On Maximum a Posteriori Estimation of Hidden Markov Processes

Armen Allahverdyan
Yerevan Physics Institute
Yerevan 375036, Armenia
aarmen@mail.yerphi.am

Aram Galstyan
USC Information Sciences Institute
Marina del Rey, CA 90292, USA
galstyan@isi.edu

Abstract

We present a theoretical analysis of Maximum a Posteriori (MAP) sequence estimation for binary symmetric hidden Markov processes. We reduce the MAP estimation to the energy minimization of an appropriately defined Ising spin model, and focus on the performance of MAP as characterized by its accuracy and the number of solutions corresponding to a typical observed sequence. It is shown that for a finite range of sufficiently low noise levels, the solution is uniquely related to the observed sequence, while the accuracy degrades linearly with increasing the noise strength. For intermediate noise values, the accuracy is nearly noise-independent, but now there are *exponentially* many solutions to the estimation problem, which is reflected in non–zero ground–state entropy for the Ising model. Finally, for even larger noise intensities, the number of solutions reduces again, but the accuracy is poor. It is shown that these regimes are different thermodynamic phases of the Ising model that are related to each other via first-order phase transitions.

1 Introduction

Hidden Markov Models (HMM) are used extensively for modeling sequential data in various areas [9, 4]: information theory, signal processing, bioinformatics, mathematical economics, linguistics, *etc.* One of the main problems underlying many applications of HMMs amounts to inferring the hidden state sequence $\mathbf{x}$ based on noise-corrupted observation sequence $\mathbf{y}$. This is often done through maximum a posteriori (MAP) approach, which finds an estimate $\hat{\mathbf{x}}(\mathbf{y})$ by maximizing the posterior probability $\Pr(\mathbf{x}|\mathbf{y})$.

The computational solution to the MAP optimization problem is readily available via the Viterbi algorithm [9]. Despite its extensive use in many applications, however, the properties of MAP estimation, and specifically, the structure of its solution space, have received surprisingly little attention. On the other hand, it is clear that choosing a single state sequence might be insufficient for adequately understanding the structure of the inferred process. To get a more complete picture, one needs to know whether there are other nearly optimal sequences, how many of them, how they compare with the optimal solution, and so on.

Generally, the structure of an inference method can be characterized by the accuracy of the estimation, and the number $\mathcal{N}(\mathbf{y})$ of solutions $\hat{\mathbf{x}}(\mathbf{y})$ that the method can produce in response to a given sequence $\mathbf{y}$. In this paper we study the structure of MAP inference for the simplest binary, symmetric HMM. As an accuracy measure we employ the moments of the estimated sequence $\hat{\mathbf{x}}$ in comparison of those of the actual sequence $\mathbf{x}$, while the number $\mathcal{N}(\mathbf{y})$ of possible estimates will be characterized by its averaged logarithm $\sum_{\mathbf{y}} \Pr(\mathbf{y}) \ln \mathcal{N}(\mathbf{y})$. The binary symmetric HMM is studied by reducing it to the Ising model in random fields, a relation well-known both in computer science [6] and statistical physics [8]. In this way, the average cost $-\sum_{\mathbf{y}} \Pr(\mathbf{y}) \Pr(\hat{\mathbf{x}}(\mathbf{y})|\mathbf{y})$ of MAP and the logarithm of the number of solutions $\sum_{\mathbf{y}} \Pr(\mathbf{y}) \ln \mathcal{N}(\mathbf{y})$ relate, respectively, to the energy and the entropy of the Ising model at the zero temperature.

Our results indicate that even for a simple process such as binary symmetric HMM, MAP yields a very rich and non-trivial solution structure. The main findings can be summarized as follows: For a small, but finite range of noise values the MAP solution is uniquely related to the observed sequence, and the accuracy of the solution degrades linearly with increasing the noise strength. For intermediate values of noise the accuracy is nearly noise-independent, but now there are *exponentially* many solutions to the estimation problem,

which is reflected in non–zero ground–state entropy for the Ising model. Finally, for larger noise intensities the number of solutions is reduced again, but the accuracy is poor. Furthermore, those regimes are the manifestation of different thermodynamic phases of the Ising model, which are related to each other via first-order phase transitions.

The rest of the paper is organized as follows: After some general discussion of MAP scheme in Section 2, we define the model studied here in Section 3. Its solution is given in Sections 4 and 5. The latter also discusses our concrete findings on the structure of MAP for the binary symmetric HMM. We conclude the paper by discussion of our results and future work.

2 Maximum a posteriori (MAP) estimation: general description

Let $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{y} = (y_1, \ldots, y_N)$ be realizations of discrete-time random processes $\mathcal{X}$ and $\mathcal{Y}$, respectively. We write their probabilities as $\mathbf{p}(\mathbf{x})$ and $\mathbf{p}(\mathbf{y})$. We assume that $\mathcal{Y}$ is the noisy observation of $\mathcal{X}$; the influence of noise is described by the conditional probability $\mathbf{p}(\mathbf{y}|\mathbf{x})$. Let us further assume that we are given an observed sequence $\mathbf{y}$, and we know the probabilities $\mathbf{p}(\mathbf{y}|\mathbf{x})$ and $\mathbf{p}(\mathbf{x})$. We do not know which specific sequence $\mathbf{x}$ generated the observation $\mathbf{y}$. MAP offers a method for estimating the generating sequence $\hat{\mathbf{x}}(\mathbf{y})$ on the ground of $\mathbf{y}$: $\hat{\mathbf{x}}$ is found by maximizing over $\mathbf{x}$ the posterior probability $\mathbf{p}(\mathbf{x}|\mathbf{y}) = \mathbf{p}(\mathbf{y}|\mathbf{x})\mathbf{p}(\mathbf{x})/\mathbf{p}(\mathbf{y})$. Since $\mathbf{p}(\mathbf{y})$ does not depend on $\mathbf{x}$, we can equally well minimize

$$- \ln[\, \mathbf{p}(\mathbf{y}|\mathbf{x})\mathbf{p}(\mathbf{x}) \,] \equiv H(\mathbf{y}, \mathbf{x}). \qquad (1)$$

The advantage of using $H(\mathbf{y}, \mathbf{x})$ is that if $\mathcal{Y}$ is ergodic (in the sense of weak law of large numbers) [2], which we assume from now on, then for $N \gg 1$, $H(\mathbf{y}, \hat{\mathbf{x}}(\mathbf{y}))$ will be independent from $\mathbf{y}$, if $\mathbf{y}$ belongs to the typical set $\Omega_N(\mathcal{Y})$ [2]. The typical set has the overall probability converging to one: $\sum_{\mathbf{y} \in \Omega_N(\mathcal{Y})} \mathbf{p}(\mathbf{y}) \to 1$. Since all elements of $\Omega_N(\mathcal{Y})$ have (nearly) equal probability, we can employ with probability one the averaged quantity $\sum_{\mathbf{y}} \mathbf{p}(\mathbf{y}) H(\mathbf{y}, \hat{\mathbf{x}}(\mathbf{y}))$ instead of $H(\mathbf{y}, \hat{\mathbf{x}}(\mathbf{y}))$.

