Chapter 1
Overview of Watanabe's Bayes

In this chapter, we will first review the basics of Bayesian statistics as a warm-up. In the latter half, assuming that knowledge alone, we will describe the full picture of Watanabe's Bayes Theory. In this chapter, we would like to avoid rigorous discussions and talk in an essay-like manner to grasp the overall picture.

From now on, we will write the sets of non-negative integers, real numbers, and complex numbers as $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{C}$, respectively.

1.1 Frequentist Statistics

For example, let's represent heads of a coin as 1 and tails as 0. If x is a variable representing heads or tails of the coin, x takes the value of 0 or 1. $\mathcal{X} = \{0, 1\}$ is the set of possible values of x. Furthermore, we represent the probability of getting heads with θ that takes values from 0 to 1. $\Theta = [0, 1]$ is the set of possible values of θ. We call θ a parameter. Here, we consider the distribution $p(x|\theta)$ of $x \in \mathcal{X}$ determined by $\theta \in \Theta$. In the case of this coin toss example,

$$p(x|\theta) = \begin{cases} \theta, & x = 1 \\ 1 - \theta, & x = 0 \end{cases} \tag{1.1}$$

can be established. In statistics, when "$p(x|\theta)$ is a distribution", $p(x|\theta)$ must be non-negative and the sum of $x \in \mathcal{X}$ must be 1 (in this case, $p(0|\theta) + p(1|\theta) = 1$).

As it is a coin toss, it might be common to assume that the parameter θ is 0.5. In this case, in statistics, the value of θ is "known". However, if the value of θ is "unknown" because we cannot assume $\theta = 0.5$, for example, due to the coin being bent, we would try tossing the coin several times and estimate the value of θ. If we toss the coin $n = 10$ times and get heads 6 times, we might guess that $\theta = 0.6$, and if we suspect something, we might toss the coin 20 or 100 times.

J. Suzuki, *WAIC and WBIC with R Stan*,
https://doi.org/10.1007/978-981-99-3838-4_1

In this way, the problem of estimating the true value of θ from the data $x_1, \ldots, x_n \in \mathcal{X}$ is called *parameter estimation*, and n (≥ 1) is called the *sample size*. The estimated parameter is often denoted as $\hat{\theta}_n = 0.6$, for example, to distinguish it from the true parameter θ. Alternatively, it can be seen as a mapping like

$$\mathcal{X}^n \ni (x_1, \ldots, x_n) \mapsto \hat{\theta}(x_1, \ldots, x_n) \in \Theta \,.$$

Statistics is a discipline that deals with problems such as parameter estimation, where the distribution generating $x_1, \ldots, x_n$ is estimated.

In the coin tossing problem, with $x_i = 0, 1$, we have

$$\hat{\theta}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1.2}$$

which is called the relative frequency. This is the ratio of the frequency of 1 to the number of data n. If $x_1, \ldots, x_n$ occur independently, as shown in Chap. 4, this value converges to the true parameter θ. This is the weak law of large numbers. However, the convergence means that the probability of $|\hat{\theta}(x_1, \ldots, x_n) - \theta|$ staying within a certain value approaches 1, as $x_1, \ldots, x_n$ vary probabilistically.

At this point, it is worth noting that there are two types of averages. The value in (1.2), which is obtained by dividing the sum of the randomly occurring $x_1, \ldots, x_n$ by the number n, is called the *sample mean*. In contrast, $0 \cdot (1 - \theta) + 1 \cdot \theta = \theta$ is called the *expected value*. In this book, when we say average, we mean the latter.

1.2 Bayesian Statistics

However, it is undeniable that estimators like (1.2) can feel awkward. When watching a baseball game, the batting average from the beginning of the season is displayed. By the second half of the season, that average seems to be close to the player's true ability and the true parameter θ. In the opening game, especially in the second at-bat, it is clear that the displayed batting average is either 0 or 1. That is, in the case of $n = 1$, the calculation in (1.2) results in $\hat{\theta}_n = 0$ or $\hat{\theta}_n = 1$. Furthermore, if the first at-bat results in a walk or something that does not count as an at-bat, $n = 0$ and the calculation in (1.2) cannot even be done. Is there a more intuitive estimator? So, considering that there are many hitters around a .250 batting average in baseball, how about estimating as follows?

$$\hat{\theta}(x_1, \ldots, x_n) = \frac{\sum_{i=1}^{n} x_i + 25}{n + 100} \,. \tag{1.3}$$

The numbers 25 and 100 may be too arbitrary, but they represent prior information and the beliefs of the person making the estimate. The framework that justifies this way of thinking is *Bayesian statistics*. How Eq. (1.3) is derived will be resolved in this chapter.

