Statistical Papers (2023) 64:117–137
https://doi.org/10.1007/s00362-022-01309-9

REGULAR ARTICLE

Convergence of estimative density: criterion for model complexity and sample size

Yo Sheena[1,2]

Received: 2 August 2021 / Revised: 19 March 2022 / Accepted: 20 March 2022 /
Published online: 25 April 2022
© The Author(s) 2022

Abstract

For a parametric model of distributions, the closest distribution in the model to the true distribution located outside the model is considered. Measuring the closeness between two distributions with the Kullback–Leibler divergence, the closest distribution is called the "information projection." The estimation risk of the maximum likelihood estimator is defined as the expectation of Kullback–Leibler divergence between the information projection and the maximum likelihood estimative density (the predictive distribution with the plugged-in maximum likelihood estimator). Here, the asymptotic expansion of the risk is derived up to the second order in the sample size, and the sufficient condition on the risk for the Bayes error rate between the predictive distribution and the information projection to be lower than a specified value is investigated. Combining these results, the "p/n criterion" is proposed, which determines whether the estimative density is sufficiently close to the information projection for the given model and sample. This criterion can constitute a solution to the sample size or model selection problem. The use of the p/n criteria is demonstrated for two practical datasets.

Keywords Kullback–Leibler divergence · Exponential family · Asymptotic risk · Information projection · Predictive density

Mathematics Subject Classification Primary 60F99 · Secondary 62F12

✉ Yo Sheena
yo-sheena@biwako.shiga-u.ac.jp

1 Faculty of Data Science, Shiga University, Hikone, Japan

2 Institute of Statistical Mathematics, Tokyo, Japan

1 Introduction

Given a certain data set, an unknown probability distribution that generates the data as the independent, identically distributed (i.i.d.) sample can be assumed. Under this assumption, if a certain parametric distribution model is adopted to "explain" the data, the first task is to find the "best" approximating distribution in the model. Because the true distribution is assumed to be outside the model (except for some rare cases), the "best" means the "closest" to the true distribution.

Consider the following parametric distribution model:

$$\mathcal{M} = \{g(x;\theta) \,|\, \theta = (\theta^1, \ldots, \theta^p) \in \Theta\},$$

where $g(x;\theta)$ is the probability density function (p.d.f.) with respect to a reference measure $d\mu$ on a measurable space. The p.d.f. of the unknown true distribution with respect to $d\mu$ is denoted by $g(x)$. If we use a certain divergence $D[\cdot \,|\, \cdot]$ to measure the closeness between $g(x)$ and $g(x;\theta)$, then the "best" approximating distribution in $\mathcal{M}$ is given by the predictive distribution $g(x;\theta_*)$, where

$$\theta_* = \underset{\theta\in\Theta}{arg\,min}\, D[g(x) \,|\, g(x;\theta)].$$

Following Csiszár (1975), we will call $g(x;\theta_*)$ the "information projection" in this paper.

Let $\hat{\theta}$ denote the maximum likelihood estimator (MLE) based on the i.i.d. sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $g(x)$. Consider the predictive density $g(x;\hat{\theta})$. Since MLE converges to θ_* in probability (see, e.g., Theorem 5.21 of van der Vaart (1998)) as the sample size, n, increases,

$$D[g(x;\theta_*) \,|\, g(x;\hat{\theta})] \tag{1}$$

also converges to zero in probability. The predictive density $g(x;\hat{\theta})$ is produced with plugged-in MLE. This type of predictive density is called "estimative density". Another common method to formulate the predictive density is Bayesian predictive density. For the asymptotic properties of Bayesian predictive density, see e.g. Komaki (1996), Hartigan (1998), Komaki (2015) and Zhang et al. (2018).

Take the expectation

$$R[g(x;\theta_*) \,|\, g(x;\hat{\theta})] = E\Big[D[g(x;\theta_*) \,|\, g(x;\hat{\theta})]\Big] \tag{2}$$

with respect to the i.i.d. sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $g(x)$. Throughout this study, the expectation under $g(x)$ is denoted by $E[\cdot]$, while the expectation under $g(x;\theta_*)$ is denoted by $E_{\theta_*}[\cdot]$. We call (2) "estimation risk" for discriminating it with the "total risk"

$$R[g(x) \,|\, g(x;\hat{\theta})] = E\Big[D[g(x) \,|\, g(x;\hat{\theta})]\Big].$$

The estimation risk converges to zero under some mild conditions. We will use this estimation risk as the measure of the closeness between $g(x;\theta_*)$ and $g(x;\hat{\theta})$.