If the noise is very weak, $\mathbf{p}(\mathbf{x}|\mathbf{y}) \simeq \delta(\mathbf{x} - \mathbf{y}) = \prod_{k=1}^{N} \delta(x_k - y_k)$ (with $\delta(x)$ being the Kronecker delta), we recover the generating sequence almost exactly. For a strong noise the estimation is dominated by the prior distribution $\mathbf{p}(\mathbf{x}|\mathbf{y}) \simeq \mathbf{p}(\mathbf{x})$, so that the estimation is not informative. When no priors are put, $\mathbf{p}(\mathbf{x}) \propto$ const, the MAP estimation reduces to the Maximum Likelihood (ML) estimation scheme. The latter also reproduces the source sequence almost exactly if the noise is weak.

According to the Viterbi algortithm, for a given $\mathbf{y}$ the mimimization of $H(\mathbf{y}, \mathbf{x})$ in (1) produces one single estimate $\hat{\mathbf{x}}(\mathbf{y})$. However, it is possible that there are other sequences $\hat{\mathbf{x}}^{[\alpha]}(\mathbf{y})$ for which $H(\mathbf{y}, \hat{\mathbf{x}}^{[\alpha]}(\mathbf{y}))$, though greater than $H(\mathbf{y}, \hat{\mathbf{x}}(\mathbf{y}))$, is almost equal to the latter in the sense of $\lim_{N \to \infty} \frac{H(\mathbf{y}, \hat{\mathbf{x}}^{[\alpha]}(\mathbf{y}))]}{N} = \lim_{N \to \infty} \frac{H(\mathbf{y}, \hat{\mathbf{x}}(\mathbf{y}))}{N}$. All such sequences are equivalent for $N \to \infty$ and we list them as possible solutions:

$$\hat{\mathbf{x}}^{[\alpha]}(\mathbf{y}), \qquad \alpha = 1, \ldots, \mathcal{N}(\mathbf{y}). \qquad (2)$$

If $\ln \mathcal{N}(\mathbf{y}) \propto N$, we repeat the above ergodicity argument and get for the logarithm of the number of solutions corresponding to a typical observed sequence

$$\Theta = \sum_{\mathbf{y}} \mathbf{p}(\mathbf{y}) \ln \mathcal{N}(\mathbf{y}). \qquad (3)$$

A finite $\frac{\Theta}{N}$ means that there are exponentially many outcomes of minimizing $H(\mathbf{y}, \mathbf{x})$ over $\mathbf{x}$. We call Θ entropy, since it relates to the entropy of the Ising model; see below.

We can calculate various moments of $\hat{\mathbf{x}}^{[\alpha]}(\mathbf{y})$, which are random variables due to the dependence on $\mathbf{y}$, and employ them for characterizing the accuracy of the estimation; see below for examples. For small noise values these moments will be close to those of the original process $\mathcal{X}$. Another useful quantity is the average overlap between the estimated sequences $\hat{\mathbf{x}}^{[\alpha]}(\mathbf{y})$, and the observed sequence $\mathbf{y}$ (definition of overlap is clarified below). A small overlap means that the estimation is not dominated by observations.

3 Binary symmetric hidden markov model (HMM).

3.1 Definition.

We consider the MAP estimation of a binary, discrete-time Markov stochastic process $\mathcal{X} = (X_1, X_2, \ldots, X_N)$. Each random variable X_k has only two realizations $x_k = \pm 1$. The Markov feature implies

$$\mathbf{p}(\mathbf{x}) = \prod_{k=2}^{N} p(x_k|x_{k-1}) p(x_1), \qquad (4)$$

where $p(x_x|x_{k-1})$ is a time-independent transition probability of the Markov process. For the considered binary symmetric situation it is parameterized by a single number $0 < q < 1$, $p(1|1) = p(-1|-1) = 1 - q$, $p(1|-1) = p(-1|1) = q$, and the stationary distribution is $p_{\text{st}}(1) = p_{\text{st}}(-1) = \frac{1}{2}$. Furthermore, the noise process is assumed to be memory-less, time-independent and unbiased:

$$\mathbf{p}(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^{N} \pi(y_k|x_k), \qquad y_k = \pm 1 \qquad (5)$$

where $\pi(-1|1) = \pi(1|-1) = \epsilon$, $\pi(1|1) = \pi(-1|-1) = 1 - \epsilon$, and ϵ is the probability of error. Here memory-less refers to the factorization in (5), time-independence refers to the fact that in (5) $\pi(...|...)$ does not depend on k, while unbiased means that the noise acts symmetrically on both realizations of the Markov process: $\pi(1|-1) = \pi(-1|1)$.

Note that the composite process $\mathcal{XY}$ with realizations (y_k, x_k) is Markov with transition probabilities

$$p(y_{k+1}, x_{k+1}|y_k, x_k) = \pi(y_{k+1}|x_{k+1})p(x_{k+1}|x_k). \quad (6)$$

However, $\mathcal{Y}$ is in general not a Markov process.

3.2 Mapping to the Ising model.

Let us represent the transition probabilities as

$$p(x_k|x_{k-1}) = \frac{e^{Jx_k x_{k-1}}}{2\cosh J}, \qquad J = \frac{1}{2}\ln\left[\frac{1-q}{q}\right]. \quad (7)$$

Likewise, we represent the noise model as

$$\pi(y_i|x_i) = \frac{e^{hy_i x_i}}{2\cosh h}, \qquad h = \frac{1}{2}\ln\left[\frac{1-\epsilon}{\epsilon}\right]. \quad (8)$$

We combine (1, 4–5) to represent the log–likelihood as

$$H(\mathbf{y}, \mathbf{x}) = -J\sum_{k=1}^{N} x_k x_{k+1} - h\sum_{k=1}^{N} y_k x_k, \quad (9)$$

where we have omitted an irrelevant additive factor. $H(\mathbf{y}, \mathbf{x})$ is the Hamiltonian of a one–dimensional (1d) Ising spin model with external random fields hy_k governed by the probability $\mathbf{p}(\mathbf{y})$ [11]. The factor J in (9) is the spin–spin interaction constant, uniquely determined from the transition probability q: If $q < \frac{1}{2}$, the constant J is positive, which refers to the *ferromagnetic* situation: the spin–spin interaction tends to align the spins. From now on we assume $J > 0, h > 0$. We note that the main difference between (9) and other random–field Ising models considered in literature [8, 1], is that in our situation the random fields are not uncorrelated random variables, but display non-Markovian correlations.

