Journal of Multidisciplinary Engineering Science Studies (JMESS)
ISSN: 2458-925X
Vol. 7 Issue 2, February - 2021

# Parameter Determination of Probability Distribution by Bayesian Inference

**Hiroshi Isshiki**
Representative
Institute of Mathematical Analysis (Osaka)
Osaka-Sayama, Japan
isshiki@dab.hi-ho.ne.jp

***Abstract*—Nowadays, the neural network is used in many fields. The neural network or deep learning has become synonymous with artificial intelligence (AI), and the idea that human intellectual work will soon be replaced by AI has been born. However, it's doubtful that deep learning is perfect. It seems to have many problems. For example, it is known that there are many problems such as "inference is a black box", "unexpected answer due to overfitting", and "large-scale network and long-time learning". Bayesian inference can provide learning and inference that is completely different from neural networks. Therefore, it may be possible to overcome the problems of neural networks. In this paper, we discuss the application of Bayesian inference to parameter estimation of probability distribution.**

***Keywords*— Bayes statistics; Bayesian inference; prior probabilities; AI; neural network; parameters of probability distribution**

## I. INTRODUCTION

Since the advent of deep learning, the neural network has brought a big innovation in the world [1, 2, 3]. However, deep learning might be far from perfect, because of "the inference is a black box", "unexpected answer due to the overfitting", and "large scale of the network and long time learning". The earliest answer to them should be given. Among them, the black box nature would be a fundamental problem.

The Bayesian inference is based on a quite different theory as the neural network [4, 5, 6]. It might be free from a few problems of the neural networks. The Bayesian inference has once almost disappeared from the world of statistics. The prior probability has a densely subjective aspect, and a part of the orthodox statisticians rejected it. However, in the decision making in the management problems, completely objective premises are almost impossible. In those cases, the Bayesian inference that can include subjectivity could be highly practical.

In neural network learning, the difference between the neuron value of the output layer and the teacher data for a certain input is regarded as an error, and the weight is adjusted so that the error is minimized. That is, we have to solve the multivariable minimum value problem. On the other hand, in Bayesian learning using Bayesian inference, it means to obtain the probability distribution of the output conditional on the input from a large number of training data. In Bayesian learning, learning is to find the frequency distribution from the learning data. However, in the case of Bayesian inference, we encounter the problem of the maximum value of multiple variables because we look for the one that maximizes the probability at the judgment stage, but the number of variables is much smaller than that of neural networks. In general, it will not raise a serious problem.

We discuss the application of Bayesian inference to the estimation of probability distribution parameters.

## II. WHAT IS BAYESIAN INFERENCE?

Let the probability of P(*Result* | *Cause*) of *Result* under *Cause* be given. The reverse probability P(*Cause* | *Result*) of the cause that brought the result is obtained by the Bayesian theorem. The Bayesian inference estimates the reverse probability using the Bayesian theorem.

The Bayesian theorem is given by

$$P(Cause \mid Result) = \frac{P(Result, Cause)}{P(Result)} = \frac{P(Result \mid Cause)P(Cause)}{P(Result)} . \quad (1)$$

This is merely a mathematical theorem. *P*(*Cause*), *P*(*Result* | *Cause*), and P(*Cause* | *Result*) are called the prior probability, likelihood function, and posterior probability, respectively. In the following, we show concretely how the Bayesian theorem is used.

Let 5% of a population be patients *Pat* of disease, and 95% be healthy people *Nor*. They are called the prior probabilities, namely

$$P(Pat) = 0.05, \quad P(Nor) = 0.95 . \quad (2)$$

The Bayesian inference was rejected for a while because of the prior probabilities. They do not cause any problem when the prior probabilities are given objectively. In some cases such as business, political and social problems, we are frequently obliged to give them subjectively. However, this property of the Bayesian inference has now become the characteristics of the Bayesian inference.

According to a test, 98% of the patients are positive (*Ptv*) and 2% of those negative (*Ntv*). On the other hand, 4% and 96% of the healthy people are positive and negative, respectively. Namely

$$P(Ptv \mid Pat) = 0.98, \quad P(Ntv \mid Pat) = 0.02, \quad P(Ptv \mid Nor) = 0.04, \quad P(Ntv \mid Nor) = 0.96. \quad (3)$$

We assume that a person who is not identified as patient or non-patient is positive to the test. The probability that a person is a patient is estimated as follows using the Bayesian theorem. In the Bayesian inference, the learning is nothing but to obtain the probabilities.

