Journal of Statistical Physics, Vol. 115, Nos. 3/4, May 2004 (© 2004)

Multi-Information in the Thermodynamic Limit

Ionas Erb[1,2] **and Nihat Ay**[2–4]

Received July 3, 2003; accepted November 5, 2003

A multivariate generalization of mutual information, multi-information, is defined in the thermodynamic limit. The definition takes phase coexistence into account by taking the infimum over the translation-invariant Gibbs measures of an interaction potential. It is shown that this infimum is attained in a pure state. An explicit formula can be found for the Ising square lattice, where the quantity is proved to be maximized at the phase-transition point. By this, phase coexistence is linked to high model complexity in a rigorous way.

KEY WORDS: Mutual information; Ising model; phase transitions; excess entropy; complexity.

1. INTRODUCTION

1.1. Why Multi-Information?

Shannon's mutual information compares the summed entropies H of two distributions $p_{\{1\}}, p_{\{2\}}$ with the entropy of their joint distribution $p_{\{1,2\}}$:

$$I(p_{\{1,2\}}) = H(p_{\{1\}}) + H(p_{\{2\}}) - H(p_{\{1,2\}}). \tag{1}$$

There are several generalizations for finite sets Λ with more than two subsystems. Keeping the two-point property of I, one can let the quantity depend on a distance between the elements of Λ.[29] There are also multivariate generalizations. Co-information is an alternating sum of entropies

[1] Bioinformatik, Institut für Informatik, University of Leipzig, Kreuzstr. 7b, 04103 Leipzig, Germany; e-mail: ierb@bioinf.uni-leipzig.de

[2] Max-Planck Institute for Mathematics, Inselstr. 22-26, 04103 Leipzig, Germany.

[3] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501; e-mail: nay@santafe.edu

[4] Mathematical Institute, Friedrich-Alexander University Erlangen-Nuremberg, Bismarckstr. 1 1/2, 91054 Erlangen, Germany.

0022-4715/04/0500-0949/0

of the marginals p_V of a distribution p_Λ for all subsystems $V \subset \Lambda$.[8] A simple multivariate generalization that is valid in any dimension is called multi-information:[41]

$$I(p_\Lambda) = \sum_{i \in \Lambda} H(p_{\{i\}}) - H(p_\Lambda) \tag{2}$$

Below we will give a motivation for this quantity coming from information geometry. Also, in Section 4 we show the relationship to excess entropy. Let us already mention that in the limit of infinite shift-invariant systems, multi-information will again be mutual information, namely between an elementary subsystem and the infinite system.

Information-theoretic measures as the ones cited above quantify stochastic interdependence in probability distributions. They are used in a variety of fields, e.g., communication theory,[39] multivariate statistics,[26] neural networks,[8] complexity measures,[16, 42] learning rules,[5, 31] to mention only a few of them.

The behaviour of a quantity like mutual information is best shown by a simple example: two units x_1, x_2 which can take values from $\{0, 1\}$. Knowing the probabilities $p_{\{1,2\}}(x_1, x_2)$ of the four configurations, mutual information is given by (1). Let us introduce an additional parameter β by which we can tune $p_{\{1,2\}}$. We define

$$p_\beta(x_1, x_2) := \frac{(p_{\{1,2\}}(x_1, x_2))^\beta}{\sum_{x'_1, x'_2 \in \{0,1\}} (p_{\{1,2\}}(x'_1, x'_2))^\beta}. \tag{3}$$

The denominator normalizes p_β. For $\beta = 0$, p_β is the equidistribution, whereas $\beta \to \infty$ gives us the Dirac measure. For a generic choice of $p_{\{1,2\}}$, the function $\beta \mapsto I(p_\beta)$, let us call it $I(\beta)$, is shown in Fig. 1. The trajectory of the curve $\beta \mapsto p_\beta$ within the simplex of all probability measures for the four configurations is shown in Fig. 2. It ranges from the barycentre to one of the corners of the simplex. The Kullback–Leibler distance (see below)

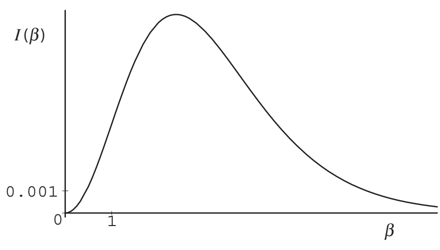

Fig. 1. Plot of $I(\beta)$ for fixed values of $p_{\{1,2\}}(x_1, x_2)$. $I(\beta)$ vanishes for $b = 0$ and $\beta \to \infty$, i.e., there are no stochastic dependencies for complete randomness or complete predictability.

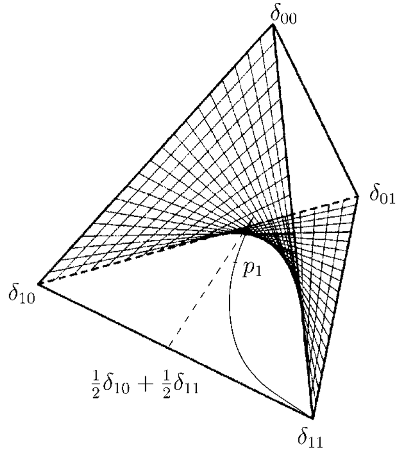

Fig. 2. The set of probability distributions for the four configurations (x_1, x_2) with the plane of factorizable distributions and the curve p_β.

of this curve from the surface of independent distributions is mutual information.

Multi-information generalizes exactly this property: It still is the Kullback–Leibler distance of p_Λ from its factorized distribution $\bigotimes_{i \in \Lambda} p_{\{i\}}$.[1,4] Interest in this quantity is motivated by finite-volume information geometry, see ref. 1 and the references therein. Entropy and multi-information have natural decompositions of a form that we will briefly describe now. Let us use a simple example for p_Λ:

$$p_\Lambda(x_\Lambda) = e^{F + \sum_{\varnothing \neq V \subset \Lambda} \Theta_V \prod_{i \in V} x_i}, \tag{4}$$

with x_Λ the collection of x_i, $i \in \Lambda$, where the x_i are from $\{0, 1\}$. F is a normalization constant (the free energy). The coefficients $\Theta_V \in \mathbb{R}$, $V \subset \Lambda$ represent the strength of direct interaction between the units i. Let us now denote by $p_\Lambda^{(k)}$ a distribution that has the same marginals as p_Λ up to kth order, but no intrinsic interactions of higher order ($\Theta_V = 0$ for all V such that $|V| > k$). It is the maximum-entropy estimate[28] of p_Λ given its kth order marginals. Now there is an "extended Pythagoras theorem"

$$D(p_\Lambda \,\|\, p_\Lambda^{(0)}) := \sum_{x_\Lambda} p_\Lambda(x_\Lambda) \ln \frac{p_\Lambda(x_\Lambda)}{p_\Lambda^{(0)}(x_\Lambda)} = \sum_{k=1}^{|\Lambda|} D(p_\Lambda^{(k)} \,\|\, p_\Lambda^{(k-1)}), \tag{5}$$

where D denotes the Kullback–Leibler distance.[14] In our example, the left-hand side (LHS) is $|\Lambda| \ln 2 - H(p_\Lambda)$, the "distance" from the equidistribution, whereas multi-information is given by $D(p_\Lambda \parallel p_\Lambda^{(1)})$. It can be decomposed into a sum like the right-hand side (RHS) without the $k=1$ term. Note that D, although not a metric, is the canonical[2] measure of distance in information geometry and that the above decomposition is non-trivial for $|\Lambda| > 2$. Note also that, e.g., covariance makes only use of correlations up to second order, whereas multi-information contains also information from all the higher-order marginals.

1.2. Statistical Mechanics

Looking at Figs. 1 and 2, the question about the maximum of I springs to mind. To give us an idea what such a maximization[4, 6] can mean, we want to define multi-information in the context of statistical mechanics. There we have a mathematical formalism that describes models of different structural richness. A simple example for a finite-volume state (Gibbs measure) is given by (4). The set of all its terms $\Theta_V \prod_{i \in V} x_i$ in the exponent is called an interaction potential. A parametrized family of such potentials is called a model. The parameters (cf. the β in (3)) can be inverse temperature, magnetic field etc.

There are (infinite-volume) potentials whose infinite-volume state is not uniquely determined. Models showing this phenomenon for certain parameter values are said to exhibit phase coexistence. According to clear hints in the literature, measures of stochastic interdependence are maximized at critical parameter values,[3, 17, 34] or at phase transitions in a less strict sense.[15, 24, 33] One can go a step further and look at the structural phenomena occurring at the phase coexistence point in standard models like the Ising square lattice: infinite-cluster formation, divergence of the correlation length. They can be seen as signs of "complex" behaviour. From this perspective it seems natural to assume that large stochastic interdependence is associated with high structural complexity. In ref. 20 this is discussed in connection with excess entropy, a quantity that, as we will show, is closely related to multi-information.