3.3 Implementation of MAP

To minimize $\sum_{\mathbf{y}} \mathbf{p}(\mathbf{y})H(\mathbf{y}, \mathbf{x})$ over $\mathbf{x}$, we introduce a non-zero temperature $T = \frac{1}{\beta} \geq 0$, and define the following conditional probability

$$\rho(\mathbf{x}|\mathbf{y}) \equiv \frac{e^{-\beta H(\mathbf{y}, \mathbf{x})}}{Z(\mathbf{y})}, \qquad Z(\mathbf{y}) \equiv \sum_{\mathbf{x}} e^{-\beta H(\mathbf{y}, \mathbf{x})}, \quad (10)$$

where $Z(\mathbf{y})$ is the partition function. In the terminology of statistical physics, $\rho(\mathbf{x}|\mathbf{y})$ gives the probability distribution of states $\mathbf{x}$ for a system with Hamiltonian

$H(\mathbf{y}, \mathbf{x})$ interacting with a thermal bath at temperature T, and with frozen (i.e., fixed for each site) random fields y_k [7]. For $T \to 0$, and a given $\mathbf{y}$, the function $e^{-\beta H(\mathbf{y}, \mathbf{x})}$ is strongly picked at those $\hat{\mathbf{x}}(\mathbf{y})$ [ground states], which minimize $H(\mathbf{y}, \mathbf{x})$. If, however, the limit $T \to 0$ is taken after the limit $N \to \infty$, we get

$$\rho(\mathbf{x}|\mathbf{y}) \to \frac{1}{\mathcal{N}(\mathbf{y})} \sum_{\alpha} \delta[\mathbf{x} - \hat{\mathbf{x}}^{[\alpha]}(\mathbf{y})], \quad (11)$$

where $\hat{\mathbf{x}}^{[\alpha]}$ and $\mathcal{N}(\mathbf{y})$ were defined in (2). From now on we understand the limit $T \to 0$ in this sense.

The average of H [average energy] in the $T \to 0$ limit will be equal to the $H(\mathbf{y}, \mathbf{x})$ minimized over $\mathbf{x}$:

$$\sum_{\mathbf{xy}} \mathbf{p}(\mathbf{y})\rho(\mathbf{x}|\mathbf{y})H(\mathbf{y}, \mathbf{x}) = \sum_{\mathbf{y}} \mathbf{p}(\mathbf{y})H(\mathbf{y}, \hat{\mathbf{x}}^{[1]}(\mathbf{y})). \quad (12)$$

where we have used the fact that all ground state configurations $\hat{\mathbf{x}}(\mathbf{y})$ have the same energy, $H(\mathbf{y}, \hat{\mathbf{x}}^{[\alpha]}) = H(\mathbf{y}, \hat{\mathbf{x}}^{[1]})$, for any α.

The average logarithm Θ of the number of MAP solutions is equal to the zero-temperature entropy

$$\Theta = -\sum_{\mathbf{xy}} \mathbf{p}(\mathbf{y})\rho(\mathbf{x}|\mathbf{y})\ln\rho(\mathbf{x}|\mathbf{y}) = \sum_{\mathbf{y}} \mathbf{p}(\mathbf{y})\ln\mathcal{N}(\mathbf{y}).$$

Let us introduce the the free energy:

$$F(J, h, T) = -T\sum_{\mathbf{y}} \mathbf{p}(\mathbf{y})\ln\sum_{\mathbf{x}} e^{-\beta H(\mathbf{y}, \mathbf{x}; J, h)}, \quad (13)$$

defined with the Ising Hamiltonian (9). The entropy Θ is expressed via the free energy as [see (3, 12)]:

$$\Theta = -\partial_T F|_{T\to 0}. \quad (14)$$

Furthermore, we define the following relevant characteristics of MAP:

$$c = \sum_{\mathbf{y}} \mathbf{p}(\mathbf{y})\rho(\mathbf{x}|\mathbf{y})\frac{1}{N}\sum_{k=1}^{N} x_k x_{k+1} = \frac{1}{N}\partial_J F,$$

$$v = \sum_{\mathbf{y}} \mathbf{p}(\mathbf{y})\rho(\mathbf{x}|\mathbf{y})\frac{1}{N}\sum_{k=1}^{N} y_k x_k = \frac{1}{N}\partial_h F, \quad (15)$$

Here c accounts for the correlations between neighbouring spins in the estimated sequence, while v measures the overlap between the estimated and the observed sequences (the average Hamming distance between the two is simply $1 - v$). In the limiting case of very weak noise, when the magnitude h of the random fields is large [see (8)], we have $v \to v_0 = 1$ (observation-dominance), while c is equal to the corresponding value c_0 of the Markov process $\mathcal{X}$:

$$c = c_0 = \sum_{x_1, x_2} x_1 x_2\, p_{\text{st}}(x_1)p(x_2|x_1) = 1 - 2q. \quad (16)$$

And for very strong noise (the probability of error ϵ is close to $\frac{1}{2}$, which means $h \to 0$), v nullifies, while c goes to the corresponding values calculated over the prior distribution $\mathbf{p}(\mathbf{x})$: $c = \text{sign}(J)$.