Before the season begins, there might be someone who imagines the distribution of the player's batting average θ to be roughly around $\theta \in \Theta$. Such a distribution, determined by prior information or the estimator's beliefs, is called the prior distribution, and is denoted by $\varphi(\theta)$. Although it is the same distribution as $p(x|\theta)$, it does not depend on $x \in \mathcal{X}$. However, since it is a distribution, not only must $\varphi(\theta) \geq 0$, but also $\int_\Theta \varphi(\theta)d\theta = 1$. As long as these conditions are met, a uniform distribution such as $\varphi(\theta) = 1, 0 \leq \theta \leq 1$ is acceptable.

The results of the first three at-bats in the opening game, represented by hits as 1 and outs as 0, can be any of the following $(x_1, x_2, x_3) \in \{0, 1\}^3$:

$$000, 001, 010, 011, 100, 101, 110, 111 \ .$$

Here, someone who clearly imagines the prior distribution $\varphi(\theta)$ can calculate the probabilities of these eight events. However, for simplicity, assume that the occurrences of $x_1, x_2, x_3 = 0, 1$ are independent. In fact, the conditional probability for a parameter value of θ is $p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)$. Furthermore, multiplying the prior probability $\varphi(\theta)$ results in $p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)\varphi(\theta)$, but actually, integration over θ is necessary. That is,

$$Z(x_1, x_2, x_3) = \int_\Theta p(x_1|\theta)p(x_2|\theta)p(x_3|\theta)\varphi(\theta)d\theta$$

is the probability of (x_1, x_2, x_3). If the prior distribution is uniform,

$$Z(0, 0, 0) = \int_{[0,1]} (1-\theta)^3 d\theta = \frac{1}{4} \ , \ Z(0, 0, 1) = \int_{[0,1]} (1-\theta)^2 \theta d\theta = \frac{1}{12}$$

$$Z(0, 1, 1) = \int_{[0,1]} (1-\theta)\theta^2 d\theta = \frac{1}{12} \ , \text{and } Z(1, 1, 1) = \int_{[0,1]} \theta^3 d\theta = \frac{1}{4}$$

can be calculated in this way. The other four cases can also be calculated similarly, and the sum of the probabilities of the eight sequences is 1. Similarly, for any general $n \geq 1$ and $\varphi(\cdot)$, we can define $Z(x_1, \ldots, x_n)$. This value is called the *marginal likelihood*.

1.3 Asymptotic Normality of the Posterior Distribution

Next, after the opening game is over and a person has seen the results of the first three at-bats (x_1, x_2, x_3), they can estimate the batting average θ more accurately. As the season progresses and a person sees 100 at-bats $(x_1, \ldots, x_{100})$, the estimation

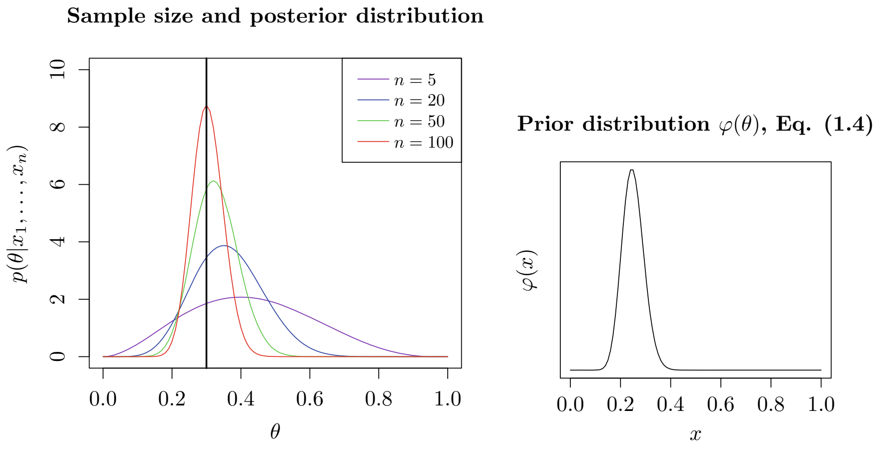

Fig. 1.1 As the sample size increases, the posterior distribution concentrates near the true parameter (left). The prior distribution of batting average (1.4) is maximized at $\theta = 0.25$ (right)

of θ becomes even more accurate. The conditional probability of θ under the data $x_1, \dots, x_n$ is called its *posterior distribution*. As the sample size n increases, the width of the posterior distribution narrows, concentrating around the true value of θ (Fig. 1.1 left).

