Statistical physical view of statistical inference in Bayesian linear regression model

Kazuaki Murayama

Department of Computer and Network Engineering, Graduate School of Informatics and Engineering,
The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan

(Received 20 November 2023; revised 8 June 2024; accepted 16 August 2024; published 9 September 2024)

This paper considers similarities between statistical physics and Bayes inference through the Bayesian linear regression model. Some similarities have been discussed previously, such as the analogy between the marginal likelihood in Bayes inference and the partition function in statistical mechanics. In particular, this paper considers the proposal to associate discrete sample size with inverse temperature [C. H. LaMont and P. A. Wiggins, Phys. Rev. E **99**, 052140 (2019)]. The previous study suggested that incorporating this similarity motivates the derivation of analogs of thermodynamic functions such as energy and entropy. The study also anticipated that those analogous functions have potential to describe Bayes estimation from physical points of view and to provide physical insights into mechanisms of estimation. This paper incorporates a macroscopic perspective as an asymptotics similar to the thermodynamic limit into the previous suggestion. Its motivation stems from the statistical mechanical concept of deriving thermodynamic functions that characterize macroscopic properties of macroscopic systems. This incorporation not only allows analogs of macroscopic thermodynamic functions to be considered but also suggests a candidate for an analog of inverse temperature with continuity, which is partly consistent with the previous proposal to associate the discrete sample size with inverse temperature. On the basis of this suggestion, we analyze analogs of macroscopic thermodynamic functions for a Bayesian linear regression model which is the basis of various machine learning models. We further investigate, through the behavior of these functions, how Bayes estimation is described from the perspective of physics and what kind of physical insight is obtained. As a result, the estimation of regression coefficients, which is the primary task of regression, appears to be described by the physical picture of balance between decreasing energy and increasing entropy as in equilibrium states of thermodynamic systems. More specifically we observe the physical view of Bayes inference as follows: the estimation succeeds where the effect of decreasing energy is dominant at low temperature. On the other hand, the estimation fails where the effect of increasing entropy is dominant at high temperature.

DOI: 10.1103/PhysRevE.110.034118

I. INTRODUCTION

Interdisciplinary studies have made progress in physics and machine learning [1,2]. The studies arise in the application of machine learning techniques to issues in physics such as detection of phase transition [3–5], neural network based approaches to quantum many-body systems [6–8], and neural network representation of potential-energy surfaces [9,10]. Another direction is to apply methods from physics to problems in machine learning and related fields such as statistical learning and information theory [11–13]. For example, we mention statistical mechanical analysis of the Hopfield model [14,15], storage capacity of neural networks [16,17], and the performance of error-correcting code [18,19]. Applications of the maximum entropy principle in physics [20,21] to machine learning and related areas have also been influential. The principle was applied to construction of the prior distribution in Bayes inference [22–25], natural language processing [26,27], and density estimation [20,28]. The task of density estimation was also addressed by a field-theoretic approach [29–33].

Studies of importing physical methods into machine learning are performed through similarities between quantities in machine learning and those in physics. In particular, previous studies have indicated similarities between statistical physics and Bayes inference, which is a framework to execute machine learning and statistical inference. For example, a frequently employed analog is between the marginal likelihood in Bayes inference and the partition function in statistical mechanics [2,12,30]. In addition, several studies have been concerned with finding an analog which plays the role of temperature in Bayes inference or models. References [25,34] proposed that sample size corresponds to inverse temperature. The correspondence was also indicated in decision making associated with immune response [35] and network models for various memory strategies [36]. Our attention is drawn to what this correspondence provides. Reference [25] anticipated it as follows: Incorporating the correspondence allows us to define analogs of thermodynamic functions such as free energy, energy, and entropy for the Bayesian model under consideration. By analyzing properties of those analogous functions, we have the prospect of gaining new insights and into the mechanisms of estimation, learning, and inference associated with the Bayesian model under consideration from physical points of view.

Given the suggestion associated with temperature, several interests arise. The first is to bring a macroscopic perspective into the previous suggestion and perform the analysis with an analog of inverse temperature. More precisely, we are concerned with investigating the analogs of thermodynamic functions with macroscopic properties of the Bayesian linear regression model. Statistical mechanics essentially deals with macroscopic systems that are huge enough to be described by thermodynamics. It especially provides the procedure for obtaining thermodynamic functions that characterize macroscopic properties of macroscopic systems from a microscopic point of view. The previous study suggested that analogous thermodynamic functions for Bayes models can be derived by using the correspondence between sample size and inverse temperature together with other similarities [25]. However it has not been taken into account that thermodynamic functions provided by statistical mechanics are physical quantities which characterize macroscopic properties, obtained as a result of applying statistical mechanics to macroscopic systems. This point motivates us to incorporate the macroscopic perspective into the previous suggestion. The second interest is the difference that the inverse temperature is a continuous quantity while the sample size is discrete. The third, together with the above interests, is to analyze thermodynamic functions with macroscopic properties of the Bayesian linear regression model, which is the basis of various machine learning methods; from the behavior of those analogous functions we would like to discuss how Bayes estimation is described from the physical perspective and what kind of physical insight is obtained.

This paper addresses these issues through a Bayesian linear regression model with the following steps: First, we scrutinize an asymptotic limit in (Bayes) statistics that suggests the model is macroscopic or an infinite system, i.e., an analog of thermodynamic limit to extract macroscopic properties of the model. For the linear regression model, simultaneous limits of the sample size and the number of parameters, with their ratio converging to some constant, was used a the context reminiscent of thermodynamic limit [37,38]. A similar limit operation is also used with various names in statistics, such as the Kolmogorov limit [39] and general asymptotics [40]. Such a limit is employed to investigate the analogous thermodynamic functions with macroscopic properties, expecting it to imply that the model is macroscopic. In that case, we indicate that the ratio emerging from the limit appears to be a candidate for analogous inverse temperature with continuity. Next, combining this analogous inverse temperature with the other similarities we calculate analogs of free energy, energy, and entropy for the Bayesian linear regression model. Then we observe properties of these analogous functions in terms of analogous temperature dependence and equations which hold between them. Finally, based on these properties, we investigate what kind of physical insight can be obtained regarding the estimation mechanism of Bayesian linear regression.

As a result of the above analysis, we see the following. In thermodynamic systems, the equilibrium state with minimum free energy at a given temperature is determined by the balance between the effect of decreasing energy and that of increasing entropy. We observe that estimation of Bayesian linear regression is also understood in a similar picture using the analogous thermodynamic functions as follows: The estimation that succeeds in the regression task is understood as an equilibrium state dominated by decrease in energy at low temperatures. Alternatively, the estimation failing in the task implies an equilibrium state dominated by increase in entropy at high temperature. Furthermore, there appears to be a singularity between these estimations in the sense that the analog of free energy becomes nondifferentiable together with discontinuity in those of energy and entropy.

The structure of the paper is as follows: In Sec. II, the Bayesian linear regression model and distribution for data are provided. In Sec. III, the similarities between Bayes inference and statistical mechanics are summarized. In particular, we explain details of the asymptotic limit called, e.g., Kolmogorov asymptotics in a context reminiscent of the thermodynamic limit. In Sec. IV, we calculate the analogs of free energy, energy, and entropy in accordance with the statistical mechanical prescriptions. In Sec. V, we see properties of those analogous functions and examine what kind of physical insight we gain for the Bayesian linear regression model from those properties. In Sec. VI, we discuss our results.

II. MODEL SETTING

A. Bayesian linear regression model

Let $\{(y_\mu, \boldsymbol{x}_\mu); \mu = 1, \ldots, M\}$ be a set consisting of independent observations with sample size M, where $y_\mu \in \mathbb{R}$ is a response variable and $\boldsymbol{x}_\mu \in \mathbb{R}^N$ represent N-dimensional explanatory variables. A linear regression models the response and explanatory variables together with regression coefficients $\boldsymbol{w} = (w_1, \ldots, w_N)^{\mathrm{T}} \in \mathbb{R}^N$ as follows:

$$y_\mu = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_\mu + \epsilon_\mu, \quad \mu = 1, \ldots, M, \tag{2.1}$$

where ϵ_μ is frequently assumed to be independently and identically distributed (i.i.d.) like $\mathcal{N}(\epsilon_\mu|0, \sigma^2)$. Here, $\mathcal{N}(\epsilon_\mu|0, \sigma^2)$ represents normal distribution with mean zero and variance σ^2. Equation (2.1) is also expressed in multivariate style as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_M)^{\mathrm{T}} \in \mathbb{R}^M$ consists of the observations of response variable and we define $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)^{\mathrm{T}} \in \mathbb{R}^{M\times N}$, termed design matrix arraying explanatory variables. The second term, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)^{\mathrm{T}}$, is an error vector.

