Journal of Physics: Conference Series **638** (2015) 012018 doi:10.1088/1742-6596/638/1/012018

Learning probability distributions from smooth observables and the maximum entropy principle: some remarks

Tomoyuki Obuchi[1,†], **Rémi Monasson**[2]

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

[2]Laboratoire de Physique Théorique de l'Ecole Normale Supérieure, associé au CNRS et Í'Université Pierre et Marie Curie, 75005 Paris, France

E-mail: [†]obuchi@sp.dis.titech.ac.jp

Abstract. The maximum entropy principle (MEP) is a very useful working hypothesis in a wide variety of inference problems, ranging from biological to engineering tasks. To better understand the reasons of the success of MEP, we propose a statistical-mechanical formulation to treat the space of probability distributions constrained by the measures of (experimental) observables. In this paper we first review the results of a detailed analysis of the simplest case of randomly chosen observables. In addition, we investigate by numerical and analytical means the case of smooth observables, which is of practical relevance. Our preliminary results are presented and discussed with respect to the efficiency of the MEP.

1. Introduction

The maximum entropy (ME) principle, which was proposed by Jaynes in 1957 to rederive statistical mechanics from an information-theoretic viewpoint [1], recently stands out as a useful working hypothesis in a wide variety of inference problems [2, 3, 4, 5, 6, 7, 8, 9, 10]. However, a better comprehension of why and when ME applies is still needed. In this respect, we have recently introduced a statistical-mechanical formulation, in which we treated the space of all probability distributions satisfying a set of moment-matching conditions [11]. Moment-matching conditions were defined through the measured observables, and provided information about the target distribution of interest. It was assumed in that work that the observables were identically and independently distributed random variables, which enabled us to perform analytical investigations and to get a precise quantitative characterization of the volume of the space of probability distributions and of the distance between certain probability distributions and the target one as the number of moment-matching conditions varies. An interesting, but disappointed outcome of the analysis was that the ME distribution was not 'closer' to the target distribution than most of the other distributions satisfying the conditions. It was however clear that the failure of the ME principle was due to the irrealistic nature of the observable statistics. In practical applications of the ME, indeed, the observables are not chosen randomly, but are designed to represent some underlying features of the target distribution.

In the present paper, we try to relax the above randomization hypothesis. This strategy leads us to introduce a bias in choosing observables. In particular, we consider the case of 'smooth'

observables, that is, taking values that cannot differ much between two nearby microscopic configurations of the system. Though we have not been able to extend the detailed analysis of [11] to this (much) more involved case, we report here some analytical and numerical results we have recently obtained.

The organization of this paper is as follows. In the Sec. 2, we present our statistical-physics framework, and discuss different distributions of observables (random and smooth); we also give an overview of our previous results on the random case. In Sec. 3, we report numerical calculations based on a Monte Carlo method in several different situations. In Sec. 4 we report some attempts at solving the problem analytically. Last of all, we discuss our findings are their possible implications for the ME. in the last Section.

2. Formalism and overview of results in the random observable case

2.1. Basic formulation

Let us consider a system described by configurations of N binary spins, $\boldsymbol{s} = \{s_i = \pm 1\}_{i=1}^{N}$, which we call target system. The probability distribution of the target system, the target distribution, is denoted by $\hat{p}_{\boldsymbol{s}}$. Trial distributions, which we will fit based on the measurements of observables (moments). will be generally written as $p_{\boldsymbol{s}}$.

We assume that a measurement gives a value of a linear combination of some moments of the target distribution. Thus we can write a measurement of an observable $\{v_{\boldsymbol{s}}\}_{\boldsymbol{s}}$ as

$$\sum_{\boldsymbol{s}} v_{\boldsymbol{s}} \hat{p}_{\boldsymbol{s}} = \boldsymbol{v} \cdot \hat{\boldsymbol{p}}. \tag{1}$$

where the summation runs over all the spin configuration and we introduce a vector notation of the 2^N dimension corresponding to all the spin configurations. The moment-matching (MM) conditions for M observables $\{\boldsymbol{v}^\mu\}_{\mu=1}^{M}$ are then written as

$$\boldsymbol{v}^\mu \cdot \boldsymbol{p} = \boldsymbol{v}^\mu \cdot \hat{\boldsymbol{p}}, \ \forall \mu. \tag{2}$$

Those equations impose that the trial distribution $\boldsymbol{p}$ reproduces the measurements of the average values observable $\boldsymbol{v}^\mu$ over the target distribution.

We are interested in the phase space of trial distributions constrained by the MM conditions (2). To investigate this, we introduce the following probability distribution on trial distributions

$$\rho\left(\boldsymbol{p}|\Gamma, E, \{\boldsymbol{v}^\mu\}_{\mu=1}^{M}, \hat{\boldsymbol{p}}\right) = \frac{1}{V} \prod_{\boldsymbol{s}} \theta(p_{\boldsymbol{s}}) \delta\left(\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} - 1\right) \exp\left\{ -\frac{1}{2E} \sum_{\mu=1}^{M} \left(\boldsymbol{v}^\mu \cdot (\boldsymbol{p} - \hat{\boldsymbol{p}})\right)^2 + \Gamma S(\boldsymbol{p}) \right\}, \tag{3}$$

where $S(\boldsymbol{p})$ is the Shannon entropy whose mathematical expression is

$$S(\boldsymbol{p}) = -\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} \log p_{\boldsymbol{s}}. \tag{4}$$

The denominator V is the normalization factor, or the partition function,

$$V\left(\Gamma, E, \{\boldsymbol{v}^\mu\}_{\mu=1}^{M}, \hat{\boldsymbol{p}}\right) = \int_0^\infty \prod_{\boldsymbol{s}} dp_{\boldsymbol{s}} \delta\left(\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} - 1\right) \exp\left\{ -\frac{1}{2E} \sum_{\mu=1}^{M} \left(\boldsymbol{v}^\mu \cdot (\boldsymbol{p} - \hat{\boldsymbol{p}})\right)^2 + \Gamma S(\boldsymbol{p}) \right\}. \tag{5}$$

The factor $\exp\left(-\frac{1}{2E} \sum_{\mu=1}^{M} \left(\boldsymbol{v}^\mu \cdot (\boldsymbol{p} - \hat{\boldsymbol{p}})\right)^2\right)$ corresponds to the MM conditions and E is the tolerance against the error on the MM conditions. If $\boldsymbol{p}$ satisfies the MM conditions, the

corresponding probability value $\rho(\boldsymbol{p})$ is large, and small otherwise. Especially in the $E \to 0$ limit, trial distributions exactly satisfying all the MM conditions are the only ones to 'survive' and to contribute to V. We call those distributions 'good' distributions.

All good distributions have an equal weight value of ρ at $\Gamma = 0$, but for large $\Gamma \gg 0$ good distributions with larger entropies are assigned larger weights. Hence Γ is a bias towards distributions with large entropies. Studying the two extreme cases $\Gamma = 0$ and $\Gamma \to \infty$, for which only the ME distribution survives, we can compare typical good distributions and the ME distribution. Actually, we observe some drastic changes for specific intermediate values of Γ, which correspond to phase transitions; hence, the whole range of $\Gamma > 0$ is of interest.

2.2. Distribution of observables

We now need to specify how the observables $\{\boldsymbol{v}^\mu\}_\mu$ are chosen. In [11], we assumed the observables were identically independently distributed (i.i.d.) Gaussian random variable

$$P_{\rm rand}(\boldsymbol{v}) = \prod_{\boldsymbol{s}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v_{\boldsymbol{s}}^2}. \tag{6}$$

The analysis of this random case, which is briefly reported in Section sec. 2.3, yields some important notions and order parameters. However, it is unrealistic as the components of observables are uncorrelated with each other. In particular, two configurations $\boldsymbol{s}$ and $\boldsymbol{s}'$ differing by a single spin, say, s_i, have totally uncorrelated observable values, $v_{\boldsymbol{s}}$ and $v_{\boldsymbol{s}'}$. This is clearly not true for most systems: changing one among N spins generically does not affect dramatically the system properties.

An inspiring idea comes from the Fourier transform in the spin-configuration space. Let us consider an arbitrary function of spin configurations $f_{\boldsymbol{s}}$ and the Fourier transform characterized by a wavenumber vector of binary components $\boldsymbol{k} = \{k_i = 0, 1\}_{i=1}^N$. The explicit expression of the Fourier transform is

$$f_{\boldsymbol{k}} = \sum_{\boldsymbol{s}} \left(\frac{1}{\sqrt{2^N}} \prod_{i=1}^N s_i^{k_i} \right) f_{\boldsymbol{s}} = \boldsymbol{w}^{\boldsymbol{k}} \cdot \boldsymbol{f}, \tag{7}$$

where we defined the $\boldsymbol{s}$-component of $\boldsymbol{w}^{\boldsymbol{k}}$ to be equal to $(\prod_{i=1}^N s_i^{k_i})/\sqrt{2^N}$. The modes $\{\boldsymbol{w}^{\boldsymbol{k}}\}_{\boldsymbol{k}}$ constitute a complete orthogonal set, and thus any vector in the phase space can be expanded on this set. Each mode has a physical meaning: a mode with a wavenumber $k = |\boldsymbol{k}| = \sum_i k_i$ corresponds to a k-spin multiplet. In particular, if $f_{\boldsymbol{s}}$ is a probability distribution, $f_{\boldsymbol{k}}$ is the average value of the k-spin multiplet (divided by $1/\sqrt{2^N}$).

If a random vector drawn from eq. (6) is expanded in the Fourier space, the corresponding weights are statistically uniform on all the modes, which implies that the corresponding MM condition includes correlations of very high-order multiplets of spins with relevant weights. This is not natural because low-order multiplets are often considered to be more important than high-order ones in a general situation. Actually in most of recent applications ranging from biological to engineering problems [2, 3, 4, 5, 6, 7, 8, 9, 10], the ME models are constructed based on observations of singlets and doublets, $|\boldsymbol{k}| = 1$ and 2, of spins. Hence, a naive extension from the random case to a more realistic situation is to bias the weights on some (low-$|\boldsymbol{k}|$'s) Fourier modes in generating observables.

A simple choice along the above idea is using the following distribution

$$P(\boldsymbol{v}) = C \exp\left\{ -\frac{1}{2} \sum_{\boldsymbol{s},\boldsymbol{t}} M_{\boldsymbol{s}\boldsymbol{t}} v_{\boldsymbol{s}} v_{\boldsymbol{t}} \right\}, \tag{8}$$

instead of eq. (6). In the distribution above, the matrix $M_{\boldsymbol{st}}$ is chosen such to make sure that components of the observable $\mathbf{v}$ corresponding to nearby configurations have similar values. In other words, the presence of M should make the observable *smooth* over the configuration space. Alternatively, and perhaps more simply, we may require that the Fourier expression of eq. (8) becomes diagonal and the variance only depends on the absolute value of $k = |\boldsymbol{k}|$

$$P(\boldsymbol{v}) = C \exp\left\{ -\frac{1}{2} \sum_{\boldsymbol{k}} \frac{v_{\boldsymbol{k}}^2}{V_k} \right\}. \tag{9}$$

Expression eq. (9) implies that the matrix entry $M_{\boldsymbol{st}}$ depends on $\boldsymbol{s}$ and $\boldsymbol{t}$ through the overlap $\sum_i s_i t_i$ only. If we choose V_k to be a rapidly-decreasing function of the modulus k, modes with large k will be suppressed. There are many different ways to introduce a k-dependence in V_k. We try several different expressions for V_k in the numerical studies reported in Section sec. 3.

