Microcanonical and Multicanonical Approaches to Bayesian Image Restoration

Yukito Iba*

Abstract: In this paper, we discuss image restoration with a Markov field prior that exhibits phase transition. First, we point out that hyperparameter estimation with maximum marginal likelihood procedure is often inadequate for natural images with large clusters (large domains). Next, we will discuss the ways to avoid the difficulty. Specifically, the use of the microcanonical distribution as a prior is useful for this purpose. We also discuss the concept of multicanonical prior and introduce a class of random field models that contains Markov field models as a special case.
Keywords: Markov Field, Phase Transition, microcanonical, multicanonical

1 Introduction

A summary of this paper is:

1. We survey the issues on image restoration with Markov field prior [1] that exhibit phase transition. Specifically, we point out that hyperparameter estimation with maximum marginal likelihood procedure is inadequate for natural images with large clusters (large domains) when the estimated value of the hyperparameter corresponds to the ordered phase of the prior. These issues are partly discussed in the references [1, 2, 3], but not well-known by statistical scientists and physicists. Here we will give a detailed discussion on this subject.

2. From the above viewpoint, we discuss the study [4] by Tanaka and Morita, which introduces "microcanonical distribution" to Bayesian image analysis. We argue that the microcanonical approach is essential to Bayesian inference with the models that exhibit a phase transition.

3. Along the line of the arguments that motivate the introduction of microcanonical prior, we introduce multicanonical prior for image restoration. It will be interesting that multicanonical ensemble that is originally introduced for computational purpose has a natural meaning in a Bayesian context. We also propose the concept "Glocal field" that is an extension of the concept of Markov random field.

In Sec. 2, we survey Bayesian image restoration with a Markov field prior with the example of an Ising prior. In Sec. 3, phase transition of Ising model is reviewed. In Sec. 4 its significance in the Bayesian image restoration is discussed. In Sec. 5, we list up the ways to avoid the difficulty explained in Sec. 4. In Sec. 6 and 7, we introduce micro- and multi- canonical priors. In Sec. 8, further perspectives on the subject is shown. General notion of micro/multi canonical ensemble is discussed in the appendix.

*The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 1068569, Japan, tel 03-5421-8759, iba@ism.ac.jp.

2 Image Restoration with a Markov Field Prior

Let us survey the framework of Bayesian image restoration. Here we consider an Ising prior for binary images $\{x_i\}$, $x_i \in \{\pm 1\}$ on a square lattice as a simplest non-trivial example for our purpose:

$$\pi_J(x) = \frac{\exp(J \sum_{(i,j)} x_i x_j)}{Z_J} \tag{1}$$

$$Z_J = \sum_x \exp(J \sum_{(i,j)} x_i x_j) \tag{2}$$

Here and hereafter $\sum_{(i,j)}$ is a summation over the lattice points j neighboring to i and $\sum_x$ denote summation over all configurations of $\{x_i\}$.

The model (1) corresponds to the prior knowledge that the pixels neighboring on the lattice tend to have the same color and form clusters (Here and hereafter we call $+1$ and -1 as "black" and "white" respectively.). For the images for which the assumption is valid, we can reconstruct the original image $\{x_i\}$ from an noisy observation $\{y_i\}$ by the use of the Bayes formula. Given the likelihood function $P(y|x)$ that describe the observation process, say,

$$P_h(y|x) = \frac{\exp(h \sum_i y_i x_i)}{\exp(h) + \exp(-h)}, \tag{3}$$

the posterior distribution is written as follows:

$$P_{J,h}(x|y) = \frac{\exp(J \sum_{(i,j)} x_i x_j + h \sum_i y_i x_i)}{Z_{pos}} \tag{4}$$

where Z_{pos} is a normalization constant.

A way to extract information from the posterior is to take its mode, i.e., the use of the MAP estimator defined by

$$x^{MAP} = \arg\max_x \left\{ J \sum_{(i,j)} x_i x_j + h \sum_i y_i x_i \right\} \qquad (5)$$

There are, however, more information contained in the posterior and we can use it for the calculation of the estimator optimal for a given loss function (see [5] and references therein) and the estimation of the error of an estimate.