A primary task of linear regression is to estimate the regression coefficients $\boldsymbol{w} \in \mathbb{R}^N$ in Eq. (2.2) given the data $(\boldsymbol{y}, \boldsymbol{X})$. Bayes inference can be used as one of methods to perform this task. The method requires a likelihood function and prior distribution for $\boldsymbol{w}$. The likelihood function is given as

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = (2\pi\sigma^2)^{-M/2} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2\right), \tag{2.3}$$

where $\|\cdot\|$ means Euclidean norm. This study equips the regression coefficients with following prior distribution:

$$p(\boldsymbol{w}) = (2\pi)^{-N/2} \exp\left(-\tfrac{1}{2}\|\boldsymbol{w}\|^2\right). \tag{2.4}$$

The posterior distribution is constructed using Bayes's formula from the above likelihood function and prior distribution as follows:

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})}, \tag{2.5}$$

where the denominator is expressed as

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int d\boldsymbol{w}\, p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w}). \tag{2.6}$$

Equation (2.6) is called evidence or marginal likelihood in Bayes inference. The primary task of Bayesian linear regression is achieved by estimating the regression coefficients $\boldsymbol{w}$ based on Eq. (2.5). We note that this paper employs the common notation $p(\cdot)$ for the likelihood function, prior distribution, and posterior distribution. The details of distribution are distinguished by arguments.

B. Probability distribution of data

We set up probability distribution associated with data $(\boldsymbol{y}, \boldsymbol{X})$ in Eqs. (2.1)–(2.3) and (2.5), considering $(\boldsymbol{y}, \boldsymbol{X})$ as random variables. This situation is not realistic in practical data analysis. However, such a setup is required to address the subject of this paper explained in Sec. I and to carry out the theoretical analysis in later sections.

In general, it is not necessarily the case that the response variable and explanatory variables are linearly connected as in Eqs. (2.1) and (2.2). However, we assume this is the case using true regression coefficients $\boldsymbol{v} \in \mathbb{R}^N$ as follows:

$$y_\mu = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}_\mu + \varepsilon_\mu, \quad \mu = 1, \dots, M, \tag{2.7}$$

where ε_μ is distributed like $\varepsilon_\mu \sim \mathcal{N}(\varepsilon_\mu|0, \varsigma^2)$. As with Eq. (2.2) from Eq. (2.1), Eq. (2.7) is denoted as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{v} + \boldsymbol{\varepsilon}. \tag{2.8}$$

Next, we set distribution for data $(\boldsymbol{y}, \boldsymbol{X})$. The notation $p(\cdot)$ is used for the distribution associated with the Bayesian *model* in Sec. II A. To distinguish from that notation, we employ $r(\cdot)$ for the distribution that represents the data generation process. A conditional distribution of $\boldsymbol{y} \in \mathbb{R}^M$ given $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{v} \in \mathbb{R}^N$ is assumed as follows:

$$r(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{v}) = (2\pi\varsigma^2)^{-M/2} \exp\left(-\frac{1}{2\varsigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{v}\|^2\right). \tag{2.9}$$

Further, we set the true regression coefficients $\boldsymbol{v} \in \mathbb{R}^N$ to be distributed as

$$r(\boldsymbol{v}) = (2\pi)^{-N/2} \exp\left(-\tfrac{1}{2}\|\boldsymbol{v}\|^2\right). \tag{2.10}$$

Finally, each row of the design matrix is assumed to distribute according to N-dimensional multivariate normal distribution with mean vector $\mathbf{0}_N$ and covariance matrix $\boldsymbol{I}_N/N$. In this case, the distribution for the design matrix $\boldsymbol{X}$ is

$$r(\boldsymbol{X}) = \prod_{\mu=1}^{M} \left(\frac{N}{2\pi}\right)^{N/2} \exp\left(-\frac{N}{2}\|\boldsymbol{x}_\mu\|^2\right). \tag{2.11}$$

Using Eqs. (2.9)–(2.11), the joint probability distribution of data $(\boldsymbol{y}, \boldsymbol{X})$ together with the true regression coefficients $\boldsymbol{v} \in \mathbb{R}^N$ is given as follows:

$$r(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v}) = r(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{v})r(\boldsymbol{X})r(\boldsymbol{v}). \tag{2.12}$$

C. Estimation error and asymptotics

As an estimation error that measures whether the task of Bayesian linear regression is successful or not, one employs the mean squared error (MSE) per N as follows:

$$\mathrm{MSE} = \lim_{N,M\to\infty} \frac{1}{N}\mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}[\|\boldsymbol{v} - \mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{w}]\|^2], \tag{2.13}$$

where $\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{w}]$ means the mean of the posterior distribution [Eq. (2.5)], which is also the Bayes estimate of the regression coefficients $\boldsymbol{w} \in \mathbb{R}^N$. We recall that $\boldsymbol{v} \in \mathbb{R}^N$ represents the true regression coefficients set in Sec. II B. Equation (2.13), also called the Bayes risk aside from the asymptotic limit, is a measure of whether the estimation of regression coefficients $\boldsymbol{w}$ as $\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{w}]$ is successful or not. The case of MSE ~ 0 means success of the regression task, while MSE $\nsim 0$ indicates failure of the task as the value of it is away from zero.

In Eq. (2.13) the following asymptotic limit is included:

$$N, M \to \infty, \quad \alpha(M, N) = M/N \to \alpha \in (0, \infty), \tag{2.14}$$

where M represents the sample size and N is the number of parameters to be estimated, that is, the number of regression coefficients. We scrutinize this asymptotic limit in statistics so that the model becomes macroscopic or infinite. A detailed explanation of Eq. (2.14) is needed because it differs from the standard asymptotic limit in which one takes the infinite limit of M with N fixed or tending to zero [40–43]. The asymptotics of Eq. (2.14) taking the infinite limit of both N and M has various names in statistics. In linear discriminant analysis, it is called Kolmogorov asymptotics [39] or double asymptotics [44,45]. Reference [40] terms it general asymptotics in the context of constructing an estimator for a covariance matrix [46]. Additionally, the asymptotic limit and data corresponding to it are also called "large M, large N" and vice versa [41,42,47–49], in contrast with "large M, small N" or "large M, fixed N" which are the settings of classical asymptotics. In statistics, the asymptotics with infinite limit for N as well as M has been used to analyze asymptotic properties where N is comparable to M [40–43]. For example, the large N scenario was adopted to analyze the asymptotic properties of robust estimates for regression problems [50]. The asymptotics was also used for the following issues in statistics in terms of random matrices: the convergence of the empirical distribution associated with eigenvalues of random matrix [51], the estimation of a covariance matrix and its eigenvalues [47,48], the establishment of a test statistic and its asymptotic distribution to test mutual independence of random vectors [52], limiting behavior of largest and smallest eigenvalues of a covariance matrix [53–55], and limiting behavior of linear spectral statistics associated with a covariance matrix [56].

III. SIMILARITIES BETWEEN BAYES INFERENCE AND STATISTICAL MECHANICS

We summarize similarities between Bayes inference and statistical mechanics used in this paper in Table I. As a primary analogy, the evidence or marginal likelihood in Bayes inference has often been used as an analog of the partition function in statistical mechanics [2,12,30], i.e.,

$$Z_{\mathrm{B}} = p(\boldsymbol{y}|\boldsymbol{X}), \tag{3.1}$$

where $p(\boldsymbol{y}|\boldsymbol{X})$ is Eq. (2.6). The use of this correspondence in the present model implies regarding $\boldsymbol{w} \in \mathbb{R}^N$ and Eq. (2.5) as

TABLE I. The analogs between Bayes inference and statistical mechanics are summarized primarily according to [25,34] except for Eq. (2.14). Equation (2.14) has been considered a counterpart of the thermodynamic limit in the context of linear regression problems, its applications, and variations associated with it [15,16,19,37,38,57,58]. There are several points to note for Eq. (2.14). In statistics, the sample size M and the number of parameters N are often denoted as n and p, respectively. This is also the case for the term "large-M, large-N". In statistics, it is often assumed that N/M rather than M/N converges to some constant. Although we adopt the latter, it does not affect results within this study.