2.3. Overview of results for random observables

We now briefly recall the results of [11] corresponding to the random case defined in eq. (6). The average value of logarithm of V in the large-system-size limit can be computed with the replica method (within the replica symmetric (RS) ansatz),

$$F = [\log V]_{\text{rand}} = \lim_{n \to 0} \frac{1}{n} \log [V^n]_{\text{rand}}\,, \tag{10}$$

where the square brackets $[\cdots]_{\text{rand}}$ denote the average over the observables (6). Our analysis showed that F depends on the target distribution only through the entropy curve

$$\sigma(\omega) = \frac{1}{N} \sum_{\boldsymbol{s}} \delta\left(\omega - \omega_{\boldsymbol{s}}\right), \tag{11}$$

where we assume the target distribution obeys an exponential scale for $N \gg 1$, *i.e.* that

$$\omega_{\boldsymbol{s}} \equiv -\frac{1}{N} \log \hat{p}_{\boldsymbol{s}}, \tag{12}$$

has a well-defined limit for $N \to \infty$. This assumption is true for common physical systems. We give a typical shape of $\sigma(\omega)$ as the left panel of Fig. 1, which corresponds to the independent spin model (ISM) with $H = 0.5$ as the target distribution

$$\hat{p}_{\boldsymbol{s}}^{\text{ISM}} = \frac{e^{H \sum_i s_i}}{(2 \cosh H)^N}. \tag{13}$$

The characteristic values of ω, ω_k with $k = 0, 1, 2$, are defined by

$$\left. \frac{d\sigma(\omega)}{d\omega} \right|_{\omega = \omega_k} = k, \tag{14}$$

and they are connected to certain phase transitions (see [11] for details of the transitions). These characteristic values of ω separate the entropy curve in three regions (the rightmost region is not relevant), each of which is called I, II, and III as shown in Fig. 1.

The analysis in the random case revealed that the following three order parameters are useful

$$Q(\{\boldsymbol{v}^\mu\}_\mu, \hat{\boldsymbol{p}}) \equiv \sum_{\boldsymbol{s}} \left\langle \left(p_{\boldsymbol{s}} - \hat{p}_{\boldsymbol{s}}\right)^2 \right\rangle, \tag{15}$$

$$R(\{\boldsymbol{v}^\mu\}_\mu, \hat{\boldsymbol{p}}) \equiv \sum_{\boldsymbol{s}} \left(\langle p_{\boldsymbol{s}} \rangle - \hat{p}_{\boldsymbol{s}}\right)^2, \tag{16}$$

$$D(\{\boldsymbol{v}^\mu\}_\mu, \hat{\boldsymbol{p}}) \equiv Q - R = \sum_{\boldsymbol{s}} \left(\left\langle p_{\boldsymbol{s}}^2 \right\rangle - \langle p_{\boldsymbol{s}} \rangle^2 \right), \tag{17}$$

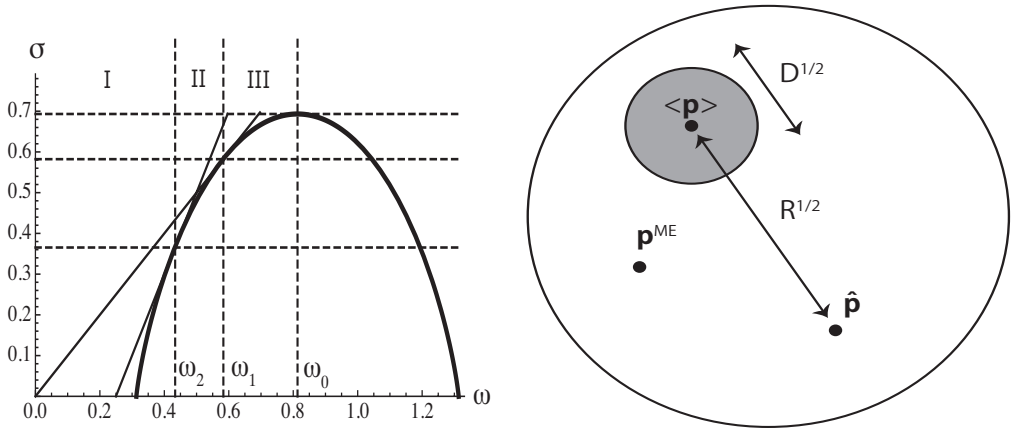

Figure 1. (Left) The entropy curve of the ISM with $H = 0.5$. There are three characteristic values of ω_k with $k = 0, 1, 2$, each of which is defined as a tangential point of the slope k (tangential lines are denoted by solid straight lines). (Right) A schematic diagram of the phase space. The large circle represents the set of all good probability vectors. The shaded area represents the typical fluctuating region of $\boldsymbol{p}$. The target distribution exists apart from $\langle \boldsymbol{p} \rangle$ by $\sqrt{R}$. The ME distribution $\boldsymbol{p}^{\rm ME}$ also exists somewhere inside the large circle.

where the angular brackets $\langle \cdots \rangle$ denote the average over $\rho(\boldsymbol{p})$. The physical significance of each order parameter is clear: Q measures the averaged square distance between $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$. R also quantifies the distance between the target distribution and the averaged $\boldsymbol{p}$ and is similar to Q, but R's meaning is clearer since it does not include the fluctuation of $\boldsymbol{p}$. On the other hand, D purely measures the fluctuation of $\boldsymbol{p}$ around the averaged value. We depict the phase space of good probability vectors with the order parameters in the right panel of Fig. 1.

The control and order parameters scale exponentially with N in the large N limit

$$M = e^{N\alpha},\ \Gamma = e^{N\gamma},\ Q = e^{Nq},\ R = e^{Nr},\ D = e^{Nd}. \tag{18}$$

Here we plot q, r and d of the ISM against α and γ in Fig. 2. Clearly $q = r$ holds, which is

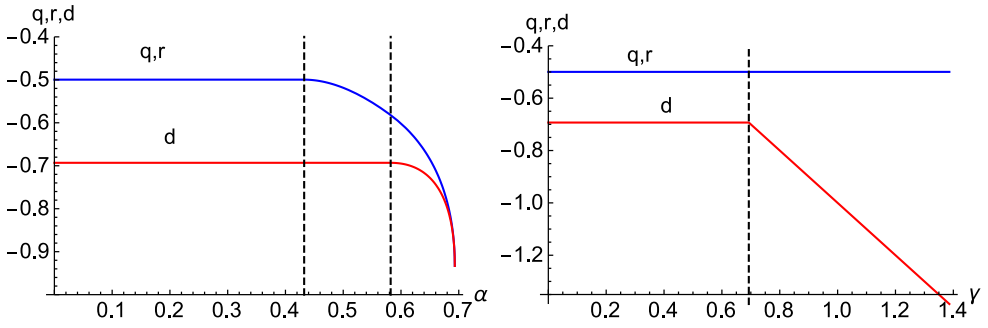

Figure 2. Plots of q, r and d against α at $\Gamma = 0$ (left) and against γ at a small value of α (right). Dotted vertical lines denote the transition points.

always true in the random case. Our crucial observation about the ME principle is that q (or r) does not depend on γ, that is, on the strength of the entropy bias at all as seen in Fig. 2, which is the case for all values of α. This means that the distances from the target distribution

to the ME one and to a typical (randomly chosen) 'good' distribution are equal. Thus the ME does not help to infer the target distribution. This is the consequence of the random nature of the observables.

An interesting finding in the analysis of the random case was the existence of the so-called 'learning edge'. The learning edge ω_E is the scale separating accurately inferred values of $\omega_{\boldsymbol{s}}$ (see eq. (12)) from the ones which cannot be inferred: for target probability values with $\omega_{\boldsymbol{s}} \leq \omega_E$ we can infer those values, and for $\omega_{\boldsymbol{s}} > \omega_E$ we cannot. The learning edge is a strictly monotonically increasing function of α (see Fig. 3 as an example for the ISM), whose behaviour reveals the existence of phase transitions. When the learning edge takes values of ω_k with $k = 0, 1, 2$, the

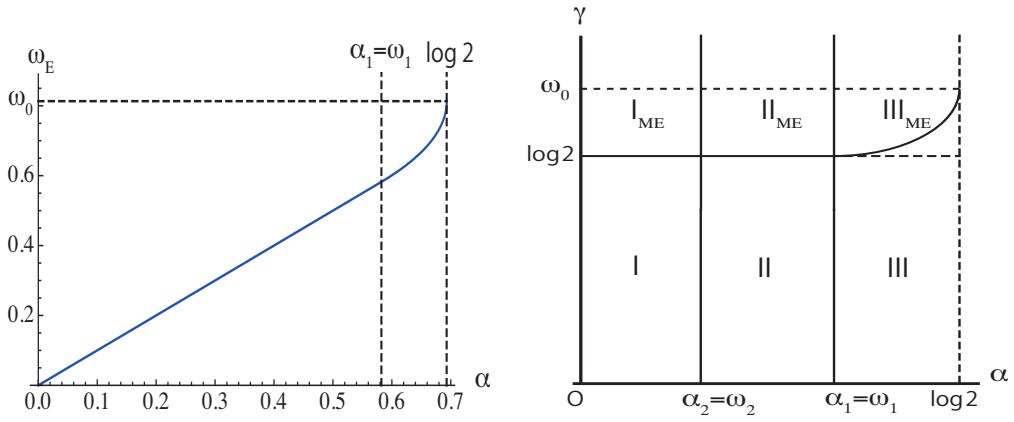

Figure 3. (Left) Plot of the learning edge ω_E against α at $\Gamma = 0$ for the ISM with $H = 0.5$. The dotted lines represent the transition points of the learning edge. At $\alpha = \log 2$, all the values of target probabilities are learned, implying the learning edge is not well defined for $\alpha > \log 2$. (Right) A shematic phase diagram of the random case. Each phase is named after the location of the learning edge.

phase transitions occur. We call each phase by the location of the learning edge: if $\omega_E < \omega_2$ then the phase is I, and phases II and III are defend as well (see the left panel of Fig. 1). The resultant phase diagram is given in the right panel of Fig. 3. In the phases subscripted by ME, the order parameter d is dominated by the ME bias Γ in contrast to the phases without the subscript.

An important finding of the calculation in the random case is that the marginal density of a small number (compared to 2^N) of probabilities of configurations factorizes a product of single-configuration probability densities $\rho(\boldsymbol{p}) \approx \prod_{\boldsymbol{s}} \rho_{\boldsymbol{s}}(p_{\boldsymbol{s}})$. Unfortunately, this nice property, at the origin of the existence of the learning edge, does not hold for the case of smooth observable, which makes the analysis much more complicated.

3. Numerical studies

3.1. Algorithm for sampling the probability distribution space

We briefly summarize our Monte Carlo (MC) algorithm. We randomly change the probability vector from $\boldsymbol{p}$ to $\boldsymbol{p}' = \boldsymbol{p} + \Delta\boldsymbol{p}$ in the phase space to perform sampling of the space of all distributions. The change of the vector, $\Delta\boldsymbol{p}$, must satisfy the following conditions

Orthogonality $(E = 0)$**:** $\Delta\boldsymbol{p} \cdot \boldsymbol{v}^{\mu} = 0, \forall\mu$.

Normalization: $\sum_{\boldsymbol{s}} \Delta p_{\boldsymbol{s}} = \Delta\boldsymbol{p} \cdot \mathbf{1} = 0$ where $\mathbf{1} = (1, 1, \cdots, 1)^t$ and the symbol ^{t} denotes the transpose.

Positivity: $p_{\boldsymbol{s}} + \Delta p_{\boldsymbol{s}} \geq 0, \forall \boldsymbol{s}$.

