entropy

Article

The Stochastic Complexity of Spin Models: Are Pairwise Models Really Simple?

Alberto Beretta [1], Claudia Battistin [2,*], Clélia de Mulatier [1,3], Iacopo Mastromatteo [4] and Matteo Marsili [1,5]

[1] The Abdus Salam International Centre for Theoretical Physics (ICTP), Strada Costiera 11, I-34014 Trieste, Italy; albertoberetta92@gmail.com (A.B.); clelidm@gmail.com (C.d.M.); marsili@ictp.it (M.M.)
[2] Kavli Institute for Systems Neuroscience and Centre for Neural Computation, Norges Teknisk-Naturvitenskapelige Universitet (NTNU), Olav Kyrres Gate 9, 7030 Trondheim, Norway
[3] Department of Physics and Astronomy, University of Pennsylvania, 209 South 33rd Street, Philadelphia, PA 19104-6396, USA
[4] Capital Fund Management, 23 rue de l'Université, 75007 Paris, France; iacopomas@gmail.com
[5] Istituto Nazionale di Fisica Nucleare (INFN) Sezione di Trieste, 34100 Trieste, Italy
* Correspondence: claudia.battistin@ntnu.no

Received: 10 August 2018; Accepted: 18 September 2018; Published: 27 September 2018

Abstract: Models can be simple for different reasons: because they yield a simple and computationally efficient interpretation of a generic dataset (e.g., in terms of pairwise dependencies)—as in statistical learning—or because they capture the laws of a specific phenomenon—as e.g., in physics—leading to non-trivial falsifiable predictions. In information theory, the simplicity of a model is quantified by the stochastic complexity, which measures the number of bits needed to encode its parameters. In order to understand how simple models look like, we study the stochastic complexity of spin models with interactions of arbitrary order. We show that bijections within the space of possible interactions preserve the stochastic complexity, which allows to partition the space of all models into equivalence classes. We thus found that the simplicity of a model is not determined by the order of the interactions, but rather by their mutual arrangements. Models where statistical dependencies are localized on non-overlapping groups of few variables are simple, affording predictions on independencies that are easy to falsify. On the contrary, fully connected pairwise models, which are often used in statistical learning, appear to be highly complex, because of their extended set of interactions, and they are hard to falsify.

Keywords: statistical inference; model complexity; minimum description length; spin models

1. Introduction

Science, as the endeavour of reducing complex phenomena to simple principles and models, has been instrumental to solve practical problems. Yet, problems such as image or speech recognition and language translation have shown that Big Data can solve problems without necessarily understanding them [1–3]. A statistical model trained on a sufficiently large number of instances can learn how to mimic the performance of the human brain on these tasks [4,5]. These models are simple in the sense that they are easy to evaluate, train, and/or to infer. They offer simple interpretations in terms of low order (typically pairwise) dependencies, which in turn afford an explicit graph theoretical representation [6]. Their aim is not to uncover fundamental laws but to "generalize well", i.e., to describe well yet unseen data. For this reason, machine learning relies on "universal" models that are apt to describe any possible data on which they can be trained [7], using suitable "regularization" schemes in order to tame parameter fluctuations (overfitting) and achieve small generalization error [8]. Scientific models, instead, are the simplest possible descriptions of experimental results. A physical model is a representation of a real system and its structure reflects the laws and symmetries of Nature. It predicts well not because it generalizes well, but rather because it captures essential features of the specific phenomena that it describes. It should depend on few parameters and is designed to provide predictions that are easy to be falsified [9]. For example, Newton's laws of motion are consistent with momentum conservation, a fact that can be checked in scattering experiments.

The intuitive notion of a "simple model" hints at a succinct description, one that requires few bits [10]. The *stochastic complexity* [11], derived within Minimum Description Length (MDL) [12,13], provides a quantitative measure for "counting" the complexity of models in bits. The question this paper addresses is: what are the features of simple models according to MDL, and are they simple in the sense surmised in statistical learning or in physics? In particular, are models with up to pairwise interactions, which are frequently used in statistical learning, simple?

We address this issue in the context of spin models, describing the statistical dependence among n binary variables. There has been a surge of recent interest in the inference of spin models [14] from high dimensional data, most of which was limited to pairwise models. This is partly because pairwise models allow for an intuitive graph representation of statistical dependencies. Most importantly, since the number of k-variable interactions grows as n^k, the number of samples is hardly sufficient to go beyond $k = 2$. For this reason, efforts to go beyond pairwise interactions have mostly focused on low order interactions (e.g., $k = 3$, see [15] and references therein). Reference [16] recently suggested that even for data generated by models with higher order interactions, pairwise models may provide a sufficiently accurate description of the data. Within the class of pairwise models, L1 regularization [17] has proven to be a remarkably efficient heuristic of model selection (but see also [18]).

Here we focus on the exponential family of spin models with interactions of arbitrary order. This class of models assumes a sharp separation between relevant observables and irrelevant ones, the expected value of which is predicted by the model. In this setting, the stochastic complexity [11] computed within MDL coincides with the penalty that, in Bayesian model selection, accounts for model's complexity, under non-informative (Jeffrey's) priors [19].

1.1. The Exponential Family of Spin Models (With Interactions of Arbitrary Order)

Consider n spin variables, $\boldsymbol{s} = (s_1, \ldots, s_n)$, taking values $s_i = \pm 1$ and the set of all product spin operators, $\phi^\mu(\boldsymbol{s}) = \prod_{i\in\mu} s_i$, where μ is any subset of the indices $\{1, \ldots, n\}$. Each operator $\phi^\mu(\boldsymbol{s})$ models the interaction that involves all the spins in the subset μ.

Definition 1. *The probability distribution of* $\boldsymbol{s}$ *under* spin model $\mathcal{M}$ *is defined as:*

$$P(\boldsymbol{s}\,|\,\boldsymbol{g}, \mathcal{M}) = \frac{1}{Z_{\mathcal{M}}(\boldsymbol{g})} e^{\,\sum_{\mu\in\mathcal{M}} g^\mu \phi^\mu(\boldsymbol{s})}\,, \tag{1}$$

$$\text{with} \qquad Z_{\mathcal{M}}(\boldsymbol{g}) = \sum_{\boldsymbol{s}} e^{\sum_{\mu\in\mathcal{M}} g^\mu \phi^\mu(\boldsymbol{s})} \tag{2}$$

being the partition function*, which ensures normalisation. The model* $\mathcal{M}$ *is identified by the set* $\{\phi^\mu(\boldsymbol{s}), \mu \in \mathcal{M}\}$ *of product spin operators* $\phi^\mu(\boldsymbol{s})$ *that it contains.*

Note that, under this definition, we consider interactions of arbitrary order (see Section SM-0 of the Supplementary Material). For instance, for pairwise interaction models, the operators $\phi^\mu(\boldsymbol{s})$ are single spins s_i or product of two spins $s_i s_j$, for $i, j \in \{1, ..., n\}$. The g^μ are the conjugate parameters [20] that modulate the strength of the interaction associated with ϕ^μ.