Since we have

$$P(Ptv, Pat) = P(Ptv \mid Pat)P(Pat) = 0.98 \times 0.05 = 0.049$$
$$P(Ptv, Nor) = P(Ptv \mid Nor)P(Nor) = 0.04 \times 0.95 = 0.038 \quad (4)$$

from (2) and (3), we obtain

$$P(Pat \mid Ptv) = \frac{P(Ptv, Pat)}{P(Ptv, Pat) + P(Ptv, Nor)} = \frac{0.049}{0.087} = 0.563,$$
$$P(Nor \mid Ptv) = \frac{P(Ptv, Nor)}{P(Ptv, Pat) + P(Ptv, Nor)} = \frac{0.038}{0.087} = 0.437. \quad (5)$$

Hence, the probability that the person is a patient is 56.3% and a non-patient 46.7%.

Since we have similarly, if the test result is negative

$$P(Ntv, Pat) = P(Ntv \mid Pat)P(Pat) = 0.02 \times 0.05 = 0.001$$
$$P(Ntv, Nor) = P(Ntv \mid Nor)P(Nor) = 0.96 \times 0.95 = 0.912, \quad (6)$$

we obtain

$$P(Pat \mid Ntv) = \frac{P(Ntv, Pat)}{P(Ntv, Pat) + P(Ntv, Nor)} = \frac{0.001}{0.913} = 0.001$$
$$P(Nor \mid Ntv) = \frac{P(Ntv, Nor)}{P(Ntv, Pat) + P(Ntv, Nor)} = \frac{0.912}{0.913} = 0.999. \quad (7)$$

The probability that the person is a patient is 0.1% and a non-patient 99.9%.

We call $P(Nor|Pos) = 0.437$ and $P(Pat|Ntv) = 0.001$ as false positive and false negative, respectively. There is a case where false positives and false negatives can't be non-negligible.

Rewriting (5), we have

$$P(Pat \mid Ptv) = \frac{P(Ptv, Pat)}{P(Ptv, Pat) + P(Ptv, Nor)} = \frac{P(Ptv \mid Pat)P(Pat)}{P(Ptv \mid Pat)P(Pat) + P(Ptv \mid Nor)P(Nor)}$$
$$P(Nor \mid Ptv) = \frac{P(Ptv, Nor)}{P(Ptv, Pat) + P(Ptv, Nor)} = \frac{P(Ptv \mid Nor)P(Nor)}{P(Ptv \mid Pat)P(Pat) + P(Ptv \mid Nor)P(Nor)}. \quad (8)$$

Hence, if the prior probabilities are equal, namely

$$P(Pat) = P(Nor), \quad (9)$$

we obtain

$$P(Pat \mid Ptv) = \frac{P(Ptv \mid Pat)}{P(Ptv \mid Pat) + P(Ptv \mid Nor)} = \frac{P(Ptv \mid Pat)}{P(Ptv)}$$
$$P(Nor \mid Ptv) = \frac{P(Ptv \mid Nor)}{P(Ptv \mid Pat) + P(Ptv \mid Nor)} = \frac{P(Ptv \mid Nor)}{P(Ptv)}. \quad (10)$$

Equation (10) is nothing but the likelihood estimation.

## Ⅲ. Estimastion of Parameters of Probability Distribtion

### A. *Bernoulli distribution*

Let $x$=1 and $x$=0 refer to the face and back in coin throw, respectively. The probability $P(x)$ is called Bernoulli distribution and given by

$$P(x \mid \mu) = \mu^x (1-\mu)^{1-x}. \quad (11)$$

The parameter μ is the probability of $x$=1. We infer the parameter $\mu$ when a sequence of random numbers $\mathbf{x}=x_1, x_2, \ldots, x_N$ consisting of 1 and 0 is given.

According to the Bayesian theorem, the reverse probability $P(\mu \mid \mathbf{x})$ is given by

$$P(\mu \mid \mathbf{x}) = \frac{P(\mathbf{x}, \mu)}{P(\mathbf{x})} = \frac{P(\mathbf{x} \mid \mu)P(\mu)}{P(\mathbf{x})}. \quad (12)$$

When the number of the candidates of the parameter $\mu$ is $I$, that is, they are $\mu_1, \mu_2, \ldots, \mu_I$, the probability of $\mu=\mu_i$ can be obtained by

$$P(\mu_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \mu_i)P(\mu_i)}{\sum_{j=1}^{I} P(\mathbf{x}, \mu_j)} = \frac{P(\mathbf{x} \mid \mu_i)P(\mu_i)}{\sum_{j=1}^{I} P(\mathbf{x} \mid \mu_j)P(\mu_j)}. \quad (13)$$

If we assume

$$P(\mu_1) = P(\mu_2) = \cdots = P(\mu_I), \quad (14)$$

(13) becomes

$$P(\mu_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \mu_i)}{\sum_{j=1}^{I} P(\mathbf{x} \mid \mu_j)}. \quad (15)$$

This is nothing but the likelihood method. The appropriateness is discussed later.

The likelihood function $P(\mathbf{x} \mid \mu_i)$ could be calculated by

$$P(\mathbf{x} \mid \mu_i) = \prod_{n=1}^{N} P(x_n \mid \mu_i). \quad (16)$$

The $\mu_i$ that makes $P(\mu_i|\mathbf{x})$ given by (15) the maximum becomes the estimation of the parameter $\mu$. This is nothing but the estimation by the maximum likelihood method.

