MATHEMATICAL BIOSCIENCES
AND ENGINEERING
Volume **11**, Number **1**, February **2014**

doi:10.3934/mbe.2014.11.149

pp. **149–165**

NETWORK INFERENCE WITH HIDDEN UNITS

Joanna Tyrcha

Department of Mathematics, Stockholm University
Kräftriket, S-106 91 Stockholm, Sweden

John Hertz

Nordita, Stockholm University and KTH
Roslagstullsbacken 23, S-106 91 Stockholm, Sweden
and
Niels Bohr Institute, Blegdamsvej 17
DK-2100 Copenhagen, Denmark

Abstract. We derive learning rules for finding the connections between units in stochastic dynamical networks from the recorded history of a "visible" subset of the units. We consider two models. In both of them, the visible units are binary and stochastic. In one model the "hidden" units are continuous-valued, with sigmoidal activation functions, and in the other they are binary and stochastic like the visible ones. We derive exact learning rules for both cases. For the stochastic case, performing the exact calculation requires, in general, repeated summations over an number of configurations that grows exponentially with the size of the system and the data length, which is not feasible for large systems. We derive a mean field theory, based on a factorized ansatz for the distribution of hidden-unit states, which offers an attractive alternative for large systems. We present the results of some numerical calculations that illustrate key features of the two models and, for the stochastic case, the exact and approximate calculations.

1. **Introduction.** Recent interest in network identification problems has been motivated by the advent of multi-electrode neural recordings and other large-scale biological data [16, 13, 12, 7]. Current inference methods, however, do not take into account the effects of units in the networks that are not recorded, though they are almost always present. This problem can be serious: For example, in cortical neural data, almost all recorded cells are excitatory, though inhibitory cells are essential in the network dynamics. In this paper we extend previous methodology to include "hidden units", presenting algorithms for inferring the strengths of connections to, from and among them.

There is a long history of work of problems of this sort. Perhaps the best know is that on "Boltzmann machines" [1]. These are symmetrically coupled networks of stochastic binary units. Their states are updated, one randomly chosen unit at a time, with the probability of being in a particular one of its two possible states given by a logistic sigmoid function of the net input from other units. Because of the symmetric coupling matrix, their dynamics satisfies detailed balance, so their equilibrium distributions are of Gibbs-Boltzmann form $Z^{-1}\exp(-E)$, where E is

2010 *Mathematics Subject Classification.* Primary: 62M45, 82C20; Secondary: 62J02.

Key words and phrases. Network inference, latent variables, kinetic Ising models, mean field theory, hidden units.

a quadratic form. This fact that simplifies their analysis considerably. The problem has also been studied in networks where the unit outputs are continuous sigmoidal functions of their inputs, for both continuous-time (asynchronous-update) and discrete-time (simultaneous-update) dynamics, extending the back-propagation algorithm used earlier for layered networks.

Applying either of these kinds of models to multineuron spike data is problematic. Real biological networks do not have symmetric connections, invalidating the first kind, while the nature of synaptic transmission and neuronal spiking calls for a stochastic binary representation, ruling out the second. In this paper we treat models in which the recorded neurons are stochastic and binary, and there is no symmetry requirement on the connections in the network. They obey a discrete-time kinetic Ising (Glauber) dynamics [6], and a value +1 represents an action potential. We study two kinds of models, in which the hidden units are deterministic or stochastic, respectively. We employ, for convenience, a discrete-time dynamics [10], though it should be straightforward to extend the treatment to continuous-time models.

2. **Continuous, deterministic hidden units.**

We examine first the deterministic case, taking the output of a hidden unit to be a sigmoidal function of its input. Though it is a big simplification of a real spiking-neuron network, this kind of model can be practical for analyzing neural data. One cannot hope to model the detailed dynamics of all the unrecorded neurons in the network of interest, because they vastly outnumber the recorded ones. However, at least as a first approximation, one can hope to describe the effect of unrecorded *populations* of neurons (for example, of inhibitory neurons when only excitatory neurons have been recorded). The values of the hidden units in our model here could represent the firing rates of those populations. The available data might capture essential features of the network dynamics, though they would never be sufficient to identify the entire network in detail.

We draw here on work in learning in analog neural networks a couple decades ago, under the names "back-propagation in time" and "recurrent back-propagation" [14, 11, 9, 21]. Our treatment differs from that work in having stochastic visible units and a likelihood-based objective function.

2.1. **Model.**

We denote the states of the visible units by $s_i(t)$, where i labels the unit and t the time bin. They can take the values ±1. (We assume the recorded spikes have been sorted into time bins small enough that there is no more than one spike per bin.) We denote the hidden unit values by $\mu_a(t)$, $1 \leq \mu_a(t) \leq 1$. To make our equations a little more transparent, we use indices $i, j, \cdots$ for visible units and $a, b, \cdots$ for hidden ones. Our model is defined by the stochastic evolution rule

$$P[s_i(t+1)|\{s(t), \mu(t)\}] = \frac{\exp[s_i(t+1)H_i(t)]}{2\cosh H_i(t)} \tag{1}$$

$$\mu_a(t+1) = \tanh B_a(t), \tag{2}$$

with

$$H_i(t) = \sum_j J_{ij} s_j(t) + \sum_b K_{ib} \mu_b(t) \tag{3}$$

$$B_a(t) = \sum_j L_{aj} s_j(t) + \sum_b M_{ab} \mu_b(t). \tag{4}$$

All $s_i(t+1)$ are assumed independent, conditional on $\{s_j(t)\}, \{\mu_b(t)\}$. The model is pictured in Fig. 1. We do not write constant bias terms in H or B here; they can be included by adding input units which are always $+1$. We will denote the number of visible units by N_v and the number of hidden ones by N_h.

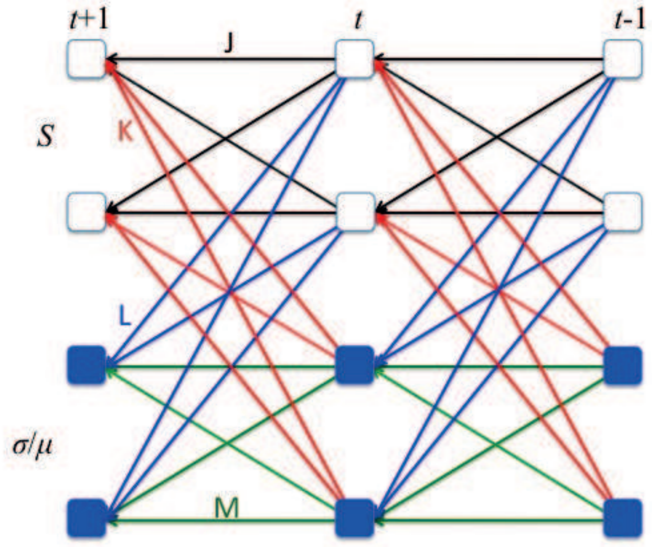

FIGURE 1. Schematic picture of the model. (Color online) White squares represent visible units s_i; blue ones, hidden units μ_a (or σ_a when they are stochastic). Visible-visible connections J_{ij} are black, hidden-to-visible ones K_{ib} are red, visible-to-hidden ones L_{aj} are blue, and hidden to hidden ones M_{ab} are green. Rows represent time steps.

