J Stat Phys (2015) 161:598–632
DOI 10.1007/s10955-015-1341-7

Learning Probabilities From Random Observables in High Dimensions: The Maximum Entropy Distribution and Others

Tomoyuki Obuchi[1] · **Simona Cocco**[2] · **Rémi Monasson**[2]

Received: 12 March 2015 / Accepted: 21 July 2015 / Published online: 4 August 2015

Abstract We consider the problem of learning a target probability distribution over a set of N binary variables from the knowledge of the expectation values (with this target distribution) of M observables, drawn uniformly at random. The space of all probability distributions compatible with these M expectation values within some fixed accuracy, called version space, is studied. We introduce a biased measure over the version space, which gives a boost increasing exponentially with the entropy of the distributions and with an arbitrary inverse 'temperature' Γ. The choice of Γ allows us to interpolate smoothly between the unbiased measure over all distributions in the version space ($\Gamma = 0$) and the pointwise measure concentrated at the maximum entropy distribution ($\Gamma \to \infty$). Using the replica method we compute the volume of the version space and other quantities of interest, such as the distance R between the target distribution and the center-of-mass distribution over the version space, as functions of $\alpha = (\log M)/N$ and Γ for large N. Phase transitions at critical values of α are found, corresponding to qualitative improvements in the learning of the target distribution and to the decrease of the distance R. However, for fixed α, the distance R does not vary with Γ, which means that the maximum entropy distribution is not closer to the target distribution than any other distribution compatible with the observable values. Our results are confirmed by Monte Carlo sampling of the version space for small system sizes ($N \leq 10$).

Keywords Probabilistic inference · Maximum entropy principle · Replica method

✉ Tomoyuki Obuchi
obuchi@sp.dis.titech.ac.jp

[1] Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama, Japan

[2] Laboratoire de Physique Statistique de l'Ecole Normale Supérieure, affilié au CNRS & à l'Université Pierre et Marie Curie, Paris, France

1 Introduction

Multi-components and strongly interacting systems, in physics and beyond, may show complex behaviours eluding simple quantitative modeling. A common strategy to describe such systems is to define probability distributions over the space of their configurations. The task is at first sight daunting. The number of unknown probabilities scales as the dimension of the configuration space, and is enormous, generally exponentially large in the number of the degrees of freedom defining the system. The selection of one probability distribution among the multitude of possibilities may be done in a Bayesian way. A class of parametrized model distributions is considered, and a good choice of the parameters is sought, *e.g.* which maximizes the likelihood of the observed data. Statisticians are generally interested in understanding how learning proceeds, that is, in the speed of convergence to the target distribution (assumed to be in the parametrized class, and to be consistent with one value of the parameters) as more and more data are made available. Yet, though the number of parameters defining the class of distributions may be very large and the inference problem may be computationally very hard, the classes of parametrized distributions are generally extremely small compared to the plethora of possible distributions over the configuration space. While some classes of distributions may appear more adequate to represent the data or more amenable to computations, those criteria are largely arbitrary.

The maximum entropy (ME) principle is an alternative approach, rejecting the notion of arbitrariness. The ME principle may be informally stated as follows: among all possible distributions compatible with what is known of the data, choose the one with ME. It was proposed as an alternative foundation of statistical mechanics [1–3]; as an illustration, the Boltzmann distribution in the canonical ensemble may be found back as the distribution with ME under the sole knowledge of the average value of the energy. The ME principle is supported by information theory [4], which argues that any other distribution would be too constrained, and, in fact, reflect additional properties of the data [5]. An alternative and somewhat colourful formulation of this argument was given in [6], where the ME distribution was shown to emerge as the most frequent distribution which an uninformed operator (monkeys in [6]) would find, upon repeated and unbiased trials compatible with the observations on the data. In other words, given an unbiased prior over the set of all possible distributions over the configuration space, the ME distribution is the most likely one compatible with our knowledge of the data. Furthermore the ME distribution enjoys important and valuable properties, *e.g.* a weak sensitivity to measurement errors [7]. To the practitioner, the most compelling argument in support of the ME distribution may actually come from the successful applications to experimental data, see for instance [8] for a clear presentation regarding biological data.

The purpose of the present work is to compare the performances of the ME distribution with the ones of the other distributions compatible with the data. While the literature on the ME principle has a rich history, we are not aware of works attempting to carry out this comparison in a quantitative and rather general setting. We consider a set of N variables, taking binary values ± 1 (the generalization to a larger number of states would be straightforward, as long as it remains finite). Any distribution over the space of configurations is entirely characterized by 2^N probabilities, which are non-negative real numbers summing to unity. Each distribution may therefore be seen as a point in the 2^N-dimensional simplex. We pick up one point in this simplex, hereafter referred to as the target distribution. Next we consider a set of M observables, each of which may be any polynomial function of the variables, and compute

their average values over the target distribution. The distributions in the simplex compatible with those data, *i.e.* such that the observables have the same average within some prescribed accuracy, are called admissible. The set of admissible distributions, called version space, contains the target distribution, the ME distributions, and many others, such as the center-of-mass distribution, which is the flat average over all admissible distributions. Our objective is to compute the main geometrical features of the version space, *e.g.* its volume, the distances from the target to the ME or to the center-of-mass distributions, the average inter-distribution distance, ...

To give a precise meaning to those quantities in a mathematically tractable framework we will assume that observables are drawn from a simple statistical ensemble. More precisely the values taken by the observables are assumed to be random and uncorrelated across the 2^N configurations. This assumption is not meant to be realistic. In most real applications, indeed, observables reflect the low-order statistics of the configurations, *e.g.* the value of the first variable, the product of the fifth and seventh variables, ... Such observables vary very smoothly over the configuration space, as they depend only a small (compared to N) number of the configuration variables, and may be adequate to provide information about smooth distributions. Our hypothesis can be considered as worst-case-like in the following sense. The inference of the target distribution from the average values of the observables may be recast as the problem of reconstructing a 2^N-dimensional non-negative vector from the knowledge of its scalar products with M vectors (corresponding to the observables). When those vectors are randomly chosen the scalar products are typically very small, of the order of $2^{-N/2}$, and are weakly informative about the direction of the target vector. In this pessimistic setting we expect that the number of scalar products (observables) necessary to reconstruct the target vector with accuracy will be of the order of the number of relevant components of the target vector, that is, of the order of the exponential of the entropy of the target distribution. While this statement is correct we will see that some important features of the target distribution are correctly inferred even with a much smaller number M of observables. In particular, the probabilities of configurations with large values of the target distribution are learned with a limited number of data, a phenomenon connected to the onset of phase transitions in the learning process.

An important aspect of our framework is that it allows us to bias the measure over the version space, in order to boost the distributions with large entropies. The magnitude of this entropic bias may be continuously tuned from zero (uniform measure over the version space) to infinity, which amounts to selecting the ME distribution alone. We study how the distance between the target distribution and center-of-mass distribution varies with the bias (for a fixed number M of observables). Our main result is that this distance does not depend on the bias, showing that the ME distribution is not better than any other distribution randomly picked up in the version space. While this result is valid for any target distribution in the case of random observables we do not expect it to apply to the more realistic case where both the observables and the target distribution are smooth.

All our results are derived within the replica symmetric framework when N and M go to infinity at fixed ratio $\alpha = (\log M)/N$, and are therefore non rigorous. We have, however, checked the local stability of our replica-symmetric solution against replica-symmetry-breaking fluctuations, the so-called replicon modes. Our results are therefore self-consistent and we expect them to be correct for large N, M. In addition we have designed a Monte Carlo algorithm to sample the version space, and applied it to small system sizes ($N \leq 10$). Remarkably, simulations show only weak finite-size effects compared to our large-N calculations; a good qualitative (and sometimes even quantitative) agreement with our analytical predictions is found.

The paper is organized to be accessible to the reader not interested in the details of our calculation. In Sect. 2, we present the necessary definitions and notations. An overview of our results, free of technicalities, is given in Sect. 3. All technical details and calculations are reported in Sect. 4 and in the Appendices. We present the sampling algorithm and the results of our numerical simulations in Sect. 5. Conclusions can be found in Sect. 6.

2 Definitions and Notations

2.1 Target Distribution

Let us consider a system consisting of N Ising spins, with configurations $\boldsymbol{s} = \{s_i = \pm 1\}_{i=1}^{N}$. The probability distributions of the system configurations, hereafter called target distribution, is denoted by $\hat{p}_{\boldsymbol{s}}$. We consider large-size systems, $N \gg 1$, and write

$$\hat{p}_{\boldsymbol{s}} \doteq e^{-N\omega_{\boldsymbol{s}}}, \tag{1}$$

where the rate $\omega_{\boldsymbol{s}} > 0$ of the configuration probability $\hat{p}_{\boldsymbol{s}}$ is introduced, and the symbol $\doteq$ stands for equality in the leading exponential-in-N term. It is convenient to introduce the entropy σ of the rates $\omega_{\boldsymbol{s}}$,

$$\sigma(\omega) = \lim_{\epsilon \to 0^+} \lim_{N \to \infty} \frac{1}{N} \log \left[\text{nb. of configurations } \boldsymbol{s} \text{ such that } e^{-N\omega} \le \hat{p}_{\boldsymbol{s}} < e^{-N(\omega-\epsilon)} \right], \tag{2}$$

The entropy curve $\sigma(\omega)$ has some remarkable features in standard physical systems. First, it is convex and bounded from above and below. Secondly, the maximum of the curve is $\sigma_0 = \log 2$; the corresponding value of ω is denoted by ω_0. Thirdly, the curve lies below the $\sigma = \omega$ line, and is tangent to this line at $\omega_1 (\le \omega_0)$. The value $\sigma_1 = \sigma(\omega_1)$ is the entropy per spin of the target distribution

$$\sigma_1 = \lim_{N \to \infty} -\frac{1}{N} \sum_{\boldsymbol{s}} \hat{p}_{\boldsymbol{s}} \log \hat{p}_{\boldsymbol{s}}. \tag{3}$$

More generally, given a real number k, we define ω_k, σ_k, and ℓ_k through

$$\left. \frac{d\sigma}{d\omega} \right|_{\omega = \omega_k} = k \ , \quad \sigma_k \equiv \sigma(\omega_k) \ , \quad \ell_k \equiv k\,\omega_k - \sigma_k \ . \tag{4}$$

Note that the range of values of k such that ω_k is well-defined is generally bounded. These quantities characterize the dominant contribution to the kth moment of $\hat{p}$:

$$\sum_{\boldsymbol{s}} (\hat{p}_{\boldsymbol{s}})^k = \int d\omega\, e^{N(\sigma(\omega) - k\,\omega)} \doteq e^{N(\sigma_k - k\,\omega_k)} = e^{-N\ell_k} \ . \tag{5}$$

As an illustration of the properties above we show in Fig. 1 the entropy curve of the independent spin model (ISM) defined as

$$\hat{p}_{\boldsymbol{s}}^{\,\mathrm{ISM}} = \frac{1}{(2\cosh H)^N} \ \exp\left(H \sum_{i=1}^{N} s_i \right), \tag{6}$$

for $H = 0.5$. The value of H will not change throughout this paper.

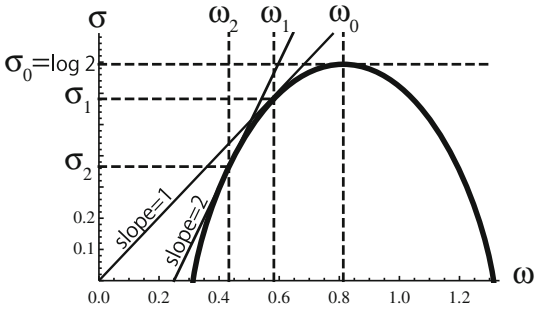

Fig. 1 Entropy σ as a function of ω for the independent spin model in Eq. (6) with $H = 0.5$. The entropy curve is obtained through a parametric representation $\omega(m) = \log(2\cosh H) - Hm$, $\sigma(m) = \log 2 - \frac{1}{2}(1+m)\log(1+m) - \frac{1}{2}(1-m)\log(1-m)$, where m is the magnetization per spin, ranging from -1 to $+1$. It is easy to show that $\omega_k = \omega(\tanh(kH))$

2.2 Observables, Version Space, and Maximum Entropy Distribution

Let $\boldsymbol{v}$ be an observable, taking value $v_{\boldsymbol{s}}$ when the system configuration is $\boldsymbol{s}$. Observables and probability distributions can be seen as vectors in a 2^N-dimensional space, with components labelled by the configurations $\boldsymbol{s}$. We assume that measurements give us access to the average value of the observable over the target distribution,

$$\sum_{\boldsymbol{s}} v_{\boldsymbol{s}}\, \hat{p}_{\boldsymbol{s}} \equiv \boldsymbol{v} \cdot \hat{\boldsymbol{p}}, \tag{7}$$

which may be simply written as the scalar product of the vectors attached to the observable and to the target distribution. Suppose we have made M measurements corresponding to M observables $\boldsymbol{v}^\mu$ ($\mu = 1, \ldots, M$). An 'admissible' probability vector, $\boldsymbol{p}$, compatible with all the measurements is such that

$$\left(\boldsymbol{p} - \hat{\boldsymbol{p}}\right) \cdot \boldsymbol{v}^\mu = 0\,, \quad \forall\, \mu = 1, \cdots, M. \tag{8}$$

We hereafter use the term 'constraints' to refer to Eq. (8) or to the attached observables $\boldsymbol{v}^\mu$. The set of vectors satisfying Eq. (8), together with the normalization and the non-negativity conditions

$$\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} = 1 \quad \text{and} \quad p_{\boldsymbol{s}} \geq 0\,, \forall \boldsymbol{s}\,, \tag{9}$$

defines the version space.

Each distribution $\boldsymbol{p}$ in the version space is characterized by its Shannon entropy,

$$S(\boldsymbol{p}) = -\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} \log p_{\boldsymbol{s}}\,, \tag{10}$$

which may be interpreted as an estimate of the logarithm of the effective number of configurations under the distribution. The maximum entropy (ME) distribution, $\boldsymbol{p}_{ME}$, is the distribution maximizing Eq. (10) in the version space, *i.e.* under the constraints listed in Eq. (8) and in Eq. (9). Using Lagrange multipliers η^μ to enforce those constraints the ME distribution may be formally written as

$$p_{\boldsymbol{s}}^{\mathrm{ME}} = C\ \exp\left(\sum_\mu \eta^\mu\, v_{\boldsymbol{s}}^\mu\right), \tag{11}$$

where C is a normalization constant. As an illustration, if we consider the set of the N single-spin observables, $v_{\boldsymbol{s}} = s_i$, $1 \leq i \leq N$, and of the $N(N-1)/2$ two-spin observables

$v_s = s_i\, s_j$, $1 \le i < j \le N$, we recover the well-known result that the Ising model is the ME model given the average values of those observables. The Lagrange multipliers η coincide with, respectively, the N fields and the $N(N-1)/2$ pairwise couplings acting on the spins.