A numerical example of the above-mentioned estimation method is shown below. Let $\mathbf{x}=x_1, x_2, \ldots, x_N$ be a random sequence of 0 and 1 with $N$=100 generated by Bernoulli distribution given by (11) with $\mu$=0.35. The random sequence is shown in Fig. 1.

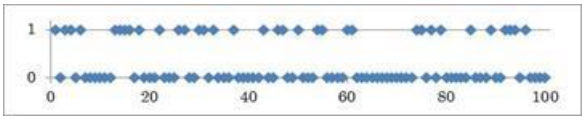



Fig. 1 A random sequence that follows Bernoulli distribution.

The calculation result is shown in Fig. 2, where the candidates of μ are given by

$$\mu_i = \frac{i}{10}, \quad i = 0,1,\cdots,10 \quad (17)$$

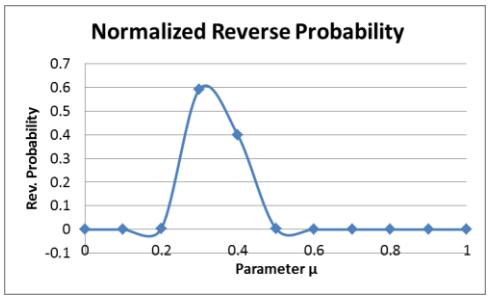



Fig. 2 *The reverse probability (The candidates of μ are coarsely set).*

If we use the finer setting for the candidates as

$$\mu_i = \frac{i}{20}, \quad i = 0,1,\cdots,20\,, \tag{18}$$

we have a result as shown in Fig. 3. Since $\mu$ =0.35 gives the maximum value, this value could be the estimation.

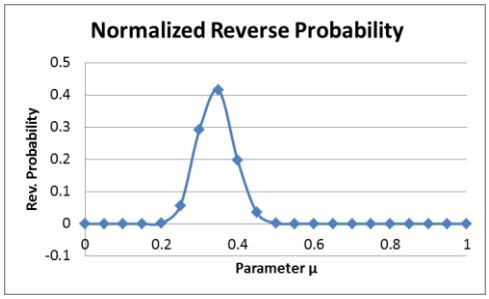
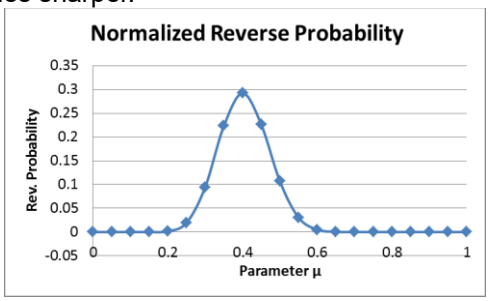
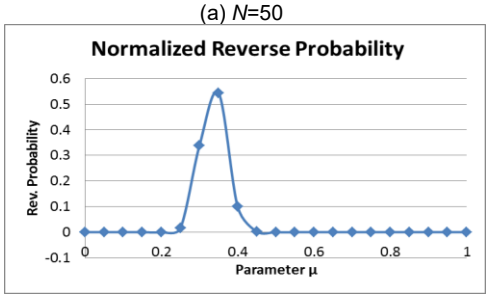



Fig. 3 *The reverse probability (The candidates of μ are finely set).*

For reference, the results of *N*=50 and *N*=200 are shown in Fig. 4 (a) and Fig. 4 (b), respectively. In the case of *N*=50, the estimation of $\mu$=0.4 is wrong. In the case of *N*=200, the result is correct and the peak becomes sharper.




(a) *N*=50




(b) *N*=200

Fig. 4 *Effects of the length of the random sequence N on the reverse probability.*

If the prior probabilities are equal, the Bayes inference is equal to the maximum likelihood method. If the prior probabilities are not equal, a result different from that of the maximum likelihood method might be obtained. However, in the present case, if we increase the number of data *N*, the result does not depend on the choice of the prior probabilities. We show this property using numerical examples below.

If the prior probability is proportional to a normal distribution with the average μ=0.5 and the standard deviation σ=0.25:

$$P(\mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu_i - \mu)^2}{2\sigma^2}\right), \tag{19}$$

the parameter $\mu_i$ estimated using (13) converges to 0.35 as *N* increases.

TABLE 1 *The Bayesian estimates of $\mu_i$ with the unequal choices of the prior probabilities.*

| Number of Data *N* | Estimated $\mu_i$ | Max of $P(\mu_i)$ |
|---|---|---|
| 25 | 0.4 | 0.21 |
| 50 | 0.4 | 0.292 |
| 100 | 0.35 | 0.415 |
| 200 | 0.35 | 0.543 |
| 400 | 0.35 | 0.684 |

If we assume that the prior probabilities are proportional to a normal distribution with the average $\mu$=0.75 and the standard deviation $\sigma$=0.125, we also obtain the same result. This means that the prior probabilities do not affect the estimate as far as we use as big *N* as sufficient. This property originates from the fact that the data are generated from a single source. If the data are generated from several sources, the prior probabilities possibly affect the posterior probabilities.