Combining the orthogonality with the initial condition $\boldsymbol{p}(0) = \hat{\boldsymbol{p}}$, we see that the vector chosen in this algorithm always satisfies the constraints $\boldsymbol{v}^\mu \cdot \boldsymbol{p} = \boldsymbol{v}^\mu \cdot \hat{\boldsymbol{p}}, \forall \mu$, which is needed since we are basically interested in $E = 0$. These orthogonality and normalization conditions restrict the possible move directions. On the other hand, the positivity condition is maintained by appropriately choosing the move width x. With a chosen move direction $\boldsymbol{w}$, the minimum and maximum values of x are written as

$$x_{\max} = \max_{\boldsymbol{s}} \min \left(\frac{-p_{\boldsymbol{s}}}{\omega_{\boldsymbol{s}}}, \frac{1 - p_{\boldsymbol{s}}}{w_{\boldsymbol{s}}} \right), \; x_{\min} = \min_{\boldsymbol{s}} \max \left(\frac{-p_{\boldsymbol{s}}}{\omega_{\boldsymbol{s}}}, \frac{1 - p_{\boldsymbol{s}}}{w_{\boldsymbol{s}}} \right). \tag{19}$$

The intermediate values between these two bounds are randomly chosen with equal weight, and thus the trial move becomes $\Delta \boldsymbol{p} = x\boldsymbol{w}$. Finally, we calculate the entropy difference $\Delta S = S(\boldsymbol{p}') - S(\boldsymbol{p})$ and determines whether the move $\boldsymbol{p} \to \boldsymbol{p}'$ is accepted or not by the probability

$$p_{\text{accept}} = \min \left(1, e^{\Gamma \Delta S}\right). \tag{20}$$

3.1.1. Observables drawn from the Fourier basis In the analytical treatment, we put $\boldsymbol{v}$ as a random vector drawn from eq. (8), each component of which can take a continuous value. However for numerical implementation, it is more convenient to choose observables from a set of vectors constituting a complete orthogonal set, because the above orthogonal and normalization conditions can be easily satisfied in that way.

We choose the Fourier basis as such a set. The random case corresponds to the situation where all the modes (except for $\boldsymbol{w}^{\boldsymbol{0}}$) are drawn uniformly at random. In the studies reported below, we bias the weights of the Fourier modes based on the values of corresponding wavenumber $k = |\boldsymbol{k}|$, see eq. (9).

3.1.2. Lessons from the random case: equilibration and finite-size effects We define one MC step by one trial of changing the probability vector $\boldsymbol{p}$. As reported in [11], the total MC steps for sufficient equilibration and sampling grow fast as the system size increases. For example for $N = 10$, we need $O(10^8)$ MC steps typically, which requires a couple of days. Due to this expensiveness, we mainly run our simulations for sizes less than or equal to $N = 10$.

This system size might sound small, but our numerical simulation on the random case showed that even for $N = 10$ numerical result could be quantitatively compared to the analytical one. As a demonstration, we display the order parameters of the random case calculated by the MC simulation for finite N in Fig. 4, which exhibits a clear indication of the phase transitions even for size $N = 10$. This result in the random case provides a guide for simulating more general cases. Our simulations given below are based on the equilibration criterion invented in the random case, and we expect that the treated sizes $N = 8$ or 10 capture the behavior in the large-size-system limit.

3.2. Exponentially-decaying weight and the smoothness parameter

Here we assume that each mode of the Fourier basis is chosen with a probability weight $e^{-b|\boldsymbol{k}|}$, and we call b smoothness parameter. The pseudo code to construct the set of observables is as follows:

(i) Prepare a set of Fourier modes consisting of all the modes except for $|\boldsymbol{k}| = 0$. We call this set "pool". Similarly one prepares an empty set, called "observables".

(ii) Calculate the normalization constant as $Z = \sum_{\boldsymbol{k} \in \text{pool}} e^{-b|\boldsymbol{k}|}$.

(iii) Choose one mode of the wavenumber vector $\boldsymbol{k}$ from the pool with probability $p_{\boldsymbol{k}} = e^{-b|\boldsymbol{k}|}/Z$ and put the chosen mode into the observables set. Remove the mode from the pool.

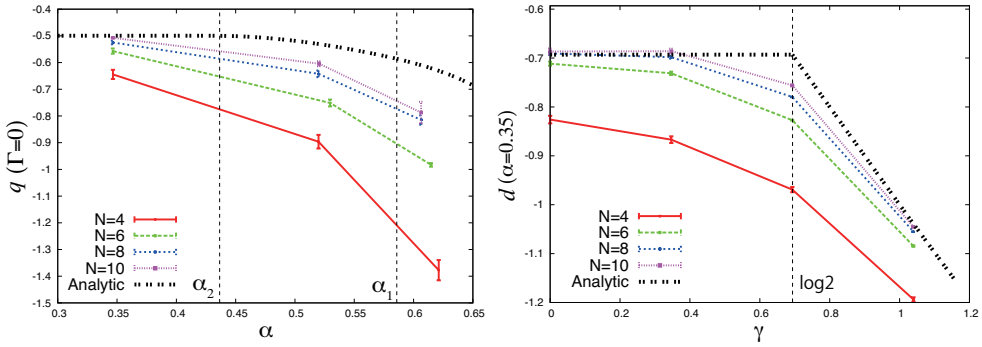

Figure 4. Numerically-evaluated order parameters q (left) and d (right) of the ISM with $H = 0.5$ plotted against α (left) and γ (right). Analytical predictions are given by black curves, which are well consistent with the numerical data. Clear indication of the phase transition from small to large Γ is seen in the right panel, and thus finite-size effects seem to be not so large.

(iv) Continue **(ii)-(iii)** until the size of the observables set becomes equal to M. At the end, what is left in the pool is our set of of possible MC-move directions.

We have to remove $\boldsymbol{w}^{\boldsymbol{0}} = \mathbf{1}/\sqrt{2^N}$ from the beginning since it is connected to the normalization condition and can neither be a nontrivial observable nor a MC-move direction. The exponentially-decaying weight in **(iii)** corresponds to putting $V_{\boldsymbol{k}}^{-1} = 2b|\boldsymbol{k}|$ in eq. (9).

As for the target model, we adopt the Boltzmann distribution $\hat{p}_{\boldsymbol{s}} = e^{-\beta\mathcal{H}}/Z$ generated from the p-spin Hamiltonian

$$\mathcal{H} = -\sum_{i_1<i_2<\cdots<i_p} J_{i_1 i_2 \cdots i_p} s_{i_1} s_{i_2} \cdots s_{i_p}, \tag{21}$$

where we assume $J_{i_1 i_2 \cdots i_p}$ is an i.i.d. random variable drawn from $\mathcal{N}\left(0, \frac{p!}{N^{p-1}}\right)$. Hereafter we call this model p-spin model (pSM). Parameter p measures the "smoothness" of the target distribution. In the large-p limit (after the large-N limit), this model converges to the random energy model (REM)[1]. The REM is a completely non-smooth model and we expect the smoothness of observables will not help inference of this model at all. On the other hand, if p is enough small, we expect that the ME distribution at large values of b preferentially has lower-spin multiplets in the effective Hamiltonian and matches to the target one, leading to an excellent performance in inference. Hence it is meaningful to compare different-p results for examining the effect of smoothness in the observables.

We below see the MC results with changing the parameters N, M, b, p and Γ. The quality of inference is quantified by the order parameter q as introduced in the random case. Error bar is estimated from the standard deviation σ_{sample} of the objective quantity among different samples as

$$\text{Error bar} = \frac{\sigma_{\text{sample}}}{\sqrt{N_{\text{sample}} - 1}}. \tag{22}$$

where N_{sample} is the number of simulated samples and are chosen to be 10 in most of the below results.

[1] The REM is defined by the Boltzmann distribution whose energies $\{E_{\boldsymbol{s}}\}_{\boldsymbol{s}}$ are drawn from an i.i.d. normal distribution $\mathcal{N}(0, 2/N)$ (see reprints in [12] for details).

3.2.1. Result in the smooth case We plot q of the 2SM in Fig. 5. We see that the larger b tends to give smaller q meaning better inference. Especially for $M = 64$ and 128 with enough large b, we find that q becomes smaller and smaller as Γ grows, implying that the ME distribution given in $\Gamma \to \infty$ realizes the perfect learning $q \to \infty$. Similar observation is found for the 3SM in Fig.

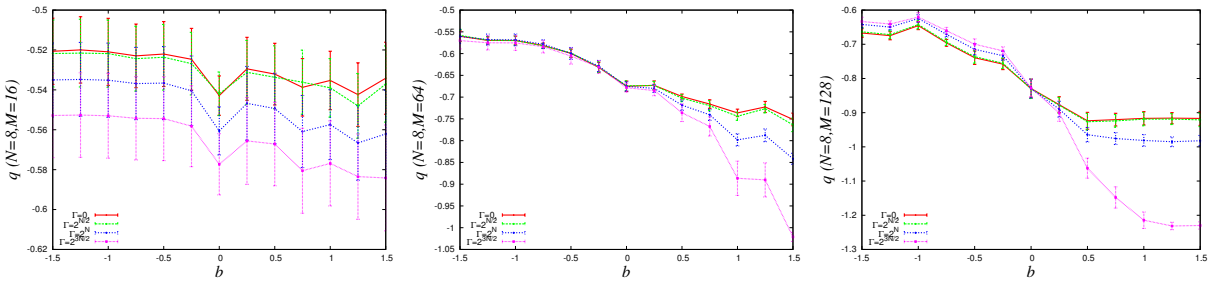

Figure 5. Plots of q for the $N = 8$ 2SM against b, for $M = 16$, 64, and 128 from left to right. Different colors correspond to different values of Γ.

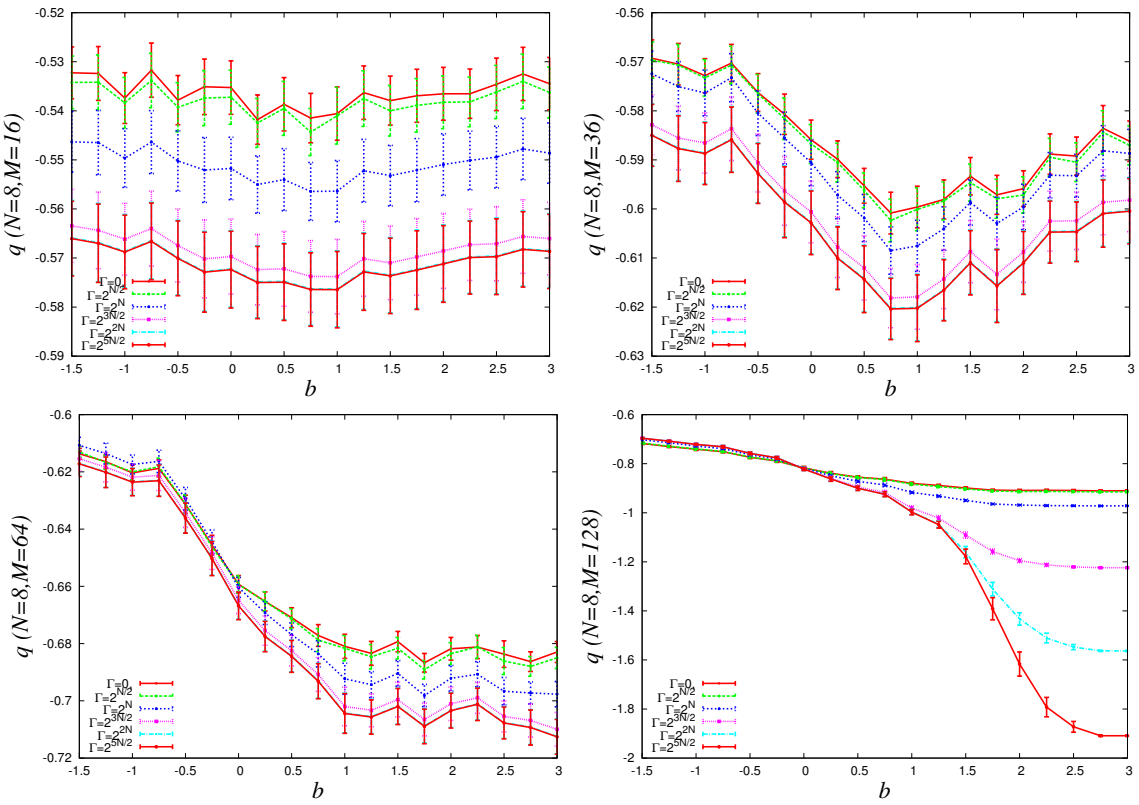

Figure 6. Plots of q for the $N = 8$ 3SM vs. b, for $M = 16, 36, 64$, and 128 from left to right and from upper to lower. The curve for $M = 36$ (right upper) shows a dip around $b = 0.8$, implying the existence of an optimal value of b.

