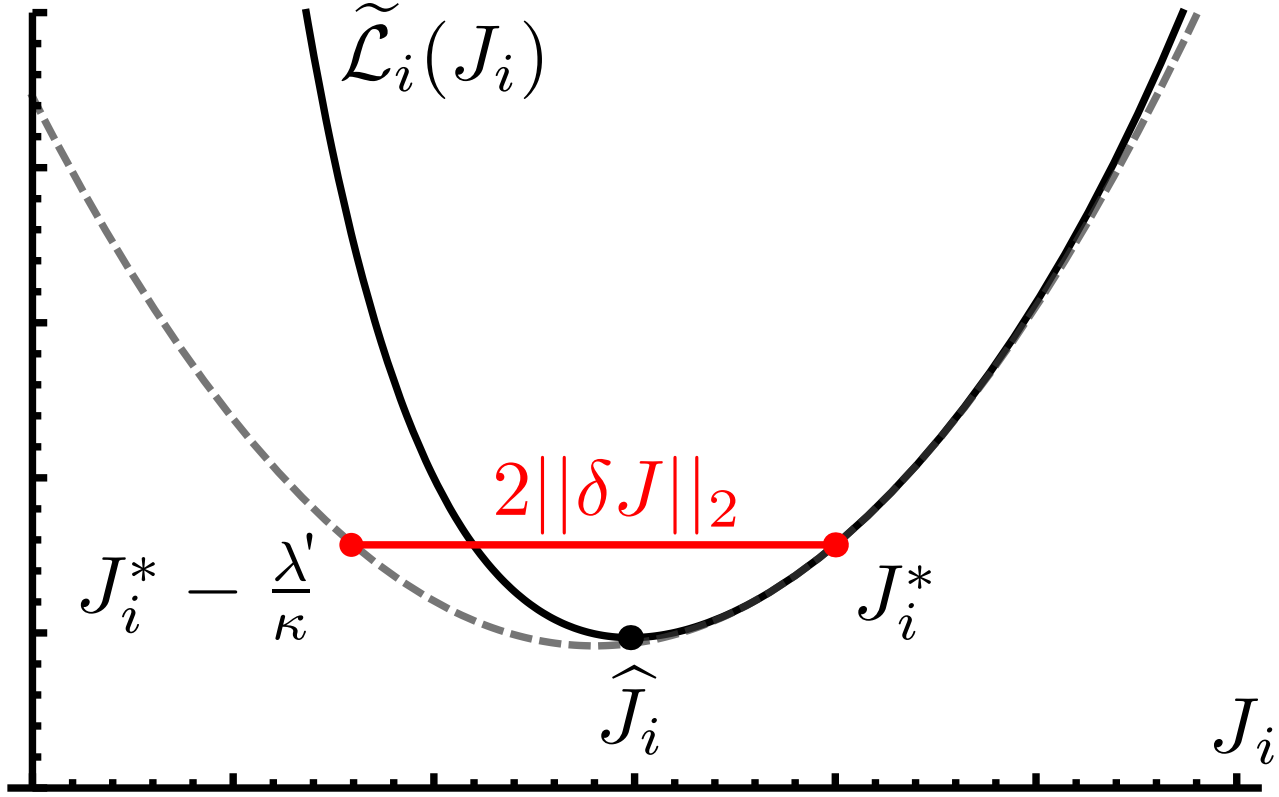








Appendix: Idea of the proof of the Theorem

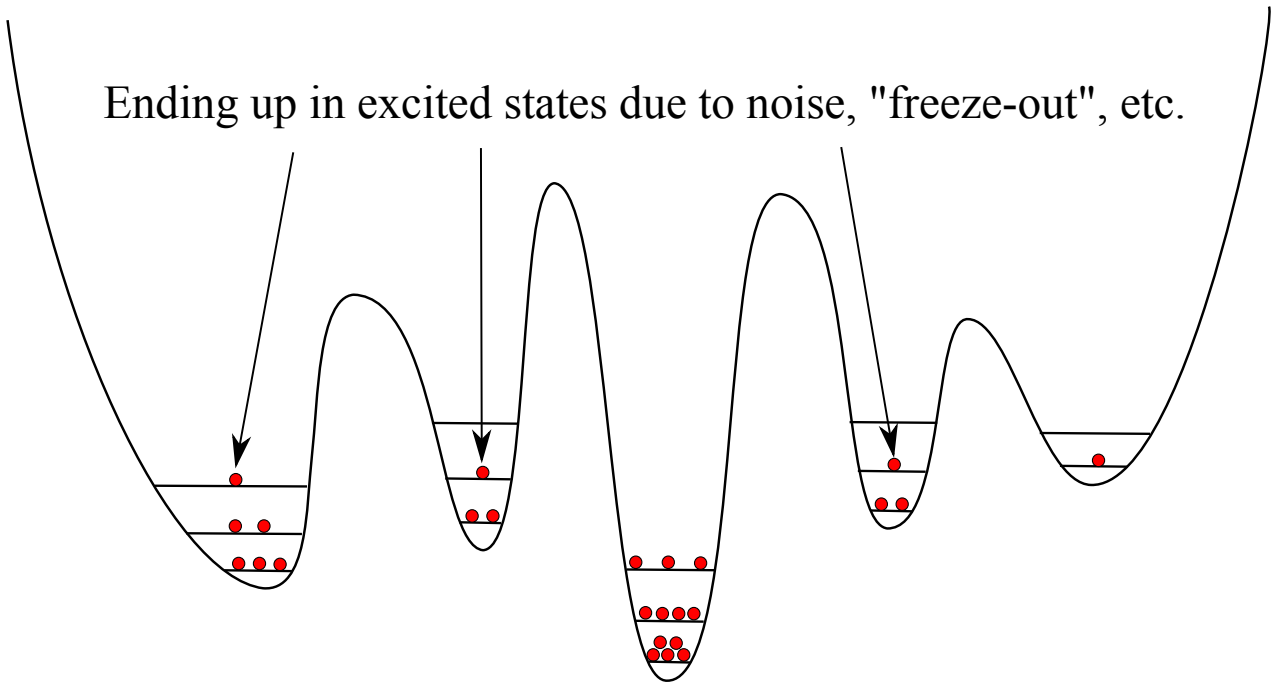


Application to the calibration of the D-Wave quantum computer

Introduction: D-wave as an efficient sampler

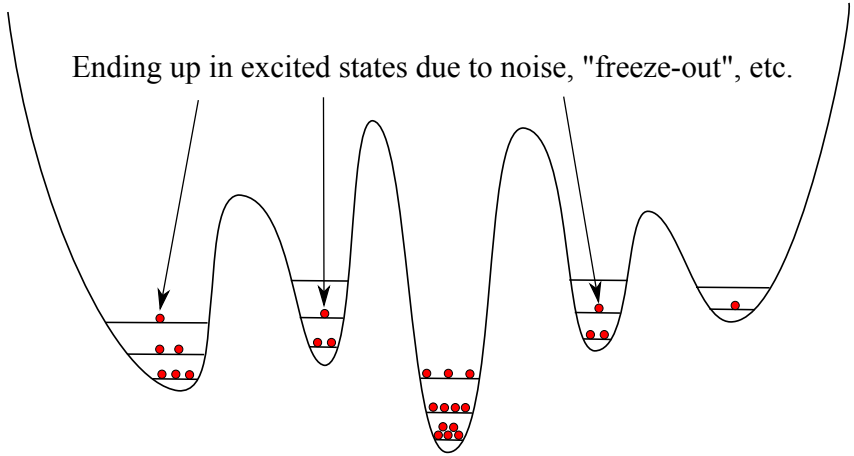
Theoretical and experimental evidence that D-wave can approximately sample from a Boltzmann distribution at some effective temperature

Ronnow *et al.*, Science (2014)
Amin, Phys. Rev. A (2015)
Perdomo-Ortiz *et al.*, Sci. Rep. (2016)
Benedetti *et al.*, Phys. Rev. A (2016)



Introduction: D-wave as an efficient sampler

Disadvantage for optimization turned into advantage for numerous applications:



✓ Restricted Boltzmann Machines (blocks for Deep Learning)

Denil & Freitas, NIPS (2011); Dumoulin *et al.*, AAAI Artificial Intelligence (2015); Benedetti *et al.*, Phys. Rev. A (2016); Amin *et al.*, "Quantum Boltzmann Machine" (2016)

✓ Producting samples in hard glassy models

Katzgraber *et al.*, Phys. Rev. X (2014 & 2015); Martin-Mayor & Hen, Sci. Rep. (2015); Venturelli *et al.*, Phys. Rev. X (2015); Zhu *et al.*, Phys. Rev. A (2016)

✓ Accurate calibration of the D-wave machine

King & McGeoch (2014) "Algorithm engineering for a quantum annealing platform"; Perdomo-Ortiz *et al.*, Sci. Rep. (2016); Raymond *et al.*, "Global warming: temperature estimation in annealers" (2016); Also example in this debrief!

Relation between input and effective Hamiltonians in D-wave

Input Hamiltonian

$$\mathcal{H} = \sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j + \sum_{i \in V} H_i\sigma_i$$

Effective Hamiltonian in D-wave