To calculate the posterior distribution, it is necessary to apply Bayes' theorem. When expressing the conditional probability of event A under event B as $P(A|B)$, *Bayes' theorem* can be written as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} .$$

Let A represent the probability of the parameter being θ, and B represent the probability of the data being $x_1, \dots, x_n$. That is, by setting $P(B|A)$ as $p(x_1|\theta) \cdots p(x_n|\theta)$, $P(A)$ as the prior probability $\varphi(\theta)$, and $P(B)$ as the marginal likelihood $Z_n(x_1, \dots, x_n)$, the posterior distribution $p(\theta|x_1, \dots, x_n)$ can be written as

$$\frac{p(x_1|\theta) \cdots p(x_n|\theta)\varphi(\theta)}{Z(x_1, \dots, x_n)} .$$

Returning to the batting average example,

$$p(\theta|0, 0, 0) = 4(1 - \theta)^3 \ , \ p(\theta|0, 0, 1) = 12(1 - \theta)^2\theta,$$

$$p(\theta|0, 1, 1) = 12(1 - \theta)\theta^2 \text{ , and } p(\theta|1, 1, 1) = 4\theta^3$$

are the result. The other four cases can also be calculated similarly, and in each of the eight series, $\int_0^1 p(\theta|x_1, x_2, x_3)d\theta = 1$ holds.

As such, defining a prior distribution and finding its posterior probability is Bayesian estimation. In contrast to traditional statistics, which provides a single estimate $\hat{\theta}_n$ for parameter estimation as in (1.2) (called *point estimation*), Bayesian statistics gives the result as a posterior distribution. Under certain conditions (regularity) regarding the relationship between the true distribution and the estimated distribution, it is known that the posterior distribution of the parameter $p(\theta|x_1, \ldots, x_n)$ follows a normal distribution even if it is not a normal distribution when the sample size n is large. This is called *asymptotic normality*.

Regularity $\Longrightarrow$ Asymptotic Normality

This theorem will be proven in Chap. 5, but in practical data analysis using Bayesian statistics, asymptotic normality is often not considered. Rather, it is primarily utilized by some Bayesian theorists to prove mathematical propositions and is often seen as a theory for the sake of theory.

In Watanabe's Bayesian theory, this theorem is generalized. Under the condition of "having a relatively finite variance" defined in Chap. 2, the posterior distribution (asymptotic posterior distribution) is derived when n is large. This posterior distribution generally does not become a normal distribution, but it does become a normal distribution when the regularity condition is added.

On the other hand, if the posterior distribution $p(\theta|x_1, \ldots, x_n)$ is known, the conditional probability $r(x_{n+1}|x_1, \ldots, x_n)$ of $x_{n+1} \in \mathcal{X}$ occurring under the series $x_1, \ldots, x_n \in \mathcal{X}$ can be calculated as

$$r(x_{n+1}|x_1 \ldots, x_n) = \int_{\Theta} p(x_{n+1}|\theta) p(\theta|x_1, \ldots, x_n) d\theta \, .$$

This is called the *predictive distribution*. For example, in the case of the batting average problem, where $\mathcal{X} = \{0, 1\}$, the predictive distribution satisfies the properties of a distribution as follows:

$$r(1|x_1, \ldots, x_n) + r(0|x_1, \ldots, x_n) = 1 \, .$$

That is, when the prior distribution $\varphi(\cdot)$ is determined, the marginal likelihood, posterior distribution, and predictive distribution are also determined. For example, if the prior distribution is

$$\varphi(\theta) = \frac{\theta^{24}(1-\theta)^{74}}{\int_0^1 \theta_1^{24}(1-\theta_1)^{74} d\theta_1} \, , \; 0 \le \theta \le 1 \tag{1.4}$$

(as shown in Fig. 1.1 on the right), it can be shown that the predictive distribution $r(1|x_1, \ldots, x_n)$ is given by (1.3) (see Sect. 2.1). The *generalization loss*

$$\mathbb{E}_X \left[-\log r(X|x_1, \ldots, x_n) \right] \tag{1.5}$$

and the *empirical loss*

$$\frac{1}{n}\sum_{i=1}^{n}\{-\log r(x_i|x_1, \ldots, x_n)\} \tag{1.6}$$

are also defined using the predictive distribution. These are the mean and arithmetic mean of $-\log r(x|x_1, \ldots, x_n)$ with respect to $x \in \mathcal{X}$, respectively (Chap. 5). The WAIC (Chap. 6) is also defined using the empirical loss.

In other words, Bayesian statistics can be said to be a statistical method that estimates the true distribution using not only samples but also prior distributions that reflect beliefs and prior information.

1.4 Model Selection

In this book, in addition to determining the posterior distribution for parameter estimation, Bayesian statistics are applied for another purpose. Here, we consider the problem of estimating which of the statistical models 0 and 1 is correct from the data sequence $x_1, \ldots, x_n$, where the statistical model (1.1) is model 1 and the statistical model with equal probabilities of 0 and 1 occurring is model 0. In conventional statistics, the details are omitted, but this would typically involve hypothesis testing.