4 Recursion relation

Let us return to the partition function (10)

$$Z(\mathbf{y}) = \sum_{x_1=\pm 1 \ldots x_N = \pm 1} e^{\beta J \sum_{k=1}^{N} x_{k+1} x_k + \beta h \sum_{k=1}^{N} y_k x_k}.$$

We apply to $Z(\mathbf{y})$ to the following transformations [1]:

$$\sum_{x_2 \ldots x_N} e^{\beta J \sum_{k=2}^{N} x_{k+1} x_k + \beta h \sum_{k=2}^{N} y_k x_k} \sum_{x_1} e^{\beta J x_1 x_2 + \beta h y_1 x_1}$$

$$= \sum_{x_2 \ldots x_N} e^{\beta J \sum_{k=2}^{N} x_{k+1} x_k + \beta h \sum_{k=3}^{N} y_k x_k + \beta \xi_2 x_2 + \beta B(\xi_1)},$$

where $\xi_2 = hy_2 + A(\xi_1)$, $\xi_1 = hy_1$, and where

$$A(u) = \frac{1}{2\beta} \ln \frac{\cosh[\beta J + \beta u]}{\cosh[\beta J - \beta u]}, \qquad (17)$$

$$B(u) = \frac{1}{2\beta} \ln \left[4 \cosh[\beta J + \beta u] \cosh[\beta J - \beta u] \right]. \quad (18)$$

Thus, once the first spin is excluded, the field acting on the second spin changes from hy_2 to $hy_2 + A(\xi_1)$. Note the zero-temperature ($\beta \to \infty$) limits ($J > 0$)

$$A(u) = u\vartheta(J - |u|) + J\vartheta(u - J) - J\vartheta(-u - J), \quad (19)$$
$$B(u) = J\vartheta(J - |u|) + u\vartheta(u - J) - u\vartheta(-u - J), \quad (20)$$

where $\vartheta(x) = 0$ for $x < 0$ and $\vartheta(x) = 1$ for $x > 0$.

Repeating the above steps we express the partition function as follows:

$$Z(\mathbf{y}) = e^{\beta \sum_{k=1}^{N} B(\xi_k)}, \qquad (21)$$

where ξ_k is obtained from the recursion relation

$$\xi_k = h\,y_k + A(\xi_{k-1}), \quad k = 1, 2, \ldots, N, \quad \xi_0 = 0. \quad (22)$$

This is a random recursion relation, since y_k are random quantities governed by the probability $p(\mathbf{y})$. Depending on the value of y_k, ξ_{k+1} can take values $h + A(\xi_{k-1})$ or $-h + A(\xi_{k-1})$. Even when y_k assumes a finite number of values, ξ_k from (22) can in principle assume an infinite number of values. Fortunately, for $T \to 0$, due to the special form (19, 20) of $A(u)$ and $B(u)$, the number of values assumed by ξ_k is finite (though it can be large). It is checked by inspection that the values taken by ξ_k are parametrized as $\zeta(n_1, n_2) = (n_1 h + n_2 J)$, where n_1 is a positive or negative integer, while n_2 can assume only three values $0, \pm 1$. It can also be seen that the states $\zeta(n_1, 0)$ are not recurrent: once ξ_k takes a value with $n_2 = \pm 1$ (note that there is a finite probability for that), it shall never return to the states $\zeta(n_1, 0)$. In the limit $N \gg 1$, we can completely disregard the states $\zeta(n_1, 0)$.

Now recall that the process $\mathcal{Y}$ with probabilities $p(\mathbf{y})$ is not Markov. To make it Markov we should enlarge it by adding the random variable $\mathbf{z}$; see (6). Here we write the realizations of this auxiliary Markov process $\mathcal{Z}$ as $\mathbf{z}$, so as not to mix them with those of original process $\mathbf{x}$. ($\mathcal{X}$ and $\mathcal{Z}$ have identical statistical characteristics, but these are different processes: $\mathcal{Z}$ is employed merely for making the composite process Markov.) Likewise, we make the process with realizations $[\xi, y]$ Markov by enlarging it to $[\xi, y, z]$. Let us denote this composite Markov process by $\mathcal{C}$. Its conditional probabilities read

$$w(\xi, y, z | \xi', y', z') = p(z|z')\pi(y|z)\varphi(\xi|\xi', y), \qquad (23)$$

where $p(z|z')$ and $\pi(y|z)$ refer to the Markov process $\mathcal{X}$ and the noise, while $\varphi(\xi|\xi', y)$ takes two values 0 and 1, depending on whether the corresponding transition is allowed or not by recursion (22). Now the task is to find all possible values of ξ_k, and then to determine $\varphi(\xi|\xi', y)$. Before turning to this task, we relate the characteristics of the studied MAP estimation to the stationary probabilities $\omega(\xi, y, z)$ of the composite Markov process $\mathcal{C}$. First we get from $\omega(\xi, y, z)$ the stationary probabilities $\omega(\xi)$. Next we return to (21) and to the definition of free energy (13). Since the composite Markov process $\mathcal{C}$ will be seen to be ergodic, the free energy can be written as [1]

$$-f(J, h) \equiv -F(J, h)/N = \sum_{\xi} \omega(\xi) B(\xi), \qquad (24)$$

where the summation is taken over all possible [for a given range of (J, h)] values of ξ. Once $f(J, h)$ is found, we can apply (3.3, 15).

As for entropy (14) we get from (18, 21)

$$F(\mathbf{y}) = -\frac{T}{2} \sum_{k=1}^{N} \sum_{s=\pm} \ln \left[2 \cosh[\beta(\xi_k + sJ)] \right]. \,(25)$$

In this expression we should now select the terms which survive $T \to 0$ *and* ∂_T:

$$-\partial_T F(\mathbf{y})|_{T \to 0} = \frac{\ln 2}{2} \left. \partial_T \left\{ T \sum_{k=1}^{N} \delta(\xi_k \pm J) \right\} \right|_{T \to 0},$$

where $\delta(.)$ is the Kronecker symbol. In the limit $N \gg 1$, $\frac{1}{N} \sum_{k=1}^{N} \delta(\xi_k \pm J)$ should – with probability one, i.e., for the elements of the typical set $\Omega(\mathcal{Y})$ – converge to $\omega(\xi = J) + \omega(\xi = -J)$, provided that the composite Markov process is ergodic. We thus get [1]

$$\theta \equiv \Theta/N = \ln 2 \left[\omega(J) + \omega(-J) \right]/2. \qquad (26)$$

The physical meaning of this formula is that the zero-temperature entropy can be extensive only when the external field ξ acting on the spin has the same energy $\xi x_k = \pm 1$ as the spin–spin coupling constant J; see (7). If this is the case, then a macroscopic amount of spins is frustrated, i.e., the factors influencing those spins compensate each other, so that their sign is not predetermined even at the zero temperature.
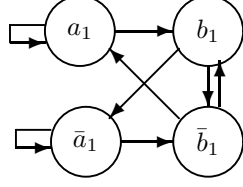
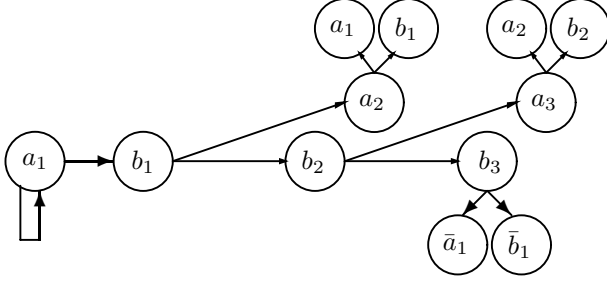Figure 1: The transition graph between various states (29) for $m = 1$; see (27).