2.3 Measures Over the Space of Distributions

In realistic situations involving certain measurement noise, it may be beneficial not to perfectly reproduce the given average value by avoiding overfitting. To take into account such a flexibility, we introduce a Gaussian-like measure with variance E over the vector space of $\boldsymbol{p}$:

$$\rho\left(\boldsymbol{p}|\{\boldsymbol{v}^\mu\}_{\mu=1}^M, \hat{\boldsymbol{p}}\right) = \frac{1}{V}\prod_{\boldsymbol{s}}\theta(p_{\boldsymbol{s}})\,\delta\left(\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} - 1\right)\exp\left\{-\frac{1}{2E}\sum_{\mu=1}^{M}\left(\boldsymbol{v}^\mu\cdot\left(\boldsymbol{p}-\hat{\boldsymbol{p}}\right)\right)^2\right\}, \tag{12}$$

where $\theta(x)$ is the step function, equal to 1 if $x \ge 0$ and to 0 if $x < 0$, and $\delta(x)$ is the Dirac delta function. We will call E tolerance hereafter. The denominator V is defined to make sure that the measure ρ is normalized:

$$V\left(E, \{\boldsymbol{v}^\mu\}_{\mu=1}^M, \hat{\boldsymbol{p}}\right) = \int_0^\infty \prod_{\boldsymbol{s}} dp_{\boldsymbol{s}}\,\delta\left(\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} - 1\right)\exp\left\{-\frac{1}{2E}\sum_{\mu=1}^{M}\left(\boldsymbol{v}^\mu\cdot\left(\boldsymbol{p}-\hat{\boldsymbol{p}}\right)\right)^2\right\}. \tag{13}$$

This normalization factor measures the volume of 'admissible' probability vectors given the constraints Eq. (8). In this probabilistic setting we will loosely use the term 'version space' to refer to the set of distributions $\boldsymbol{p}$ associated to 'large' measure values $\rho(\boldsymbol{p})$. Note that, while $\rho(\boldsymbol{p})$ defines the joint-measure of the 2^N-configuration probabilities, we will also consider below the marginal measure for a single-configuration probability,

$$\rho_{\boldsymbol{s}}(p_{\boldsymbol{s}}|\{\boldsymbol{v}^\mu\}_{\mu=1}^M, \hat{\boldsymbol{p}}) = \int_0^\infty \prod_{\boldsymbol{t}(\neq\boldsymbol{s})} dp_{\boldsymbol{t}}\,\rho\left(\boldsymbol{p}|\{\boldsymbol{v}^\mu\}_{\mu=1}^M, \hat{\boldsymbol{p}}\right). \tag{14}$$

It is possible to consider other measures of the space of $\boldsymbol{p}$ for the purpose of studying the performances of the ME distribution. To favor the probability vectors $\boldsymbol{p}$ with large Shannon entropies $S(\boldsymbol{p})$, see Eq. (10), it is natural to introduce the new measure

$$\rho\left(\boldsymbol{p}|\Gamma, \{\boldsymbol{v}^\mu\}_{\mu=1}^M, \hat{\boldsymbol{p}}\right) \propto \rho\left(\boldsymbol{p}|\{\boldsymbol{v}^\mu\}_{\mu=1}^M, \hat{\boldsymbol{p}}\right)\; e^{\Gamma\, S(\boldsymbol{p})}, \tag{15}$$

where Γ is the strength of this entropic bias. The normalization of this new measure ρ implicitely defines the following expression for the volume in the presence of an entropic bias,

$$V\left(\Gamma, E, \{\boldsymbol{v}^\mu\}_{\mu=1}^M, \hat{\boldsymbol{p}}\right) = \int_0^\infty \prod_{\boldsymbol{s}} dp_{\boldsymbol{s}}\,\delta\left(\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} - 1\right) \exp\left\{-\frac{1}{2E}\sum_{\mu=1}^{M}\left(\boldsymbol{v}^\mu\cdot\left(\boldsymbol{p}-\hat{\boldsymbol{p}}\right)\right)^2 + \Gamma S(\boldsymbol{p})\right\}. \tag{16}$$

For $\Gamma = 0$ we recover the 'unbiased' measure in Eq. (12), while the limit $\Gamma \to \infty$ retains the ME distribution only.

2.4 Randomization of Observables

Hereafter we assume that the M constraints $\{\boldsymbol{v}^\mu\}_{\mu=1}^M$ are randomly and independently chosen from the following Gaussian distribution over the space of 2^N-dimensional vectors $\boldsymbol{v}$:

$$P(\boldsymbol{v}) = \prod_s \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v_s^2}. \tag{17}$$

As mentioned in the introduction we do not pretend that this assumption is realistic. Indeed in practice, observables are often chosen to be low-order moments of the target distribution such as magnetizations or pairwise spin correlations, which is quite distinct from the Gaussian distribution above.

In contrast to the constraints $\{\boldsymbol{v}^\mu\}_\mu$, which are randomly drawn, we do not choose any particular statistical ensemble for the target distribution $\hat{\boldsymbol{p}}$; The only properties of $\hat{\boldsymbol{p}}$ we need for the analysis is the existence of the entropy curve discussed above. We stress that these constraints $\{\boldsymbol{v}^\mu\}_\mu$, appearing in the measure ρ over the distribution space, are quenched random variables, drawn once for all.

Since each constraint contributing a multiplicative factor $e^{-(\boldsymbol{v}^\mu\cdot(\boldsymbol{p}-\hat{\boldsymbol{p}}))^2/2E}$ to the volume V in Eq. (16), we expect that the logarithm of this volume will be self-averaging in the large-M limit. We will therefore calculate the average value of $\log V$ over the constraints, using the replica method.

3 Overview of Results

We hereafter report the main outcomes of our replica calculation in an informal way. All technical details are postponed to Sect. 4.

3.1 Order Parameters: Interpretation and Scaling with N

Given the set of constraints $\{\boldsymbol{v}^\mu\}_{\mu=1}^M$ and the target distribution $\hat{\boldsymbol{p}}$ we may draw a schematic picture of the version space such as the one shown in Fig. 2. In addition to the target distribution, two distributions of interest in the version space are the ME distribution, $\boldsymbol{p}^{\mathrm{ME}}$, and the center-of-mass distribution, $\langle \boldsymbol{p} \rangle$, where the angular brackets denote the average over the measure ρ:

$$\langle \boldsymbol{p} \rangle = \int d\boldsymbol{p}\ \rho(\boldsymbol{p})\ \boldsymbol{p}\ . \tag{18}$$

The following quantities are useful to characterize the 'distances' between the distributions in the version space, see Fig. 2:

$$R(\{\boldsymbol{v}^\mu\}_\mu, \hat{\boldsymbol{p}}) \equiv \sum_s \left(\langle p_s \rangle - \hat{p}_s\right)^2\ , \quad D(\{\boldsymbol{v}^\mu\}_\mu, \hat{\boldsymbol{p}}) \equiv \sum_s \left(\langle p_s^2 \rangle - \langle p_s \rangle^2\right)\ . \tag{19}$$

R measures the L_2-squared distance between the target and the center-of-mass distributions. D gives the size of the squared fluctuations of $\boldsymbol{p}$ around the center of mass. We will also consider hereafter

$$Q(\{\boldsymbol{v}^\mu\}_\mu, \hat{\boldsymbol{p}}) \equiv R + D = \sum_s \left\langle \left(p_s - \hat{p}_s\right)^2 \right\rangle\ , \tag{20}$$

which measures the averaged square distance between $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$ (Fig. 2).

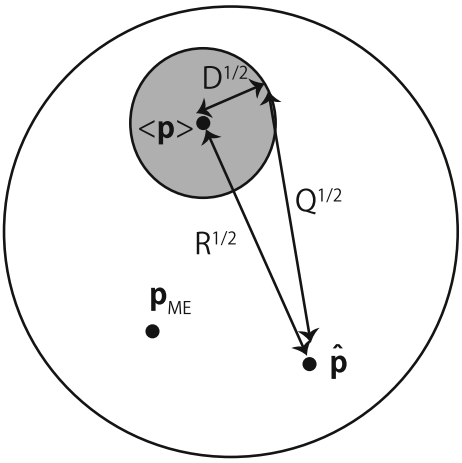

Fig. 2 Schematic view of the space of distributions. The *large circle* represents the version space of all admissible probability vectors, see Eq. (8), which depends on the constraints $\{\boldsymbol{v}^\mu\}_{\mu=1}^M$. The *shaded area* represents the typical fluctuations of $\boldsymbol{p}$ around the center-of-mass distribution $\langle \boldsymbol{p}\rangle$, of magnitude $\sqrt{D}$. The distance between the target distribution $\hat{\boldsymbol{p}}$ and $\langle \boldsymbol{p}\rangle$ is $\sqrt{R}$, while $\sqrt{Q}$ is the square root of the averaged squared distance between $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$. These three order parameters measure the lengths of the sides of the *rectangular triangle* shown in the figure, see Eq. (20). The ME distribution, $\boldsymbol{p}^{\mathrm{ME}}$, lies inside the version space

D, Q, R are intimately related to the statistics of the generalization error, that is, the error on the prediction of the average value of a new observable. Assume we have a measure ρ over the version space defined from a set of constraints $\{\boldsymbol{v}^\mu\}_{\mu=1}^M$. Let us now consider a new observable, $\boldsymbol{v}'$. The error on the average value of this observable $\boldsymbol{v}'$ computed with a distribution $\boldsymbol{p}$ is simply given by

$$\Delta(\boldsymbol{p}) = \boldsymbol{v}' \cdot \left(\boldsymbol{p} - \hat{\boldsymbol{p}}\right). \tag{21}$$

Suppose now that $\boldsymbol{p}$ is randomly chosen according to measure ρ and that the components v'_s are independent and normal random variables, with zero means and unit variances. Let us denote the average over $\boldsymbol{v}'$ by $[\cdots]_{\boldsymbol{v}'}$. It is easy to show that

$$Q = \left[\langle \Delta^2\rangle\right]_{\boldsymbol{v}'}\ , \quad R = \left[\langle\Delta\rangle^2\right]_{\boldsymbol{v}'} = \left[\langle\Delta\rangle^2\right]_{\boldsymbol{v}'} - \left[\langle\Delta\rangle\right]^2_{\boldsymbol{v}'}\ , \quad D = \left[\langle\Delta^2\rangle - \langle\Delta\rangle^2\right]_{\boldsymbol{v}'}. \tag{22}$$

Hence, Q represents the averaged square error on a new observable, R quantifies the observable-to-observable squared fluctuations of the error, and D measures the distribution-to-distribution squared fluctuations of the error. The squared error on the prediction of the average value of the new observable, Q, is the sum of two contributions. The first one, R, is due to the choice of observable. The second one, D, reflects the distribution-to-distribution fluctuations with the version space measure.

To understand the scaling of D, Q, R with N let us consider the simple case of no constraint at all, $M = 0$, and no bias on the entropy, $\Gamma = 0$. In this case, due to the permutation symmetry over the 2^N configurations, all the configuration probabilities p_s obey the same marginal distribution, $\rho_s(p_s)$, see Eq. (14). It is easy to convince oneself, that ρ_s becomes a pure exponential when $N \gg 1$, with average value $\langle p_s\rangle = 2^{-N}$. Indeed, the volume in Eq. (13) is given by

$$V = \int_0^\infty \prod_s dp_s\, \delta\left(\sum_s p_s - 1\right) = \int_{-i\infty}^{i\infty} d\Lambda\, e^{\Lambda} \int_0^\infty \prod_s \left(dp_s\, e^{-\Lambda p_s}\right) = \int_{-i\infty}^{i\infty} d\Lambda\, \frac{e^{\Lambda}}{\Lambda^{2^N}}, \tag{23}$$

where we have used an integral representation of the Dirac delta function. This can be directly integrated but we here use the saddle-point method, valid for large N for later convenience. The saddle point is located at

$$\Lambda = 2^N\,, \tag{24}$$

giving $\log V \doteq 2^N(1 - N\log 2)$, which agrees with the leading behaviour of the volume of the 2^N-dimensional simplex, $V = 1/[(2^N)!]$. In addition we see from Eqs. (14) and (23) that ρ_s is an exponential distribution with mean value $1/\Lambda = 2^{-N}$, as announced above. Hence,

$$D \doteq e^{Nd}\,,\ Q \doteq e^{Nq}\,,\ R \doteq e^{Nr}\,,\ \Lambda \doteq e^{N\lambda}\,, \tag{25}$$

with $q = r = -\ell_2$, see Eq. (4), and $d = -\lambda = \log 2$. Our calculation with the replica method shows that, in the presence of constraints, *i.e.* for $M \geq 1$, the exponential-in-N scaling of D, Q, R, Λ will still hold after averaging over the constraints, even if a bias Γ over the entropy is imposed. Note that the rates d, q, r, λ will then depend on M and Γ.

3.2 Description of the Learning Process

We first focus on the effect of increasing the number of measured observables, M, and set the entropic bias Γ to zero. The effect on non-zero biases will be reported in Sect. 3.3. We assume in Sects. 3.2.1 and 3.2.2 that the tolerance E is negligible; the dependence of the results on the value of E is exposed in Sect. 3.2.3.

3.2.1 The Learning Edge

One major result of the replica calculation is that the marginal measure over the single-configuration probabilities, ρ_s, Eq. (14), may have two distinct behaviours, sketched in Fig. 3:

– The marginal measure over the probability p_s of a configuration $\boldsymbol{s}$ having a *large* value $\hat{p}_s$ in the target distribution is concentrated at a value close to this target probability:

$$\rho_s(p_s) \approx A\, e^{-B\,(p_s - \hat{p}_s - \delta_s)^2/2}\,, \tag{26}$$

where the values of A and B depend on the parameters M, E, and will be specified later. The corresponding curve of ρ_s is shown in Fig. 3, left. The shift δ_s is small compared

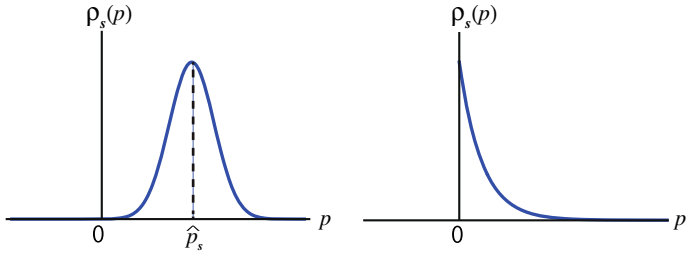

Fig. 3 Schematic pictures of the marginal measure $\rho_s(p_s)$ for large (*left*) and small (*right*) configuration probabilities $\hat{p}_s$

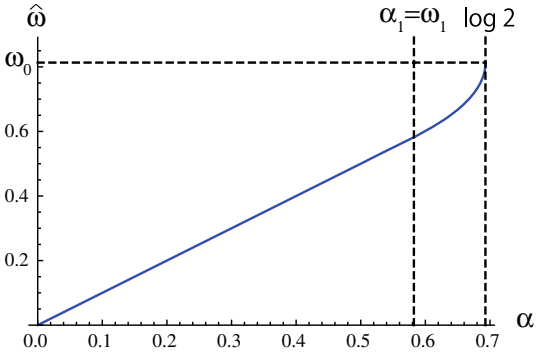

Fig. 4 Learning edge $\hat{\omega}$ as a function of α for the ISM (6) with $H = 0.5$ and for small E

to the target probability $\hat{p}_s$, as we will see in Sect. 4. In other words, the configuration probability $\hat{p}_s$ is correctly 'learned' from the values of the constraints.

– The marginal measure over the probability p_s of a configuration s having a *small* value $\hat{p}_s$ in the target distribution is a decaying exponential:

$$\rho_s(p_s) \approx \frac{1}{A'} e^{-p_s/A'} . \tag{27}$$

The corresponding curve of ρ_s is shown in Fig. 3, right. The average value A' of the configuration probability p_s does not depend on $\hat{p}_s$ in the dominant scaling with N. A' is a function of the parameters M, E only, and will be specified later. In other words, the configuration probability is not at all inferred.

The concepts of *large* and *small* target probabilities are defined as follows:

$$\hat{p}_s \doteq e^{-N\omega_s} \text{ is large if } \omega_s < \hat{\omega} , \text{ and is small if } \omega_s > \hat{\omega} . \tag{28}$$

The boundary $\hat{\omega}$ is hereafter referred to as the *learning edge*, as it separates the set of probabilities p_s into the ones which can be correctly learned and the ones which remain essentially unknown, despite the knowledge of the observable values. Its value depends on the parameters M and E. The representative curve of the learning edge $\hat{\omega}$ is an increasing function of the rate

$$\alpha = \frac{\log M}{N} , \tag{29}$$

and is shown in Fig. 4 for the ISM and a negligible tolerance E. We find qualitative changes of the curve $\hat{\omega}(\alpha)$, taking place at critical values of the ratio α and corresponding to the onset of phase transitions in the learning process.