We remark that the model defined in Equation (1) belongs to the *exponential family of spin models*. In other words, it can be derived as the maximum entropy distributions that are consistent with the requirement that the model reproduces the empirical averages of the operators $\phi^\mu(\boldsymbol{s})$ for all $\mu \in \mathcal{M}$ on a given dataset [21,22]. In other words, empirical averages of $\phi^\mu(\boldsymbol{s})$ are sufficient statistics, i.e., their values are enough to compute the maximum likelihood parameters $\hat{\boldsymbol{g}}$. Therefore the choice of the operators ϕ^μ in $\mathcal{M}$ inherently entails a sharp separation between relevant variables (the sufficient statistics) and irrelevant ones, which may have important consequences in the inference process. For example, if statistical inference assumes pairwise interactions, it might be blind to relevant patterns in the data resulting from higher order interactions. Without prior knowledge, all models $\mathcal{M}$ should be compared. According to MDL and Bayesian model selection (see Section SM-0 of the Supplementary Material), models should be compared on the basis of their maximum (log)likelihood penalized by their complexity. In other words, simple models should be preferred a priori.

1.2. Stochastic Complexity

The complexity of a model can be defined unambiguously within MDL as the number of bits needed to specify a priori the parameters $\hat{\boldsymbol{g}}$ that best describe a dataset $\hat{\boldsymbol{s}} = (\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(N)})$ consisting of N samples independently drawn from the distribution $P(\boldsymbol{s} \,|\, \boldsymbol{g}, \mathcal{M})$ for some unknown $\boldsymbol{g}$ (see Section SM-0 of the Supplementary Material). Asymptotically for $N \to \infty$, for systems of discrete variables, the *MDL complexity* is given by [23,24]:

$$\log \sum_{\hat{\boldsymbol{s}}} P(\hat{\boldsymbol{s}} \,|\, \hat{\boldsymbol{g}}, \mathcal{M}) \simeq \frac{|\mathcal{M}|}{2} \log\left(\frac{N}{2\pi}\right) + c_{\mathcal{M}}. \tag{3}$$

The first term in the right hand, which is the basis of the Bayesian Information Criterion (BIC) [25,26], captures the increase of the complexity with the number $|\mathcal{M}|$ of model's parameters and with the number N of data points. This accounts for the fact that the uncertainty in each parameter $\hat{\boldsymbol{g}}$ decreases with N as $N^{-1/2}$, so its description requires $\sim \frac{1}{2} \log N$ bits. The second term $c_{\mathcal{M}}$ quantifies the statistical dependencies between the parameters, and encodes the intrinsic notion of simplicity we are interested in. The sum of these two terms, in the right hand side of (3), is generally referred as stochastic complexity [11,26]. However, to distinguish these two terms, we will refer to $c_{\mathcal{M}}$ as the *stochastic complexity* and the other as BIC term.

Definition 2. *For models of the exponential family, the* stochastic complexity $c_{\mathcal{M}}$ *in Equation* (3) *is given by [23,24]*

$$c_{\mathcal{M}} = \log \int d\boldsymbol{g} \sqrt{\det \mathbb{J}(\boldsymbol{g})}\,, \tag{4}$$

where $\mathbb{J}(\boldsymbol{g})$ *is the Fisher Information matrix with entries*

$$J_{\mu\nu}(\boldsymbol{g}) = \frac{\partial^2}{\partial g^\mu \partial g^\nu} \log Z_{\mathcal{M}}(\boldsymbol{g}). \tag{5}$$

For the exponential family of models, the MDL criterium (3) coincides with the Bayesian model selection approach, assuming Jeffreys' prior over the parameters $\boldsymbol{g}$ [26–28] (see Section SM-0 of the Supplementary Material). Notice, however, that we take $c_{\mathcal{M}}$ as an information theoretic measure of model complexity, and abstain from entering into the debate on whether Jeffreys' prior is an adequate choice in Bayesian inference (see e.g., [29]).

Within a fully Bayesian approach, the model that maximises its posterior given the data $\hat{\boldsymbol{s}}$, $P(\mathcal{M}|\hat{\boldsymbol{s}})$, is the one to be selected. Therefore, if two models have the same number of parameters (same BIC term), the simplest one, i.e., the one with the lowest stochastic complexity $c_{\mathcal{M}}$, has to be chosen a priori. However, the number of possible interactions ϕ^μ among n spins is $2^n - 1$, and therefore the number of spin models is 2^{2^n-1}. The super-exponential growth of the number of models with the number of spins n makes selecting the best model unfeasible even for moderate n. Our aim is then to understand how the stochastic complexity depends on the structure of the model $\mathcal{M}$ and eventually provide guidelines for the search of simpler models in such a huge space.

2. Results

2.1. Gauge Transformations

Let's start by showing that low order interactions do not have a privileged status and are not necessarily related to low complexity $c_{\mathcal{M}}$, with the following argument: Alice is interested in finding which model $\mathcal{M}$ best describes a dataset $\hat{\mathbf{s}}$; Bob is interested in the same problem, but his dataset $\hat{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}^{(1)}, \ldots, \boldsymbol{\sigma}^{(N)})$ is related to Alice's dataset by a *gauge transformation*.