2.2. **Objective function and learning rules.** We assume we are given the data $\{s_i(t)\}$ and that we know the number of hidden units. The task is to learn the connections $\{J_{ij}\}$, $\{K_{ia}\}$, $\{L_{aj}\}$, and $\{M_{ab}\}$, and our objective function is the log likelihood of the observed visible history:

$$\mathcal{L} = \sum_{it} [s_i(t+1)H_i(t) - \log 2\cosh H_i(t)]. \tag{5}$$

We consider the simplest form of gradient-based learning, where the parameters are adjusted proportional to the derivative of the log likelihood with respect to them. For $\{J_{ij}\}$ and $\{K_{ia}\}$, this is straightforward:

$$\begin{aligned} \Delta J_{kl} &= \sum_{it}[s_i(t+1)-\tanh(H_i(t))]\frac{\partial H_i(t)}{\partial J_{kl}} = \sum_t \epsilon_k(t+1)s_l(t), & (6)\\ \Delta K_{kb} &= \sum_{it}[s_i(t+1)-\tanh(H_i(t))]\frac{\partial H_i(t)}{\partial K_{kb}} = \sum_t \epsilon_k(t+1)\mu_b(t), & (7)\end{aligned}$$

with $\epsilon_k(t+1) = s_i(t) - \tanh H_i(t)$, the observed error on unit i at $t+1$ under the model with the current parameters, given its state at t. This is standard error × input learning, as in networks without hidden units [14, 12].

For the connections that lead to hidden units, the derivatives of $H_i(t)$ with respect to $\{L_{aj}\}$ and $\{M_{ab}\}$ are through its dependence on the $\mu_b(t)$; as in

$$\Delta L_{al} = \sum_{it}\epsilon_i(t+1)\frac{\partial H_i(t)}{\partial L_{al}} = \sum_{it}\epsilon_i(t+1)\sum_b K_{ib}\frac{\partial \mu_b(t)}{\partial L_{al}}. \quad (8)$$

Furthermore, the derivatives of the $\mu_j(t)$ have terms proportional to derivatives of all the μs at the previous time step:

$$\frac{\partial \mu_b(t)}{\partial L_{al}} = (1-\mu_b^2(t))\left[\delta_{ab}s_l(t-1) + \sum_c M_{bc}\frac{\partial \mu_c(t-1)}{\partial L_{al}}\right]. \quad (9)$$

These equations can be iterated starting from the initial condition $\partial\mu_c(0)/\partial L_{al} = 0$. The solution can be written relatively compactly:

$$\frac{\partial \mu_b(t)}{\partial L_{al}} = X_{bb}(t)\left\{\delta_{ab}s_l(t-1) + \sum_{q=1}^{t-1}\left[\prod_{r=1}^{q}[\mathsf{MX}(t-r)]\right]_{ba} s_l(t-q-1)\right\}, \quad (10)$$

where

$$X_{ab}(t) = (1-\mu_a^2(t))\delta_{ab} \quad (11)$$

and we make the convention that the product over r is equal to 1 when $q=0$. The learning rule for L_{al} can then be written as

$$\Delta L_{al} = \sum_t\sum_{q=0}^{t-1}\sum_i \epsilon_i(t+2+q)\left[\mathsf{KX}(t+1+q)\left(\prod_{r=1}^{q}\mathsf{MX}(t+r)\right)\right]_{ia} s_l(t), \quad (12)$$

Exactly the same procedure for the derivative with respect to M_{ab} gives

$$\Delta M_{ab} = \sum_t\sum_{q=0}^{t-1}\sum_i \epsilon_i(t+2+q)\left[\mathsf{KX}(t+1+q)\left(\prod_{r=1}^{q}\mathsf{MX}(t+r)\right)\right]_{ia} \mu_b(t), \quad (13)$$

which differs from (12) only in the last factor.

This all has a nice graphical interpretation. The effective error is the sum over all paths starting at future visible units (time $t+2+q$) and propagating back through the hidden units at intermediate times until it reaches the receiving unit a at time $t+1$. For each such path, we pick up a factor $\epsilon_i(t+q+2)$ at the visible error source, a factor of a K_{ib} for backpropagating from the source unit to a hidden unit b, factors of elements of M for the hidden–to-hidden connections on the path, and factors of $X_{cc} = 1-\mu_c^2$ at every hidden unit c that it passes through. This is just the standard prescription for back-propagation of errors in layered networks. Fig. 2 shows a typical path for $q=2$.

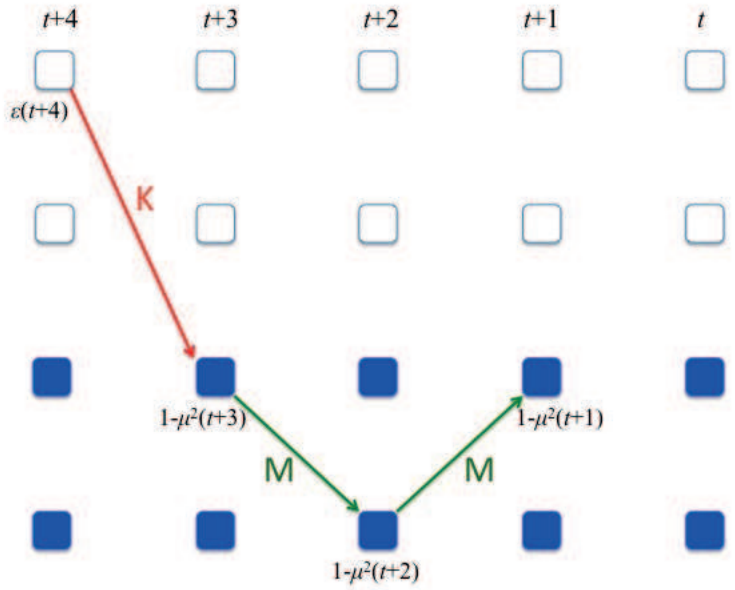

Figure 2. Back-propagation of errors from the future through the hidden units. The example path here starts at a visible unit i where the output error $\epsilon_i(t+4)$ is measured. It is then propagated back in time, first to a hidden unit at time $t+3$, then through another hidden unit at $t+2$ and finally to the one at $t+1$ which is the receiving unit on the connection being evaluated. It gives a change in that connection strength equal to the product of $\epsilon_i(t+4)$, all the connection strengths on the path, and factors of $1-\mu_b^2(t)$ for each hidden unit on the path. The total connection strength change is a sum over all such paths from all visible units in the future.