6. In this case, we calculate a wider range values of b and Γ. The results of $M = 16, 64$, and 128 are very similar to the $2SM$ case. A characteristic dip is observed in $M = 36$, implying the existence of an optimal value of b, though the difference from other values of b is not so strong.

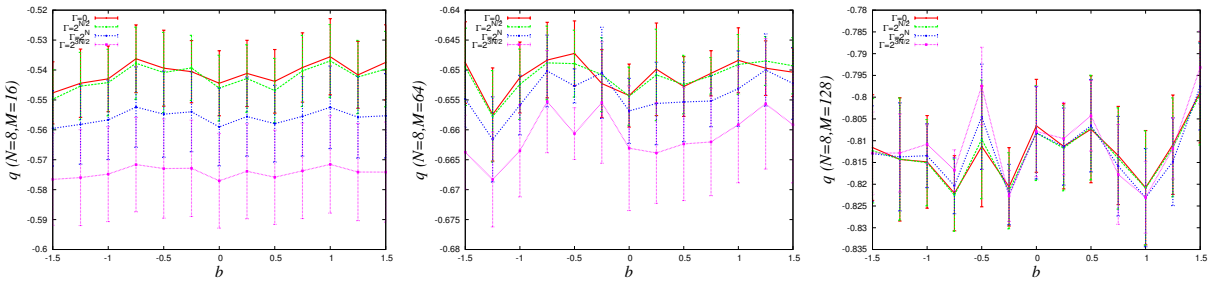

Figure 7. Plots of q for the $N = 8$ REM vs. b, for $M = 16$, 64, and 128 from left to right.

Fig. 7 shows the result for the REM. We see no clear dependence of q on b for any value of M, meaning that the smoothness does not help the inference at all, as expected.

The above results for the 2SM, 3SM, and REM clearly exhibit a certain compatibility between the observables and the target distribution. For further examining this, the 3SM result, especially the existence of an optimal value of b, can be a good starting point. To interpret this result, we introduce $N_k\left(\{\boldsymbol{v}_\mu\}_\mu\right)$ representing the number of modes with the wavenumber k in a given set of observables $\{\boldsymbol{v}_\mu\}_\mu$, and define the average occupation ratio of the k-modes by

$$u_k = \left[N_k\left(\{\boldsymbol{v}_\mu\}_\mu\right)\right]_b \binom{N}{k}^{-1}, \tag{23}$$

where the brackets $[\cdots]_b$ denote the average over the distribution with the smoothness b stated above. This quantity is easy to calculate because this is independent of the target distribution and moreover the MC simulation is not needed. The occupation ratios for $M = 16, 36, 64$, and 128 are shown in Fig. 8. A remarkable quantity for examining the 3SM result is u_3 which shows the following behavior

- For $M = 36$, it exhibits a clear peak around $b = 0.8$.
- For $M = 64$, it increases monotonically as b grows up to $b = 4$.
- For $M = 128$, it saturates to unity for $b \gtrsim 2$.

These observations are probably connected to the characteristic behavior of q: the dip for $M = 36$, the monotonic decrease of q with b for $M = 64$, and the significant decrease of q as Γ grows at large values of b for $M = 128$. It is also interesting to see how q behaves as Γ changes when $u_3 = 1$ is satisfied, which is demonstrated in the right panel of Fig. 9. For $b = 3$, where $u_3 = 1$, q seems to decrease as a linear function of γ in the large γ region, while at small values of b the dependence on γ is quite weak and can become positive (decreasing q) and negative (increasing q). For comparison, we also give the same plot of $M = 64$ in the left panel, in which no significant decreasing of q occur.

3.2.2. Consequences of the smooth-case results. The above results certainly demonstrate some situations where the ME distribution, and more generally, the introduction of an entropic bias helps to infer the target distribution. However, the interpretation based on u_k implies the close connection between target distribution and observables. If we know the "smoothness" of a target distribution, we can determine the appropriate observables for applying to the target system, but in practical situations the nature of the target distribution is unclear.

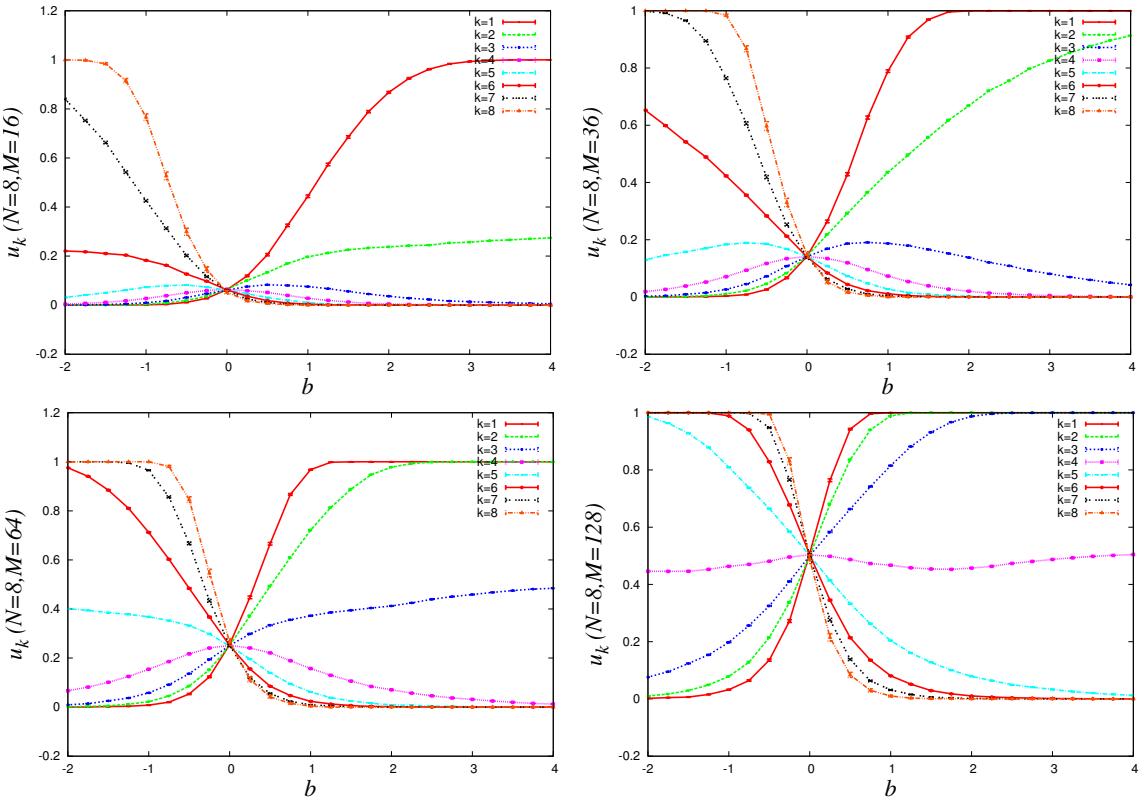

Figure 8. The occupation ratio of Fourier modes with wavenumber k for $M = 16, 36, 64$, and 128 in $N = 8$. Different colors correspond to different k.

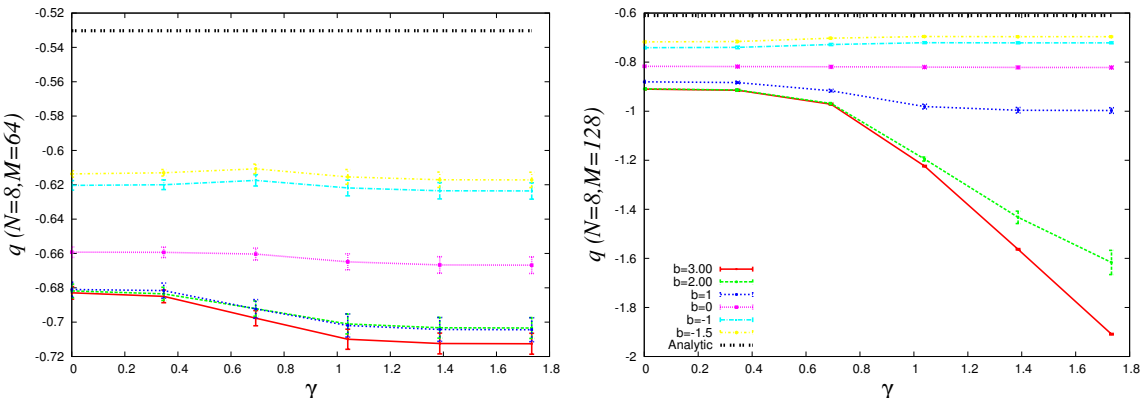

Figure 9. Plots of q for the $N = 8$ 3SM against γ, for $M = 64$ (left), and 128 (right). Black dotted lines denote the analytic solution derived in the random case. For $M = 128$ at $b = 3$, q seems to decrease linearly as γ grows at large values of γ.

3.3. On control of the ratio of "correct" observables.
Here, we examine a quantitative relation between q and the ratio of "correct" observables. A "correct" observable here is such that the corresponding interaction in the target energy, see eq. (21), is nonzero. For example in the 2SM, only observables with the wavenumber $k = 2$ are correct. Let us consider the 2SM and $M = \binom{N}{2}$ observables. We assume that the observables are a mixture only of $k = 2$ and 3 modes. We treat u_2 as a control parameter on this setup, in contrast to the smooth case where u_k is the function of b and M.

In Fig. 10, we give the plot of q against u_2 of the $N = 10$ 2SM. To examine the size dependence of q, we also show the u_2-q plots for $N = 6, 8$ and 10 at $\Gamma = 0$ and $2^{5N/2}$ in Fig. 11. In Fig.

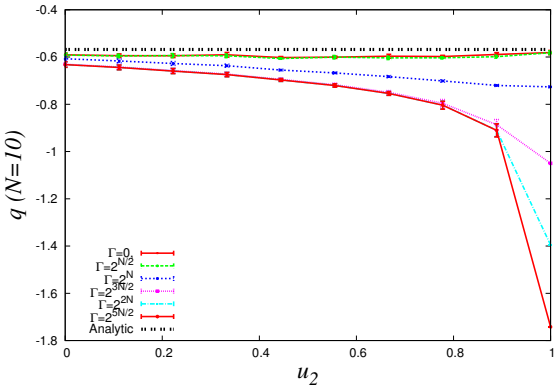

Figure 10. The order parameter q is plotted against u_2 in the $N = 10$ 2SM. Different colors correspond to different Γ values. At u_2=1, q will go to $-\infty$ in the limit $\Gamma \to \infty$.

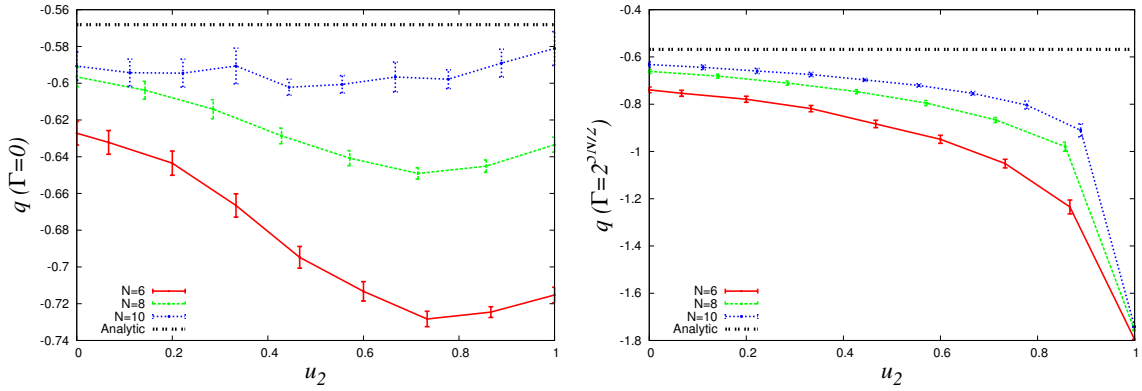

Figure 11. The u_2-q plots at $\Gamma = 0$ (left) and $2^{5N/2}$ (right). Different colors represent different system sizes. Black dotted lines denote the analytic solution in the random case.