Given the data and the model, we need to know whether $g(x;\hat{\theta})$ is sufficiently close to the information projection. Thus, with a certain threshold C, the following criterion is considered.

$$\hat{R}[g(x;\theta_*) \mid g(x;\hat{\theta})] < C, \tag{3}$$

where the left hand side is the estimator of the estimation risk.

This criterion gives a solution to the following two problems.

- Sample size problem: With the model fixed, it indicates exactly how much sample size n is needed for $g(x;\hat{\theta})$ to be close to the information projection. If the criterion is not satisfied, we need to collect more sample.
- Model selection problem: With the sample size fixed, it tells us whether a model is simple enough (especially the dimension of the parameter p is small enough) to guarantee that $g(x;\hat{\theta})$ is close to the information projection. Unless the criterion is satisfied, simplifying the model could be a remedy.

As seen later in the manuscript, the estimation risk is mainly determined by p/n when the information projection is close to the true distribution, and we will call this criterion "p/n criterion" hereafter.

In this paper, as the divergence, Kullback–Leibler divergence is taken, that is,

$$D[g(x) \mid g(x;\theta)] = \int g(x) \log\Big(g(x)/g(x;\theta)\Big) d\mu.$$

Note that for this divergence, the information projection is given by

$$E\left[\frac{\partial}{\partial\theta^i} \log g(X;\theta)\right] = 0, \quad i = 1,\ldots,p, \tag{4}$$

and its solution θ^* is naturally estimated via the MLE, which is the solution of

$$\sum_{t=1}^{n} \frac{\partial}{\partial\theta^i} \log g(X_t;\theta) = 0, \quad i = 1,\ldots,p.$$

For the other divergences, the information projection is more complicated, and its natural estimator is not as simple as MLE.

This paper aims to present a simple and practical criterion (3), and proceeds as follows;

1. The asymptotic expansion of the estimation risk is derived.
2. The asymptotic expansion combined with the estimated moments gives the estimator of the estimation risk.
3. The reasonable (persuasive) threshold C is proposed.

An overview of the contents of each section is now provided. First, the asymptotic expansion of the estimation risk is given for both the general model (Sect. 2.1) and an exponential family model (Sect. 2.2). The estimator of the estimation risk is given in Sect. 2.3. Next, the concrete threshold C is proposed in view of the Bayes error rate. With these results combined, p/n criterion is proposed in an explicit form (Sect. 3.1). As an application of p/n criterion, the bin number problem in a multinomial distribution or a histogram is considered (Sect. 3.2). In Sect. 3.3, the algorithm for calculating the p/n criterion in the case of an exponential family is described. In Sect. 3.4, the use of the p/n criterion is demonstrated for two practical examples.

2 Estimation risk for general case and exponential family

In this section, the asymptotic expansion with respect to n of the estimation risk (2) is presented up to the first-order term for a general distribution, and up to the second-order term for an exponential family distribution.

Hartigan (1998) derives the asymptotic expansion of the estimation risk (2) up to the second order under the assumption $g(x)$ belongs to $\mathcal{M}$. The result here is the extension of his result in the sense that the true distribution is not necessarily located in $\mathcal{M}$.

On the risk of an exponential family, the most relevant work is that of Barron and Sheu (1991). They consider the convergence rate of the K–L divergence (not the risk, but the divergence itself) for an exponential family on a compact set. Their interest lies in the closeness between $g(x)$ and $g(x;\hat{\theta})$, while this research focuses on the closeness between $g(x;\theta_*)$ and $g(x;\hat{\theta})$

2.1 Estimation risk for general case

Taylor expansion of

$$D[g(x;\theta_*)\,|\,g(x;\hat{\theta})] = \int g(x;\theta_*)\log(g(x;\theta_*)/g(x;\hat{\theta}))d\mu$$

as a function of $\hat{\theta}$ around θ_* is considered:

$$\begin{aligned}
&D[g(x;\theta_*)\,|\,g(x;\hat{\theta})]\\
&\quad= -\sum_i \int \frac{\partial}{\partial\theta^i} g(x;\theta)\Big|_{\theta=\theta_*} d\mu\ (\hat{\theta}^i-\theta_*^i)\\
&\qquad+\frac{1}{2}\sum_{i,j}\int g(x;\theta_*)\Big(\frac{\partial}{\partial\theta^i}\log g(x;\theta)\Big|_{\theta=\theta_*}\Big)\Big(\frac{\partial}{\partial\theta^j}\log g(x;\theta)\Big|_{\theta=\theta_*}\Big)d\mu\\
&\qquad\times(\hat{\theta}^i-\theta_*^i)(\hat{\theta}^j-\theta_*^j)\\
&\qquad-\frac{1}{2}\sum_{i,j}\int \frac{\partial^2}{\partial\theta^i\partial\theta^j} g(x,\theta)\Big|_{\theta=\theta_*} d\mu\ (\hat{\theta}^i-\theta_*^i)(\hat{\theta}^j-\theta_*^j)
\end{aligned}$$

$$-\frac{1}{3!}\sum_{i_1,i_2,i_3}\int g(x;\theta_*)\frac{\partial^3 \log g(x,\theta)}{\partial\theta^{i_1}\partial\theta^{i_2}\partial\theta^{i_3}}\bigg|_{\theta=\tilde{\theta}_*} d\mu(\hat{\theta}^{i_1}-\theta_*^{i_1})(\hat{\theta}^{i_2}-\theta_*^{i_2})(\hat{\theta}^{i_3}-\theta_*^{i_3}),$$

where $\tilde{\theta}_*$ is a point between θ_* and $\hat{\theta}$. Because

$$\int \frac{\partial}{\partial\theta^i} g(x;\theta)d\mu = 0, \qquad \int \frac{\partial^2}{\partial\theta^i\partial\theta^j} g(x,\theta)d\mu = 0, \qquad \forall\theta\in\Theta,$$

it turns out that

$$\begin{aligned} R[g(x;\theta_*)\,|\,g(x;\hat{\theta})] = {} & \frac{1}{2}\sum_{i,j} g^*_{ij}(\theta_*)E[(\hat{\theta}^i-\theta_*^i)(\hat{\theta}^j-\theta_*^j)] \\ & -\frac{1}{3!}\sum_{i_1,i_2,i_3} E[\tau_{i_1,i_2,i_3}(\hat{\theta}^{i_1}-\theta_*^{i_1})\cdots(\hat{\theta}^{i_3}-\theta_*^{i_3})]. \end{aligned}$$

Here,

$$\tau_{i_1,i_2,i_3} = \int g(x;\theta_*)\frac{\partial^3 \log g(x,\theta)}{\partial\theta^{i_1}\partial\theta^{i_2}\partial\theta^{i_3}}\bigg|_{\theta=\hat{\theta}_*} d\mu$$

and g^*_{ij} indicates the components of the Fisher metric matrix on $\mathcal{M}$, given by

$$g^*_{ij}(\theta_*) = (G^*(\theta_*))_{ij} = E_{\theta_*}\Big[\Big(\frac{\partial}{\partial\theta^i}\log g(x;\theta)\Big|_{\theta=\theta_*}\Big)\Big(\frac{\partial}{\partial\theta^j}\log g(x;\theta)\Big|_{\theta=\theta_*}\Big)\Big].$$

As θ_* is the solution of equation (4) and $\hat{\theta}$ is its empirical solution (i.e., the M-estimator), the following result holds (see, e.g., Theorem 5.21 of van der Vaart (1998)).

$$\sqrt{n}\big(\hat{\theta}-\theta_*\big) \xrightarrow{d} N_p(0,\tilde{G}^{-1}G\tilde{G}^{-1}),$$

where

$$\begin{aligned} g_{ij}(\theta_*) &= (G(\theta_*))_{ij} = E\Big[\Big(\frac{\partial}{\partial\theta^i}\log g(X;\theta)\Big|_{\theta=\theta_*}\frac{\partial}{\partial\theta^j}\log g(X;\theta)\Big|_{\theta=\theta_*}\Big)\Big], \\ \tilde{g}_{ij}(\theta_*) &= (\tilde{G}(\theta_*))_{ij} = -E\Big[\frac{\partial^2}{\partial\theta^j\partial\theta^i}\log g(X;\theta)\Big|_{\theta=\theta_*}\Big]. \end{aligned}$$

For a general distribution, the estimation risk is asymptotically given as follows;

Theorem 1

$$R[g(x;\theta_*)\,|\,g(x;\hat{\theta})] = (2n)^{-1}\mathrm{tr}\Big(\tilde{G}(\theta_*)^{-1}G(\theta_*)\tilde{G}(\theta_*)^{-1}G^*(\theta_*)\Big) + O(n^{-2}). \quad (5)$$

Because the n^{-2}-order term is prohibitively lengthy, if it is incorporated into the p/n criterion, the result is not suitable for practical use. Hence, it is omitted here. (For interested readers, Theorem 1 of Sheena (2021) is being referred to. You can also find the proof of the whole expansion there.)