In Bayesian statistics, the value obtained by applying the negative logarithm to the marginal likelihood, $-\log Z(x_1, \ldots, x_n)$, is called the *free energy*. Free energy is used for *model selection*, which estimates which statistical model the sample sequence $x_1, \ldots, x_n$ follows. Under certain conditions, selecting the model with a smaller free energy value results in a correct choice (called *consistency*) as the sample size $n \to \infty$. If the prior probability is uniform at $n = 3$, the marginal likelihood for model 1 can be calculated as

$$-\log Z(0, 0, 0) = \log 4\ , \ -\log Z(0, 0, 1) = \log 12$$

$$-\log Z(0, 1, 1) = \log 12\ , \text{and}\ -\log Z(1, 1, 1) = \log 4\ .$$

In the case of Model 0, regardless of $(x_1, x_2, x_3) \in \mathcal{X}^3$, the free energy becomes $\log 8$, so Model 1 is chosen when $(x_1, x_2, x_3) = (0, 0, 0), (1, 1, 1)$, and Model 0 is chosen otherwise. If we toss a coin three times and the result is biased toward either 0 or 1, rather than a mix of 0 and 1, our intuition tells us that Model 0 is suspicious. The value of the free energy depends on the choice of the prior distribution, but as the sample size n increases, this dependence disappears. In other words, the influence of actual evidence becomes relatively more significant than prior information or the estimator's beliefs.

Next, let's examine how the estimate in (1.2) is obtained. Here, we derive (1.2) from the criterion of maximizing the likelihood. The likelihood is a quantity defined by

$$p(x_1|\theta) \cdots p(x_n|\theta)$$

when data $x_1, \dots, x_n \in \mathcal{X}$ are obtained. The θ that maximizes this value is called the *maximum likelihood estimator*. The likelihood of (1.2) is $\theta^k(1-\theta)^{n-k}$ when the number of i with $x_i = 1$ is k and the number of i with $x_i = 0$ is $n-k$. We could maximize this value, but instead, we take advantage of the fact that $f(x) = \log x$ is a monotonically increasing function and differentiate

$$k \log \theta + (n-k) \log(1-\theta)$$

with respect to θ and set it to 0, resulting in

$$\frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \, .$$

Solving this equation, we find that (1.2) is obtained.

Furthermore, we assume that the parameter set Θ is a subset of the d-dimensional Euclidean space $\mathbb{R}^d$. Roughly speaking, assuming regularity, it is known that when the sample size n is large, the free energy can be written as

$$\sum_{i=1}^{n} - \log p(x_i|\hat{\theta}_n) + \frac{d}{2} \log n \ ,$$

where we have omitted constant terms and written the maximum likelihood estimator obtained from $x_1, \dots, x_n$ as $\hat{\theta}_n$. This value is called the *BIC*. Also, the dimension d of the parameter space can be interpreted as the number of independent parameters.

Furthermore, replacing the second term $\frac{d}{2} \log n$ with d, we obtain

$$\sum_{i=1}^{n} - \log p(x_i|\hat{\theta}_n) + d \ , \tag{1.7}$$

which we call the *AIC*. The details of AIC and BIC will be discussed in Chap. 6. In any case, these criteria are used to select models with smaller values. AIC, BIC, and free energy are examples of quantities used for this purpose, which we call *information criteria*.

1.5 Why are WAIC and WBIC Bayesian Statistics?

As mentioned in the preface, this book is intended for the following readers:

1. Those with a comprehensive knowledge of mathematical statistics.
2. Those who have used WAIC or WBIC but want to understand their essence.
3. Those with a basic understanding of university-level mathematics such as linear algebra, calculus, and probability statistics.

Readers in categories 2 and 3 should be able to approach the level of reader 1 in mathematical statistics by reading up to this point. On the other hand, we often receive questions like the following, particularly from readers in category 2. The purpose of this book is to answer these questions, but we would like to provide an overview of the answers here.

Why are WAIC and WBIC used instead of AIC and BIC in non-regular cases?

Why are WAIC and WBIC considered Bayesian statistics?

Why is algebraic geometry necessary for understanding this?

Information criteria such as AIC, BIC, *WAIC*, and *WBIC* (Chap. 6) can be calculated from the data sequence $x_1, \ldots, x_n$. Here,

$$WAIC = (\text{empirical loss}) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{V}(x_i) \ , \tag{1.8}$$

where $\mathcal{V}(\cdot)$ is a quantity defined in Chap. 5 and, like empirical loss, is calculated from the posterior distribution $p(\cdot|x_1, \ldots, x_n)$. Calculating this value using R or Python is not difficult.

In any case, by feeding the data sequence into pre-prepared functions, one can calculate the values of such information criteria. However, many people use WAIC instead of AIC in non-regular cases without understanding why, just believing that "in non-regular cases, use WAIC instead of AIC". Just as one would want to understand "why AIC", one would also want to understand "why WAIC". We would like to discuss this in more detail below.