10, we see that q decreases as u_2 increases at large values of Γ. Especially at $u_2 = 1$, the ME distribution appearing in the limit $\Gamma \to \infty$ recovers the target distribution and thus q should go to $-\infty$ in that limit. This tendency is actually observed in Fig. 10. On the other hand from Fig. 11, at $\Gamma = 0$ the u_2-q curves seem to converge to the analytical solution in the random case for any u_2 as the system size grows. A similar behaviour is expected even for large Γ if u_2 is not too large. These observations suggest to focus on u_2 close to, but slightly smaller than 1.

Comparing the size dependence of q at $u_2 = 1$ and around $u_2 = 0.9$ in Fig. 11, we find that the size dependence at $u_2 = 1$ seems to converge already for $N = 10$ and q in the limit $N \to \infty$ is expected to be different from the random-case one, while the one around $u_2 = 0.9$ is still significant and q seems to converge to the random-case value as N grows. This observation yields a pessimistic description: in the $N \to \infty$ limit, only the $u_2 = 1$ point is special and for all other values of $u_2 < 1$ the order parameter q converges to the random-case value irrespectively of the ME bias Γ.

These results imply that the ME principle does not work even as an approximation, at least in the present scale of the analysis, unless we exactly know what kind of interactions are needed to describe the target distribution. This consequence is only based on a limited result, and hence in the next subsection we reexamine it from another viewpoint.

3.4. Comparison of low- and high-$|\boldsymbol{k}|$ modes

As stated above, in realistic situations low-k modes are thought to be more important than high-k modes, and $k = 1$ and 2 modes are used in most applications. We now examine the possible difference in approximation performance between low-k and high-k modes. We choose 3SM as the target system. Due to computational limitations, we fix the system size to $N = 8$. The number of observables is fixed as $M = 36$ which is equal to $\binom{N}{1} + \binom{N}{2}$ with $N = 8$. Hence, we can include all $k = 1$ and 2 modes, the common choice in the applications, in a set of observables. Let us label this set by $a = 1$. We also define sets of observables labelled by $a = 2, 3, 4$, and 5. The sets $a = 2, 3$, and 4 consists only of $k = 3, 4$, and 5 modes, respectively, and the $a = 5$ one is a mixture of all the $k = 6$ and 7 modes. We compare the values of q among those different sets. Since the target system is the 3SM, the set $a = 2$ ($k = 3$) gives a better performance than the others, which is trivial. An interesting possibility is whether the results for $a = 1$ ($k = 1, 2$) are better than the ones for $a = 3, 4$, and 5. If so, it could give support to why the ME distribution constructed with $k = 1$- and 2–observables works well in many applications. Unfortunately, we do not see any obvious difference in performance between $a = 1$ and $a \geq 3$ in Fig. 12.

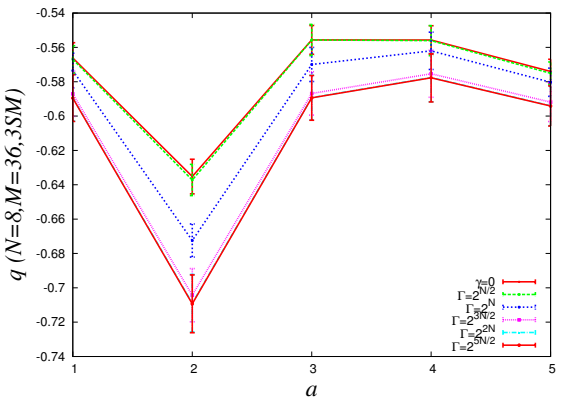

Figure 12. The order parameter q is plotted against u_2 in the $N = 10$ 2SM. Different colors correspond to different Γ values. At u_2=1, q will go to $-\infty$ in the limit $\Gamma \to \infty$.

4. Analytical treatment of the smooth-observables case: some results

In this section, we provide an analytical solution in some limited cases. We start from stating some assumptions and approximations for making the analysis amenable.

4.1. Assumptions and approximations

4.1.1. Prior over the target distribution and the annealed approximation. We have introduced the prior distribution of observables (8) with the matrix $M_{\boldsymbol{st}}$ which controls the smoothness of observables. Similarly, we here introduce the prior over the target distribution. This was irrelevant in the random case due. In the presence of nontrivial $M_{\boldsymbol{st}}$, however, we have to treat correlations among different spin configurations, which requires to specify the properties of the target distribution. The prior distribution of the target distribution is chosen as the following quadratic form

$$\hat{P}\left(\hat{\boldsymbol{p}}\right) = \hat{C}\prod_{\boldsymbol{s}}\theta(\hat{p}_{\boldsymbol{s}})\int_{-i\infty}^{i\infty} d\hat{\Lambda}\; e^{\hat{\Lambda}\sum_{\boldsymbol{s}}(\hat{p}_{\boldsymbol{s}}-1)-\frac{1}{2}\sum_{\boldsymbol{s},\boldsymbol{t}}\hat{M}_{\boldsymbol{st}}\hat{p}_{\boldsymbol{s}}\hat{p}_{\boldsymbol{t}}}. \tag{24}$$

The matrix $\hat{M}_{\boldsymbol{st}}$ controls the smoothness of the target distribution as $M_{\boldsymbol{st}}$ in eq. (8). The integration variable Λ keeps the normalization condition $\sum_{\boldsymbol{s}}\hat{p}_{\boldsymbol{s}} = 1$. The factor $\hat{C}$ is just the normalization.

As eq. (10) in the random case, we calculate the averaged logarithm of the partition function, F, by using the replica method with the RS ansatz. After some calculations, we get

$$\log\left[V^n\right]_{\rm smooth} = \underset{\Omega}{\rm Extr}\Bigg\{-\frac{M}{2}\left\{\log\left(1+\frac{Q+(n-1)R)}{E}\right)+(n-1)\log\left(1+\frac{Q-R}{E}\right)\right\}$$
$$+\frac{1}{2}nQ(Q'-R')-\frac{1}{2}n(n-1)RR'+\log\int\frac{\sqrt{\det M_{\boldsymbol{st}}}}{\sqrt{2\pi}^{2^N}}\left(\prod_s dz_s\right)e^{-\frac{1}{2}\sum_{\boldsymbol{s},\boldsymbol{t}}M_{\boldsymbol{st}}z_{\boldsymbol{s}}z_{\boldsymbol{t}}}X,\Bigg\} \tag{25}$$

where the brackets $[\cdots]_{\rm smooth}$ denote the average over eqs. (8,24), Ω represents the set of order parameters $\Omega = \{Q, Q', R, R', \Lambda, \Lambda'\}$, and

$$X = \int_0^\infty\prod_{\boldsymbol{s}}d\hat{p}_{\boldsymbol{s}}\hat{P}(\hat{\boldsymbol{p}})Y^n, \tag{26}$$

$$Y = \int_0^\infty\prod_{\boldsymbol{s}}dp_{\boldsymbol{s}}\exp\Bigg\{\sqrt{R'}\sum_{\boldsymbol{s}}z_{\boldsymbol{s}}(p_{\boldsymbol{s}}-\hat{p}_{\boldsymbol{s}})-\Lambda\sum_{\boldsymbol{t}}(p_{\boldsymbol{t}}-\hat{p}_{\boldsymbol{t}})$$
$$-\frac{1}{2}Q'\sum_{\boldsymbol{s},\boldsymbol{t}}M^{-1}_{\boldsymbol{st}}(p_{\boldsymbol{s}}-\hat{p}_{\boldsymbol{s}})(p_{\boldsymbol{t}}-\hat{p}_{\boldsymbol{t}})-\Gamma\sum_{\boldsymbol{s}}p_{\boldsymbol{s}}\log p_{\boldsymbol{s}}\Bigg\}. \tag{27}$$

Now, we should in principle take the quenched limit $n\to 0$, but unfortunately, the computation in this limit is far from being an easy task. Instead, we work in the annealed case $n = 1$, and the resultant formula becomes much simpler

$$\log\left[V\right]_{\rm smooth} = \underset{\Omega'}{\rm Extr}\left\{-\frac{M}{2}\log\left(1+\frac{Q}{E}\right)+\frac{1}{2}QQ'+\log X_{\rm anneal}\right\}, \tag{28}$$

where $\Omega' = \{Q, Q', \Lambda, \Lambda'\}$

$$X_{\rm anneal} = \int_0^\infty\prod_{\boldsymbol{s}}d\hat{p}_{\boldsymbol{s}}\hat{P}(\hat{\boldsymbol{p}})\int_0^\infty\prod_{\boldsymbol{s}}dp_{\boldsymbol{s}}\exp\Bigg\{-\Lambda\sum_{\boldsymbol{t}}(p_{\boldsymbol{t}}-\hat{p}_{\boldsymbol{t}})$$
$$-\frac{1}{2}Q'\sum_{\boldsymbol{s},\boldsymbol{t}}M^{-1}_{\boldsymbol{st}}(p_{\boldsymbol{s}}-\hat{p}_{\boldsymbol{s}})(p_{\boldsymbol{t}}-\hat{p}_{\boldsymbol{t}})-\Gamma\sum_{\boldsymbol{s}}p_{\boldsymbol{s}}\log p_{\boldsymbol{s}}\Bigg\}. \tag{29}$$

4.1.2. A simple choice of $M_{\boldsymbol{st}}$ and $\hat{M}_{\boldsymbol{st}}$. As noted above, $M_{\boldsymbol{st}}$ should only depend on the overlap $\sum_i s_i t_i$. Here we consider one of the simplest choice obeying this constraint:

$$M_{\boldsymbol{st}}^{-1} = \frac{e^{B\sum_i s_i t_i}}{(2\cosh B)^N}. \tag{30}$$

The parameter $B \in \mathbb{R}$ controls the smoothness of observables and the random case is reproduced in the limit $B \to \infty$. Note that B is defined from the inverse of $M_{\boldsymbol{st}}$. The matrix $M_{\boldsymbol{st}}$ itself can be written in the same form, but the parameter corresponding B becomes complex in general.

How about the matrix $\hat{M}_{\boldsymbol{st}}$? We do not want to bias a specific spin and thus $\hat{M}_{\boldsymbol{st}}$ should have the permutation symmetry of the spin index, meaning that $\hat{M}_{\boldsymbol{st}}$ is a function of the overlap $\sum_i s_i t_i$ and the two magnetizations $\sum_i s_i$ and $\sum_i t_i$ only. The symmetry between the two magnetizations should be also maintained. Here we choose

$$\hat{M}_{\boldsymbol{st}} = \hat{C} e^{\hat{B}\sum_i s_i t_i + \hat{H}(s_i + t_i)}. \tag{31}$$

As seen below, this functional form leads to the ISM as the target distribution.

4.1.3. A saddle-point approximation of the integration over $\boldsymbol{p}$. A notable point of the present formulation is that all the variables obey the exponential scale. Thanks to this, we can approximate the integration over $\boldsymbol{p}$ by a saddle-point (SP) technique. Actually, this approximation gives the exact result, which was demonstrated in [11] in the random case. We expect the same is true in the present case. Hence, instead of integrating out over $\boldsymbol{p}$ and $\hat{\boldsymbol{p}}$, we take the SP condition on the exponent of X_{anneal}. The result is

$$-\Lambda - Q' \sum_{\boldsymbol{t}} M_{\boldsymbol{st}}^{-1} p_{\boldsymbol{t}} + Q' \sum_{\boldsymbol{t}} M_{\boldsymbol{st}}^{-1} \hat{p}_{\boldsymbol{t}} - \Gamma(\log p_{\boldsymbol{s}} + 1) = 0, \tag{32}$$

$$\hat{\Lambda} + Q' \sum_{\boldsymbol{t}} M_{\boldsymbol{st}}^{-1} (p_{\boldsymbol{t}} - \hat{p}_{\boldsymbol{t}}) - \sum_{\boldsymbol{t}} \hat{M}_{\boldsymbol{st}}^{-1} \hat{p}_{\boldsymbol{t}} = 0, \tag{33}$$

Taking the extremization condition with respect to Q in eq. (28), we see Q' is proportional to M. Hence, for small M, we find that the SP solution of $\hat{\boldsymbol{p}}$ gives back the ISM

$$\hat{p}_{\boldsymbol{s}}^* = \hat{\Lambda} \sum_{\boldsymbol{t}} \hat{M}_{\boldsymbol{st}} = \frac{e^{H^* \sum_i s_i}}{(2\cosh H^*)^N}, \tag{34}$$

where H^* is a function of $\hat{B}$ and $\hat{H}$ in eq. (31). We expect the smallness of the second term in eq. (33) to hold for general M. Therefore, we hereafter choose the ISM as the target distribution. This is natural since the target distribution should not be changed as M increases. Thanks to this assumption, we can concentrate on solving only eq. (32) for $\hat{\boldsymbol{p}}$ defined by the ISM, see eq. (13).