Definition 3. *We define a* gauge transformation *as a bijective transformation between n spin variables* $\mathbf{s} = (s_1, \cdots, s_n) \in \{\pm 1\}^n$ *and n spin variables* $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_n) \in \{\pm 1\}^n$ *that corresponds to a bijection from the set of all operators to itself,* $\phi^{\mu}(\mathbf{s}) \to \phi^{\mu'}(\boldsymbol{\sigma})$ *(see Section SM-1 of the Supplementary Material). Gauge transformations preserve the structure of the set of all operators. For examples of gauge transformations see Figure 1.*

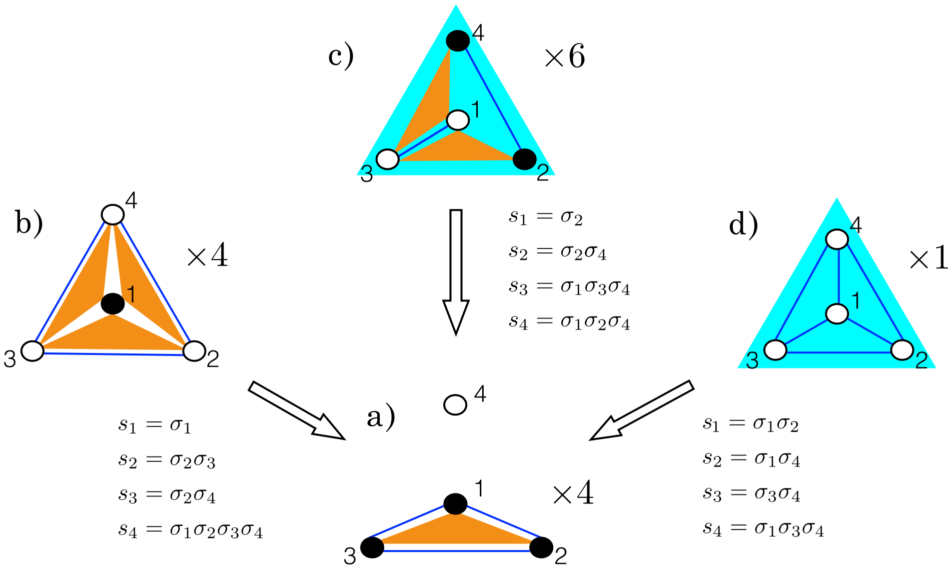

Figure 1. Example of gauge transformations between models with $n = 4$ spins. Models are represented by diagrams (color online): spins are full dots • in presence of a local field, empty dots ○ otherwise; blue lines are pairwise interactions ($\phi^{\mu} = s_i s_j$); orange triangles denote 3-spin interactions ($\phi^{\mu} = s_i s_j s_k$); and the 4-spin interaction ($s_1 s_2 s_3 s_4$) is a filled blue triangle. Note that model a) has all its interactions grouped on 3 spins; the gauge transformations leading to this model are shown along the arrows. All the models belong to the same complexity class, with $|\mathcal{M}| = 7$, $\lambda = 4$ and a number of independent operators $n_{\mathcal{M}} = 3$ (e.g., s_1, s_2 and s_3 in model a)—see tables in Section SM-6 of the Supplementary Material. The class contains in total 15 models, which are grouped, with respect to the permutation of the spins, behind the 4 representatives shown here with their multiplicity ($\times m$). Models a), b), c) and d) are respectively identified by the following sets of operators: (**a**) $\mathcal{M} = \{s_1, s_2, s_3, s_1s_2, s_1s_3, s_2s_3, s_1s_2s_3\}$; (**b**) $\mathcal{M} = \{s_1, s_2s_3, s_3s_4, s_4s_2, s_1s_2s_3, s_1s_3s_4, s_1s_2s_4\}$; (**c**) $\mathcal{M} = \{s_2, s_4, s_2s_4, s_1s_3, s_1s_2s_3, s_1s_3s_4, s_1s_2s_3s_4\}$; and (**d**) $\mathcal{M} = \{s_1s_2, s_1s_3, s_1s_4, s_2s_3, s_2s_4, s_3s_4, s_1s_2s_3s_4\}$.

This gauge transformation induces a bijective transformation between Alice's models and those of Bob, as shown in Figure 1, that preserves the number of interactions $|\mathcal{M}|$. Whatever conclusion Bob draws on the relative likelihood of models can be translated into Alice's world, where it has to coincide with Alice's result. It follows that two models, $\mathcal{M}$ and $\mathcal{M}'$, related by a gauge transformation, must also have the same complexity $c_{\mathcal{M}} = c_{\mathcal{M}'}$. In particular, pairwise interactions can be mapped to interactions of any order (see Figure 1), and, consequently, low order interactions are not necessarily simpler than higher order ones.

Proposition 1. *Two models related by a gauge transformation have the same complexity.*

Proposition 2. *The stochastic complexity $c_{\mathcal{M}}$ of a model is not defined by the order of its interactions. Models with low order interactions don't necessarily have a low complexity, and, reciprocally, high order interactions don't necessarily imply large complexity.*

Observe that models connected by gauge transformations have remarkably different structures. In Figure 1, model a) has all the possible interactions concentrated on 3 spins, having the properties of a simplicial complex [30,31]; however, its gauge-transformed counterparties are not simplicial complexes. Model d) is invariant under any permutations of the four spins, whereas the other models have a lower degree of symmetry under permutations (see the different multiplicities in Figure 1).

Gauge transformations are discussed in more details in Section SM-1 of the Supplementary Material. One can also see them as a change of the basis $\boldsymbol{s} \to \boldsymbol{\sigma}$ in which the operators are expressed. Counting the number of possible bases then gives us the number of gauge transformations (see Section SM-1 of the Supplementary Material).

Proposition 3. *The total number of gauge transformations for a system of n spin variables is:*

$$\mathcal{N}_{GT}(n) = 2^{n^2} \prod_{k=1}^{n} \left(1 - 2^{-k}\right). \tag{6}$$

Notice that the number of gauge transformations, (6) is much smaller than the number $2^n!$ of possible bijections of the set of 2^n states into itself. Indeed, a generic bijection between the state spaces of $\boldsymbol{s}$ and $\boldsymbol{\sigma}$ maps each product operator to one of the binary functions $f : \boldsymbol{\sigma} \to \{+1, -1\}$, which does not necessarily correspond to a product operator $\phi^{\mu}(\boldsymbol{\sigma})$.

2.2. Complexity Classes

Gauge transformations define equivalence relations, which partition the set of all models into equivalence classes. Models belonging to the same class are related to each other by a gauge transformation, and thus, from Proposition 1, have the same complexity $c_{\mathcal{M}}$, which leads us to introduce the notion of *complexity classes*.

Definition 4. *A* complexity class *is an equivalence class of models defined by gauge transformations.*

This classification suggests the presence of "quantum numbers" (invariants), in terms of which models can be classified. These invariants emerge explicitly when writing the cluster expansion of the partition function [32–34] (see Section SM-2 of the Supplementary Material):

$$Z_{\mathcal{M}}(\boldsymbol{g}) = 2^n \left(\prod_{\mu \in \mathcal{M}} \cosh(g^{\mu}) \right) \sum_{\ell \in \mathcal{L}} \prod_{\mu \in \ell} \tanh(g^{\mu}). \tag{7}$$

The sum runs on the set $\mathcal{L}$ of all possible *loops* ℓ that can be formed with the operators $\mu \in \mathcal{M}$, including the empty loop $\ell = \emptyset$.

