The Nishimori line and Bayesian statistics

1999 J. Phys. A: Math. Gen. 32 3875

(http://iopscience.iop.org/0305-4470/32/21/302)

J. Phys. A: Math. Gen. **32** (1999) 3875–3888. Printed in the UK PII: S0305-4470(99)97628-X

The Nishimori line and Bayesian statistics

Yukito Iba†

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu Minato-ku, Tokyo 106-8569, Japan

Received 15 September 1998, in final form 25 January 1999

Abstract. The 'Nishimori line' is a line or hypersurface in the parameter space of systems with quenched disorder, where simple expressions of the averages of physical quantities over the quenched random variables are obtained. It has been playing an important role in the theoretical studies of the random frustrated systems since its discovery in around 1980. In this paper, an interpretation of the Nishimori line from the viewpoint of statistical information processing is developed. Our main aim is the reconstruction of the whole theory of the Nishimori line from the viewpoint of Bayesian statistics, or, almost equivalently, from the viewpoint of the theory of error-correcting codes. As a byproduct of the interpretation, counterparts of the Nishimori line in models without gauge invariance are given. We also discussed the issues on the 'finite-temperature decoding' of error-correcting codes and clarify the role of gauge invariance in this topic.

1. Introduction

There are not many rigorous results that are useful for the study of random frustrated systems. Among them, theorems related to the Nishimori line of random spin models form an important family. There have been many papers [1–5] about the Nishimori line after the seminal paper [1] of Nishimori. There is, however, still a mystery about the Nishimori line, i.e., its physical meaning and the motivation behind the proof are not yet clear.

The purpose of this paper is to develop an interpretation on the Nishimori line from the viewpoint of statistical information processing, more specifically from Bayesian statistics or from the coding theory. This interpretation has two advantages. First, it gives an interesting example of an unexpected relation between two different areas, rigorous arguments in the statistical physics and Bayesian statistics, and elucidates the meaning of the trick in the derivation of the Nishimori line. Secondly, it gives some new results on the analogue of the Nishimori line without *gauge invariance* in the sense of Toulouse [6].

Our arguments are closely related to the works on the 'the optimality of finite-temperature decoding' of error-correcting codes [7–10] . These works, however, mostly focused on the decoding of error-correcting codes itself. Our aim here is to develop the idea suggested in these studies and discuss the whole theory of the Nishimori line from the viewpoint of statistical inference. We will also give a comprehensive treatment on the finite-temperature decoding in the latter part of the paper.

In this paper, we make efforts to give a self-contained description of this material. No special knowledge on Bayesian statistics, error-correcting codes and gauge invariance of spin glass is assumed.

† E-mail address: iba@ism.ac.jp

0305-4470/99/213875+14$19.50

2. Bayesian framework

In this section, we give basic notions and terminology of Bayesian statistics. We also discuss identities and inequalities that naturally arise from the Bayesian framework. Although the motivation for these formulae as well as their proofs are quite simple, they are essential in the derivation of the properties of the Nishimori line.

Let us assume that the data y is generated by a probability distribution $p(y|x)$, which is parametrized by the value of an unknown variable x. In the Bayesian framework, we also assume that the parameter x is, in itself, a random sample from a *prior distribution* $\pi(x)$. With these assumptions, the probability distribution of the parameter x conditioned on given data y is

$$p(x|y) = \frac{p(y|x)\pi(x)}{\sum_x p(y|x)\pi(x)}. \tag{1}$$

Here $\sum_x$ means the summation or integral over the possible values of x. This distribution, the *posterior distribution*, is the source of knowledge with given data y in the Bayesian formalism.

A similar formalism is also used in a seemingly different branch of information science: the theory of error-correcting codes. Consider a noisy channel and a set of messages. We encode and send a message x through the noisy channel and someone at the other end of the channel tries to infer the original message x from the output y. If we assume that the probability $p(y|x)$ of an output y with the input x and the distribution $\pi(x)$ of the average frequencies of input messages, the conditional probability $p(x|y)$ of an input x with the output y is given by (1). Note that the probability $p(y|x)$ represents the coding scheme as well as the noise of the channel in this formalism.

We will introduce notations that indicate the averages over different types of distributions. Here the symbol $A(x)$ denotes a function of the parameter x and $B(y)$ denotes a function of the data y . First we define the average over the prior distribution of x,

$$[A(x)]_{\pi(x)} = \sum_x A(x)\pi(x). \tag{2}$$

We also define the average over the posterior distribution of x,

$$\langle A(x)\rangle_{p(x|y)} = \sum_x A(x)p(x|y). \tag{3}$$

Finally we define the average over the probability distribution $p(y|x)$ of data y with the given parameter x,

$$[B(y)]_{p(y|x)} = \sum_y B(y)p(y|x). \tag{4}$$

These notations are not common in the literatures on Bayesian statistics. They are introduced to contrast the analogy to the statistical physics of systems with quenched disorder. Later we show that the averages [ ] correspond to the quenched average over the configuration of the impurities and $\langle\ \rangle$ corresponds to the thermal average.

Let us consider relations among these averages. First we note that an identity

$$[\langle A(x')\rangle_{p(x'|y)}]_{p(y|x)}]_{\pi(x)} = [A(x)]_{\pi(x)} \tag{5}$$

holds. The posterior average $\langle A(x')\rangle_{p(x'|y)}$ can be regarded as an estimate of $A(x)$ from the data y. It is a random variable dependent on y and the identity (5) shows that the average of it over the possible values of the data y and the original parameter x coincides with the prior average $[A(x)]_{\pi(x)}$. The proof of the formula (5) is straightforward. When we substitute the left-hand side of (5) for the definition of the averages (3), (4), (2), we obtain the expression,

$$[\langle A(x')\rangle_{p(x'|y)}]_{p(y|x)}]_{\pi(x)} = \sum_x \sum_y \frac{\sum_{x'} A(x')p(y|x')\pi(x')}{\sum_{x'} p(y|x')\pi(x')} \cdot p(y|x)\pi(x). \tag{6}$$

By changing the order of the summation and a dummy index, we can show that the factors $\sum_x p(y|x)\pi(x)$ in the numerator and denominator cancel each other. Using $\sum_y p(y|x') = 1$ and $\sum_{x'} A(x')\pi(x') = [A(x)]_{\pi(x)}$, the proof of (5) is completed.