Note that, if $g(x)$ exists within the model, then $G = \tilde{G} = G^*$. Hence, the first-order term equals $p/(2n)$ (for more general result for the well-specified model, see Sheena (2018)). Thus, the first-order term is mainly determined by p if $g(x;\theta_*)$ is close to $g(x)$.

2.2 Estimation risk for exponential family

This subsection investigates the estimation risk when the parametric model is an exponential family (for general references on exponential families, see Brown (1986), Barndorff-Nielsen (2014) and Sundberg (2019)). In the case of the exponential family, the n^{-2}-order term in the asymptotic expansion of the estimation risk has a simpler form.

Let the model $\mathcal{M}$ be given by

$$\mathcal{M} = \Big\{ g(x;\theta) = \exp\Big(\sum_{i=1}^{p} \theta^i \xi_i(x) - \Psi(\theta)\Big) \,\Big|\, \theta \in \Theta \Big\}. \tag{6}$$

where $\Psi(\theta)$ is the cumulant-generating function of the ξ terms, such that,

$$\Psi(\theta) = \log \int \exp\Big(\sum_{i=1}^{p} \theta^i \xi_i(x)\Big) d\mu.$$

The "dual coordinate" η is defined as

$$\eta_i(\theta) = \frac{\partial \Psi(\theta)}{\partial \theta^i} = E_\theta[\xi_i], \quad i = 1, \ldots, p.$$

In particular, from the definition of θ_* (see (4)),

$$\eta_i^* = \eta_i(\theta_*) = E_{\theta_*}[\xi_i] = E[\xi_i], \quad i = 1, \ldots, p.$$

The last equation requires the means of ξ_i to coincide under $g(x)$ and $g(x;\theta_*)$. It is known that $g(x;\theta_*)$ maximizes the Shannon entropy among all probability distributions of $(\xi_1, \ldots, \xi_p)$ with a given $E[\xi_i]$, $i = 1, \ldots p$ (the "entropy maximization property" of an exponential family; see, e.g., Wainwright and Jordan (2008)). The K–L divergence is the difference between the cross-entropy and Shannon entropy.

The η coordinate is easily estimated. In fact, $\hat{\eta}$, the MLE for η, is the sample mean of ξ. Hence,

$$\hat{\eta}_i = \frac{\partial \Psi}{\partial \theta_i}(\hat{\theta}) = \bar{\xi}_i \Big(= n^{-1} \sum_{t=1}^{n} \xi_i(X_t)\Big). \tag{7}$$

In contrast, $\hat{\theta}$ is difficult to obtain explicitly because Ψ or its derivative cannot be theoretically obtained for a complex model. This could pose a serious obstacle to application of an exponential family model to a practical problem, and is discussed in Sect. 3.3.

Let the matrix $\ddot{\Psi}(\theta)$ be defined by

$$(\ddot{\Psi}(\theta))_{ij} = \frac{\partial^2 \Psi(\theta)}{\partial\theta^i \partial\theta^j} = E_\theta[(\xi_i - \eta_i)(\xi_j - \eta_j)], \quad 1 \le i, j \le p.$$

Thus, $\ddot{\Psi}$ is a covariance matrix of the ξ_i terms under $g(x;\theta)$; hence, it is positive definite. Therefore, $\Psi(\theta)$ is a convex function. The notable property

$$g^*_{ij}(\theta) = \tilde{g}_{ij}(\theta), \quad 1 \le i, j \le p, \quad \forall\theta$$

is proven by the fact that both sides are equal to $(\ddot{\Psi}(\theta))_{ij}$.