Quantity in Bayes inference	Name or concept in Bayes inference	Analog in statistical mechanics
$\boldsymbol{w} \in \mathbb{R}^N$ in Eq. (2.5)	regression coefficients	microstate or configuration
Eq. (2.5)	posterior distribution	canonical ensemble
Eq. (2.6)	marginal likelihood or evidence	partition function
Eq. (2.12)	population distribution or true distribution for data	quenched randomness
Eq. (2.14)	Kolmogorov asymptotics [39]	thermodynamic limit
ditto	double asymptotics [44,45]	ditto
ditto	general asymptotics [40]	ditto
ditto	large M, large N [41,42,47,48]	ditto
ditto	limit in issues of statistics with random matrices [51–56]	ditto

microstate or configuration and ensemble, respectively. The subscript in the left-hand side of Eq. (3.1), the initial of *Bayes*, is meant to emphasize that it is just an analog corresponding to the partition function.

An additional similarity is that one treats the Bayesian linear regression model as a random system quenched by the random variables associated with data; that is, $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v})$ is regarded as quenched randomness. In the present case, Eq. (2.12) is regarded as the distribution of quenched randomness.

Furthermore, a similarity between sample size and inverse temperature was pointed out in [25,34], and Ref. [25] proposed that this similarity can be used to derive analogs of thermodynamic functions. In the previous study [25], the concept that statistical mechanics provides thermodynamic functions which characterize macroscopic properties was not taken into consideration. As mentioned in Sec. I, this motivates us to bring the macroscopic perspective into Bayes inference and to perform the analysis with analogous inverse temperature.

Equation (2.14) was considered to be similar to the thermodynamic limit in the context of linear regression problems [37,38]. The asymptotic limit has also been used, in a style similar to the thermodynamic limit, for other models and applications of linear regression models [15,16,19,57,58]. Accordingly, this paper employs Eq. (2.14) to investigate analogs of thermodynamic functions with macroscopic properties, expecting the asymptotic limit to imply that the model becomes macroscopic. Reference [39] in statistics indicated that the converged value of ratio α in Eq. (2.14) becomes a new parameter of asymptotic theory. The parameter α is expected to play some role as an analog in statistical mechanics, considering the indication by Ref. [39] and that we are focusing on similarities between Bayes inference and statistical mechanics. This paper suggests that the parameter appears to be a candidate for an analog of inverse temperature with continuity. The suggestion fulfills the second interest, as mentioned in Sec. I, of seeking an alternative analog of inverse temperature with continuity instead of discrete sample size.

We explain reasons leading up to our suggestion in detail. One expects that the parameter α is treated as a continuous variable in $(0,\infty) \subset \mathbb{R}$ with the following explanations. Although $\alpha(M,N) = M/N \in \mathbb{Q}_{>0} \subset (0,\infty)$ before taking the limit is discrete as a positive rational number for finite positive integers $M, N \in \mathbb{Z}_{>0}$, the set of values of $\alpha(M,N)$ would become dense in $(0,\infty)$ as $M, N \to \infty$. In this case, we would treat α as being a continuous variable in $(0,\infty)$. This property does not contradict the property of inverse temperature in statistical mechanics. Furthermore, the ratio $\alpha(M,N) = M/N$ before the limit measures the sample size M per parameters N, which is partly consistent with the previous proposal to associate the sample size with inverse temperature [25,34]. For these reasons, adopting Eq. (2.14) considered as the counterpart of thermodynamic limit suggests that α is the candidate expected to play the analogous role of inverse temperature.

We further explain the physical motivation for introducing the analogous inverse temperature. When one sees the minus log-likelihood per sample size as an analogous Hamiltonian, i.e., $\mathcal{H} = -M^{-1} \ln p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$, M appears in the argument of the exponential function as follows:

$$Z_{\mathrm{B}} = \int d\boldsymbol{w}\, \exp\left(-M\,\mathcal{H}\right) p(\boldsymbol{w}) \tag{3.2}$$

in the correspondence shown in Eq. (3.1). References [25,34] explained that this appearance of M in Eq. (3.2) is the physical motivation to propose it as analogous inverse temperature. In the case of incorporating the macroscopic perspective as analogous thermodynamic limit, the argument of the exponential function should (asymptotically) be of the form "N times something" like

$$Z_{\mathrm{B}} = \int d\boldsymbol{w}\, \exp\left[N \times \{-\alpha(M,N)\,\mathcal{H}\}\right] p(\boldsymbol{w}). \tag{3.3}$$

This is for the thermodynamic limit of analogous free energy per N to exist, shown in Eq. (3.4) shortly. The appearance of $\alpha(M,N)$ as in Eq. (3.3) means that its convergence, α, appears (asymptotically) to be the candidate for analogous inverse temperature instead of M.

The above analogs, summarized in Table I, between Bayes inference and statistical mechanics lead us to realize an equivalent of free energy (density) as follows:

$$f_{\mathrm{B}}(\alpha) = -\lim_{N,M\to\infty} \frac{1}{\alpha N}\mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}[\ln Z_{\mathrm{B}}] \tag{3.4}$$

where $\mathbb{E}_{r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})}[\cdot]$ expresses that the expectation value is taken with respect to $r(\boldsymbol{y},\boldsymbol{X},\boldsymbol{v})$ [Eq. (2.12)] [59].

This paper has focused on the sample size ↔ inverse temperature correspondence. We note that another analog of inverse temperature using an external parameter that extends Bayesian model was also proposed (see the Appendix).

IV. ANALOGS OF THERMODYNAMIC FUNCTIONS AND ESTIMATION ERROR

A. Analog of free energy

We calculate Eq. (3.4) using a replica method with replica symmetry. References [11,57,58] are referred to for this calculation and that of Eq. (2.13). For the sake of simplicity, we focus on the calculation result in the case of $\sigma, \varsigma \sim 0$ in Eqs. (2.3) and (2.9). The result is as follows:

$$f_{\mathrm{B}}(\alpha) = \min_{Q,q,m,\tilde{Q},\tilde{q},\tilde{m}} \mathsf{f}_{\mathrm{B}}(Q,q,m,\tilde{Q},\tilde{q},\tilde{m}), \tag{4.1}$$

where f_{B} on the right side in the sense of nonminimum analogous free energy is

$$\mathsf{f}_{\mathrm{B}}(Q,q,m,\tilde{Q},\tilde{q},\tilde{m}) = \frac{1}{2}\left\{\ln[2\pi(Q-q)] + \frac{1-2m+q}{Q-q}\right\} - \frac{1}{\alpha}\left\{\frac{1}{2}\tilde{Q}Q + \frac{1}{2}\tilde{q}q - \tilde{m}m - \frac{1}{2}\ln(\tilde{Q}+\tilde{q}+1) + \frac{\tilde{q}+\tilde{m}^2}{2(\tilde{Q}+\tilde{q}+1)}\right\}. \tag{4.2}$$