Phase transitions thus seem to mark the "border of maximum complex behaviour" between complete randomness and absolute predictability. It is one of the objectives of the present work to give an example where this kind of statement can be made rigorous. To do so, we generalize multi-information to a quantity in the thermodynamic limit that takes into account the non-uniqueness of infinite-volume Gibbs measures for certain interaction potentials. Using the example of the Ising square lattice, we can connect phase coexistence with maximum distance from factorizability in a formal way.

2. MULTI-INFORMATION IN STATISTICAL MECHANICS

2.1. Notation

Our systems take discrete values on the points of an infinite lattice. Let S be a finite set (the spin space), and let $x\colon \mathbb{Z}^d \to S$, $i \mapsto x_i$ be configurations on the d-dimensional lattice of integers $\mathbb{Z}$. To make it a measurable space, the space of configurations $\Omega := S^{(\mathbb{Z}^d)}$ is equipped with the product sigma algebra $\mathscr{F}$, which contains the cylinder sets $\{x \in \Omega : X_\Lambda(x) = x_\Lambda\}$, where $x_\Lambda := (x_i)_{i \in \Lambda}$ is a configuration on the finite[5] set $\Lambda \subset\subset \mathbb{Z}^d$ and $X_\Lambda\colon \Omega \to S^\Lambda$, $x \mapsto x_\Lambda$ the natural projection onto a finite configuration. Thus the projection X_Λ yields finite measurable spaces $(\Omega_\Lambda, \mathscr{F}_\Lambda)$, where $\Omega_\Lambda := S^\Lambda$ denotes the set of x_Λ and $\mathscr{F}_\Lambda$ its power set.

We will first define multi-information on these finite spaces, for p_Λ a probability measure on $(\Omega_\Lambda, \mathscr{F}_\Lambda)$. At this point, the form of the measure is of no importance.

Definition 2.1. Let p_Λ be a probability measure on $(\Omega_\Lambda, \mathscr{F}_\Lambda)$, where Λ is a finite set. The *multi-information* of p_Λ is defined by

$$I(p_\Lambda) := \sum_{i \in \Lambda} H(p_{\{i\}}) - H(p_\Lambda). \tag{6}$$

Here, $H(p_\Lambda) := -\sum_{x_\Lambda \in \Omega_\Lambda} p_\Lambda(x_\Lambda) \ln p_\Lambda(x_\Lambda)$ denotes the Shannon entropy and $p_{\{i\}}(x_i) := \sum_{x_{\Lambda\setminus\{i\}}} p(x_{\Lambda\setminus\{i\}}, x_i)$ are the marginal distributions of the elementary subsystems in Λ.

2.2. Thermodynamic Limit

To define multi-information for distributions on the infinite measurable space $(\Omega, \mathscr{F})$, our starting point are measures p_Λ on finite spaces $(\Omega_\Lambda, \mathscr{F}_\Lambda)$, $\Lambda \subset\subset \mathbb{Z}^d$. These we consider as being obtained from a translation invariant measure p on $(\Omega, \mathscr{F})$ by defining its marginal distributions $p_\Lambda(x_\Lambda) := p(X_\Lambda = x_\Lambda)$. Translation invariance of p is defined by

$$p(\{(x_{i+j})_{j \in \mathbb{Z}^d} \mid (x_j)_{j \in \mathbb{Z}^d} \in A\}) = p(A) \qquad \forall A \in \mathscr{F}, \quad \forall i \in \mathbb{Z}^d. \tag{7}$$

Existence and properties of the van Hove limit[38] of multi-information follow in straightforward fashion from well-known results for entropy (see the appendix for a proof). Notice that the set of translation invariant measures is a simplex and thus convex.[22, 40]

[5] We denote finiteness of subsets by $\subset\subset$.

Theorem and Definition 2.2. Let p be a translation invariant probability measure on $(\Omega, \mathcal{F})$. Then the van Hove limit $\lim_{\Lambda \nearrow \mathbb{Z}^d} \frac{1}{|\Lambda|} I(p_\Lambda) =: I(p)$ exists and $I(p) \in [0, \ln |S|]$. The function $p \mapsto I(p)$ is concave and lower-semicontinuous (w.r.t. the weak* topology).

The quantity $I(p)$ depends on the *state* of a system. In statistical mechanics, however, models are defined via the *interaction* between their constituents (spins, particles). In the following, we want to obtain a definition which directly depends on the interaction potential.

2.3. Phase Coexistence

The construction of measures in infinite volume[18, 32] can yield non-uniqueness for a given interaction, so the description of phase coexistence becomes possible. For an interaction-dependent definition of multi-information we have to choose from a set of possible measures now. To introduce the necessary notation and to make our point clear, we give a brief description of the standard construction of infinite-volume Gibbs measures. All the results stated in this section can be found in this or a similar form in refs. 22 and 40, for short descriptions of the subject see also refs. 25 and 37.

From finite-volume statistical mechanics one knows the form of the conditional probabilities for a finite configuration given an exterior configuration that the measure p on $(\Omega, \mathcal{F})$ should have.[25] Specifying these, one obtains a condition that possible infinite-volume Gibbs measures should fulfill. For this, we need to define interaction potentials in infinite volume.

Definition 2.3. A *potential* Φ on $\mathbb{Z}^d$ is a family of functions $\{\Phi_V\}_{V \subset\subset \mathbb{Z}^d}$ from Ω to $\mathbb{R}$ with

(i) Φ_V is X_V-measurable for all $V \subset\subset \mathbb{Z}^d$.

(ii) The series $E^\Phi_\Lambda(x_\Lambda, y_{\Lambda^c}) := \sum_{V \subset\subset \mathbb{Z}^d : V \cap \Lambda \neq \varnothing} \Phi_V(x_\Lambda, y_{\Lambda^c})$ converges for all $\Lambda \subset\subset \mathbb{Z}^d$ and for all $(x_\Lambda, y_{\Lambda^c}) := x \in \Omega$ (where Λ^c denotes the complement of Λ in $\mathbb{Z}^d$).

$E^\Phi_\Lambda(x_\Lambda, y_{\Lambda^c})$ is the *energy* of x_Λ with boundary condition y_{Λ^c}.

This definition enables us to specify the Gibbsian conditional probabilities for the desired measures. Since we know nothing about the existence of these measures, we can only fix probability kernels (i.e., loosely speaking, conditional probabilities "waiting for a measure"). Let $\Omega_{\Lambda^c} = S^{\mathbb{Z}^d \setminus \Lambda}$.

Using a definition from ref. 22, a *specification* for finite S is given by a family $\{k_\Lambda^\Phi\}_{\Lambda \subset\subset \mathbb{Z}^d}$ of probability kernels from $(\Omega_{\Lambda^c}, \mathscr{F}_{\Lambda^c})$ to $(\Omega, \mathscr{F})$ where

$$A \mapsto k_\Lambda^\Phi(A \mid y_{\Lambda^c}) := \sum_{x_\Lambda : (x_\Lambda, y_{\Lambda^c}) \in \mathscr{A}} \frac{e^{-E_\Lambda^\Phi(x_\Lambda, y_{\Lambda^c})}}{\sum_{x'_\Lambda \in \Omega_\Lambda} e^{-E_\Lambda^\Phi(x'_\Lambda, y_{\Lambda^c})}}. \tag{8}$$

Here, Φ is a potential, $\Lambda \subset\subset \mathbb{Z}^d$, $A \in \mathscr{F}$, and $y_{\Lambda^c} \in \Omega_{\Lambda^c}$. Such specifications fulfill consistency conditions analogous to those of conditional probabilities.

The set of DLR measures is now defined as the solution set of $p(A \mid \mathscr{F}_{\Lambda^c}) = k_\Lambda^\Phi(A \mid \cdot)$ p-a.s. for all finite volumes Λ and events A. Here, $p(A \mid \mathscr{F}_{\Lambda^c})$ is the conditional expectation of 1_A, i.e., of the indicator function for an event A, given the sigma algebra of events outside Λ. For the definition of conditional expectations given a sub-sigma algebra, see, e.g., ref. 7. The properties of the set of DLR measures are well known.