Figure 2: Transitions between various states (29) for $m = 3$; see (27). This is one half of the real transition graph. The second half is obtained from the above one by adding bars to all above symbols: $a \to \bar{a}$, $b \to \bar{b}$; see (28, 29).

4.1 Stationary states of the recursion

For given J and h define an integer m as

$$2J/(m-1) > h > 2J/m, \quad m = 1, 2, \ldots \quad (27)$$

Note that the case $h > 2J$ (and there is no upper limit on h) corresponds to $m = 1$. One can check that for each integer m the recurrent states $[\xi, y]$ assumed by the recursion (22) can be parametrized as

$$\{a_i, b_i, \bar{a}_i, \bar{b}_i\}_{i=1}^m, \quad (28)$$
$$a_i = [(2-i)h + J, 1] \equiv [\alpha_i, 1], \quad \bar{a}_i \equiv [-\alpha_i, -1],$$
$$b_i = [-ih + J, -1] \equiv [\beta_i, -1], \quad \bar{b}_i \equiv [-\beta_i, 1]. \quad (29)$$

Note the symmetry $\bar{a}_i = -a_i$ and $\bar{b}_i = -b_i$. The transitions between these states—which via the binary function φ determine the transition matrix in (23)—are illustrated in Figs. 1 and 2 for $m = 1$ and $m = 3$, respectively. The reader can easily generalize the latter graph to an arbitrary m.

We are now prepared to write down from (23, 28, 29) and Fig. 1 the following transition matrix for the composite Markov process $\mathcal{C}$ with $m = 1$

$$W = \left[\begin{array}{cc|cc} w_{a_1|a_1} & 0 & 0 & w_{a_1|\bar{b}_1} \\ w_{b_1|a_1} & 0 & 0 & w_{b_1|\bar{b}_1} \\ \hline 0 & w_{\bar{a}_1|b_1} & w_{\bar{a}_1|\bar{a}_1} & 0 \\ 0 & w_{\bar{b}_1|b_1} & w_{\bar{b}_1|\bar{a}_1} & 0 \end{array} \right]. \quad (30)$$

This is a block matrix composed of 2×2 matrices (hence the actual size of W in (30) is 8×8): 0 means the 2×2 matrix with all its elements equal to 0, and

$$w_{\ldots|a_1} = w_{\ldots|\bar{b}_1} = P, \quad w_{\ldots|b_1} = w_{\ldots|\bar{a}_1} = M, \quad (31)$$
$$P_{xx'} \equiv \pi(+1|x)p(x|x'), \quad M_{xx'} \equiv \pi(-1|x)p(x|x'), \quad (32)$$

where $x, x' = \pm 1$. Note that $P + M$ is equal to the transition matrix of the Markov process $\mathcal{X}$. Once the 8×1 stationary probability vector w of W is found from $Ww = w$, we get $\omega(\alpha_1) = w_1 + w_2$, $\omega(\beta_1) = w_3 + w_4$, $\omega(\bar{\alpha}_1) = w_5 + w_6$, and $\omega(\bar{\beta}_1) = w_7 + w_8$.

For a general m the following $m \times m$ matrices serve as building blocks for the matrix W

$$L = \{L_{i,i+1} = 1, i = 1, \ldots, m-1\}, \quad E = \{E_{1,1} = 1\},$$
$$U = \{U_{i+1,i} = 1, i = 1, \ldots, m-1\}, \quad S = \{S_{1,m} = 1\},$$

where all not indicated elements are equal to zero. The transition matrix for a general m reads [see Figs. 1, 2]

$$W = \left[\begin{array}{cc} M & 0 \\ P & 0 \end{array} \right] \otimes \left[\begin{array}{cc} 0 & 0 \\ 0 & E \end{array} \right] \quad (33)$$
$$+ \left[\begin{array}{cc} 0 & P \\ 0 & M \end{array} \right] \otimes \left[\begin{array}{cc} L & 0 \\ 0 & 0 \end{array} \right] + \left[\begin{array}{cc} 0 & M \\ 0 & P \end{array} \right] \otimes \left[\begin{array}{cc} 0 & 0 \\ S & L \end{array} \right]$$
$$+ \left[\begin{array}{cc} P & 0 \\ M & 0 \end{array} \right] \otimes \left[\begin{array}{cc} U & 0 \\ 0 & 0 \end{array} \right] + \left[\begin{array}{cc} M & 0 \\ P & 0 \end{array} \right] \otimes \left[\begin{array}{cc} 0 & 0 \\ 0 & U \end{array} \right]$$
$$+ \left[\begin{array}{cc} P & 0 \\ M & 0 \end{array} \right] \otimes \left[\begin{array}{cc} E & 0 \\ 0 & 0 \end{array} \right] + \left[\begin{array}{cc} 0 & P \\ 0 & M \end{array} \right] \otimes \left[\begin{array}{cc} 0 & S \\ 0 & 0 \end{array} \right].$$

The left matrices of each tensor product is a block matrix; each block consists of one 2×2 matrix. The right matrices of each tensor product are also block matrices; now each block consists of one $m \times m$ matrix. The zero $m \times m$ matrix is written as 0. The overall size of W is $8m \times 8m$, since each state in (28) is augmented by two realizations of the hidden Markov process.

Note that going from one value of m to another amounts to changing the dimension of the matrices E, U, S and L. Since these matrices are sparse, efficient numerical algorithms of treating them are available, even for larger values of m.