Among these, for WAIC and WBIC, in order to calculate the average of $f : \Theta \to \mathbb{R}$

$$I = \int_{\Theta} f(\theta) p(\theta|x_1, \ldots, x_n) d\theta \tag{1.9}$$

with respect to the posterior distribution $p(\theta|x_1, \ldots, x_n)$, it is necessary to generate random numbers $\theta = a_1, \ldots, a_m \in \Theta$ $(m \geq 1)$ according to $p(\theta|x_1, \ldots, x_n)$, $\theta \in \Theta$

and calculate the approximate value of I:

$$\hat{I} = \frac{1}{m}\sum_{j=1}^{m} f(a_j)\ .$$

Except for special cases, it is considered difficult to calculate the integral of (1.9) mathematically. In this book, we assume the use of specialized software Stan for this purpose. In Chap. 3, we explain how to use it with examples.

Since WAIC and WBIC are calculated by generating random numbers according to the posterior distribution using Stan, it can be inferred that they are Bayesian quantities.

In the following, we denote the operation of taking the average over $x \in \mathcal{X}$ as $\mathbb{E}_X[\cdot]$. For example, the average of $-\log p(x|\hat{\theta}_n)$ over $x \in \mathcal{X}$ is written as $\mathbb{E}_X[-\log p(X|\hat{\theta}_n)]$. Here, we write the variable for which the average is taken in capital letters, such as X. On the other hand, AIC, BIC, and the maximum likelihood estimate $\hat{\theta}_n$ are calculated from the n data points $x_1, \ldots, x_n \in \mathcal{X}$, which actually occur randomly. Therefore, we denote the average of the AIC values taken over these as

$$\mathbb{E}_{X_1\cdots X_n}[AIC(X_1, \ldots, X_n)] \tag{1.10}$$

or simply as $\mathbb{E}_{X_1\cdots X_n}[AIC]$. Also, the value of $\mathbb{E}_X[-\log p(X|\hat{\theta}_n)]$ varies depending on the values of $x_1, \ldots, x_n \in \mathcal{X}$ for $\hat{\theta}_n$. Taking the average over these gives

$$\mathbb{E}_{X_1\cdots X_n}\left[\mathbb{E}_X[-\log p(X|\hat{\theta}(X_1, \ldots, X_n))]\right]\ , \tag{1.11}$$

which we will simply write this as $\mathbb{E}_{X_1\cdots X_n}\mathbb{E}_X[-\log p(X|\hat{\theta}_n)]$. Hirotsugu Akaike, who proposed the AIC, considered the value obtained by averaging the maximum log-likelihood $-\log p(x|\hat{\theta}(x_1, \ldots, x_n))$ over both the training data $x_1, \ldots, x_n$ and the test data x (1.11) to be an absolute quantity. He justified the AIC by showing that the AIC averaged over the training data (1.10) matched (1.11).

Sumio Watanabe found it difficult to remove the assumption of regularity as long as the maximum likelihood estimate used in the first term of AIC (1.7) was applied. In Chap. 6, we will prove that the maximum likelihood estimate may not converge to its original value without assuming regularity. In WAIC (1.8), the empirical loss of (1.6) was introduced as a substitute.

Why is WAIC Bayesian?

Breaking away from AIC's maximum likelihood estimation, which does not work in non-regular cases, replacing its first term with the empirical loss was the first step in developing WAIC. As a result, there was a need to explore the derivation of a posterior distribution without assuming regularity.

We have already mentioned that the justification for AIC is that AIC and $\mathbb{E}_X[-\log p(X|\hat{\theta}_n)]$ coincide when averaged over the training data $x_1, \ldots, x_n$. In

Watanabe's Bayesian theory, just as the negative log-likelihood is replaced with the empirical loss, $\mathbb{E}_X[-\log p(X|\hat{\theta}_n)]$ is replaced with the generalization loss of (1.5). Then, with or without regularity, excluding terms that can be ignored when n is large,

$$\mathbb{E}_{X_1\cdots X_n}[WAIC] = \mathbb{E}_{X_1\cdots X_n}[\text{generalization loss}]$$

holds (strictly speaking, the equality does not hold, but the difference becomes negligible). We will discuss the details in Chap. 6 (regular cases) and Chap. 8 (general cases).

Justification of AIC and WAIC

$$\mathbb{E}_{X_1\cdots X_n}[AIC] = \mathbb{E}_{X_1\cdots X_n}\mathbb{E}_X[-\log p(X|\hat{\theta}_n)]$$
$$\mathbb{E}_{X_1\cdots X_n}[WAIC] = \mathbb{E}_{X_1\cdots X_n}[\text{generalization loss}]$$

In regular cases, $WAIC = AIC$

Abandoning maximum likelihood estimation and introducing generalization loss and empirical loss was the starting point of Sumio Watanabe's journey into his novel theory.