### *B. Normal distribution*

We assume two parameters of a normal distribution are the mean $\mu$=0 and standard deviation $\sigma$=1. Let the candidate of μ and σ be discretized as

$$\begin{aligned} \mu_i &= -3.0 + 0.25i, \quad i = 0,1,\cdots,24 \\ \sigma_j &= 0.125 + 0.125\,j, \quad j = 0,1,\cdots,24 \end{aligned}. \tag{20}$$

Equation (13) in case of Bernoulli distribution is replaced by

$$\begin{aligned} P(\mu_i,\sigma_j \mid \mathbf{x}) &= \frac{P(\mathbf{x} \mid \mu_i,\sigma_j)P(\mu_i)P(\sigma_j)}{\sum_{i=1}^{I}\sum_{j=1}^{J} P(\mathbf{x},\mu_i,\sigma_j)} \\ &= \frac{P(\mathbf{x} \mid \mu_i,\sigma_j)P(\mu_i)P(\sigma_j)}{\sum_{i=1}^{I}\sum_{j=1}^{J} P(\mathbf{x} \mid \mu_i,\sigma_j)P(\mu_i)P(\sigma_j)} \end{aligned}. \tag{21}$$

With respect to the prior probabilities, if we assume

$$\begin{aligned} P(\mu_1) = P(\mu_2) = \cdots = P(\mu_I) \\ P(\sigma_1) = P(\sigma_2) = \cdots = P(\sigma_I) \end{aligned}, \tag{22}$$

(21) becomes

$$P(\mu_i,\sigma_j \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \mu_i,\sigma_j)}{\sum_{i=1}^{I}\sum_{j=1}^{J} P(\mathbf{x} \mid \mu_i,\sigma_j)}. \tag{23}$$

This is nothing but the likelihood method.

The likelihood function $P(\mathbf{x}|\mu_i,\sigma_j)$ could be calculated by

$$P(\mathbf{x}\,|\,\mu_i,\sigma_j)=\prod_{n=1}^{N}P(x_n\,|\,\mu_i,\sigma_j)\,. \tag{24}$$

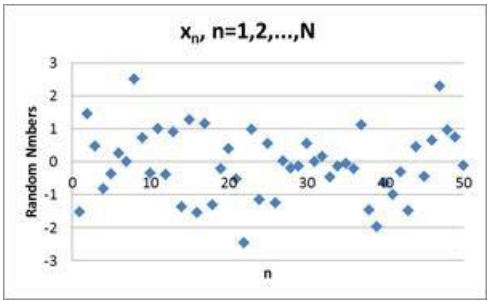
Parameters $\mu_i$ and $\sigma_j$ making the posterior probability maximum become the estimates of the parameters $\mu$ and $\sigma$.

Numerical examples applying the above-mentioned estimation method are given below. A random number sequence $\boldsymbol{x}=x_1, x_2, \ldots, x_N$ of the length $N$=50 is generated from a normal distribution with the parameters of $\mu$=0、$\sigma$=1:

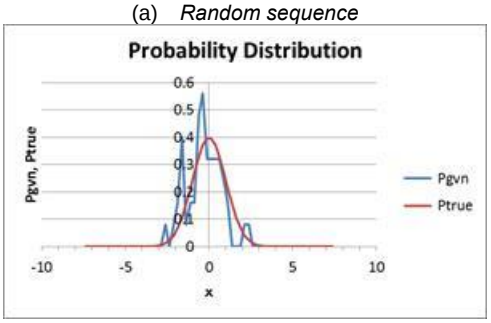
$$P(x\,|\,\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{25}$$

Fig. 5 shows the random sequence and a comparison of the approximate and true probability distributions.

The values of the probability calculated by (23) are shown in Table 2. Since the probability takes the maximum at $\mu$=0 and $\sigma$=1, we consider the values as the estimates. The results are correct.




(a) *Random sequence*




(b) *Probability distribution*

Fig. 5 *A random sequence and probability distribution from normal distribution*.