It is easy to generalize (5) to an identity

$$[\langle C(x', y)\rangle_{p(x'|y)}]_{p(y|x)}]_{\pi(x)} = [[C(x, y)]_{p(y|x)}]_{\pi(x)}. \tag{7}$$

Here $C(x, y)$ is a function of the data (the output of the channel) y as well as the parameter x. The proof of the relation (7) is essentially the same as that of (5). The only difference from (5) is that the average $[\ ]_{p(y|x)}$ in the right-hand side cannot be removed.

In these arguments, we assume that the 'true' distributions $p(y|x)$ and $\pi(x)$ behind given data are exactly known. They are, however, often unknown in a real world example. A way to fill this gap is to include 'hyperparameters' α and γ in the expression of $p(y|x)$ and $\pi(x)$ and estimate them from the data. Hereafter we use the notation $p_\alpha(y|x)$ and $\pi_\gamma(x)$ to indicate the distributions that contain hyperparameters. An approach to estimate hyperparameters α and γ from the data y is the minimization of a free-energy-like quantity,

$$F(\alpha, \gamma) = -\log \sum_x p_\alpha(y|x)\pi_\gamma(x). \tag{8}$$

Note that the procedure based on the *marginal likelihood* $\sum_x p_\alpha(y|x)\pi_\gamma(x)$ is successfully used by many authors in practical problems [11–18]. It is known by many different terms, e.g. the maximization of *type II likelihood* [11, 12], the minimization of *ABIC* [13], the maximization of *evidence* [14, 15, 25, 26, 43], and, simply, the maximization of the likelihood of α and γ [16, 18]†.

At the moment, we assume that the form of the distribution $\pi_\gamma(x)$ and $p_\alpha(y|x)$ is correctly known except the values of the hyperparameters. Even in this case, the hyperparameters (α, γ) that minimize (8) are random variables dependent on the data y and they fluctuate around the true values (α_0, γ_0) of (α, γ). On the other hand, if we consider the average $[[F(\alpha, \gamma)]_{p_{\alpha_0}(y|x)}]_{\pi_{\gamma_0}(x)}$ of $F(\alpha, \gamma)$ over the true distributions $p_{\alpha_0}(y|x)$ and $\pi_{\gamma_0}(x)$, a set of (α, γ) that minimize the average coincides with the true value (α_0, γ_0). That is, the inequality

$$[[F(\alpha_0, \gamma_0)]_{p_{\alpha_0}(y|x)}]_{\pi_{\gamma_0}(x)} \leqslant [[F(\alpha, \gamma)]_{p_{\alpha_0}(y|x)}]_{\pi_{\gamma_0}(x)} \tag{9}$$

holds for any value of α and γ.

If the right-hand side of (9) is a sufficiently smooth function of (α, γ), the derivatives of F at $(\alpha, \gamma) = (\alpha_0, \gamma_0)$ should be zero. For example, the following relations are direct consequences of (9):

$$\left[\left[\frac{\partial}{\partial\alpha}F(\alpha, \gamma)\right]_{p_{\alpha_0}(y|x)}\right]_{\pi_{\gamma_0}(x)}\Bigg|_{(\alpha,\gamma)=(\alpha_0,\gamma_0)} = 0 \tag{10}$$

$$\left[\left[\frac{\partial}{\partial\gamma}F(\alpha, \gamma)\right]_{p_{\alpha_0}(y|x)}\right]_{\pi_{\gamma_0}(x)}\Bigg|_{(\alpha,\gamma)=(\alpha_0,\gamma_0)} = 0. \tag{11}$$

Here, the derivatives $\partial F/\partial\alpha$ and $\partial F/\partial\gamma$ should be interpreted as the derivatives $\partial F/\partial\alpha_k$ and $\partial F/\partial\gamma_k$ with each component of $\alpha = \{\alpha_k\}$ and $\gamma = \{\gamma_k\}$, when α and γ are vectors with more

† When the expression of the probability $p(y|x)$ is considered as a function of x with given data y, it is called 'likelihood of the parameter x'. This terminology is preferred by the non-Bayesians who do not treat the parameter x as a random variable, but is also used by Bayesians. The mixture distribution $\sum_x p_\alpha(y|x)\pi_\gamma(x)$ at the right-hand side of (8) can be regarded as the likelihood of the hyperparameters α and γ.

than one components. The conditions on the second derivatives are also derived from (9) by using positive semi-definiteness of the Hessian, say,

$$\left[\left[\frac{\partial^2}{\partial^2\alpha}F(\alpha,\gamma)\right]_{p_{\alpha_0}(y|x)}\right]_{\pi_{\gamma_0}(x)}\Bigg|_{(\alpha,\gamma)=(\alpha_0,\gamma_0)}\geqslant 0 \tag{12}$$

which ensure that (α_0, γ_0) is a relative minimum of the right-hand side of (9).

A simple way to prove (9) is the use of an inequality,

$$\sum_z P(z)\log\frac{Q(z)}{P(z)}\leqslant 0 \tag{13}$$

where $P(z)$ and $Q(z)$ are arbitrary functions that satisfy the relations $0\leqslant P(z), Q(z)\leqslant 1$ and $\sum_z Q(z)=\sum_z P(z)=1$. If we set $P(y)=\sum_x p_{\alpha_0}(y|x)\pi_{\gamma_0}(x)$ and $Q(y)=\sum_x p_\alpha(y|x)\pi_\gamma(x)$, it is easy to verify the requirement of the inequality (13). Then it follows that, for any α and γ,

$$\left[\left[\log\frac{\sum_x p_\alpha(y|x)\pi_\gamma(x)}{\sum_x p_{\alpha_0}(y|x)\pi_{\gamma_0}(x)}\right]_{p_{\alpha_0}(y|x)}\right]_{\pi_{\gamma_0}(x)}\leqslant 0. \tag{14}$$

This proves (9) and its corollaries (10)–(12). We can also prove (10)–(12) through direct calculations similar to that for (5).

The Bayesian framework is an important language in wide areas of the science of information processing, such as time-series analysis, image restoration, inference with neural networks and artificial intelligence. An earlier remark on the analogy between Bayesian statistics and statistical mechanics is found, for example, in Iba [23]. Sourlas [24] seems to be the first to have discussed the relation between coding theory and spin glasses. For recent works dealing with the relation between statistical mechanics and Bayesian statistics (or error-correcting codes), see [25–31, 43].