1.6 What is "Regularity"

In statistics, assuming the true distribution is $q(x)$ and the statistical model is $p(x|\theta)$, $\theta \in \Theta$, we often use $\mathbb{E}_X[\log \dfrac{q(X)}{p(X|\theta)}]$ to represent the discrepancy between the two. This quantity is called the Kullback-Leibler (KL) information, and it becomes 0 when $p(x|\theta)$ and $q(x)$ match. We will discuss its definition and properties in Chap. 2. Let's denote the value of $\theta \in \Theta$ that minimizes this value as θ_*. Then, the KL information between $p(x|\theta_*)$ and $p(x|\theta)$ is defined as

$$K(\theta) = \mathbb{E}_X[\log \frac{p(X|\theta_*)}{p(X|\theta)}].$$

If this value becomes 0, θ coincides with θ_*, and $p(x|\theta_*)$ is closest to the true distribution $q(x)$. In the following, we will denote the set of such θ_* as Θ_*.

The *regularity* condition (3 conditions), which we have mentioned several times, will be discussed in detail in Chap. 2, but firstly, it requires that Θ_* consists of exactly one element. Secondly, the θ_* contained in Θ_* minimizes $K(\theta)$, but θ_* must not be an endpoint of Θ, and it must not be 0 when differentiated with respect to θ, where Θ is included in the Euclidean space $\mathbb{R}^d$. When we differentiate $K(\theta)$ twice with respect to each of $\theta_1, \ldots, \theta_d$, we denote the matrix with elements $-\dfrac{\partial^2 K(\theta)}{\partial\theta_i \partial\theta_j}$ as

$J(\theta)$. The value of $J(\theta)$ at $\theta = \theta_*$, $J(\theta_*)$, has all non-negative eigenvalues (non-negative definite, details in Chap. 4), but the third condition is that they all take positive values. In regular cases, as n increases, the posterior distribution approaches a normal distribution, and it is shown that the inverse of the covariance matrix is $nJ(\theta_*)$ (Chap. 5). If $J(\theta_*)$ does not have an inverse, the covariance matrix does not exist.

For example, let's assume that the three types of readers mentioned above follow some distribution for each of the three variables: statistics, WAIC/WBIC, and mathematics, and overall, the distribution is the sum of the three distributions divided by 3. The mixture of normal distributions discussed in Chap. 2 corresponds to this case. In this case, it is known that there are multiple θ_* values for which $K(\theta_*) = 0$. Although regular distributions are common in high school and university statistics courses, it is said that most real-world data are non-regular.

1.7 Why is Algebraic Geometry Necessary for Understanding WAIC and WBIC?

In Watanabe's Bayesian theory, the derived posterior distribution is described using the concept of algebraic geometry, specifically the real log canonical threshold λ. It would be impossible to discuss Watanabe's Bayesian theory without using algebraic geometry.

However, Bayesian statistics and algebraic geometry (Chap. 7) are independent academic fields. To solve the problem of generalizing the posterior distribution, a formula called the state density is used to connect the two. The climax could be said to be there.

In Chap. 5, we define $B_n := \{\theta \in \Theta | K(\theta) < n^{-1/4}\}$ and prove that the posterior distribution of $\theta \in \Theta$ not included in B_n can be ignored, regardless of whether it is regular or not. Therefore, the generalization of the posterior distribution to non-regularity in Chap. 8 is performed only for θ included in B_n, i.e., θ close to θ_*.

The state density formula in Chap. 8 seeks an integration formula when $n \to \infty$ and the volume of a certain B_n is sufficiently small (Sect. 8.2). This allows us to express the posterior distribution and free energy using λ and its multiplicity m. The definitions of λ and m are discussed below.

1.8 Hironaka's Desingularization, Nothing to Fear

Many readers may take time to learn the concept of *manifolds* rather than *desingularization*. However, the algebraic geometry covered in Chap. 7 is intended to provide the prerequisite knowledge for Chap. 8 and beyond, and from the perspective of algebraic geometry as a whole, Watanabe's Bayesian theory uses only a very small portion of it.

Hironaka's theorem is a theorem that claims that even if $K(\theta)$ has singular points, there exists a manifold that appears as if there are no singular points. Heisuke Hironaka says the following:

> When you look at the shadow cast by the track of a roller coaster on the ground, an extremely complex figure is drawn there. Lines intersect in various ways, and in some parts, the shape becomes pointed. What was actually a smooth curve appears as intersecting lines and pointed shapes when focusing on the shadow. In algebraic geometry, such points where lines intersect or become pointed are called "singular points".

Let's give two examples of manifolds.