TABLE 2 *Calculation results of the reverse probability*.

| | σ=0.875 | σ=1 | σ=1.125 | σ=1.25 | σ=1.375 | σ=1.5 | σ=1.625 | σ=1.75 |
|---|---|---|---|---|---|---|---|---|
| μ=-0.75 | 0 | 0.000003 | 0.000022 | 0.000035 | 0.000021 | 0.000007 | 0.000002 | 0 |
| μ=-0.5 | 0.000156 | 0.002803 | 0.005469 | 0.00307 | 0.000841 | 0.000152 | 0.000022 | 0.000003 |
| μ=-0.25 | 0.023711 | 0.131224 | 0.114232 | 0.035999 | 0.006432 | 0.00084 | 0.000093 | 0.00001 |
| μ=0 | 0.06083 | 0.269952 | 0.201982 | 0.057119 | 0.00942 | 0.001157 | 0.000122 | 0.000012 |
| μ=0.25 | 0.002634 | 0.0244 | 0.030235 | 0.012266 | 0.002642 | 0.000398 | 0.000049 | 0.000006 |
| μ=0.5 | 0.000002 | 0.000097 | 0.000383 | 0.000356 | 0.000142 | 0.000034 | 0.000006 | 0.000001 |
| μ=0.75 | 0 | 0 | 0 | 0.000001 | 0.000001 | 0.000001 | 0 | 0 |

### *C. Gamma distribution*

Gamma distribution has also two positive parameters, that is, the shape parameter $\kappa$=1 and scale parameter $\theta$ =1.5 :

$$P(x\,|\,\kappa,\theta)=\frac{1}{\Gamma(\kappa)\theta^{\kappa}}x^{\kappa-1}\exp^{-x/\theta}\quad \text{for}\ \ x>0\,. \tag{26}$$

Let the candidates of the parameters be, for example

$$\begin{aligned}\kappa_i&=0.125+0.125i,\quad i=0,1,\cdots,24\\ \theta_j&=0.125+0.125\,j,\quad j=0,1,\cdots,24\end{aligned}\,. \tag{27}$$

If we make similar assumptions, (23) for normal distribution becomes

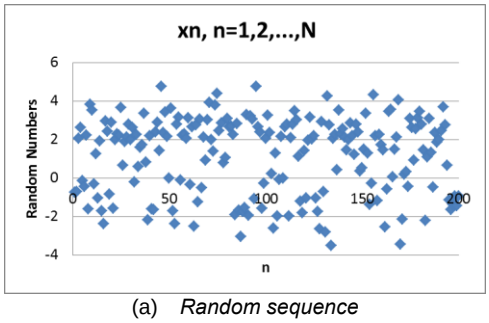
$$P(\kappa_i,\theta_j\,|\,\mathbf{x})=\frac{P(\mathbf{x}\,|\,\kappa_i,\theta_j)}{\sum_{i=1}^{I}\sum_{j=1}^{J}P(\mathbf{x}\,|\,\kappa_i,\theta_j)}\,. \tag{28}$$

The likelihood function $P(\mathbf{x}|\kappa_i,\theta_j)$ can be calculated by

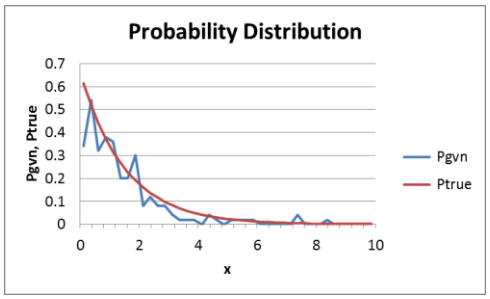
$$P(\mathbf{x}\,|\,\kappa_i,\theta_j)=\prod_{n=1}^{N}P(x_n\,|\,\kappa_i,\theta_j)\,. \tag{29}$$

Parameters $\kappa_i$ and $\theta_j$ making (28) maximum become the estimates of the parameter $\kappa$ and $\theta$. This is nothing but the maximum likelihood method.

A numerical example of the above-mentioned estimation method is shown below. Suppose that a random sequence $\boldsymbol{x}=x_1, x_2, \ldots, x_N$ be generated by gamma distribution given by (26) with $\kappa$=1, $\theta$=1.5 and $N$=200. Fig. 6 shows the random sequence and a comparison of the approximate and true probability distributions.




(a) *Random sequence*




(b) Probability distribution

Fig. 6 *A random sequence and probability distribution from gamma distribution*.

The values of the probability calculated by (28) are shown in Table 3. Since the probability takes the maximum at $\kappa$=1 and$\theta$=1.5, we consider the values as the estimates. The results are correct.

TABLE 3 *Calculation results of the reverse probability*.

| | θ=1.25 | θ=1.375 | θ=1.5 | θ=1.625 | θ=1.75 | θ=1.875 | θ=2.0 | θ=2.125 |
|---|---|---|---|---|---|---|---|---|
| κ=0.75 | 0 | 0 | 0 | 0.000018 | 0.000196 | 0.000749 | 0.001292 | 0.001205 |
| κ=0.875 | 0.000001 | 0.000267 | 0.007728 | 0.04341 | 0.072789 | 0.049475 | 0.017012 | 0.003484 |
| κ=1 | 0.002883 | 0.074423 | 0.245057 | 0.186109 | 0.048935 | 0.005927 | 0.000406 | 0.000018 |
| κ=1.125 | 0.04973 | 0.118464 | 0.044303 | 0.004549 | 0.000188 | 0.000004 | 0 | 0 |
| κ=1.25 | 0.011016 | 0.002422 | 0.000103 | 0.000001 | 0 | 0 | 0 | 0 |
| κ=1.375 | 0.000057 | 0.000001 | 0 | 0 | 0 | 0 | 0 | 0 |
| κ=1.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