5 MAP inference

Let us indicate how the quantities of interest are expressed via the stationary probability ω of the Markov process $\mathcal{C}$ (obtained from (33)). Recall that since the estimated process is unbiased, we are interested in the second moment c, overlap v and entropy θ. The former two quantities have to be obtained via the free energy. To this end, we trace out the redundant variables in the stationary probability of W to obtain the following probabilities $(i = 1, \ldots, m)$:

$$\omega_m(\alpha_i) = \omega_m(\bar{\alpha}_i), \quad \omega_m(\beta_k) = \omega_m(\bar{\beta}_k), \quad (34)$$

where the equalities in (34) are due to the symmetry of the unbiased situation. We add a lower index to relevant quantities (e.g., to ω's) to indicate the specific value of m. Recall that, e.g., $\omega_1(\alpha_1)$ and $\omega_2(\alpha_1)$ are in general different quantities, since they belong to different Markov processes $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

Due to (34), we shall need only the probabilities $\omega_m(\alpha_k)$ and $\omega_m(\beta_k)$ that normalize to one-half:

$$\sum_{k=1}^{m} \left[\, \omega_m(\alpha_k) + \omega_m(\beta_k)\,\right] = 1/2. \qquad (35)$$

The free energy then reads (see (20, 24))

$$-\frac{f_m}{2} = \sum_{k=1}^{m} \left[\,\omega_m(\alpha_k)B(\alpha_k) + \omega_m(\beta_k)B(\beta_k)\,\right]$$
$$= h[\,\omega_m(\alpha_1) + m\omega_m(\beta_m)\,] + J[\,\frac{1}{2} - 2\omega_m(\beta_m)\,]. \quad (36)$$

Now we make use of the fact that free energy is a continuous function of its parameters [1], which in our case implies

$$f_m = f_{m+1} \quad \text{at} \quad h = 2J/m, \quad m = 1, 2, \dots \quad (37)$$

This leads from (36) to

$$\omega_m(\alpha_1) = \omega_{m+1}(\alpha_1) + \omega_{m+1}(\beta_m). \qquad (38)$$

One can confirm (38) from (43, 44, 45). Note that (38) will hold for all values of ϵ and q, since it does not depend on h and/or J (the formalism holds without requiring any specific relation between h, J and ϵ, q). Combining (36) with (15) from Section 3.3, we obtain for the second moment c_m of the estimated sequence and the overlap v_m

$$c_m = 1 - 4\omega_m(\beta_m), \quad v_m = 2\omega_m(\alpha_1) + 2m\omega_m(\beta_m). \ (39)$$

As seen from (28, 29), if the relations (27) hold [recall that they are strict inequalities], there are only two realizations $\alpha_2 = J$ and $\bar{\alpha}_2 = -J$, which, according to (26), contribute into the entropy. Recalling also (34), we get ($m = 1, 2, \dots$)

$$\theta = \omega_m(\alpha_2) \ln 2, \quad \text{for} \quad \frac{2J}{m-1} > h > \frac{2J}{m}. \qquad (40)$$

This relation holds for $m = 1$, if we assume $\omega_1(\alpha_2) = 0$.

At the transition points $h = \frac{2J}{m}$ between the various regimes (27), there are more states that contribute into the entropy. The reader can verify that

$$\theta = [\omega_m(\alpha_2) + \omega_m(\beta_m)] \ln 2, \quad \text{if} \quad h = 2J/m. \quad (41)$$

[1]Outside phase transitions free energy is smooth, while at the phase-transition points it has to be at least continuous, since, besides being the generating function for calculating various averages, free energy is also a measure of dynamic stability, and at the phase-transition points both phases are equally stable by definition (see [7] for more details).

This equation is written down assuming that the value of θ at $h = \frac{2J}{m}$ does not depend on whether the latter point is reached as $h \to \frac{2J}{m} + 0$ or as $h \to \frac{2J}{m} - 0$. This assumptions leads from (41) to:

$$\omega_m(\alpha_2) + \omega_m(\beta_m) = \omega_{m+1}(\alpha_2) + \omega_{m+1}(\beta_m). \quad (42)$$

This relation has the same origin as the continuity of the free energy.

5.1 The regime $m = 1$ or $h > 2J$.

We deduce for the stationary probabilities from (33)

$$\omega_1(\alpha_1) \equiv \omega_1(h + J) = \frac{1 - q}{2} + \epsilon(1 - \epsilon)(2q - 1), \quad (43)$$
$$\omega_1(\beta_1) \equiv \omega_1(-h + J) = \frac{q}{2} - \epsilon(1 - \epsilon)(2q - 1). \quad (44)$$

This implies from (39) $v_1 = 1$, $c_1 = (1 - 2q)(1 - 2\epsilon)^2$. This is in fact the Maximum Likelihood (ML) regime: the noise is so small (or h is so large) that the estimation is completely governed by the observations: $v_1 = 1$. The second moment c_1 of the estimated sequence in this regime is given by the original value $c_0 = 1 - 2q$ (see (16)) times the squared error probability $(1 - 2\epsilon)^2$. The entropy in this regime is zero (see (40)): $\theta_1 = 0$. In this sense the estimation is uniquely determined by observations. We stress that the ML and MAP schemes agree with each other not only for very small, but also for finite noises.

5.2 The regimes $m = 2$ and $m = 3$.

For more compact presentation of the probabilities, let us introduce separate notations for the noise strength and the Markov correlator $g = \epsilon(1 - \epsilon)$, $u = 1 - 2q$, where $0 < g < \frac{1}{4}$ and $0 < u < 1$; see (16). The probabilities obtained from (33) are written as

$$\omega_2(\alpha_1) = \frac{\frac{1}{2} + \left(\frac{1}{2} - 3g\right)u}{3 - (1 + 2g)u}, \quad \omega_2(\beta_1) = \frac{\frac{1}{2} + \left(g - \frac{1}{2}\right)u}{3 - (1 + 2g)u},$$
$$\omega_2(\alpha_2) = \frac{\frac{1}{4} + \frac{1}{2}(3g - 1)u + \frac{1}{4}(1 - 2g(4g + 1))u^2}{3 - (1 + 2g)u},$$
$$\omega_2(\beta_2) = \frac{\frac{1}{4} - \frac{gu}{2} + \frac{1}{4}\left(8g^2 + 2g - 1\right)u^2}{3 - (1 + 2g)u}. \qquad (45)$$

We skip a tedious analytical expressions for ω_3.