Definition 5. *A* loop *is any subset $\ell \subseteq \mathcal{M}$ such that $\prod_{\mu \in \ell} \phi^{\mu}(\boldsymbol{s}) = 1$ for any value of $\boldsymbol{s}$, i.e., such that each spin s_i occurs zero or an even number of times in this product.*

The structure of $Z_{\mathcal{M}}(\boldsymbol{g})$ in (7) depends on few characteristics of the model $\mathcal{M}$: The number $|\mathcal{M}|$ of operators (or, equivalently, of parameters) and the structure of its set of loops $\mathcal{L}$ (which operator is involved in which loop). The invariance under gauge transformation of the complexity in (4) reveals itself in the fact that the partition function of models related by a gauge transformation have the same functional dependence on their parameters up to relabeling.

Let us focus on the loop structure of models belonging to the same class. The set $\mathcal{L}$ of loops of any model $\mathcal{M}$ has the structure of a finite Abelian group: if $\ell_1, \ell_2 \in \mathcal{L}$, then $\ell_1 \oplus \ell_2$ is also a loop of $\mathcal{M}$, where $\oplus$ is the symmetric difference [35] of two sets (see Section SM-3 of the Supplementary Material). As a consequence, for each model $\mathcal{M}$ one can identify a *minimal generating set* of λ loops, such that any loop in $\mathcal{L}$ can be uniquely expressed as the symmetric difference of loops in the minimal generating set. Note that the choice of the generating set is not unique, though all choices have the same cardinality λ; Figure 2 gives examples of this decomposition for the models of Figure 1. Note also that $\ell \oplus \ell = \varnothing$ for each loop $\ell \in \mathcal{L}$. As a consequence, the cardinality of the loop group is $|\mathcal{L}| = 2^\lambda$ (including the empty loop $\varnothing$). We found that λ is related to the number $|\mathcal{M}|$ of operators of the model by $\lambda = |\mathcal{M}| - n_\mathcal{M}$ (see Section SM-3 of the Supplementary Material), where $n_\mathcal{M}$ is the number of *independent operators* of a model $\mathcal{M}$, i.e., the maximal number of operators that can be taken in $\mathcal{M}$ without forming any loop. This implies that λ attains its minimal value, $\lambda = 0$, for models with only independent operators ($|\mathcal{M}| = n_\mathcal{M}$), and its maximal value, $\lambda = 2^n - 1 - n$, for the *complete model* $\overline{\mathcal{M}}$, that contains all the $|\overline{\mathcal{M}}| = 2^n - 1$ possible operators. The number of independent operators, $n_\mathcal{M}$, is preserved by gauge transformation, and, as the total number of operators, $|\mathcal{M}|$, is also an invariant of the class, so is the cardinality of the minimal generating set λ. For example, all models in Figure 1 have $n_\mathcal{M} = 3$ independent operators and $\lambda = 4$ (see Figure 2). It can also be shown that gauge transformations imply a duality relation, that associates to each class of models with $|\mathcal{M}|$ operators a class of models with the $2^n - 1 - |\mathcal{M}|$ complementary operators (see Section SM-3 of the Supplementary Material).

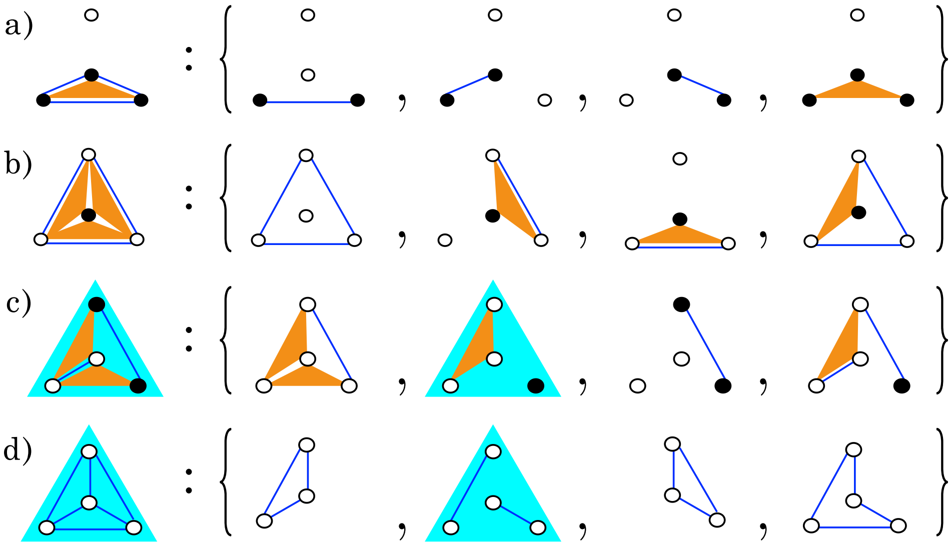

Figure 2. Example of a minimal generating set of loops for each model of Figure 1. As these models belong to the same class, their (respective) sets of loops have the same cardinality, 2^λ, where $\lambda = 4$ is the number of generators (as shown here). For model a), one can easily check that the 4 loops of the set are independent, as each of them contains at least one operator that doesn't appear in the other 3 loops (see Section SM-3 of the Supplementary Material). Within each column, on the r.h.s., loops are related by the same gauge transformation morphing models into one another on the l.h.s. (i.e., the transformations displayed in Figure 1). This shows that the loops of these 4 generating sets have the same structure, which implies that the loop structure of the 4 models is the same. Any loop of a model can finally be obtained by combining a subset of its generating loops. Note that the choice of the generating set is not unique. (**a**–**d**) refer to the same models as in Figure 1.

To summarize, the distinctive features of a complexity class are given in the following Proposition.

