# Inverse Ising Inference Using All the Data

Erik Aurell*
*ACCESS Linnaeus Centre, KTH, Stockholm, Sweden and Department Computational Biology, AlbaNova University Centre, 106 91 Stockholm, Sweden*

Magnus Ekeberg†
*Engineering Physics Program, KTH Royal Institute of Technology, 100 77 Stockholm, Sweden*
(Received 29 September 2011; revised manuscript received 12 December 2011; published 1 March 2012)

We show that a method based on logistic regression, using all the data, solves the inverse Ising problem far better than mean-field calculations relying only on sample pairwise correlation functions, while still computationally feasible for hundreds of nodes. The largest improvement in reconstruction occurs for strong interactions. Using two examples, a diluted Sherrington-Kirkpatrick model and a two-dimensional lattice, we also show that interaction topologies can be recovered from few samples with good accuracy and that the use of $l_1$ regularization is beneficial in this process, pushing inference abilities further into low-temperature regimes.

DOI: 10.1103/PhysRevLett.108.090201 PACS numbers: 02.50.Tt, 05.10.−a, 75.10.Nr

*Introduction.*—When analyzing systems of interacting elements from data, disentangling direct from indirect interactions is an intrinsically complex task. Versions of this problem come about naturally in biology, sociology, neuroscience, and many other fields, and are bound to become more and more important as the amount and diversity of data on large systems continue to grow. In the Ising model, which has served as a basic starting point for studying such situations in applications [1–3], a set of binary variables $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_N\}$, $\sigma_i = \pm 1$, has the distribution

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\beta \sum_i h_i \sigma_i + \beta \sum_{i<j} J_{ij} \sigma_i \sigma_j\right), \qquad (1)$$

where $Z$ is the partition function, $\beta = 1/T$ the inverse temperature, $h_i$ are the external fields, and $J_{ij}$ the pairwise couplings (representing direct interactions). Given magnetizations $m_i = \langle \sigma_i \rangle$ and pairwise correlations $c_{ij} = \langle \sigma_i \sigma_j \rangle - m_i m_j$ the probability distribution which maximizes the entropy has the Ising model form. The standard *inverse Ising problem* means to compute (approximately, efficiently, or according to other criteria) the parameters $h_i$ and $J_{ij}$ from observed $m_i$ and $c_{ij}$. The practical interest in inverse Ising, in the context of the present and future data-rich world, is to use it as an information extraction tool superior to measuring correlations. For example, Ising models can explain the higher order correlations observed in networks of neurons [4] and, extending the number of states from two to 21, spectacular success has been achieved in predicting protein structure by inferring directly interacting residues (amino acids) [5–7]. In this Letter, we address the following two questions: (i) can one do better by keeping all the data for reconstruction and not only empirical pairwise correlation functions, and (ii) can such a method be implemented in a computationally efficient manner? The answer is positive on both accounts, using a method inspired by the regularized logistic regression of Wainwright, Ravikumar, and Lafferty [8]. We show, in particular, that keeping all the data greatly improves reconstruction of an Ising model in the important parameter region of strong interactions.

*Maximum likelihood and computability.*—Given $B$ independent observations $\{\boldsymbol{\sigma}^{(k)}\}_{k=1}^B$ all drawn from (1), the log-likelihood function is

$$l(\{h_i\}, \{J_{ij}\}; \{\boldsymbol{\sigma}^{(k)}\}_{k=1}^B) = \beta \sum_i h_i m_i^{(B)} + \beta \sum_{i<j} J_{ij} (m_i^{(B)} m_j^{(B)} + c_{ij}^{(B)}) - \log Z, \qquad (2)$$