Note that variable $p_{\boldsymbol{s}}$ is a probability value and cannot be negative, a fact ignored in deriving eq. (32). We give the name "SAT" to the set of spin configurations $\mathbf{s}$ whose probability values derived from the SP equation becomes positive, and "UNS" to the other configurations. Thus the correct SP equation is

$$\begin{cases} -\Lambda - Q' \sum_{\boldsymbol{t} \in \text{SAT}} M_{\boldsymbol{st}}^{-1} p_{\boldsymbol{t}} + Q' \sum_{\boldsymbol{t}} M_{\boldsymbol{st}}^{-1} \hat{p}_{\boldsymbol{t}} - \Gamma(\log p_{\boldsymbol{s}} + 1) = 0, & (\boldsymbol{s} \in \text{SAT}) \\ p_{\boldsymbol{s}} = f_{\boldsymbol{s}}(\Gamma), & (\boldsymbol{s} \in \text{UNS}) \end{cases} \tag{35}$$

This is the extension of the modified SP approximation used in [11]. For UNS, the function $f_{\boldsymbol{s}}(\Gamma)$ is expected to be a simple function of Γ. For example in the random case, $f_{\boldsymbol{s}}(\Gamma) = 0$ for $0 \le \Gamma < 2^N$ and $f_{\boldsymbol{s}}(\Gamma) = 1/2^N$ for $\Gamma \ge 2^N$.

The difficulty in solving this set of coupled equations is the classification of the SAT and UNS configurations. In the random case $B \to \infty$, target probabilities are learned in descending order of their values, and the structure in the spin-configuration space is completely irrelevant: the classification of the SAT and UNS configurations is reduced to the self-consistent determination of the boundary value, namely, the learning edge, between the learned, large probabilities and the unlearned, small ones. However in the present case, due to the presence of the nontrivial matrix $M_{\boldsymbol{st}}$, we have to treat the spin-configuration space itself, and the classification of the SAT and UNS configurations is very complicated.

In the following subsection, we assume that the learning edge can be uniquely determined for finite but still enough large B, which enables us to conduct the analysis. The applicable range of this assumption is also discussed.

4.2. Perturbation from large B at $\Gamma = 0$

In the discussion below, we put $\Gamma = 0$ in eq. (76) and use the following truncation with a fixed $\Xi \in \mathbb{N}$

$$\sum_{\boldsymbol{t}} M^{-1}_{\boldsymbol{st}}(\cdots) \approx \sum_{\{\boldsymbol{t}|Q_{\boldsymbol{st}}=N,N-2,\cdots,N-2\Xi\}} M^{-1}_{\boldsymbol{st}}(\cdots), \tag{36}$$

where $Q_{\boldsymbol{st}} = \sum_i s_i t_i$ and Ξ is the truncation range. Let us write the solution of the SP equation as

$$p^*_{\boldsymbol{s}} = \hat{p}_{\boldsymbol{s}} - \frac{\Lambda}{Q'} + \Phi_{\boldsymbol{s}}. \tag{37}$$

For $\boldsymbol{s} \in$ SAT the residual term $\Phi_{\boldsymbol{s}}$ is zero at the random case and we expect Φ is small for enough large B, while for $\boldsymbol{s} \in$ UNS $\Phi_{\boldsymbol{s}} = \Lambda/Q' - \hat{p}_{\boldsymbol{s}}$ since $p^*_{\boldsymbol{s}} = 0$.

4.2.1. The $\Xi = 1$ case. Let us put $T_B = e^{BN}/(2\cosh B)^N$. Remembering that our target distribution is the ISM and depends only on the magnetization $M_{\boldsymbol{s}} = \sum_i s_i = N m_{\boldsymbol{s}}$, we get from the third term in the upper line of eq. (76)

$$\sum_{\{\boldsymbol{t}|Q_{\boldsymbol{st}}=N,N-2\}} M^{-1}_{\boldsymbol{st}} \hat{p}_{\boldsymbol{t}} = T_B \left\{ \hat{p}_{M_{\boldsymbol{s}}} + e^{-2B} \left(D_{\boldsymbol{s}} \hat{p}_{M_{\boldsymbol{s}}+2} + U_{\boldsymbol{s}} \hat{p}_{M_{\boldsymbol{s}}-2} \right) \right\}, \tag{38}$$

where $U_{\boldsymbol{s}}(D_{\boldsymbol{s}})$ is the number of up (down) spins of the configuration $\boldsymbol{s}$. The variables $M_{\boldsymbol{s}}, U_{\boldsymbol{s}}$, and $D_{\boldsymbol{s}}$ are extensive and proportional to N. Hence, to validate the above truncation, we need

$$\epsilon \equiv N e^{-2B} \ll 1. \tag{39}$$

Inserting the solution form eq. (37) into eq. (76), we get

$$\epsilon \frac{(1-m_s)}{2} \Phi_{M_s+2} + \Phi_{M_s} + \epsilon \frac{(1+m_s)}{2} \Phi_{M_s-2} = \frac{\Lambda}{Q'} \frac{1 - T_B(1+\epsilon)}{T_B}. \tag{40}$$

The term in the righthand side appears due to the modification from the normalization

$$0 = 1 - \sum_t M^{-1}_{st} \approx 1 - T_B(1+\epsilon). \tag{41}$$

So, we can neglect the righthand side of eq. (40).

$M_{\boldsymbol{s}}$-dependent solution. As $\Phi = \text{const}$ is an inappropriate solution to eq. (40), we look for a $M_{\boldsymbol{s}}$-dependent solution. To lighten notations, we change the index of $\hat{p}$ from the magnetization $M_{\boldsymbol{s}}$ into the number of up spins, $U_{\boldsymbol{s}}$, and carry out the same change of notations on the other variables. The equation we want to solve is now

$$\epsilon(1-u_{\boldsymbol{s}})\Phi_{U_{\boldsymbol{s}}+1} + \Phi_{U_{\boldsymbol{s}}} + \epsilon u_{\boldsymbol{s}}\Phi_{U_{\boldsymbol{s}}-1} = 0, \tag{42}$$

where $u_{\boldsymbol{s}} = U_{\boldsymbol{s}}/N$. The general solution of this equation is not simple due to the $u_{\boldsymbol{s}}$-dependence of the coefficients. Thus we work on the large N limit and assume the scale $\Phi(U_{\boldsymbol{s}}) = e^{N\phi(u_{\boldsymbol{s}})}$. Correspondingly,

$$\Phi_{U_{\boldsymbol{s}}\pm 1} = e^{N\phi(u_{\boldsymbol{s}}\pm 1/N)} \approx e^{N\phi(u_{\boldsymbol{s}})} e^{\pm \frac{\partial\phi}{\partial u}\big|_{u=u_{\boldsymbol{s}}}}. \tag{43}$$

The SP equation becomes quadratic of $y = e^{\frac{\partial\phi}{\partial u}\big|_{u=u_{\boldsymbol{s}}}}$ and the solution is

$$y = \frac{-1 \pm \sqrt{1-4\epsilon^2 u_{\boldsymbol{s}}(1-u_{\boldsymbol{s}})}}{2\epsilon(1-u_{\boldsymbol{s}})}. \tag{44}$$

In the limit $\epsilon \to 0$, we expect $\Phi(U_{\boldsymbol{s}}) = 0$ if $U_{\boldsymbol{s}}$, implying $y = 0$. Thus we need to take the plus sign in eq. (44). Using this solution and integrating $\log y$ with respect to u, we get

$$\begin{aligned}\phi(u) = \phi_0 + u\log\left\{\frac{-1+\sqrt{1-4\epsilon^2 u(1-u)}}{2\epsilon(1-u)}\right\} - \frac{1}{2\epsilon}\log\left\{-\epsilon + 2\epsilon u + \sqrt{1-4\epsilon^2 u(1-u)}\right\} \\ + \frac{1}{2}\log\left\{1 - 2\epsilon^2 + 2u\epsilon^2 + \sqrt{1-4\epsilon^2 u(1-u)}\right\} \equiv \phi_0 + \phi_c(u).\end{aligned} \tag{45}$$

Since our target distribution is the ISM, there is the edge value of U, $U_E = Nu_E$, corresponding to the learning edge ω_E. The constant $\Phi_0 = e^{N\phi_0}$ is determined from the boundary condition at $U_{\boldsymbol{s}} = U_E$. We can put $\Phi_{U_E-1} = \hat{p}_{U_E-1} - \frac{\Lambda}{Q'}$, hence from the SP equation at $U_{\boldsymbol{s}} = U_E$, we get

$$\Phi(u) = -y(u_E)\left(\hat{p}_{U_E-1} - \frac{\Lambda}{Q'}\right) e^{N(\phi_c(u)-\phi_c(u_E))}, \tag{46}$$

where we used a simple relation

$$\frac{2\epsilon u_E}{1+\sqrt{1-4\epsilon^2 u_E(1-u_E)}} = -y(u_E). \tag{47}$$

The corresponding p^*_U thus becomes

$$p^*_U = \hat{p}_U - \frac{\Lambda}{Q'} - y(u_E)\left(\hat{p}_{U_E-1} - \frac{\Lambda}{Q'}\right) e^{N(\phi_c(u)-\phi_c(u_E))}. \tag{48}$$

It is easy to confirm that $\text{sign}(y(u)) = -1$ and $0 < |y(u)| < 1$ for any $0 < \epsilon < 1$ and $0 < u < 1$. Hence, eq. (48) has a good interpretation. The factor $y(u_E)e^{N(\phi_c(u)-\phi_c(u_E))}$ controls the changing speed when we change U around U_E. Namely,

$$p^*_{U_E+K} \approx \hat{p}_{U_E+K} - \frac{\Lambda}{Q'} + (\text{sign}(y(u_E)))^{K+1}|y(u_E)|e^{K\log|y(u_E)|}\left(\frac{\Lambda}{Q'} - \hat{p}_{U_E-1}\right). \tag{49}$$

If the increasing speed $e^{K\log|y(u_E)|}$ is slower than the one of $\hat{p}_{U_E+K} = \hat{p}_{U_E}e^{-\frac{\partial\omega}{\partial u}K}$, we can see all $U \geq U_E$ are fine and the corresponding $\{p^*_U\}_{U\geq U_E}$ take positive values, then the learning

edge is well-defined. This is the case for $\epsilon < 1$ since $\log|y|$ is always negative and smaller than $-\frac{\partial\omega}{\partial u} = \frac{H^*}{2} > 0$. Thus our analysis is consistent with the assumptions.