3. The Nishimori line

Now we discuss the relation between the results in the previous section and the Nishimori line of spin glasses. To see this in the simplest case of $\pm J$ Ising spin glass, we set the distributions as follows:

$$p_\alpha(y|x)=\frac{1}{Z_\alpha}\exp(-E_\alpha(x,y)) \tag{15}$$

$$-E_\alpha(x,y)=\alpha\sum_{(i,j)}y_{ij}x_ix_j \tag{16}$$

$$Z_\alpha=\sum_y\exp(-E_\alpha(x,y))=(\exp(\alpha)+\exp(-\alpha))^M \tag{17}$$

and

$$\pi(x)=\frac{1}{2^N}\qquad\text{(the uniform distribution)} \tag{18}$$

where each of the component x_i ($i\in\{1\ldots N\}$) of the parameter x takes the value of ± 1. The component x_i is defined on the vertices i of a graph G, say a square lattice or a random network, of degree N. The data $y=\{y_{ij}\}$ is defined on the edges (i, j) of G and the summation $\sum_{(i,j)}$ runs over them. We denote the number of the edges of G as M, which is also the number of the data.

This probability $p_\alpha(y|x)$ corresponds to a binary symmetric channel where a set $\{y_{ij}^{in}\}$ $((i, j) \in G)$ of the pair product $y_{ij}^{in} = x_i x_j$ of the inputs is sent as an error-correcting code [8, 10, 24, 28]. Here, 'binary symmetric' means that the output of the channel y_{ij} is given by the formula

$$\begin{aligned} y_{ij} &= +y_{ij}^{in} \qquad \text{with probability} \quad 1-q \\ y_{ij} &= -y_{ij}^{in} \qquad \text{with probability} \quad q. \end{aligned} \tag{19}$$

If we assume that the data y_{ij} is generated by $p_{\alpha_0}(y|x)$ and $\pi(x)$ defined by (15) and (18), the noise q of the channel is related to the hyperparameter α_0 by

$$q = \frac{\exp(-\alpha_0)}{\exp(\alpha_0) + \exp(-\alpha_0)}. \tag{20}$$

Although the 'pair product code' $y_{ij}^{in} = x_i x_j$ defined above looks rather artificial, recent works [28, 29] on error-correcting codes suggest that its generalization might have practical importance†.

The posterior distribution of the model with data $\{y_{ij}\}$ and hyperparameter α is

$$p_\alpha(x|y) = \frac{1}{Z_{pos}} \exp(-E_\alpha(x, y)) \tag{21}$$

$$Z_{pos} = \sum_x \exp(-E_\alpha(x, y)). \tag{22}$$

This is the Gibbs distribution of a random-bond Ising model with coupling constants $\{y_{ij}\}$ defined on the graph G. A derivative of the function F defined by (8) is

$$\frac{\partial}{\partial\alpha} F(\alpha, \gamma) = \frac{1}{\alpha} \langle E_\alpha(x', y) \rangle_{p_\alpha(x'|y)} + M \cdot \tanh\alpha \tag{23}$$

where $\langle\ \rangle_{p_\alpha(x|y)}$ indicates the canonical average with the energy $E_\alpha(x, y)$. (Here and hereafter, we assume that we are working at the unit temperature $T = 1$ and α is treated as a (hyper)parameter of the model but not the temperature.) The term $M \cdot \tanh\alpha$ comes from the derivative of the logarithm of the normalization factor $(\exp(h) + \exp(-h))^M$ of the probability $p_\alpha(y|x)$.

In general, a mis-specification of hyperparameter α in (21) and (23) is possible. In such cases, the corresponding average of the energy $[\langle E_\alpha(x', y) \rangle_{p_\alpha(x'|y)}]_{p_{\alpha_0}(y|x)}]_{\pi(x)}$ is not easy to calculate. If we consider the case where we know the 'true' value α_0 used in the generation of the data $\{y_{ij}\}$ and set $\alpha = \alpha_0$, or, equivalently,

$$\frac{\exp(-\alpha)}{\exp(\alpha) + \exp(-\alpha)} = q \tag{24}$$

in the expression of the average, we have an identity

$$-[\langle E_\alpha(x', y) \rangle_{p_\alpha(x'|y)}]_{p_\alpha(y|x)}]_{\pi(x)} = \alpha M \cdot \tanh\alpha. \tag{25}$$

† Another interesting interpretation of the probability $p_\alpha(y|x)$ is given by a problem [32] that arises in the analysis of social network data [33]. With this interpretation, the index i indicates a person and the binary variable $y_{ij} \in \{\pm 1\}$ $(y_{ij} = y_{ji})$ indicates a social relation between persons i and j, for example, whether they have an acquaintance or not. Each person is assumed to belong to one of the social groups A or B, and the problem is to infer the group structure from the data $\{y_{ij}\}$. We set the indicator $x_i = 1$ when $i \in A$ and $x_i = -1$ when $i \in B$ and assume the following property:

> If a pair of the persons i and j is in the same social group, $y_{ij} = 1$ with a probability q and -1 with a probability $1-q$. Else if they are in different groups, $y_{ij} = 1$ with probability q' and -1 with a probability $1-q'$.

Then we get the probability $p_\alpha(y|x)$ in the text as a special case where $q' = 1 - q$.

from the identity (10) and the expression (23).

So far, the average over the bond randomness

$$[[\cdots]_{p_\alpha(y|x)}]_{\pi(x)} \tag{26}$$

has a rather complicated structure. There are two steps in the generation of the quenched random variables $\{y_{ij}\}$, which are described by $\pi(x)$ and $p_\alpha(y|x)$, respectively. In this particular case, we can simplify it using *gauge invariance* of the problem. The gauge transformation group of this model is defined by the family of transformations,

$$V_z : \{y_{ij}\} \to \{z_i \cdot y_{ij} \cdot z_j\} \tag{27}$$

$$U_z : \{x_i\} \to \{x_i \cdot z_i\} \tag{28}$$

parametrized by $z = \{z_i\}$, $z_i \in \{\pm 1\}$. This set of transformations consist of one-to-one onto-mappings (permutations) of their domain and U_z satisfy a transitive property, i.e., there exists z with which $U_z(x) = x'$ for any pair of x and x' in the domain. It is easy to show the following relations:

$$p_\alpha(U_z(x)|V_z(y)) = p_\alpha(x|y) \tag{29}$$

$$\pi(U_z(x)) = \pi(x) \tag{30}$$

$$p_\alpha(V_z(y)|U_z(x)) = p_\alpha(y|x) \tag{31}$$

$$E_\alpha(U_z(x), V_z(y)) = E_\alpha(x, y). \tag{32}$$

They are an example of the gauge invariance (or gauge covariance) in a random-spin system ([1, 2, 6, 34], see [5] for a comprehensive treatment with applications to the Nishimori line). By these formulae, we can show that the left-hand side of the expression (25) is written as a simpler average