$$\mathcal{H}_{\text{eff}} = \sum_{\langle i,j \rangle} J'_{ij}\sigma_i\sigma_j + \sum_{i \in V} H'_i\sigma_i$$

Relation between input and effective Hamiltonians in D-wave

Input Hamiltonian

$$\mathcal{H} = \sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j + \sum_{i \in V} H_i\sigma_i$$

Effective Hamiltonian in D-wave

$$\mathcal{H}_{\text{eff}} = \sum_{\langle i,j \rangle} J'_{ij}\sigma_i\sigma_j + \sum_{i \in V} H'_i\sigma_i$$

Let us write $J'_{ij} = \beta(J_{ij} + \Delta J_{ij})$, $H'_i = \beta(H_i + \Delta H_i)$, where

$T = 1/\beta$: effective temperature

ΔJ_{ij}, ΔH_i : possible biases

Correspondence $\mathcal{H} \leftrightarrow \mathcal{H}_{\text{eff}}$ by solving the reconstruction problem of learning β, ΔJ_{ij}, ΔH_i from samples produced by D-wave with $\mathcal{H}_{\text{eff}}$

Effective temperature

Where does the effective temperature come from? Let us look at the annealing procedure with $\tau = t/t_{\text{annealing}}$:

$$\mathcal{H}(\tau) = A(\tau)\left(-\sum_{i\in V}\sigma_i^x\right) + B(\tau)\left(\sum_{\langle i,j\rangle} J_{ij}\sigma_i^z\sigma_j^z + \sum_{i\in V} H_i\sigma_i^z\right)$$

Monotonic functions A and B satisfy $A(0) \gg B(0)$ and $A(1) \ll B(1)$.

Effective temperature

Where does the effective temperature come from? Let us look at the annealing procedure with $\tau = t/t_{\text{annealing}}$:

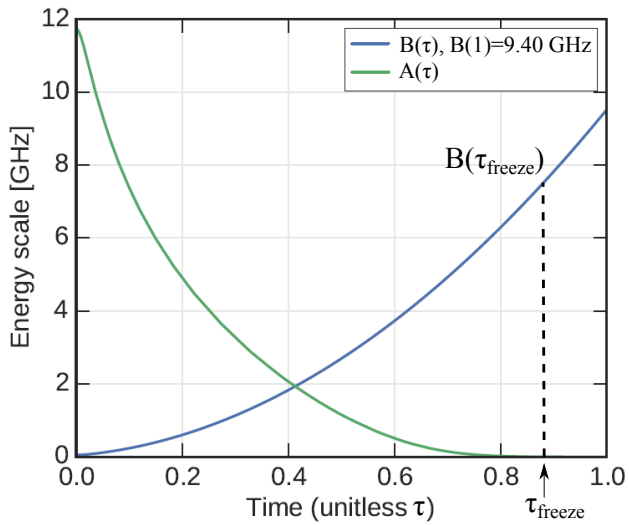
$$\mathcal{H}(\tau) = A(\tau)\left(-\sum_{i\in V}\sigma_i^x\right) + B(\tau)\left(\sum_{\langle i,j\rangle} J_{ij}\sigma_i^z\sigma_j^z + \sum_{i\in V} H_i\sigma_i^z\right)$$

Monotonic functions A and B satisfy $A(0) \gg B(0)$ and $A(1) \ll B(1)$.