Proposition and Definition 2.4. Given a potential Φ, the set of infinite-volume Gibbs states (DLR measures) is defined by

$$\mathscr{G}(\Phi) := \{p \text{ on } (\Omega, \mathscr{F}) : p(A \mid \mathscr{F}_{\Lambda^c}) = k_\Lambda^\Phi(A \mid \cdot) \ p\text{-a.s. } \forall A \in \mathscr{F}, \Lambda \subset\subset \mathbb{Z}^d\}.$$

$\mathscr{G}(\Phi)$ is a compact, convex set (more precisely: a simplex). Depending on the potential Φ, there are the following possibilities for its cardinality:

$$|\mathscr{G}(\Phi)| = 0, \tag{9}$$

$$|\mathscr{G}(\Phi)| = 1, \tag{10}$$

$$|\mathscr{G}(\Phi)| = \infty. \tag{11}$$

The set of Gibbs measures is always non-empty if the potential is translation invariant, i.e., if it fulfills

$$\Phi_{V+i}((x_{j-i})_{j \in \mathbb{Z}^d}) = \Phi_V(x) \qquad \forall x \in \Omega, \quad \forall V \subset\subset \mathbb{Z}^d, \quad \forall i \in \mathbb{Z}^d. \tag{12}$$

Notice that even translation-invariant potentials need not have only translation-invariant states. Since the thermodynamic limit of multi-information was obtained for translation-invariant states, we will actually need the set of translation-invariant Gibbs measures $\mathscr{G}_I(\Phi)$, i.e., the intersection of all translation-invariant measures on $(\Omega, \mathscr{F})$ with the set of Gibbs measures of translation invariant potentials. The set $\mathcal{G}_I(\Phi)$ is also compact and convex and its cardinality can be 1 or infinity.

2.4. Multi-Information of a Potential

We are now in the position to define multi-information as a function of the interaction potential of a statistical-mechanics model. To extract the minimum stochastic dependence, we define

Definition 2.5. Multi-information given a translation invariant potential Φ is defined by

$$I(\Phi) := \inf_{p \in \mathcal{G}_I(\Phi)} I(p), \tag{13}$$

where $I(p)$ is given by Proposition and Definition 2.2.

Remark 2.6. Because of lower-semicontinuity of $I(p)$ (Theorem 2.2) and compactness of $\mathcal{G}_I$ it follows that the infimum is indeed attained. This follows from general statements about extrema of semicontinuous functions over compact sets, cf. Theorem 25.9 in ref. 13.

The non-uniqueness expressed by (11) is called *phase coexistence.* Phases are the extreme points of the simplex $\mathcal{G}(\Phi)$, which are also just the physically realized states.[6] These so-called pure states have fluctuation-free macroscopic quantities. On the other hand, we can construct convex combinations of them, which do not stand for physically realized states but which express our uncertainty about the state we are in, see ref. 22. That is why the following proposition helps motivating our choice of defining $I(\Phi)$.

Theorem 2.7. Let $\mathrm{ex}(\mathcal{G}_I)$ be the set of extreme points of $\mathcal{G}_I$. We have

$$I(\Phi) = \inf_{p \in \mathcal{G}_I(\Phi)} I(p) = \inf_{p \in \mathrm{ex}(\mathcal{G}_I)(\Phi)} I(p). \tag{14}$$

Thus the infimum is attained in a physically relevant state. To illustrate Definition 2.5 and Proposition 2.7, Fig. 3 shows $I(p)$ over the set of infinite-volume Gibbs measures in the case of the two-dimensional Ising model.

[6] Also, the property of ergodicity is equivalent with being an extreme point of the simplex of translation-invariant probability measures.

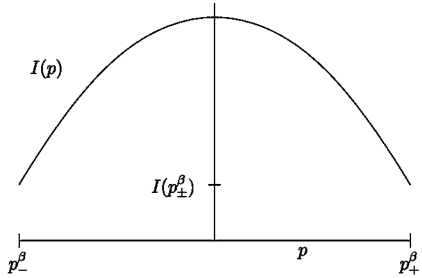

Fig. 3. Schematic view of multi-information depending on p in the 2d Ising model.

3. ISING SQUARE LATTICE

3.1. Multi-Information for the Model

Taking advantage of the wealth of exact results for the two-dimensional Ising model, we can find an explicit expression for multi-information. Definition 2.5 is applied to the Ising potential

$$\Phi^\beta_V(x) = -\beta x_i x_j \qquad \text{if} \quad V = \{i, j\} \subset \mathbb{Z}^2 \quad \text{where } |i-j| = 1, \tag{15}$$

and $\Phi^\beta_V(x) = 0$ for all other sets V, the spin space $S = \{\pm 1\} \ni x_i$ and $\beta \in \mathbb{R}^+$. The parameter β is the inverse temperature and stands for the strength of interaction between spins. We use existing results for free energy and magnetization, critical temperature and the known set of Gibbs measures, for a list of references see ref. 22.

Let us first present a visualization of the main result of this paper: A plot of multi-information of the potential (15) as a function of inverse temperature (see Fig. 4). What one can see is a sharp isolated global maximum at the point of phase transition. The analytic result will be given in 3.2.

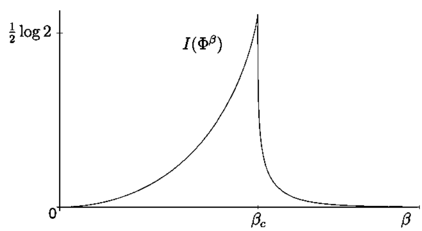

Fig. 4. Multi-information of the Ising square lattice.

It is well known that below a critical temperature the set of infinite-volume Gibbs measures is the convex hull of two extreme probability measures:

$$\mathcal{G}(\Phi^{\beta}) = \{t p_{-}^{\beta} + (1-t)\, p_{+}^{\beta} : t \in [0, 1]\} \tag{16}$$

where the two extreme points $p_{\pm}^{\beta}$ are connected by a spin-flip symmetry that can be written as

$$p_{+}^{\beta}(X_{\Lambda} = x_{\Lambda}) = p_{-}^{\beta}(X_{\Lambda} = -x_{\Lambda}) \qquad \forall \Lambda \subset\subset \mathbb{Z}^{d}. \tag{17}$$

Moreover, for the single-spin expectations (the magnetization) we have $p_{-}^{\beta}(X_0) = -p_{+}^{\beta}(X_0)$. It is essential that these *order parameters* are non-zero for $\beta > \beta_c$. The Yang formula (a rigorous result, see ref. 40, p. 153) is

$$m_{\beta} := p_{+}^{\beta}(X_0) = \begin{cases} (1 - \sinh^{-4} 2\beta)^{\frac{1}{8}} & \text{if} \quad \beta > \beta_c, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The essential feature of the model is a continuous phase transition at a critical temperature β_c:

$$\sinh 2\beta_c = 1, \qquad \text{i.e.} \quad \beta_c = \tfrac{1}{2} \ln(1 + \sqrt{2}). \tag{19}$$

We will also need the entropy (per unit volume)

$$\begin{aligned} h(\beta) &:= h(p_{\pm}^{\beta}) \\ &= \ln(\sqrt{2} \cosh 2\beta) + \frac{1}{\pi} \int_0^{\frac{\pi}{2}} \ln\{1 + \sqrt{1 - \kappa_{\beta}^2 \sin^2 \omega}\}\, d\omega \\ &\quad - 2\beta \tanh 2\beta - \beta \frac{\sinh^2 2\beta - 1}{\sinh 2\beta \cosh 2\beta} \left[\frac{2}{\pi} \int_0^{\frac{\pi}{2}} \frac{d\omega}{\sqrt{1 - \kappa_{\beta}^2 \sin^2 \omega}} - 1 \right], \end{aligned} \tag{20}$$

where

$$\kappa_{\beta} = \frac{2 \sinh 2\beta}{\cosh^2 2\beta}. \tag{21}$$

This expression can be found using the results for free energy $f(\beta)$ and energy $e(\beta)$, see, e.g., ref. 43, because of

$$h(\beta) = \beta e(\beta) - \beta f(\beta). \tag{22}$$

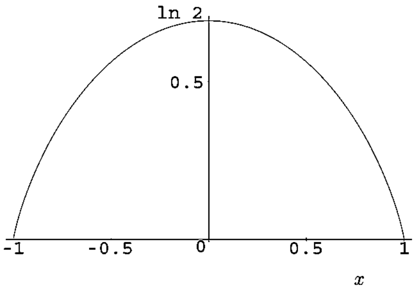

Fig. 5. The function $s(x)$.

Theorem 3.1. Let m_β and $h(\beta)$ be defined by (18) and (20). Also, let

$$s(x)=-\frac{1+x}{2}\ln\frac{1+x}{2}-\frac{1-x}{2}\ln\frac{1-x}{2},\qquad x\in[-1,1], \tag{23}$$

(see Fig. 5).[7] Multi-information of the Ising square lattice is given by

$$I(\Phi^\beta)=s(m_\beta)-h(\beta). \tag{24}$$

Remark 3.2. Notice that similar expressions can be found for all translation-invariant models with binary spin space.

3.2. The Maximum of Multi-Information

Putting some effort into bounding the terms in (24), one can obtain analytic results connecting the phase transition with maximum multi-information:

Theorem 3.3. In the two-dimensional Ising model, multi-information as a function $\beta\mapsto I(\Phi^\beta)$ of inverse temperature attains its isolated global maximum at the point of phase transition $\beta=\beta_c$. At this point, the left-sided derivative goes to $+\infty$, the right-sided one to $-\infty$.