2.3. **Numerical results.** In this calculations reported in this paper we restrict ourselves to networks with no hidden-to-hidden connections ($\mathsf{M}=0$). This simplifies the learning algorithm considerably: There are no backpropagation paths longer than two steps. Fig. 3 shows an example of learning for a network with 18 visible and 2 hidden units, based on $T=10000$ time steps of data. The top left panel shows how the cost function (the negative log-likelihood of the data per time step) falls smoothly to a minimum. The top right panel shows the evolution of the errors in the couplings J_{ij}, K_{ib} and L_{aj} under learning. The apparent poor performance can be understood by comparing the middle panels, which show the coupling matrix elements of the model that generated the data (left) and the inferred couplings (right), respectively. It is apparent that the input connection strengths L_{2j} to the second hidden unit (unit 20 in these plots) are negatives of each other in the two panels. The same is true of the outgoing connections K_{i2} from that unit, though it is hard to see in these graphs. These two inversions have no effect on the visible units, so the "true" model and the one with the flipped signs of L_{2j} and K_{i2} are equivalent: there is no way we can know from the visible data alone which one was the true model. The bottom left panel shows how, if we bias the initial random values of the couplings to have the right sign, results close to the true model are obtained. The equivalence of the two inferred models is apparent from the fact (bottom right panel) that the final values of the cost function are exactly the same.

In this example it was easy to see the relation between the inferred and true connections. However, in general there is a $2^{N_h} \times N_h!$-fold degeneracy (the signs of the connections to and from every hidden unit could be flipped, and the labels on the hidden units can be permuted arbitrarily.) Thus, for large N_h, most likely one will infer one of the models equivalent to the true one, but not the true one itself.

In all our calculations in this paper, we used "batch learning" (updating based on the gradient of the total log likelihood). We always found optimal convergence (maximum speed with monotonic cost function) with learning rates of order η/T, where T is the number of time steps. In the above calculations we used $\eta = 0.5$; otherwise we used $\eta = 1$.

3. **Stochastic hidden units.** The case where all units in the model, including the hidden ones, are stochastic is more difficult, but it is the more interesting one from a theoretical point of view. Furthermore, it is not irrelevant to data analysis. While complete inference of the couplings to, from, and among a set of unobserved neurons much more numerous than the recorded ones is not practically possible, performing the inference assuming a much smaller $N_h \sim N_v$ can still give some insight into what "hidden" neurons are doing.

Denoting the hidden units by $\sigma_a(t)$, the dynamics are now given by

$$P[s_i(t+1), \sigma_a(t+1)|\{s(t), \sigma(t)\}] = \frac{\exp[s_i(t+1)H_i(t)]}{2\cosh H_i(t)} \frac{\exp[\sigma_a(t+1)B_a(t)]}{2\cosh B_a(t)} \tag{14}$$

with

$$\begin{aligned} H_i(t) &= \sum_j J_{ij} s_j(t) + \sum_b K_{ib}\sigma_b(t) && (15) \\ B_a(t) &= \sum_j L_{aj} s_j(t) + \sum_b M_{ab}\sigma_b(t). && (16) \end{aligned}$$

We restrict our treatment to networks with weak dense random connections, $J_{ij}, L_{aj} = \mathcal{O}(1/\sqrt{N_v})$, $K_{ib}, M_{ab} = \mathcal{O}(1/\sqrt{N_h})$, so that $H_i(t)$ and $B_a(t)$ are of order 1.

The likelihood of the history of the full system is

$$P[s, \sigma] = \prod_{tia} P[s_i(t+1), \sigma_a(t+1)|\{s(t), \sigma(t)\}], \tag{17}$$

and the likelihood of the visible history is

$$P[s] = \sum_\sigma P[s, \sigma]. \tag{18}$$

The distribution of the σ, conditional on the observed data, is

$$P[\sigma|s] = \frac{P[s, \sigma]}{P[s]}. \tag{19}$$

This has the form of a Gibbs distribution $Z_s^{-1}\exp(-E_s[\sigma])$, with

$$E_s[\sigma] = \log P[s, \sigma] \tag{20}$$

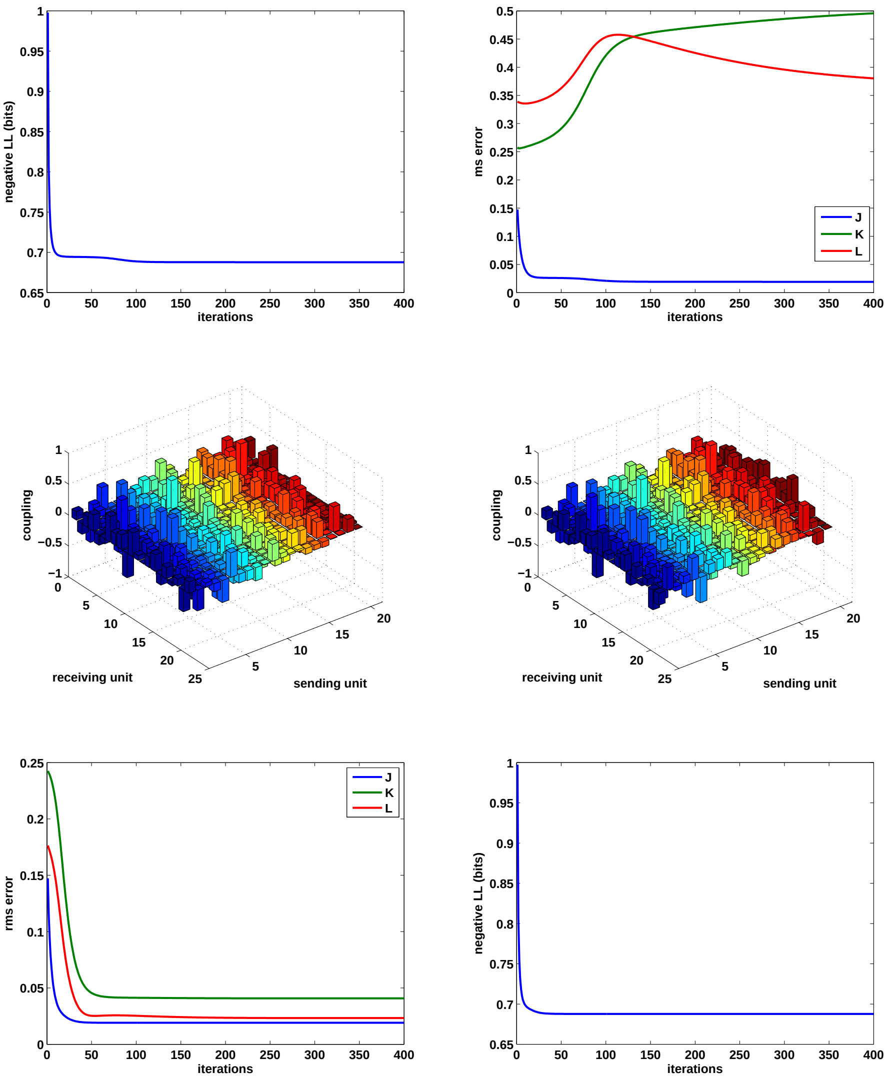

FIGURE 3. Learning example: network of 18 visible and 2 hidden units (no hidden-hidden connections). Top left: iterative minimization of the cost function $-\mathcal{L}/T$ for a data set of length $T = 10000$. Top right: rms errors on J_{ij}, K_{ib}, and L_{aj} as functions of the number of iterations of the learning algorithm when it is started at small random values of the couplings. Middle panels: true (left) and inferred coupling strengths. The hidden units are number 19 and number 20. Bottom panels: rms errors (left) and cost function (right) when the initial parameter values have the correct signs. A learning rate equal to $1/T$ was used in both cases.