### *D. How to treat when the calculated values become too small because of the product of too many probabilities*

We must calculate many products of the probabilities in the Bayesian estimation. Since the probability is less than or equal to 1, the products of many probabilities make underflows. The denominator of the reverse probability does not affect the order of the size of the reverse probabilities. Hence, if we compare the logarithm of the likelihood function given by (16), (24) and (29), the maximum of the reverse function can be determined. We can prevent the underflow by taking the logarithm of the probabilities. at $\mu$=0 and $\sigma$=1, we consider the values as the estimates. The results are correct.

As an example, we consider the problem of determining the parameters of normal distribution discussed in section 3.B. We consider the logarithm of the likelihood function given by (24):

$$\log P(\mathbf{x} \mid \mu_i, \sigma_j) = \sum_{n=1}^{N} \log P(x_n \mid \mu_i, \sigma_j) \,. \tag{30}$$

In the following calculation, a function $f$ given by (31):

$$f(\mathbf{x} \mid \mu_i, \sigma_j) = \exp\left[ \log\left\{ \prod_{n=1}^{N} \frac{P(x_n \mid \mu_i, \sigma_j)}{\max_{i',j'} \left( P(x_{n'} \mid \mu_{i'}, \sigma_{j'}) \right)} \right\} \right]$$

$$= \exp\left[ \sum_{n=1}^{N} \left\{ \log P(x_n \mid \mu_i, \sigma_j) - \max_{i',j'} \left( \log P(x_n \mid \mu_i, \sigma_j) \right) \right\} \right] \tag{31}$$

is used instead of (30). The calculation results for the same numerical example as in section 3.B are shown in Table 1. Since $\mu$=0 and $\sigma$=1 make $f$ maximum, these values is considered the estimate of the parameter $\mu$ and $\sigma$.

TABLE 4 *Results f given by (31)*.

| | σ=0.75 | σ=0.875 | σ=1 | σ=1.125 | σ=1.25 | σ=1.375 | σ=1.5 |
|---|---|---|---|---|---|---|---|
| μ=-1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| μ=-0.75 | 0 | 0 | 0.00001 | 0.000082 | 0.000131 | 0.000078 | 0.000025 |
| μ=-0.5 | 0.000001 | 0.000578 | 0.010382 | 0.02026 | 0.011374 | 0.003116 | 0.000563 |
| μ=-0.25 | 0.000477 | 0.087833 | 0.486102 | 0.423157 | 0.133353 | 0.023828 | 0.003111 |
| μ=0 | 0.00172 | 0.225336 | 1 | 0.748213 | 0.21159 | 0.034896 | 0.004287 |
| μ=0.25 | 0.000024 | 0.009758 | 0.090386 | 0.112 | 0.045436 | 0.009787 | 0.001473 |
| μ=0.5 | 0 | 0.000007 | 0.000359 | 0.001419 | 0.00132 | 0.000526 | 0.000126 |
| μ=0.75 | 0 | 0 | 0 | 0.000002 | 0.000005 | 0.000005 | 0.000003 |

### *E. Compound distribution*

For the solution to the problem in the present section, we need a large number of data. We face difficulties in the numerical calculations since the underflows discussed in section 3.D occur and the calculations can't be continued. When the data number $N$ is smaller than or equal to 200, we calculate the likelihood function using the conventional method. However, when N is bigger than 200, we take the logarithm of the likelihood function as discussed in the previous

section, since the large or small relationship does not change, if we take logarithm.

We consider a compound distribution consisting of several probability distributions. As an example, we consider a compound distribution of two normal distributions. Let the parameters of the two distributions be $(\mu_1,\sigma_1)$ and $(\mu_2,\sigma_2)$, and the mixing ratio be $a$ :

$$P(x\,|\,\mu_1,\sigma_1,\mu_2,\sigma_2,a)=a\frac{1}{\sqrt{2\pi\sigma_1^{\,2}}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^{\,2}}\right)+(1-a)\frac{1}{\sqrt{2\pi\sigma_2^{\,2}}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^{\,2}}\right). \quad (32)$$