This subsection is devoted to the proof of the theorem. Some technical results are needed. Using the shorthand notation

$$\Theta(\beta):=\frac{\sinh^2 2\beta-1}{\sinh 2\beta\cosh 2\beta}\left[\frac{2}{\pi}\int_0^{\frac{\pi}{2}}\frac{d\omega}{\sqrt{1-\kappa_\beta^2\sin^2\omega}}-1\right], \tag{25}$$

[7] $0\ln 0:=0$.

we have the following bounds:

Lemma 3.4. Let $\beta \geqslant \beta_c$. Then

$$\beta\Theta(\beta) \leqslant \min\left\{\frac{\sinh 2\beta - 1}{2}\ln\frac{\sinh 2\beta + 1}{\sinh 2\beta - 1}, \frac{\beta}{\sinh 2\beta\cosh 2\beta}\right\}. \tag{26}$$

Moreover, $\Theta(\beta_c) = 0$.

$$\begin{aligned} &-\ln(\sqrt{2}\cosh 2\beta) + 2\beta\tanh 2\beta \\ &\qquad \leqslant \min\left\{2\beta_c(\beta - \beta_c) + \sqrt{2}\,\beta_c - \ln 2, \frac{-\beta}{\sinh 2\beta\cosh 2\beta} + \ln\sqrt{2}\right\}, \end{aligned} \tag{27}$$

$$s(m_\beta) \leqslant \ln 2 - \frac{(1 - \sinh^{-4} 2\beta)^{\frac{1}{4}}}{2} - \frac{(1 - \sinh^{-4} 2\beta)^{\frac{1}{2}}}{12}, \tag{28}$$

$$\frac{ds(m_\beta)}{d\beta} \leqslant -\frac{\sinh^{-4} 2\beta}{\tanh 2\beta}\, m_\beta^{-6}. \tag{29}$$

Lemma 3.5. For $0 \leqslant y \leqslant 1/2$ we have

$$\frac{(1 - (1 + y^2)^{-4})^{\frac{1}{4}}}{2} > \frac{y^2}{2}\ln\frac{2 + y^2}{y^2}. \tag{30}$$

Proof of Theorem 3.3. Multi-information is considered in four different regimes. Let us start with the high-temperature case:

(A) $\beta \leqslant \beta_c$. We consider the monotonicity of $I(\Phi^\beta)$. Here, the order parameter m_β vanishes. Thus, as an immediate consequence of Theorem 3.1 we obtain

$$I(\Phi^\beta) = \ln 2 - h(\beta), \qquad \beta \leqslant \beta_c. \tag{31}$$

Using (22) and $d(\beta f(\beta))/d\beta = e(\beta)$,[38] the β derivative is

$$\frac{dI(\Phi^\beta)}{d\beta} = -\frac{dh(\beta)}{d\beta} = -\beta\frac{de(\beta)}{d\beta} = -\beta\frac{d^2(\beta f(\beta))}{d\beta^2} \geqslant 0. \tag{32}$$

This relation follows from the convexity of $-\beta f(\beta)$, see again ref. 38. Hence, the monotonicity up to the critical point is known.

(B) $\beta = \beta_c$. Let us now prove the cusp at the critical temperature. Some care especially for the right-sided derivative is necessary since we have antagonist terms going to infinity. For the left-sided derivative we use the second equality in (32). The divergence of the specific heat is known from the literature:

$$\frac{de(\beta)}{d\beta} = \frac{8}{\pi} \ln |\beta - \beta_c| + \text{bounded terms}, \tag{33}$$

see ref. 40 (p. 152). Since this expression goes to $-\infty$ as $\beta \nearrow \beta_c$, the left-sided derivative of $I(\Phi^{\beta})$ goes to $+\infty$.

Above β_c, the derivative of the first term in (24) comes into play. To make (33) depend on m_β, we use $(1 - \sinh^{-4} 2\beta) \leqslant 8\sqrt{2}\,(\beta - \beta_c)$, which follows from the concavity of the LHS. Together with (18) we have

$$\ln(\beta - \beta_c) \geqslant 8 \ln m_\beta - \ln(8\sqrt{2}). \tag{34}$$

With this and (29) we find

$$\begin{aligned} \lim_{m_\beta \searrow 0} \frac{dI(\Phi^{\beta})}{d\beta} &\leqslant \lim_{m_\beta \searrow 0} \left[-\frac{\sinh^{-4} 2\beta}{\tanh 2\beta} m_\beta^{-6} - \frac{8}{\pi} 8 \ln m_\beta + \text{b.t.} \right] \\ &= \lim_{y \to \infty} y^6 \left[-\sqrt{2} + \frac{64 \ln y}{\pi y^6} + \frac{\text{b.t.}}{y^6} \right] = -\infty, \end{aligned} \tag{35}$$

where we used the substitution $y := 1/m_\beta$ and $\sinh^{-4} 2\beta_c / \tanh 2\beta_c = \sqrt{2}$.

(C) $\beta_c < \beta$. For the remaining β domain we only show $I(\Phi^{\beta}) < I(\Phi^{\beta_c})$ for $\beta > \beta_c$. Together with Theorem 3.1 this becomes

$$s(m_\beta) - h(\beta) < \ln 2 - h(\beta_c), \qquad \beta > \beta_c. \tag{36}$$

With (19), (20), the entropy at β_c is found to be

$$h(\beta_c) = \ln 2 - \sqrt{2}\,\beta_c + \frac{1}{\pi} \int_0^{\frac{\pi}{2}} \ln[1 + \cos\omega]\, d\omega. \tag{37}$$

(We used $\Theta(\beta_c) = 0$ from Lemma 3.4 and $\cosh 2\beta_c = \sqrt{2}$.) The relation to be shown, (36), thus becomes

$$\begin{aligned} s(m_\beta) - \ln(\sqrt{2} \cosh 2\beta) + \frac{1}{\pi} \int_0^{\frac{\pi}{2}} \ln \frac{1 + \cos\omega}{1 + \sqrt{1 - \kappa_\beta^2 \sin^2 \omega}}\, d\omega \\ + 2\beta \tanh 2\beta + \beta\Theta(\beta) - \sqrt{2}\,\beta_c < 0, \qquad \beta > \beta_c. \end{aligned} \tag{38}$$

The integral in (38) is smaller or equal to zero since $\cos\omega \leqslant \sqrt{\ldots}$. It thus suffices to no longer consider this term in the following. For an additional partitioning of the domain above β_c we use

$$\bar{\beta} := \tfrac{1}{2}\,\mathrm{arsinh}(1-K^4)^{-\frac{1}{4}}, \qquad K := 2(\sqrt{2}\,\beta_c - \tfrac{3}{2}\ln 2). \tag{39}$$

(C1) $\beta_c < \beta \leqslant \bar{\beta}$. If we feed the corresponding terms from Lemma 3.4 into (38), we obtain the following inequality whose proof suffices to prove (38):

$$-\frac{(1-\sinh^{-4}2\beta)^{\frac{1}{4}}}{2} - \frac{(1-\sinh^{-4}2\beta)^{\frac{1}{2}}}{12} + 2\beta_c(\beta-\beta_c) + \frac{\sinh 2\beta - 1}{2}\log\frac{\sinh 2\beta + 1}{\sinh 2\beta - 1} < 0. \tag{40}$$

We now show that the sum of the first and last terms of the LHS, as well as the sum of the two terms in-between them are negative in the required range β_c up to $\bar{\beta}$. For first and last term we define y by $\sinh 2\beta =: 1 + y^2$ and use Lemma 3.5. Also using

$$\sinh 2\bar{\beta} < 1 + (1/2)^2 \tag{41}$$

(which can be checked using (39), $\sinh 2\bar{\beta} \approx 1.18$) it is clear that the sum of first and last terms of the LHS of (40) are smaller zero in the required range. For the middle terms, we square the corresponding inequality to obtain

$$4\beta_c^2(\beta-\beta_c)^2 < \frac{1-\sinh^{-4}2\beta}{144}, \qquad \beta_c < \beta \leqslant \bar{\beta}. \tag{42}$$

Here, we do the following: At $\beta = \beta_c$ both sides are equal to zero. Taking the second derivatives shows that the LHS is a concave function, the RHS a convex one. If the inequality holds for the point $\bar{\beta}$, it also holds for the entire interval $(\beta_c, \bar{\beta}]$. Using (39) one calculates for $\beta = \bar{\beta}$

$$4\beta_c^2(\bar{\beta}-\beta_c)^2 < \frac{K^4}{144}, \tag{43}$$

from which we obtain (taking the square root, shifting terms and applying the hyperbolic sine)

$$\sinh 2\bar{\beta} < \sinh\left[\frac{K^2}{12\beta_c} + 2\beta_c\right] = 1.19471\ldots. \tag{44}$$

Putting in the value of $\sinh \bar{\beta}$ using (39) shows that this relation indeed holds. Hence (42) holds in the required β range including $\bar{\beta}$, and thus (38) holds.