To minimize Eq. (4.2) in Eq. (4.1), $Q, q, m, \tilde{Q}, \tilde{q}, \tilde{m}$ satisfy the following conditions:

$$\tilde{Q}_* = \alpha\left\{\frac{1}{Q_*-q_*} - \frac{1-2m_*+q_*}{(Q_*-q_*)^2}\right\}, \tag{4.3}$$

$$\tilde{q}_* = \alpha\left\{-\frac{1}{Q_*-q_*} + \frac{1-2m_*+Q_*}{(Q_*-q_*)^2}\right\}, \tag{4.4}$$

$$\tilde{m}_* = \frac{\alpha}{Q_*-q_*}, \tag{4.5}$$

$$Q_* = \frac{1}{\tilde{Q}_*+\tilde{q}_*+1} + \frac{\tilde{q}_*+\tilde{m}_*^2}{(\tilde{Q}_*+\tilde{q}_*+1)^2}, \tag{4.6}$$

$$q_* = \frac{1}{\tilde{Q}_*+\tilde{q}_*+1} + \frac{-\tilde{Q}_*-1+\tilde{m}_*^2}{(\tilde{Q}_*+\tilde{q}_*+1)^2}, \tag{4.7}$$

$$m_* = \frac{\tilde{m}_*}{\tilde{Q}_*+\tilde{q}_*+1}. \tag{4.8}$$

B. Analogs of energy and entropy

The analogs of energy and entropy are derived from differentiating Eq. (4.1) with respect to the analogous inverse temperature. In Ref. [25], the differentiation with respect to sample size M was replaced with finite difference due to its discreteness. Alternatively, the differentiation of $f_{\mathrm{B}}(\alpha) = \mathsf{f}_{\mathrm{B}}(Q_*, q_*, m_*, \tilde{Q}_*, \tilde{q}_*, \tilde{m}_*)$ with respect to α can be performed just as it is thanks to continuity of α. We note that Q_*, q_*, m_*, $\tilde{Q}_*$, $\tilde{q}_*$, and $\tilde{m}_*$ which satisfy Eqs. (4.3)–(4.8) are regarded as functions of α. The notation that these are functions of α, e.g., $Q_*(\alpha)$, is omitted for simplicity. We obtain the analogs of energy density and entropy density as follows:

$$\begin{aligned} u_{\mathrm{B}}(\alpha) &= \frac{\partial\{\alpha f_{\mathrm{B}}(\alpha)\}}{\partial\alpha} \\ &= \frac{1}{2}\left\{\ln[2\pi(Q_*-q_*)] + \frac{1-2m_*+q_*}{Q_*-q_*}\right\}, \end{aligned} \tag{4.9}$$

$$\begin{aligned} s_{\mathrm{B}}(\alpha) = \alpha^2\frac{\partial f_{\mathrm{B}}(\alpha)}{\partial\alpha} &= \frac{1}{2}\tilde{Q}_*Q_* + \frac{1}{2}\tilde{q}_*q_* - \tilde{m}_*m_* \\ &\quad - \frac{1}{2}\ln(\tilde{Q}_*+\tilde{q}_*+1) + \frac{\tilde{q}_*+\tilde{m}_*^2}{2(\tilde{Q}_*+\tilde{q}_*+1)}. \end{aligned} \tag{4.10}$$

We refer to what Eqs. (3.4), (4.1), and (4.9)–(4.10) mean in terms of Bayesian statistics. The logarithm of evidence in Eq. (3.4) is known as a measure for performing model selection and determination of hyperparameters in a prior distribution based on a guideline called evidence maximization [60]. The guideline says that one performs model selection and determination of hyperparameters so that the evidence is maximized. For example, Bayesian information criteria (BIC) [61] which forms the expansion of log-evidence is representative for performing model selection based on the guideline. The meaning of Eqs. (3.4) and (4.1) in Bayesian statistics is a macroscopic kind of measure of evidence maximization, up to the multiplicative factor of $-\alpha^{-1}$ and expectation with respect to Eq. (2.12). The maximization of the evidence is translated into the minimization of analogous free energy f_{B}. Therefore, f_{B} is the object to be minimized in terms of model selection and determination of hyperparameters based on evidence maximization [62]. Apart from this minimization aspect, one needs to minimize the analogous free energy f_{B} in Eqs. (4.1) and (4.2) such that $(Q, q, m, \tilde{Q}, \tilde{q}, \tilde{m})$ do not necessarily satisfy Eqs. (4.3)–(4.8). Its minimization stems from considering the macroscopic system with Eq. (2.14), and more specifically from Laplace's method [63] in calculating Eq. (3.4), which is also used in deriving BIC mentioned above. The minimization in Eq. (4.1) which occurs in such a way shares a similar direction with the concept of evidence maximization in that free energy is minimized.

In terms of Bayes statistics, the analogous energy and entropy calculated using the sample size ↔ inverse temperature correspondence have meant a measure of model performance associated with cross-validation [64] and the logarithmic number of models consistent with data, respectively [25]. The meanings of Eqs. (4.9) and (4.10) in Bayesian statistics could be considered macroscopic varieties of those meanings with respect to Eq. (2.14).

C. Estimation error

The concreteness of Q in Eq. (4.2) is a macrostate associated with the sum of w_i^2 from $i = 1$ to N per

N, i.e., $N^{-1}\sum_{i=1}^{N}\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},X)}[w_i^2]$. By the same token, macrostates of q and m in Eq. (4.2) are associated with $N^{-1}\sum_{i=1}^{N}\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},X)}[w_i]^2$ and $N^{-1}\sum_{i=1}^{N}\mathbb{E}_{p(\boldsymbol{w}|\boldsymbol{y},X)}[v_i w_i]$, respectively. Hence, results obtained by Eqs. (4.3)–(4.8) that minimize Eq. (4.2) are the (analogous) equilibrium values of those macrostates together with Eq. (2.14) and the expectation by Eq. (2.12). This means that Eq. (2.13) can be expressed as

$$\mathrm{MSE}(\alpha) = 1 - 2m_* + q_*, \tag{4.11}$$

where $N^{-1}\mathbb{E}_{r(\boldsymbol{y},X,\boldsymbol{v})}[\sum_{i=1}^{N} v_i^2] = N^{-1}\sum_{i=1}^{N}\mathbb{E}_{r(\boldsymbol{v})}[v_i^2] = 1$ is applied. Note that $\mathbb{E}_{r(\boldsymbol{v})}[v_i^2]$ is unity since it expresses the variance of Eq. (2.10).

V. PROPERTIES OF ANALOGOUS THERMODYNAMIC FUNCTIONS AND ESTIMATION ERROR

In Secs. V A and V B, we investigate properties of analogous thermodynamic functions and inverse temperature, i.e., Eqs. (4.1), (4.9), (4.10), and α. The points to be investigated are equations which hold between them in Sec. V A and analogous temperature dependence of thermodynamic functions in Sec. V B. In Sec. V C, we explore, based on Secs V A and V B, what physical insights can be gained for estimation associated with the Bayesian linear regression model.

A. Equations

We observe equations that hold between the analogs, i.e., f_{B}, u_{B}, s_{B}, and α. These analogs are not thermodynamic functions given to some thermodynamic systems, which are standard objects for analysis in statistical mechanics and thermodynamics. However one can eventually recognize that properties of analogs are consistent with those of genuine thermodynamics functions which are not analogs. These analogs are derived on the basis of the correspondences in Sec. III and statistical mechanical prescriptions for deriving thermodynamic functions in Sec. IV. Thus, the well-known expressions that hold between free energy, energy, entropy, and inverse temperature which are not analogs in thermodynamics also work for their analogs, i.e., f_{B}, u_{B}, s_{B}, and α. For example, we confirm

$$f_{\mathrm{B}}(\alpha) = u_{\mathrm{B}}(\alpha) - \frac{1}{\alpha}s_{\mathrm{B}}(\alpha) \tag{5.1}$$

and

$$\alpha\frac{\partial u_{\mathrm{B}}(\alpha)}{\partial\alpha} = \frac{\partial s_{\mathrm{B}}(\alpha)}{\partial\alpha}. \tag{5.2}$$

Equations (5.1) and (5.2) are reminiscent of the equations, in thermodynamics, connecting free energy, energy, and entropy whose arguments are (inverse) temperature.

Additionally, the expression of inverse temperature with energy and entropy in thermodynamics is reproduced by their analogs α, u_{B}, and s_{B} as follows:

$$\frac{\partial s_{\mathrm{B}}(u_{\mathrm{B}})}{\partial u_{\mathrm{B}}} = \alpha, \tag{5.3}$$

where analogous entropy s_{B} of argument u_{B} instead of α in the left-hand side is obtained with the Legendre-Fenchel transform of $f_{\mathrm{B}}(\alpha)$ with respect to $1/\alpha$, i.e.,

$$s_{\mathrm{B}}(u_{\mathrm{B}}) = \min_{1/\alpha}\left\{\frac{u_{\mathrm{B}} - f_{\mathrm{B}}(\alpha)}{1/\alpha}\right\}. \tag{5.4}$$

We see Eqs. (5.1)–(5.3) from the viewpoint of Bayes statistics. As mentioned in Sec. IV B, f_{B}, u_{B}, and s_{B} could be considered to have Bayesian statistical meanings as the macroscopic measure of evidence maximization, that of model performance, and that of log-number of models consistent with data, respectively. Additionally, the identity of α in Bayes statistics is the convergence of sample size per the number of regression coefficients. We realize that f_{B}, u_{B}, s_{B}, and α which are Bayesian quantities with those meanings are connected through Eqs. (5.1)–(5.3), disguised free energy, energy, entropy, and inverse temperature, respectively.