We should clarify the magnitude relation between Λ/Q' and $\hat{p}_{U_E}$. We expect the righthand side of eq. (49) to be positive if $U \geq U_E$, and zero or negative for $U \leq U_E - 1$. The definition of y leads to a relation $e^{N\{\phi_c(u_E - 1/N) - \phi_c(u_E)\}} \approx y^{-1}(u_E)$, which simplifies the inequalities for $U = U_E$ and $U = U_E - 1$:

$$\hat{p}_{U_E} \geq \frac{\Lambda}{Q'}\{1 - y(u_E)\} + y(u_E)\hat{p}_{U_E-1}, \tag{50}$$

$$\hat{p}_{U_E-1} \leq \frac{\Lambda}{Q'}\left\{1 - y(u_E)y^{-1}(u_E)\right\} + y(u_E)y^{-1}(u_E)\hat{p}_{U_E-1} = \hat{p}_{U_E-1}. \tag{51}$$

Hence the SP equation at $U_{\boldsymbol{s}} = U_E - 1$ is marginally satisfied, implying $p^*_{U_E-2} < 0$. These inequalities imply

$$\hat{p}_{U_E-2} < \frac{\Lambda}{Q'} < \hat{p}_{U_E}. \tag{52}$$

Effective number of observables. In the usual situation, we calculate the learning edge ω_E for a given number of observables M, but here we perform the opposite analysis, namely we calculate the effective number of observables $M_{\rm eff}$ for a given ω_E. If $M_{\rm eff}(\omega_E)$ is smaller than the one in the random case with the same value of ω_E, it means that the inference improves as a result of the introduction of the nontrivial matrix $M_{\boldsymbol{st}}$.

From the extremization conditions with respect to Q and Q' in eq. (28) with negligibly small E, we get

$$Q = \left\langle \sum_{\boldsymbol{s},\boldsymbol{t}} M^{-1}_{\boldsymbol{st}}(p_{\boldsymbol{s}} - \hat{p}_{\boldsymbol{s}})(p_{\boldsymbol{t}} - \hat{p}_{\boldsymbol{t}}) \right\rangle, \tag{53}$$

$$Q'Q = M. \tag{54}$$

The summation over $\sum_{\boldsymbol{s},\boldsymbol{t}}$ in the first equation can be divided into three categories

$$\sum_{\boldsymbol{s},\boldsymbol{t}}(\cdots) = \left(\sum_{\boldsymbol{s},\boldsymbol{t}\in\mathrm{SAT}} + \sum_{\boldsymbol{s},\boldsymbol{t}\in\mathrm{UNS}} + 2 \sum_{\boldsymbol{s}\in\mathrm{SAT},\boldsymbol{t}\in\mathrm{UNS}} \right)(\cdots). \tag{55}$$

Let us assume the average over ρ, $\langle\cdots\rangle$, can be replaced by the SP values of $\boldsymbol{p}$. Then we get

$$\sum_{\boldsymbol{s},\boldsymbol{t}\in\mathrm{SAT}} M^{-1}_{\boldsymbol{st}}(p^*_{\boldsymbol{s}} - \hat{p}_{\boldsymbol{s}})(p^*_{\boldsymbol{t}} - \hat{p}_{\boldsymbol{t}}) = \sum_{\boldsymbol{s},\boldsymbol{t}\in\mathrm{SAT}} M^{-1}_{\boldsymbol{st}} \left(\Phi_{\boldsymbol{s}} - \frac{\Lambda}{Q'}\right)\left(\Phi_{\boldsymbol{t}} - \frac{\Lambda}{Q'}\right) \doteq e^{N\hat{\sigma}} \left(\frac{\Lambda}{Q'}\right)^2, \tag{56}$$

where the symbol $\doteq$ denotes the equality in the exponential scale. To derive this, we assume $\Lambda/Q' > |\Phi_{\boldsymbol{s}}|$ which is true in the random case and is strongly expected in the present case too. Similarly, we have

$$\sum_{\boldsymbol{s},\boldsymbol{t}\in\mathrm{UNS}} M^{-1}_{\boldsymbol{st}}(p^*_{\boldsymbol{s}} - \hat{p}_{\boldsymbol{s}})(p^*_{\boldsymbol{t}} - \hat{p}_{\boldsymbol{t}}) \approx \sum_{\boldsymbol{s},\boldsymbol{t}\in\mathrm{UNS}} M^{-1}_{\boldsymbol{st}}\hat{p}_{\boldsymbol{s}}\hat{p}_{\boldsymbol{t}} \doteq \sum_{U_{\boldsymbol{s}}<U_E} \hat{p}^2_{U_{\boldsymbol{s}}}, \tag{57}$$

$$\sum_{\boldsymbol{s}\in\mathrm{SAT},\boldsymbol{t}\in\mathrm{UNS}} M^{-1}_{\boldsymbol{st}}(p^*_{\boldsymbol{s}} - \hat{p}_{\boldsymbol{s}})(p^*_{\boldsymbol{t}} - \hat{p}_{\boldsymbol{t}}) \doteq e^{N\sigma_E} \left(\frac{\Lambda}{Q'}\right)^2, \tag{58}$$

where $\sigma_E = \sigma(\omega_E)$. Collecting and comparing all these terms, we see the dominant term is

$$Q \doteq \sum_{\boldsymbol{s}\in \mathrm{UNS}} \hat{p}_{\boldsymbol{s}}^2 \doteq \begin{cases} e^{N(\sigma(\omega_2)-2\omega_2)} & (\omega_E \le \omega_2) \\ e^{N(\sigma_E-2\omega_E)} & (\omega_E > \omega_2) \end{cases} = e^{Nq(\omega_E)}. \tag{59}$$

We expect $\Lambda = 2^N$ in an interesting situation. Hence, we can calculate M_{eff} from a given value of $\Lambda/Q' = \mathcal{L}$ and from eq. (54) as

$$M_{\mathrm{eff}} = \frac{2^N}{\mathcal{L}} Q \doteq e^{N(\log 2 - \omega_E + q(\omega_E))}. \tag{60}$$

The last inequality comes from $\mathcal{L} \doteq e^{-N\omega_E}$ derived from eq. (52). This effective M_{eff} is identical to the one in the random case for a given ω_E, implying that the inference is not improved at least in the exponential scale, unfortunately.

4.2.2. General $\Xi = O(1)$. Next, we proceed to the case of larger values of ϵ. Now, we need higher orders in the truncation. We assume Ξ is a $O(1)$ constant and is not extensive. As in the $\Xi = 1$ case, the SP equation is expressed as

$$\sum_{i=0}^{\Xi} \epsilon^i \sum_{j=0}^{i} \frac{u_{\boldsymbol{s}}^j}{j!} \frac{(1-u_{\boldsymbol{s}})^{i-j}}{(i-j)!} \Phi_{U_{\boldsymbol{s}}+i-2j} = 0. \tag{61}$$

To derive this, we neglect sub-leading factors in the large-size limit as

$$\binom{U_{\boldsymbol{s}}}{j} \approx N^j \frac{u_{\boldsymbol{s}}^j}{j!}. \tag{62}$$

This truncation is justified if eq. (39) is satisfied. Fortunately, we can solve this equation for general $\Xi = O(1)$. We state the outline of deriving the solution.

First we need to solve the characteristic equation

$$\sum_{i=0}^{\Xi} \epsilon^i \sum_{j=0}^{i} \frac{u_{\boldsymbol{s}}^j}{j!} \frac{(1-u_{\boldsymbol{s}})^{i-j}}{(i-j)!} y(u_{\boldsymbol{s}})^{i-2j} = 0. \tag{63}$$

This has 2Ξ different roots. The appropriate roots are the ones going to zero as $\epsilon \to 0$. As far as we searched, the appropriates roots consist of some real-valued ones and some complex-conjugate pairs. For example in $\Xi = 3$, we found that the appropriate roots are constituted by one real-valued root y_1 and one pair of complex roots y_2 and y_3 which are in the complex conjugate relation. Using these roots, we can write

$$\Phi(u) = \sum_{i=1}^{\Xi} A_i e^{N\phi_i(u)}, \tag{64}$$

where

$$\phi_i(u) = \int^u dx \log y_i(x). \tag{65}$$

Now, the coefficients $\{A_i\}$ are determined from the boundary conditions $U_{\boldsymbol{s}} = U_E, \cdots, U_{E+\Xi-1}$, leading to

$$\frac{\Lambda}{Q'} - \hat{p}_{U_E-\xi} = \Phi(U_E - \xi) \approx \sum_{i=1}^{R} A_i y_i^{-\xi} e^{N\phi_i(u_E)} \ \ (\xi = 1, \cdots \Xi). \tag{66}$$

Thus means $p^*_{U_E-\xi} = 0$, $(\xi = 1, \cdots, \Xi)$ is maintained by the boundary conditions. From the boundary conditions, we also find the magnitude relation between Λ/Q' and $\hat{p}_{U_E}$

$$\hat{p}_{U_E-\Xi-1} \leq \frac{\Lambda}{Q'} \leq \hat{p}_{U_E}. \tag{67}$$

These are the construction of the solution of eq. (61).

We confirmed up to a reasonable value of Ξ that the real part of the suitable $\{y_i(u)\}$ are decreasing function with respect to u for a certain moderate range of $\epsilon > 0$, and thus $\phi_i(u)$ too. Hence, $\Phi(U_E + K)$ is expected to decrease as K grows by the factor $e^{K \log |y_i(u_E)|}$, and the learning edge is well-defined.

The effective number of observables can be also evaluated in this case. Unfortunately, as long as $\Xi = O(1)$, $M_{\text{eff}}(\omega_E)$ cannot be different in the exponential scale from $M(\omega_E)$ in the random case, since both Q and $\mathcal{L} = \Lambda/Q' \doteq e^{-N\omega_E}$ are essentially the same as the ones of the random case. This naturally motivates us to treat $\Xi = O(N)$, but that is not easy since it becomes even nontrivial to write down the SP equation. We leave this as a future work. Instead in the next subsection, we try to examine the effect of the ME bias Γ in the same setting.

4.3. Perturbation from large B at finite Γ

4.3.1. Review of the random case. We start by reviewing the random case with relatively small M. The SP equation is

$$p_{\boldsymbol{s}} = \hat{p}_{\boldsymbol{s}} - \frac{\Lambda}{Q'} - \frac{\Gamma}{Q'}\left(1 + \log p_{\boldsymbol{s}}\right). \tag{68}$$

For large $\Gamma \geq 2^N$, the order parameter Λ behaves as

$$\Lambda = (N \log 2 - 1)\Gamma, \tag{69}$$

and the solution of the SP equation is

$$p^*_{\boldsymbol{s}} = \begin{cases} \hat{p}_{\boldsymbol{s}} - \frac{\Gamma \log 2^N \hat{p}_{\boldsymbol{s}}}{Q'}, & (\boldsymbol{s} \in \text{SAT}) \\ p_0 + \Delta_{\boldsymbol{s}}, & (\boldsymbol{s} \in \text{UNS}) \end{cases}, \tag{70}$$

where

$$p_0 = e^{-1-\frac{\Lambda}{\Gamma}} = \frac{1}{2^N},\ \Delta_{\boldsymbol{s}} = \frac{Q' p_0}{Q' p_0 + \Gamma}\left(\hat{p}_{\boldsymbol{s}} - p_0\right). \tag{71}$$

The expression of $\Delta_{\boldsymbol{s}}$ can be derived from the perturbation with respect to Δ in eq. (68). Inequalities

$$p_0 > \Delta_{\boldsymbol{s}},\ \Gamma > Q' \hat{p}_{\boldsymbol{s}}\ (^\forall \boldsymbol{s} \in \text{UNS}) \tag{72}$$

are needed to justify this perturbation. To make the above expressions consistent, the learning edge should be

$$\omega_E = q' - \gamma = q' - \lambda. \tag{73}$$

For $\Gamma < 2^N \doteq \Lambda(\Gamma = 0)$, the situation is essentially identical to the case $\Gamma = 0$. This can be understood by comparing each term of the righthand side in eq. (68). For $\boldsymbol{s} \in$ SAT the dominant term is the first term $\hat{p}_{\boldsymbol{s}}$ and for $\boldsymbol{s} \in$ UNS the second one is dominating. The last term proportional to Γ/Q' is always smaller than the other two in the exponential scale and hence essentially does not change the SP equation.

4.3.2. General $\Xi = O(1)$. A lesson from sec. 4.2.2 tells that the introduction of the nontrivial $M_{\boldsymbol{st}}$ does not change the scale of the problem if $\Xi = O(1)$, and another lesson from sec. 4.3.1 says that the effect of the ME bias Γ appears when Γ becomes comparable with Λ in the scale. These two lessons imply that the ME bias and $M_{\boldsymbol{st}}$ cannot provide a meaningful effect on the inference in the present setting.