and

$$Z_s = P[s] = \sum_{\sigma} P[s,\sigma]. \tag{21}$$

(Z_s also depends on all the model parameters $\{J_{ij}, K_{ib}, L_{aj}, M_{ab}\}$, but to save some space we do not write that explicitly.) To show the nature of the interactions in the energy $E_s[\sigma]$, we write it out explicitly:

$$\begin{aligned} E_s[\sigma] \quad = \quad & -\sum_t \Bigg\{ \sum_{ij} s_i(t+1) J_{ij} s_j(t) + \sum_{ib} s_i(t+1) K_{ib} \sigma_b(t) \\ + \quad & \sum_{aj} \sigma_a(t+1) L_{aj} s_j(t) + \sum_{b} \sigma_a(t+1) M_{ab} \sigma_b(t) \\ - \quad & \sum_i \log 2\cosh \Bigg[\sum_j J_{ij} s_j(t) + \sum_b K_{ib} \sigma_b(t) \Bigg] \\ - \quad & \sum_a \log 2\cosh \Bigg[\sum_j L_{aj} s_j(t) + \sum_b M_{ab} \sigma_b(t) \Bigg] \Bigg\}. \end{aligned} \tag{22}$$

The first term is just a constant (independent of the σs), the next two are like external fields acting on the $\sigma_a(t)$ from the visible data $s_i(t \pm 1)$ one time step in the future and past, respectively, and the fourth term represents interactions between σs at successive time steps. The final two terms are interactions among all the σs at one time (but these terms do not couple σs at different times). Their non-polynomial form leads to important features in this problem that are not present in Boltzmann machines.

3.1. **Exact learning algorithm.** Just as for Boltzmann machines, we can derive an exact learning algorithm for the model parameters by gradient ascent on $\log Z_s$, the log likelihood of the visible history. It can be written

$$\Delta J_{ij} \propto \frac{\partial \log Z_s}{\partial J_{ij}} = \sum_t [s_i(t+1) - \langle \tanh H_i(t) \rangle_{\sigma|s}] s_j(t) \tag{23}$$

$$\Delta K_{ib} \propto \frac{\partial \log Z_s}{\partial K_{ib}} = \sum_t \langle [s_i(t+1) - \tanh H_i(t)] \sigma_b(t) \rangle_{\sigma|s} \tag{24}$$

$$\Delta L_{aj} \propto \frac{\partial \log Z_s}{\partial L_{aj}} = \sum_t \langle \sigma_a(t+1) - \tanh B_a(t) \rangle_{\sigma|s} s_j(t) \tag{25}$$

$$\Delta M_{ab} \propto \frac{\partial \log Z_s}{\partial M_{ab}} = \sum_t \langle [\sigma_b(t+1) - \tanh B_a(t)] \sigma_b(t) \rangle_{\sigma|s} \tag{26}$$

The averages $\langle \cdots \rangle_{\sigma|s}$ are over all hidden histories $\sigma(t)$, weighted by the probability $P[\sigma|s] = Z_s^{-1} \exp\{-E_s[\sigma]\}$ that they produce the known visible history s. In each learning rule, the first term comes from differentiating the terms in the first two lines of (22) and the second from differentiating one of the $\log 2\cosh$ terms.

When there are no hidden-to-hidden connections M_{ab}, $P[\sigma|s]$ becomes a product of independent terms, one for each t. The averages over $P[\sigma|s]$ in (23-26) then involve sums over 2^{N_h} terms, where N_h is the number of hidden units. For small networks, they can be computed exactly in a reasonable time.

3.2. **Numerical results.** Fig. 4 shows, for a model with $N_h = N_v = 10$ (and, again, no hidden-to-hidden connections), how the cost function ($-\mathcal{L}/T$, the negative log-likelihood of the data per time step) converges to its asymptotic value as the number of steps in the data set is increased. All the couplings in this example were i.i.d. and normal with variance 0.1. In addition to $-\mathcal{L}/T$ evaluated on the training data, we plot it evaluated on an independently-generated test data set. It is evident that these converge to a common value for large T. We also plot the values of the Akaike and Bayesian information criteria, based on the training cost function. The Akaike information criterion penalizes the estimated log likelihood (i.e., increases the cost) by the number of parameters N, and the Bayesian information criterion penalizes it by $N \log N$. Thus these statistics are equal to the training cost function plus N/T and $(N \log N)/T$, respectively, so they also approach the training set cost function as $T \to \infty$.

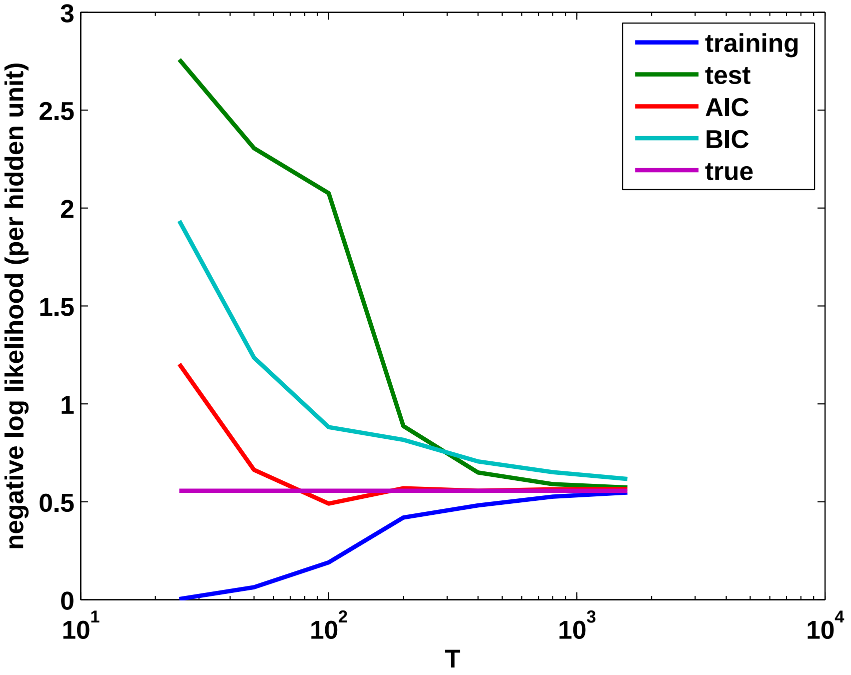

FIGURE 4. Cost functions for learning in a network of 10 visible and 10 hidden units, as functions of the data length T. (Color online) Blue: evaluated on training data. Green: evaluated on independent test data. Red: Akaike information criterion (AIC [2]). Cyan: Bayesian information criterion (BIC [17]). Purple: $T \to \infty$ limiting value. A learning rate equal to $1/T$ was used in all calculations.

For networks larger than ~ 10, one has to resort to Monte Carlo to estimate the averages. When there are hidden-to-hidden connections, the number of states to sum over becomes $2^{N_h T}$, where T is the number of time steps in the data. In this case, exact calculations are never possible, even for just one hidden unit, and even Monte Carlo becomes impractical for moderate numbers of hidden units.