As the candidates of the parameters, we consider

$$\mu_{1i}=-3+\frac{6}{P}i,\quad \sigma_{1i}=0.125+\frac{3}{P}i,\quad i=0,1,\cdots,P; \quad (33a)$$

$$\mu_{2i}=-3+\frac{6}{P}i,\quad \sigma_{2i}=0.125+\frac{3}{P}i,\quad i=0,1,\cdots,P; \quad (33a)$$

$$a_i=+\frac{1}{P}i,\quad i=0,1,\cdots,P. \quad (33a)$$

If we assume that the prior probabilities are all equal, we then have an expression of the reverse probability similar to (23) in the case of a single normal distribution:

$$P(\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m\,|\,\mathbf{x})=\frac{P(\mathbf{x}\,|\,\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m)}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}P(\mathbf{x}\,|\,\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m)}. \quad (34)$$

The likelihood function $P(\mathbf{x}\,|\,\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m)$ is given by

$$P(\mathbf{x}\,|\,\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m)=\prod_{n=1}^{N}P(x_n\,|\,\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m). \quad (35)$$

The parameter $\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m$ that makes (34) maximum becomes the estimate. This is nothing but the maximum likelihood estimation.

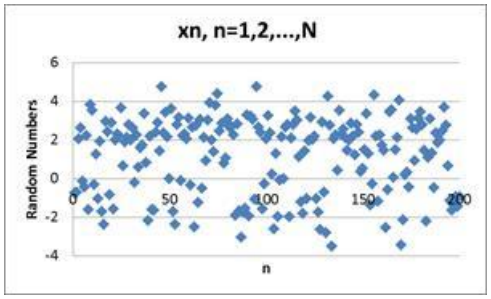
First, we show a numerical result with $N$=200 below. We set the parameters $\mu_1=-1.2,\sigma_1=1.025,\mu_2=2.4,\sigma_2=0.875$ and $a=0.3$. A random sequence $\boldsymbol{x}=x_1, x_2, \ldots, x_N$ is generated from the compound distribution. In Fig. 7(a) , the random sequence is shown. In Fig. 7(b), the true and approximate probability distribution is shown. The approximate probability distribution means the distribution calculated from the random sequence. The maximums of the reverse probability occurred at

$$\mu_1=-1.2,\sigma_1=1.025,\mu_2=2.4,\sigma_2=0.875,a=0.35$$

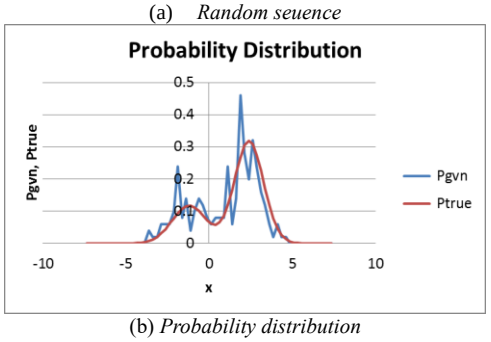
and

$$\mu_1=2.4,\sigma_1=0.875,\mu_2=-1.2,\sigma_2=1.025,a=0.65.$$

The parameters of each probability distribution are estimated correctly, but the estimate of the mixing ratio is not accurate. The correct estimation of the mixing ratio seems difficult.




(a) *Random seuence*




(b) *Probability distribution*

Fig. 7 *A random sequence and the probability distribution*.

We must conduct the numerical calculations with $N$ ＞200 in order to estimate the mixing ratio $a$ correctly. For this purpose, we apply the method of using a logarithm of the likelihood function instead of the likelihood function itself as discussed in section 3.D. If we take the logarithm of (35), we have

$$\log P(\mathbf{x}\,|\,\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m)=\sum_{n=1}^{N}\log P(x_n\,|\,\mu_{1i},\sigma_{1j},\mu_{2k},\sigma_{2l},a_m). \quad (36)$$

Table 5 gives the results. When $N$=800, the correct result is given. A probability distribution obtained approximately from the frequency distribution with $N$=800 and the true probability distribution are given in Fig 8.

TABLE 5 *Estimation results using (36).*

| N | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | a | logP(**x**\|…) | |
|---|---|---|---|---|---|---|---|
| | | | | | | Biggest | 2nd Biggest |
| 200 | -1.2 | 1.025 | 2.4 | 0.875 | 0.35 | -385.932 | -386.873 |
| 300 | -1.2 | 1.025 | 2.4 | 0.875 | 0.35 | -581.68 | -583.072 |
| 400 | -1.2 | 1.025 | 2.4 | 0.875 | 0.35 | -774.149 | -775.885 |
| 500 | -1.2 | 1.025 | 2.4 | 0.875 | 0.35 | -965.733 | -967.112 |
| 600 | -1.2 | 1.025 | 2.4 | 0.875 | 0.35 | -1154.51 | -1156.63 |
| 700 | -1.2 | 1.025 | 2.4 | 0.875 | 0.35 | -1329.84 | -1330.53 |
| 800 | -1.2 | 1.025 | 2.4 | 0.875 | 0.3 | -1515.98 | -1518.77 |

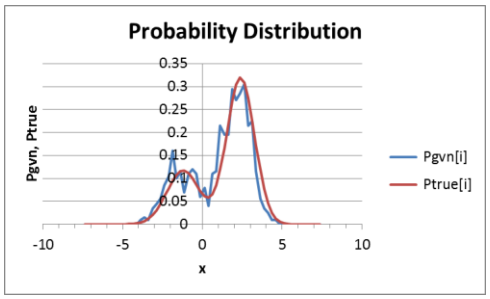



Fig. 8 *Probability distribution (N=800).*

### *F. Search of probability maximum using the mountain-climbing method*

In the above discussion, we obtained the estimation by choosing the parameters making the reverse probability maximum among the candidates of the parameters set beforehand. However, we can obtain the maximum without setting the candidates beforehand.