The "freeze-out" phenomenon: the evolution stops at the point τ_{freeze}:

$$T_{\text{eff}} = T_{\text{D-wave}}\frac{B(1)}{B(\tau_{\text{freeze}})}$$

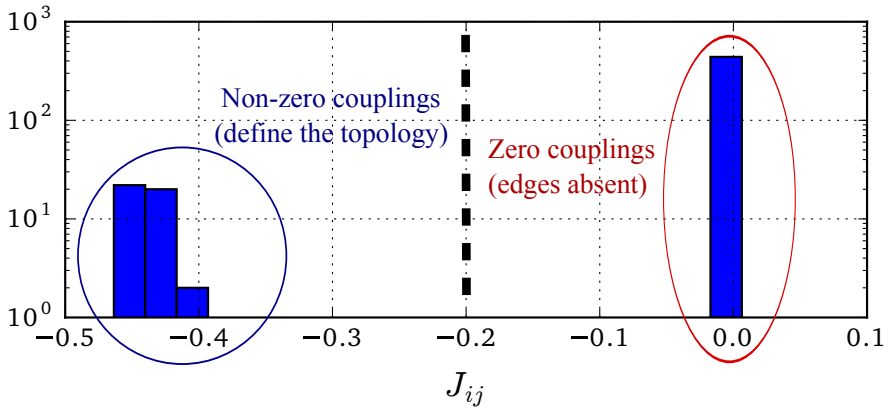
Benedetti *et al.*, Phys. Rev. A (2016)
Raymond *et al.*, "Global warming: temperature estimation in annealers" (2016)

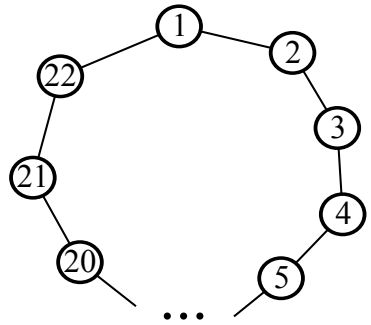


- ✓ **No unique T_{eff}:** β is the function of the input Hamiltonian
- ✓ **"Single qubit freeze-out":** τ_{freeze} can vary for different spins

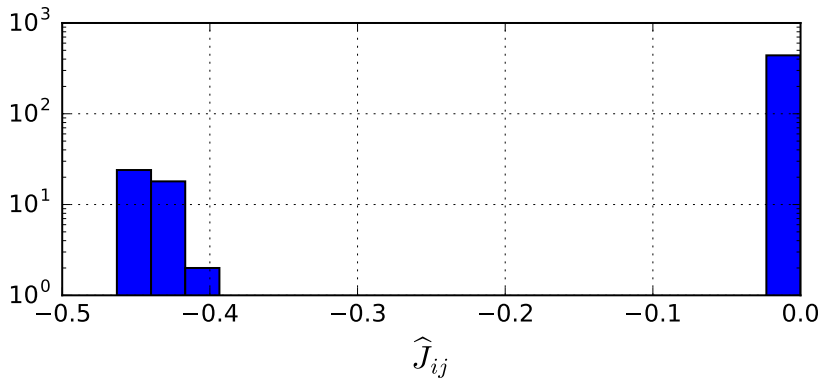
Illustration: estimating the effective temperature

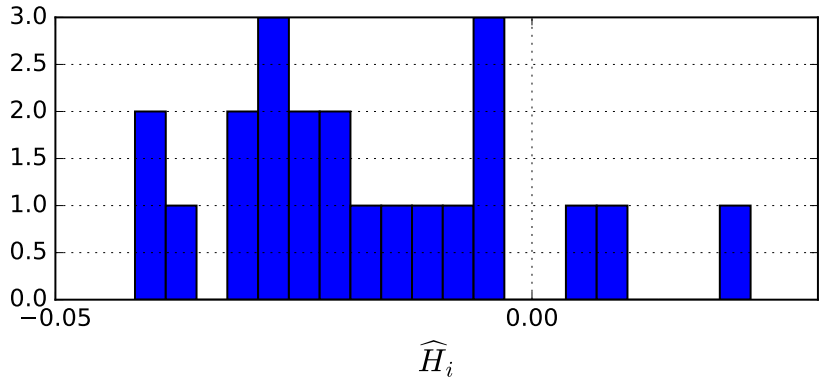
Data set: embedded closed circles of $N = 22$ spins with different values of $J_{i,i+1}$ and $H_i = 0$ (diverse realizations, $t_{\text{annealing}}$, *etc.*). Example for $M = 7250$ and $J_{i,i+1} = -0.0625$ $\forall(i, i+1)$.





Refined $\{J'_{ij}, H'_i\}$. Neglecting H'_i and biases, $\beta_{\text{eff}} \approx 7$ since $\overline{J'} = -0.44$.

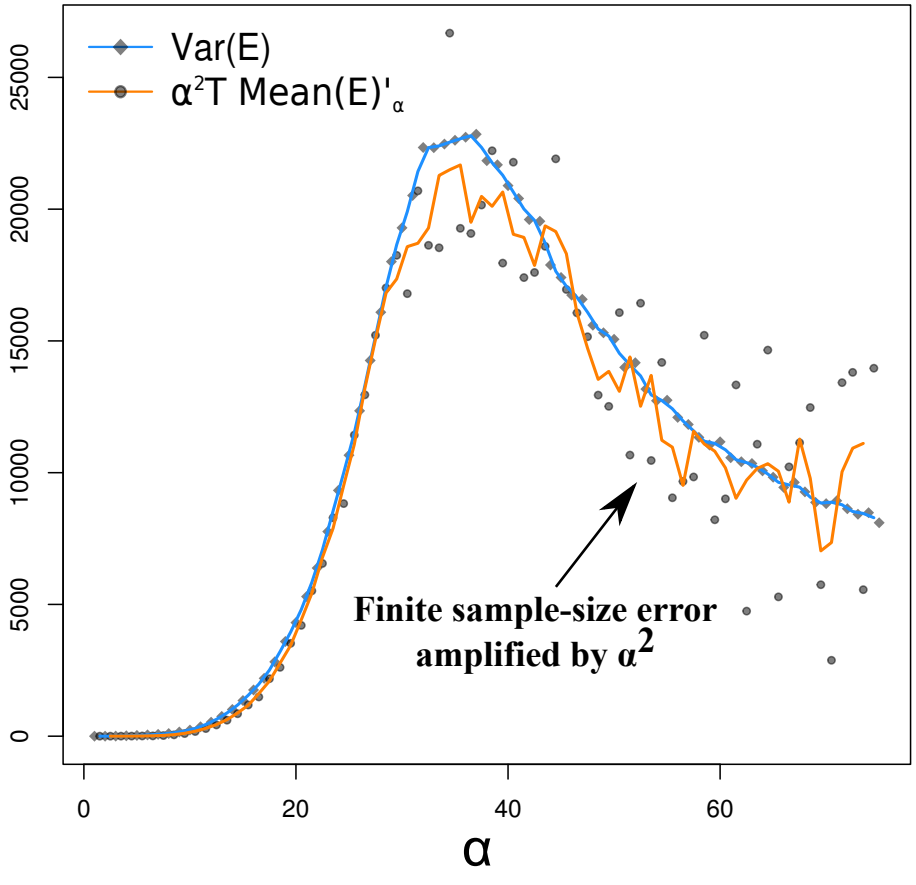




What about biases?