4. Mean field theory for stochastic hidden units.

An attractive approximate alternative is mean field theory. It can be formulated variationally [3]: One seeks the best approximation to $P[\sigma|s]$ that factorizes over the different $\sigma_a(t)$. Each such factor is parametrized by a single number: the probability that $\sigma_a(t) = +1$. Equivalently (and conventionally), one can use the "magnetization", denoted $\mu_a(t)$, which is the difference between the probabilities to be $+1$ and -1. The entire factorizable distribution is then parametrized by the set of magnetizations $\{\mu_a(t)\}$. The learning proceeds in an EM fashion [18, 4], iterating the two steps: (1) For given coupling parameters, find the $\mu_a(t)$ that maximize the factorized $\log Z_s$, and (2), for these $\mu_a(t)$, improve the estimates of the coupling parameters as in rules (23-26) but with the averages computed under the factorized approximate $P[\sigma|s]$.

4.1. Derivation of mean-field theory.

Under the factorizability assumption, the likelihood of the visible data $\{s_i(t)\}$, given $\langle\sigma_a(t)\rangle = \mu_a(t)$, is

$$P_{MF}[\mu, s] = \exp\{S[\mu] - E_s[\mu]\} \equiv \exp A[\mu, \{J_{ij}, K_{ib}, L_{aj}, M_{ab}\}], \tag{27}$$

where

$$S[\mu] = -\sum_{at}\left[\frac{1+\mu_a(t)}{2}\log\left(\frac{1+\mu_a(t)}{2}\right) + \frac{1-\mu_a(t)}{2}\log\left(\frac{1-\mu_a(t)}{2}\right)\right] \tag{28}$$

is the entropy: the average log of the probability of magnetizations $\mu_a(t)$. In (27) we indicate explicitly that A depends on the parameters $\{J_{ij}, K_{ib}, L_{aj}, M_{ab}\}$ (through E_s). Thus, the EM learning procedure involves repeatedly maximizing over μ for fixed parameters (the "E-step") and taking uphill steps on A (equivalently, downhill steps on E_s) in parameter space for fixed μ (the "M-step").

The prescription for obtaining the average energy by the replacement $\sigma_a(t) \to \mu_a(t)$ in E_s is based on the independence of different $\sigma_a(t)$ under the factorized distribution. For example, if E_s contains a term like $\sigma_a(t)\sigma_b(t)$, then $\langle\sigma_a(t)\sigma_b(t) = \langle\sigma_a(t)\rangle\langle\sigma_b(t)\rangle = \mu_a(t)\mu_b(t)$. Thus, one might think that we get the $E_s[\mu]$ to use in (27) by simply substituting μ for σ in (22). Then maximizing $A[\mu]$ would lead to the equations

$$\begin{aligned}\tanh^{-1}\mu_a(t) &= \textstyle\sum_j L_{aj}s_j(t-1) + \sum_b M_{ab}\mu_b(t-1)\\ &+\textstyle\sum_i \left\{s_i(t+1) - \tanh\left[\sum_j J_{ij}s_j(t) + \sum_b K_{ib}\mu_b(t)\right]\right\}K_{ia}\\ &+\textstyle\sum_b \left\{\mu_b(t+1) - \tanh\left[\sum_j L_{bj}s_j(t) + \sum_c M_{bc}\mu_c(t)\right]\right\}M_{ba}\end{aligned} \tag{29}$$

for the $\mu_a(t)$. This equation has a nice interpretation: The first two terms are just the inputs from visible and hidden units, respectively, at the previous time step, and the last two terms are just the back-propagated errors from visible and hidden units one time step later.

However appealing this equation looks, it is wrong. One has to be careful in the $\log 2\cosh$ terms in $E_s[\sigma]$. Expanding it in powers of the K_{ib}, we get a second order term proportional to $\sum_{ab} K_{ia}K_{ib}\sigma_a\sigma_b$. The double sum includes terms with $a = b$, and for these terms we should make the replacement $\sigma_a^2 = 1$, not $\sigma_a^2 = \mu_a^2$. (This situation does not arise for the usual Ising energy $-\sum_{i<j} J_{ij}s_is_j$, since the $i = j$ term is explicitly excluded from the sum.) The same problem comes up in all higher-order terms in the expansion whenever there are repeated indices in the

sums over hidden unit indices. This problem was noticed already by Saul *et al* [15], who tried to deal with it by introducing an extra set of variational parameters. Here, we make a treatment for a particular ensemble of models that is exact (within the factorization approximation) in the limit of large N_h. In these models, all the couplings are zero-mean independent random numbers with variances proportional to $1/N$. This makes the net inputs $H_i(t)$ and $B_a(t)$ Gaussian (for large N), with variances of order unity.

Writing the nth order term in the expansion of $\log 2\cosh H$ (we drop the visible unit index i temporarily here, for simplicity) as

$$\alpha_n = \frac{c_n}{n!}\sum_{a_1\cdots a_n} K_{a_1}\cdots K_{a_n}\sigma_{a_1}\cdots\sigma_{a_n}, \tag{30}$$

consider first the terms in every term of (30) where two of the indices are equal. There are $n(n-1)/2$ such pairs, so the correction to this subset of the nth order terms is

$$\gamma_n^{(2)} = \tfrac{1}{2}\frac{c_n}{(n-2)!}\sum_{a_1,a_2,\cdots a_{n-2}} K_{a_1}\cdots K_{a_{n-2}}\sigma_{a_1}\cdots\sigma_{a_{n-2}}\left[\sum_a K_a^2(1-\mu_a^2)\right]. \tag{31}$$

because naive substitution of μ_a for σ_a would have given μ_a^2 instead of 1. But what multiplies the sum on a here is just half the $n-2$nd term in the expansion of the second derivative of $\log 2\cosh H$, i.e., $1-\tanh^2 H$. So we can sum all such terms over n, yielding a correction

$$E_2 = \tfrac{1}{2}(1-\tanh^2 H)\sum_a K_a^2(1-\mu_a^2). \tag{32}$$

Thus, at this level of approximation, we should use an energy $E_s[\mu]$ in which $\sum_i \log 2\cosh H_i(t)$ is replaced by

$$\sum_i \log 2\cosh H_i(t) + \tfrac{1}{2}\sum_{ia}[1-\tanh^2 H_i(t)]K_{ia}^2[1-\mu_a^2(t)] \tag{33}$$

(now with the substitution $\sigma \to \mu$ in the first term). This looks like the TAP term in the free energy for the usual Ising model, but with the opposite sign. The same argument applies to the $\log 2\cosh B$ term in E_s, which should be replaced by

$$\sum_a \log 2\cosh B_a(t) + \tfrac{1}{2}\sum_{ab}[1-\tanh^2 B_a(t)]M_{ab}^2[1-\mu_b^2(t)]. \tag{34}$$

These corrections will lead to new terms in the MF equations for $\mu_a(t)$ and in the learning rule for the K_{ia} and M_{ab}. Note also that, for the models we are considering, these correction terms are of order 1 (per visible or hidden unit, respectively) since they contain sums of N_h terms and each term is of order $1/N_h$.