(C2) $\bar{\beta} < \beta$. Lemma 3.4 again makes (38) an inequality whose proof suffices to prove (38):

$$\frac{3}{2}\ln 2 - \sqrt{2}\,\beta_c - \frac{(1-\sinh^{-4} 2\beta)^{\frac{1}{4}}}{2} < 0, \qquad \bar{\beta} < \beta. \tag{45}$$

The corresponding equality is just solved by $\bar{\beta}$, cf. (39). As the LHS is monotonically decreasing, the inequality holds above $\bar{\beta}$. ▮

3.3. Discussion

Since $I = H(p_{\{0\}}) - h(p)$ (see (50)), two sorts of uncertainty or knowledge play a role: In Fig. 6 we show the information about the single spin that stems from the entire system, $\ln 2 - h(p)$ (information per site, or redundancy) and the information about the single spin that comes from a single spin only, $\ln 2 - H(p_{\{0\}})$ (cf. also, e.g., refs. 9 and 10 for a discussion of entropies on different scales). Multi-information is the difference between these terms and can be seen as the average information the single spin carries about the full system and vice versa. We will show this in terms of mutual information for the one-dimensional case.

Interestingly, if one plots correlation functions of near neighbours and the squared magnetization, one gets a very similar picture to Fig. 6, see Fig. 8.10 in ref. 35. The difference then would be covariance, which, although only accounting for second-order interactions, behaves similar to multi-information in the Ising case, cf. also ref. 29.

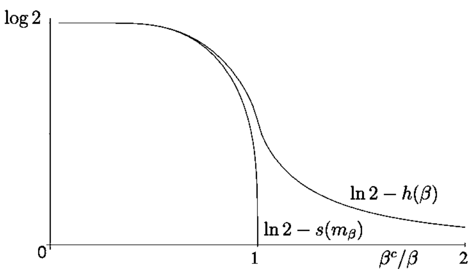

Fig. 6. 2d Ising: Full-system and single-spin information versus normalized temperature. Their difference is multi-information.

Figure 4 fits nicely into the universal picture discussed in ref. 3, Fig. 1. Clearly, the cusp is due to the phase transition, which is also characterized by non-analyticity of the free energy.

Let us also briefly mention the approaches to stochastic dependence in the Ising square lattice that can, to our knowledge, be found in the literature. In ref. 34, the mutual information between two spins depending on their distance and on temperature is considered. The relation to magnetization and correlation function is shown. Plots very similar to ours are obtained by simulation. In ref. 3, a similar plot, here of excess entropy (see next section) is obtained by simulation. There are different choices for a definition of excess entropy in two dimensions. In ref. 21 this problem is tackled, and three different definitions (seen as equivalent in one dimension) are presented in the 2d case. Plots for nearest and next-nearest neighbour Ising model are obtained using simulations.

Since multi-information coincides with excess entropy for the Ising chain (see next section), we can consider it *another definition of excess entropy* for the Ising square lattice.

Only in the last of the above cited articles is the problem of multiple excess entropies stated. They appear because of non-uniqueness of the Gibbs measure. Note that Definition 2.5 takes this fact into account.

4. ONE-DIMENSIONAL SYSTEMS

4.1. Multi-Information and Excess Entropy

One-dimensional systems on $(S^{\mathbb{Z}}, \mathscr{F})$ are readily interpreted as time-dependent. Translation invariance is seen as stationarity of a stochastic process. In this section it is convenient to leave out the dependence on p and to use the notation $\dots X_{-1}, X_0, X_1, X_2 \dots$ for the infinite chain of random variables that take values from the finite alphabet S, so, e.g., $H(X_1,\dots X_n) = H(p_{\{1,\dots,n\}})$ denote the entropy of n consecutive random variables. In the following we want to discuss the relationship of multi-information to standard quantities defined in the time context, so we make the following

Definition 4.1. Let the *conditional entropies* be denoted by $h_n := H(X_n \mid X_{n-1},\dots, X_1) = H(p_{\{1,\dots,n\}}) - H(p_{\{1,\dots,n-1\}})$ where $n > 0$ and $h_1 := H(X_1) =: H_0$, the *entropy rate* by $h := \lim_{n\to\infty} H(p_{\{1,\dots,n\}})/n$. The *mutual information* between X_n and $X_1,\dots, X_{n-1}$ is $MI(X_n; X_1,\dots, X_{n-1}) := H_0 - h_n$. The *excess entropy* (or effective measure complexity) is defined by

$$E := \sum_{n=1}^{\infty} (h_n - h). \tag{46}$$

It is well known that h_n converges to h for $n \to \infty$. Depending on the speed of convergence, E can be finite or infinite. It is known to measure the total information needed for optimal predictions,[23] see ref. 17 and references there. On the other hand h_n is the average unpredictability[19] of X_n given the values of the $n-1$ preceding variables in the chain.

Using the chain rule for entropy, finite-volume multi-information can be written as

$$I_L(X_1,\ldots, X_L) = \sum_{n=1}^{L} (H(X_n) - h_n) = \sum_{n=1}^{L} (H_0 - h_n). \tag{47}$$

Defining a finite-volume excess entropy $E_L(X_1,\ldots X_L) := \sum_{n=1}^{L} (h_n - h)$, we observe that E_L and I_L are two sides of the same coin, both summing up differences of h_n from a fixed quantity. These summands are called measure complexities[19, 23] for E_L and mutual informations for I_L. In the spirit of refs. 17 and 20 the similarity between E_L and I_L is schematically illustrated in Fig. 7. Note however that in the limit both quantities are in general no longer similar.

Theorem 4.2. For p a translation-invariant probability measure on $(S^{\mathbb{Z}}, \mathscr{F})$, we have the following expressions for I and E:

(i) $I = \lim_{n\to\infty} MI(X_n; X_1,\ldots, X_{n-1})$,

(ii) $E = I + \sum_{n=2}^{\infty} (h_n - h)$, especially $E = I$ if p is Markov.

The proof follows immediately from the definitions of the involved quantities. Multi-information in the limit is thus the average information that the past carries about the next variable in the chain. Excess entropy is

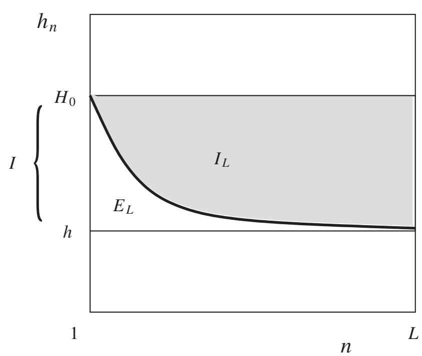

Fig. 7. The area below the h_n curve converges to excess entropy, the area above the curve is finite-volume multi-information.

often discussed in the context of mutual information between the past and the future, i.e., between two semi-infinite blocks in the chain. Also, it is well-known that the convergence behaviour of h_n informs us about important system properties.[9, 10] However, from the size of E alone we cannot in general tell how fast h_n converges: Different values of I give different convergence behaviour for the same E. Thus E and I complement each other as regards information about the convergence of h_n. Moreover, being the first summand in (46), I is a lower bound on E. We have $E = I$ for Markov chains (nearest-neighbour spin chains) since $h_n = h$ for all $n > 1$ in this case. This leads us to the Ising model.

4.2. Ising Model as a Markov Chain

The usual description of Markov chains by one-sided conditional probabilities is closely related to a description by means of the Gibbsian probability kernels. Indeed, S assumed finite, we have a one-to-one correspondence between the set of all positive transition matrices of a Markov chain and the shift-invariant potentials for which $\Phi_V = 0$ if $V \neq \{i\}$ or $\{i, i+1\}$ (homogeneous nearest-neighbour).[22, 25] The set of Gibbs measures of such potentials only contains the unique measure defining the Markov chain whose transition matrix can be calculated explicitly from Φ.

There is a complete solution for the nearest-neighbour Ising chain in a magnetic field b. Our interaction potential depends on this additional parameter now:

$$\begin{aligned} \Phi_V^{\beta, b}(x) &= -\beta x_i x_{i+1} \qquad &\text{if} \quad V = \{i, i+1\}, \\ \Phi_V^{\beta, b}(x) &= -\beta b x_i \qquad &\text{if} \quad V = \{i\}, \end{aligned} \tag{48}$$

and $\Phi_V^{\beta, b}(x) = 0$ for all other $V \subset \mathbb{Z}$. In analogy to (24) we can write down a formula for multi-information (for entropy and magnetization see, e.g., ref. 36) and plot it depending on β and b (Fig. 8). Alternatively, one can get an expression for I from the transition matrix of the corresponding Markov chain.[30] This matrix is[22]

$$P_{\beta, b} = \begin{pmatrix} e^{-2\beta b} q_{\beta, b}^{-1} & 1 - e^{-2\beta b} q_{\beta, b}^{-1} \\ 1 - q_{\beta, b}^{-1} & q_{\beta, b}^{-1} \end{pmatrix} \tag{49}$$

where $q_{\beta, b} := e^{-\beta b}(\cosh \beta b + \sqrt{e^{-4\beta} + \sinh^2 \beta b})$. For $b = 0$ the two ground state configurations are described by the unique Gibbs measure that gives equal weight to the respective Dirac measures (cf. also the dotted line in Fig. 2). We have symmetry in this case, the probability that the direction of

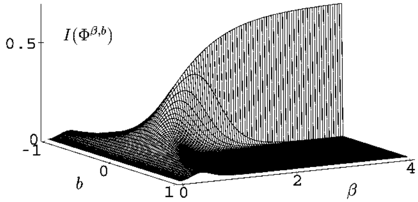

Fig. 8. Multi-information (=excess entropy) in the Ising chain.

the spin stays the same after the next time step is $1/(1+e^{-2\beta})$. H_0 keeps its maximal value ln 2 through all temperatures, so multi-information is completely determined by the entropy rate. It means that we can predict that the spin will not flip with higher and higher certainty when lowering temperature, but that there are equal chances for both possibilities regarding the actual value of X_1. For $b>0$ the symmetry is broken. We have a preference for the value of X_1 in the direction of the field, H_0 is reduced. There is competition between the order creating and destroying influences of field and temperature. The past carries most information about X_1 if these influences balance out. Although no phase transition takes place, this maximization of I can be seen as a related phenomenon.