3.2.2 Phase Transitions and Typical Distances Between Distributions

Given the value of the learning edge $\hat{\omega}$ we define, according to Eq. (28), the sets of configurations s having large and small probabilities as, respectively, L and S. Our replica calculation shows that the order parameters Q and R are given by, to the dominant order in N,

$$Q \doteq R \doteq \sum_{s\in S} \hat{p}_s^2 . \tag{30}$$

If M is small, so is the learning edge $\hat{\omega}$, most configurations are found in S. This implies $\sum_{s\in S} \hat{p}_s^2 \doteq \sum_s \hat{p}_s^2 = e^{-N\ell_2}$. As M increases, the learning edge grows, and reaches $\hat{\omega} = \omega_2$.

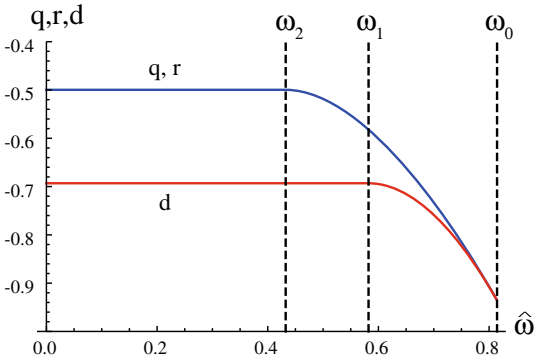

Fig. 5 Order parameters q, r, d as functions of the learning edge $\hat{\omega}$ for the ISM (6) with $H = 0.5$ and for small E. The critical values of ω are $\omega_0 \approx 0.81$, $\omega_1 \approx 0.58$, and $\omega_2 \approx 0.43$, and the corresponding entropies are $\sigma_0 = \log 2 \approx 0.69$, $\sigma_1 \approx 0.58$, and $\sigma_2 \approx 0.37$, respectively

For larger values of M, the dominant point $\omega = \omega_2$ is now not in $\boldsymbol{s} \in S$ and hence the rates q and r change from $-\ell_2$. This leads non-analyticities in Q and R, and the transition point is at $\hat{\omega} = \omega_2$. Similarly, a switch of the dominant term in the normalization condition $1 = \sum_{\boldsymbol{s}\in S} \hat{p}_{\boldsymbol{s}} + \sum_{\boldsymbol{s}\in L} \hat{p}_{\boldsymbol{s}}$ from the S to the L occurs at $\hat{\omega} = \omega_1$. The Lagrange multiplier Λ is non analytic at this point, which defines a second phase transition.

Hence, we have three distinct phases, separated by specific values of the learning edge. We label the phases by 'I' when $\hat{\omega} < \omega_2$, 'II' when $\omega_2 \leq \hat{\omega} < \omega_1$, and 'III' when $\omega_1 \leq \hat{\omega}$. The value of the ratio α corresponding to the transition point $\hat{\omega} = \omega_2$ is denoted as α_2, and the one corresponding to $\hat{\omega} = \omega_1$ is called α_1. In phases I and II with small E, $\hat{\omega}$ turns out to be equal to α. Thus $\alpha_2 = \omega_2$ and $\alpha_1 = \omega_1$.

To get insights about the typical distances between the distributions of interest in the version space we plot the order parameters q, r and d of the ISM in Fig. 5. Those order parameters continuously change at the transition points. At $\alpha = \log 2$, the learning edge becomes equal to ω_0; q and r reach the same value as d. The agreement between r and d, the distance of the center of mass to the target distribution and the typical fluctuations, implies that the volume largely shrinks and scales with N in a different manner at this point. Our calculation, restricted to the leading order in the volume, is not informative for larger ratios, i.e. $\alpha > \log 2$.

3.2.3 Effect of the Tolerance E

Let us now consider the case of a non-negligible tolerance E. We give the corresponding phase diagram in Fig. 6. It is natural to scale the tolerance as $E \doteq e^{N\epsilon}$. As ϵ grows from very negative values, we expect that the quality of the inference becomes worse, as the version space include more distributions, which do not exactly reproduce the target values of the observables. Indeed, for large values of the tolerance E, there emerge new phases where the learning edge decreases as ϵ grows. We call these phases with large tolerance I_{LT}, $\mathrm{II}_{\mathrm{LT}}$, and $\mathrm{III}_{\mathrm{LT}}$, in agreement with the denomination chosen above based on the different ranges of possible values for the learning edge.

The transition from small to large tolerances takes place at $E \doteq D$. This result is easy to interpret: deviations in the values of observables from the target ones smaller than the scale of the intrinsic fluctuations D will be masked by those fluctuations, and cannot degrade the inference performances. The results shown in Sects. 3.2.1 and 3.2.2 are therefore valid as long as $\epsilon < d$.

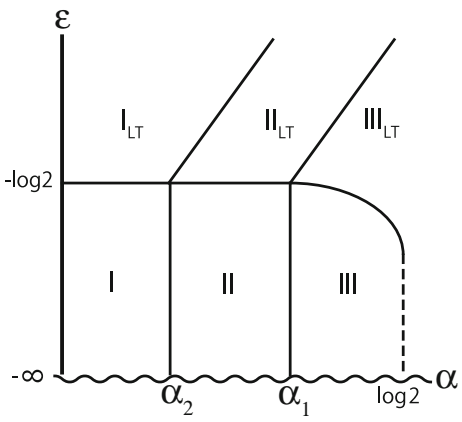

Fig. 6 Phase diagram in the α-ϵ plane in the absence of any entropic bias

In the present study, we do not consider any measurement error: the measured values of the observables correspond exactly to the projection of the observable vectors onto the target distribution. If measurement errors were considered, the role of the tolerance parameter could possibly change; an appropriate amount of tolerance E, of the order of the measurement error, would lead to better inference by avoiding overfitting.

3.3 Effects of the Entropic Bias

We now consider the effect of an entropic bias $\Gamma > 0$ on the learning performances. As Γ grows, the measure ρ gives more and more weight to the distributions $\boldsymbol{p}$ with large entropies. In the $\Gamma \to \infty$ limit ρ singles out the ME distribution. The Γ-dependence of the quantities of interest, in particular, the learning edge, is of key importance. Similarly to the other parameters we assume that the entropic bias scales $\Gamma = e^{N\gamma}$.

We first consider the case of negligible $E \approx 0$. Our replica calculation shows that the order parameters Q, R and the learning edge $\hat{\omega}$ do not depend at all on the value $\Gamma \geq 0$. An immediate consequence is that the ME distribution is not closer to the target distribution than any randomly picked up distribution (with the unbiased measure) in the version space. Nevertheless the presence of an entropic bias affects the fluctuations of $\boldsymbol{p}$, measured by the order parameter D. For $\alpha < \alpha_1$, our replica calculation shows that $D \doteq 2^{-N}$ for small Γ, while $D \doteq \Gamma^{-1}$ for large Γ. The transition between these two scalings takes place at $\gamma_c = \log 2$. The same transition for D is observed when $\alpha > \alpha_1$, though the transition point γ_c now depends on α. The order parameter Λ shows non-analyticities at those transition points, contrary to Q, R, and the learning edge, which remain unaffected as mentioned above. The non-analyticities in D and Λ discriminate the phases for $\gamma \leq \gamma_c$ from the ones for $\gamma > \gamma_c$; we denote the phases with large entropic bias by I_{ME}, $\mathrm{II}_{\mathrm{ME}}$, and $\mathrm{III}_{\mathrm{ME}}$, in agreement with the denomination based on the ranges of values of the learning edge. The corresponding phase diagram is shown in Fig. 7, left. On the right panel of the same figure, the order parameter d is plotted against γ for the ISM with $H = 0.5$ and $\alpha < \alpha_1$ as an illustration.

The phases above change when the value of the tolerance $E \doteq e^{N\epsilon}$ is not negligible any longer. As in the $\Gamma = 0$ case the tolerance matters if $E > D$. As the fluctuations scale as $D \doteq \Gamma^{-1}$ for large Γ with $\alpha < \alpha_1$, this implies that a regime takes place for $\gamma > -\epsilon$. This large tolerance, large entropic bias regime shows unusual properties. For instance the learning edge becomes smaller and smaller, that is, the learning performances become worse

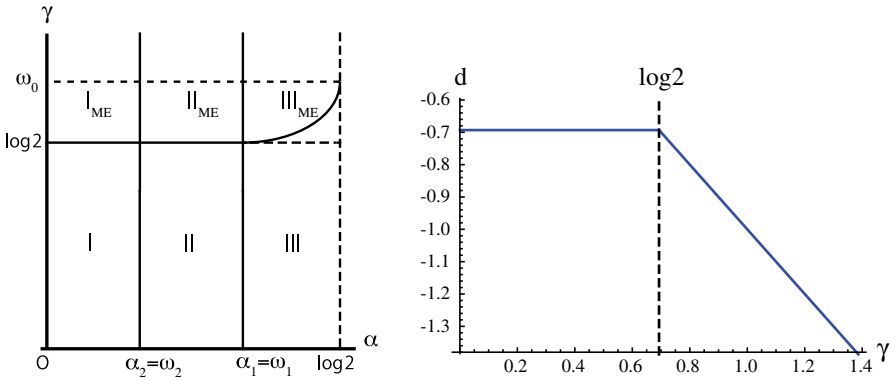

Fig. 7 *Left* Phase diagram in the α-γ plane for very small tolerance E. The fluctuation order parameter $D \doteq e^{Nd}$ depends on γ in the phases with large entropic biases, I_{ME}, $\mathrm{II}_{\mathrm{ME}}$, and $\mathrm{III}_{\mathrm{ME}}$, while it keeps the same value as for $\Gamma = 0$ in I, II, and III. *Right* order parameter d as a function of γ for the ISM and $\alpha < \alpha_1$. A non-analyticity appears at $\gamma = \gamma_c (= \log 2)$, which signals the onset of the phase with large entropic bias for $\gamma > \gamma_c$

and worse, as Γ grows, in contrast to the case of negligible $E \approx 0$ where the learning edge has no dependence on Γ. This seemingly-strange behaviour is actually expected. We see from Eq. (16) that, for very large Γ and finite fixed E, the constraints become irrelevant. Hence, in this limit, the measure concentrates around the uniform distribution $p_s^{\mathrm{ME}} = 1/2^N$, maximizing the entropy irrespective of the constraints. To get an meaningful ME distribution, the tolerance E must be enough small compared to the fluctuations D governed by Γ in the large Γ limit. In conclusion, this large tolerance, large entropic bias regime is irrelevant and will not be studied further.

4 Analytical Calculations

In this section, we present our replica calculation of the average value of the logarithm of the volume V over the constraints. We introduce the notations

$$\operatorname*{Tr}_{\boldsymbol{p}}(\cdots) \equiv \int_0^\infty \prod_s dp_s \delta\left(\sum_s p_s - 1\right)(\cdots), \tag{31}$$

and $Dz \equiv dz e^{-\frac{1}{2}z^2}/\sqrt{2\pi}$. We also write undefined integrals as a shorthand notation for the domains of integration $[-\infty, \infty]$ or $[-i\infty, i\infty]$.

4.1 Replica Calculation of the Average Logarithm of the Volume

We start by unraveling the squared terms in the exponent of volume (16) through auxiliary Gaussian integrations, or the so-called Hubbard-Stratonovich transformations:

$$V = \int \prod_{\mu=1}^{M} Dz_\mu \operatorname*{Tr}_{\boldsymbol{p}} \exp\left\{ i \sum_\mu \sum_s \frac{z_\mu}{\sqrt{E}} v_s^\mu (p_s - \hat{p}_s) + \Gamma S(\boldsymbol{p}) \right\}. \tag{32}$$

According to the replica method we calculate the nth moment of the volume of $n \in \mathbb{N}$

$$
\begin{aligned}
[V^n] &= \prod_{a=1}^{n}\left(\int \prod_{\mu} Dz_\mu^a \operatorname*{Tr}_{\boldsymbol{p}^a}\right)\left[\exp\left\{i\sum_{\mu}\sum_{s}\sum_{a}\frac{z_\mu^a}{\sqrt{E}}v_s^\mu(p_s^a-\hat{p}_s)+\Gamma\sum_{a}S(\boldsymbol{p}^a)\right\}\right] \\
&= \prod_{a=1}^{n}\left(\int \prod_{\mu} Dz_\mu^a \operatorname*{Tr}_{\boldsymbol{p}^a}\right)\exp\left\{-\sum_{\mu}\sum_{a,b}\frac{z_\mu^a z_\mu^b}{2E}Q_{ab}+\Gamma\sum_{a}S(\boldsymbol{p}^a)\right\} \\
&= \prod_{a=1}^{n}\operatorname*{Tr}_{\boldsymbol{p}^a}\exp\left\{-\frac{M}{2}\log\det\left(1+\frac{Q}{E}\right)+\Gamma\sum_{a}S(\boldsymbol{p}^a)\right\},
\end{aligned}
\tag{33}
$$

where the square brackets denote the average over the observables $\{\boldsymbol{v}^\mu\}_{\mu=1}^M$ as in Sect. 3.1, and we define

$$
Q_{ab} = \sum_{s}(p_s^a-\hat{p}_s)(p_s^b-\hat{p}_s). \tag{34}
$$

To perform the integrations $\operatorname{Tr}_{\boldsymbol{p}^a}$, we make use of the following identity

$$
\begin{aligned}
1 &= \int \prod_{a\le b} dQ_{ab}\prod_{a\le b}\delta\left(Q_{ab}-\sum_{s}(p_s^a-\hat{p}_s)(p_s^b-\hat{p}_s)\right) \\
&= C\int \prod_{a\le b} dQ_{ab}dQ'_{ab}\exp\left\{\frac{1}{2}\sum_{a\le b}Q_{ab}Q'_{ab}-\frac{1}{2}\sum_{a\le b}\sum_{s}Q'_{ab}(p_s^a-\hat{p}_s)(p_s^b-\hat{p}_s)\right\},
\end{aligned}
\tag{35}
$$

where the normalization constant C is irrelevant and will be discarded hereafter. Similarly, we rewrite $\operatorname{Tr}_{\boldsymbol{p}^a}$ as

$$
\operatorname*{Tr}_{\boldsymbol{p}^a} = \int d\Lambda_a \int_0^\infty \prod_{s} dp_s^a \; e^{-\Lambda_a\sum_s(p_s^a-\hat{p}_s)}, \tag{36}
$$

where we used the normalization identity $1=\sum_s \hat{p}_s$. We then obtain

$$
\begin{aligned}
[V^n] &= \int \prod_{a\le b} dQ_{ab}dQ'_{ab}\int\prod_{a} d\Lambda_a\int_0^\infty\prod_{a}\prod_{s}dp_s^a \\
&\times\exp\left(-\sum_{a}\Lambda_a\sum_{s}(p_s^a-\hat{p}_s)-\frac{1}{2}\sum_{a\le b}\sum_{s}Q'_{ab}(p_s^a-\hat{p}_s)(p_s^b-\hat{p}_s)+\Gamma\sum_{a}S(\boldsymbol{p}^a)\right) \\
&\times\exp\left\{\frac{1}{2}\sum_{a\le b}Q_{ab}Q'_{ab}-\frac{M}{2}\log\det\left(1+\frac{Q}{E}\right)\right\}.
\end{aligned}
\tag{37}
$$