$$\begin{aligned}
[[\langle E_\alpha(x', y)\rangle_{p_\alpha(x'|y)}]_{p_\alpha(y|x)}]_{\pi(x)} &= \sum_x \sum_y \sum_{x'} E_\alpha(x', y) \cdot p_\alpha(x'|y) \cdot p_\alpha(y|x) \cdot \pi(x) \\
&= \sum_x \sum_y \sum_{x'} E_\alpha(U_z(x'), V_z(y)) \cdot p_\alpha(U_z(x')|V_z(y)) \cdot p_\alpha(V_z(y)|U_z(x)) \cdot \pi(x) \\
&= \sum_y \sum_{x'} E_\alpha(x', y) \cdot p_\alpha(x'|y) \cdot p_\alpha(y|x^*) \\
&= [\langle E_\alpha(x', y)\rangle_{p_\alpha(x'|y)}]_{p_\alpha(y|x^*)}
\end{aligned} \tag{33}$$

where x^* is a ferromagnetic state $\{x_i^*\}$ $(\forall i\ x_i^* = 1)$ and z is a function of x that satisfy $x^* = U_z(x)$. The existence of such z is secured by the transitive property of U_z and the change of the dummy index, say, from $V_z(y)$ to y in the summation $\sum_y$, is justified by the one-to-one onto property of V_z and U_z.

The expression

$$p_\alpha(y|x^*) = \frac{\exp(\alpha \sum_{(i,j)} y_{ij})}{Z_\alpha} \tag{34}$$

defines a joint distribution of $\{y_{ij}\}$, but it is easy to see that the components y_{ij} are mutually independent. Each component y_{ij} is a sample from the distribution

$$\Pr(y_{ij}) = q \cdot \delta(y_{ij} + 1) + (1 - q) \cdot \delta(y_{ij} - 1) \tag{35}$$

where the relation between q and α is defined in (24). Here and hereafter, we denote the average over the distribution (34) or (35) by $[\ ]_q$. By using (33) and these notations, the formula (25) is reduced to the identity

$$-[\langle E(x, y)\rangle_{p_\alpha(x|y)}]_q = \alpha M \cdot \tanh\alpha \tag{36}$$

on the average of the energy of $\pm J$ spin glass with the coupling constants $\{y_{ij}\}$ from the distribution (35). This is nothing but a result reported in the first paper [1] on the Nishimori line.

The relation (24) between (hyper)parameter α in the canonical average and the noise level q in the quenched average is essential and known as the definition of the *Nishimori line*† of the model. In the present derivation, it arises from the condition $\alpha = \alpha_0$ in formulae (9) and (10). This means that the model $p_\alpha(y|x)$ assumed in the analysis of the data coincides with the 'true' probability $p_{\alpha_0}(y|x)$ used in the generation of the data. In terms of the coding theory, it corresponds to the situation where the decoder knows exactly the property of the channel, the coding, and the relative frequencies of the possible messages.

A similar argument with the substitution of the second derivative $\frac{\partial^2}{\partial^2 \alpha} F(\alpha, \gamma)$ of F into the expression (12) leads to the inequality

$$[[\langle E_\alpha^2(x', y)\rangle_{p_\alpha(x'|y)} - \langle E_\alpha(x', y)\rangle^2_{p_\alpha(x'|y)}]_{p_\alpha(y|x)}]_{\pi(x)} \leqslant \frac{\alpha^2 M}{\cosh^2 \alpha}. \tag{37}$$

With gauge invariance of the model, we can derive the inequality

$$[\langle E_\alpha(x, y)^2\rangle_{p_\alpha(x|y)} - \langle E_\alpha(x, y)\rangle^2_{p_\alpha(x|y)}]_q \leqslant \frac{\alpha^2 M}{\cosh^2 \alpha} \tag{38}$$

from (37). This is an inequality on the fluctuation of the energy (the specific heat) on the Nishimori line, which is also discussed in [1]. Some of the other relations that hold on the Nishimori line of the model is derived from the identity (7) and the gauge invariance of the model. For example, the distribution of the internal fields at the vertex i_0 [3] is reproduced, when we set $C(x, y) = \sum_j y_{i_0 j} x_j$, where j runs over the set of vertices neighbouring to i_0, i.e., $(i_0, j) \in G$. The expression of the gauge invariant correlation function [1] is also derived, when we set $C(x, y) = x_k \cdot (\Pi_{(i,j)\in\Gamma} y_{ij}) \cdot x_l$, where Γ denotes a path that connects the vertices k and l.

Here we discuss a statistical model defined by (15) and (18), which leads to the Nishimori line of the $\pm J$ spin glass model. Our argument is, however, general and can be applied to the Nishimori line of other models, say, spin glasses with a Gaussian distribution of the coupling [2], models with multiple spin interactions [5, 8, 24, 28–30], and the gauge glasses [5, 9]. For each model, we can consider the corresponding statistical model (or a noisy channel) and derive the properties of the Nishimori line from the relations (10), (12), (7) with additional arguments on the gauge invariance. A problem corresponding to gauge glass might have practical importance in the analysis of the data from optical measurements where differences of the phase between neighbouring points are observed with noise [35].

4. What is new?

Now, careful readers may ask *what is really new* in our approach. Once the left-hand side of (25)

$$\sum_x \sum_y \left(\frac{\sum_{x'} (\alpha \sum_{(i,j)} y_{ij} x_i' x_j') \cdot \exp(\alpha \sum_{(i,j)} y_{ij} x_i' x_j')}{\sum_{x'} \exp(\alpha \sum_{(i,j)} y_{ij} x_i' x_j')} \cdot \frac{\exp(\alpha \sum_{(i,j)} y_{ij} x_i x_j)}{Z_\alpha} \right) \tag{39}$$

is derived by the gauge invariance from that of (36), it is not difficult to show the relation (25) by direct inspection of the expression. If we combined these steps of the proof in this order,

† In general cases, where more than one (component of) hyperparameter is contained in the model, it is actually a 'Nishimori hypersurface'. The term *Nishimori temperature* is also used by statistical physicists. It seems, however, inadequate terminology in the context of information processing, because the notion of temperature has no specific meaning in the problems of the statistics and the coding theory.

it is nothing but a conventional proof [1, 2] of the property of the Nishimori line. The same is true for the derivations with the identity (7). In this sense, our argument is not a re-derivation of the Nishimori line but a *reformulation* or *re-interpretation* of the original derivation.