The following notation is used for the third- or fourth-order cumulant:

$$\begin{aligned}
\kappa_{ijk} &= E[(\xi_i - \eta^*_i)(\xi_j - \eta^*_j)(\xi_i - \eta^*_k)] \\
\kappa^*_{ijk} &= E_{\theta_*}[(\xi_i - \eta^*_i)(\xi_j - \eta^*_j)(\xi_i - \eta^*_k)] = \frac{\partial^3 \Psi(\theta_*)}{\partial\theta^i \partial\theta^j \partial\theta^k} \\
\kappa^*_{ijkl} &= E_{\theta_*}[(\xi_i - \eta^*_i)(\xi_j - \eta^*_j)(\xi_i - \eta^*_k)(\xi_i - \eta^*_l)] \\
&\quad - E_{\theta_*}[(\xi_i - \eta^*_i)(\xi_j - \eta^*_j)]E_{\theta_*}[(\xi_k - \eta^*_k)(\xi_l - \eta^*_l)] \\
&\quad - E_{\theta_*}[(\xi_i - \eta^*_i)(\xi_k - \eta^*_k)]E_{\theta_*}[(\xi_j - \eta^*_j)(\xi_l - \eta^*_l)] \\
&\quad - E_{\theta_*}[(\xi_i - \eta^*_i)(\xi_l - \eta^*_l)]E_{\theta_*}[(\xi_j - \eta^*_j)(\xi_k - \eta^*_k)] = \frac{\partial^4 \Psi(\theta_*)}{\partial\theta^i \partial\theta^j \partial\theta^k \partial\theta^l}
\end{aligned}$$

for $1 \le i, j, k, l \le p$.

Next theorem states the asymptotic expansion of the estimation risk for an exponential family distribution. In the case of an exponential family, the second-order term is relatively simple and can be practically used if it is incorporated into the p/n criterion proposed in the next section.

In the theorem, for brevity, Einstein notation is used and the dependency on θ_* is omitted; e.g., G for $G(\theta_*)$ and $\tilde{g}^{ij}$ for $\tilde{g}^{ij}(\theta_*)$.

Theorem 2 *If the parametric model is an exponential family, the estimation risk is given by*

$$\begin{aligned}
R[g(x;\theta_*) \mid g(x;\hat{\theta})] &= \frac{1}{2n}\mathrm{tr}\left(\tilde{G}^{-1}G\right) \\
&+ \frac{1}{24n^2}\Big[-8\tilde{g}^{uk}\tilde{g}^{ls}\tilde{g}^{mt}\kappa_{kst}\kappa^*_{lmu} \\
&+ 9\tilde{g}^{ko}\tilde{g}^{lu}\tilde{g}^{sv}\tilde{g}^{tw}\tilde{g}^{hm}\kappa^*_{lmo}\kappa^*_{sth}(g_{ku}g_{vw} + g_{kv}g_{uw} + g_{kw}g_{uv}) \\
&- 3\tilde{g}^{kw}\tilde{g}^{ls}\tilde{g}^{mu}\tilde{g}^{tv}\kappa^*_{lmtw}(g_{ks}g_{uv} + g_{ku}g_{sv} + g_{kv}g_{su})\Big] + O(n^{-3}).
\end{aligned} \tag{8}$$

Proof The calculation is carried out straightforwardly from the expansion for the general distribution. See Sheena (2021) for the proof. □

The estimation risk up to the second-order term is determined by the moments of the ξ_i terms, g_{ij}, and κ_{ijk} under $g(x)$, as well as their moments under $g(x;\theta_*)$, $\tilde{g}^{ij}$, κ^*_{ijk}, and κ^*_{ijkl}.

2.3 Estimator of estimation risk

We will use Theorem 1 and 2 for the approximation of the estimation risk. In order to establish the criterion (3), we need the estimator of the (approximated) estimation risk. The moments contained in (5) or (8) needs to be estimated; The second moments (Fisher information metric)

$$G^* = (g^*_{ij}), \quad \tilde{G} = (\tilde{g}_{ij}), \quad G = (g_{ij})$$

and cumulant

$$\kappa_{ijk}, \quad \kappa^*_{ijk}, \quad \kappa^*_{ijkl}, \qquad 1 \le i, j, k, l \le p.$$

Naive estimators of these properties (denoted by the "hat" mark: $\hat{G}$, $\hat{\kappa}_{ijk}$, etc.) are gained by replacing θ_* with MLE $\hat{\theta}$, and the expectation $E[\cdot]$ with the empirical mean.

First the estimator of the second moments are given as follows;

$$\begin{aligned}
(\hat{G})_{ij} &= n^{-1}\sum_{t=1}^{n} \frac{\partial}{\partial\theta^i}\log g(X_t;\theta)\Big|_{\theta=\hat{\theta}} \frac{\partial}{\partial\theta^j}\log g(X_t;\theta)\Big|_{\theta=\hat{\theta}} \\
(\hat{\tilde{G}})_{ij} &= -n^{-1}\sum_{t=1}^{n} \frac{\partial^2}{\partial\theta^i\partial\theta^j}\log g(X_t;\theta)\Big|_{\theta=\hat{\theta}} \\
(\hat{G}^*)_{ij} &= \int g(x;\hat{\theta})\Big(\frac{\partial}{\partial\theta^i}\log g(x;\theta)\Big|_{\theta=\hat{\theta}}\Big)\Big(\frac{\partial}{\partial\theta^j}\log g(x;\theta)\Big|_{\theta=\hat{\theta}}\Big)d\mu.
\end{aligned}$$

Now we have the p/n criterion for a general distribution with a given C.