where $m_i^{(B)}$ and $c_{ij}^{(B)}$ are the empirical first and second moments. In (1), the averages of the functions multiplying the model parameters are sufficient statistics [9–11], which in the case at hand means that inference of the biases $h_i$ and the interaction strengths $J_{ij}$ cannot be done better using all the $B$ samples ($NB$ data points), than by observing just $m_i^{(B)}$ and $c_{ij}^{(B)}$ ($\frac{N(N+1)}{2}$ data points). The optimal estimates (in a maximum likelihood sense) are given by $\partial_{h_i} \log Z = \beta m_i^{(B)}$ and $\partial_{J_{ij}} \log Z = \beta [m_i^{(B)} m_j^{(B)} + c_{ij}^{(B)}]$. Boltzmann learning [12] uses Monte Carlo (MC) sampling to compute these gradients of the partition function with, in principle, unbounded accuracy, but is computationally tractable only for very small systems (although faster versions of this procedure have been introduced, see, e.g., [13] or [14]). A whole series of approximations, reviewed in [15], have therefore been developed by expanding in high temperature (small interactions), large external fields, or other parameters cf. (naive) mean field (nMF) [16], Thouless-Anderson-Palmer inversion [16], small-correlation expansion (SCE) [17] and

 

have been further extended using the fluctuation-dissipation theorem [18–20]. It is well-established that all these approximate methods are not accurate when the number of samples is small, nor when the interactions are strong (temperature is low). However, a recent method based on expansion of the system into "clusters" (the contributions of which to the estimates of $\{\mathbf{h}, \mathbf{J}\}$ are included or discarded depending on their entropy share) manages to select correctly the parameters from few samples in various low-temperature settings [21], questioning these limitations. Another promising candidate called minimum probability flow, recently introduced in [22], has been shown to very efficiently recover Ising parameters for a two-dimensional grid. Its performance on more general systems, in particular, strongly correlated ones, is an interesting and open question.

*Pseudolikelihood maximization (without regularization).*—The conditional probability of one variable $\sigma_r$ given all the others $\boldsymbol{\sigma}_{\setminus r} = (\sigma_1, \ldots, \sigma_{r-1}, \sigma_{r+1}, \ldots, \sigma_N)$ is

$$P_{\{\mathbf{h},\mathbf{J}\}}(\sigma_r|\boldsymbol{\sigma}_{\setminus r}) = \frac{1}{1 + e^{-2\beta\sigma_r[h_r + \sum_{i\neq r} J_{ir}\sigma_i]}}, \tag{3}$$

where we take $J_{ir}$ to mean $J_{ri}$ when $i > r$. If $\sigma_r$ by itself is considered a dependent variable, and the complementary set $\boldsymbol{\sigma}_{\setminus r}$ is taken as independent variables, then the maximum likelihood estimates of the parameters $h_r$ and $\mathbf{J}_r = \{J_{ir}\}_{i\neq r}$, given $B$ samples, minimize

$$f_r(h'_r, \mathbf{J}'_r) = -\frac{1}{B}\sum_{k=1}^{B} \ln P_{\{\mathbf{h}',\mathbf{J}'\}}(\sigma_r^{(k)}|\boldsymbol{\sigma}_{\setminus r}^{(k)}). \tag{4}$$

Minimizing these functions $f_r$ for all $r$ simultaneously, which we call pseudolikelihood maximization (PLM), is not the same as maximizing the total log-likelihood (2). For example, it typically gives different estimates $J_{ij}^{*,i}$ and $J_{ij}^{*,j}$ depending on if $\sigma_i$ or $\sigma_j$ is considered the dependent variable. We will for definiteness sake always take $J_{ij}^{*} = \frac{1}{2}(J_{ij}^{*,i} + J_{ij}^{*,j})$. Alternatively, one could minimize the sum of all $f_r$ while requiring $J_{ij}^{*,i} = J_{ij}^{*,j}$. When the number of samples is large, we can substitute sample average with ensemble average, and write

$$\begin{aligned} f_r(h'_r, \mathbf{J}'_r) &\approx \langle -\ln[P_{\{\mathbf{h}',\mathbf{J}'\}}(\sigma_r|\boldsymbol{\sigma}_{\setminus r})]\rangle \\ &= \sum_{\boldsymbol{\sigma}} \ln(1 + e^{-2\beta\sigma_r[h'_r + \sum_{i\neq r} J'_{ir}\sigma_i]}) P_{\{\mathbf{h},\mathbf{J}\}}(\boldsymbol{\sigma}), \end{aligned} \tag{5}$$