There are, however, two major advantages of this approach. First, the present interpretation elucidates the meaning of the Nishimori line. It is a line on which we make inference (or decoding) using the 'true' probability structure that generates the data (or codes). The coincidence of the encoding and decoding scheme gives drastic simplifications of the averages of various kind of physical quantities. Prototypes of this interpretation are found in the studies of finite-temperature decoding [7–10]. In this paper, we further developed the idea and represent the whole theory of the Nishimori line with this interpretation. Specifically, we present a novel interpretation to the identity (36) of the energy and the inequality (38). They are essentially necessary conditions that the average of the marginal likelihood takes the maximum value on the Nishimori line.

In the conventional derivation [1, 2] of the Nishimori line, the insertion of the variable x to the left-hand side of (36) looks a rather artificial procedure and the expression (39), which is defined in the configuration space enlarged by a gauge transformation, seems to have no definite meaning. This lacks of the interpretation is a reason why the derivation of the Nishimori line looks somewhat mysterious, even though the manipulation of the formula required in the proof is quite simple and elegant. We believe that the present interpretation will contribute to make this point clear.

The second advantage of the present approach is that it suggests the existence of correspondence of the Nishimori line in the models *without gauge invariance*. Sourlas [10] argued it in the case of optimal decoding. The notion of the Nishimori line without gauge invariance is, however, more general. The relations (10), (12), and (7), which are used in the derivation of the Nishimori line, are obtained without the gauge invariance of the models. By using them, we can prove the identity of the energy, inequality of the specific heat, and the expression of the distribution of internal fields etc, which hold on the 'Nishimori line' of the models without gauge invariance.

For example, we consider a Bayesian model

$$p_\alpha(y|x) = \frac{1}{Z_\alpha}\exp(-E_\alpha(x, y)) \tag{40}$$

$$-E_\alpha(x, y) = \alpha \sum_i y_i x_i \tag{41}$$

$$Z_\alpha = \sum_y \exp(-E_\alpha(x, y)) = (\exp(\alpha) + \exp(-\alpha))^N \tag{42}$$

and

$$\pi_\gamma(x) = \frac{1}{Z_\pi}\exp(-E_\gamma(x)) \tag{43}$$

$$-E_\gamma(x) = \gamma \sum_{(i,j)} x_i x_j \tag{44}$$

$$Z_\gamma = \sum_x \exp(-E_\gamma(x)). \tag{45}$$

In this case, we assume that the unknown parameters $\{x_i\}$ and the data $\{y_i\}$ are defined on the vertices of a graph G with the degree N. When G is a two- or three-dimensional lattice and $x_i, y_i \in \{\pm 1\}$, this model corresponds to an image restoration problem with a prior knowledge on images that is well described by the Ising prior (43). (For image restoration with Ising and

Potts priors, see [19–22, 26, 27].) The posterior distribution of the model is

$$p_{\alpha\gamma}(x|y) = \frac{1}{Z_{\alpha\gamma}} \exp(-E_{\alpha\gamma}(x, y)) \tag{46}$$

$$-E_{\alpha\gamma}(x, y) = \alpha \sum_i y_i x_i + \gamma \sum_{(i,j)} x_i x_j \tag{47}$$

$$Z_{\alpha\gamma} = \sum_x \exp(-E_{\alpha\gamma}(x, y)). \tag{48}$$

This is the Gibbs distribution of an Ising model with an inhomogeneous external field $\{\alpha \cdot y_i\}$.

Let us consider the cases where the data generation mechanism is exactly described by the probabilities (40) with $\alpha = \alpha_0$ and (43) with $\gamma = \gamma_0$, i.e., the pattern of the random field $\{y_i\}$ is given by the following process: (i) Generate a sample pattern $\{y_i^{in}\}$ from the Gibbs distribution of the Ising model (43) with the coupling constant γ_0. (ii) Flip each component $\{y_i^{in}\}$ with the probability

$$q = \frac{\exp(-\alpha_0)}{\exp(\alpha_0) + \exp(-\alpha_0)} \tag{49}$$

(a binary symmetric channel). Then, we can interpret the posterior (46) as a Gibbs distribution of a random field Ising model (RFIM). Note that external fields on the sites of this model are not mutually independent random variables, but correlated with a way specified by (i) and (ii). The 'Nishimori line' of this model is defined as a surface where the parameters (α_0, γ_0) in the definition of the quenched randomness coincide with (α, γ) in the canonical average. Equivalently,

$$\frac{\exp(-\alpha)}{\exp(\alpha) + \exp(-\alpha)} = q \tag{50}$$

$$\gamma = \gamma_0 \tag{51}$$

where α and γ are the parameters in (47), and q and γ_0 are the parameters in (i) and (ii). Then, we can prove identities and inequalities that hold on the Nishimori line of this model. For example, (10) and (5) with $A(x) = x_i x_j$ give the following identities:

$$\left[\left[\left\langle \sum_i y_i x_i \right\rangle_{p_{\alpha\gamma}(x|y)}\right]_{p_\alpha(y|x)}\right]_{\pi_\gamma(x)} = N \cdot \tanh\alpha \tag{52}$$

$$[[\langle x_i x_j \rangle_{p_{\alpha\gamma}(x|y)}]_{p_\alpha(y|x)}]_{\pi_\gamma(x)} = \langle x_i x_j \rangle_\gamma^{pure}. \tag{53}$$

Here and hereafter, the average $\langle \cdots \rangle_\gamma^{pure}$ is the canonical average with the 'pure' Ising model with homogeneous couplings of the strength γ on the same graph G. The expression (53) shows that the quenched average of the correlation of spins in the RFIM with a correlated random field is just the same as that of the corresponding pure Ising model. Furthermore, an identity between the order parameters is obtained if we consider a set of systems of the fixed boundary condition with which $x_i = 1$ for the all spins at the boundary [2]. Assuming that the site j belongs to the boundary and the site i is located far from the boundary, the relation

$$[[\langle x_i \rangle_{p_{\alpha\gamma}(x|y)}]_{p_\alpha(y|x)}]_{\pi_\gamma(x)} = m_\gamma^{pure} \tag{54}$$

is derived from (53), where m_γ^{pure} is the bulk magnetization per spin of the corresponding pure system.