B. Analogous temperature dependence and estimation error

Continuing from the previous Sec. V A, we examine properties of analogous thermodynamic functions. In this subsection, we investigate how $f_{\mathrm{B}}(\alpha)$, $u_{\mathrm{B}}(\alpha)$, and $s_{\mathrm{B}}(\alpha)$ depend on the analogous temperature, $1/\alpha$. For various values of $1/\alpha$ except $1/\alpha = 1$, those of $\tilde{Q}_*$, $\tilde{q}_*$, $\tilde{m}_*$, Q_*, q_*, and m_* are obtained by iteratively calculating Eqs. (4.3)–(4.8) with initial condition $(Q, q, m) = (1.0, 0.001, 0.001)$. Substituting those values for Eqs. (4.1), (4.9), and (4.10), we acquire the values of $f_{\mathrm{B}}(\alpha)$, $u_{\mathrm{B}}(\alpha)$, and $s_{\mathrm{B}}(\alpha)$ for various values of $1/\alpha$. The results are plotted in Fig. 1.

In Fig. 1(a), f_{B} seems to be concave with respect to $1/\alpha$. We observe that $u_{\mathrm{B}}(\alpha)$ and $s_{\mathrm{B}}(\alpha)$ appear to increase with respect to $1/\alpha$, as shown in Figs. 1(b) and 1(c). These $1/\alpha$ dependencies of $f_{\mathrm{B}}(\alpha)$, $u_{\mathrm{B}}(\alpha)$, and $s_{\mathrm{B}}(\alpha)$ do not contradict those of free energy, energy, and entropy which are not analogs in thermodynamics. Additionally, one encounters the singular behavior of analogous thermodynamic functions in the sense that f_{B} becomes nondifferentiable at $1/\alpha \sim 1$ together with discontinuity in u_{B} and s_{B}.

We additionally obtain the values of $\mathrm{MSE}(\alpha)$ for various values of $1/\alpha$ by substituting q_* and m_* determined above for Eq. (4.11). The result is plotted in Fig. 2. In Fig. 2, we observe MSE ~ 0 in the range of $1/\alpha < 1$ and MSE $\not\sim 0$ in $1/\alpha > 1$. This abrupt change in the behavior of $\mathrm{MSE}(\alpha)$ at $1/\alpha \sim 1$ is related to the statistical meaning of $1/\alpha$. As shown in Eq. (2.14), α is the convergence of sample size M per the number of parameters N, i.e., M/N. In the case of $1/\alpha < 1$, M is (asymptotically) larger than N, which means an easy situation to estimate. The case of $1/\alpha > 1$ represents (asymptotically) $M < N$, which makes the estimation difficult.

C. Physical insights into Bayesian linear regression

In thermodynamic systems, the effect of decreasing energy and that of increasing entropy are competing in achieving an equilibrium state with minimum free energy. We observe that estimation of Bayesian linear regression could also be understood in a similar picture. To explain it, we introduce following quantities:

$$\mathsf{u}_{\mathrm{B}}(Q, q, m) = \frac{1}{2}\left\{\ln[2\pi(Q - q)] + \frac{1 - 2m + q}{Q - q}\right\} \tag{5.5}$$

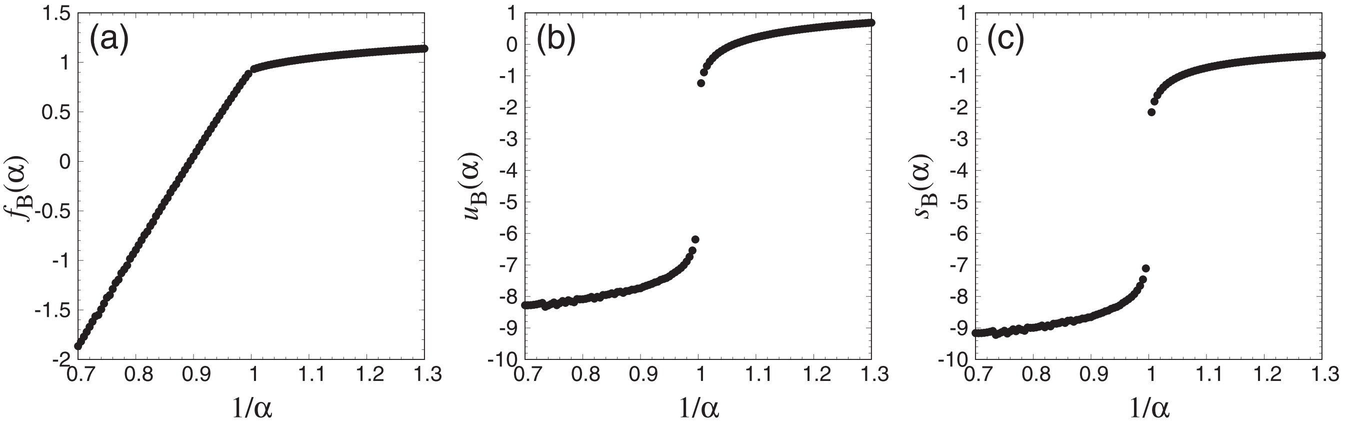

FIG. 1. Analogs of thermodynamic functions, Eqs. (4.1), (4.9), and (4.10), are plotted versus the analogous temperature, $1/\alpha$. (a) $f_{\rm B}(\alpha)$-$1/\alpha$. (b) $u_{\rm B}(\alpha)$-$1/\alpha$. (c) $s_{\rm B}(\alpha)$-$1/\alpha$.

and

$$\begin{aligned}\mathsf{s}_{\rm B}(Q,q,m,\tilde{Q},\tilde{q},\tilde{m}) = {}& \frac{1}{2}\tilde{Q}Q+\frac{1}{2}\tilde{q}q-\tilde{m}m\\ &-\frac{1}{2}\ln(\tilde{Q}+\tilde{q}+1)\\ &+\frac{\tilde{q}+\tilde{m}^2}{2(\tilde{Q}+\tilde{q}+1)},\end{aligned} \tag{5.6}$$

in the sense that these represent the analogous energy and entropy for which $(Q,q,m,\tilde{Q},\tilde{q},\tilde{m})$ do not necessarily satisfy Eqs. (4.3)–(4.8), respectively. Such an introduction comes from the forms of right-hand sides in Eqs. (4.9) and (4.10). Regarding the nonminimum analogous free energy, $\mathsf{f}_{\rm B}=\mathsf{u}_{\rm B}-\alpha^{-1}\mathsf{s}_{\rm B}$ holds using Eqs. (4.2), (5.5), and (5.6). Thus Eq. (4.1) with the sense of minimization of $\mathsf{f}_{\rm B}$ is also expressed as follows:

$$f_{\rm B}(\alpha)=\min_{\mathcal{A}}\{\mathsf{u}_{\rm B}(Q,q,m)\}-\frac{1}{\alpha}\max_{\mathcal{A}}\{\mathsf{s}_{\rm B}(Q,q,m,\tilde{Q},\tilde{q},\tilde{m})\}, \tag{5.7}$$

where $\mathcal{A}=(Q,q,m,\tilde{Q},\tilde{q},\tilde{m})$. Equations (5.1) and (5.7) show that $\mathsf{u}_{\rm B}$ decreases, while $\mathsf{s}_{\rm B}$ increases to achieve Bayes inference with minimum $\mathsf{f}_{\rm B}$. According to Figs. 1(b) and 1(c), both $u_{\rm B}$ and $s_{\rm B}$ simultaneously take lower values in $1/\alpha<1$ and higher values in $1/\alpha>1$, in the images of them, respectively. In $1/\alpha>1$, it seems that $u_{\rm B}$ cannot be lower in its image while $s_{\rm B}$ is higher in its image. In $1/\alpha<1$, $s_{\rm B}$ cannot be higher while $u_{\rm B}$ is lower. These properties suggest that the effect of decreasing $\mathsf{u}_{\rm B}$ and that of increasing $\mathsf{s}_{\rm B}$ are competing in achieving Bayes inference so that $\mathsf{f}_{\rm B}$ is minimized. In other words, both effects of $\min\{\mathsf{u}_{\rm B}\}$ and $\max\{\mathsf{s}_{\rm B}\}$ in Eq. (5.7) are balanced. Using this picture, Figs. 1(b) and 1(c) provide not only $1/\alpha$ dependence of $u_{\rm B}$ and $s_{\rm B}$ but also an aspect of balance between them as follows; the effect of decreasing $\mathsf{u}_{\rm B}$ is dominant in the range of $1/\alpha<1$, while that of increasing $\mathsf{s}_{\rm B}$ is dominant in $1/\alpha>1$. These two balanced states seem to switch to each other at $1/\alpha\sim 1$.