Proposition 4. *The number of operators, $|\mathcal{M}|$, the number of independent operators, $n_\mathcal{M}$, and the structure of the set of loops, $\mathcal{L}$ (through its generators), fully characterize a complexity class.*

3. Discussion: How Do Simple Models Look Like?

3.1. Fewer Independent Operators, Shorter Loops

Coming to the quantitative estimate of the complexity, $c_{\mathcal{M}}$ generally depends on the extent to which ensemble averages of the operators $\phi^{\mu}(\boldsymbol{s})$ in the model $\mu \in \mathcal{M}$ constrain each other. This appears explicitly by rewriting (4) as an integral over the ensemble averages of the operators, $\boldsymbol{\varphi} = \{\langle \phi^{\mu} \rangle, \mu \in \mathcal{M}\}$, exploiting the bijection between the parameters $\boldsymbol{g}$ and their dual parameters $\boldsymbol{\varphi}$ and re-parameterization invariance [28,36]:

$$c_{\mathcal{M}} = \log \int_{\mathcal{F}} \mathrm{d}\boldsymbol{\varphi} \sqrt{\det \mathbb{J}(\boldsymbol{\varphi})}\,, \tag{8}$$

where $\mathbb{J}(\boldsymbol{\varphi})$ is the Fisher Information Matrix in the $\boldsymbol{\varphi}$-coordinates. The new domain, $\mathcal{F}$, of integration is over the values of $\boldsymbol{\varphi}$ that can be realized in any empirical sample drawn from the model $\mathcal{M}$ (known in this context as *marginal polytope* [37,38]) and is related to the mutual constraints between the ensemble averages, φ^{μ}, (see Section SM-4 of the Supplementary Material for more details).

Proposition 5. *The complexity of a model without loops, i.e.,* $\mathcal{L} = \{\emptyset\}$*, and* $n_{\mathcal{M}} = |\mathcal{M}|$ *independent operators is* $c_{\mathcal{M}} = |\mathcal{M}| \log \pi$.

Indeed, if the model contains no loop, then $J_{\mu\nu}(\boldsymbol{\varphi}) = [1 - (\varphi^{\mu})^2]^{-1}\delta_{\mu\nu}$ is diagonal: The integral in (8) factorizes and Proposition 5 follows. In this case, the variables φ^{μ} are not constrained at all and the domain of integration is $\mathcal{F} = [-1,1]^{|\mathcal{M}|}$. If instead the model contains loops, the variables φ^{μ} become constrained and the marginal polytope, $\mathcal{F}$, is reduced. For example, for a model with a single loop of length three (e.g., $\phi^1 = s_1$, $\phi^2 = s_2$, and $\phi^3 = s_1 s_2$), the values of $\boldsymbol{\varphi}$ in $[-1,1]^3$ are not all attainable, indeed $\mathcal{F} = \{\boldsymbol{\varphi} \in [-1,1]^3 : \ |\varphi^1 + \varphi^2| - 1 \le \varphi^3 \le 1 - |\varphi^1 - \varphi^2|\}$ is reduced, which decreases the complexity.

Proposition 6. *The complexity of models with a fixed number of operators and a single (non-empty) loop increases with the length of the loop.*

The complexity, $c_{\mathcal{M}}(k)$, of models with a fixed number, $|\mathcal{M}|$, of parameters and a single (non-empty) loop of length k is shown in Figure 3 (see Section SM-6 of the Supplementary Material): $c_{\mathcal{M}}(k)$ increases with k and saturates at $|\mathcal{M}| \log \pi$, which is the value one would expect if all operators were unconstrained. This is consistent with the expectation that longer loops induce weaker constraints among the operators. Note that the number of independent operators is kept constant here, equal to $n_{\mathcal{M}} = |\mathcal{M}| - 1$.

Proposition 7. *At fixed number of operators, the complexity of a model increases with the number of independent operators.*

The single loop calculation allows computing the complexity of models with non-overlapping loops ($\ell \cap \ell' = \emptyset$ for all $\ell, \ell' \in \mathcal{L}$), for which $c_{\mathcal{M}} = \sum_{\ell \in \mathcal{L}} c_{\ell}$ is the sum over the complexity, c_{ℓ}, associated to each loop. In the general case of models with more complex loop structures, the explicit calculation of $c_{\mathcal{M}}$ is non-trivial. Yet, the argument above suggests that, at fixed number of parameters $|\mathcal{M}|$, $c_{\mathcal{M}}$ should increase with the number $n_{\mathcal{M}}$ of independent operators. Figure 4 summarises the results for all models with $n = 4$ spins and supports this conclusion: For a given value of $|\mathcal{M}|$, classes with lower values of $n_{\mathcal{M}}$ (i.e., with less independent operators) are less complex.

Proposition 8. *At fixed number of operators, complete models on a subset of spins and their equivalents are the simplest models. We refer to these models as* sub-complete models. *Classes of sub-complete models are the* classes of minimal complexity.

A surprising result of Figure 4 is that $c_{\mathcal{M}}$ is not monotonic with the number, $|\mathcal{M}|$, of operators of the model, increasing first with $|\mathcal{M}|$ and then decreasing. Complete models $\overline{\mathcal{M}}$ turn out to be the simplest (see the dashed curve in Figure 4). As a consequence, for a given $|\mathcal{M}|$, models that contain a complete model on a subset of spins are generally simpler than models where operators have support on all the spins. For instance, the complexity class displayed in Figure 1 is the class of models with $|\mathcal{M}| = 7$ operators that has the lowest complexity (see green triangle on the dashed curve in Figure 4).

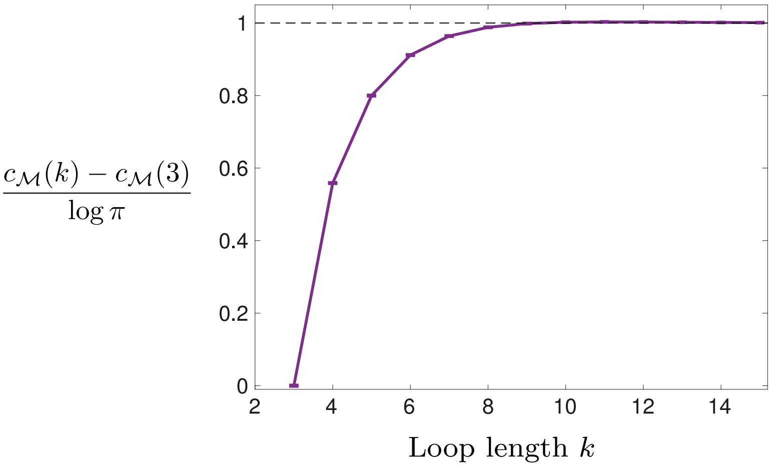

Figure 3. Complexity, $c_{\mathcal{M}}(k)$, of models with a single loop of length k, and $|\mathcal{M}| - k$ *free* operators, i.e., not involved in any loop. For $k = 3$, $c_{\mathcal{M}}(3) = (|\mathcal{M}| - 1) \log \pi$ can be computed analytically from (4). Values of $c_{\mathcal{M}}$ are averaged over 10^3 numerical estimates of the integral in (4), using 10^6 Monte Carlo samples each. Error bars correspond to their standard deviation.