Criterion for a general distribution

$$C \ge \frac{1}{2n}\mathrm{tr}\Big(\hat{\tilde{G}}^{-1}\hat{G}\hat{\tilde{G}}^{-1}\hat{G}^*\Big) \tag{9}$$

Next the criterion for the exponential family is considered. $\hat{G}$ equals the sample covariance matrix of the ξ_i terms, $\hat{\Sigma}$:

$$\hat{G} = \hat{\Sigma}, \quad \hat{g}_{ij} = \big(\hat{\Sigma}\big)_{ij}, \quad \big(\hat{\Sigma}\big)_{ij} = n^{-1}\sum_{t=1}^{n}(\xi_i(X_t) - \bar{\xi}_i)(\xi_j(X_t) - \bar{\xi}_j), \tag{10}$$

where $\bar{\xi}_i = n^{-1}\sum_t \xi_i(X_t)$. Similarly, the estimator of the true third-order cumulant is given by the sample third-order cumulant:

$$\hat{\kappa}_{ijk} = n^{-1}\sum_{t=1}^{n}(\xi_i(X_t)-\bar{\xi}_i)(\xi_j(X_t)-\bar{\xi}_j)(\xi_k(X_t)-\bar{\xi}_k). \tag{11}$$

Further,

$$\hat{\ddot{G}} = \ddot{\Psi}(\hat{\theta}), \quad \hat{\ddot{g}}_{ij} = \big(\ddot{\Psi}(\hat{\theta})\big)_{ij} \tag{12}$$

$$\hat{\kappa}^*_{ijk} = \frac{\partial^3}{\partial\theta^i\partial\theta^j\partial\theta^k}\Psi(\theta)\Big|_{\theta=\hat{\theta}} \tag{13}$$

$$\hat{\kappa}^*_{ijkl} = \frac{\partial^4}{\partial\theta^i\partial\theta^j\partial\theta^k\partial\theta_l}\Psi(\theta)\Big|_{\theta=\hat{\theta}}. \tag{14}$$

Consequently, for an exponential family, the p/n criterion is given as follows.

Criterion for an exponential family

$$\begin{aligned} C \geq{}& \frac{1}{2n}\mathrm{tr}\Big(\hat{\Sigma}(\ddot{\Psi}(\hat{\theta}))^{-1}\Big) \\ &+ \frac{1}{24n^2}\Big[-8\hat{\ddot{g}}^{uk}\hat{\ddot{g}}^{ls}\hat{\ddot{g}}^{mt}\hat{\kappa}_{kst}\hat{\kappa}^*_{lmu} + 9\hat{\ddot{g}}^{ko}\hat{\ddot{g}}^{lu}\hat{\ddot{g}}^{sv}\hat{\ddot{g}}^{tw}\hat{\ddot{g}}^{hm}\hat{\kappa}^*_{lmo}\hat{\kappa}^*_{sth}(\hat{g}_{ku}\hat{g}_{vw}+\hat{g}_{kv}\hat{g}_{uw} \\ &\qquad + \hat{g}_{kw}\hat{g}_{uv}) - 3\hat{\ddot{g}}^{kw}\hat{\ddot{g}}^{ls}\hat{\ddot{g}}^{mu}\hat{\ddot{g}}^{tv}\hat{\kappa}^*_{lmtw}(\hat{g}_{ks}\hat{g}_{uv}+\hat{g}_{ku}\hat{g}_{sv}+\hat{g}_{kv}\hat{g}_{su})\Big]. \end{aligned} \tag{15}$$

How to determine C in (9) or (15) is studied in the next section. Once C is determined, we can use these criterion for the two problems, that is, the sample size problem and the model selection problem, as introduced in Sect. 1.