Although these results are dependent on the special features of the model, similar arguments are applicable in other models without gauge invariance and leads to identities and inequalities on the Nishimori line of the model. An example is provided by the posterior distribution corresponding to a binary asymmetric channel, which is already discussed in

Sourlas [10] in the context of optimal decoding. It is related to models with a special type of site/bond randomness.

There are, however, some intrinsic limitations on the utility of the notion of the Nishimori line without gauge invariance. First, we cannot simplify the definition of the quenched average $[[\cdots]_{p_\alpha(y|x)}]_{\pi_\gamma(x)}$ on the Nishimori line without gauge invariance. Then, the results usually contain a complicated quenched average, which often lacks a clear correspondence to that in physical systems. The two-stage process (i) and (ii) of the generation of quenched randomness in the RFIM described above is a typical example of this. Another important remark is that not all of the properties of the Nishimori line with gauge invariance are applicable to the models without gauge invariance. For example, the identity

$$[\langle x_i\rangle_{p_\alpha(x|y)}\cdot\langle x_i\rangle_{p_{\alpha'}(x|y)}]_q=[\langle x_i\rangle_{p_{\alpha'}(x|y)}]_q \tag{55}$$

valid for the $\pm J$ spin glass model [2] has no correspondence in a model without gauge invariance. Here α and q is related by the condition (24) of the Nishimori line and α' takes an arbitrary value. The identity (55) is important because the upper bound

$$|[\langle x_i\rangle_{p_{\alpha'}(x|y)}]_q|\leqslant[|\langle x_i\rangle_{p_\alpha(x|y)}|]_q \tag{56}$$

of the order parameter is derived from it. If we substitute $C(x,y)=x_i\cdot\langle x_i'\rangle_{p_{\alpha'}(x'|y)}$ in (7), we can prove the relation

$$[[\langle x_i'\rangle_{p_\alpha(x'|y)}\cdot\langle x_i'\rangle_{p_{\alpha'}(x'|y)}]_{p_\alpha(y|x)}]_{\pi_\gamma(x)}=[[x_i\cdot\langle x_i'\rangle_{p_{\alpha'}(x'|y)}]_{p_\alpha(y|x)}]_{\pi_\gamma(x)} \tag{57}$$

which apparently corresponds to (55) (here we assume an arbitrary model with binary variables $\{x_i\}$, $x_i\in\{\pm1\}$). However, further simplification of the right-hand side is not possible without gauge invariance. Unfortunately, the expression (57) gives little information on the shape of the boundaries in phase diagram and seems not as useful as (55).

5. Finite-temperature decoding

The notion of the optimality of 'finite-temperature decoding' was introduced to the community of statistical physicists by Ruján [7] and discussed by Nishimori [8, 9] and Sourlas [10]. Recently, it has again drawn the attention of researchers of this field, because the development in statistical mechanics of error-correcting codes [28] enables a quantitative tackling of the problem with analytical methods.

Roughly speaking, 'the optimality of finite-temperature decoding' means that the estimator that maximizes the posterior probability (maximum *a posteriori* estimator (MAP estimator)) is not always the best estimator. The best estimator is dependent on the purpose of inference (or decoding) and often defined with averages over the posterior distribution. If we call the MAP estimator, which is defined as a 'ground state' of the corresponding physical system, the 'zero-temperature decoder', it is natural to call an estimator defined with the posterior averages a 'finite-temperature decoder' or '$T=1$ decoder' [10].

This fact has been well known in the study of the statistics and pattern recognition. For example, Marroquin [37] (see also [21, 22, 26]) discussed an estimator ('MPM estimator') in image restoration problems, which is just the same as the one proposed by Ruján [7]. Moreover, this was not the first work to use the estimator in this field†. General arguments on the optimality of the estimator in the Bayesian framework is already found in the textbooks [12, 38–41] of statistics. The branch of statistics that discusses optimal decisions with uncertain information is known as *statistical decision theory*.

† See, for example, [36] and section 2.4 of [19]. A recent paper on the optimal estimator in image restoration is by Rue [42].

Here, we will briefly discuss the basic results on optimal estimators. Our treatment is not very different from the arguments in Sourlas [10] and those in the textbooks of statistics [12, 38, 40, 41]. It is, however, useful to give a coherent derivation with the notations in the earlier sections, because no comprehensive treatment of this subject seems available in the literature of physics.

To give a formal definition of optimal estimators, we introduce the notion of a *loss function* $L(x, \hat{x})$ that gives a measure of distance‡ between the original parameter x and an estimate $\hat{x}$ of x. Then we define an optimal estimator $\hat{x}(y)$ for a loss function L as a function of y that minimize the expected loss

$$[[L(x, \hat{x}(y))]_{p(y|x)}]_{\pi(x)}. \tag{58}$$

Here and hereafter, we assume that we know exactly about the data generation process and omit the subscripts that indicate hyperparameters α, γ in the expressions, say, $p(y|x)$, $\pi(x)$ and $\langle\ \rangle$ (i.e., we set the values of the hyperparameters to their 'true' values). Note that the optimality of an estimator defined here is a very strong notion. It means that $\hat{x}(y)$ has better or equal average performance against any function of the data y, provided that the data generation scheme (or the set of the channel and the frequencies of the messages) is exactly described by the given probability $p(y|x)$ and $\pi(x)$. The optimality of a $T = 1$ estimator in a set of estimators defined at different 'temperatures' T is a consequence from the optimality defined here.

A basic result on optimal estimators is the following lemma.

Lemma. *An optimal estimator $\hat{x}(y)$ for a loss function L is an estimator that minimize the posterior average $\langle L(x, \hat{x}(y))\rangle_{p(x|y)}$ for each y.*

Proof. The proof of the lemma is as follows [40]†. When we set $C(x, y) = L(x, \hat{x}(y))$, the identity (7) gives

$$[[\langle L(x', \hat{x}(y))\rangle_{p(x'|y)}]_{p(y|x)}]_{\pi(x)} = [[L(x, \hat{x}(y))]_{p(y|x)}]_{\pi(x)}. \tag{59}$$

Note that the estimator $\hat{x}(y)$ is an arbitrary function of y and we can freely attribute its value at each y. On the other hand, the average $[[\cdots]_{p(y|x)}]_{\pi(x)}$ in the left-hand side of (59) is an average over y with non-negative weights. With these properties, we can see that the minimizer of the left-hand side of (59) is the minimizer of the posterior average $\langle L(x', \hat{x}(y))\rangle_{p(x'|y)}$ for each y. Thus, the lemma is proved. □