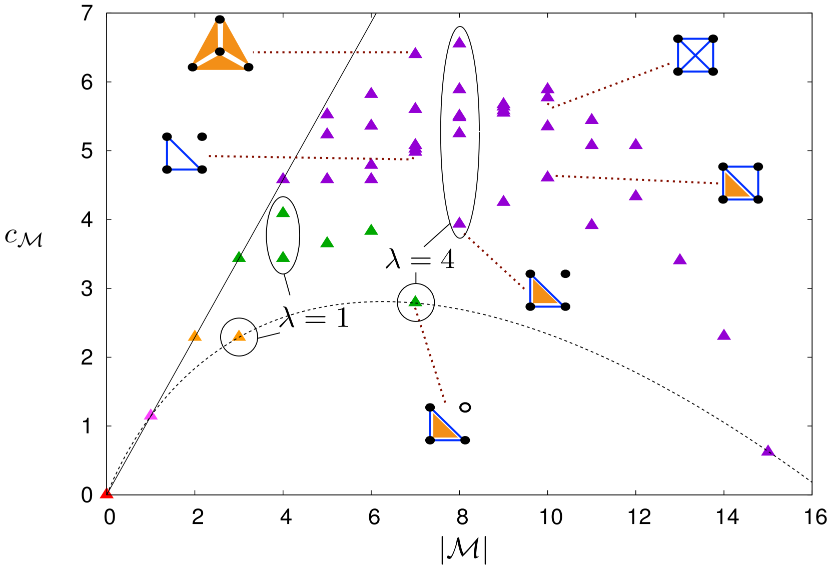

Figure 4. (color online) Complexity of models for $n = 4$ as a function of the number $|\mathcal{M}|$ of operators: Each triangle represents a class of complexity, which contains one or more models (see Section SM-6 of the Supplementary Material). For each class, the value of $c_{\mathcal{M}}$ was obtained from a representative of the class; some of them are shown here with their corresponding diagram (same notation as in Figure 1). The triangle colors indicate the values of $n_{\mathcal{M}}$: Violet for $n_{\mathcal{M}} = 4$, green for 3, yellow for 2, pink for 1, and red for 0 (model with no operator). Models on the black line have only independent operators ($|\mathcal{M}| = n_{\mathcal{M}}$) and complexity, $c_{\mathcal{M}} = |\mathcal{M}| \log \pi$; models on the dashed curve are complete models, the complexities of which are given in (9). Complexity classes with the same values of $|\mathcal{M}|$ and $n_{\mathcal{M}}$ have the same value of $\lambda = |\mathcal{M}| - n_{\mathcal{M}}$, i.e., the same number of loops, $|\mathcal{L}|$, but a different loop structure.

Figure 4 also confirms that pairwise models are not simpler than models with higher order interactions. Indeed, for instance for $|\mathcal{M}| = 7$, $c_{\mathcal{M}}$ increases drastically when changing model a) of Figure 1 into a pairwise model by turning the 3-spin interaction into an external field acting on s_4. Likewise, the model with all six pairwise interactions for $|\mathcal{M}| = 10$ is more complex than the one where one of them is turned into a 3-spin interaction.

3.2. Complete and Sub-Complete Models

It is possible to compute explicitly the complexity of a complete model, $\overline{\mathcal{M}}$, with n spins using the mapping $g^\mu = 2^{-n} \sum_{s} \phi^\mu(s) \log p(s)$ between the $2^n - 1$ parameters g^μ of $\overline{\mathcal{M}}$ and the 2^n probabilities $p(s)$ constrained by their normalization [39]. The complexity in (4) being invariant under reparametrization [36], one can easily re-write this integral in terms of the variables $p(s)$. Finally, using that $\det \mathbb{J}(\mathbf{p}) = \prod_s 1/p(s)$, we obtain the following proposition (see Section SM-5 of the Supplementary Material).

Proposition 9. *The complexity of a complete model* $\overline{\mathcal{M}}$ *with n spins is:*

$$\begin{aligned} c_{\overline{\mathcal{M}}} &= \log \int_0^1 d\boldsymbol{p}\, \delta\left(\sum_s p(s) - 1\right) \prod_s \frac{1}{\sqrt{p(s)}}\,, \\ &= 2^{n-1} \log \pi - \log \Gamma(2^{n-1})\,. \end{aligned} \tag{9}$$

Note that, for $n > 4$, $c_{\overline{\mathcal{M}}}$ becomes negative (for $n = 6$, $c_{\overline{\mathcal{M}}} \simeq -41.5$). This suggests that the class of least complex models with $|\mathcal{M}|$ interactions is the one that contains the model where the maximal number of loops are concentrated on the smallest number of spins. This agrees with our previous observations on single loop models and sub-complete models. On the contrary, models where interactions are distributed uniformly across the variables (e.g., models with only single spin operators for $n \geq |\mathcal{M}|$ or with non-overlapping sets of loops) have higher complexity.

We finally note that complete models are also extremely simple to infer from empirical data. Indeed, the maximum likelihood estimates of the parameter are trivially given by $\hat{g}^\mu = 2^{-n} \sum_s \phi^\mu(s) \log \hat{p}(s)$, where $\hat{p}(s)$ is the empirical distribution. By contrast, learning the parameters of pairwise interacting models can be a daunting task [14].

3.3. Maximally Overlapping Loops

This finally leads us to conjecture that stochastic complexity is related to the localization properties of the set of loops $\mathcal{L}$ (i.e., its group structure) rather than to the order of the interactions: Models where the loops, $\ell, \ell' \in \mathcal{L}$, have a "large" overlap, $\ell \cap \ell'$, are simple, whereas models with an extended homogeneous network of interactions (e.g., fully connected Ising models with up-to pairwise interaction) have many non-overlapping loops, $\ell \cap \ell' = \emptyset$, and therefore are rather complex. It is interesting to note that the former (simple models) lend themselves to predictions on the independence of different groups of spins. These predictions suggest "fundamental" properties of the system under study (i.e., invariance properties, spin permutation symmetry breaking) and are easy to falsify (i.e., it is clear how to devise a statistical test for these hypotheses to any given confidence level). On the contrary, complex models (e.g., fully connected pairwise Ising models) are harder to falsify as their parameters can be adjusted to fit reasonably well any sample, irrespectively of the system under study.