One of the most important advantages of a Bayesian interpretation is that we can estimate hyperparameters using maximum marginal likelihood procedure (minimum ABIC procedure [1], maximum type II likelihood procedure, maximum evidence procedure)

$$l(J,h) = \log Z_{pos} = \log \sum_x P_h(y|x)\pi_J(x) \to \max . \qquad (6)$$

For the computation with Markov Chain Monte Carlo, it is convenient to use the derivatives of $l(J,h)$:

$$\frac{\partial l(J,h)}{\partial J} = \langle \sum_{(i,j)} x_i x_j \rangle_{pos} - \langle \sum_{(i,j)} x_i x_j \rangle_{\pi}, \qquad (7)$$

$$\frac{\partial l(J,h)}{\partial h} = \langle \sum_i y_i x_j \rangle_{pos} - \tanh h \qquad (8)$$

where $\langle \cdots \rangle_\pi$ and $\langle \cdots \rangle_{pos}$ are averages over the prior (1) and the posterior (4) respectively.

These formulae are also useful to discuss the nature of the maximum marginal likelihood procedure. If we set $\frac{\partial l(J,h)}{\partial J} = 0$, we obtain

$$\langle \sum_{(i,j)} x_i x_j \rangle_{pos} = \langle \sum_{(i,j)} x_i x_j \rangle_{\pi}. \qquad (9)$$

A set of values of hyperparameters J, h that satisfy the equation (9) and an equation derived from $\frac{\partial l(J,h)}{\partial h} = 0$ is chosen by the maximum marginal likelihood procedure. The relation (9) has a clear interpretation: A sample from the posterior is similar to a sample from the prior when we measure the similarity of the images by the value of $\sum_{(i,j)} x_i x_j$.

In a Bayesian framework, the posterior distribution contains all information required for the calculation of MAP or other estimators. On the other hand, for maximum marginal likelihood estimation, we need *prior averages* $\langle \cdots \rangle$ of the statistics. Thus, knowledge on the properties of the prior is required for the application of the procedure. In the following section, we will review the property of the Ising prior.

[1] The "minimum ABIC procedure" also contains a model selection procedure based on the comparison of marginal likelihoods with an AIC-like correction.

3 Phase Transition

Here we survey the phase transition and related properties of the Ising model. Mathematically, phase transition is a concept defined in the infinite-size-limit, but it also significantly affects the properties of finite size systems.

For the Ising model (1) with $J > 0$ defined on a two- or more- dimensional lattice, there are a value of coupling J called "critical coupling" J_c characterized by the following property:

- Disordered Phase $J < J_c$: The marginal distribution of the "magnetization" (the difference of the number of white and black pixels) $M = \sum_i x_i$ is a single Gaussian-like distribution around $M = 0$.

- Ordered Phase $J > J_c$: The marginal distribution of M has a double peak structure. It is approximated by a mixture of Gaussian-like distributions, whose components are located at $\pm M^*$.

This means that in the ordered phase $J > J_c$, a sample from the Ising model of a sufficiently large size N is nearly "white" ($M \sim -M^*$) or nearly "black" ($M \sim M^*$) image. In Fig. 1 and Fig. 2, images generated by the Ising model with $J < J_c$ and $J > J_c$ is shown respectively.

The value of J_c and M^* defined above is a function of system-size N and the critical coupling J_c^∞ in infinite-system-size limit is used to characterize the model. The width of a peak of the marginal of $m = 1/N \sum_i x_i$ decreases as the system-size N increases. It converges to the delta measure on the origin for $J < J_c^\infty$ and splits to two disjoint delta measures at $\pm m^\infty$ for $J > J_c^\infty$ when $N \to \infty$.

It is also useful to discuss the properties of the model $J > J_c$ with a bias term (an external field) $H \sum_i x_i$:

$$\pi_{J,H}(x) = \frac{\exp(J \sum_{(i,j)} x_i x_j + H \sum_i x_i)}{Z_{J,H}}. \qquad (10)$$

In this case, the ratio of the weights of peaks of the marginal is sensitive to the change of H, while locations of the peaks are relatively insensitive to H for a small value of H. This means that most of the samples from the distribution (10) becomes "nearly white" with $H < 0$ and "nearly black" with $H > 0$. It is impossible to realize the states $M/N \sim 0$ by the introduction of the term $H \sum_i x_i$ in the ordered phase.

Phase transition is a phenomenon occurred in non-Gaussian or discrete state space models defined in more than two-dimension. It does not occur in the Gaussian models and non-Gaussian time-series models, where the maximum marginal likelihood method is successfully applied. In the next section, we will discuss the way the existence of phase transition spoils the soundness of the Bayesian interpretation and the maximum marginal likelihood method.