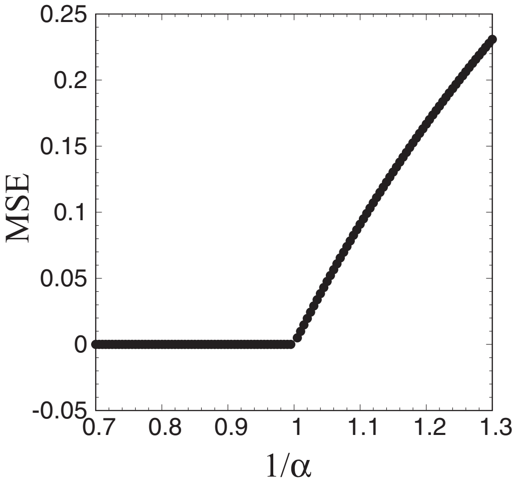

FIG. 2. Mean squared error (MSE), Eqs. (2.13) and (4.11), is plotted versus the analog of temperature $1/\alpha$. MSE is one of the measures that represent whether the regression task is successful or not. The case of MSE ~ 0 in $1/\alpha<1$ means success of the regression task, while that of MSE $\nsim 0$ in $1/\alpha>1$ means failure as the value of it is away from zero.

Moreover, the two balanced states seem to be associated with whether or not the task of Bayesian linear regression is successful. To understand this, we call attention to the macroscopic behavior of mean squared error (MSE) shown in Fig. 2. We recall that the case of MSE ~ 0 means success of the regression task, while MSE $\nsim 0$ indicates failure of the task as the value of it is away from zero.

Here Fig. 2 and the above description in the first paragraph of this Sec. V C lead us to see the following physical view: On the one hand, Bayes inference that succeeds in the regression task (MSE ~ 0) is understood to be a low-temperature analogous state in $1/\alpha<1$ dominated by the decreasing effect of $\mathsf{u}_{\rm B}$. We interpret $\boldsymbol{w}$ to take configurations close to $\boldsymbol{v}$, considering that MSE shown in Eq. (2.13) measures the square of radius of a hypersphere centered on $\boldsymbol{v}$. On the other hand, the inference that fails in the task (MSE $\nsim 0$) is understood as a high-temperature analogous state in $1/\alpha>1$ favoring the effect of increasing $\mathsf{s}_{\rm B}$. It is interpreted that $\boldsymbol{w}$ scatteringly takes configurations not necessarily close to $\boldsymbol{v}$. The Bayes inference at the boundary point in both cases, $1/\alpha\sim 1$, is singular, as mentioned in Sec. V B. In other words, the present model experiences the singularity when both the Bayes inferences are switched to each other with change in analogous temperature $1/\alpha$.

D. Discussion on nature of phase transition

We refer to the singular behavior of analogous thermodynamic functions at $1/\alpha \sim 1$ in Fig. 1, i.e., nondifferentiability of $f_{\rm B}$ and discontinuity in $u_{\rm B}$ and $s_{\rm B}$. In the context of signal reconstruction which is performed with an application of linear regression models, behavior similar to the first-order phase transition was reported under the condition called Bayes optimal [58]. Its behavior was observed from the logarithm of evidence corresponding to that of analogous partition function, without using analogous inverse temperature. The model treated in this paper falls under that condition. It is not clear whether such a phase-transition-like phenomenon can be treated in the same way as phase transitions in genuine thermodynamic systems. However, the singular behavior in Fig. 1 is in agreement with the report in Ref. [58], within the Ehrenfest classification of phase transitions, in the sense that the discontinuity has appeared in $u_{\rm B}$ and $s_{\rm B}$ related to the first derivative of $f_{\rm B}$. We note that the Ehrenfest scheme classifies a singularity in which the nth derivative of free energy exhibits discontinuity as an nth-order phase transition [65,66].

A way to further discuss the nature of this phase transition is to enter into the equivalence and nonequivalence of ensembles [67,68]. If there exists an analog of the microcanonical ensemble whose thermodynamic potential is $s_{\rm B}(u_{\rm B})$, the nature of that phase transition would be explained as follows. The singular behavior in Fig. 1(a) means that the phase coexistence occurs as the first-order phase transition with analogous latent heat $\Delta u_{\rm B} = \lim_{\alpha\uparrow 1} u_{\rm B}(\alpha) - \lim_{\alpha\downarrow 1} u_{\rm B}(\alpha)$. This phenomenon can arise in two cases: either $s_{\rm B}(u_{\rm B})$ is affine, meaning linear in $\Delta u_{\rm B}$, or $s_{\rm B}(u_{\rm B})$ is nonconcave in $\Delta u_{\rm B}$. In the first case where $s_{\rm B}(u_{\rm B})$ is concave with the linear part (nonstrictly concave), Eq. (2.5) and the analogous microcanonical ensemble are equivalent, since their thermodynamic potentials, namely $f_{\rm B}(\alpha)$ and $s_{\rm B}(u_{\rm B})$, are related through the Legendre-Fenchel transform as in Eq. (5.4). In the second case, the two ensembles are not equivalent since the nonconcave part of $s_{\rm B}(u_{\rm B})$ cannot be obtained from the Legendre-Fenchel transform, i.e., $u_{\rm B}$ that makes $s_{\rm B}(u_{\rm B})$ nonconcave does not exist in Eq. (2.5). The nature of phase transition encountered in Fig. 1 would fall into one of these two cases. However, it is not possible to tell which one is from the canonical angle of Eq. (2.5) and $f_{\rm B}(\alpha)$ alone. To solve this problem, we need to analyze $s_{\rm B}(u_{\rm B})$ from the microcanonical angle, which leads us to further research exploring the analogous microcanonical ensemble in response to Eq. (2.5): canonical correspondence.

VI. SUMMARY AND DISCUSSION

The subject of this paper concerns the similarities between statistical physics and Bayes inference through the Bayesian linear regression model. We have concentrated on the proposal to associate discrete sample size with inverse temperature [25,34] and on the derivation of analogous thermodynamic functions using its correspondence.

This paper suggests incorporating the macroscopic perspective as analogous thermodynamic limit into the previous suggestion. Its motivation comes from the statistical mechanical concept of calculating thermodynamic functions which characterize macroscopic properties. We have adopted the asymptotic limit in statistics, called by various names shown in Table I, where the sample size and the number of parameters both increase at once. In this asymptotics, the convergence of the ratio of sample size and the number of parameters becomes a variable to control asymptotic behavior [39]. This paper indicates that the converged quantity appears to be a candidate for the analog of inverse temperature with continuity. Combining this analogous inverse temperature with the other similarities we calculated the analogs of free energy, energy, and entropy for the Bayesian linear regression model, and those properties were investigated. On the basis of those properties, we considered the physical view of Bayes inference.

As a result, Bayes estimation of regression coefficients could be observed as the statistical mechanical or thermodynamic picture of an equilibrium state, i.e., the balance between the effect of decreasing energy and that of increasing entropy. The low-temperature analogous state dominated by the former effect was shown to be associated with the success in estimation, while the failure in estimation is understood as the high-temperature analogous state favoring the latter effect.