The logarithm of $[V^n]$ can be approximated by the saddle-point value of

$$
\phi(n) \equiv \frac{1}{2}\sum_{a\le b}Q_{ab}Q'_{ab}-\frac{M}{2}\log\det\left(1+\frac{Q}{E}\right)+\sum_{s}\log\Theta_s, \tag{38}
$$

over the matrices Q and Q', and where

$$\Theta_s = \int_0^\infty \prod_a dp_s^a \exp\left(-\sum_a \Lambda_a (p_s^a - \hat{p}_s) - \frac{1}{2}\sum_{a\le b} Q'_{ab}(p_s^a - \hat{p}_s)(p_s^b - \hat{p}_s) - \Gamma \sum_a p_s^a \log p_s^a \right). \tag{39}$$

We now assume that the order parameter matrices are invariant under permutation symmetry of the replica indices, and write

$$Q_{ab} = R + (Q-R)\delta_{ab}, \quad Q'_{ab} = -2R' + (Q'+R')\delta_{ab}, \quad \Lambda_a = \Lambda, \tag{40}$$

where δ_{ab} is the Kronecker delta. Thus

$$\frac{1}{2}\sum_{a\le b} Q_{ab}Q'_{ab} = \frac{1}{2}nQ(Q'-R') - \frac{1}{2}n(n-1)RR', \tag{41}$$

$$\log\det\left(1+\frac{Q}{E}\right) = \log\left(1+\frac{Q+(n-1)R)}{E}\right) + (n-1)\log\left(1+\frac{Q-R}{E}\right), \tag{42}$$

and,

$$\frac{1}{2}\sum_{a\le b} Q'_{ab}(p_s^a - \hat{p}_s)(p_s^b - \hat{p}_s) = \frac{1}{2}Q'\sum_a (p_s^a - \hat{p}_s)^2 - \frac{1}{2}R'\left(\sum_a (p_s^a - \hat{p}_s)\right)^2. \tag{43}$$

Using the Hubbard-Stratonovich transformation again Eq. (39) becomes

$$\Theta_s = \int Dz \int_0^\infty \prod_a dp_s^a e^{-\left(\Lambda - z\sqrt{R'}\right)(p_s^a - \hat{p}_s) - \frac{1}{2}Q'(p_s^a - \hat{p}_s)^2 - \Gamma p_s^a \log p_s^a} \equiv \int Dz X_s^n. \tag{44}$$

Hence,

$$\begin{aligned}\phi(n) = {} & \frac{1}{2}nQ(Q'-R') - \frac{1}{2}n(n-1)RR' \\ & - \frac{M}{2}\left\{\log\left(1+\frac{Q+(n-1)R)}{E}\right) + (n-1)\log\left(1+\frac{Q-R}{E}\right)\right\} \\ & + \sum_s \log\int Dz X_s^n. \end{aligned} \tag{45}$$

We finally obtain the expression for $\left[\log V\right] = \lim_{n\to 0}\phi(n)/n$,

$$\begin{aligned}\left[\log V\right] = {} & \frac{1}{2}Q(Q'-R') + \frac{1}{2}RR' - \frac{M}{2}\left\{\frac{R}{E+Q-R} + \log\left(1+\frac{Q-R}{E}\right)\right\} \\ & + \sum_s \int Dz \log X_s, \end{aligned} \tag{46}$$

where

$$X_s = \int_0^\infty dp\, e^{-\left(\Lambda - z\sqrt{R'}\right)(p-\hat{p}_s) - \frac{1}{2}Q'(p-\hat{p}_s)^2 - \Gamma p\log p}. \tag{47}$$

The integration over p in Eq. (47) has to be done with care. We assume that the conjugated order parameters, Q' and R', obey an exponential scaling with N,

$$Q' \doteq e^{Nq'}, \quad R' \doteq e^{Nr'}. \tag{48}$$

Therefore, all the parameters appearing in the integration diverge or vanish when $N \to \infty$. We need to find out which order parameters diverge or not, consistently with the equations of state derived below. These procedures require involved case analyses. The outcome is that the equations of state are consistent, to the leading order in N, if the following conditions are met:

$$0 < \lambda < q' \leq 2\lambda, \ 0 < \lambda < r' \leq 2\lambda, \ q = r < 0, \ \epsilon < -\gamma \leq 0. \tag{49}$$

4.2 $\Gamma = 0$ Case

In this case, the integral in Eq. (47) can be evaluated in two different ways. The first one is more direct, but the latter one is useful to clarify the physical significance of the solutions and to treat the finite-Γ case.

4.2.1 Direct Integration and Expansion

The first strategy to evaluate X_s is to directly integrate Eq. (47) in $\Gamma = 0$. The result is

$$X_s = \int_0^\infty dp \, e^{-\left(\Lambda - z\sqrt{R'}\right)(p-\hat{p}_s) - \frac{1}{2}Q'(p-\hat{p}_s)^2} = \sqrt{\frac{2\pi}{Q'}} e^{\frac{1}{2}\frac{(\Lambda - z\sqrt{R'})^2}{Q'}} H\left(y_s - z\sqrt{\frac{R'}{Q'}}\right), \tag{50}$$

where we put $y_s = (\Lambda - \hat{p}_s Q')/\sqrt{Q'}$ and we define the complementary error function

$$H(y) = \int_y^\infty Dz. \tag{51}$$

Thus

$$\begin{aligned}\left[\log V\right] = &\frac{1}{2}Q(Q' - R') + \frac{1}{2}RR' - \frac{M}{2}\left\{\frac{R}{E+Q-R} + \log\left(1 + \frac{Q-R}{E}\right)\right\} \\ &+ 2^N\left(\frac{\Lambda^2 + R'}{2Q'} + \frac{1}{2}\log 2\pi - \frac{1}{2}\log Q'\right) + \sum_s \int Dz \log H\left(y_s - z\sqrt{\frac{R'}{Q'}}\right).\end{aligned} \tag{52}$$

The definition of y_s allows us to introduce the learning edge in a natural way. Under assumption (49), $y_s = (\Lambda - \hat{p}_s Q')/\sqrt{Q'}$ diverges. If the target probability $\hat{p}_s \doteq e^{-N\omega_s}$ is small enough, *i.e.* if $\omega_s > q' - \lambda$, the dominant term in y_s is $\Lambda/\sqrt{Q'}$, which goes to $+\infty$ in the thermodynamic limit. On the contrary we find $y_s \to -\infty$ if $\omega_s < q' - \lambda$. This sharp difference defines the learning edge and the corresponding entropy as

$$\hat{\omega} = q' - \lambda, \ \hat{\sigma} = \sigma(\hat{\omega}) \tag{53}$$

Furthermore, based on assumption (49), $|y_s|$ increases faster with $N \gg 1$ than $\sqrt{R'/Q'}$, entailing that the argument of the complementary error function H in Eq. (50) is dominated

by y_s. Thus, the complementary error function can be replaced with its asymptotic behavior, depending on the sign of y_s,

$$H(y) \to \begin{cases} \frac{1}{\sqrt{2\pi}} \frac{1}{y} e^{-\frac{1}{2}y^2}, & (y \to \infty) \\ 1, & (y \to -\infty) \end{cases}. \tag{54}$$

Using this, we get

$$\begin{aligned} &\sum_s \int Dz \log H\left(y_s - z\sqrt{\frac{R'}{Q'}}\right) \\ &= \int Dz \sum_{s \in S} \left\{ -\frac{1}{2}\left(y_s - z\sqrt{\frac{R'}{Q'}}\right)^2 - \frac{1}{2}\log 2\pi - \log\left(y_s - z\sqrt{\frac{R'}{Q'}}\right)\right\} + \sum_{s \in L} 0. \end{aligned} \tag{55}$$

The logarithmic term in the above formula can be expanded according to Eq. (49) as

$$\log\left(y_s - z\sqrt{\frac{R'}{Q'}}\right) \approx \log \Lambda - \frac{1}{2}\log Q' - \frac{\hat{p}_s Q' + \sqrt{R'}z}{\Lambda} - \frac{1}{2}\frac{(\hat{p}_s Q' + \sqrt{R'}z)^2}{\Lambda^2}. \tag{56}$$

Linear terms with respect to z vanish upon Gaussian integration. We are left with

$$\begin{aligned} [\log V] \approx{}& \frac{1}{2}Q(Q' - R') + \frac{1}{2}RR' - \frac{M}{2}\left\{\frac{R}{E + Q - R} + \log\left(1 + \frac{Q - R}{E}\right)\right\} \\ &+ e^{N\hat{\sigma}}\left(\frac{\Lambda^2 + R'}{2Q'} + \frac{1}{2}\log 2\pi - \frac{1}{2}\log Q'\right) - 2^N\left(\log \Lambda - \frac{1}{2}\frac{R'}{\Lambda^2}\right) \\ &+ \left(\Lambda + \frac{Q'}{\Lambda}\right)\sum_{s \in S}\hat{p}_s - \frac{1}{2}\left(Q' - \frac{Q'^2}{\Lambda^2}\right)\sum_{s \in S}\hat{p}_s^2, \end{aligned} \tag{57}$$

where we have used

$$2^N - \sum_{s \in S} 1 = \sum_{s \in L} 1 \doteq e^{N\hat{\sigma}}, \ \sum_{s \in S} 1 \doteq 2^N. \tag{58}$$

These formulas are correct as long as $\hat{\omega} \leq \omega_0$.

4.2.2 Saddle-Point Approximation

Our second method to evaluate X_s in Eq. (50) is based on the saddle-point approximation. Writing the exponent as $f_s(p) = -\left(\Lambda - z\sqrt{R'}\right)(p - \hat{p}_s) - \frac{1}{2}Q'(p - \hat{p}_s)^2$ in Eq. (50) and differentiating with respect to p we get the solution of the saddle-point equation $f_s'(p^*) = 0$:

$$p^* = \hat{p}_s - \frac{\Lambda - z\sqrt{R'}}{Q'}. \tag{59}$$

As we know from the previous calculation $\frac{1}{N}\log(\Lambda/Q')$ determines the learning edge. If $\hat{p}_s > \Lambda/Q'$ the dominant contribution to p^* is $\hat{p}_s$ itself and becomes positive, implying the saddle point is in the feasibility domain:

$$X_s \approx \int_{-\infty}^{\infty} dx\, e^{f_s(p^*) - \frac{1}{2}|f_s''(p^*)|x^2} = e^{f_s(p^*)}\sqrt{\frac{2\pi}{|f_s''(p^*)|}} = e^{\frac{1}{2}\frac{(\Lambda - z\sqrt{R'})^2}{Q'}}\sqrt{\frac{2\pi}{Q'}}. \tag{60}$$

This result coincides with Eq. (50) with $H(y) \to 1, \quad y \to -\infty$. To check the validity of the saddle-point approximation we can compare the peak location $p^* \doteq \hat{p}_s$ to the width of the Gaussian, $\sigma = 1/\sqrt{|f_s''(p^*)|} = 1/\sqrt{Q'}$. Under assumption (49) we see that $\omega_s \leq \hat{\omega} = q' - \lambda < q'/2$ and thus $p^* > \sigma$ to dominant exponential-in-N order. The peak height $f_s(p^*)$ also diverges rapidly with N, and the saddle-point approximation is justified.

In the case of configurations $s \in S$ with small probabilities, $\omega_s > \hat{\omega}$, the dominant term to the right hand side of Eq. (59) is $-\Lambda/Q'$; p^* becomes negative and lies outside the domain of integration. Hence, the true saddle-point value is $p^* = 0$. We expand $f_s(p)$ around $p = 0$ up to the first order, and approximate the integral as

$$X_s \approx \int_0^\infty dp\, e^{f_s(0)+f_s'(0)p} = \frac{e^{(\Lambda - z\sqrt{R'})\hat{p}_s - \frac{1}{2}Q'\hat{p}_s^2}}{\Lambda - Q'\hat{p}_s - z\sqrt{R'}}. \tag{61}$$

This is again identical to Eq. (50) in the limit $H(y) \approx e^{-y^2/2}/(\sqrt{2p}y)$ with $y = y_s - z\sqrt{R'/Q'}$.

In addition, the saddle-point calculations above give us the expressions for the marginal measure ρ_s given in Sect. 3.2. Indeed, the marginal measure of the probability of the configuration s reads

$$\rho_s(p_s) = \frac{1}{X_s}\, e^{f_s(p_s)}. \tag{62}$$

Therefore ρ_s is approximately an exponentially-decaying function (27) for small target probabilities, and a Gaussian centered close to the corresponding target probability (26) for large target probabilities.

4.2.3 Equations of State and Their Solutions

Here we derive the equations of state (EOSs). Taking the derivatives of Eq. (57) with respect to the order parameters, we get

$$R' = \frac{MR}{(E+Q-R)^2},\ Q' = \frac{M}{E+Q-R}, \tag{63}$$

$$Q = \left(1 - 2\frac{Q'}{\Lambda^2}\right)\sum_{s\in S}\hat{p}_s^2 - \frac{2}{\Lambda}\sum_{t\in S}\hat{p}_s + e^{N\hat{\sigma}}\left(\frac{\Lambda^2 + R' + Q'}{Q'^2}\right), \tag{64}$$

$$D = Q - R = \frac{2^N}{\Lambda^2} + \frac{e^{N\hat{\sigma}}}{Q'}, \tag{65}$$

$$e^{N\hat{\sigma}}\frac{\Lambda}{Q'} - \frac{2^N}{\Lambda} + \left(1 - \frac{Q'}{\Lambda^2}\right)\sum_{s\in S}\hat{p}_s - \frac{Q'^2}{\Lambda^3}\sum_{s\in S}\hat{p}_s^2 = 0. \tag{66}$$

The order parameter D defined in Sect. 3 naturally appears. According to assumption (49), we know that all the terms with the factor Q'/Λ^2 are not dominant in Eqs. (63)–(66); the same statement applies to the terms with the factor $e^{N\hat{\sigma}}$ in Eqs. (64)–(66). Matching the dominant contributions to the remaining terms we obtain

$$Q \doteq \sum_{s\in S}\hat{p}_s^2\,, \quad \frac{2^N}{\Lambda} \doteq \sum_{s\in S}\hat{p}_s\,, \tag{67}$$

$$D \doteq \frac{2^N}{\Lambda^2}\,. \tag{68}$$

The physical meaning of the first equation is clear. The order parameter $Q = \sum_s \left\langle \left(p_s - \hat{p}_s\right)^2 \right\rangle$ quantifies the average squared distance of a distribution $\boldsymbol{p}$ (chosen with measure ρ) to the target distribution. The contributions coming from the large probabilities are negligible because, as we have seen in Sect. 4.2.2, the marginal measures for large probabilities are centered close to the target values, which makes $\sum_{s\in L} \left\langle \left(p_s - \hat{p}_s\right)^2 \right\rangle$ very small. However, for the configurations for small target probabilities, $\langle p_s^2 \rangle$ is much smaller than $\hat{p}_s^2$. Thus Q is, to the leading order in N, equal to $\sum_{s\in S} \hat{p}_s^2$. We see, in addition, from Eq. (67) that Λ ensures the normalization condition, and is equal to 2^N as long as most configurations are in the set S. According to Eq. (67), we find

$$q = r = \begin{cases} -\ell_2 & (\hat{\omega} < \omega_2) \\ \hat{\sigma} - 2\hat{\omega} & (\omega_2 \le \hat{\omega}) \end{cases}, \quad \lambda = \begin{cases} \log 2 & (\hat{\omega} < \omega_1) \\ \log 2 - \hat{\sigma} + \hat{\omega} & (\omega_1 \le \hat{\omega}) \end{cases}, \quad d = \log 2 - 2\lambda. \tag{69}$$

The other order parameters q', r' and $\hat{\omega}$ are computed accordingly.