We can also sum up terms with 2 pairs of indices equal, 3 pairs of indices, equal, etc. Consider first the terms where two pairs of indices are equal. In the nth order term α_n (30), there are $n!/[4!(n-4)!]$ ways of picking the 4 indices and 3 ways to pair them. The correction is

$$\gamma_n^{(4)} = \frac{3}{4!}\frac{c_n}{(n-4)!}\sum_{a_1,a_3,\cdots a_{n-4}} K_{a_1}\cdots K_{a_{n-4}}\sigma_{a_1}\cdots\sigma_{a_{n-4}}\cdot\left[\sum_a K_a^2(1-\mu_a^2)\right]^2 \tag{35}$$

The sum over n of these terms is just

$$E_4 = \frac{3}{4!}\frac{\partial^4(\log 2\cosh H)}{\partial H^4}\left[\sum_a K_a^2(1-\mu_a^2)\right]^2. \tag{36}$$

Like (32), (35) is of order 1.

Extending this argument to the general term with $j/2$ pairs of coincident indices, in the nth order term, there are $n!/[j!(n-j)!]$ ways to pick our the j indices, and the number of ways to pair them is $(j-1)!! \equiv (j-1)(j-3)\cdots 3\cdot 1$. Thus, we get a correction

$$E_j = \frac{(j-1)!!}{j!}\frac{\partial^j(\log 2\cosh H)}{\partial H^j}\left[\sum_a K_a^2(1-\mu_a^2)\right]^j. \tag{37}$$

Again, all these terms are all of order 1.

On the other hand, terms we have not considered, with 3 or more indices equal, are negligible in the mean-field limit $N_h \to \infty$. (Consider terms with p equal indices. They involve the sum $\sum_a K_a^p$, which is of order $N_h^{1-p/2}$ and therefore negligible for $p > 2$ as $N_h \to \infty$.

Now we can sum all the E_j over j, exploiting the fact that $(j-1)!!$ is the jth moment of a zero-mean univariate normal distribution. The result of all these manipulations is simply the replacement

$$\log 2\cosh H_i(t) \longrightarrow \int Dx \log 2\cosh[H_i(t) + \Delta_i(t)x], \tag{38}$$

where Dx means $(2\pi)^{-1/2}\mathrm{e}^{-x^2/2}dx$ and

$$\Delta_i^2(t) = \sum_a K_{ia}^2[1-\mu_a^2(t)]. \tag{39}$$

Thus, the effect of all these corrections can be described in terms of an effective Gaussian noise. The same arguments apply to the $\log 2\cosh B$ term, with the final result that the effective energy can be written, exactly in the limit $N_h \to \infty$, as

$$\begin{aligned} E_s[\mu] &= -\sum_t \Bigg\{ \sum_{ij} s_i(t+1)J_{ij}s_j(t) + \sum_{ib} s_i(t+1)K_{ib}\mu_b(t) \\ &+ \sum_{aj}\mu_a(t+1)L_{aj}s_j(t) + \sum_b \mu_a(t+1)M_{ab}\mu_b(t) \\ &- \sum_i \int Dx \log 2\cosh\left[\sum_j J_{ij}s_j(t) + \sum_b K_{ib}\mu_b(t) + \Delta_i(t)x\right] \\ &- \sum_a \int Dy \log 2\cosh\left[\sum_j L_{aj}s_j(t) + \sum_b M_{ab}\mu_b(t) + \Gamma_a(t)y\right]\Bigg\}, \end{aligned} \tag{40}$$

with

$$\Gamma_a^2(t) = \sum_b M_{ab}^2[1-\mu_b^2(t)]. \tag{41}$$

We note that this form could have been motivated heuristically: In (15) and (16), the $\sigma_b(t)$ are fluctuating variables of variance $1-\mu_b^2(t)$. Since K_{ib} and M_{ab} are assumed to be independent random variables, $H_i(t)$ and $B_a(t)$ are normally distributed with variances $\Delta_i^2(t)$ and $\Gamma_a^2(t)$ given by (39) and (41), respectively.

4.2. **Learning algorithm.** The resulting equations for the E-step are then

$$\begin{aligned}\tanh^{-1}\mu_a(t) &= \textstyle\sum_j L_{aj}s_j(t-1) + \sum_b M_{ab}\mu_b(t-1)\\
&+ \textstyle\sum_i \left\{ s_i(t+1) - \int Dx \tanh\left[\sum_j J_{ij}s_j(t) + \sum_b K_{ib}\mu_b(t) + \Delta_i(t)x\right]\right\} K_{ia}\\
&+\mu_a \textstyle\sum_i \left\{1 - \int Dx \tanh^2\left[\sum_j J_{ij}s_j(t) + \sum_b K_{ib}\mu_b(t) + \Delta_i(t)x\right]\right\} K_{ia}^2\\
&+ \textstyle\sum_b \left\{\mu_b(t+1) - \int Dy \tanh\left[\sum_j L_{bj}s_j(t) + \sum_c M_{bc}\mu_c(t) + \Gamma_b(t)y\right]\right\} M_{ba}\\
&+\mu_a \textstyle\sum_b \left\{1 - \int Dy \tanh^2\left[\sum_j L_{bj}s_j(t) + \sum_c M_{bc}\mu_c(t) + \Gamma_b(t)y\right]\right\} M_{ba}^2. \quad (42)\end{aligned}$$

They differ from the naive equations (29) in that the tanh terms in the second and fourth lines are averaged over the Gaussian noises and in the presence of the new terms on the third and fifth lines. The latter have the form of cavity field corrections [8]: The effect of μ_a itself on the expected $s_i(t+1)$ and $\mu_b(t+1)$ should not be counted in calculating the $\tanh H$ terms in the second and fourth lines.

For the M-step, the learning rules for J_{ij} and L_{aj} are

$$\Delta J_{ij} \propto -\frac{\partial E_s}{\partial J_{ij}} = \sum_t \left\{ s_i(t+1) - \int Dx \tanh\left[H_i(t) + \Delta_i(t)x\right]\right\} s_j(t) \quad (43)$$

$$\Delta L_{aj} \propto -\frac{\partial E_s}{\partial L_{aj}} = \sum_t \left\{ \mu_a(t+1) - \int Dy \tanh\left[B_a(t) + \Gamma_a(t)y\right]\right\} s_j(t), \quad (44)$$

differing from those we would find in the naive mean field theory only in the averaging of the tanh's over the Gaussian noises. The rules for K_{ib} and M_{ab},

$$\begin{aligned}\Delta K_{ib} &\propto -\frac{\partial E_s}{\partial K_{ib}} = \sum_t \left\{ \left(s_i(t+1) - \int Dx \tanh\left[H_i(t) + \Delta_i(t)x\right]\right)\mu_b(t)\right.\\
&- \left.\left[1 - \int Dx \tanh^2[H_i(t) + \Delta_i(t)x]\right] K_{ib}[1-\mu_b^2(t)]\right\} \quad (45)\\
\Delta M_{ab} &\propto -\frac{\partial E_s}{\partial M_{ab}} = \sum_t \left\{ \left(\mu_a(t+1) - \int Dy \tanh\left[B_a(t) + \Gamma_a(t)y\right]\right)\mu_b(t)\right.\\
&- \left.\left[1 - \int Dy \tanh^2[B_a(t) + \Gamma_a(t)y]\right] M_{ab}[1-\mu_b^2(t)]\right\} \quad (46)\end{aligned}$$

have extra terms that come from the dependence of $\Delta_i(t)$ and $\Gamma_a(t)$ on K_{ia} and M_{ab} in (39) and (41), respectively.