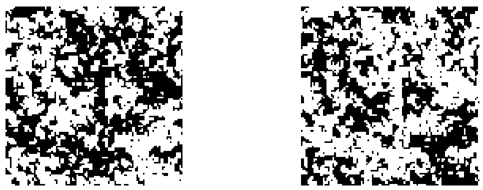

Figure 1: An example of the sample image with $J < J_c$. The negative $\{-x_i\}$ of the image $\{x_i\}$ in the left panel is shown in the right panel. Both samples appear with the same probability according to (1). They are qualitatively very similar.

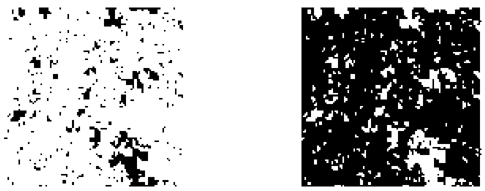

Figure 2: An example of the sample image with $J > J_c$. Again, the negative $\{-x_i\}$ of the image $\{x_i\}$ in the left panel is shown in the right panel. In this case, they looks very different: White pixels dominate in the left and black pixels dominate in the right.

4 A Difficulty in Ordered Phase

If we consider an experiment that uses images generated by Ising prior as original images, there seems no objection to the use of the maximum marginal likelihood procedure. An example of successful restoration without knowledge of J and h is found in [6, 7].

On the other hand, natural images that consist of large clusters (domains) are often well restored by the Ising prior with $J > J_c$. A problem is that the use of the maximum marginal likelihood procedure can be justified even in such a case. The answer will be negative — Even a Bayesian interpretation of the problem is doubtful in the case.

To make the situation simpler, we will consider the case that we estimate J from a noiseless image (teacher) $\{x_i^t\}$. Then the equation (9) becomes

$$\sum_{(i.j)} x_i^t x_j^t = \langle \sum_{(i,j)} x_i x_j \rangle_\pi. \tag{11}$$

The sufficient statistics $\sum_{(i.j)} x_i x_j$ represent the total length of the boundaries of white/black clusters. Thus, the requirement expressed by the equation (11) is "The boundary length of typical samples from the prior is equal to those in the natural image $\{x_i^t\}$".

This statement, however, does not mean the similarity of the samples from the prior and an image given as a teacher, because the sample images generated by the prior with $J > J_c$ never show a structure with large clusters. The size of the clusters in the image generated by the prior (1) increase as J increased, when $J < J_c$. However, when the value of the J exceeds J_c, a cluster in a color dominate in a sample and small "islands" of the other color is seen on this background. In this phase, the boundary length of the clusters is mostly controlled by boundaries between a large cluster ("sea") and small islands, instead of boundaries between several large clusters.

Fig. 3 and Fig. 4 show the comparison of images generated by the Ising model (1) and checkerboard patterns with the same total length of boundaries. It is useful to understand the above-mentioned difficulty (Here, the images with checkerboard patterns are regarded as examples of natural images that we want to treat.).

Apparently, such a difficulty, caused by the imbalance of the number of black and white pixels, will be cured by the introduction of a penalty term to the black/white ratio. However, as already discussed in the previous section, a simple exponential-family-type modification (10) with an "external field" $H \sum_i x_i$ is not effective for our purpose. The addition of H causes a discontinuous jump between white-dominant phase and black-dominant phase and does not realize desired images.

If we start a Markov chain Monte Carlo simulation for the model (1)in the ordered phase from a random

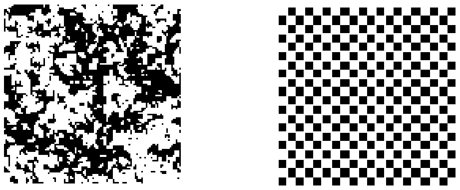

Figure 3: In the left panel, an image generated by the Ising model (1) with $J < J_c$ is shown. In the right panel, an image with a checkerboard pattern, whose total length of boundaries equals to that of the image in the left panel within the accuracy of a few percent.