It is interesting to investigate the nature of this non-critical "transition." See ref. 16 for a similar picture as Fig. 8 (for b held fixed). There, excess entropy (which, as mentioned, equals multi-information in this case) was calculated using transfer matrices.

5. PROOFS OF LEMMAS AND THEOREMS

Let us first state a lemma which will be needed for the proof of Theorem 2.2:

Lemma 5.1. Let p, q be probability measures on $(\Omega, \mathscr{F})$. We have

(i) $0 \leqslant H(p_{\{0\}}) \leqslant \ln |S|$,

(ii) $H((tp+(1-t)\,q)_{\{0\}}) \geqslant tH(p_{\{0\}})+(1-t)\,H(q_{\{0\}})\ \forall t \in [0, 1]$,

(iii) $H(p_{\{0\}})$ is continuous for the weak* topology.

Proof. For (i) and (ii) see ref. 14. In our case the measures are marginals of p, q, but (ii) follows immediately from the affinity of the projection of p onto $p_{\{0\}}$, i.e., $(tp+(1-t)\,q)_{\{0\}} = tp_{\{0\}}+(1-t)\,q_{\{0\}}$.

(iii) follows from the continuity of entropy w.r.t. $p_{\{0\}}$, see ref. 12. It remains to show that the projection π_0 of p onto $p_{\{0\}}$ is continuous, i.e., that from $p_n \to p$ follows $\pi_0(p_n) \to \pi_0(p)$. Continuity for the weak* topology on our topological space Ω means that $p_n \to p$ is equivalent to $p_n(f) \to p(f)\ \forall f \in C(\Omega)$ (being the space of continuous functions for the product topology). For f we choose the indicator function $1_{\{X_0 = x_0\}}$ (which is continuous since the inverse images of 1 and 0 are open sets). Now we have

$$\pi_0(p_n)(x_0) = p_n(1_{\{X_0 = x_0\}}) \to p(1_{\{X_0 = x_0\}}) = \pi_0(p)(x_0) \qquad \forall x_0 \in S. \quad \blacksquare$$

Proof of Theorem 2.2. We use the existence of the van Hove limit, upper-semicontinuity and affinity of the entropy $\lim_{\Lambda \nearrow \mathbb{Z}^d} \frac{1}{|\Lambda|} H(p_\Lambda) =: h(p) \in [0, \ln |S|]$ (cf. refs. 27 and 38, these properties follow immediately from the proof for a more generally defined entropy not requiring finite S) and Lemma 5.1. Similarly to $I(p_\Lambda)$, we can split $I(p) =$

$$\lim_{\Lambda \nearrow \mathbb{Z}^d} \frac{1}{|\Lambda|} \left[\sum_{i \in \Lambda} H(p_{\{i\}}) - H(p_\Lambda) \right] = H(p_{\{0\}}) - h(p), \tag{50}$$

where the second equality follows from translation invariance. Since we have all the required properties for $h(p)$ and $H(p_{\{0\}})$, the theorem follows. $\blacksquare$

Proof of Theorem 2.7. We have to show that the Infimum is always attained in an extreme point of $\mathcal{G}_I$. This follows from compactness and convexity of $\mathcal{G}_I$ as well as lower-semicontinuity and concavity of $I(p)$ according to Theorem 25.9 in Vol. 2 of ref. 13. $\blacksquare$

Proof of Theorem 3.1. We show:

(i) $I(\Phi^\beta) = I(p^\beta_\pm)$,

(ii) $I(p^\beta_\pm) = s(m_\beta) - h(\beta)$.

(i) According to (16), the $p^\beta_\pm$ are the only extreme Gibbs measures. First we show that $I(p)$ is symmetric around $(p^\beta_- + p^\beta_+)/2$. For this we use the measures $p = (1-t)\, p^\beta_- + t p^\beta_+$ and $p' = t p^\beta_- + (1-t)\, p^\beta_+$ for $t \in [0, 1]$. Because of the spin-flip symmetry (17), for $\Lambda \subset\subset \mathbb{Z}^d$ we have $H(p_\Lambda) = H(p'_\Lambda)$. Taking the limit yields $h(p) = h(p')$, with (50) we also have $I(p) = I(p')$. By Theorem 2.7 $I(\Phi^\beta) = \inf_{p \in \{p^\beta_-, p^\beta_+\}} I(p)$, because of the above symmetry we have $I(p^\beta_-) = I(p^\beta_+)$ (see Fig. 3).

(ii) We have $p(X_0) = \sum_{x_0 = \pm 1} p(X_0 = x_0)\, x_0 = p(X_0 = 1) - p(X_0 = -1)$. Also using $\sum_{x_0 = \pm 1} p(x_0) = 1$, we obtain for the single-spin probability

$$p(X_0 = x_0) = \frac{1 + x_0\, p(X_0)}{2}. \tag{51}$$

Hence

$$H_0(p_{\pm}^{\beta}) = -\sum_{x_0=\pm 1} \frac{1+x_0\, p_{\pm}^{\beta}(X_0)}{2} \ln \frac{1+x_0\, p_{\pm}^{\beta}(X_0)}{2} = s(p_{+}^{\beta}(X_0)). \tag{52}$$

Since s is even, both expectations give the same result. From (50) follows (ii). ▮

Proof of Lemma 3.4.

Equation (26).

(1) $\beta\Theta(\beta) \leqslant \frac{\sinh 2\beta - 1}{2} \log \frac{\sinh 2\beta + 1}{\sinh 2\beta - 1}$, $\Theta(\beta_c) = 0$

Taking the -1 out of the square brackets in (25) and into the integral, the resulting numerator in the integral can be modified like this:

$$1 - \sqrt{1 - \kappa_\beta^2 \sin^2 \omega} \leqslant 1 - (1 - \kappa_\beta^2 \sin^2 \omega) \leqslant \kappa_\beta^2 \sin \omega. \tag{53}$$

Estimating the resulting integral[11] yields

$$\int_0^{\frac{\pi}{2}} \frac{\sin \omega}{\sqrt{1 - \kappa^2 \sin^2 \omega}}\, d\omega = \frac{1}{2\kappa} \ln \frac{1+\kappa}{1-\kappa}, \qquad |\kappa| < 1. \tag{54}$$

We use (21) and

$$\frac{1+\kappa_\beta}{1-\kappa_\beta} = \left(\frac{\sinh 2\beta + 1}{\sinh 2\beta - 1}\right)^2 \tag{55}$$

to find

$$\Theta(\beta) \leqslant \frac{\sinh^2 2\beta - 1}{\cosh^3 2\beta} \frac{4}{\pi} \ln \frac{\sinh 2\beta + 1}{\sinh 2\beta - 1}. \tag{56}$$

We still have to show that

$$\beta \frac{\sinh^2 2\beta - 1}{\cosh^3 2\beta} \frac{4}{\pi} \leqslant \frac{\sinh 2\beta - 1}{2}, \tag{57}$$

or

$$\frac{\sinh 2\beta + 1}{\cosh^3 2\beta} \leqslant \frac{\pi}{8\beta}. \tag{58}$$

Note that for $\beta > \beta_c$ we have $\sinh 2\beta > 1$, so the LHS can be bounded like

$$\frac{\sinh 2\beta+1}{\cosh^3 2\beta} \leqslant \frac{\sinh^2 2\beta+1}{\cosh^3 2\beta} = \frac{\cosh^2 2\beta}{\cosh^3 2\beta} \leqslant \frac{1}{1+(2\beta)^2/2}, \tag{59}$$

which follows from the series expansion of the hyperbolic cosine. Returning to (58), we have to show

$$\frac{1}{1+(2\beta)^2/2} \leqslant \frac{\pi}{8\beta} \tag{60}$$

or

$$0 \leqslant \beta^2 - \frac{4}{\pi}\beta + \frac{1}{2}, \tag{61}$$

which is fulfilled for $\beta = 0$. Since the corresponding equation has no real zeroes, the relation also holds for all the other β, and thus (57) is proven.