with equality expected in the limit. Necessary maximum likelihood conditions (for one of the conditional probabilities) are then

$$\frac{\partial f_r}{\partial J'_{sr}}(h'_r, \mathbf{J}'_r) = \sum_{\boldsymbol{\sigma}} \frac{-2\beta\sigma_s\sigma_r}{e^{2\beta\sigma_r[h'_r + \sum_{i\neq r} J'_{ir}\sigma_i]} + 1} P_{\{\mathbf{h},\mathbf{J}\}}(\boldsymbol{\sigma}) = 0, \tag{6}$$

and similarly for the variation with respect to an external field. At the true parameters these equations hold, since

$$\begin{aligned} \frac{\partial f_r}{\partial J'_{sr}}(h_r, \mathbf{J}_r) &= \frac{-\beta}{Z\{\mathbf{h}, \mathbf{J}\}} \sum_{\boldsymbol{\sigma}} \sigma_s\sigma_r \frac{e^{\beta\sum_{i\neq r} h_i\sigma_i + \beta\sum_{\substack{i<j \\ i,j\neq r}} J_{ij}\sigma_i\sigma_j}}{\cosh(\beta\sigma_r[h_r + \sum_{i\neq r} J_{ir}\sigma_i])} \\ &= 0, \end{aligned} \tag{7}$$

where the expressions vanish because each state for which $\sigma_r = 1$ has exactly one opposing state for which $\sigma_r = -1$, contributing equally in size. Assuming this stationary point is a minimum we can locate, the pseudolikelihood approach to inferring an Ising model is exact in the limit of large sample size, and is in this sense qualitatively different from other approximate inverse Ising schemes.

*Pseudolikelihood maximization with $l_1$-regularization.*—Ravikumar, Wainwright, and Lafferty in [8] introduced a $l_1$-regularized version of the pseudolikelihood approach, i.e., one where the functions to be minimized are $[f_r(h'_r, \mathbf{J}'_r) + \lambda||\mathbf{J}'_r||_1]$ with some penalty parameter $\lambda > 0$. $l_1$ (absolute value) regularization is widely used to recover sparse signals [23–25], in situations where a large fraction of parameters is known to be zero, but not which parameters. The numerical minimization can be done efficiently using convex programming, such as the interior point method of Koh, Kim, and Boyd [26], which we have used below.

*Results for high-quality data.*—We minimized (4) using Newton decent for several values of $B$ in the setting of the dilute Sherrington-Kirkpatrick (SK) model [27], a commonly used test bench for comparing performances of inverse Ising solvers. Every $J_{ij}$ is thus nonzero with probability $p$, and if so drawn from a Gaussian distribution with zero mean and variance $1/c$, $c = pN$. External fields were first taken as zero. Reconstruction error was measured by $\Delta = \frac{1}{1/\sqrt{N}}\langle(J_{ij}^{*} - J_{ij})^2\rangle^{1/2}$. For comparison we also applied nMF, Thouless-Anderson-Palmer inversion, and two versions of the small-correlation expansion: the general result of [17], as well as their higher order zero-magnetization version. The latter is currently the best performing approximate method tested on the zero-field SK model. Figures 1(a) and 1(b) show simulation results for $N = 64$ compared to nMF, i.e., $J_{ij}^{\mathrm{nMF}} = -\frac{1}{\beta}(\mathbf{c}^{-1})_{ij}$. We also include one curve (for $10^8$ samples) for the best performing competing method, which indeed turned out to be the higher order SCE. The curves are the averages of five different parameter sets, yielding error bars small enough to be omitted. MC sampling was performed using a warm-up time of $10^7$ $N$ spin updates and a sampling frequency of one observation every 10 $N$ updates. Evidently, PLM outperforms nMF and SCE in the low-temperature region. As $T$ approaches one from above (towards the spin glass phase), nMF and SCE start to perform poorly, while our logistic regression algorithm appears unaffected. Lowering the temperature further to $T = 0.5$, where indeed all approximate methods tested on this example to date are unusable, PLM continues to