3.4. Summary

We find that at fixed number, $|\mathcal{M}|$, of operators, simpler models are those with fewer independent operators (i.e., smaller $n_{\mathcal{M}}$). For the same value of $n_{\mathcal{M}}$, models can still have different complexities. The simpler ones are then those with a loop structure that will impose the most constraints between the operators of the model. More generally, we show that the complexity of a model is not defined by the order of the interactions involved, but is, instead, intimately connected to its internal geometry, i.e., how interactions are arranged in the model. The geometry of this arrangement implies mutual dependencies between interactions that constrain the states accessible to the system. More complex models are those that implement fewer constraints, and can thus account for broader types of data. This result is consistent with the information geometric approach of Reference [26], which studies model complexity in terms of the geometry of the space of probability distributions [40]. The contribution of this paper clarifies the relation between the information geometric point of view and the specific structure of the model, i.e., the actual arrangement of its interactions. We remark that these results apply to non-degenerate spin models. In the broader class of degenerate models [20], arrangement of the operators and degeneracy of the parameters may interact non trivially in terms of complexity. Our preliminary numerical investigation of single loop models (see Section SM-7 of the Supplementary Material) indicates that degeneracy decreases complexity by constraining the parameters and therefore the statistics of the operators. Also, Reference [41] discusses model selection on mixture models of spin variables, and shows that they can be cast in the form of Equation (1), where g^{μ} are subject to linear constraints. For these models, Bayesian model selection can be performed exactly without resorting to the approach discussed here.

A rough estimate of the number, N, of data samples beyond which the complexity term becomes negligible in Bayesian inference can be obtained with the following argument: An upper bound for the complexity of models with n spins and m parameters is given by $m \log \pi$, i.e., when all operators are independent. As a lower bound, we take Equation (9) with $m = 2^n - 1$. This implies that an upper bound for the variation of the complexity is given by $\Delta c = \frac{m-1}{2} \log \pi + \log \Gamma \left(\frac{m+1}{2} \right)$. When this is much smaller than the BIC term, the stochastic complexity can be neglected. For large m this implies $N \gg m$, which may be relevant for the applicability of fully connected pairwise models ($m \simeq n^2/2$) in typical cases, for instance when samples cannot be considered as independent observations from a stationary distribution (see [18]).

4. Conclusions

As pointed out by Wigner [42] long ago, the unreasonable effectiveness of mathematical models relies on isolating phenomena that depend on few variables, whose mutual variation is described by simple models and is independent of the rest. Remarkably we find that, for a fixed number of spin variables and parameters, simple models, according to MDL, are precisely of this form: Statistical dependencies are concentrated on the smallest subset of variables and these are independent of all the rest. Such simple models are not optimal for generalizing, i.e., to describe generic statistical dependencies, rather they are easy to falsify. They are designed for spotting independencies that may hint at deeper principles (e.g., symmetries or conservation laws) that may "take us beyond the data" [43], meaning that they can hint at hypotheses (on e.g., symmetries or other regularities) that can be tested in future experiments. On the contrary, fully connected pairwise models, which provide simple interpretation of statistical dependencies in terms of direct interactions, appears to be rather complex. This, we conjecture, is the origin of pairwise sufficiency [16] that makes them so successful to describe a wide variety of data from neural tissues [44] to voting behaviour [45].

Our results, however, show that any model that can be obtained from a pairwise model via a gauge transformation has the same complexity and hence the same generalisation power, but has higher order interactions. Hence, gauge transformations can be used to compare pairwise models with models in the same complexity class, in order to quantitatively assess when a dataset is genuinely described by pairwise interactions. Notice that this comparison can be done directly on the basis of their maximum likelihood alone.

This is only one of the possible applications of gauge transformations, which are one of the main results of this paper. In loose words, these transformations preserve the topology of models, i.e., the manner in which interactions are mutually arranged, but change the "basis" of the operators that embody these interactions. Besides model selection within the same complexity class, as in the example of pairwise models above, we can think of selecting the appropriate complexity class according to the availability of data. One particularly interesting avenue of future research is to perform model selection among models of minimal complexity (i.e., sub-complete models). Model comparison between and within these classes may be relevant given the high degree of clustering and modularity in neural, social, metabolic, and regulatory networks [46]. These models offer the simplest possible explanation of a dataset, not necessarily the most accurate one (e.g., in terms of generalisation error), but the one that can potentially reveal regularities and symmetries in the data. Interestingly, parameter inference for these models is also remarkably simple.

In conclusion, when data are scarce and high dimensional, Bayesian inference should privilege simple models, i.e., those with small stochastic complexity, over more complex ones, such as fully connected pairwise models that are often used [14,44,45]. A full Bayesian model selection approach is hampered by the calculation of the stochastic complexity that is a daunting task. Developing approximate heuristics for accomplishing this task is a challenging future avenue of research.

Supplementary Materials: The Supplementary Material is available online at http://www.mdpi.com/1099-4300/20/10/739/s1.

Author Contributions: Conceptualization, I.M. and M.M.; Formal analysis, A.B., C.B. and C.d.M.; Methodology, A.B., C.B. and C.d.M.; Software, A.B.; Writing-original draft, C.B., C.d.M. and M.M.

Funding: C.B. acknowledges financial support from the Kavli Foundation and the Norwegian Research Council's Centre of Excellence scheme (Centre for Neural Computation, grant number 223262). A.B. acknowledges financial support from International School for Advanced Studies (SISSA).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References and Notes

1. Mayer-Schonberger, V.; Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work and Think*; John Murray Publishers: London, UK, 2013.
2. Anderson, C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, 2008. Wired. Available online: https://www.wired.com/2008/06/pb-theory/(accessed on 20 September 2018)
3. Cristianini, N. Are we there yet? *Neural Netw.* **2010**, *23*, 466–470. [CrossRef] [PubMed]
4. LeCun, Y.; Kavukcuoglu, K.; Farabet, C. Convolutional networks and applications in vision. In Proceedings of the 2010 IEEE International Symposium on Circuits and Systems, Paris, France, 30 May–2 June 2010; pp. 253–256.
5. Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenge, R.; Satheesh, S.; Sengupta, S.; Coates, A.; Ng, A. Deep Speech: Scaling up end-to-end speech recognition. *arXiv* **2014**, arXiv:cs.CL/1412.5567.
6. Bishop, C. *Pattern Recognition and Machine Learning*; (Information Science and Statistics); Springer-Verlag: New York, NY, USA, 2006.
7. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [CrossRef]
8. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: http://www.deeplearningbook.org (accessed on 20 September 2018).
9. Popper, K. *The Logic of Scientific Discovery (Routledge Classics)*; Taylor & Francis: London, UK, 2002.
10. Chater, N.; Vitányi, P. Simplicity: A unifying principle in cognitive science? *Trends Cogn. Sci.* **2003**, *7*, 19–22. [CrossRef]
11. Rissanen, J. Stochastic complexity in learning. *J. Comput. Syst. Sci.* **1997**, *55*, 89–95. [CrossRef]
12. Rissanen, J. Modeling by shortest data description. *Automatic* **1978**, *14*, 465–471. [CrossRef]
13. Grünwald, P. *The Minimum Description Length Principle*; (Adaptive Computation and Machine Learning); MIT Press: Cambridge, MA, USA, 2007.
14. Chau Nguyen, H.; Zecchina, R.; Berg, J. Inverse statistical problems: From the inverse Ising problem to data science. *arXiv* **2017**, arXiv:cond-mat.dis-nn/1702.01522.