For small K_{ia} and M_{ab} (i.e., at the level of the corrections (33) and (34), the E-step equations reduce to

$$\begin{aligned}\tanh^{-1}\mu_a(t) &= \sum_j L_{aj}s_j(t-1) + \sum_b M_{ab}\mu_b(t-1)\\
&+ \sum_i \left\{[s_i(t+1) - \tanh H_i(t)]K_{ia} + [1-\tanh^2 H_i(t)]K_{ia}^2\mu_a(t)\right.\\
&+ \left.\tanh H_i(t)[1-\tanh^2 H_i(t)]K_{ia}\sum_b K_{ib}^2[1-\mu_b^2(t)]\right\}\end{aligned}$$

$$+ \sum_b \Big\{[\mu_b(t+1) - \tanh B_b(t)]M_{ba} + [1-\tanh^2 B_b(t)]M_{ba}^2 \mu_a(t)$$

$$+ \tanh B_b(t)[1-\tanh^2 B_b(t)]M_{ba} \sum_c M_{bc}^2[1-\mu_c^2(t)]\Big\}, \qquad (47)$$

and the learning rules are

$$\Delta J_{ij} \propto \sum_t \big\{s_i(t+1) - \tanh H_i(t)[1-(1-\tanh^2 H_i(t))\Delta_i(t)]\big\}\, s_j(t) \qquad (48)$$

$$\Delta L_{aj} \propto \sum_t \big\{\mu_a(t+1) - \tanh B_a(t)[1-(1-\tanh^2 B_a(t))\Gamma_a(t)]\big\}\, s_j(t), \qquad (49)$$

$$\Delta K_{ib} \propto \sum_t \big\{\big(s_i(t+1) - \tanh H_i(t)[1-(1-\tanh^2 H_i(t))\Delta_i(t)]\big)\,\mu_b(t)$$

$$- \big[1-\tanh^2 H_i(t)\big]\, K_{ib}[1-\mu_b^2(t)]\big\} \qquad (50)$$

$$\Delta M_{ab} \propto \sum_t \big\{\big(\mu_a(t+1) - \tanh B_a(t)[1-(1-\tanh^2 B_a(t))\Gamma_a(t)]\big)\,\mu_b(t)$$

$$- [1-\tanh^2 B_a(t)]M_{ab}[1-\mu_b^2(t)]\big\} \qquad (51)$$

A few final remarks are in order. The reader might notice that the lowest-order corrections in (33), (34), and (47) resemble Thouless-Anderson-Palmer (TAP) corrections in spin glasses [20]. However, there the TAP equations come from the first corrections to the factorized-distribution approximation, whereas ours here come from evaluating the average energy within that approximation. We expect that for our model here, as for spin glasses, to get an exact theory for large N_h, TAP corrections analogous to theirs should also be included. We do not try to do that here, working entirely within the factorized-distribution ansatz. In problems like ours for networks without hidden units, this is sometimes called “naive mean field theory” [12].

4.3. **Numerical results.** We have carried out mean-field inference computations for some models with no hidden-to-hidden connections ($M_{ab} = 0$), using the lowest-order mean-field equations (47-51). Fig. 5 shows how the mean square errors of J_{ij}, K_{ib} and L_{aj} depend on the data set length T for two networks with 80 visible units. The left-hand panel shows the case where the number of hidden units $N_h = 80$, and the right-had panel shows the case where $N_h = 20$. For the smaller N_h, all three mean square errors fall off like $1/T$, as we would expect to find if we could do this calculation exactly. However, for the larger N_h, while the errors on the visible-to-visible couplings also fall off with T in this way, the errors on the couplings to and from the hidden units are larger and fall off much more slowly.

We can get a little insight into this behavior by doing the mean-field calculations for small N_h, where it is also possible to do the exact calculations, as described in Sect. 2.3. Fig. 6 shows the results of both kinds of calculations for $N_v = N_h = 5$ and 8. In these cases we can see that for small T the mean-field and exact calculations nearly coincide. The T-dependence is in this region is qualitatively like that for the mean-field results at $N_v = N_h = 80$. However, at larger T, the mean-field errors all fall less rapidly. At the same time, the exact calculation gives errors on the Js which continue to fall off like $1/T$, and those on the Ks and Ls also start to fall more rapidly at the largest Ts studied. This behavior is consistent with the expectation that, as for models with no hidden units, all exact-method errors should fall off asymptotically like $1/T$, while the mean-field errors should approach limits

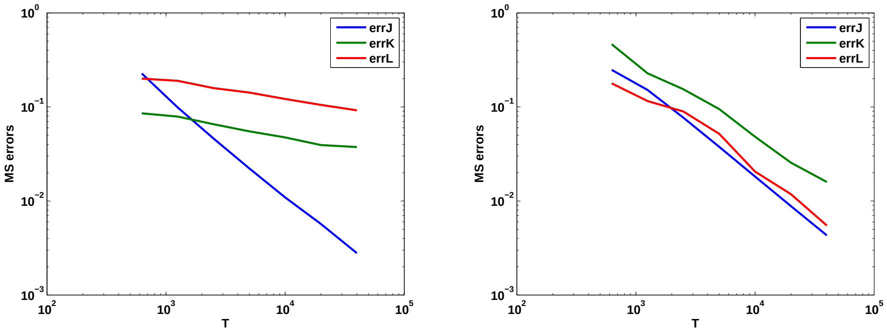

FIGURE 5. Mean square errors on Js (blue), Ks (green) and Ls (red) computed in mean field theory as functions of data set length T. (Color online) Left panel: $N_v = N_h = 80$. Right panel: $Nv = 80$, $N_h = 20$. All couplings are i.i.d. normal, with variance $1/N_v$ for J_{ij} and L_{aj} and $1/N_h$ for K_{ib}. Learning rates: $1/T$ in all cases.

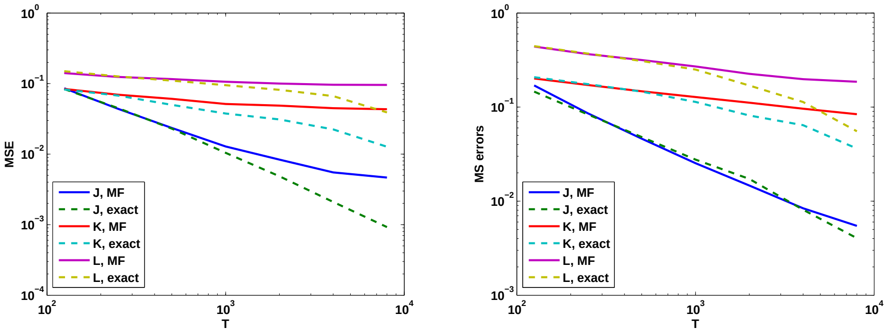

FIGURE 6. Mean square errors on Js (blue), Ks (green) and Ls (red) as functions of data set length for small networks. (Color online) Left Panel: $N_v = N_h = 5$. Right Panel: $N_v = N_h = 8$. Mean-field results are solid lines; exact results are dashed. Couplings chosen as in Fig. 5. (All learning rates $= 1/T$.)