Small-E case. We can ignore E when it is smaller than $D = Q - R$. In this case we get the EOSs for the parameters q' and r':

$$r' = \alpha + r - 2d\,, \quad q' = \alpha - d, \tag{70}$$

and the learning edge is self-consistently determined by Eq. (53). Solving theses equations, we obtain

Phase I ($\hat{\omega} < \omega_2$):

$$\begin{aligned} &\lambda = \log 2,\ \hat{\omega} = \alpha,\ q' = \alpha + \log 2, \\ &q = r = -\ell_2,\ r' = \alpha + 2\log 2 - \ell_2,\ d = -\log 2. \end{aligned} \tag{71}$$

Phase II ($\omega_2 \le \hat{\omega} < \omega_1$):

$$\begin{aligned} &\lambda = \log 2,\ \hat{\omega} = \alpha,\ q' = \alpha + \log 2,\ q = r = \sigma(\alpha) - 2\alpha, \\ &r' = \sigma(\alpha) - \alpha + 2\log 2,\ d = -\log 2. \end{aligned} \tag{72}$$

Phase III ($\omega_1 < \hat{\omega}$):

$$\begin{aligned} &\lambda = \log 2 + \hat{\omega} - \alpha,\ \hat{\omega} = \sigma^{-1}(\alpha),\ q' = \log 2 + 2\hat{\omega} - \alpha,\ q = r = \sigma(\hat{\omega}) - 2\hat{\omega}, \\ &r' = 2(\log 2 + \hat{\omega} - \alpha) = 2\lambda,\ d = -\log 2 + 2\hat{\sigma} - 2\hat{\omega}. \end{aligned} \tag{73}$$

The critical ratio separating phases I and II is given by $\alpha_2 = \omega_2$, while the critical ratio associated to the transition from II to III is $\alpha_1 = \omega_1$. We can easily check that those solutions satisfy the assumptions (49), and are therefore self-consistent.

Large-E case. If $E \doteq e^{N\epsilon}$ is larger than $D \doteq e^{Nd}$, *i.e.* if $\epsilon > d$, the EOSs of the parameters q' and r' are modified as

$$r' = \alpha + r - 2\epsilon\,, \quad q' = \alpha - \epsilon\,. \tag{74}$$

These large-tolerance EOSs become valid beyond the critical values $\hat{\epsilon}(\alpha)$:

$$\hat{\epsilon}(\alpha) = -\log 2 \text{ (Phase I, II)},\ \hat{\epsilon}(\alpha) = -\log 2 + 2\hat{\sigma}(\alpha) - 2\hat{\omega}(\alpha) \text{ (Phase III)}. \tag{75}$$

To distinguish the solutions above from the ones of the small-E case we denote phase I with large tolerance by I_{LT} as explained in Sect. 3.2.3. Phases $\mathrm{II}_{\mathrm{LT}}$ and $\mathrm{III}_{\mathrm{LT}}$ are defined in the same way. The solutions of the large-tolerance case become

Phase I_{LT} ($\hat{\omega} < \omega_2$):

$$\lambda = \log 2,\ \hat{\omega} = \alpha - \epsilon - \log 2,\ q' = \alpha - \epsilon = \hat{\omega} + \log 2,\ q = r = -\ell_2,$$
$$r' = \alpha - 2\epsilon - \ell_2 = 2\log 2 + 2\hat{\omega} - \ell_2 - \alpha,\ d = -\log 2. \tag{76}$$

Phase $\mathrm{II}_{\mathrm{LT}}$ ($\omega_2 \le \hat{\omega} < \omega_1$):

$$\lambda = \log 2,\ \hat{\omega} = \alpha - \epsilon - \log 2,\ q' = \alpha - \epsilon = \hat{\omega} + \log 2,\ q = r = \hat{\sigma} - 2\hat{\omega},$$
$$r' = \alpha + \hat{\sigma} - 2\hat{\omega} - 2\epsilon = \hat{\sigma} + 2\log 2 - \alpha,\ d = -\log 2. \tag{77}$$

Phase $\mathrm{III}_{\mathrm{LT}}$ ($\omega_1 \le \hat{\omega}$):

$$\lambda = \log 2 + \hat{\omega} - \hat{\sigma},\ \hat{\sigma} - 2\hat{\omega} = \log 2 + \epsilon - \alpha,\ q' = \alpha - \epsilon,\ q = r = \hat{\sigma} - 2\hat{\omega},$$
$$r' = \alpha + \hat{\sigma} - 2\hat{\omega} - 2\epsilon = 2\log 2 - \alpha - \hat{\sigma} + 2\hat{\omega},\ d = -\log 2 + 2\hat{\sigma} - 2\hat{\omega}. \tag{78}$$

It is instructive to examine the validity of the condition $\epsilon > d$ in phase $\mathrm{III}_{\mathrm{LT}}$. To do so we need to know the rate of increase of $\hat{\sigma} - \hat{\omega}$ with ϵ at fixed α. Differentiating the equation determining the learning edge in Eq. (78), we get

$$\frac{\partial(\hat{\sigma} - \hat{\omega})}{\partial\epsilon} = 1 + \frac{\partial\hat{\omega}}{\partial\epsilon}. \tag{79}$$

This equation can be written as

$$-1 \le \frac{\partial\hat{\omega}}{\partial\epsilon} = \left(\frac{\partial\hat{\sigma}}{\partial\hat{\omega}} - 2\right)^{-1} \le -\frac{1}{2}. \tag{80}$$

The inequalities come from the condition $\partial\hat{\sigma}/\partial\hat{\omega} < 1$, which holds in phase III. This implies that d increases more slowly than ϵ itself in $\mathrm{III}_{\mathrm{LT}}$, since $\partial d/\partial\epsilon = 2\left(1 + \partial\hat{\omega}/\partial\epsilon\right) \le 1$. Thus the necessary condition $\epsilon > d$ is satisfied in phase $\mathrm{III}_{\mathrm{LT}}$, as it should. This condition will be useful for the study of the stability of the RS ansatz.

4.3 Large-Γ Case

4.3.1 Saddle-Point Approximation

We use the saddle-point approximation in the integral defining X_s (47) as in Sect. 4.2.2. We call $f_s(p)$ the exponent in the integral, and write down the derivatives

$$f_s(p) = -\frac{1}{2}Q'(p - \hat{p}_s)^2 - (\Lambda - z\sqrt{R'})(p - \hat{p}_s) - \Gamma p\log p, \tag{81}$$
$$f_s'(p) = -Q'(p - \hat{p}_s) - (\Lambda - z\sqrt{R'}) - \Gamma(\log p + 1), \tag{82}$$
$$f_s''(p) = -Q' - \frac{\Gamma}{p}. \tag{83}$$

The main difference with the $\Gamma = 0$ case is the presence of the term $p\log p$, singular at $p = 0$. As a result the functional behavior around $p = 0$ changes completely, and there is always a peak in $[0, \infty]$. This implies that the saddle-point approximation can be applied, irrespective of the target probability value. The saddle-point equation $f_s(p^*) = 0$ can be written as

$$p^* = \hat{p}_s - \frac{\Lambda - z\sqrt{R'} + \Gamma(1 + \log p^*)}{Q'}. \tag{84}$$

Let us make the following assumption, correct when Γ is large enough,

$$\frac{\Lambda}{\Gamma} = O(N). \tag{85}$$

This relation, in turns, defines the meaning of 'large' Γ. This assumption, combined with Eq. (49), gives us the value of the learning edge from Eq. (84): $\hat{\omega} = q' - \lambda = q' - \gamma$. For configurations with large target probabilities the dominant term is $\hat{p}_s$ in Eq. (84), and an iterative substitution yields

$$p_s^* \approx \hat{p}_s - \frac{\Lambda - z\sqrt{R'} + \Gamma(1 + \log \hat{p}_s)}{Q'}, \quad (s \in L). \tag{86}$$

Expanding $f_s(p)$ up to the second order, we may evaluate X_s.

For configurations corresponding to small target probabilities, the saddle point is very different. We *a priori* postulate the expression of the solution

$$p_s^* \approx \tilde{p} + \Delta_s, \quad (s \in S), \tag{87}$$

where the dominant term $\tilde{p}$ is

$$\tilde{p} = e^{-1-\left(\Lambda - Q'\hat{p}_s - z\sqrt{R'}\right)/\Gamma} < \hat{p}_s, \quad (\forall s \in L)\,. \tag{88}$$

Appendix 1 gives a reasoning of this form. $\tilde{p}$ is required to take this value to satisfy the normalization condition. According to the saddle-point calculation above we have

$$1 = \left\langle \sum_s p_s \right\rangle \approx \sum_{s \in S} \tilde{p} + \sum_{s \in L} \hat{p}_s. \tag{89}$$

For small learning edge values, $\hat{\omega} < \omega_1$, configurations with small probabilities dominate, implying that $\sum_{s \in S} \tilde{p} \doteq 2^N \tilde{p} \approx 1 \Rightarrow \tilde{p} \doteq 2^{-N}$. Thus, $\tilde{p}$ is automatically determined by the normalization condition. Moreover, the correction term Δ_s in Eq. (87) is determined by expanding Eq. (84) with respect to Δ_s to the first order and solving the resulting equation. We find

$$\Delta_s = -\frac{Q'\tilde{p}^2}{Q'\tilde{p} + \Gamma}. \tag{90}$$

This expansion is valid if $\tilde{p} > |\Delta_s|$, yielding an trivial inequality $1 > Q'\tilde{p}/(Q'\tilde{p} + \Gamma)$.

4.3.2 Equations of State and Validity of the Saddle-Point Approximation

It is convenient to compute the derivatives with respect to the order parameters of the expression in Eq. (46). The result is

$$Q = \sum_s \int Dz \left\langle (p - \hat{p}_s)^2 \right\rangle_{X_s}, \tag{91}$$

$$D = Q - R = \sum_s \int Dz \left\{ \left\langle (p - \hat{p}_s)^2 \right\rangle_{X_s} - \left\langle p - \hat{p}_s \right\rangle_{X_s}^2 \right\}, \tag{92}$$

$$0 = \sum_s \int Dz \left\langle p - \hat{p}_s \right\rangle_{X_s} \tag{93}$$

where we define

$$\langle(\cdots)\rangle_{X_s} = \frac{1}{X_s}\int_0^\infty dp\ (\cdots)e^{f_s(p)}. \tag{94}$$

The EOSs for R' and Q' are identical to Eq. (63), and are thus omitted above. The average value of p is replaced with the saddle-point value p_s^*. We need to quantify the fluctuations around the saddle point to estimate the terms in Eq. (92), that is,

$$\left\langle \left(p-\hat{p}_s\right)^2\right\rangle_{X_s} - \left\langle p-\hat{p}_s\right\rangle_{X_s}^2 \approx \frac{e^{f_s(p_s^*)}\int dx\ x^2\, e^{-\frac{f_s''(p_s^*)}{2}x^2}}{e^{f_s(p_s^*)}\int dx\ e^{-\frac{|f_s''(p_s^*)|}{2}x^2}} = \frac{1}{|f_s''(p_s^*)|} \doteq \begin{cases} Q'^{-1} & (\boldsymbol{s}\in L) \\ \frac{\tilde{p}}{\Gamma} & (\boldsymbol{s}\in S) \end{cases}. \tag{95}$$

Note that, for $\boldsymbol{s}\in S$, we can write $p_s^* - \hat{p}_s \doteq \tilde{p} - \hat{p}_s$. These considerations lead us to

$$\begin{aligned} Q &= \sum_{s\in S}\int Dz(\tilde{p}-\hat{p}_s)^2 + \sum_{s\in L}\int Dz\left\{\left(-\frac{\Lambda+\Gamma+\Gamma\log\hat{p}_s}{Q'} - z\frac{\sqrt{R'}}{Q'}\right)^2\right\} \\ &\doteq 2^N\tilde{p}^2 - 2\tilde{p}\sum_{s\in S}\hat{p}_s + \sum_{s\in S}\hat{p}_s^2 + e^{N\hat{\sigma}}\left(\frac{\Lambda+\Gamma}{Q'}\right)^2\left\{\left(1-\frac{N\hat{\omega}\Gamma}{\Lambda+\Gamma}\right)^2 + \frac{R'}{(\Lambda+\Gamma)^2}\right\}. \end{aligned} \tag{96}$$

$$D = \sum_{s\in S}\int Dz\frac{\tilde{p}}{\Gamma} + \sum_{s\in L}\int Dz(Q')^{-1} \doteq 2^N\frac{\tilde{p}}{\Gamma} + \frac{e^{N\hat{\sigma}}}{Q'}, \tag{97}$$

$$\begin{aligned} 0 &= \sum_{s\in S}\int Dz\left(\tilde{p}-\hat{p}_s\right) + \sum_{s\in L}\int Dz\left(-\frac{\Lambda+\Gamma+\Gamma\log\hat{p}_s}{Q'} - z\frac{\sqrt{R'}}{Q'}\right) \\ &\doteq 2^N\tilde{p} - \sum_{s\in S}\hat{p}_s - e^{N\hat{\sigma}}\left(\frac{\Lambda+\Gamma}{Q'}\right)\left(1-\frac{N\hat{\omega}\Gamma}{\Lambda+\Gamma}\right). \end{aligned} \tag{98}$$

The dominant terms turn out to be

$$Q \doteq \sum_{s\in S}\hat{p}_s^2, \tag{99}$$

$$D \doteq 2^N\frac{\tilde{p}}{\Gamma}, \tag{100}$$

$$2^N\tilde{p} \doteq \sum_{s\in S}\hat{p}_s, \tag{101}$$

with the corresponding exponents,

$$q = r = \begin{cases} -\ell_2 & (\hat{\omega}<\omega_2) \\ \hat{\sigma}-2\hat{\omega} & (\omega_2\le\hat{\omega}) \end{cases},\quad d = \begin{cases} -\gamma & (\hat{\omega}<\omega_1) \\ \hat{\sigma}-\hat{\omega}-\gamma & (\omega_1\le\hat{\omega}) \end{cases}, \tag{102}$$

$$\lambda = \gamma + \frac{\log N}{N} + \frac{1}{N}\begin{cases} \log\log 2 & (\hat{\omega}<\omega_1) \\ \log(\log 2+\hat{\omega}-\hat{\sigma}) & (\omega_1\le\hat{\omega}) \end{cases}. \tag{103}$$

Small-E case. Again we can neglect $E(\ll D)$. The EOSs for q' and r' and the equation for the learning edge are fully identical to Eqs. (70) and (53). Each phase is then characterized as follows:

Phase $\mathrm{I_{ME}}$ ($\hat{\omega} < \omega_2$):

$$\lambda = \gamma + \frac{\log(N \log 2)}{N},\ \hat{\omega} = \alpha,\ q' = \alpha + \gamma,\ q = r = -\ell_2,$$
$$r' = \alpha + 2\gamma - \ell_2,\ d = -\gamma. \tag{104}$$

Phase $\mathrm{II_{ME}}$ ($\omega_2 \le \hat{\omega} < \omega_1$):

$$\lambda = \gamma + \frac{\log(N \log 2)}{N},\ \hat{\omega} = \alpha,\ q' = \alpha + \gamma,\ q = r = \sigma(\alpha) - 2\alpha,$$
$$r' = \sigma(\alpha) - \alpha + 2\gamma,\ d = -\gamma. \tag{105}$$

Phase $\mathrm{III_{ME}}$ ($\omega_1 < \hat{\omega}$):

$$\lambda = \gamma + \frac{\log(N(\log -\alpha + \hat{\omega})2)}{N},\ \hat{\omega} = \sigma^{-1}(\alpha),\ q' = \hat{\omega} + \gamma,\ q = r = \alpha - 2\hat{\omega},$$
$$r' = 2\gamma,\ d = \hat{\sigma} - \hat{\omega} - \gamma. \tag{106}$$

It is easy to check that the assumption Eq. (49) is satisfied by these solutions. Note that the learning edge does not depend on γ.