15. Margolin, A.; Wang, K.; Califano, A.; Nemenman, I. Multivariate dependence and genetic networks inference. *IET Syst. Biol.* **2010**, *4*, 428–440. [CrossRef] [PubMed]
16. Merchan, L.; Nemenman, I. On the Sufficiency of Pairwise Interactions in Maximum Entropy Models of Networks. *J. Stat. Phys.* **2016**, *162*, 1294–1308. [CrossRef]
17. Ravikumar, P.; Wainwright, M.J.; Lafferty, J.D. High-dimensional Ising model selection using $\ell 1$-regularized logistic regression. *Ann. Stat.* **2010**, *38*, 1287–1319. [CrossRef]
18. Bulso, N.; Marsili, M.; Roudi, Y. Sparse model selection in the highly under-sampled regime. *J. Stat. Mech. Theor. Exp.* **2016**, *2016*, 093404. [CrossRef]
19. Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput.* **1997**, *9*, 349–368. [CrossRef]
20. There is a broader class of models, where subsets $\mathcal{V} \subseteq \mathcal{M}$ of operators have the same parameter, i.e., $g^{\mu} = g^{\mathcal{V}}$ for all $\mu \in \mathcal{V}$ or g^{μ} are subject to linear constrains. These *degenerate models* are rarely considered in the inference literature. Here we confine our discussion to non-degenerate models and refer the reader to Section SM-7 of the Supplementary Material for more discussion.
21. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
22. Tikochinsky, Y.; Tishby, N.Z.; Levine, R.D. Alternative approach to maximum-entropy inference. *Phys. Rev. A* **1984**, *30*, 2638–2644. [CrossRef]
23. Rissanen, J.J. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **1996**, *42*, 40–47. [CrossRef]
24. Rissanen, J. Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Trans. Inf. Theo.* **2001**, *47*, 1712–1717. [CrossRef]
25. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
26. Myung, I.J.; Balasubramanian, V.; Pitt, M.A. Counting probability distributions: Differential geometry and model selection. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 11170–11175. [CrossRef] [PubMed]
27. Jeffreys, H. An Invariant Form for the Prior Probability in Estimation Problems. *Proc. R. Soc. Lond. A Math. Phys. Eng. Sci.* **1946**, *186*, 453–461. [CrossRef]
28. Amari, S. *Information Geometry and Its Applications*; (Applied Mathematical Sciences); Springer: Tokyo, Japan, 2016.
29. Kass, R.E.; Wasserman, L. The selection of prior distributions by formal rules. *J. Am. Stat. Assoc.* **1996**, *91*, 1343–1370. [CrossRef]
30. A *simplicial complex* [31], in our notation, is a model such that, for any interaction $\mu \in \mathcal{M}$, any interaction that involves any subset $\nu \subseteq \mu$ of spins is also contained in the model (i.e., $\nu \in \mathcal{M}$).
31. Courtney, O.T.; Bianconi, G. Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. *Phys. Rev. E* **2016**, *93*, 062311. [CrossRef] [PubMed]
32. Landau, L.; Lifshitz, E. *Statistical Physics*, 3rd ed.; Elsevier Science: London, UK, 2013.
33. Kramers, H.A.; Wannier, G.H. Statistics of the Two-Dimensional Ferromagnet. Part II. *Phys. Rev.* **1941**, *60*, 263–276. [CrossRef]
34. Pelizzola, A. Cluster variation method in statistical Physics and probabilistic graphical models. *J. Phys. A Math. Gen.* **2005**, *38*, R309–R339. [CrossRef]
35. The *symmetric difference* of two sets ℓ_1 and ℓ_2 is defined as the set that contains the elements that occur in ℓ_1 but not in ℓ_2 and viceversa: $\ell_1 \oplus \ell_2 = (\ell_1 \cup \ell_2) \setminus (\ell_1 \cap \ell_2)$. It corresponds to the XOR operator between the operators of the two loops.
36. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; (Translations of mathematical monographs); American Mathematical Society: Providence, RI, USA, 2007.
37. Wainwright, M.J.; Jordan, M.I. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends® Mach. Learn.* **2008**, *1*, 1–305. [CrossRef]
38. Wainwright, M.J.; Jordan, M.I. Variational inference in graphical models: The view from the marginal polytope. In Proceedings of the Forty-First Annual Allerton Conference on Communication, Control, and Computing, Monticello, NY, USA, 1–3 October 2003; Volume 41, pp. 961–971.
39. Mastromatteo, I. On the typical properties of inverse problems in statistical mechanics. *arXiv* **2013**, arXiv: cond-mat.stat-mech/1311.0190.
40. In information geometry [28,36], a model $\mathcal{M}$ defines a manifold in the space of probability distributions. For exponential models (1), the natural metric, in the coordinates g^{μ}, is given by the Fisher Information (5), and the stochastic complexity (4) is the volume of the manifold [26].

41. Gresele, L.; Marsili, M. On maximum entropy and inference. *Entropy* **2017**, *19*, 642. [CrossRef]
42. Wigner, E.P. The unreasonable effectiveness of mathematics in the natural sciences. *Commun. Pure Appl. Math.* **1960**, *13*, 1–14. [CrossRef]
43. In his response to Reference [2] on `edge.org`, W.D. Willis observes that "Models are interesting precisely because they can take us beyond the data".
44. Schneidman, E.; Berry, M.J., II; Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **2006**, *440*, 1007–1012. [CrossRef] [PubMed]
45. Lee, E.; Broedersz, C.; Bialek, W. Statistical mechanics of the US Supreme Court. *J. Stat. Phys.* **2015**, *160*, 275–301. [CrossRef]
46. Albert, R.; Barabási, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47–97. [CrossRef]