$\propto 1/N_h$ [12]. However, apparently one has to go to very large data sets (roughly $T > 10^3 N_h$) to see this.

5. **Discussion.** We have derived learning rules for two kinds of stochastic binary networks with hidden units. These networks differ from Boltzmann machines in that (1) the units in them are updated synchronously rather than asynchronously, and (2) the connection strengths are allowed to be asymmetric. Because of these differences, the usual kind of Gibbs equilibrium does not hold, and a new kind of treatment is required.

The first kind of network has deterministic, continuous-valued hidden units. The learning rules for it are very similar to those in the back-propagation-in-time approach for recurrent networks where all the units are deterministic and continuous-valued.

Units in the second kind of network are binary and stochastic, like the visible units. Here the learning problem is harder, but we have showed that one can always put it into the form of an equilibrium statistical mechanical problem with a non-polynomial energy function. The learning rules involve averages over the Gibbs distribution for this problem. For small systems and in the absence of hidden-to-hidden couplings, the problem can be solved exactly numerically, but otherwise one must resort to Monte Carlo methods or other approximations. We explored in detail one such approximation: mean field theory. A careful analysis revealed that the naive way one might write this theory was wrong, but we were able to construct a version of mean field theory that was exact for weak, dense connectivity in the limit of a large number of hidden units (the analog of the Sherrington-Kirkpatrick model of spin glasses [19]).

We also performed some numerical calculations to illustrate and to begin to explore some of the features of the different kinds of networks and learning rules. A general feature is that when the number of hidden units is large (i.e., comparable to the number of visible units), the errors in determining the couplings to and from the hidden units are much larger than those on the couplings among the visible units. This is true for both kinds of networks and for both exact learning algorithms and mean field theory. This should not be surprising, since the information about the connections to and from hidden units is only available indirectly, through the statistics of the visible units. On the other hand, it is noteworthy that even a rather poor estimation of the connections to and from the hidden units does not spoil the good estimation of the couplings among the visible ones.

Another point worth mentioning is that for small data lengths mean field theory is as good as doing the full exact calculation, which would take prohibitively long for N_h much bigger than 10 or so. For large N_h the errors on connections to and from hidden units can be rather large and fall off very slowly with T, but the results on small systems seem to show that doing the exact calculation instead of mean field theory (even if this were feasible), would not help except at very large T.

We have only scratched the surface of this problem in our numerical calculations. It would be useful to know, for example, what the asymptotic errors on the Ks and Ls are for the mean-field algorithm in the limit of large data sets and at what T the approach to these values begins, as functions of N_h and N_v. We leave this and other questions to future work. The theory presented here provides a foundation for those investigations and, we hope, will point the way toward other questions that will be interesting to study.

Dunn and Roudi [5] have derived similar results in a different way for weak coupling. (In the context of the models considered by both them and us, where, e.g., $K_{ia} = \kappa \mathcal{N}(0,1)/\sqrt{N_h}$, "weak coupling" means small κ, and analogously for the other couplings.) When there are no hidden-to-hidden couplings ($\mathsf{M} = 0$), their results agree with ours in this limit (i.e., at the level of eqns. (47)-(51)), but our more general results (42)-(45) are not restricted to small κ. On the other hand, they are also able (for weak coupling) to learn hidden-to-hidden couplings M_{ab}. This requires TAP corrections which are (as remarked above) beyond the naive mean-field theory that we treat here, but which appear naturally in their approach.

Acknowledgments. We would like to thank Yasser Roudi and Ben Dunn for useful discussions.

REFERENCES

[1] D. Ackley, G. E. Hinton and T. J. Sejnowski, *A learning algorithm for Boltzmann machines*, Cogn. Sci., **9** (1985), 147–169.

[2] H. Akaike, *A new look at the statistical model identification. System identification and time-series analysis*, IEE Transactions on Automatic Control, **AC-19** (1974), 716–723.

[3] D. Barber, "Bayesian Reasoning and Machine Learning," chapter 11, Cambridge Univ. Press, 2012.

[4] A. P. Dempster, N. M. Laird and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm. With discussion*, J. Roy. Stat. Soc. B, **39** (1977), 1–38.

[5] B. Dunn and Y. Roudi, *Learning and inference in a nonequilibrium Ising model with hidden nodes*, Phys. Rev. E, **87** (2013), 022127.

[6] R. J. Glauber, *Time-dependent statistics of the Ising model*, J. Math. Phys., **4** (1963), 294–307.

[7] J. Hertz, Y. Roudi and J. Tyrcha, *Ising models for inferring network structure from spike data*, in "Principles of Neural Coding" (eds. S. Panzeri and R. R. Quiroga), CRC Press, (2013), 527–546.

[8] M. Mézard, G. Parisi and M. Virasoro, "Spin Glass Theory and Beyond," chapter 2, World Scientific Lecture Notes in Physics, **9**, World Scientific Publishing Co., Inc., Teaneck, NJ, 1987.

[9] B. A. Pearlmutter, *Learning state space trajectories in recurrent neural networks*, Neural Computation, **1** (1989), 263–269.

[10] P. Peretto, *Collective properties of neural networks: A statistical physics approach*, Biol. Cybern., **50** (1984), 51–62.

[11] F. J. Pineda, *Generalization of back-propagation to recurrent neural networks*, Phys. Rev. Lett., **59** (1987), 2229–2232.

[12] Y. Roudi and J. Hertz, *Mean-field theory for nonequilibrium network reconstruction*, Phys. Rev. Lett., **106** (2011), 048702.

[13] Y. Roudi, J. Tyrcha and J. Hertz, *The Ising model for neural data: Model quality and approximate methods for extracting functional connectivity*, Phys. Rev. E, **79** (2009), 051915.

[14] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning Internal Representations by Error Propagation*, in "Parallel Distributed Processing" (eds. D. E. Rumelhart and J. L. McClelland), Vol. 1, Chapter 8, MIT Press, 1986.

[15] L. K. Saul, T. Jaakkola and M. I. Jordan, *Mean field theory for sigmoid belief networks*, J. Art. Intel. Res., **4** (1996), 61–76.

[16] E. Schneidman, M. J. Berry, R. Segev and W. Bialek, *Weak pairwise correlations imply strongly correlated network states in a neural population*, Nature, **440** (2006), 1007–1012.

[17] G. E. Schwarz, *Estimating the dimension of a model*, Annals of Statistics, **6** (1978), 461–464.

[18] R. Sundberg, *Maximum likelihood theory for incomplete data from an exponential family*, Scand. J. Statistics, **1** (1974), 49–58.

[19] D. Sherrington and S. Kirkpatrick, *Solvable model of a spin-glass*, Phys. Rev. Lett., **35** (1975), 1792–1796.

[20] D. J. Thouless, P. W. Anderson and R. G. Palmer, *Solution of "soluble model of a spin glass,"* Philos. Mag., **92** (1974), 272–279.

[21] R. J. Williams and D. Zipser, *A learning algorithm for continually running fully recurrent networks*, Neural Comp., **1** (1989), 270–280.

Received December 11, 2012; Accepted May 22, 2013.

E-mail address: joanna@math.su.se
E-mail address: jhertz@nordita.org