First, it can be easily verified that the set $\mathbb{P}^1$ of the ratio $[x : y]$ of x and y becomes the union of the set U_x of elements that can be written as $[x : 1]$ and the set U_y of elements that can be written as $[1 : y]$. The elements of $U_x \cap U_y$ can be written as both $[x : 1]$ and $[1 : y]$, so we assume a relationship where $xy = 1$. In each case, we have bijections (one-to-one mappings to the top) with $\mathbb{R}^1$ as follows:

$$\begin{cases} \phi_x : U_x \ni [x : 1] \mapsto x \in \mathbb{R}^1 \\ \phi_y : U_y \ni [1 : y] \mapsto y \in \mathbb{R}^1 \end{cases} .$$

Next, notice that the pairs of elements of $\mathbb{P}^1$ and the entire set of real numbers $\mathbb{R}$, as $x, y, z \in \mathbb{R}$, can be written as either $([x : 1], z)$ or $([1 : y], z)$. If we write each set as U_x, U_y, we have bijections with $\mathbb{R}^2$ as follows:

$$\begin{cases} \phi_x : U_x \ni ([x : 1], z) \mapsto (x, z) \in \mathbb{R}^2 \\ \phi_y : U_y \ni ([1 : y], z) \mapsto (y, z) \in \mathbb{R}^2 \end{cases} .$$

If $\mathbb{P}^1$ and $\mathbb{P}^1 \times \mathbb{R}$ satisfy several other conditions, then they are considered to form manifolds of dimensions 1 and 2, respectively. Then, we call the x, y in the first example and the $(x, z), (y, z)$ in the second example the respective *local variables* and their coordinates the *local coordinates*. In both cases, the manifold M is a union of two open sets, but in general, there can be any number of them, and we describe them as a set like $(U_1, \phi_1), (U_2, \phi_2), \ldots$.

From now on, we consider the mapping from $M = \mathbb{P}^1 \times \mathbb{R}$ to $\mathbb{R}^2$

$$g : M \ni ([x : y], z) \mapsto (xz, yz) \in \mathbb{R}^2$$

to be written as $g : (x, z) \mapsto (zx, z)$ when $y \neq 0$, and $g : (y, w) \mapsto (w, wy)$ when $x \neq 0$.

At this stage, let's return to the topic of statistics. Suppose we can express the function $K(\theta)$ for the two-dimensional parameter $\theta = (\theta_x, \theta_y)$ as $\theta_x^2 + \theta_y^2$. When

$\theta_y \neq 0$, we can express (θ_x, θ_y) as (zx, z) and when $\theta_x \neq 0$, we can express (θ_x, θ_y) as (w, wy). In each case, we have

$$K(g(x, z)) = K(zx, z) = z^2(1 + x^2) \ , \ K(g(y, z)) = K(w, wy) = w^2(1 + y^2) \ . \tag{1.12}$$

Also, for the first case, we obtain

$$\frac{\partial \theta_x}{\partial x} = z \ , \ \frac{\partial \theta_x}{\partial z} = x \ , \ \frac{\partial \theta_y}{\partial x} = 0 \ , \ \frac{\partial \theta_y}{\partial z} = 1$$

and the Jacobian is z for the first case and it is $-w$ for the second case. The absolute values of these are $|z|$ and $|w|$, respectively.

The precise statement of Hironaka's theorem will be discussed in Chap. 7, but it asserts the existence of a manifold M and a mapping $g : M \to \Theta$ such that, for each local coordinate with local variables $u = (u_1, \ldots, u_d)$,

$$K(g(u)) = u_1^{2k_1} \cdots u_d^{2k_d} \tag{1.13}$$

$$|g'(u)| = b(u)|u_1^{h_1} \cdots u_d^{h_d}|. \tag{1.14}$$

The form of (1.13) is called *normal crossing*. The absolute value of $g'(u)$ in (1.14) is the Jacobian, and $b(u)$ is a function that always takes positive values. In relation to (1.12), if we replace the local variables (x, z) with (x, v) where $v = z\sqrt{1 + x^2}$, we get

$$K(x, v) = v^2$$

$$|g'(x, v)| = \frac{1}{1 + x^2}|v|.$$

In fact, we can calculate

$$\begin{bmatrix} \frac{\partial \theta_x}{\partial x} & \frac{\partial \theta_x}{\partial v} \\ \frac{\partial \theta_y}{\partial x} & \frac{\partial \theta_y}{\partial v} \end{bmatrix} = \begin{bmatrix} \frac{1}{(1 + x^2)^{3/2}} u & \frac{x}{(1 + x^2)^{1/2}} \\ -\frac{x}{(1 + x^2)^{3/2}} u & \frac{1}{(1 + x^2)^{1/2}} \end{bmatrix} .$$

On the other hand, if we set $r = w\sqrt{1 + x^2}$, we obtain $K(y, r) = r^2$ and $|g'(y, r)| = \frac{1}{1+y^2}|r|$.