Simple test: if $P(\underline{\sigma}) \propto e^{-\mathcal{H}(\underline{\sigma})/(\alpha T)}$, then $\alpha^2 T \frac{\partial}{\partial \alpha} \langle H \rangle = \langle H^2 \rangle - \langle H \rangle^2$

Checkerboard pattern with magnetic fields



Checkerboard pattern without magnetic fields

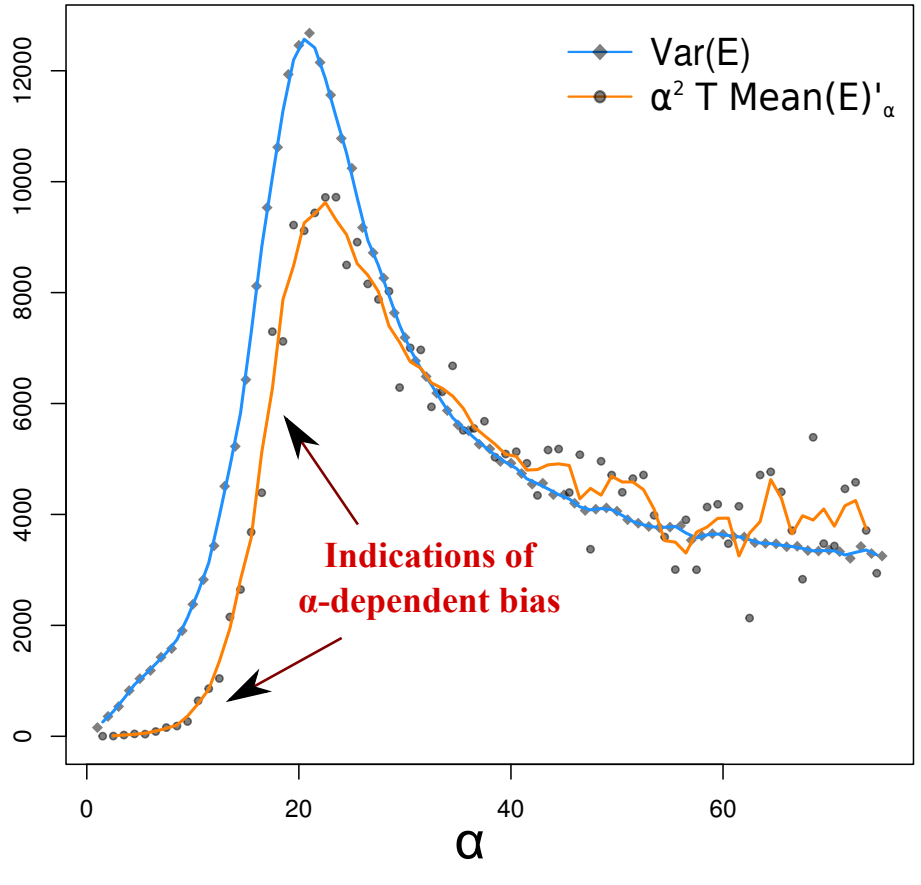
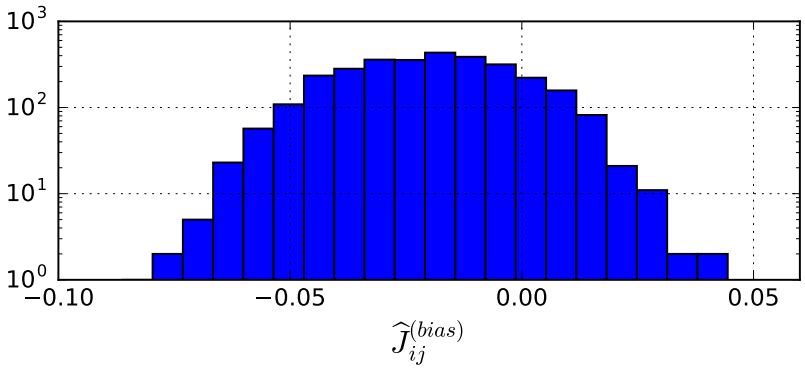
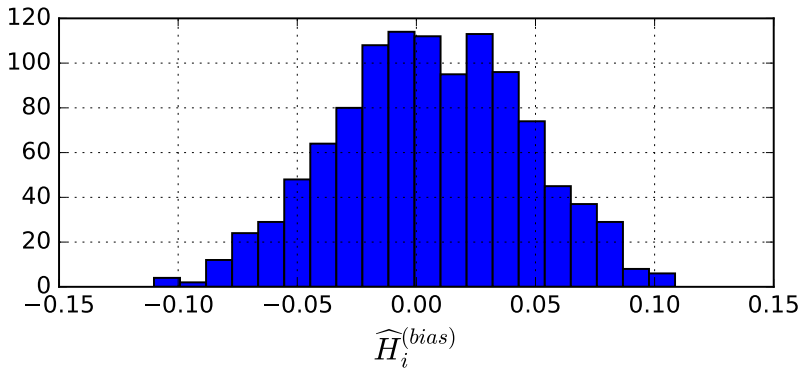


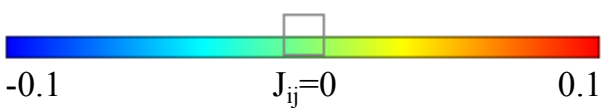
Illustration: detecting and correcting biases

Example of the input $\mathcal{H} = 0$ over the entire Chimera graph

Although D-wave comes with a software for correcting biases, they are still present and persist. Example from the Burnaby machine on Sep 15:







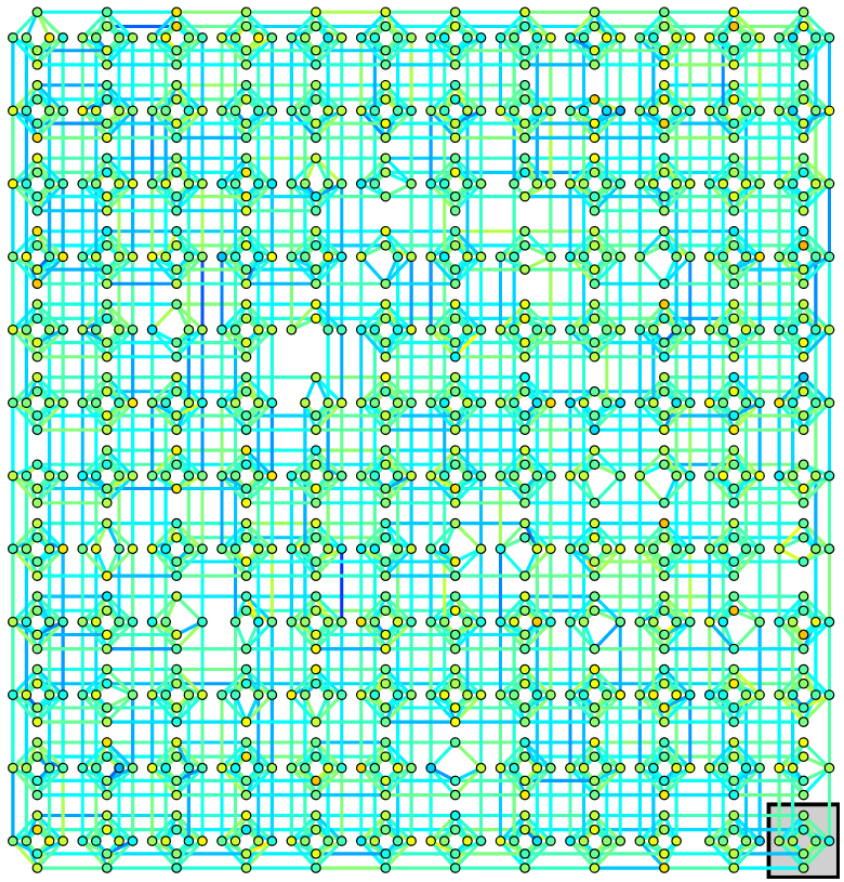
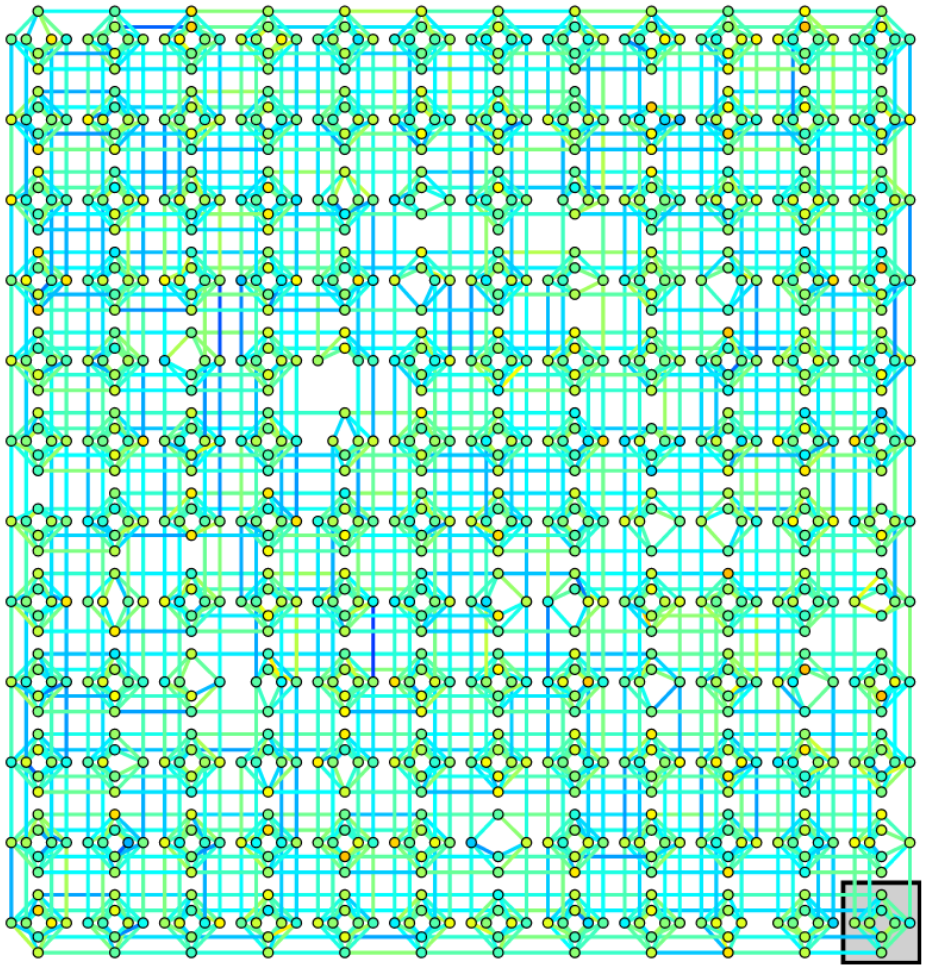
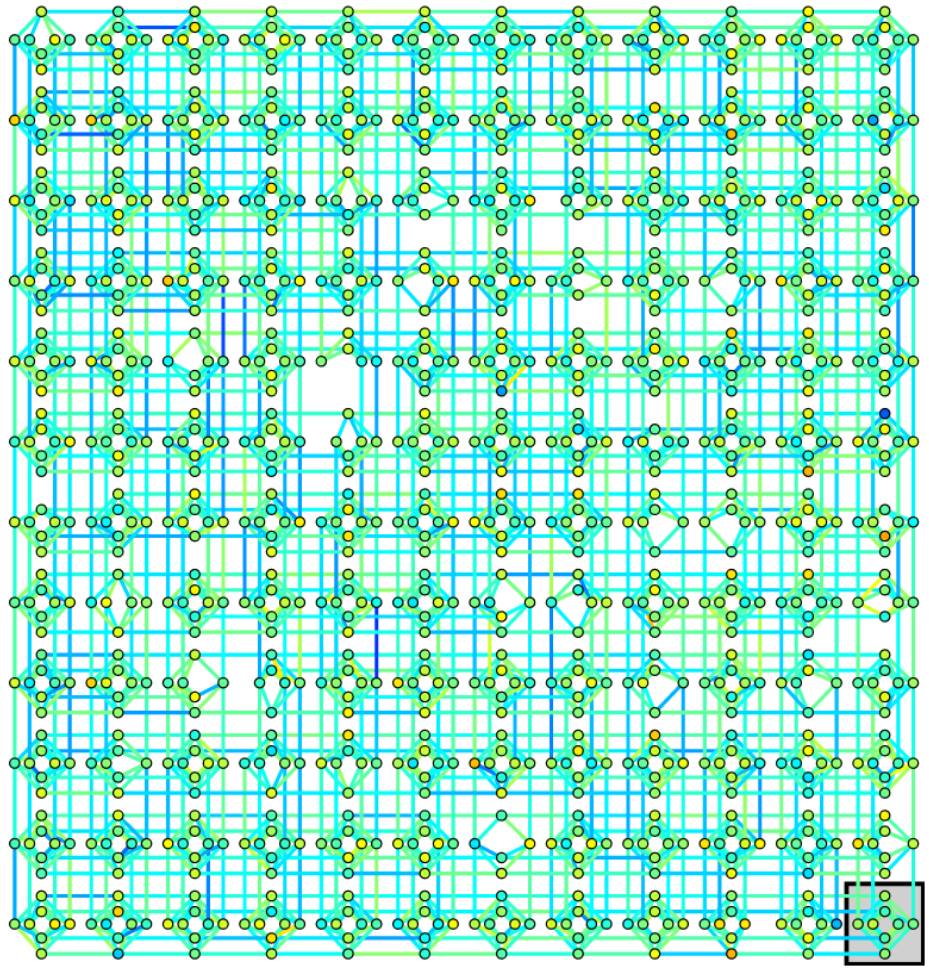