For example, consider the case where the distance L between the binary sequence $x = \{x_i\}$ and $\hat{x} = \{\hat{x}_i\}$ $(x_i, \hat{x}_i \in \{\pm 1\})$ is measured by the overlap $\sum_i \hat{x}_i x_i$ of the pattern, i.e., $L(x, \hat{x}) = -\sum_i \hat{x}_i x_i$. With this loss function,

$$\langle L(x, \hat{x}(y))\rangle_{p(x|y)} = -\sum_i \hat{x}_i(y)\langle x_i\rangle_{p(x|y)} \tag{60}$$

where $\hat{x}_i(y)$ is the ith component of an estimator $\hat{x}(y)$. Then, the optimal estimator $\hat{x}_i(y) \in \{\pm 1\}$, which minimizes the right-hand side of (60), is given by

$$\hat{x}_i(y) = \frac{\langle x_i\rangle_{p(x|y)}}{|\langle x_i\rangle_{p(x|y)}|}. \tag{61}$$

This expression coincides with the result in [7, 8, 10, 18, 21, 22, 37]. Examples of loss functions and the corresponding optimal estimators are shown in table 1. By using the lemma, we can easily derive them.

‡ It is *not* necessary to satisfy the axiom of the distance.

† Essentially the same, but slightly simpler proof is found in [38]. The present style of the proof has the advantage that it suggests an estimate of the loss with an estimator $\hat{x}(y)$, i.e., the identity (59) justifies the use of $\langle L(x, \hat{x}(y))\rangle_{p(x|y)}$ as an estimator of the loss.

Table 1. Loss functions and corresponding optimal estimators. If there are no special comments in the table, a component of parameters x_i takes its value in a subset of **R** and and an estimate $\hat{x}_i$ of x_i is assumed to take its value in **R**. The symbol $\langle\ \rangle$ indicates the average over the posterior distribution. The expression $\arg\max_z f(z)$ indicates a value of z that maximizes $f(z)$ and Kronceker delta $\delta_{w,z}$ is defined as usual, i.e., $\delta_{w,z} = 1$ if $w = z$, else $\delta_{w,z} = 0$.

Loss function (L)	Optimal estimator ($\hat{x}$)	Comments
$\sum_i (x_i - \hat{x}_i)^2$	$\hat{x}_i = \langle x_i \rangle$	
$\sum_i \lvert x_i - \hat{x}_i \rvert$	$\hat{x}_i =$ median of $p_i(x_i)$[a]	
$1 - \prod_i \delta_{x_i,\hat{x}_i}$	$\{\hat{x}_i\} = \arg\max_x p(x\lvert y)$[b]	x_i: a discrete variable
$\sum_i (1 - \delta_{x_i,\hat{x}_i})$	$\hat{x}_i = \arg\max_{x_i} p_i(x_i)$[a]	x_i: a discrete variable
$-\sum_i x_i \hat{x}_i$	$\hat{x}_i = \frac{\langle x_i \rangle}{\lvert\langle x_i \rangle\rvert}$	$\hat{x}_i \in \pm 1$
$-\sum_i \{x_i \log \hat{x}_i + (1 - x_i)\log(1 - \hat{x}_i)\}$	$\hat{x}_i = \langle x_i \rangle$	$0 < x_i, \hat{x}_i < 1$

[a] Here $p_i(x_i)$ indicates the marginal distribution $p_i(x_i) = \sum_{\{x_j\}j\neq i} p(x|y)$ of x_i, where $\sum_{\{x_j\}j\neq i}$ means the summation over x with a fixed value of the ith component x_i.
[b] It is often called the 'MAP (maximum *a posteriori*) estimator'.

So far, our discussion in this section does not depend on the notion of gauge invariance. Correspondences between loss functions and optimal estimators shown in table 1 are independent of the existence of gauge invariance of the model. With gauge invariance, we can prove an additional result. Let us assume that the model is gauge invariant and the following properties of the loss function L and the estimator $\hat{x}(y)$

$$L(U_z(x), U_z(\hat{x})) = L(x, \hat{x}) \tag{62}$$

$$U_z(\hat{x}(y)) = \hat{x}(V_z(y)) \tag{63}$$

are satisfied for all z (the mappings V_z and U_z are defined in section 3). Then, we can show that the expected loss $[L(x, \hat{x}(y))]_{p(y|x)}$ with any fixed x is independent of the value of x. The proof ([12] p 396, [41] p 168) is as follows:

$$\begin{aligned}
[L(x, \hat{x}(y))]_{p(y|x)} &= \sum_y L(x, \hat{x}(y)) \cdot p(y|x) \\
&= \sum_y L(x^*, U_z(\hat{x}(y))) \cdot p(V_z(y)|x^*) \\
&= \sum_y L(x^*, \hat{x}(V_z(y))) \cdot p(V_z(y)|x^*) \\
&= \sum_y L(x^*, \hat{x}(y)) \cdot p(y|x^*) \\
&= [L(x^*, \hat{x}(y))]_{p(y|x^*)}
\end{aligned} \tag{64}$$

where x^* is an arbitrary chosen 'standard' configuration, say a ferromagnetic state, and z is chosen to satisfy the relation $x^* = U_z(x)$. The result (64) means that the estimator performs equally well for any value of the original parameter x. In terms of statistics [41], an estimator that is optimal within the class of the estimators with such uniformity is called a minimum risk equivariant estimator (MRE)†. The case discussed in Ruján [7] and Nishimori [8] corresponds to a special example of MRE.

In fact, we can remove the assumption (63) on the estimator, if the estimator is optimal and the optimal estimator is known to be unique. This means that if the loss function is gauge

† The term 'invariant' is also used. The author prefers 'covariant', but does not know whether it has been used by statisticians. Here we restrict ourselves within the special form of U_z and V_z induced by the gauge transformation group of Ising spin glass. See [12,41] for definitions and results with an arbitrary group of transitive transformations.

invariant, the corresponding optimal estimator is automatically gauge covariant and satisfies (63)†. The proof is easy, if we note that the estimator defined by

$$\hat{x}_z(y) = U_z^{-1}(\hat{x}(V_z(y))) \tag{65}$$

is an estimator of the equal performance to the original estimator $\hat{x}(y)$, i.e.,

$$[[L(x, \hat{x}_z(y))]_{p(y|x)}]_{\pi(x)} = [[L(x, \hat{x}(y))]_{p(y|x)}]_{\pi(x)}. \tag{66}$$

The relation (66) is confirmed by the calculation similar to that in the proof of (64) under the assumption of (62) and the gauge covariance of $p(y|x)$ and $\pi(x)$. Thus, with the assumption of the uniqueness of the optimal estimator, $\hat{x}_z(y)$ should coincide with $x(y)$ for any value of z. It proves the relation (63).