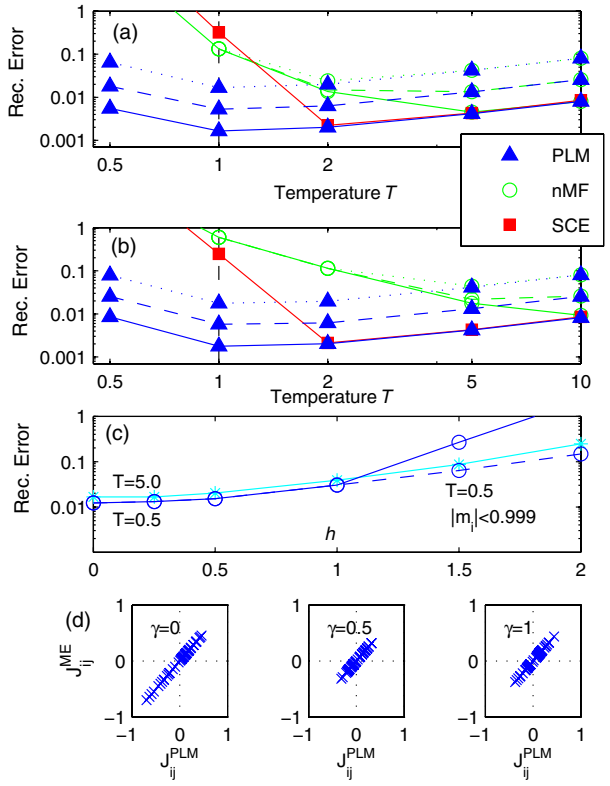


FIG. 1 (color online). (a) and (b) show reconstruction errors of PLM, nMF, and SCE versus temperature for (a) fully ($p = 1$) and (b) sparsely ($p = 0.1$) connected SK systems of size $N = 64$. The number of MC samples used are $10^6$ (dotted lines), $10^7$ (dashed lines) and $10^8$ (continuous lines). (c) Reconstruction errors of PLM as functions of external field strength for a SK system of size $N = 10$ using $10^6$ samples for two different temperatures. The dashed curve is obtained for $T = 0.5$ by excluding parameter sets where one or more empirical $|m_i| > 0.999$. (d) Comparison of parameter estimates between Boltzmann learning ($J_{ij}^{\mathrm{ME}}$) and PLM ($J_{ij}^{\mathrm{PLM}}$) using data generated from a distribution with Hamiltonian $-(1-\gamma)\sum_{i<j} J_{ij}\sigma_i\sigma_j - \gamma\sum_{i<j<k} J_{ijk}\sigma_i\sigma_j\sigma_k$. The system parameters used are $N = 10$, $B = 10^6$, $T = 2$, and all interaction parameters are drawn from a $N(0, \frac{1}{\sqrt{N}})$ distribution.

function. This makes it the first tested method to reconstruct successfully in this strongly correlated region of the SK model. As the temperature increases, performance is limited by the finiteness of $B$ rather than by the method choice. In this limit [right part of the curves in Figs. 1(a) and 1(b)] reconstruction error $\Delta$ follows $\sim \frac{1}{\sqrt{B}}$ for all methods, but for PLM this seems to hold for all temperatures [parallel curves in Figs. 1(a) and 1(b)]. The switch to sparse $\mathbf{J}$ clearly worsens the performance of nMF, but does not seem to affect the pseudolikelihood scheme or SCE much, if at all. The results for system sizes $N = 16$ and $N = 128$ are similar (data not shown).