3 Criterion for model complexity and sample size

In this section, we complete p/n criterion by providing reasonable threshold C for (9) or (15) (Sect. 3.1). As an immediate application of the criterion, we deal with the bin number problem in a multinomial distribution or a histogram (Sect. 3.2). We also state the algorithm for the calculation of the n^{-2}-order term in (15) (Sect. 3.3). In the end, the use of the p/n criterion is demonstrated for two practical examples (Sect. 3.4).

3.1 Choice of threshold

Because the value of the divergence (1) or the risk (2) does not have an absolute standard by itself, we relate it to another reasonable standard. One of the often used measures of the closeness between the two distributions is the error rate, which is more intuitive than the divergence and is suitable for setting a threshold. Let $g_i(x)$, $i = 1, 2$

be the p.d.f. If both $g_i(x),\ i = 1, 2$, are known, the Bayes discriminant rule (with the noninformative prior) is as follows.

For the sample X from either $g_1(x)$ or $g_2(x)$,

$$\frac{g_{i_1}(X)}{g_{i_2}(X)} > 1 \Longleftrightarrow \text{ Judge that } X \text{ is generated from } g(x; \theta_{i_1})$$

The Bayes error rate, Er, i.e., the probability that this rule gives an error, is formally defined by

$$Er[g_1(x) \mid g_2(x)] = \frac{1}{2}\int \min\Big(g_1(x), g_2(x)\Big)d\mu.$$

The next theorem states the relation between Er and the K–L divergence.

Theorem 3 *If* $D[g_1(x) \mid g_2(x)] \leq \delta$, *then*

$$Er[g(x; \theta_1) \mid g(x; \theta_2)] \geq \min\{t \mid (x, t) \in A(\delta)\},$$

where

$$\begin{aligned} A(\delta) = \Big\{(x, t) \Big| & \ x \log\Big(\frac{1-2t}{x} + 1\Big) + (1-x)\log\Big(\frac{2t-1}{1-x} + 1\Big) \\ & = -\delta, \quad 0 < x < 2t < 1\Big\}. \end{aligned}$$

Proof See Appendix. □

Corollary 1 *Let* $\delta = D[g(x; \theta_1) \mid g(x; \theta_2)]$ *and* α *be a certain small positive number (e.g.* $\alpha = 0.05, 0.01$*). If*

$$\min\{t \mid (x, t) \in A(\delta)\} \geq 1/2 - \alpha, \tag{16}$$

then

$$Er[g(x; \theta_1) \mid g(x; \theta_2)] \geq 1/2 - \alpha.$$

Analytical calculation of $\min\{t \mid (x, t) \in A(\delta)\}$ is difficult. The approximation when t is close to $1/2$ is given here. As $\log(1 + x) \doteqdot x - x^2/2$ around $x = 0$,

$$\begin{aligned} & x \log\Big(\frac{1-2t}{x} + 1\Big) + (1-x)\log\Big(\frac{2t-1}{1-x} + 1\Big) \\ & = x\Big(\frac{1-2t}{x}\Big) - \frac{x}{2}\Big(\frac{1-2t}{x}\Big)^2 + (1-x)\frac{2t-1}{1-x} - \frac{(1-x)}{2}\Big(\frac{2t-1}{1-x}\Big)^2 = -\frac{1}{2}\frac{(1-2t)^2}{x(1-x)}. \end{aligned}$$

Therefore, $A(\delta)$ is approximated by

$$A^*(\delta) = \Big\{(x, t) \Big| \ t = \frac{1}{2}\Big(1 - \sqrt{2\delta x(1-x)}\Big), \quad 0 < x < 2t < 1\Big\}.$$

Note that

$$\min\{t \mid (x,t) \in A^*(\delta)\} \geq \min_{0<x<1} \frac{1}{2}\left(1 - \sqrt{2\delta x(1-x)}\right) = \frac{1}{2} - \sqrt{\delta/8},$$

Hence, the condition $\sqrt{\delta/8} \leq \alpha$ or, equivalently, $\delta \leq 8\alpha^2$ is approximately sufficient for (16). Let the solution of δ denoted by C_α for the equation

$$\min\{t \mid (x,t) \in A(\delta)\} = 1/2 - \alpha,$$

or more simply, let C_α be given by

$$C_\alpha = 8\alpha^2. \tag{17}$$

In the latter case, if $\alpha = 0.05(0.01)$, then $C_\alpha = 1/50(1/1250)$. The final form of p/n criterion is given by substituting C in (9) or (15) with C_α.