6. Summary

In this paper, we presented a reconstruction of the theory of the Nishimori line from the viewpoint of Bayesian statistics, or, from the viewpoint of the theory of error-correcting codes. We have developed the idea suggested in the studies of finite-temperature decoding of error-correcting codes [7–10] and explicitly shown that the properties of the Nishimori line are coherently understood with this interpretation. As a byproduct of the interpretation, counterparts of the Nishimori line in models without gauge invariance are given. We also discussed the issues on the 'finite-temperature decoding' of error-correcting codes and clarified the role of gauge invariance.

Acknowledgments

The author would like to thank Y Kabashima, D Saad and H Nishimori for fruitful discussions and kind advice.

References

[1] Nishimori H 1980 *J. Phys. C: Solid State Phys.* **13** 4071–6
[2] Nishimori H 1981 *Prog. Theor. Phys.* **66** 1169–81
[3] Nishimori H 1986 *Prog. Theor. Phys.* **76** 305–6
[4] Nishimori H 1992 *J. Phys. Soc. Japan* **61** 1011–12
[5] Ozeki Y and Nishimori H 1993 *J. Phys. A: Math. Gen.* **26** 3399–429
[6] Toulouse G 1977 *Commun. Phys.* **2** 115–19
(Reprinted 1987 *Spin Glass Theory and Beyond* ed M Mezard, G Parisi and M A Virasoro (Singapore: World Scientific))
[7] Ruján P 1993 *Phys. Rev. Lett.* **70** 2968–71
[8] Nishimori H 1993 *J. Phys. Soc. Japan* **62** 2973–5.
[9] Nishimori H 1994 *Physica* A **205** 1–14
[10] Sourlas N 1994 *Europhys. Lett.* **25** 159–64
[11] Good I J 1965 *The Estimation of Probabilities* (Cambridge, MA: MIT)
[12] Berger J O 1985 *Statistical Decision Theory and Bayesian Analysis* 2nd edn (New York: Springer)
[13] Akaike H 1980 *Bayesian Statistics* ed J M Bernardo *et al* (Valencia: University Press) p 143
[14] MacKay D J C 1992 *Neural Comput.* **4** 415–47
[15] MacKay D J C 1992 *Neural Comput.* **4** 448–72

† It is not true without an additional assumption on the uniqueness. A counter-example is given by a binary symmetric channel with extreme noise $q = \frac{1}{2}$, which transmits no information. For this example, any estimator is 'optimal' for $L(x, \hat{x}) = -\sum_i \hat{x}_i x_i$. Some of them, say an estimator that returns a constant as estimates, are evidently not a MRE. Note that the gauge invariance of the prior $\pi(x)$, which is not used in the proof of (64), is also essential here.

[16] Kitagawa G and Gersch W 1996 *Smoothness Priors Analysis of Time Series (Lecture Notes in Statistics 116)* (New York: Springer)
[17] Geman S and McClure D E 1987 *Proc. 46th Session of the ISI (Bulletin of the ISI 52)* (Bulletin of the International Statistical Institute) pp 5–21
[18] Devijver P A and Dekesel M M 1987 *Pattern Recognition Theory and Applications (NATO ASI Series F30)* ed P A Devijver and J Kittler (Berlin: Springer) p 141
[19] Besag J 1986 *J. R. Stat. Soc.* B **48** 259–302 (with discussion)
[20] Geman S and Geman D 1984 *IEEE Trans. Pattern Anal. Machine Intell.* **6** 721–41
[21] Marroquin J, Mitter S and Poggio T 1987 *J. Am. Stat. Assoc.* **82** 76–89
[22] Winkler G 1995 *Image analysis, Random fields and Dynamic Monte Carlo Methods, A Mathematical Introduction* (Berlin: Springer)
[23] Iba Y 1989 *Cooperative Dynamics in Complex Physical Systems* ed H Takayama (Berlin: Springer) pp 235–6
[24] Sourlas N 1989 *Nature* **339** 693–5
[25] Bruce A D and Saad D 1994 *J. Phys. A: Math. Gen.* **27** 3355–63
[26] Pryce J M and Bruce A D 1995 *J. Phys. A: Math. Gen.* **28** 511–32
[27] Tanaka K and Morita T 1995 *Phys. Lett.* A **203** 122–8
[28] Kabashima Y and Saad D 1999 *Europhys. Lett.* **45** 97–103
[29] Kabashima Y and Saad D 1998 *Europhys. Lett.* **44** 668–74
[30] Dress C, Amic E and Luck J M *J. Phys. A: Math. Gen.* **28** 135–47
[31] Opper M and Winther O 1996 *Phys. Rev. Lett.* **76** 1964
[32] Iba Y 1996 *Proc. Inst. Stat. Math.* **44** 49–84 (in Japanese with English summary)
[33] Wang Y J and Wong G Y 1987 *J. Am. Stat. Assoc.* **82** 8–19
[34] Fradkin E, Huberman B A and Shenker S H 1978 *Phys. Rev.* B **18** 4789–814
[35] Takajo H and Takahashi T 1988 *J. Opt. Soc. Am.* A **5** 416–25
[36] Derin H, Elliott H, Cristi R and Geman D 1984 *IEEE Trans. Pattern Anal. Machine Intell.* **6** 707–20
[37] Marroquin J L 1985 *MIT AI Memo* No 839
[38] Chernoff H and Moses L E 1959 *Elementary Decision Theory* (New York: Wiley) see appendix E_8 for the derivation of the lemma in this paper
[39] Wonnacott T H and Wonnacott R J 1977 *Introductory Statistics* 3rd edn (New York: Wileys) see ch 20 on the Bayesian decision theory
[40] Matsubara N 1992 *Analysis of Scientific Data: Basic Statistics* vol 3, (in Japanese) ed Statistics Section, Department of Social Sciences, College of Arts and Sciences, University of Tokyo (Tokyo: University of Tokyo Press) see ch 9 on the Bayesian decision
[41] Lehmann E L 1983 *Theory of Point Estimation* (CA: Wadsworth) reprinted in 1991
[42] Rue H 1995 *J. Am. Stat. Assoc.* **90** 900–8
[43] Marion G and Saad D 1996 *J. Phys. A: Math. Gen.* **29** 5387–404