In Watanabe's Bayesian theory, the operation of finding the normal crossing of $K(\theta)$ for each local coordinate is called "resolving the singularity". In other words, for each local coordinate, we obtain two sequences of non-negative integers of length d, $(k_1, \ldots, k_d)$ and $(h_1, \ldots, h_d)$. We define the value

$$\lambda^{(\alpha)} = \min_{1 \le i \le d} \frac{h_i + 1}{2k_i}$$

as the (local-coordinate-wise) real log canonical threshold, and the number of i's achieving the minimum value as the (local-coordinate-wise) multiplicity, $m^{(\alpha)}$.

Then, the minimum of $\lambda^{(\alpha)}$ over all local coordinates, $\lambda = \min_\alpha \lambda^{(\alpha)}$, is called the real log canonical threshold, and the maximum of $m^{(\alpha)}$ among local coordinates achieving $\lambda^{(\alpha)} = \lambda$ is called the multiplicity.

In the example above, for the first local coordinate corresponding to (x, v), the values (k_i, h_i) are (0, 0) and (2, 1), respectively, resulting in $\lambda^{(\alpha)} = 1/2$ and $m^{(\alpha)} = 1$. The same is true for the other local coordinate. Therefore, $\lambda = 1/2$ and $m = 1$.

That is, we can determine λ and m from $K(\theta)$. At this point, the role of algebraic geometry is finished.

Even when using Hironaka's theorem, we are only calculating the normal crossing for each local coordinate. It may not be impossible to understand Watanabe's Bayesian theory by exploring the relationship between statistical regularity and algebraic geometry's singularity, but this book takes the following position:

> In Watanabe's Bayesian theory,
>
> the regularity of Bayesian statistics and the non-singularity of algebraic geometry are assumed to be unrelated.

In particular, readers of type 2 should read through Chap. 7 and beyond without the preconception that it is "difficult".

1.9 What is the Meaning of Algebraic Geometry's λ in Bayesian Statistics?

The real log canonical threshold λ in algebraic geometry is also called the *learning coefficient* in Watanabe's Bayesian theory.

Readers of types 1 and 2 who have attempted to study Watanabe's Bayesian theory may have various thoughts about the meaning of λ and how to concretely determine it.

In Chap. 9, we prove that when the system is regular, the normal crossing can actually be calculated, and $\lambda = d/2$. We have mentioned that AIC justifies its validity by satisfying (1.10), but this condition does not hold when the system is not regular. In fact, the expected AIC, $\mathbb{E}_{X_1 \cdots X_n}[AIC]$, becomes smaller. However, even in that case, (1.11) holds. If the system is regular, the second term of WAIC becomes on average equal to d, but when not regular, it becomes 2λ. In fact,

$$\mathbb{E}_{X_1 \cdots X_n}[\text{generalization loss}] = \mathbb{E}_{X_1 \cdots X_n}[\text{empirical loss} + 2\lambda$$

holds (Chap. 8).

However, it is difficult to derive the value of the learning coefficient λ mathematically, and only a few cases have been understood so far (Chap. 9). Among them, Miki Aoyagi's analysis of shrinkage rank regression is famous. In this chapter, we hope that more results will follow, and we have included the proof in the appendix. The proof in the original paper is long, so the author has rewritten it to be simpler and easier to understand the essence.

Instead of the learning coefficient itself, there are research results that seek the upper bound of the learning coefficient. Additionally, there is a method to determine the learning coefficient λ from the WBIC value. The WBIC, using our notation so far, is

$$\sum_{i=1}^{n} - \log p(x_i|\theta)$$

averaged over Θ with respect to the posterior distribution. Watanabe's Bayesian theory generalizes the posterior distribution using an inverse temperature $\beta > 0$. In that case, the WBIC is

$$\int_{\Theta} \sum_{i=1}^{n} - \log p(x_i|\theta) p_\beta(\theta|x_1, \ldots, x_n) d\theta \ ,$$

where

$$p_\beta(\theta|x_1, \ldots, x_n) = \frac{\varphi(\theta) \prod_{i=1}^{n} p(x_i|\theta)^\beta d\theta}{\int_\Theta \varphi(\theta') \prod_{i=1}^{n} p(x_i|\theta')^\beta d\theta'} \ .$$

WBIC, when regular, exhibits values similar to BIC, and even when not regular, it calculates values close to the free energy. After all, researchers in statistical physics have been attempting to calculate the free energy for a long time, and it is known that the calculations are extensive. WBIC has value as a means to calculate free energy alone.

If one understands the algebraic geometry in Chap. 7 and the state density formula in Sect. 8.1, the WBIC theory is not as complicated as the WAIC theory.

WAIC and WBIC are bifocal glasses

substitutes for AIC and BIC when not regular

Finally, for those who have experienced frustration and want to conquer the challenging Watanabe's Bayesian theory, we have described the important and hard-to-notice essence in blue throughout each chapter. If this helps reach those hard-to-reach spots, we would be delighted.