In applications, external fields typically are not zero, and there will usually also be some interdependence between samples. Thus, it is natural to ask whether the good results for $T < 1$ are maintained when relaxing these assumptions. To assess the robustness of PLM against biases, we let all $\beta h_i = h$ and observed the reconstruction error as $h$ was increased. Figure 1(c) shows results for a SK model with $N = 10$ using $10^6$ samples in the cases $T = 5$ and $T = 0.5$. For both temperatures, a moderate change in the fields has no effect on performance. For weak interactions ($T = 5$), strong magnetizations yield only a modest increase in error (for the $h = 2$ case all means are $m_i \approx 0.96$). When both couplings and fields are strong, the situation is more delicate. An examination of the data shows that the decrease in performance for $T = 0.5$ and $h > 1$ stems from a few inference runs having huge errors. For instance, at $h = 1.5$, over 90% of the runs still show errors $\Delta < 0.1$, but a few have $\Delta$ as high as 10 or higher. An explicit check of these instances shows that the cause appears to be freezing spins ($|m_i| \to 1$). When removing the few runs having any empirical magnetizations $|m_i| > 0.999$, we saw the same modest increase in error as for weak interactions [dashed curve in Fig. 1(c)]. Interestingly, even for runs where the error explodes, most parameters are still correctly identified; i.e., a few parameters diverging does not appear to destroy the reconstruction throughout the rest of the system.

In the default setup, we sampled the configurations at time steps spaced $10\,N$ MC steps apart. To assay robustness against such correlations, we have also used a spacing $k\,N$ and lowered $k$ successively in the strongly correlated case $N = 64$, $T = 0.5$, $10^6$ samples from Fig. 1(a) (data not shown). We then saw practically no effect lowering $k$ from 10 to 1, while we did see effects at $k \approx 0.5$ and smaller.

The consistency result (7) assumes an Ising model as the true underlying distribution, a premise which is, at best, only approximately true for any real data set. An interesting question is therefore: will our method deviate markedly from an exact maximum-entropy inference, as found by Boltzmann learning, when the Ising assumption is not true? The answer seems to be no. To investigate, we imposed an Ising model on a small system (thus feasible for Boltzmann learning) which had both second and third order interactions. Figure 1(d) shows that, even when the true system has no pairwise interactions at all, PLM and Boltzmann learning estimate very similar "would be" $J_{ij}$:s. In conclusion, the algorithm appears to be tolerant towards approximate distribution assumptions, high biases, as well as data interdependency.

*Results for low-quality data.*—Rebuilding the sign-sparsity pattern of $\mathbf{J}$ from few samples using the PLM idea has been done numerically for various sparsity types in [8,28]. We provide here some additional results, specifically regarding the advantages of using a regularization term. Taking $\lambda > 0$ after all makes the optimization problem considerably harder computationally. A simpler approach would be to minimize (4) with $\lambda = 0$ and declare all couplings for which $|J_{ij}| < \delta$ to be zero (for some

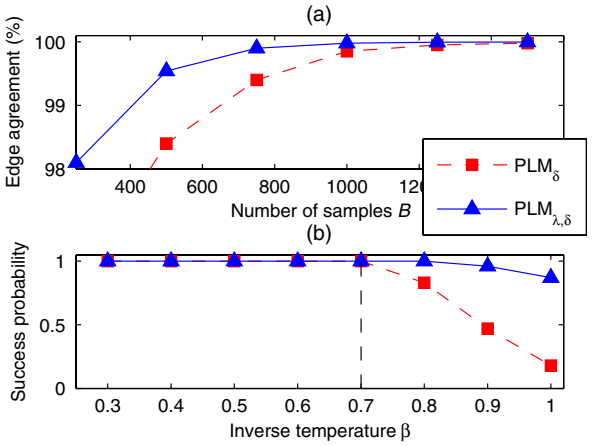


FIG. 2 (color online). (a) Edge agreement versus sample size in a binary SK model of size $N = 100$ and sparsity $p = 0.05$ for $\mathrm{PLM}_\delta$ and $\mathrm{PLM}_{\delta,\lambda}$. $T = 2$ for all data points. (b) Probability of 100% edge agreement versus inverse temperature for $\mathrm{PLM}_\delta$ and $\mathrm{PLM}_{\delta,\lambda}$ using $B = 4500$ on $7 \times 7$ nearest-neighbor grids ($N = 49$) with 30% dilution.