We discuss the behavior of analogous thermodynamic functions with respect to analogous temperature shown in Fig. 1. The behavior of $s_{\rm B}(\alpha) < 0$ seems to be, at first glance, contradictory in the context of statistical mechanics and thermodynamics. There are two possible explanations for this behavior. The previous study mentioned that the analogous entropy derived from sample size ↔ inverse temperature correspondence can have negative values [25]. Although α in Eq. (2.14) is used in place of sample size, we may have found an example of that, shown in Fig. 1(c). Another possibility is to require an additive constant to be the interpretation of $s_{\rm B}(\alpha) < 0$. Thermodynamic entropy has the property of being invariant up to an additive constant. If this is also the case for analogous $s_{\rm B}$, we could select an additive constant such that $s_{\rm B}(\alpha) > 0$. Other properties shown in Fig. 1, such as the concavity of $f_{\rm B}$ and the increase of $u_{\rm B}$ and $s_{\rm B}$ with respect to $1/\alpha$, do not appear to contradict temperature dependence of genuine thermodynamic functions which are not analogs.

The properties such as temperature dependence of analogous thermodynamic functions were found to be nearly consistent with those of genuine thermodynamics functions. The singularity in the analogous ones is consistent with the report in Ref. [58], at least within the Ehrenfest scheme. These points may support, in physical perspective, the proposal by Refs. [25,34] to associate sample size with inverse temperature, though we have employed α instead of sample size.

The analysis in this paper was performed through the specific Bayesian model as linear regression. This model is the basis of various Bayesian machine learning models such as classification, the kernel method, and neural networks [69]. However, advantages of Bayesian estimation might not be fully demonstrated if we remain within the linear regression and its estimation task. Apart from Bayesian analysis, there are even more convenient approaches to linear regression such as the weighted least squares method which assumes that the noise random variables do not have the same variance. Therefore, one of the future directions is to apply the analysis to applied models based on linear regression where advantages of Bayesian estimation can be fully utilized.

In this study, $\varsigma, \sigma \sim 0$ were assumed, which may implicitly cause unexpected behavior of sample size, e.g., high sample size. If this is true, the results for $1/\alpha > 1$ would be unreliable. One way to approach this problem is to explore the case of $\varsigma, \sigma \neq 0$. This future direction of research is also supported in the sense that $\varsigma, \sigma \neq 0$ is a practical situation.

We hope that this work will enrich the interdisciplinary science between physics, machine learning, and statistical inference.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grants No. JP21J22128 and No. JP22KJ1358 (Grant-in-Aid for Research Fellows of Japan Society for the Promotion of Science). The author is grateful to Yukiyasu Ozeki, Shuichi Kawano, and Jun Suzuki for their support.

APPENDIX: ANOTHER ANALOG OF INVERSE TEMPERATURE IN THE BAYES MODEL

Apart from the sample size ↔ inverse temperature correspondence, another analog of inverse temperature was proposed previously. It is associated with an external parameter to extend the likelihood function in the Bayes model. Although another analog using an external parameter is beyond the scope of this study, we describe details of it and a difference between these analogs. Specifically, the likelihood function is extended to it to the power of an external parameter β as follows:

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})^{\beta} p(\boldsymbol{w}), \tag{A1}$$

$$Z_{\mathrm{B}} = \int d\boldsymbol{w}\, p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})^{\beta} p(\boldsymbol{w}). \tag{A2}$$

Several references proposed to regard this β as an analogous inverse temperature [30,31,70–78]. In the extended form of Eqs. (A1) and (A2), the original Bayes inference is obtained only for the specific case of $\beta = 1$. This extended form appears practical and efficient for sampling from the posterior distribution and computing marginal likelihood [30,31,70–72,75,76,78,79].

Reference [25] acknowledged the practicality of the extended form for sampling. However, beyond its practical aspect, the previous study had some concerns about regarding β as inverse temperature, from the view of what the role of inverse temperature is in Bayes inference, as follows. The parameter β is not a preexisting quantity inherent in the Bayesian model. The case of $\beta \neq 1$ no longer corresponds to Bayes inference. Reference [80] also expressed a concern similar to that of Ref. [25] about the Bayes model extended with external parameters, which was introduced in signal inference problems in information field theory [72]. In that extended model, in addition to the above β extension, an extension using another external variable termed a moment generating source J was also added. Reference [80] commented that the extended Bayes form in Ref. [72] violates Bayes's theorem in the case of $(\beta, J) \neq (1, 0)$, while Ref. [81] explained that this concern in Ref. [80] does not affect the methodology developed in Ref. [72].

In contrast to recognizing the external parameter β as inverse temperature, the correspondence between the sample size and inverse temperature uses a quantity which originally appears in the Bayesian model, namely sample size M. In Sec. III, we have suggested that the convergence of sample size M per the number of parameters N is a candidate for analogous inverse temperature with continuity. This converged quantity consists of M and N, both of which are inherent in the Bayesian linear regression model. In this sense, our suggestion shares the style of Refs. [25,34] which observes the similarity using a quantity inherent in the Bayesian model, rather than the style of considering an external parameter as analogous inverse temperature.

[1] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[2] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. **810**, 1 (2019).

[3] L. Wang, Discovering phase transitions with unsupervised learning, Phys. Rev. B **94**, 195105 (2016).

[4] E. P. Van Nieuwenburg, Y.-H. Liu, and S. D. Huber, Learning phase transitions by confusion, Nat. Phys. **13**, 435 (2017).

[5] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).

[6] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada, Restricted Boltzmann machine learning for solving strongly correlated quantum systems, Phys. Rev. B **96**, 205152 (2017).

[7] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).

[8] G. Carleo, Y. Nomura, and M. Imada, Constructing exact representations of quantum many-body systems with deep neural networks, Nat. Commun. **9**, 5322 (2018).

[9] S. Lorenz, A. Groß, and M. Scheffler, Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks, Chem. Phys. Lett. **395**, 210 (2004).

[10] J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. **98**, 146401 (2007).

[11] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, Lecture Notes in Physics Vol. 9 (World Scientific, Singapore, 1987).

[12] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An introduction* (Clarendon Press, Oxford, 2001).

[13] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).

[14] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982).

[15] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Storing infinite numbers of patterns in a spin-glass model of neural networks, Phys. Rev. Lett. **55**, 1530 (1985).

[16] E. Gardner, The space of interactions in neural network models, J. Phys. A: Math. Gen. **21**, 257 (1988).

[17] E. Gardner and B. Derrida, Optimal storage properties of neural network models, J. Phys. A: Math. Gen. **21**, 271 (1988).

[18] N. Sourlas, Spin-glass models as error-correcting codes, Nature (London) **339**, 693 (1989).

[19] Y. Kabashima and D. Saad, Statistical mechanics of error-correcting codes, Europhys. Lett. **45**, 97 (1999).

[20] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (1957).

[21] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Principles of maximum entropy and maximum caliber in statistical physics, Rev. Mod. Phys. **85**, 1115 (2013).

[22] E. S. Soofi, Principal information theoretic approaches, J. Am. Stat. Assoc. **95**, 1349 (2000).

[23] F. A. Palmieri and D. Ciuonzo, Objective priors from maximum entropy in data classification, Inf. Fusion **14**, 186 (2013).

[24] A. Caticha and R. Preuss, Maximum entropy and Bayesian data analysis: Entropic prior distributions, Phys. Rev. E **70**, 046127 (2004).

[25] C. H. LaMont and P. A. Wiggins, Correspondence between thermodynamics and inference, Phys. Rev. E **99**, 052140 (2019).

[26] A. Ratnaparkhi, A maximum entropy model for part-of-speech tagging, in Conference on empirical methods in natural language processing, 1996, https://aclanthology.org/W96-0213.

[27] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, A maximum entropy approach to natural language processing, Comput. Linguist. **22**, 39 (1996).

[28] G. R. Terrell, The maximal smoothing principle in density estimation, J. Am. Stat. Assoc. **85**, 470 (1990).

[29] W. Bialek, C. G. Callan, and S. P. Strong, Field theories for learning probability distributions, Phys. Rev. Lett. **77**, 4693 (1996).

[30] J. C. Lemm, *Bayesian Field Theory* (JHU Press, Baltimore, 2003).