Illustration: detecting and correcting biases

September 15

October 4





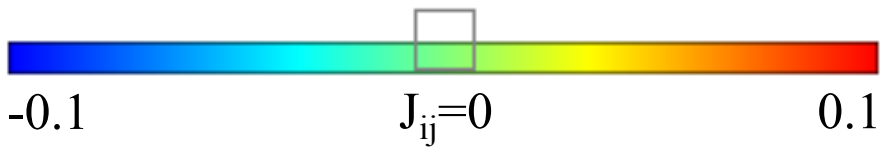
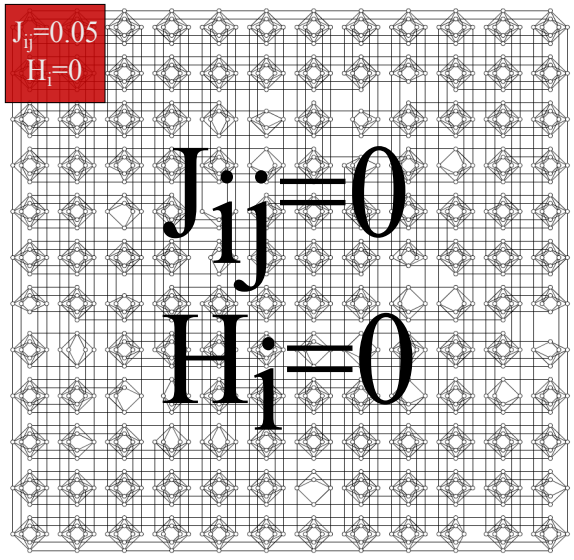
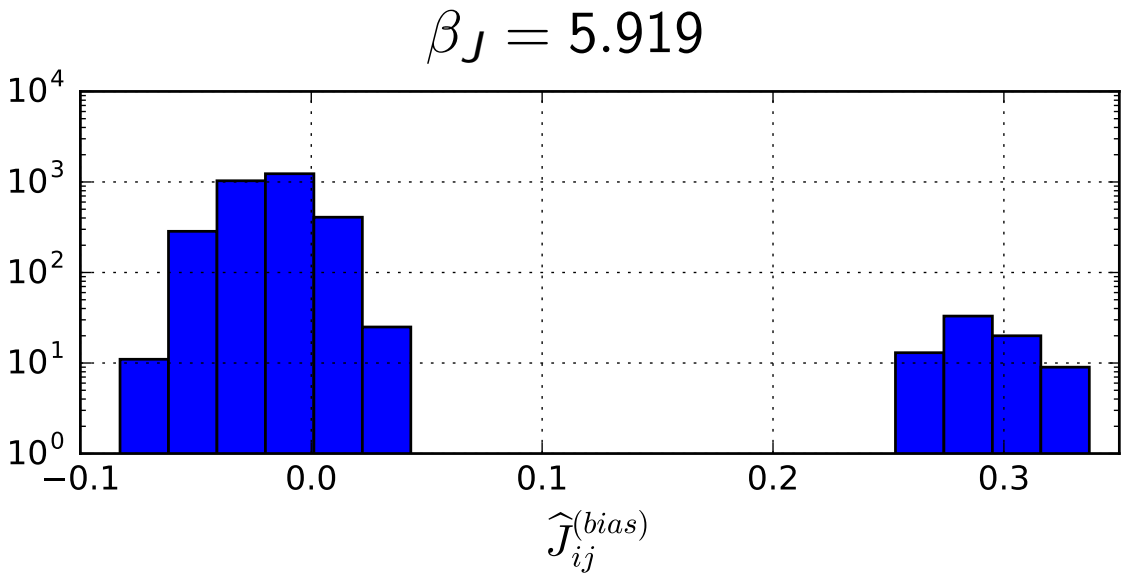
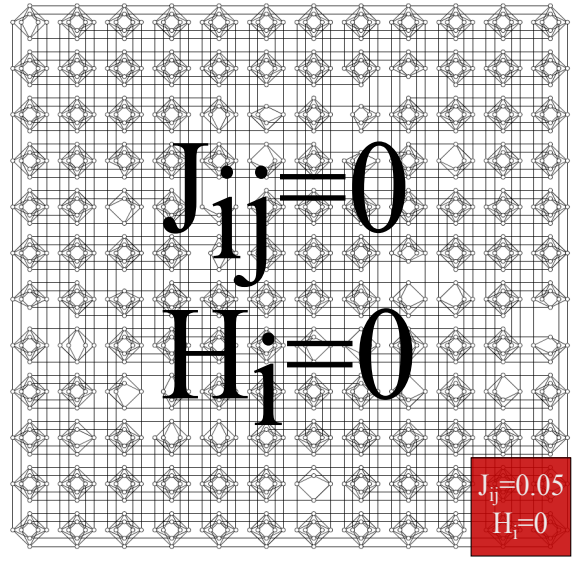