The values of c and v deduced from (39, 45) are shown in Figs. 3(a) and 3(b). We compare those values with the results obtained by actually finding the MAP estimate via the Viterbi algorithm, and calculating those quantities directly. It is seen that at the regime change points $h = 2J$ and $h = J$, v and c experience sudden jumps, or first-order phase transitions. Remarkably, those features are perfectly reproduced in the simulations, as shown in Figure 3(a)

m	4	5	6	7
ϵ	0.3700	0.3910	0.4100	0.421
$\theta/\ln 2$	0.07308	0.06587	0.05925	0.05349

Table 1: Regular values of entropy $\frac{\theta}{\ln 2}$ for $q = 0.24$; see (40). This table continues Fig. 3(c) towards larger values of the noise strength ϵ.

| h	$2J$	J	$2J/3$	$J/2$	$2J/5$
ϵ	0.0907	0.2400	0.3598	0.3867	0.4051
$\theta/\ln 2$	0.1629	0.1462	0.1220	0.0992	0.0831

Table 2: The special values of entropy $\frac{\theta}{\ln 2}$ for $q = 0.24$.

and 3(b). For instance, in the ML regime $h > 2J$ ($0 < \epsilon < 0.09068$), the overlap $v = 1$ indicating that the estimation is governed solely by observations. At $h = 2J$ it jumps sharply, and then monotonically decreases in the regime $2J > h > J$. More generally, v decays, both monotonously and via jumps, towards the prior-dominated value $v = 0$.

Consider the second moment c of the estimated sequence shown in Figure 3(b). We see that c is nearly a constant for each given $m \geq 2$. This is the main virtue of MAP scheme as compared to the ML scheme: While the latter predicts a c that quickly decays with the noise as $c_{ML} = (1-2q)(1-2\epsilon)^2$ (the dotted line in the plot), the proper MAP value of c is not far from its noise-free value $c_0 = 1 - 2q$, and is nearly a constant for a finite range of noise strength ϵ. This advantage of MAP over ML is due to supporting the estimation process by the priors. Indeed, the values of the overlap indicate that the estimated sequence is not completely driven by the observations, though it is still not very far from them. Upon increasing ϵ towards its maximal value $\epsilon = \frac{1}{2}$, c experiences jumps during each regime change. For larger m these jumps are smaller and more frequent, leading c to its prior-dominated value 1.

Now let us focus on the entropy: It naturally nullifies in the ML regime ($\theta_1 = 0$), while in the regime $m = 2$ the entropy θ_2 is monotonously increasing with ϵ, as shown in Figure 3(c). At $h = 2J$ (the phase transition point, where m changes from 1 to 2) θ_2 experiences a jump, which is again usual for first-order phase-transitions. θ maximizes at an intermediate value of ϵ, and then decays to zero for $\epsilon \to \frac{1}{2}$, see Table 1; at this point the present approach reduces to a ferromagnetic 1d Ising model without magnetic fields. This model has a trivial ground-state structure and hence zero entropy. We also note that right at the transition points $h = \frac{2J}{m}$ the values of θ is different; see Table 2. The largest value is attained for $h = 2J$.
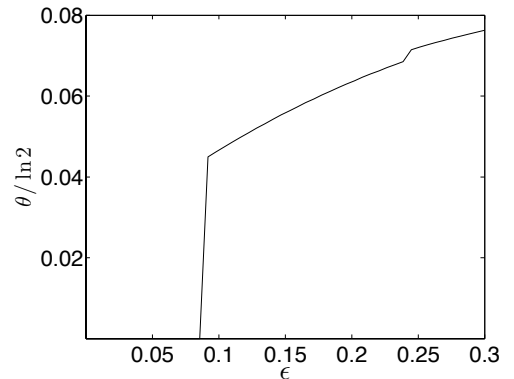
Finally, we would like to note that the second moment
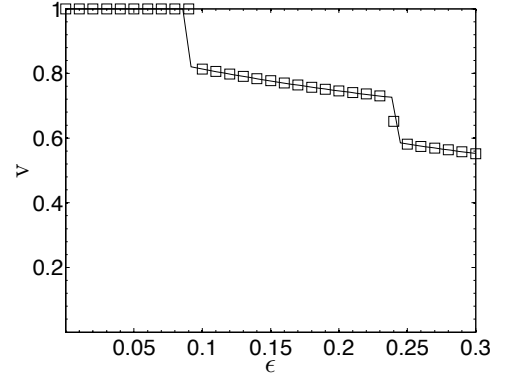
(a)
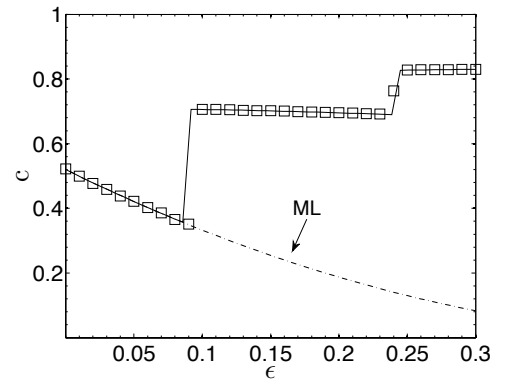
(b)

(c)

Figure 3: MAP characteristics versus the noise intensity in the regimes $m = 1, 2, 3$ for $q = 0.24$: (a) Overlap v; (b) the second moment c; (c) Entropy $\frac{\theta}{\ln 2}$. In (a) and (b) the open squares represent simulation results, obtained by running the Viterbi algorithm and calculating the respective quantities directly. We used sequences of size 10^4, and averaged the results over 100 random trials.

of the estimated sequence, c, is an indirect measure of accuracy. In practice, one is restricted to use such indirect measures as information about the true sequence

might not available. In Figure 4 we present the average error rate for the MAP estimation, which is given by the normalized Hamming distance between the true and Viterbi–decoded sequences, plotted against the noise intensity. Also shown is the average error rate of ML estimation, which is simply ϵ. For vanishing noise, both MAP and ML yield the same average error. Upon increasing the noise intensity, the MAP estimation error behaves differently depending on the parameter q: For small values of q, MAP is always superior to ML for a wide range of noise intensities. For larger values of q, however, the situation is more complicated: Although both methods perform similarly, there are some differences and crossovers between the two at intermediate noise intensities, as shown in Figure 4.
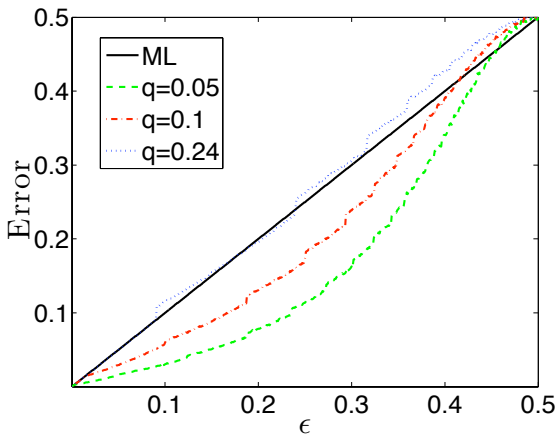

Figure 4: The average error rate given by the normalized Hamming distance between the true and the estimated sequences.