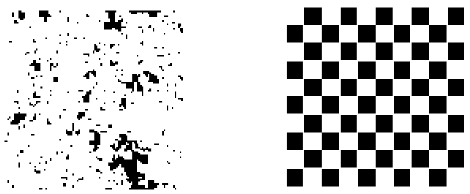

Figure 4: In the left panel, an image generated by the Ising model (1) with $J > J_c$ is shown. In the right panel, an image with a checkerboard pattern, whose total length of boundaries equals to that of the image in the left panel within the accuracy of a few percent.

initial pattern, images with several large clusters are often appeared as transient states [8]. For a large image, the mixing of Markov chain can be very slow and a huge number of MCMC sweeps are required before we get typical samples from the prior. However, the use of these "nonequilibrium" images, which is rarely generated except in the transient, is not justified in a Bayesian context. If we consider them as samples from the prior, the use of the Bayes formula is not justified.

Clearly, it is an essential problem for an Bayesian approach: The prior distribution (10) does not generate samples that are similar to the natural images that we want to reconstruct at any value of J and H.

5 *To be a Bayesian or Not to be?*

How can we circumvent the difficulty? Possible solutions are the following three:

1. Giving up Bayesian interpretations.
   In fact, we can do so at a minimum price. The MAP estimator defined by (5) can be interpreted as "maximum penalized likelihood estimator" with a log-likelihood $\log P_h(y|x) = h\sum_i x_i$ and a penalty term $J\sum_{(i,j)} x_i x_j$. With this interpretation, however, the maximum marginal likelihood procedure is not justified, and we should introduce other procedures for the estimation of J, h and the error of the MAP estimate.
2. The use of priors without phase transition.
   Another solution is to maintain the Bayesian framework and change the model. An attempt in this direction with a unilateral Markov field models is seen in [2]. It is also possible to assume the models where the probabilities of a white/black pixel at each lattice point is parameters, where a Gaussian smoothness prior to the logit of the probabilities is assumed [9].
3. To be microcanonical (or multicanonical).
   Finally, there is a way that we keep both: a model with Ising-type constraints between pixels and the Bayesian framework. In this way, we introduce the notion of microcanonical prior. Multicanonical prior is also defined as a mixture of microcanonical priors. In the following, we will discuss this approach in detail.

6 Microcanonical Prior

In a microcanonical approach introduced by Morita and Tanaka [4], we replace the "canonical" prior (10) by the corresponding microcanonical prior [2]

$$\tilde{P}_{X,M}(x) = \frac{1}{D(X,M)}\delta(\sum_{(i,j)} x_i x_j - X)\cdot\delta(\sum_i x_i - M). \tag{12}$$

Here $\delta(a-a') = \delta_{a,a'}$ is the Kronecker delta and the constant $D(X,M)$ is the number of the states x that satisfy $\sum_{(i,j)} x_i x_j = X$ and $\sum_i x_i = M$. In terms of statistical physics, the normalization constant $D(X,M)$ is called "density of states".

If we use the microcanonical prior, the difficulty discussed in the previous section disappears, because we can separately control the values of the boundary length $\sum_{(i,j)} x_i x_j$ and the ratio of black/white pixels $\sum_i x_i$ by the choice of the hyperparameters X and M.

A problem is how can we estimate the hyperparameters X and M from a noisy image. Morita and Tanaka [4] have proposed an ingenious method based on a moment estimator and a minimum modification of the model, which enable us hyperparameter estimation from a noisy image without restoration of the image.

On the other hand, a maximum marginal likelihood approach to the hyperparameter estimation requires the introduction of a "soft-version" of the microcanonical distribution. For example, we can use $\tilde{P}^*_{X,M}(x)$

[2] According to the original terminology in statistical physics, we should call it "microcanonical-micromagnetical prior" because the distribution specified by the value of M is called "micromagnetical distribution". Here and hereafter, we use the term "miicrocanonical" in a wider sense . It is also the case with "multicanonical".

defined by

$$\frac{1}{Z_{X,M}} \exp\left(-\frac{(\sum_{(i,j)} x_i x_j - X)^2}{2s_x^2} - \frac{(\sum_i x_i - M)^2}{2s_m^2} \right) \quad (13)$$

where $Z_{X,M}$ is the normalization constant and s_x, s_m is the width parameters. With this prior, we can calculate derivatives of marginal likelihood, say,

$$\frac{\partial l(X,M)}{\partial X} \propto \langle \sum_{(i,j)} x_i x_j \rangle_{pos} - \langle \sum_{(i,j)} x_i x_j \rangle_{\pi} \qquad (14)$$

where $\langle \cdots \rangle_\pi$ denote an average over the prior $\tilde{P}^*_{X,M}(x)$ and $\langle \cdots \rangle_{pos}$ denote an average over the posterior defined with the prior. However, the estimation of hyperparameter with the combination of these expressions and Markov chain Monte Carlo might be difficult, because it will be suffered from the slow mixing of the Markov chain.