tolerance $\delta$). Intuitively, inclusion of a regularization term should allow for better utilization of sample information than the simpler tolerance approach. As a test case we looked at a version of the SK model where the couplings were not Gaussian but binary, $J_{ij} = \pm \frac{1}{\sqrt{pN}}$ (with equal probability). The inference quality was measured as the percentage of pairs $(i, j)$ where the interaction strength was identified correctly as "+," "0," or "−." PLM using tolerance only and PLM using regularization (as well as a tolerance limit) will be referred to as $\mathrm{PLM}_\delta$ and $\mathrm{PLM}_{\delta,\lambda}$ respectively. Figure 2(a) shows that for $N = 100$, $p = 0.05$, and $T = 2$, $\mathrm{PLM}_{\delta,\lambda}$ fits the edges more accurately and gives perfect reconstruction for fewer samples than $\mathrm{PLM}_\delta$. Note that in this example guessing $J^*_{ij} = 0$ for all pairs would result in a 95% edge agreement on average. Optimal values of $\delta$ and $\{\delta, \lambda\}$ for each $B$ were determined empirically and used on 20 new parameter sets to yield the averages.

For several sparsity structures the performance of PLM has been shown to drop as the temperature goes below some $T_{\mathrm{crit}}$ even if $B$ is quite large [28]. One such example is $B = 4500$ on $7 \times 7$ nearest-neighbor grids with positive couplings, where each edge in the grid is removed with probability 0.3 and the remaining couplings are set to one. The "failure" occurs close to the known critical point for the Ising model on such grids [28], $\beta_{\mathrm{crit}} \approx 0.7$ [29]. We applied $\mathrm{PLM}_{\delta,\lambda}$ and $\mathrm{PLM}_\delta$ to this problem to see whether combined regularization and tolerance can boost performance at low temperatures. Figure 2(b) shows the outcome, where optimal $\delta$ and $\{\delta, \lambda\}$ for each $\beta$ were again found empirically and probabilities estimated using 200 new grids. A breakdown is indeed seen for $\mathrm{PLM}_\delta$ around $\beta = 0.7$, but the effect on $\mathrm{PLM}_{\delta,\lambda}$ is much less pronounced. Perfect edge recovery, using the latter, is experienced with high probability far into the low-temperature region. The complete data output (not reported) shows that including the tolerance threshold in $\mathrm{PLM}_{\delta,\lambda}$ (as opposed to trusting the regularization term alone to force suitable estimates of $J_{ij}$ to zero), becomes necessary at low temperatures. MC samples in this case were generated using a warm-up time of $10^7$ $N$ spin updates and a sampling frequency of one observation every 2000 $N$ updates.

*Discussion.*—Pseudolikelihood and approximate maximum entropy should be considered alternative approaches, where in both cases exact inference (likelihood or maximum entropy) has been traded in for computability. Our results suggest that the pseudolikelihood approach allows for accurate inference in Ising models even for large strongly coupled systems, a capability which appears to be maintained even when the amount of data is severely limited. Tolerating high external fields, dependence among samples, sparseness as well as strong correlations, in addition to being robust when distribution assumptions are but approximate, the method provides a very complete and real alternative to current approaches that typically fail in one or several of these respects. Our results also confirm that including an $l_1$-regularization term is helpful in retrieving sign sparsity from few samples, allowing for complete graph reconstruction even in low-temperature regions.

The PLM objective function has one term per sampled configuration, so running time is heavily dependent on sample size. For instance, the $N = 64$ cases with $10^8$ samples took hours on a (one-core) standard home desktop computer using Newton decent, with almost all of the time spent evaluating the Hessian of the objective function (which depends on all $10^8$ samples). We note for future work that one may alternatively use a quasi-Newton or a conjugate gradient method, i.e., algorithms that do not use (exact) Hessians (initial trials suggest that these work also in the strongly correlated cases). When the number of samples is small, however, the algorithm naturally runs fast. Thus, in the region where PLM is likely most interesting (small sample size), it is also computationally efficient and competitive. Moreover, a practical convenience of PLM is that the $N$ subproblems can be solved completely independently, allowing for straightforward parallel execution.