Illustration: detecting and correcting biases







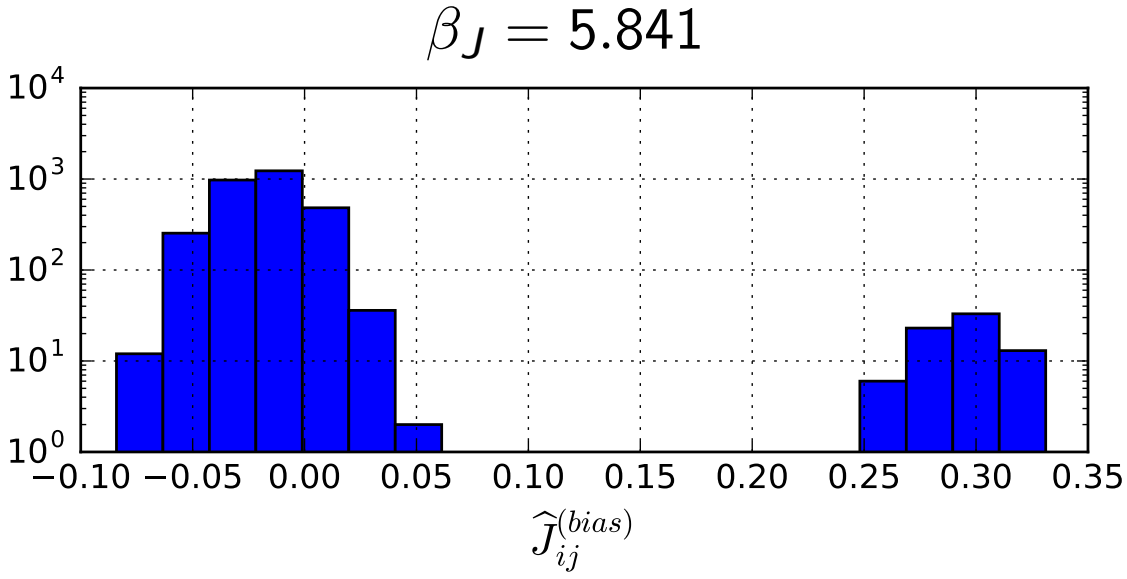
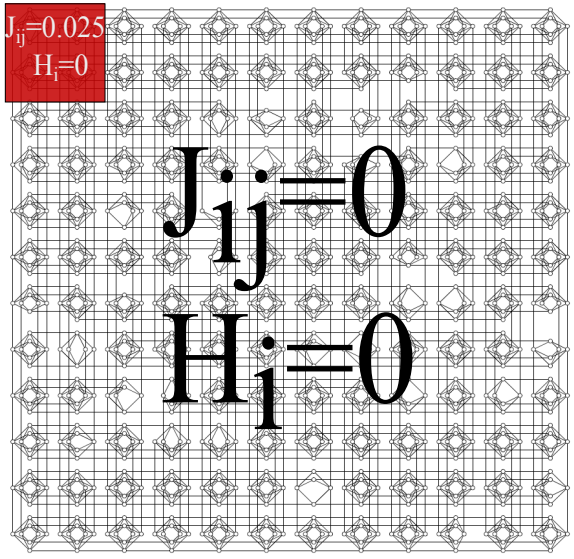
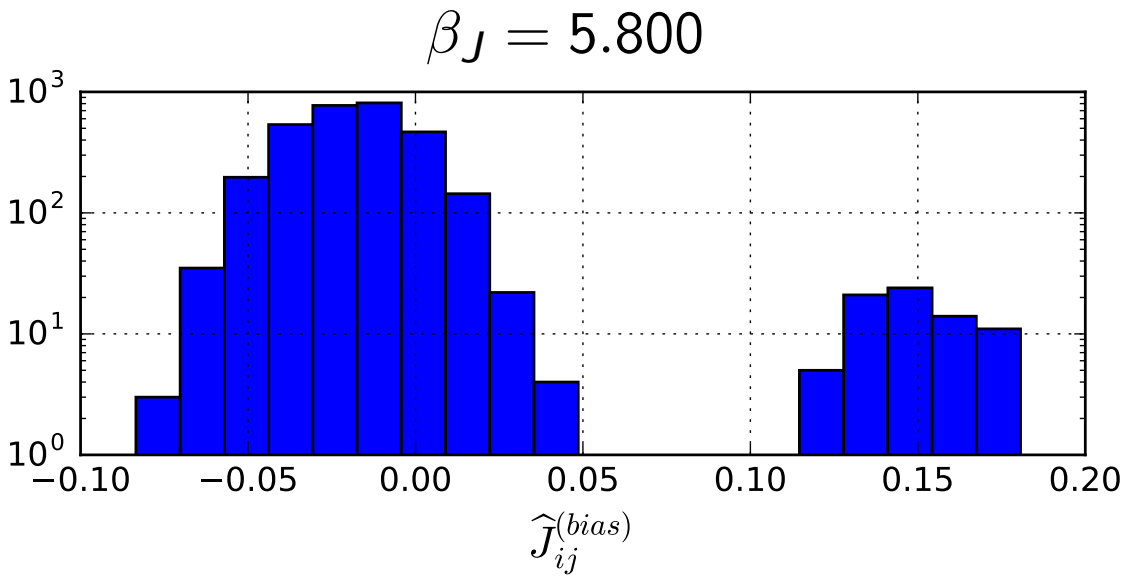
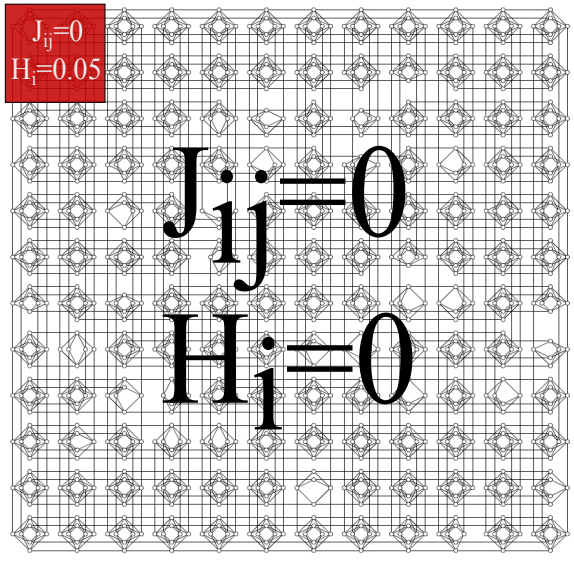