7 Multicanonical Prior

Another approach to the microcanonical prior is to consider the mixture of the microcanonical priors with different hyperparameters. If we assume the uniform prior on J and h, the corresponding mixture $\hat{P}(x)$ of the microcanonical distributions (12) is calculated as $\sum_X \sum_M \tilde{P}_{X,M}(x)$, which gives

$$\hat{P}(x) \propto \frac{1}{D\left(\sum_{(i,j)} x_i x_j,\ \sum_i x_i\right)}. \qquad (15)$$

This hyperparameter-free distribution is nothing but the multicanonical distribution corresponding to the Ising model (10).

A problem in this approach is the way we compute the density of the states $D(X,M)$. A method is the use of "multicanonical Monte Carlo" algorithms [10, 11, 12], where the values of $D(X,M)$ is automatically learned with the repetition of preliminary runs of Markov chain Monte Carlo algorithm. In each run, we use the approximate values of $D(X,M)^{-1}$ estimated by the previous runs as a weight and update them until we attain a sufficient approximation.

Multicanonical distribution (multicanonical ensemble) is originally introduced by Berg and coworkers in statistical physics [10, 11] for a purely numerical purpose, i.e. for the design of efficient Markov chain Monte Carlo algorithm [3]. It is very interesting that it has a natural meaning in a Bayesian context. On the other hand, multicanonical Monte Carlo will also be useful for the calculation of posterior averages in image restoration problems, where it is used as an efficient numerical technique.

The calculation of $D(X,M)$ is computationally expensive, even with multicanonical Monte Carlo. Thus, it is also interesting to study approximations to the density of state $D(X,M)$. An idea is the use of the combinatorial techniques for entropy estimation that used in Takagi, Kikuchi, and other cluster variation methods [14]. These methods have been applied to image restoration problems as an method to approximate canonical averages [15, 16]. Our proposal is, however, very different from those in the references, because our target is a hyperparameter-free quantity $D(X,M)$. A Test of the idea is now in progress and will be reported elsewhere.

[3]The idea of the introduction of a marginalized quantity for computational purpose is discussed in [13] in a Bayesian context. There are also references that uses ideas related to multicanonical Monte Carlo in the literature of statistics. As far as I know, however, the relation to phase transition is not discussed in them.

8 Multicolor Models, Glocal Priors, Exact Tests, ...

The perspectives given in this paper can be extended in various directions. First we note that the microcanonical and multicanonical approach is also useful in other Markov field models with phase transition, e.g., Potts models [4]. For Potts models with the number of the states per a pixel (color) $q > 4$, the nature of phase transition is different from the Ising model and the difficulty discussed in this paper becomes more serious.

Next we point out that there are an interesting class of stochastic models on a lattice, here we tentatively call "Glocal"(global+local) field models. In these models, the probability is written in the form

$$P(x) = \frac{\exp(\, f(E(x))\,)}{Z} \qquad (16)$$

where $E(x)$ is an energy function of a Markov field model and f is an arbitrary function. A microcanonical model, the model (13) and a multicanonical model are examples of this class of the distributions. It will be interesting to systematically explore the spatial patterns generated by (16) with various E and f. We have a plan of the study in this direction.

Finally, we note that the idea of microcanonical prior is not entirely new in traditional statistics. For example, "Fisher's exact test" for a contingency table seems to correspond to a microcanonical approach, while a χ^2 test corresponds to the canonical approach. It will be interesting to reconsider the statistical theory around these subjects from a novel viewpoint, such as phase transition and multicanonical Monte Carlo.

References

[1] R. C. Dubes and A. K. Jain, "Random field models in image analysis," *Journal of Applied Statistics,* vol. 16, pp.131-164, 1989.

[4]A treatment of Potts model is already found in [4].

[2] A. J. Gray, J. W. Kay and D. M. Titterington, "An empirical study of the simulation of various models used for images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. PAMI-16, pp.507-513, 1994.

[3] Y. Iba, "Gakusyu to Kaisou" (Learning and Hierarchical Structures), *Proceedings of Information Integration Workshop (IIW-95)*, Real World Computing Partnership, pp.189-198, 1995 (in Japanese). Reproduced in Bussei-Kennkyu (Kyoto), vol. 65-5, pp.657-677.