We now examine the validity of the saddle-point approximation for X_s, to determine when the solutions above are correct. To do so we compare the width of the Gaussian with the location of its peak value, and check that the peak height diverges. For example, in the phase $\mathrm{I_{ME}}$,

Peak location:

$$p_s^* \doteq \begin{cases} \hat{p}_s \doteq e^{-N\omega_s}, & (\boldsymbol{s} \in L) \\ \tilde{p} \doteq e^{-N \log 2}, & (\boldsymbol{s} \in S) \end{cases}. \tag{107}$$

Height:

$$f_s(p_s^*) \doteq \begin{cases} e^{N(\gamma + \hat{\omega} - 2\omega_s)}, & (\boldsymbol{s} \in L) \\ e^{N(\gamma - \log 2)}, & (\boldsymbol{s} \in S) \end{cases}. \tag{108}$$

Width:

$$\sigma_s = |f_s''(p_s^*)|^{-1/2} = \sqrt{\frac{p_s^*}{Q' p_s^* + \Gamma}} \doteq \begin{cases} (Q')^{-1/2} \doteq e^{-N\frac{\gamma + \hat{\omega}}{2}}, & (\boldsymbol{s} \in L) \\ \sqrt{\frac{\tilde{p}}{\Gamma}} \doteq e^{-N\frac{\gamma + \log 2}{2}}, & (\boldsymbol{s} \in S) \end{cases}. \tag{109}$$

According to these relations the saddle-point approximation is correct if $\gamma \ge \log 2 \equiv \gamma_c$. The same approach can be applied to $\mathrm{II_{ME}}$ and $\mathrm{III_{ME}}$, with the resulting critical values: $\gamma_c = \log 2$ in $\mathrm{II_{ME}}$ and $\gamma_c = \log 2 + \hat{\omega} - \hat{\sigma}$ for $\mathrm{III_{ME}}$. These critical values are consistent with Eq. (85), as Λ is a continuous function of Γ.

This continuous change also implies that, in the $0 < \Gamma < \Gamma_c = e^{N\gamma_c}$ region, all solutions coincide with the ones found for $\Gamma = 0$. Consider for instance the peak value of the marginal measure over p_s for $\boldsymbol{s} \in S$. Assuming the order parameters are given by their expressions in the $\Gamma = 0$ case, we see that the condition Eq. (49) is satisfied, which implies that Eq. (87) is valid. The saddle point for $\boldsymbol{s} \in S$, $\tilde{p} = e^{-1-(\Lambda - Q'\hat{p}_s - z\sqrt{R'})/\Gamma}$, decays in a double-exponential manner with respect to N, as Λ is exponentially larger than Γ. Similarly, the peak height $f_s(p_s^*)$ rapidly goes to zero for $\boldsymbol{s} \in S$. The peak location, $\tilde{p}$, decreases faster than the width $\sigma_s \doteq \sqrt{\tilde{p}/\Gamma}$. As a consequence the functional form $e^{f_s(p)}$ rapidly converges to a pure exponential distribution, as in the $\Gamma = 0$ case.

Large-E case. The EOSs for q' and r' coincide with Eq. (74). The critical values of E are determined by $E \doteq D$:

$$\hat{\epsilon}(\alpha,\gamma) = -\gamma \text{ (Phase I, II)}, \;\; \hat{\epsilon}(\alpha,\gamma) = \hat{\sigma}(\alpha) - \hat{\omega}(\alpha) - \gamma \text{ (Phase III)}. \tag{110}$$

New solutions can appear for $\epsilon > \hat{\epsilon}(\alpha,\gamma)$, but should be discarded as they violate the condition $E < \Gamma$, see discussion in Sect. 3.3.

4.4 Stability of the Replica Symmetry

We study the de Almeida-Thouless stability of the replica-symmetric (RS) Ansatz, see Eq. (40), against fluctuations in the replica space [9]. Detailed calculations are reported in the Appendix 2. The outcome of the calculation is the following stability condition against replicon fluctuations:

$$\frac{M}{(E+Q-R)^2}\left\{\sum_s \int Dz \left(\frac{\partial^2 \log X_s}{\partial \Lambda^2}\right)^2\right\} \leq 1. \tag{111}$$

The term in the brackets is asymptotically equivalent to $e^{N\hat{\sigma}}/Q'^2$. This can be shown in the small-Γ case ($\Gamma \leq \Gamma_c$) using Eqs. (60), (61), and in the large-Γ case ($\Gamma > \Gamma_c$) with the relation $\partial^2 \log X_s/\partial\Lambda^2 = \langle (p-\hat{p})^2\rangle_{X_s} - \langle (p-\hat{p})\rangle^2_{X_s}$ and Eq. (95). From Eq. (63), we obtain a transparent interpretation of the stability condition

$$\frac{M}{(E+Q-R)^2}\left(\frac{e^{N\hat{\sigma}}}{Q'^2}\right) = \frac{e^{N\hat{\sigma}}}{M} \doteq e^{N(\hat{\sigma}-\alpha)} \leq 1. \tag{112}$$

Therefore the RS solution becomes unstable if the number of large-probability configurations exceeds the number of constraints. Inserting the solutions for $\hat{\sigma}$ in the small E regime, see Eqs. (71)–(73) and (104)–(106), we find that, irrespective of the value of Γ, the RS ansatz is stable in phases I and II but is only marginally stable in phase III. Marginal stability means here that $\hat{\sigma}$ is equal to α in phase III. Our calculation, limited to the leading order in N, cannot decide whether phase III is actually stable.

In the large tolerance E regime, the RS solution is stable across all phases, even in phase III_{LT}. Simple calculation based on Eq. (78) yields $\hat{\sigma} - \alpha = d - \epsilon < 0$, where the last inequality is proved at the end of Sect. 4.2.3.

To summarize, the RS ansatz is stable in all phases, but only marginally in phases III and III_{ME}. This result may be related to the 'simple' structure of the version space. Consider the case of zero tolerance, $E = 0$, and two distribution vectors, $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, in the version space. Any linear combination of these two vectors, $\boldsymbol{p}_t = t\boldsymbol{p}_1 + (1-t)\boldsymbol{p}_2$ with $t \in [0,1]$, is a normalized distribution and lies in the version space. Hence the version space is convex and connected. The instability of RS ansatz, which is usually related to the appearance of many disconnected and far apart components, may therefore not take place.

5 Numerical Simulations

We now present a numerical procedure to sample the space of distributions with the measure ρ. Due to the exponential growth of the version space with N this procedure is applied in practice to small values of $N \leq 10$. However the results confirm the analytical calculations reported above, and provide insights about finite-size effects. We restrict to the case of zero tolerance ($E = 0$) throughout this section.

5.1 Sampling Algorithm

We resort to a Monte Carlo (MC) sampling method, in which the distribution $\boldsymbol{p}$ is updated at discrete time steps. Each step corresponds to a random change of the current probability vector from $\boldsymbol{p}$ to $\boldsymbol{p}' = \boldsymbol{p} + \Delta\boldsymbol{p}$. The move $\Delta\boldsymbol{p}$ must satisfy the following conditions:
Orthogonality to observable vectors: $\Delta\boldsymbol{p} \cdot \boldsymbol{v}^{\mu} = 0, \forall\mu$.
Normalization: $\sum_s \Delta p_s = \Delta\boldsymbol{p} \cdot \mathbf{1} = 0$ where $\mathbf{1} = (1, 1, \ldots, 1)$.
Positivity: $p_s + \Delta p_s \geq 0, \forall s$.

The orthogonality condition ensures that the constraints keep being satisfied at all steps provided they are fulfilled by the initial value of the distribution. The same statement applies to the normalization condition; we will specify below how the initial condition is chosen. The orthogonality and normalization conditions restrict the possible directions $\boldsymbol{w}$ (normalized to unity) of the move. Once this direction $\boldsymbol{w}$ is chosen we determine the range $[x_{min}; x_{max}]$ of the allowed amplitudes x of the move $\Delta\boldsymbol{p} = x\ \boldsymbol{w}$ to fulfill the positivity constraint:

$$x_{\min} = \min_s \max\left(\frac{-p_s}{w_s}, \frac{1-p_s}{w_s}\right), \quad x_{\max} = \max_s \min\left(\frac{-p_s}{w_s}, \frac{1-p_s}{w_s}\right). \tag{113}$$

Any intermediate values between these two bounds may be chosen uniformly at random. Next, we calculate the entropy difference $\Delta S = S(\boldsymbol{p}') - S(\boldsymbol{p})$ and accept the move $\boldsymbol{p} \to \boldsymbol{p}'$ according to the Metropolis rule, *i.e.* with probability

$$p_{\text{accept}} = \min\left(1, e^{\Gamma\Delta S}\right), \tag{114}$$

and reject it (leave $\boldsymbol{p}$ unchanged) with probability $1 - p_{\text{accept}}$. A picture illustrating one Monte Carlo step is shown in Fig. 8.

To implement the algorithm, it is convenient not to assume that the observables are Gaussianly distributed, as in the analytical treatment. Instead we consider the Fourier modes $\boldsymbol{w}^{\boldsymbol{k}}$ on the N-dimensional hypercube, where $\boldsymbol{k} = (k_1, k_2, ..., k_N)$, with $k_i = 0, 1$ for each i, denote the wave-number configurations. The 2^N components of those Fourier modes are given by

$$\left(\boldsymbol{w}^{\boldsymbol{k}}\right)_s = \frac{1}{\sqrt{2^N}} \prod_{i=1}^{N} (s_i)^{k_i}. \tag{115}$$

The Fourier modes $\{\boldsymbol{w}^{\boldsymbol{k}}\}_{\boldsymbol{k}}$ form a complete orthonormal basis of the 2^N-dimensional vector space. Note that $\boldsymbol{w}^{\mathbf{0}} = \mathbf{1}/\sqrt{2^N}$, where $\mathbf{0}$ is the all-zero wave-number configuration.

In the analytical calculation of Sect. 4 we chose the observable $\boldsymbol{v}$ to be random vectors, with Gaussian components. Any two randomly chosen vectors have very small scalar product

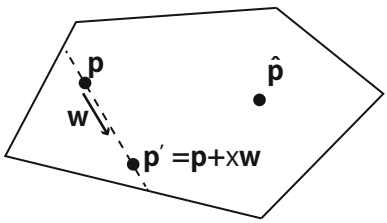

Fig. 8 Schematic picture of the version space and of one Monte Carlo step. The version space is convex, and includes the target distribution $\hat{\boldsymbol{p}}$. From a distribution $\boldsymbol{p}$, a direction $\boldsymbol{w}$ inside the solution space is randomly chosen. We choose a point on the segment along this direction (dashed line) uniformly at random, and $\boldsymbol{p}$ is updated to the chosen point $\boldsymbol{p}'$ with probability p_{accept}, see Eq. (114)

with high probability in the large-N limit. In our numerical simulation we rather choose the M observables uniformly at random over the discrete set of $2^N - 1$ Fourier modes $\boldsymbol{w}^{\boldsymbol{k}}$, with $\boldsymbol{k} \neq \boldsymbol{0}$. This choice is convenient since the Fourier modes are orthogonal by construction, implying that the orthogonality condition mentioned above is automatically satisfied as soon as we choose the direction of the move $\boldsymbol{w}$ as one of the Fourier modes outside the set of observables. The normalization condition is easily satisfied as long as we exclude $\boldsymbol{w}^{\boldsymbol{0}}$ from the set of possible directions for the move. We are thus left with a set of $2^N - M - 1$ Fourier modes, each of which is a possible direction for the Monte Carlo move. Clearly our Monte Carlo Markov Chain satisfies detailed balance and is irreducible.

5.2 Results

Using the above sampling procedure we calculate several quantities of interest including order parameters, histograms of p_s, and spin-spin correlations. The target distribution we consider is again the ISM (6) with $H = 0.5$.

5.2.1 Check of Equilibration

We have run simulations for ten different samples ($N_{\text{sample}} = 10$) for sizes $N = 4, 6$ and 8, and for two samples ($N_{\text{sample}} = 2$) for size $N = 10$ to compute the values of the order parameters. The error bars are estimated through the standard deviation σ_{sample} across the samples:

$$\text{Error bar} = \frac{\sigma_{\text{sample}}}{\sqrt{N_{\text{sample}} - 1}}. \tag{116}$$

A Monte Carlo step is defined as one attempted move, irrespective of its acceptance. Fig. 9 shows the plot of the order parameter q against the number of Monte Carlo steps, indicating how the system approaches equilibrium.

From the figure we see that thermalization becomes drastically faster for larger Γ. For example, simulations with $M = 32$ constraints and $N = 10$ spins require at least $N_{\text{MC}} \approx 2 \times 10^8$ steps for thermalization when $\Gamma = 0$, while $N_{\text{MC}} = 2 \times 10^7$ steps seem to be sufficient when $\Gamma \geq 2^{10}$. This trend is easy to understand, as an increase in the entropic bias concentrates more and more the measure around ME distribution, hence shrinking the space to be sampled.

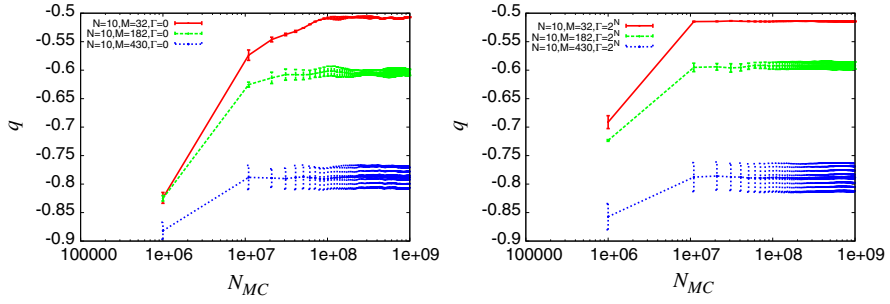

Fig. 9 Plots of q versus the number of Monte Carlo (MC) steps for $N = 10$ spins and for different values of the number M of constraints. The *left panel* corresponds to $\Gamma = 0$ (no entropic bias), and the *right panel* to $\Gamma = 2^N = 1024$

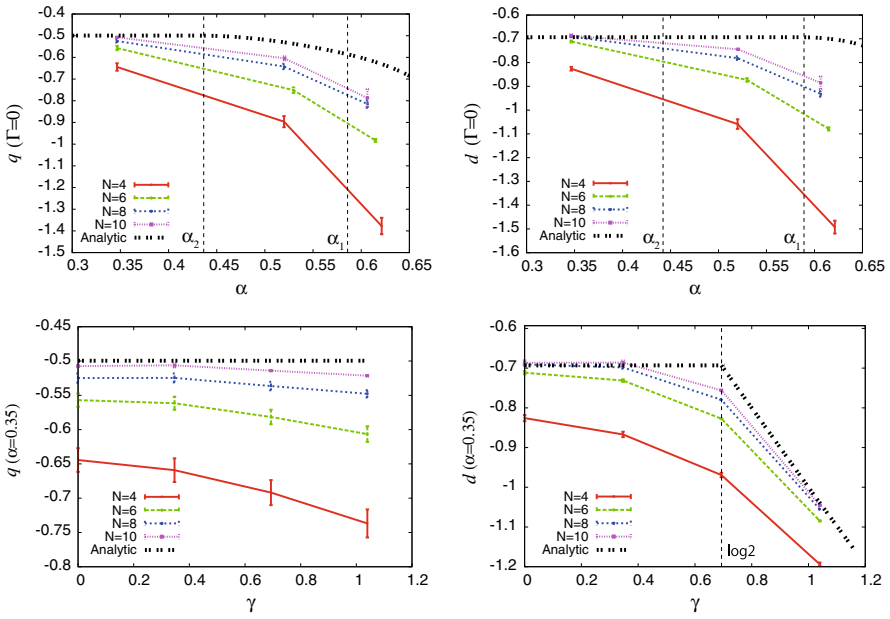

Fig. 10 Order parameters q (*left*) and d (*right*) of the ISM with $H = 0.5$, computed from Monte Carlo simulations and plotted versus α (*upper row*) and γ (*lower row*). Analytical predictions are shown with the *black curves*. The phase transition from small to large Γ can be best guessed in the *lower*, *right panel*. Finite-size effects seem to be stronger for larger α

The actual number of Monte Carlo steps used in our simulation varied depending on the quantities we wanted to estimate. To compute the values of the order parameters we chose $N_{\mathrm{MC}} = 1 \times 10^7$ for $N = 4$ and 6, $N_{\mathrm{MC}} = 3 \times 10^7$ for $N = 8$, and $N_{\mathrm{MC}} = 3 \times 10^8$ for $N = 10$. One third of those steps are discarded in the computation of the averages. To obtain accurate histogram of the distributions of the order parameters, however, we chose typically 5 times more MC steps.