Illustration: detecting and correcting biases







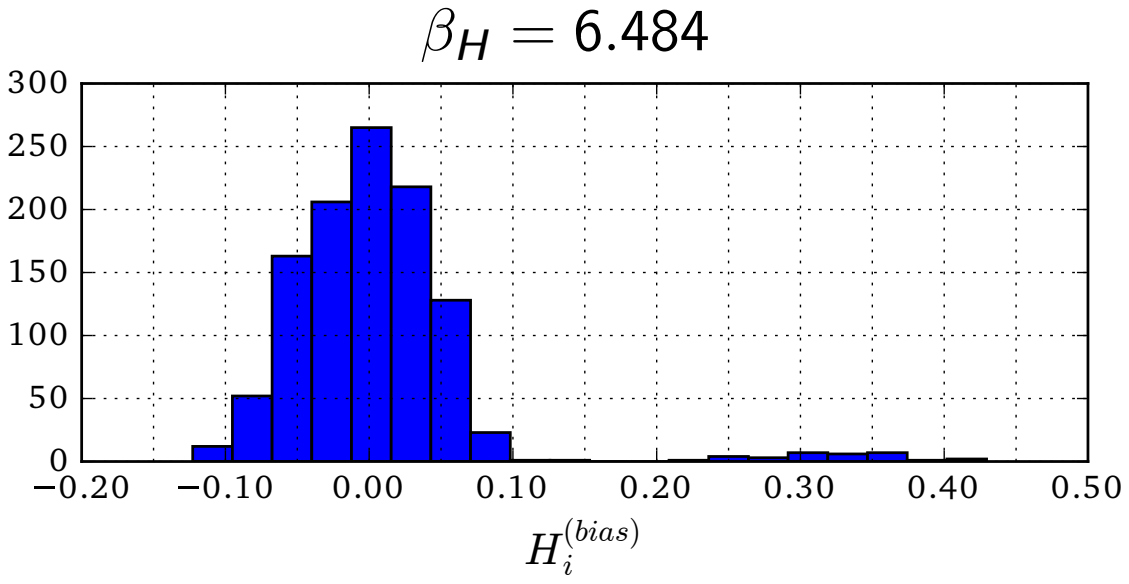
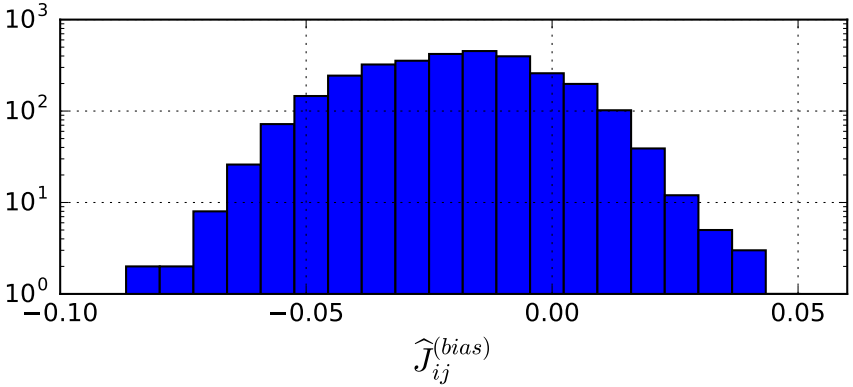
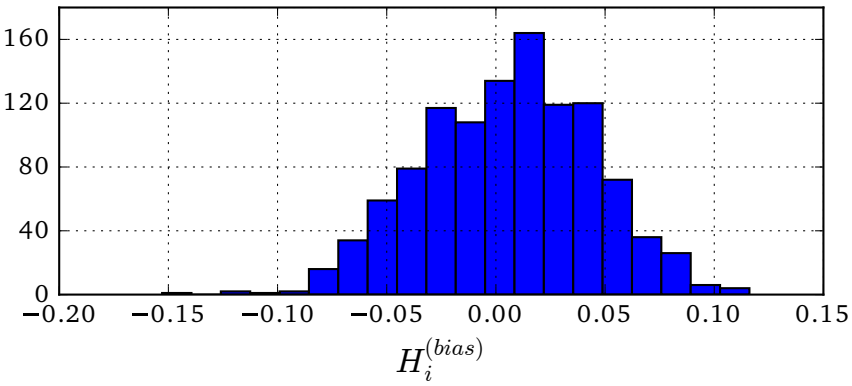


Illustration: detecting and correcting biases

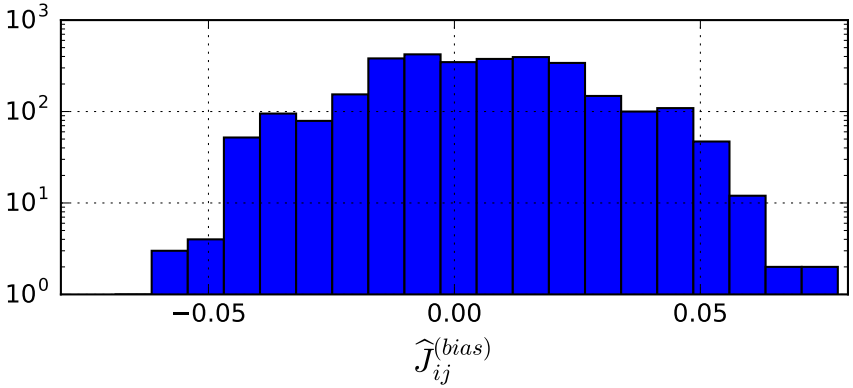
Corrections: inputting $\mathcal{H} = -\frac{1}{\beta_J} \sum_{\langle i,j \rangle} J_{ij}^{(bias)} \sigma_i \sigma_j - \frac{1}{\beta_H} \sum_{i \in V} H_i^{(bias)} \sigma_i$

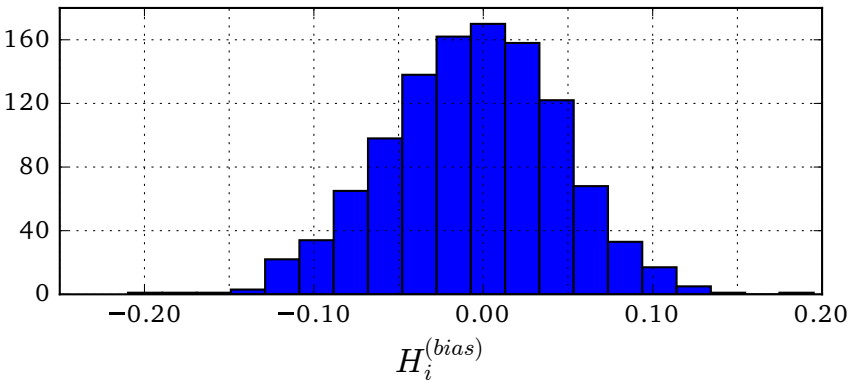
$\mathcal{H}_{\text{eff}}$ before corrections





$\mathcal{H}_{\text{eff}}$ after corrections





Symmetrized and more squeezed distributions with a single iteration

Path forward: efficient calibration of the D-wave machine

The calibration issue addressed in several recent papers with heuristic methods: King & McGeoch (2014); Perdomo-Ortiz *et al.*, Sci. Rep. (2016); ...

As shown in the previous examples, **we can do much better!**

✓ Iteratively correcting the biases for the target $\mathcal{H}_T$:

$$i)\ \frac{\mathcal{H}_T}{\beta} \longrightarrow \mathcal{H}_T + \Delta(\mathcal{H}_T)$$

$$ii)\ \frac{\mathcal{H}_T - \Delta(\mathcal{H}_T)}{\beta} \longrightarrow \mathcal{H}_T - \Delta(\mathcal{H}_T) + \Delta(\mathcal{H}_T - \Delta(\mathcal{H}_T))$$

$$\approx \mathcal{H}_T - \Delta'(\mathcal{H}_T)\Delta(\mathcal{H}_T)$$

$$iii)\ \frac{\mathcal{H}_T - \Delta(\mathcal{H}_T) + \Delta'(\mathcal{H}_T)\Delta(\mathcal{H}_T)}{\beta} \longrightarrow \dots$$

✓ **Machine learning task:** learn the functional form of $\Delta(\mathcal{H}_T)$ with the linear response theory; start directly at the point (ii)

✓ Include the **higher-order interaction** terms in the reconstruction problem to capture the effect of inactive spins