In the following, we consider the same problem as discussed in section 3.E. However, for simplicity, we assume the parameters of the two probability distributions are given as

$$\mu_1 = -1.2, \sigma_1 = 1.025, \mu_2 = 2.8, \sigma_2 = 1.025,$$

but the mixing ratio $a$ alone is unknown.

Furthermore, we use (36) instead of (35). If we differentiate (36), we have

$$\begin{aligned} &\frac{d}{da}\log P(\mathbf{x} \mid \mu_{1i}, \sigma_{1j}, \mu_{2k}, \sigma_{2l}, a) \\ &= \sum_{n=1}^{N} \frac{d}{da}\log P(x_n \mid \mu_{1i}, \sigma_{1j}, \mu_{2k}, \sigma_{2l}, a) \qquad (37) \\ &= \sum_{n=1}^{N} \frac{\frac{d}{da}P(x_n \mid \mu_{1i}, \sigma_{1j}, \mu_{2k}, \sigma_{2l}, a)}{P(x_n \mid \mu_{1i}, \sigma_{1j}, \mu_{2k}, \sigma_{2l}, a)}. \end{aligned}$$

The calculation results are shown in Fig. 9. When the number of data *N* is increased, the mixing ratio $a$ approaches the correct value 0.3.

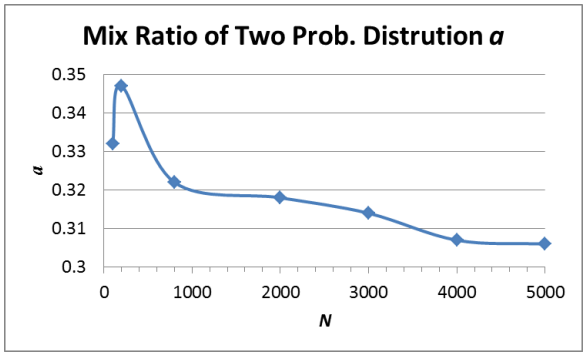



Fig. 9 *The calculation result of the mixing ratio a using the mountain-climbing method.*

In the above discussion, analytical differentiation is used. However, even if we use a numerical differentiation, we could obtain the same result. The numerical differentiation makes the calculation much easier.

## Ⅳ Conclusions

A big innovation has been brought to the world by deep learning. However, deep learning might be far from perfect, because of “the inference is a black box”, “unexpected answer due to the overfitting”, and “large scale of the network and long time learning”. The earliest answer to them should be given. Among them, the black box nature would be a fundamental problem.

The Bayesian inference is based on a quite different theory as the neural network. The learning is quite different. The learning in the Bayesian estimation is nothing but obtaining the probability distribution of the result due to the cause. And the inference is to obtain the probability of cause due to the result using the Bayesian theorem. The Bayesian inference might be free from a few problems of the neural networks.

In the present study, we apply the Bayesian inference to the parameter-estimation of the several probability distributions such as Bernoulli distribution, normal distribution, and gamma distribution. Furthermore, we applied the method to a compound probability distribution consisting of two normal distributions. According to the numerical calculations, satisfactory results are obtained.

If the prior probabilities are taken equal, the Bayesian inference becomes identical to the maximum likelihood method. If the data is generated from a single source as in the present case, the estimated values converge to the same values as shown in the present numerical results, as the number of the data increases. However, if the data are generated from several sources, the posterior probabilities change as the prior probabilities change.

## References

[1] A. Krizhevsky, I. Sutskeverl, and G. Hinton. “ImageNet classification with deep convolutional neural networks,” In Proc. Advances in Neural Information Processing Systems 25 1090–1098 2012.

[2] Yann LeCun1, Yoshua Bengio & Geoffrey Hinton, “Deep learning,” NATURE | VOL 521 | 28 MAY 2015.

[3] M. Taki, Introduction to Deep Learning, Kodansha 2017 in Japanese.

[4] N. Matsubara, Introduction to Bayesian Statistics, Tokyo Tosho 2008 in Japanese.

[5] A. Suyama, Introduction to Mchine Learning by Bayesian Inference, Kodansha 2017 in Japanese.

[6] H. Isshiki, “Pattern Recognition by Bayesian Inference,” The 63rd Joint Meeting of Automatic Control 2020.