5.2.2 Order Parameters, Marginal Measures, and Spin Correlations

To confirm our analytical prediction and to estimate finite-size effects we compute the order parameters for various values of M, Γ and N, and report the results in Fig. 10. Those figures clearly show that the analytical prediction, black curves, are consistent with the numerical results for sizes as small as $N = 10$. The phase transitions are well captured. In particular, the phase transition to the ME phase is clearly seen in the behaviour of d (right lower panel). This agreement show that our analytical findings, derived in the infinite-N limit, are numerically accurate even for small sizes.

We have also computed the histograms of the single-configuration probabilities p_s, with the results shown in Figs. 11 and 12. Due to the symmetry of the ISM the target probabilities $p_{\boldsymbol{s}}$ depend only on the number of, say, -1 spins in the configuration $\boldsymbol{s}$, which we call N_s^-. We checked that this symmetry is approximately recovered in the histograms despite the noise introduced by the Monte Carlo sampling. Therefore we show only one among the p_s with the same N_s^- in those figures.

Figure 11 shows the histograms corresponding to different N_s, for $N = 8$ and $\Gamma = 0$, and two values of the number M of constraints. Two characteristic shapes of the histograms

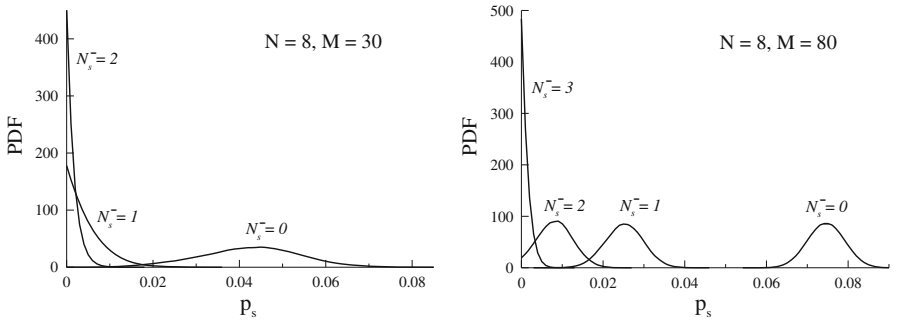

Fig. 11 Histograms of single-configuration probabilities $p_{\boldsymbol{s}}$ for $N = 8$ and $\Gamma = 0$; the numbers N_s^- of negative spins in the configurations are indicated on the figure. The *left panel* corresponds to $M = 30$, and the *right panel* to $M = 80$

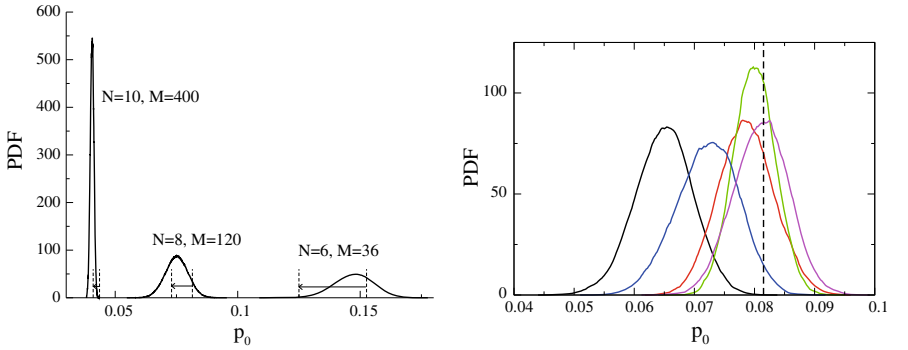

Fig. 12 (*Left*) System-size dependence of histograms of p_0 (corresponding to the configuration **1** all plus spins), for fixed $\alpha \approx 0.598$ and $\Gamma = 0$. For each histogram, the rightmost *dashed line* show the location of the target value $\hat{p}_{\mathbf{1}}$, while the leftmost *dashed line* shows $\hat{p}_{\mathbf{1}} - 1/M$. (*Right*) Sample-to-sample fluctuations of the histogram of p_0 for $N = 8$, $M = 80$ and $\Gamma = 0$. The *dashed vertical line* represents the target value

emerge. One looks like roughly Gaussian, while the other typical behaviour resembles an exponential distribution. These shapes are in very good agreement with the two possible functional dependences of $\rho_{\boldsymbol{s}}(p_{\boldsymbol{s}})$ upon $p_{\boldsymbol{s}}$ given in Fig. 3. A larger number of Gaussian-like histograms is found for $M = 80$ than for $M = 30$, implying that more target probabilities are learned in the former case than in the latter, as expected from the theory. The size and sample dependence of the histograms of p_0 (corresponding to the configuration **1** with $N_s^- = 0$ minus spins) are examined in Fig. 12. The corresponding target probability $\hat{p}_{\mathbf{1}}$ is the largest one of the ISM and is accurately learned for the values of M corresponding to the figures. From the left panel of Fig. 12, we see that our analytical prediction regarding $\rho_{\boldsymbol{s}}(p_{\boldsymbol{s}})$ being centered in $\hat{p}_{\boldsymbol{s}} - 1/M$, becomes more and more accurate as the system size increases. An excellent agreement with the the prediction is reached for $N = 10$. Sample-to-sample fluctuations of the histogram of $p_{\mathbf{1}}$ are shown in the right panel. The choice of the constraints produces moderate fluctuations in the location of the peak height of the Gaussian-like histograms.

Last of all we show in Fig. 13 the histograms of multi-spin correlations, for specific subsets of spins. We consider in particular

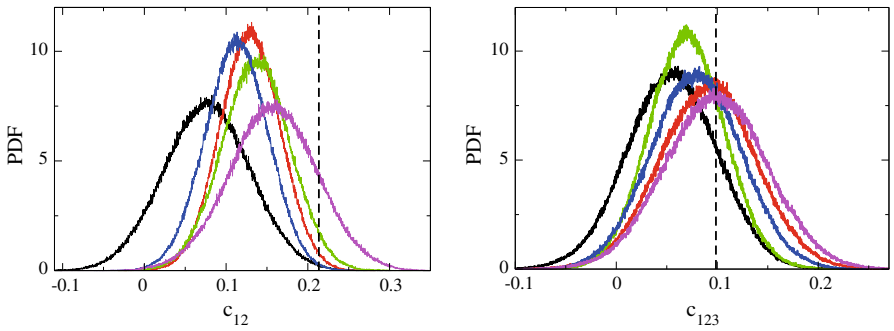

Fig. 13 Histograms of pairwise (c_{12}, *left*) and triplet (c_{123}, *right*) correlations for a system of $N = 8$ spins. *Vertical lines* indicate the target distribution values. Each histogram corresponds to one sample of $M = 80$ observables

$$c_{12} = \sum_{\boldsymbol{s}} p_{\boldsymbol{s}}\, s_1 s_2, \quad c_{123} = \sum_{\boldsymbol{s}} p_{\boldsymbol{s}}\, s_1 s_2 s_3. \tag{117}$$

We see that the predicted values of c are in fair agreement with the target distribution values, but both the thermal and the sample-to-sample fluctuations are rather large, compared to the results of Fig. 12.

6 Conclusion

In this paper we have investigated the properties of the space of probability distributions (over a large set of discrete variable configurations), constrained to reproduce many average values of observables, computed according to a target distribution $\hat{\boldsymbol{p}}$. For zero tolerance $E = 0$, the space of admissible distributions $\boldsymbol{p}$ defines the version space. The version space contains many distributions of interest, in addition to the target distribution $\hat{\boldsymbol{p}}$, such as the center-of-mass distribution, $\langle \boldsymbol{p} \rangle$, which is the flat average over all admissible distributions, and the maximum entropy distribution, $\boldsymbol{p}_{ME}$. In the case of finite tolerance $E > 0$ the version space acquires a probabilistic meaning. We introduce a probability measure ρ over the space of distributions, giving more weight to the distributions $\boldsymbol{p}$ in good agreement with the target values of the observables. The measure ρ may furthermore be biased to favour distributions $\boldsymbol{p}$ with large entropies $S(\boldsymbol{p})$. To do so we introduce a multiplicative factor in the measure ρ, growing exponentially with $\Gamma\ S(\boldsymbol{p})$. The coefficient Γ acts as an inverse temperature: $\Gamma = 0$ allows us to find back the unbiased measure, while, for $\Gamma \to \infty$, the measure becomes fully concentrated around $\boldsymbol{p}_{ME}$. Varying the value of Γ allows us to understand how effective is the maximum entropy principle to approximate the target distribution $\hat{\boldsymbol{p}}$.

To compute analytically the volume of the version space and its main properties we have assumed that observables were quenched random variables, and we have ignored correlations between different values of the observable in different variable configurations. This assumption is not realistic, as physical observables are generally rather smooth functions of the configurations. As a result of this simplifying assumption, we have been able to compute the typical distances between the target, the centre-of-mass, and the maximum entropy distributions, as well as the typical fluctuations around the centre of mass. The calculation was done within the replica symmetric framework, and we have checked that our solution was locally stable against replica-symmetry-breaking fluctuations (replicon modes). A major

outcome of our calculation is the notion of learning edge, which separates learned from as-yet-unknown configuration probabilities. Learned probabilities correspond to large target probabilities, while the probabilities of configurations associated to small target probabilities remain unknown. As the number of observables increases the learning edge moves to smaller and smaller probability values, implying that learning proceeds. The decrease of the learning edge is not generally accompanied by a decrease of the distance between the centre-of-mass and the target distribution, unless it hits the probabilities contributing most to that distance at specific critical points.

Numerical simulations were performed to confirm our asymptotic and analytical predictions, and to quantify finite-size effects. Due to the exponential growth of the dimension of the version space with the number N of variables we are limited to very small sizes, in practice $N \leq 10$. Nevertheless results are in good agreement with our analytical results and finite-size effects do not seem to alter our asymptotic results in a significant way. In particular our predictions for the existence of a learning edge and for the functional forms of the marginal measures over the probabilities of configurations are remarkably confirmed by the numerics.

We have also studied the role of the entropic bias for a given number of measured observables. Our calculation shows that the maximum entropy distribution is not closer to the target distribution than any other randomly chosen distribution in the version space. This negative result is due to our choice of fully uncorrelated observables. It is quite likely that introducing an entropic bias becomes efficient and improves learning performances if we require that both the observables and the target distribution satisfy some smoothness criteria. Some preliminary results, obtained without assuming that the observable values are uncorrelated from configuration to configuration, are reported in [10] and seem to support this guess. Unfortunately the analytical calculations with 'correlated' observables are very involved, and we have not been able to make substantial progresses so far. The theoretical importance and the practical relevance of the issues addressed here, such as how the number of constraints should depend on the smoothness of the observables and of the target distribution, or whether the phase transitions at distinct steps of the learning process found in the random uncorrelated case studied here exist in the correlated case, are strong incentives to extend this study to realistic observable ensembles. Another valuable direction for further research would be to search for 'good' sets of observables. While we have restricted here to the case of M independently drawn observables, it would be interesting to optimize the choice of those observables to make the volume of the version space as small as possible, and to get as close as possible to the target distribution. In this regard, we have extended the present analysis to the case of a finite number of replicas, $n \neq 0$, the result of which will be reported soon.

Acknowledgments The authors are grateful to U. Ferrari, Y. Kabashima, J. Lebowitz, and T. Mora for fruitful discussions. T. O. acknowledges the support by Grant-in-Aid for JSPS Fellows, as well as the JSPS Core-to-Core Program "Non-equilibrium dynamics of soft matter and information". S. C. and R. M. acknowledge financial support from the [EU-]FP7 FET OPEN project Enlightenment 284801 and Agence Nationale de la Recherche Coevstat project (ANR-13-BS04-0012-01). Numerical calculations were partly carried out on the TSUBAME2.5 supercomputer in the Tokyo Institute of Technology.

Appendix 1: Lesson From No-constraint Case for Finite Γ

If Γ is finite in the no-constraint case, the volume is written as

$$V(M=0,\Gamma) = \int_{-i\infty}^{i\infty} d\Lambda\, e^{\Lambda} \int_0^{\infty} \prod_{\boldsymbol{s}} \left(dp_{\boldsymbol{s}}\, e^{-\Lambda p_{\boldsymbol{s}} - \Gamma p_{\boldsymbol{s}} \log p_{\boldsymbol{s}}} \right). \tag{118}$$

Now, the saddle point with respect to p_s is quite simple

$$p_s^* = \tilde{p} = e^{-1-\Lambda/\Gamma}. \tag{119}$$

This is the origin of the solution assumed in Sect. 4.3.1. Assuming the saddle-point approximation is correct, we can write the integration as

$$\int dp_s \, e^{-\Lambda p_s - \Gamma p_s \log p_s} = \sqrt{2\pi}\sqrt{\frac{e^{-1-\frac{\Lambda}{\Gamma}}}{\Gamma}} e^{\Gamma e^{-1-\frac{\Lambda}{\Gamma}}} \tag{120}$$

and the log volume is

$$F = \Lambda + 2^N \left\{ \Gamma e^{-1-\frac{\Lambda}{\Gamma}} + \frac{1}{2}\left(\log(2\pi) - 1 - \frac{\Lambda}{\Gamma} - \log\Gamma \right) \right\}. \tag{121}$$

Taking a variation with respect to Λ, we get

$$\frac{1}{2^N} = e^{-1-\frac{\Lambda}{\Gamma}} + \frac{1}{2\Gamma}. \tag{122}$$

We know $\Lambda = 2^N$ at $\Gamma = 0$. Thus, it is natural to think that the saddle point $\tilde{p} = e^{-1-\Lambda/\Gamma}$ takes meaningful value only for $\Gamma \geq 2^N$, and otherwise $\tilde{p}$ rapidly becomes zero and the result comes back to the case $\Gamma = 0$. Hence, Eq. (122) is satisfied in the leading scale as $2^{-N} \doteq e^{-1-\Lambda/\Gamma}$. Hence, $\Lambda = (N \log 2 - 1)\Gamma$ and the term $1/2\Gamma$ becomes subleading one. Substituting this, we get

$$F \approx (N \log 2)\Gamma - 2^N \left(N \log 2 + \log\Gamma + \frac{1}{2}(1 - \log 2\pi) \right). \tag{123}$$

The leading scale of the log volume is thus dominated by Γ.