As an example, let us see the SP equation for the truncation range $\Xi = 1$ with Γ which can be written as

$$\Phi(u_{\boldsymbol{s}})\left(\epsilon(1-u_{\boldsymbol{s}})+1+\epsilon u_{\boldsymbol{s}}\right) = -\frac{\Gamma}{Q'}\left(1+\log p_{\boldsymbol{s}}\right), \tag{74}$$

We found that $\Phi(u_{\boldsymbol{s}}) = p^*_{u_{\boldsymbol{s}}} - \hat{p}_{u_{\boldsymbol{s}}} + \frac{\Lambda}{Q'}$ is the decreasing function of u, and the leading-order terms in Φ are $p^*_{\boldsymbol{s}} \doteq \hat{p}_{\boldsymbol{s}}$ for $\boldsymbol{s} \in$ SAT and Λ/Q' for $\boldsymbol{s} \in$ UNS. The scale of Φ itself is smaller than $p^*_{\boldsymbol{s}} \doteq \hat{p}_{\boldsymbol{s}}$ for $\boldsymbol{s} \in$ SAT due to the cancellation in $p^*_{\boldsymbol{s}} - \hat{p}_{\boldsymbol{s}}$. If Γ/Q' is smaller than any of these two terms, the situation will be identical to $\Gamma = 0$ as the random case with $\Gamma < 2^N$. This is consistent to the leading order of eq. (74). As Γ grows and $\Gamma \doteq \Lambda(\Gamma = 0) \doteq 2^N$, the situation is changed: Λ scales as Γ for $\Gamma > 2^N$ and the learning edge determined by the balance between Λ/Q' and $\hat{p}_{\boldsymbol{s}}$ starts to decrease since Q' is kept to be constant against Γ. Thus introducing the ME bias worsens the inference performances, as found with the annealed approximation in the random case.

This situation seems to remain the same as long as $\Xi = O(1)$ where the coefficient of Φ is $O(1)$. Hence, from the annealed calculations, we cannot find any good effect on inference produced by the ME bias.

4.3.3. Some comments on the quenched case. A crucial difference between the quenched and annealed cases is the behavior of Q'. The extremization condition over Q' in eq. (25) after taking the $n \to 0$ limit becomes

$$Q' = \frac{M}{D}, \tag{75}$$

and D is scaled as $D = \Gamma^{-1}$ for $\Gamma > 2^N$ in the random case. Hence, Q' increases with the growth of Γ as Λ. This is good for the ME principle since the learning edge $\omega_E = q' - \lambda$ does not change as Γ grows.

On the other hand, the SP condition with respect to $\boldsymbol{p}$ becomes

$$\begin{cases} -\Lambda - Q' \sum_{\boldsymbol{t}\in \mathrm{SAT}} M^{-1}_{\boldsymbol{st}} p_{\boldsymbol{t}} + Q' \sum_{\boldsymbol{t}} M^{-1}_{\boldsymbol{st}} \hat{p}_{\boldsymbol{t}} + \frac{\sqrt{R'}}{Q'} z_{\boldsymbol{s}} - \Gamma(\log p_{\boldsymbol{s}} + 1) = 0, & (\boldsymbol{s} \in \mathrm{SAT}) \\ p_{\boldsymbol{s}} = f_{\boldsymbol{s}}(\Gamma), & (\boldsymbol{s} \in \mathrm{UNS}) \end{cases} . \tag{76}$$

The only difference with the annealed case is the presence of the term $\frac{\sqrt{R'}}{Q'} z_{\boldsymbol{s}}$ and we expect that the nature of the SP equation does not change so much.

4.4. A semi-analytical treatment: numerical evaluation of the SP equation

In the above subsections, we have concentrated on studying the truncated SP equation. For general B, $\epsilon = Ne^{-2B}$ can be large so the truncation cannot be justified. To examine this case, we numerically evaluate (76). Actually, directly evaluation of (76) is not easy, and instead of that, we minimize the following cost function

$$L = \frac{1}{2}\sum_{\boldsymbol{s},\boldsymbol{t}} M^{-1}_{\boldsymbol{st}}(p_{\boldsymbol{s}} - \hat{p}_{\boldsymbol{s}})(p_{\boldsymbol{t}} - \hat{p}_{\boldsymbol{t}}) + \frac{\Lambda}{Q'}\sum_{\boldsymbol{t}}(p_{\boldsymbol{t}} - \hat{p}_{\boldsymbol{t}}) + \frac{\Gamma}{Q'}\sum_{\boldsymbol{s}} p_{\boldsymbol{s}} \log p_{\boldsymbol{s}}, \tag{77}$$

for given Λ/Q' and Γ/Q' under the condition $p_{\boldsymbol{s}} \geq 0$. This is a convex optimization problem with constraints and can be solved by certain known algorithms such as the interior-point method. By this minimization, we can obtain the solution of the SP equation, $\boldsymbol{p}^*$, and thus related quantities can be evaluated. In solving this, we assumed the symmetry that $p^*_{\boldsymbol{s}}$ having the same magnetization value share the same probability value as well as $\hat{p}_{\boldsymbol{s}}$, which reduces a lot of computational costs.

An interesting finding might be the absence of the unique learning edge, which is demonstrated in Fig. 13. In the figure, the absolute values of the lefthand side of eq. (32)

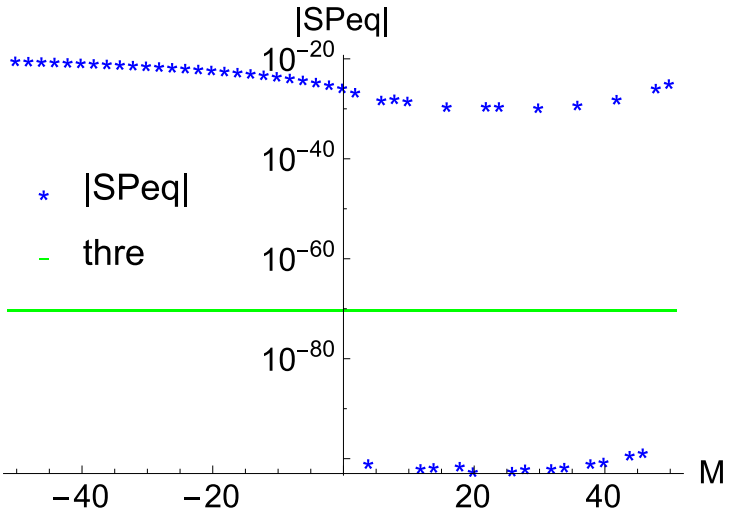

Figure 13. The absolute values of the lefthand side of eq. (32) of the upper line are plotted against the magnetization $M_{\boldsymbol{s}} = \sum_i s_i$. The target model is the ISM with $H = 0.5$, and other parameters are $N = 50, B = 0.2, H = 0.5, \Gamma = 0$ and $\Lambda/Q' = 5.0 \times 10^{-20}$. If $\boldsymbol{s} \in$ SAT (UNS), the corresponding value denoted by a blue point becomes smaller (larger) than the chosen threshold value 10^{-70} denoted by the green line. We see that the blue points locating below the threshold value are distributed in a patchy fashion, implying the absence of the learning edge.

of the upper line are plotted against the magnetization with the parameters: $N = 50, B = 0.2, H = 0.5, \Gamma = 0$ and $\Lambda/Q' = 5.0 \times 10^{-20}$. Namely, if $\boldsymbol{s} \in$ SAT, the corresponding value should be enough small (here we set the threshold value as 10^{-70}). In the random case $B \to \infty$, there exists an edge value of M, M_E corresponding to the learning edge ω_E, and all the SP-equation values for $M \geq M_E$ are smaller than the threshold, while the ones for $M < M_E$ are larger. In the present case $B = 0.2$, however, we see that the blue points locating below the threshold value are distributed in a patchy fashion and we cannot find such a clear edge value for M. This results implies that no learning edge can be defined, which makes the resolution of the SP equation directly (76) very difficult, as pointed out in sec. 4.1.3. Note that this calculation requires a very multiple precision because our variables obey the exponential scale and can be very large and small. For $N = 50$, we used 170 digits to express a real number.

5. Discussion and summary

In the present work, we have investigated the space of probability distributions constrained by smooth observables. We have first introduced a Monte Carlo procedure to sample the space of probability distribution over the set of configurations made of a small (in practice ≤ 10) number of spins. In this numerical study, we have used the order parameter q measuring the

distance between the target distribution and the averaged one to quantify the quality of inference. Our numerical investigations have revealed that the ME principle can work better than other typical distributions, in particular when the the target distribution and the observables have comparable smoothnesses. As a consequence, once we know the nature of the interactions present in the target distribution, we can construct the ME distribution precisely describing the target system. However, our numerical simulations also show that a few errors in specifying the type of interactions leads to a significant worsening of the ME distribution. This suggests that the ME principle is not good in approximating a target distribution. In practical situations, it does not seem easy to precisely know the type of the interactions of the target system, and compatibility is a serious issue.

To get deeper insight, we have also carried out some analytical investigations. We have assigned a specific functional form to the matrix controlling the smoothness of the observables and introduced a prior distribution over the target distribution. In this setting, we have analyzed the perturbative regime from the random case, combining this approach with the annealed approximation and the saddle-point approximation. We were able to construct a solution in this case, but the solution was not informative. The annealed approximation, indeed, corresponds to 'bad' choices of the constraints, making the volume V increase compared to its typical value for quenched, random observables.

One of the main outcomes of the present work is that, in the presence of smooth observables, no learning edge, separating small from large target probabilities cannot be defined, see Fig. 13, as used to be the case in the non-smooth case [11]. This result highlight the complexity of learning in the smooth case, and its collective nature. While the probabilities of the different spin configuration were essentially learned independently of each other in the non-smooth case, this is clearly not the case anymore. How to describe in a mathematically controlled and precise way this collective phenomenon remains a challenge. We hope that the remarks presented here will be useful for further progresses in this direction.

Acknowledgments

T. O. is grateful to E. Aurell for a fruitful discussion at KITPC in China. He also 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". A part of numerical calculations were carried out on the TSUBAME2.5 supercomputer in the Tokyo Institute of Technology. R.M. was partly funded by the Agence Nationale de la Recherche Coevstat project (ANR-13-BS04-0012-01) and the [EU-]FP7 FET OPEN project Enlightenment 284801.

References

[1] Jaynes E T 1957 *Phys. Rev.* **106** 620
[2] Shneidman E, Berry M J, Segev R, Bialek W 2006 *Nature* **440** 1007
[3] Tkačik G, Marre O, Mora T, Amodei D, Berry II M J, Bialek W 2013 *J. Stat. Mech.* P03011
[4] Cocco S, Leibler S, Monasson R 2009 *Proc. Nat. Acad. Sci. USA* **106** 14058
[5] Seno F, Trovato A, Banavar J R, Maritan A 2008 *Phys. Rev. Lett.* **100** 078102
[6] Weigt M, White R A, Szurmant H, Hoch J A, Hwa T 2009 *Proc. Nat. Acad. Sci. USA* **106** 67
[7] Mora T, Walczak A M, Bialek W, Callan C G 2010 *Proc. Nat. Acad. Sci. USA* **107** 5405
[8] Cocco S, Monasson R, Weigt M 2013 *PLOS Comput. Biol.* **9** e1003176; *J. Phys. Conference Series* **473** 012010
[9] Bialek W, Cavagna A, Giardina I, Mora T, Silvestri, E, Viale M, Walczak A 2012 *Proc. Nat. Acad. Sci. USA* **109** 4786
[10] Kataoka S, Yasuda M, Furtlehner C, Tanaka K 2014 *Inverse Problems* **30** 025003
[11] Obuchi T, Monasson R, and Cocco S, 2015 Learning probabilities from random observables in high dimensions: the maximum entropy distribution and others, submitted to *J. Stat. Phys.*
[12] Mézard M, Parisi G, Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[13] Bishop C M 2006 *Pattern Recognition and Machine Learning* (New York: Springer)