We still have to show that $\Theta(\beta_c) = 0$. Clearly, $\Theta(\beta) \geqslant 0$ for $\beta \geqslant \beta_c$. In part (C1) of the proof of Theorem 3.3 we have moreover shown that $\beta\Theta(\beta) \leqslant (1-\sinh^{-4} 2\beta)^{\frac{1}{4}}/2$, given the just proven first statement of the lemma. Hence we have

$$0 \leqslant \Theta(\beta_c) \leqslant \frac{(1-\sinh^{-4} 2\beta_c)^{\frac{1}{4}}}{2\beta_c} = 0. \tag{62}$$

(2) $\Theta(\beta) \leqslant 1/\sinh 2\beta \cosh 2\beta$

For $\beta \geqslant \beta_c$ we have

$$\sqrt{1-\kappa_\beta^2} = \frac{\sinh^2 2\beta - 1}{\cosh^2 2\beta}. \tag{63}$$

Using this, (25) becomes

$$\Theta(\beta) = \coth 2\beta \sqrt{1-\kappa_\beta^2} \left[\frac{2}{\pi} \int_0^{\frac{\pi}{2}} \frac{d\omega}{\sqrt{1-\kappa_\beta^2 \sin^2\omega}} - 1 \right]. \tag{64}$$

We take the root into the square brackets and obtain the integrand

$$\sqrt{\frac{1-\kappa^2}{1-\kappa^2 \sin^2 \omega}} - \sqrt{1-\kappa^2}. \tag{65}$$

This is a continuous function in ω which has the value zero at $\omega = 0$ and the value $1-\sqrt{1-\kappa^2}$ at $\omega = \pi/2$. Connecting these two points, one obtains the diagonal of a rectangle with area $(1-\sqrt{1-\kappa^2})\,\frac{\pi}{2}$. The integral can be bounded by half of the area of this rectangle. We show that the part of the area A of the rectangle above the integrand is greater or equal to the part B of the area below the integrand (the integral itself). Instead of comparing A and B, we compare their respective integrands. We obtain the integrand of A by twice reflecting the integrand of B: once in the vertical line through $\pi/4$, once in the horizontal line through $\frac{1-\sqrt{1-\kappa^2}}{2}$. The resulting inequality is

$$\sqrt{\frac{1-\kappa^2}{1-\kappa^2 \sin^2\omega}} - \sqrt{1-\kappa^2} \leqslant 1 - \sqrt{\frac{1-\kappa^2}{1-\kappa^2 \cos^2\omega}}, \tag{66}$$

or put differently,

$$\left\{\sqrt{\frac{1-\kappa^2}{1-\kappa^2 \sin^2\omega}} + \sqrt{\frac{1-\kappa^2}{1-\kappa^2 \cos^2\omega}}\right\} \leqslant 1 + \sqrt{1-\kappa^2}. \tag{67}$$

The expression in curly brackets is symmetric around $\pi/4$ because of $\cos^2\omega = \sin^2(\pi/2-\omega)$. In order to prove the inequality, we just have to show that the expression is monotonic decreasing up to $\pi/4$ (for $\omega = 0$ we have equality), which is easily seen by looking at its derivative. Thus $B \leqslant A$ and

$$\int_0^{\frac{\pi}{2}} \left[\sqrt{\frac{1-\kappa^2}{1-\kappa^2 \sin^2\omega}} - \sqrt{1-\kappa^2}\right] d\omega \leqslant \frac{1}{2}\,(1-\sqrt{1-\kappa^2})\,\frac{\pi}{2}. \tag{68}$$

Continuing with (64) we find

$$\Theta(\beta) \leqslant \coth 2\beta\, \frac{2}{\pi}\, \frac{1}{2}\,(1-\sqrt{1-\kappa_\beta^2})\,\frac{\pi}{2}. \tag{69}$$

Again using (63), we obtain the desired bound.

Equation (27).

(1) $-\ln(\sqrt{2}\cosh 2\beta) + 2\beta \tanh 2\beta \leqslant 2\beta_c(\beta-\beta_c) + \sqrt{(2)}\,\beta_c - \ln 2$

Expanding the LHS into a Taylor series around β_c, we observe that the second derivative is smaller zero for $4\beta \tanh 2\beta > 1$ (which holds for $\beta > \beta_c$), so for an upper bound the series can be truncated after the first term.

(2) $-\ln(\sqrt{2}\cosh 2\beta)+2\beta\tanh 2\beta \leqslant \frac{-\beta}{\sinh 2\beta\cosh 2\beta}+\ln\sqrt{2}$

Factoring out $e^{2\beta}$ from the definition of $\cosh 2\beta$, the LHS can be equated to $\beta(2\tanh 2\beta-2)+\ln\frac{\sqrt{2}}{1+e^{-4\beta}}$. Moreover,

$$2\tanh 2\beta-2+\frac{1}{\sinh 2\beta\cosh 2\beta}=\frac{e^{-4\beta}}{\sinh 2\beta\cosh 2\beta}. \tag{70}$$

It follows that

$$\begin{aligned}&-\ln(\sqrt{2}\cosh 2\beta)+2\beta\tanh 2\beta-\frac{\beta}{\sinh 2\beta\cosh 2\beta}\\&\leqslant \beta\frac{e^{-4\beta}}{\sinh 2\beta\cosh 2\beta}-\ln[1+e^{-4\beta}]+\ln\sqrt{2}\leqslant\ln\sqrt{2}.\end{aligned} \tag{71}$$

The last relation was obtained using the fact that the sum of the first two terms does not exceed zero. To show this, we modify the first term as follows:

$$\beta\frac{e^{-4\beta}}{\sinh 2\beta\cosh 2\beta}=\frac{2\beta e^{-4\beta}}{\sinh 4\beta}\leqslant\frac{2\beta e^{-4\beta}}{4\beta}=\frac{e^{-4\beta}}{2}. \tag{72}$$

Now we have the inequality

$$\frac{e^{-4\beta}}{2}\leqslant\ln[1+e^{-4\beta}]=\frac{e^{-4\beta}}{2}+\frac{e^{-4\beta}}{2}-\frac{e^{-8\beta}}{2}+\sum_{n=2}^{\infty}\left[\frac{1}{e^{4\beta(2n-1)}(2n-1)}-\frac{1}{e^{4\beta(2n)}2n}\right], \tag{73}$$

since on the RHS the terms after the first $\exp(-4\beta)/2$ are pairwise greater 0 (we expanded $\ln(1+x)$, cf. ref. 11). Thus (71) holds.

Equation (28). The function $s(x)$ can be rewritten as follows:

$$\begin{aligned}s(x)&=-\frac{1+x}{2}\ln\frac{1+x}{2}-\frac{1-x}{2}\ln\frac{1-x}{2}\\&=\ln 2-\frac{1}{2}[(1+x)\ln(1+x)+(1-x)\ln(1-x)].\end{aligned} \tag{74}$$

The expression in square brackets is expanded (see again ref. 11) and bounded below:

$$
\begin{aligned}
[\,] &= (1+x) \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} - (1-x) \sum_{n=1}^{\infty} \frac{x^n}{n} \\
&= \sum_{n=1}^{\infty} \left[(-1)^{n+1} \frac{x^n}{n} - \frac{x^n}{n} \right] + x \sum_{n=1}^{\infty} \left[(-1)^{n+1} \frac{x^n}{n} + \frac{x^n}{n} \right] \\
&= -\sum_{n=1}^{\infty} \frac{x^{2n}}{n} + 2 \sum_{n=1}^{\infty} \frac{x^{2n}}{2n-1} = \sum_{n=1}^{\infty} \frac{x^{2n}}{2n^2-n} \geqslant x^2 + \frac{x^4}{6}.
\end{aligned}
\tag{75}
$$

This bound is possible since all the coefficients in the sum are positive. Together with (18) for m_β we thus obtain

$$
s(m_\beta) \leqslant \ln 2 - \frac{(1-\sinh^{-4} 2\beta)^{\frac{1}{4}}}{2} - \frac{(1-\sinh^{-4} 2\beta)^{\frac{1}{2}}}{12}.
\tag{76}
$$

Equation (29). We have

$$
\frac{ds(m_\beta)}{d\beta} = \frac{\sinh^{-4} 2\beta}{\tanh 2\beta}\, m_\beta^{-7}\, \frac{1}{2} \ln \frac{1-m_\beta}{1+m_\beta}.
\tag{77}
$$

Expanding the logarithm into a series (see, e.g., ref. 11), we obtain the following bound:

$$
m_\beta^{-1}\, \frac{1}{2} \ln \frac{1-m_\beta}{1+m_\beta} = -1 - \sum_{n=1}^{\infty} \frac{m_\beta^{2n}}{2n+1} \leqslant -1.
\tag{78}
$$