Appendix 2: Stability Analysis

The replica generating function is given by

$$\phi(n) = \frac{1}{2}\sum_{a\leq b} Q_{ab} Q'_{ab} - \frac{M}{2} \operatorname{Tr} \log (E + Q) + \frac{M}{2} n \log E + \sum_s \log \Theta_s, \tag{124}$$

Here Tr denotes the trace of matrix. We write the exponent in Θ_s in Eq. (39) as

$$f_s(\{p_s^a\}_a) = -\sum_a \Lambda_a (p_s^a - \hat{p}_s) - \frac{1}{2}\sum_{a\leq b} Q'_{ab}(p_s^a - \hat{p}_s)(p_s^b - \hat{p}_s) - \Gamma \sum_a p_s^a \log p_s^a. \tag{125}$$

Let us consider some small fluctuations around the RS saddle-point order parameters:

$$Q_{ab} = Q_{ab}^{\mathrm{RS}} + x_{ab}, \;\; Q'_{ab} = Q'^{\mathrm{RS}}_{ab} + 2\,\hat{x}_{ab}. \tag{126}$$

Then,

$$\frac{1}{2}\sum_{a\leq b} Q_{ab} Q'_{ab} \approx \frac{1}{2}\sum_{a\leq b} Q_{ab}^{\mathrm{RS}} Q'^{\mathrm{RS}}_{ab} + \sum_{a\leq b} x_{ab}\hat{x}_{ab}\,, \tag{127}$$

$$\operatorname{Tr} \log (E + Q) \approx \operatorname{Tr} \log \left(E + Q^{\mathrm{RS}}\right) - \frac{1}{2}\operatorname{Tr}(Ax)^2\,. \tag{128}$$

where we define $A = \left(E + Q^{\mathrm{RS}}\right)^{-1}$, and

$$f_s = f_s^{\mathrm{RS}} - \sum_{a \le b} \hat{x}_{ab} \bar{p}_s^a \bar{p}_s^b = f_s^{\mathrm{RS}} + \Delta f_s, \tag{129}$$

with $\bar{p}_s^a = p_s^a - \hat{p}_s$. Thus

$$\log \Theta_s \approx \log \mathrm{Tr}\, e^{f_s^{\mathrm{RS}}} \left(1 + \Delta f_s + \frac{1}{2}\Delta f_s^2\right) \approx \log \Theta^{\mathrm{RS}} + \frac{1}{2}\left\langle \Delta f_s^2 \right\rangle - \frac{1}{2}\left\langle \Delta f_s \right\rangle_s^2, \tag{130}$$

where we have introduced the notation

$$\langle \cdots \rangle_s = \frac{1}{\Theta_s^{\mathrm{RS}}} \mathrm{Tr}\, e^{f_s^{\mathrm{RS}}} (\cdots). \tag{131}$$

Note that we have omitted all the first-order terms, which vanish at the saddle point. Thus we get $\phi(n) = \phi^{\mathrm{RS}} + \Delta$ with

$$2\Delta = 2\sum_{a \le b} x_{ab}\hat{x}_{ab} + \frac{M}{2}\mathrm{Tr}(Ax)^2 + \sum_{a \le b}\sum_{c \le d} \hat{x}_{ab}\hat{x}_{cd} \sum_s \left(\left\langle \bar{p}_s^a \bar{p}_s^b \bar{p}_s^c \bar{p}_s^d \right\rangle_s - \left\langle \bar{p}_s^a \bar{p}_s^b \right\rangle \left\langle \bar{p}_s^c \bar{p}_s^d \right\rangle_s \right). \tag{132}$$

We write $A_{aa} \equiv X$ and $A_{ab} \equiv Y$ $(a \neq b)$, and those components are easily calculated

$$X = \frac{E + Q + (n-2)R}{(E + Q + (n-1)R)\,(E + Q - R)}, \; Y = -\frac{R}{(E + Q + (n-1)R)\,(E + Q - R)}, \tag{133}$$

$$X + (n-1)Y = \frac{1}{(E + Q + (n-1)R)}, \; X - Y = \frac{1}{(E + Q - R)}. \tag{134}$$

Equation (132) is a quadratic form with respect to $\{x_{ab}\}$ and $\{\hat{x}_{ab}\}$ and can be written as $2\Delta = \boldsymbol{V}^t G \boldsymbol{V}$, where $\boldsymbol{V}$ is a column vector with components $\{x_{ab}\}$ and $\{\hat{x}_{ab}\}$, and $\boldsymbol{V}^t$ is its transpose. We now want to determine whether the Hessian matrix G has unstable modes. The most likely candidate lies in the replicon eigenspace, which is spanned by the vectors whose components $(x_{ab}, \hat{x}_{ab})$ depend only on whether their replica indices are equal or different to two fixed values $a = \theta$ and $b = \eta$ [9]. In the replicon space, diagonal $(a = b)$ fluctuations tend to be irrelevant since other (transverse, longitudinal) modes span the n-dimensional space $(x_{aa}, \hat{x}_{aa})$. Hence we set hereafter $x_{aa} = \hat{x}_{aa} = 0$.

We arrange the components of $\boldsymbol{V}$ by ordering x as

$$\boldsymbol{v} = (x_{12}, x_{13}, \cdots, x_{1n}, x_{23}, \cdots, x_{n-1,n}), \tag{135}$$

and $\hat{x}$ as well. The Hessian G has the following form

$$G = \left(\begin{array}{ccc|ccc} S & T\cdots T & U\cdots U & & & \\ & \ddots & & & I & \\ T\cdots T & U\cdots U & S & & & \\ \hline & & & \hat{S} & \hat{T}\cdots\hat{T} & \hat{U}\cdots\hat{U} \\ & I & & & \ddots & \\ & & & \hat{T}\cdots\hat{T} & \hat{U}\cdots\hat{U} & \hat{S} \end{array}\right), \tag{136}$$

where I is the identity matrix, and

$$\hat{S} = \sum_s \left(\left\langle (\bar{p}_s^a \bar{p}_s^b)^2 \right\rangle_s - \left\langle \bar{p}_s^a \bar{p}_s^b \right\rangle_s^2 \right), \tag{137}$$

$$\hat{T} = \sum_s \left(\left\langle (\bar{p}_s^a)^2 \bar{p}_s^b \bar{p}_s^c \right\rangle_s - \left\langle \bar{p}_s^a \bar{p}_s^b \right\rangle_s \left\langle \bar{p}_s^a \bar{p}_s^c \right\rangle_s \right), \tag{138}$$

$$\hat{U} = \sum_s \left(\left\langle \bar{p}_s^a \bar{p}_s^b \bar{p}_s^c \bar{p}_s^d \right\rangle_s - \left\langle \bar{p}_s^a \bar{p}_s^b \right\rangle_s \left\langle \bar{p}_s^c \bar{p}_s^d \right\rangle_s \right), \tag{139}$$

$$S = M(X^2 + Y^2), \quad T = MY(X + Y), \quad U = 2MY^2. \tag{140}$$

Let us find the eigenvectors of this matrix. The first eigenvector $\boldsymbol{V}_1$ is obtained by assuming $x_{ab} = a$ and $\hat{x}_{ab} = \hat{a}$ for any λ, ν. The upper half (in the matrix drawn in Eq. (136)) of the eigenvalue equation $G\boldsymbol{V}_1 = \lambda_1 \boldsymbol{V}_1$ gives

$$\left(S + 2(n-2)T + \frac{1}{2}(n-2)(n-3)U \right) a + \hat{a} = \lambda_1 a, \tag{141}$$

and the lower half yields

$$a + \left(\hat{S} + 2(n-2)\hat{T} + \frac{1}{2}(n-2)(n-3)\hat{U} \right) \hat{a} = \lambda_1 \hat{a}. \tag{142}$$

These equations have the eigenvalue λ_1 with non-vanishing $a, \hat{a}$, which must satisfies the following relation

$$\lambda_1^2 - (C_1 + \hat{C}_1)\lambda_1 + (C_1 \hat{C}_1 - 1) = 0, \tag{143}$$

where $C_1 = S + 2(n-2)T + (n-2)(n-3)U/2$ and $\hat{C}_1 = \hat{S} + 2(n-2)\hat{T} + (n-2)(n-3)\hat{U}/2$. This equation says that this mode spans a two-dimensional space.

The next type of solution $\boldsymbol{V}_2$ is obtained by selecting a replica index, say, $\theta = 1$. This solution $\boldsymbol{V}_2$ has $x_{ab} = b$ and $\hat{x}_{ab} = \hat{b}$ when λ or ν is equal to 1, $y^{\lambda\nu} = c$ and $\hat{y}^{\lambda\nu} = \hat{c}$ otherwise. The first row of the eigenvalue equation $G\boldsymbol{V}_2 = \lambda_2 \boldsymbol{V}_2$ gives

$$(S + (n-2)T)b + \left((n-2)T + \frac{1}{2}(n-2)(n-3)U \right) c + \hat{b} = \lambda_2 b, \tag{144}$$

and the first row of the lower half of matrix G in Eq. (136) yields

$$b + \left(\hat{S} + (n-2)\hat{T} \right) \hat{b} + \left((n-2)\hat{T} + \frac{1}{2}(n-2)(n-3)\hat{U} \right) \hat{c} = \lambda_2 \hat{b}. \tag{145}$$

We now impose the orthogonality condition over $\boldsymbol{V}_2$ and $\boldsymbol{V}_1$. This leads to

$$(n-1)b + \frac{1}{2}(n-1)(n-2)c = 0, \tag{146}$$

and the same relation holds for $\hat{b}$ and $\hat{c}$. Substituting these conditions, we get

$$(S + (n-4)T - (n-3)U)\, b + \hat{b} = \lambda_2 b, \quad \left(\hat{S} + (n-4)\hat{T} - (n-3)\hat{U} \right) \hat{b} + b = \lambda_2 \hat{b}, \tag{147}$$

which leads to

$$\lambda_2^2 - (C_2 + \hat{C}_2)\lambda_2 + (C_2 \hat{C}_2 - 1) = 0, \tag{148}$$

where $C_2 = S + (n-4)T - (n-3)U$ and $\hat{C}_2 = \hat{S} + (n-4)\hat{T} - (n-3)\hat{U}$. As there are n possible choices for the replica index θ and two eigenvalues/eigenvectors $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ for each choice, this particular subspace is of dimension $2n$.

The third mode $\boldsymbol{V}_3$ is obtained by treating two replicas θ, ω as special ones. This solution $\boldsymbol{V}_3$ has $x_{\theta\omega} = d$ and $\hat{x}_{\theta\omega} = \hat{d}$, $x_{\theta a} = e$ and $\hat{x}_{\omega a} = \hat{e}$, and $x_{ab} = f$ and $\hat{x}_{ab} = \hat{f}$ otherwise. The orthogonality condition with $\boldsymbol{V}_2$ is

$$\begin{aligned} &b(d + (n-2)e) + c\left((n-2)e + \frac{1}{2}(n-2)(n-5)f\right) \\ &= b\left(d + (n-4)e - (n-5)f\right) = 0, \end{aligned} \tag{149}$$

where we use Eq. (146). The one with $\boldsymbol{V}_1$ is

$$d + 2(n-2)e + \frac{1}{2}(n-2)(n-5)f = 0. \tag{150}$$

These relations mean

$$e = -\frac{d}{n-2}, \quad \frac{1}{2}(n-2)(n-5)f = -d - 2(n-2)e = d. \tag{151}$$

Similar relations hold for the hat variables. According to these relations, the first rows of the upper and lower halves of the eigenvalue equation $G\boldsymbol{V}_3 = \lambda_3 \boldsymbol{V}_3$ yield

$$dS + 2(n-2)eT + \frac{1}{2}(n-2)(n-5)fU + \hat{d} = (S - 2T + U)d + \hat{d} = \lambda_3 d, \tag{152}$$

$$d + (\hat{S} - 2\hat{T} + \hat{U})\hat{d} = \lambda_3 \hat{d}, \tag{153}$$

Thus we obtain

$$\lambda_3^2 - (C_3 + \hat{C}_3)\lambda_3 + (C_3\hat{C}_3 - 1) = 0, \tag{154}$$

where $C_3 = S - 2T + U$ and $\hat{C}_3 = \hat{S} - 2\hat{T} + \hat{U}$. This solution spans a $n(n-3)$-dimensional space, implying that no eigenmode are left.

For the stability of the saddle point, all of the eigenvalues must be non-negative. This condition corresponds to

$$\forall i, \ C_i \ \hat{C}_i \leq 1. \tag{155}$$

This condition takes into account the fact that Q'_{ab} is originally a pure imaginary variable, which means that $\delta Q_{ab} \delta Q'_{ab}$ is associated with a multiplicative factor i and $\delta Q'_{ab} \delta Q'_{ab}$ acquires a factor -1. Hence, if we change variable from Q'_{ab} to $i\, Q'_{ab}$, the diagonal block in the lower half part of G gets a factor -1, and the off-diagonal part becomes $i\, I$, which leads to the positivity condition in Eq. (155).

The replicon mode corresponds to λ_3 and $\boldsymbol{V}_3$. Thus, the stability condition is

$$(S - 2T + U)(\hat{S} - 2\hat{T} + \hat{U}) \leq 1. \tag{156}$$

A straightforward calculation shows that

$$S - 2T + U = M(X - Y)^2 = \frac{M}{(E + Q - R)^2}. \tag{157}$$

We now turn to the calculation of $\hat{S} - 2\hat{T} + \hat{U}$. Within the RS assumption the average $\langle O \rangle_s$, where $O = O_a O_b \cdots O_c$, can be expressed as

$$\langle O \rangle_s = \frac{1}{\int Dz X_s^n} \int Dz X_s^n \langle O_a \rangle_{X_s} \cdots \langle O_c \rangle_{X_s} \ , \tag{158}$$

where X_s was defined in Eq. (47). Thus,

$$\hat{S} - 2\hat{T} + \hat{U} = \sum_s \left\{ \left\langle (\bar{p}_s^a)^2 (\bar{p}_s^b)^2 \right\rangle_s - 2 \left\langle (\bar{p}_s^a)^2 \bar{p}_s^b \bar{p}_s^c \right\rangle_s - \left\langle \bar{p}_s^a \bar{p}_s^b \bar{p}_s^c \bar{p}_s^d \right\rangle_s \right\}$$
$$= \sum_s \frac{1}{\int Dz X_s^n} \int Dz X_s^n \left(\langle \bar{p}_s^2 \rangle_{X_s} - \langle \bar{p}_s \rangle_{X_s}^2 \right)^2 . \tag{159}$$

Noticing that

$$\langle \bar{p}_s^2 \rangle_{X_s} - \langle \bar{p}_s \rangle_{X_s}^2 = \frac{\partial^2 \log X_s}{\partial \Lambda^2} , \tag{160}$$

the stability condition may be written, in the $n \to 0$ limit, under the form shown in Eq. (111).

References

1. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**, 620–630 (1957)
2. Jaynes, E.T.: Information theory and statistical mechanics II. Phys. Rev. **108**, 171–190 (1957)
3. Balian, R.: Statistical mechanics and the maximum entropy method. In: Grassberger, P., Nadal, J.P. (eds.) From Statistical Physics to Statistical Inference and Back. NATO ASI series 428, pp. 11–43. Kluwer Academics Publisher, New York (1994)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley series in telecommunications and signal processing. Wiley, New York (1991)
5. Jaynes, E.T.: On the rationale of maximum entropy methods. Proc. IEEE **70**, 939–952 (1982)
6. Jaynes, E.T.: Monkeys, kangaroos and *N*. In: Justice, J.H. (ed.) Maximum-Entropy and Bayesian Methods in Applied Statistics, pp. 26–58. Cambridge University Press, Cambridge (1986)
7. Tikochinsky, Y., Tishby, N.Z., Levine, R.D.: Alternative approach to maximum-entropy inference. Phys. Rev. A **30**, 2638–2644 (1984)
8. Bialek, W.: Biophysics: Searching for Principles. Princeton University Press, Princeton (2012)
9. De Almeida, J.R.L., Thouless, D.J.: Stability of the Sherrington-Kirkpatrick solution of a spin glass model. J. Phys. A **11**, 983–990 (1978)
10. Obuchi, T., Monasson, R.: J. Phys. Conf. Ser (2015) (Submitted to)