[4] T. Morita and K. Tanaka," Determination of parameters in an Image recovery by statistical-mechanical means," *Physica A,* vol. 223A, pp.244-262, 1996.

[5] Y. Iba, "The Nishimori line and Bayesian Statistics," *Journal of Physics A*, vol. 32, pp. 3875-3888, 1999.

[6] Y. Iba, "Macroscopic Parameter Estimation from Incomplete Data with Metropolis-type Monte Carlo Algorithm," *Proceedings of the Institute of Statistical Mathematics*, vol. 39, pp. 1-21, 1991 (in Japansese). — An English version is found in Chap. 3 of the Ph.D thesis by the author (Y. Iba, Osaka Univ., 2000).

[7] A. Possolo (ed.), *Spatial Statistics and Imaging,* LECTURE NOTES-MONOGRAPH SERIES vol.20, Institute of Mathematical Statistics, Hayward, California, 1991.

[8] S. Miyashita and H. Takano, "Dynamical nature of the phase transition of the two-dimensional kinetic Ising model," Progress of Theoretical Physics, vol. 73, pp. 1122-1140, 1985.

[9] Y. Sakamoto and M. Ishiguro, "Bayesian binary regression involving two explanatory variables," *Annals of the Institute of Statistical Mathematics*, vol. 35B, pp.115-137, 1985.

[10] B. A. Berg and T. Neuhaus, "Multicanonical ensemble: a new approach to simulate first-order phase transitions," *Physical Review Letters*, vol. 68 pp. 9-12, 1992.

[11] B. A. Berg and T. Celik, "New approach to spin-glass simulations,", *Physical Review Letters,* vol. 69 (1992) pp. 2292-, 1992.

[12] I. Shteto, J. Linares and F. Varret, "Monte Carlo entropic sampling for the study of metastable states and relaxation paths," *Physical Review E*, vol. 56 p. 5128-, 1997.

[13] J. S. Liu, "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem", *Journal of the American Statistical Association*, vol. 89, pp. 958-966.

[14] T. Oguchi, *jiseitai no toukei-riron* (Statistical Theory of Magnets), Mokabou, Tokyo, (1970, 1982), (in Japanese).

[15] K. Tanaka and T. Morita, "Cluster variation method and image restoration problem," *Physics Letters*, vol. A203, pp.122-128, 1995.

[16] K. Tanaka, in this proceedings.

Appendix

Here we give a definition of the microcanonical distribution corresponding to a given exponential family. For an exponential family [5],

$$P_\gamma(x) = \frac{\exp(\sum_k \gamma_k \cdot t_k(x))}{Z_\gamma} \tag{17}$$

where $\{t_k(x)\}$, $k = 1, \ldots, K$ are the sufficient statistics and $\gamma = \{\gamma_k\}$ are the corresponding parameters. Then we define microcanonical distribution as

$$\tilde{P}_\mu(x) = \frac{1}{D_\mu} \prod_k \delta(t_k(x) - \mu_k) \tag{18}$$

where $\mu = \{\mu_k\}$ is parameters of the microcanonical distribution and $D_\mu = D(\mu_1, \mu_2, \ldots, \mu_K)$ is the normalization constant (the density of states). Here $\delta(a - a') = \delta_{a,a'}$ is the Dirac or Kronecker's delta. It is also convenient to consider a uniform distribution in the region defined by $\mu_k \le t_k(x) < \mu_k + \triangle_k$ with the width parameters $\{\triangle_k\}$ as a practical version of microcanonical distributions. The multicanonical distribution is defined as is the equally weighted mixture of the microcanonical distributions (18)

$$\hat{P}_\mu(x) = \frac{1}{D(t_1(x), t_2(x), \ldots, t_K(x))}. \tag{19}$$

In the exponential family (17), the natural parameters are $\{\gamma_i\}$, while the expectation parameters are $\{\mathrm{E}(t_i(x))\}$. There is one-to-one transformation between both parameterizations in a finite-size system. In some case, however, the marginal of $t_i(x)$ becomes multi-peaked and the expectation $\mathrm{E}(t_i(x))$ does not represent the frequency of $t_i(x)$. In that case, the inference with the microcanonical distribution is not equivalent to the one with the original exponential family. As we have seen in the main text, such a situation arises with a model with phase transition, where the anomaly in the infinite-system-size limit also appears.

[5] Here we generalize the term "exponential family" to indicate distributions with discrete parameters given as a limit of an exponential family.