From this the lemma follows. ∎

Proof of Lemma 3.5. In order to show that

$$
\frac{(1-(1+y^2)^{-4})^{\frac{1}{4}}}{2} > \frac{y^2}{2} \ln \frac{2+y^2}{y^2}, \qquad 0 \leqslant y \leqslant \frac{1}{2}
\tag{79}
$$

we show that the LHS is greater than $\frac{3}{5}\, y$, the RHS is smaller than $\frac{3}{5}\, y$. So for the RHS we have to show

$$
\frac{5}{6}\, y \ln \frac{2+y^2}{y^2} < 1.
\tag{80}
$$

We only need $y \leqslant 1/2$. In this case $2+y^2 \leqslant 9/4$, and thus we also have

$$\frac{5}{6}\, y \ln \frac{2+y^2}{y^2} \leqslant \frac{5}{6}\, y \ln \frac{9}{4y^2} = -\frac{5}{3}\, y \ln \frac{2}{3}\, y. \tag{81}$$

By equating the first derivative to zero we obtain the maximum of the function $-y\ln(2y/3)$ at $3/(2\mathrm{e})$. Hence

$$-\frac{5}{3}\, y \ln \frac{2}{3}\, y \leqslant \frac{5}{3}\frac{3}{2\mathrm{e}} < 1. \tag{82}$$

With this, (80) is shown for $y \leqslant 1/2$. For the LHS of (79) one has to prove:

$$\frac{(1-(1+y^2)^{-4})^{\frac{1}{4}}}{y} > \frac{6}{5}. \tag{83}$$

The LHS is modified as follows:

$$= \sqrt[4]{\frac{(1+y^2)^4-1}{y^4(1+y^2)^4}} = \frac{1}{(1+y^2)}\sqrt[4]{\frac{y^8+4y^6+6y^4+4y^2}{y^4}} \geqslant \frac{\sqrt[4]{4y^{-2}}}{1+y^2}. \tag{84}$$

Since in the last expression the denominator is monotonically decreasing, the numerator increasing, for a lower bound it suffices to evaluate the expression for the greatest y:

$$\frac{\sqrt[4]{4y^{-2}}}{1+y^2} \geqslant \frac{\sqrt[4]{4(\frac{1}{2})^{-2}}}{1+\frac{1}{4}} = \frac{8}{5} > \frac{6}{5}, \qquad y \leqslant \frac{1}{2}. \quad \blacksquare \tag{85}$$

ACKNOWLEDGMENTS

This work was mainly done at the Max-Planck Institute for Mathematics in Leipzig. I.E. thanks Jürgen Jost for general support and Ulrich Steinmetz for discussions. N.A. also thanks the Santa Fe Institute. We are grateful to one of the anonymous referees for considerably improving the manuscript by pointing out the importance of excess entropy and specifying relevant literature.

REFERENCES

1. S. Amari, Information geometry on hierarchy of probability distributions, *IEEE Trans. Inform. Theory* **47**:1701–1711 (2001).
2. S. Amari and H. Nagaoka, in Methods of information geometry, *AMS Translations of Mathematical Monographs*, Vol. 191 (Oxford University Press, Oxford, 2000).

3. D. V. Arnold, Information-theoretic analysis of phase transitions, *Complex Systems* **10**:143–155 (1996).
4. N. Ay, An information geometric approach to a theory of pragmatic structuring, *Ann. Probab.* **30**:416–436 (2002).
5. N. Ay, Locality of global stochastic interaction in directed acyclic networks, *Neural Comput.* **14**:2959–2980 (2002).
6. N. Ay and A. Knauf, *Maximizing Multi-Information*, Max-Planck-Institute of Mathematics in the Sciences, Preprint No. 42/2003 (2003).
7. H. Bauer, *Probability Theory* (de Gruyter, Berlin/New York, 1996).
8. A. J. Bell, *The Co-Information Lattice*, preprint Redwood Neuroscience Institute RNI-TR-02-1 (2002).
9. W. Bialek, I. Nemenman, and N. Tishby, Predictability, complexity, and learning, *Neural Comput.* **13**:2409–2463 (2001).
10. P. M. Binder and J. Plazas, Multiscale analysis of complex systems, *Phys. Rev. E* **63**: 065203R (2001).
11. I. N. Bronstein and K. A. Semendjajew, *Taschenbuch der Mathematik* (Teubner, Leipzig, 1989).
12. D. E. Catlin, *Estimation, Control, and the Discrete Kalman Filter* (Springer-Verlag, New York, 1989).
13. G. Choquet, *Lectures on Analysis*, Vols. I–III (Benjamin, London, 1969).
14. T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
15. J. P. Crutchfield and K. Young, Computation at the onset of chaos, in *Complexity, Entropy, and the Physics of Information*, W. H. Zurek, ed. (Addison–Wesley, Reading, 1990).
16. J. P. Crutchfield and D. P. Feldman, Statistical complexity of simple one-dimensional spin systems, *Phys. Rev. E* **55**:R1239–R1242 (1997).
17. J. P. Crutchfield and D. P. Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, *Chaos* **13**:25–54 (2003).
18. R. L. Dobrushin, The description of a random field by means of conditional probabilities and condition of its regularities, *Theory Probab. Appl.* **13**:458–486 (1968).
19. W. Ebeling, Prediction and entropy of nonlinear dynamical systems and symbolic sequences with LRO, *Physica D* **109**:42–52 (1997).
20. D. P. Feldman and J. P. Crutchfield, Discovering noncritical organization: Statistical mechanical, information theoretic, and computational views of patterns in one-dimensional spin systems, Santa Fe Institute Working Paper 98-04-026.
21. D. P. Feldman and J. P. Crutchfield, Structural information in two-dimensional patterns: Entropy convergence and excess entropy, *Phys. Rev. E* **67**:051104 (2003).
22. H.-O. Georgii, *Gibbs Measures and Phase Transitions* (de Gruyter, Berlin/New York, 1988).
23. P. Grassberger, Toward a quantitative theory of self-generated complexity, *Internat. J. Theoret. Phys.* **25**:907–938 (1986).
24. E. Greenfield and H. Lecar, Mutual information in a dilute, asymmetric neural network model, *Phys. Rev. E* **63**:041905 (2001).
25. D. Griffeath, Introduction to random fields, in *Denumerable Markov Chains*, J. G. Kemeney, J. L. Snell, and A. W. Knapp, eds. (Springer, New York/Heidelberg/Berlin, 1976).
26. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (Wiley, 2001).
27. R. B. Israel, *Convexity in the Theory of Lattice Gases* (Princeton University Press, Princeton, 1979).
28. E. T. Jaynes, Information theory and statistical mechanics, *Phys. Rev.* **106**:620–630 (1957).

29. W. Li, Mutual information versus correlation functions, *J. Stat. Phys.* **60**:823–837 (1990).
30. W. Li, On the relationship between complexity and entropy for Markov chains and regular languages, *Complex Systems* **5**:381–399 (1991).
31. R. Linsker, Self-organization in a perceptual network, *IEEE Comput.* **21**:105–117 (1988).
32. O. E. Lanford and D. Ruelle, Observables at infinity and states with short range correlations in statistical mechanics, *Comm. Math. Phys.* **13**:194–215 (1969).
33. B. Luque and A. Ferrera, Measuring mutual information in random Boolean networks, *Complex Systems* **12**:241–246 (2001).
34. H. Matsuda, K. Kudo, R. Nakamura, O. Yamakawa, and T. Murata, Mutual information of Ising systems, *Internat. J. Theoret. Phys.* **35**:839–845 (1996).
35. B. M. McCoy and T. T. Wu, *The Two-Dimensional Ising Model* (Harvard University Press, Cambridge, Massachusetts, 1973).
36. R. K. Pathria, *Statistical Mechanics* (Butterworth–Heinemann, Oxford, 1996).
37. D. Petritis, Thermodynamic formalism of neural computing, in *Dynamics of Complex Interacting Systems*, E. Goles and M. Servet, eds. (Kluwer Academic Publishers, Dordrecht/Boston/London, 1996).
38. D. Ruelle, *Statistical Mechanics—Rigorous Results* (World Scientific, Singapore, 1999).
39. C. E. Shannon, A mathematical theory of communication, *Bell System Tech. J.* **27**: 379–423, 623–656 (1948).
40. B. Simon, *The Statistical Mechanics of Lattice Gases* (Princeton University Press, Princeton, 1993).
41. M. Studeny and J. Vejnarova, The multiinformation function as a tool for measuring stochastic dependence, in *Learning in Graphical models*, M. I. Jordan, ed. (Kluwer, Dordrecht, 1998).
42. G. Tononi, O. Sporns, and G. Edelman, A measure for brain complexity: Relating functional segregation and integration in the nervous system, *Proc. Natl. Acad. Sci. USA* **91**:5033–5037 (1994).
43. G. H. Wannier, *Statistical Physics* (Dover, New York, 1987).