6   Discussion

We theoretically examined Maximum a Posteriori (MAP) estimation for hidden Markov sequences, and found that MAP yields a non–trivial solution structure even for the simple binary and unbiased hidden Markov process considered here. We demonstrated that for a finite range of noise intensities, there is no difference between MAP and Maximum Likelihood (ML) estimations, as the solution is observation–dominated. While it was expected that the two methods agree for a vanishing noise, *the fact of their exact agreement for a finite range of the noise is non-trivial.* Furthermore, upon increasing the noise intensity the MAP solution switches between different operational regimes that are separated by first–order phase transitions. In particular, a first-order phase-transition separates the regime where MAP and ML agree exactly. At this transition point the influence of the prior information becomes comparable to the influence of

observations. In the vicinity of the first-order phase-transitions the performance of MAP (e.g., characteristics of the estimated sequence) changes abruptly. This means that a small change in the noise intensity may lead to a large change in the performance. In other words, the phase-transition points should be avoided in applications.

For practical applications of HMM (e.g., in speech recognition, or machine translation) it is not enough to know the single solution that provides the largest posteriori probability [5]. At the very least, one should also know how many sequences have a posteriori probabilities sufficiently close to the optimal one. Motivated by this fact, we studied the number $\mathcal{N}$ of MAP-solutions that have (for long sequences $N \to \infty$) almost equal logarithms of the posterior probability. A finite $\theta = \frac{1}{N}\ln\mathcal{N}$ means that there is an *exponential* number of solutions with posterior probabilities slightly less than the optimal. We found that θ is finite whenever MAP differs from ML. We believe that this theoretical result might have practical implications as well. For instance, in applications such as statistical machine translation, one usually considers top K solutions to the inference problem, and then chooses one according to some heuristics. Our result suggests that one needs to be careful with this practice whenever θ is non–zero, as one might discard a large number of nearly optimal solutions if K is not chosen sufficiently large.

We also note that our work is directly related to the notion of *trackability*, which can be intuitively defined as one's ability to (accurately) track certain stochastic processes [3, 10]. In fact, a similar phase–transition in the number of solutions was reported by Crespi et. al. [3] for so called *weak models*, where the entries in the HMP transition and emission matrices are either 0 or 1. For more general stochastic processes, an information-theoretical characterization of trackability was suggested in [10]. Within this approach, the accuracy is characterized by the probability $\Pr[\hat{\mathbf{x}} \neq \mathbf{x}]$ of the estimated sequence $\hat{\mathbf{x}}$ not being equal to the actual one, while the structure of the solution space is described via the number of elements $|\Omega|$ in the (conditional) typical set Ω of $\mathbf{x}$ sequences given an observed sequence $\mathbf{y}$ (complexity). Both these quantities relate to the conditional entropy $-\sum_{\mathbf{x},\mathbf{y}}\Pr(\mathbf{x},\mathbf{y})\ln\Pr(\mathbf{x}|\mathbf{y})$. We note that whereas the accuracy and the complexity measures of [10] deteriorate even for a small (but generic) noise intensity, our approach of defining trackability in terms of the zero–temperature entropy of the Ising Hamiltonian (Equation 9) suggests that a process can be trackable in the MAP sense even in the presence of moderate noise.

Finally, we would like to note that another interest-

ing feature of the MAP estimation is that its characteristics (c and v) change only slightly in between the phase-transition points. In contrast to ML estimation, which deteriorates (at least linearly) when increasing the noise, MAP estimation is stable for a finite range of noise intensities. Thus, although MAP estimation may be less accurate compared to ML, it can be still useful as far as its stability is concerned, provided that its range of application is selected carefully.

There are several directions for further developments. First, we intend to obtain analytical results for the average error rate to complement the empirical analysis presented in Figure 4. Furthermore, one can think of a semi–supervised MAP estimation, where one has (possibly noisy) knowledge about the states of the hidden process at particular times. Remarkably, the framework presented here allows a natural generalization to this case. Indeed, one simply needs to modify the Ising energy function by adding quenched fields at the corresponding locations in the chain. Finally, it will be interesting to generalize the analysis presented here beyond the binary hidden Markov processes considered here. In this case, the MAP optimization problem can be mapped to a Potts model. We would like to note that the behavior observed in the simple binary model can be explained by the emergence of a finite fraction of "frustrated" spins, where the frustration can be attributed to two competing tendencies – accommodating observations on one hand, and the hidden (Markovian) dynamical model on the other. Since this mechanism is rather general, we believe that most features of the MAP scheme uncovered here via an exact analysis of the simplest binary model will survive in more general situations.

References

[1] U. Behn and V.A. Zagrebnov, *One-dimensional Markovian-field Ising model: physical properties and characteristics of discrete stochastic mapping*, J. Phys. A **21**, 2151 (1988).

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (Wiley, New York, 1991).

[3] V. Crespi, G. Cybenko, G., and G. Jiang, *The theory of trackability with applications to sensor networks*, ACM Trans. Sen. Netw. **4**, 3 (2008).

[4] Y. Ephraim and N. Merhav, *Hidden Markov processes*, IEEE Trans. Inf. Th., **48**, 1518-1569, (2002).

[5] L. A. Foreman, *Generalization of the Viterbi algorithm*, IMA J. Math. Appl. Bus. Ind., **4**, 351 (1993).

[6] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images*, IEEE Trans. Pattern Analysis Machine Intell. **6**, 721 (1984).

[7] L.D. Landau and E.M. Lifshitz, *Statistical Physics, I*, (Pergamon Press, Oxford, 1978).

[8] J.M. Pryce and A.D. Bryce, *Statistical mechanics of image restoration*, J. Phys. A **28**, 511 (1995).

[9] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE, **77**, 257-286, (1989).

[10] Y. Sheng, G. Cybenko, V. Crespi, and G. Jiang, *Trackability analysis of multiple processes using multi-distributed agents*, International Conference on Integration of Knowledge Intensive Multi-Agent Systems, (2005).

[11] O. Zuk, I. Kanter and E. Domany, *The Entropy of a Binary Hidden Markov Process*, J. Stat. Phys. **121**, 343 (2005).