[31] T. A. Enßlin, M. Frommert, and F. S. Kitaura, Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis, Phys. Rev. D **80**, 105005 (2009).

[32] J. B. Kinney, Unification of field theory and maximum entropy methods for learning probability densities, Phys. Rev. E **92**, 032107 (2015).

[33] W.-C. Chen, A. Tareen, and J. B. Kinney, Density estimation on small data sets, Phys. Rev. Lett. **121**, 160605 (2018).

[34] V. Balasubramanian, Statistical inference, occam's razor, and statistical mechanics on the space of probability distributions, Neural Comput. **9**, 349 (1997).

[35] O. H. Schnaack and A. Nourmohammad, Optimal evolutionary decision-making to store immune memory, eLife **10**, e61346 (2021).

[36] O. H. Schnaack, L. Peliti, and A. Nourmohammad, Learning and organization of memory for evolving patterns, Phys. Rev. X **12**, 021063 (2022).

[37] M. Ramezanali, P. P. Mitra, and A. M. Sengupta, Critical behavior and universality classes for an algorithmic phase transition in sparse reconstruction, J. Stat. Phys. **175**, 764 (2019).

[38] J. W. Rocks and P. Mehta, Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models, Phys. Rev. Res. **4**, 013201 (2022).

[39] V. I. Serdobolskii, *Multivariate Statistical Analysis: A High-Dimensional Approach*, Theory and Decision Library B Vol. 41 (Springer, Dordrecht, 2000).

[40] O. Ledoit and M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivar. Anal. **88**, 365 (2004).

[41] B. G. Lindsay, J. Kettenring, and D. O. Siegmund, A report on the future of statistics, Stat. Sci. **19**, 387 (2004).

[42] I. M. Johnstone and D. M. Titterington, Statistical challenges of high-dimensional data, Philos. Trans. R. Soc. A **367**, 4237 (2009).

[43] Z. Bai and J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, Springer Series in Statistics Vol. 20 (Springer, New York, 2010).

[44] Š. Raudys and D. M. Young, Results in statistical discriminant analysis: A review of the former Soviet Union literature, J. Multivar. Anal. **89**, 1 (2004).

[45] A. Zollanvari, U. M. Braga-Neto, and E. R. Dougherty, Analytic study of performance of error estimators for linear discriminant analysis, IEEE Trans. Signal Process. **59**, 4238 (2011).

[46] In the general asymptotics, the ratio of M and N is bound by some constant [40] rather than converging like Eq. (2.14). Reference [40] describes that such general asymptotics is related to Kolmogorov asymptotics.

[47] N. El Karoui, Spectrum estimation for large dimensional covariance matrices using random matrix theory, Ann. Stat. **36**, 2757 (2008).

[48] N. El Karoui, Operator norm consistent estimation of large-dimensional sparse covariance matrices, Ann. Stat. **36**, 2717 (2008).

[49] A. Nishimura and M. A. Suchard, Prior-preconditioned conjugate gradient method for accelerated Gibbs sampling in "large n, large p" Bayesian sparse regression, J. Am. Stat. Assoc. **118**, 2468 (2023).

[50] P. J. Huber, Robust regression: Asymptotics, conjectures and Monte Carlo, Ann. Stat. **1**, 799 (1973).

[51] J. W. Silverstein, Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices, J. Multivar. Anal. **55**, 331 (1995).

[52] G. Pan, J. Gao, and Y. Yang, Testing independence among a large number of high-dimensional random vectors, J. Am. Stat. Assoc. **109**, 600 (2014).

[53] Z. D. Bai and Y. Q. Yin, Limit of the smallest eigenvalue of a large dimensional sample covariance matrix, Ann. Probab. **21**, 1275 (1993).

[54] Y.-Q. Yin, Z.-D. Bai, and P. R. Krishnaiah, On the limit of the largest eigenvalue of the large dimensional sample covariance matrix, Probab. Theory Relat. Fields **78**, 509 (1988).

[55] S. Geman, A limit theorem for the norm of random matrices, Ann. Probab. **8**, 252 (1980).

[56] Z. D. Bai and J. W. Silverstein, CLT for linear spectral statistics of large-dimensional sample covariance matrices, Ann. Probab. **32**, 553 (2004).

[57] Y. Kabashima, T. Wadayama, and T. Tanaka, A typical reconstruction limit for compressed sensing based on L_p-norm minimization, J. Stat. Mech. (2009) L09003.

[58] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices, J. Stat. Mech. (2012) P08009.

[59] It could be possible to replace α included in the denominator in the right-hand side with $\alpha(M, N)$.

[60] D. J. MacKay, Bayesian interpolation, Neural Comput. **4**, 415 (1992).

[61] G. Schwarz, Estimating the dimension of a model, Ann. Statist. **6**, 461 (1978).

[62] We do not set a situation where the model selection and determination of hyperparameters are explicitly necessary for the present Bayesian linear regression model, because they are beyond the purpose and scope of this paper. It does not affect results and discussions within this study.

[63] L. Tierney and J. B. Kadane, Accurate approximations for posterior moments and marginal densities, J. Am. Stat. Assoc. **81**, 82 (1986).

[64] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Series B (Methodol.) **36**, 111 (1974).

[65] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. Newman, *The Theory of Critical Phenomena: An Introduction to the Renormalization Group* (Oxford University Press, Oxford, 1992).

[66] G. Jaeger, The Ehrenfest classification of phase transitions: Introduction and evolution, Arch. Hist. Exact Sci. **53**, 51 (1998).

[67] H. Touchette, R. S. Ellis, and B. Turkington, An introduction to the thermodynamic and macrostate levels of nonequivalent ensembles, Physica A **340**, 138 (2004).

[68] H. Touchette, The large deviation approach to statistical mechanics, Phys. Rep. **478**, 1 (2009).

[69] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, Information Science and Statistics Vol. 4 (Springer, New York, 2006).

[70] S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, Bayesian ensemble approach to error estimation of interatomic potentials, Phys. Rev. Lett. **93**, 165501 (2004).

[71] N. Friel and A. N. Pettitt, Marginal likelihood estimation via power posteriors, J. R. Stat. Soc. Series. B **70**, 589 (2008).

[72] T. A. Enßlin and C. Weig, Inference with minimal Gibbs free energy in information field theory, Phys. Rev. E **82**, 051112 (2010).

[73] S. Watanabe, Asymptotic learning curve and renormalizable condition in statistical learning theory, J. Phys.: Conf. Ser. **233**, 012014 (2010).

[74] S. Watanabe, Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, J. Mach. Learn. Res. **11**, 3571 (2010).

[75] A. Allahverdyan, Observational nonidentifiability, generalized likelihood and free energy, Int. J. Approx. Reason. **125**, 118 (2020).

[76] Y. Kurniawan, C. L. Petrie, K. J. Williams, M. K. Transtrum, E. B. Tadmor, R. S. Elliott, D. S. Karls, and M. Wen, Bayesian, frequentist, and information geometric approaches to parametric uncertainty quantification of classical empirical interatomic potentials, J. Chem. Phys. **156**, 214103 (2022).

[77] S. Tokuda, K. Nagata, and M. Okada, Intrinsic regularization effect in Bayesian nonlinear regression scaled by observed data, Phys. Rev. Res. **4**, 043165 (2022).

[78] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, Marginal likelihood computation for model selection and hypothesis testing: An extensive review, SIAM Rev. **65**, 3 (2023).

[79] D. Carlson, P. Stinson, A. Pakman, and L. Paninski, Partition functions from Rao-Blackwellized tempered sampling, in *Proceedings of the 33rd International Conference on Machine Learning*, edited by M. F. Balcan and K. Q. Weinberger, Proceedings of Machine Learning Research Vol. 48 (PMLR, New York, 2016), pp. 2896–2905.

[80] D. Iatsenko, A. Stefanovska, and P. V. E. McClintock, Comment on “Inference with minimal Gibbs free energy in information field theory”, Phys. Rev. E **85**, 033101 (2012).

[81] T. A. Enßlin and C. Weig, Reply to “Comment on ‘Inference with minimal Gibbs free energy in information field theory’ ”, Phys. Rev. E **85**, 033102 (2012).