E. A. thanks Martin Wainwright, Toshiyuki Tanaka, and Michael Hörnqvist for useful discussions. This work was supported by the Academy of Finland as part of its Finland Distinguished Professor program, Project No. 129024/ Aurell.

*Also at Aalto University School of Science, Helsinki, Finland.
eaurell@kth.se
†ekeb@kth.se

[1] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature (London) **440**, 1007 (2006).

[2] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, Proc. Natl. Acad. Sci. U.S.A. **103**, 19033 (2006).
[3] S. Cocco, S. Leibler, and R. Monasson, Proc. Natl. Acad. Sci. U.S.A. **106**, 14058 (2009).
[4] G. Tkacik, E. Schneidman, M. J. B. II, and W. Bialek, arXiv:0912.5409.
[5] A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, and H. Szurmant, Proc. Natl. Acad. Sci. U.S.A. **106**, 22124 (2009).
[6] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67 (2008).
[7] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U.S.A. **108** E1293 (2011).
[8] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, Ann. Stat. **38**, 1287 (2010).
[9] G. Darmois, C. R. Acad. Sci. Paris Ser. IV **200**, 1265 (1935).
[10] E. Pitman, Math. Proc. Cambridge Philos. Soc. **32**, 567 (1936).
[11] B. Koopman, Trans. Am. Math. Soc. **39**, 399 (1936).
[12] H. Ackley, E. Hinton, and J. Sejnowski, Cogn. Sci. **9**, 147 (1985).
[13] T. Broderick, M. Dudik, G. Tkacik, R. E. Schapire, and W. Bialek, arXiv:0712.2437.
[14] M. A. Carreira-Perpinan and G. E. Hinton, in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, edited by R. G. Cowell and Z. Ghahramani (Society for Artificial Intelligence and Statistics, Barbados, 2005), p. 33, http://www.cs.toronto.edu/~hinton/absps/cdmiguel.pdf.
[15] Y. Roudi, J. A. Hertz, and E. Aurell, Front. Comput. Neurosci. **3** 022 (2009).
[16] H. Kappen and F. B. Rodriguez, Neural Comput. **10**, 1137 (1998).
[17] V. Sessak and R. Monasson, J. Phys. A **42** 055001 (2009).
[18] M. Mézard and T. Mora, J. Phys. II (France) **103**, 107 (2009).
[19] E. Marinari and V. V. Kerrebroeck, J. Stat. Mech. (2010) P02008.
[20] E. Aurell, C. Ollion, and Y. Roudi, Eur. Phys. J. B **77** 587 (2010).
[21] S. Cocco and R. Monasson, Phys. Rev. Lett. **106**, 090601 (2011).
[22] J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese, Phys. Rev. Lett. **107**, 220601 (2011).
[23] R. Tibshirani, J. Roy. Stat. Soc., Ser. B **58**, 267 (1996).
[24] D. Donoho and M. Elad, Proc. Natl. Acad. Sci. U.S.A. **100**, 2197 (2003).
[25] M. Gustafsson, M. Hörnquist, J. Lundström, J. Björkegren, and J. Tegnér, Ann. N.Y. Acad. Sci. **1158**, 265 (2009).
[26] K. Koh, S. Kim, and S. Boyd, J. Mach. Learn. Res. **8**, 1519 (2007), http://jmlr.csail.mit.edu/papers/volume8/koh07a/koh07a.pdf.
[27] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[28] J. Bento and A. Montanari, Adv. Neural Inf. Process. Syst. **22** 1303 (2009), http://books.nips.cc/papers/files/nips22/NIPS2009_0870.pdf.
[29] D. Zobin, Phys. Rev. B **18**, 2387 (1978).