3.2 p/n Criterion for multinomial distribution

In this section, we present a formula for the bin number of a multinomial distribution using the p/n criterion. The bin number problem in a histogram can be treated similarly. Although several formulas have been proposed on the bin number (or the bin width) in the histogram such as Sturges' formula, Freedman-Diaconis' formula (see the Chapter 3 of Scott (2015)), the formula here is derived from a new perspective.

In view of the true distribution $g(x)$ and the information projection $g(x;\theta_*)$, a multinomial distribution can be seen as the approximation by the step function model. Let

$$\mathcal{M} = \{g(x;m) \mid m = (m^0, \ldots, m^p)\}$$

with

$$g(x;m) = \sum_{i=0}^{p} I(x \in S_i)\frac{m_i}{Vol(S_i)},$$

where $S_i, i = 0, 1, \ldots, p$ is the partition of the range of x with volume

$$Vol(S_i) = \int_{S_i} 1 d\mu(x),$$

and $I(x \in S_i)$ is an indicator function of S_i. In this case, from (4), the information projection $g(x;m^*)$ is given by $m_i^* = P(X \in S_i | g(x))$. The step-function model is not an exponential family. However, we easily notice that Kullback–Leibler divergence between the two step functions (where $d\mu$ is the continuous measure) is equal to the divergence between the two corresponding multinomial distributions (where $d\mu$ is

the counting measure) . Hence, the argument of the estimation risk can be deduced from that of the multinomial distribution model. It is notable that, if X is originally a discrete random variable, the model always contains $g(x)$.

Consider a multinomial distribution with $p+1$ possible values $x_i, i = 0, \ldots, p$, with the corresponding probabilities $m = (m_0, \ldots, m_p)$. This is an exponential family (6), where

$$\theta^i = \log(m_i/m_0), \quad i = 1, \ldots, p,$$
$$\xi_i(x) = \begin{cases} 1, & \text{if } x = x_i, \\ 0, & \text{otherwise,} \end{cases} \quad i = 1, \ldots, p$$

and $d\mu$ is the counting measure on $\{x_1, \ldots, x_p\}$. Here,

$$\Psi(\theta) = \log\Big(\sum_{i=0}^{p} \exp(\theta_i)\Big) = -\log m_0 = -\log\Big(1 - \sum_{i=1}^{p} m_i\Big).$$

The asymptotic expansion of the estimation risk up to the second order can be derived as follows (this corresponds to equation (41) of Sheena (2018) with $\alpha = -1$).

$$R[g(x;\theta) \mid g(x;\hat{\theta})] = \frac{p}{2n} + \frac{1}{12n^2}(M-1) + O(n^{-3}), \qquad M = \sum_{i=0}^{p} {m_i}^{-1}, \quad (18)$$

where $\theta = (m_1, \ldots, m_p)$ is the true-distribution free parameter. Note that if some m_i's are close to zero, the convergence speed reduces considerably.

If we combine the first-order approximation in (18) with the threshold (17), p/n criterion becomes

$$\frac{p}{n} \leq 16\alpha^2.$$

If we adopt $\alpha = 0.05(0.01)$, then the sample size n or the bin number $p+1$ is determined by the formula;

Simple criterion for the sample size or the bin number

$$\frac{p}{n} \leq 1/25(1/625). \quad (19)$$

The second-order approximation gives the following p/n criterion:

$$96n^2\alpha^2 - 6np - (\hat{M} - 1) > 0,$$

where

$$\hat{M} = \sum_{i=0}^{p} \hat{m}_i^{-1}$$

and $\hat{m}_i$ is the MLE, the sample relative frequency, for each i. Applying the criterion for n determination gives the formula

$$n \geq \frac{3p + \sqrt{9p^2 + 96\alpha^2(\hat{M} - 1)}}{96\alpha^2}. \tag{20}$$

In contrast, if the criterion is used for the bin number problem, the formula is given by

$$6np + \hat{M} < 96n^2\alpha^2 + 1.$$

Use of these criteria for practical examples is discussed in Sect. 3.4.

3.3 Algorithm for *p*/*n* criterion of exponential family

This section describes calculation of the right-hand side of (15). If we can calculate the function $\Psi(\theta)$ analytically, the algorithm is simply the following.

Step 1 Calculate $\hat{\eta}_i = \bar{\xi}_i,\ i = 1, \ldots, p$ from the sample.
Step 2 Solve the simultaneous equations w.r.t. θ in (7) to give $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$:

$$\hat{\eta}_i = \eta_i(\hat{\theta}) = \frac{\partial \Psi}{\partial \theta_i}(\hat{\theta}), \qquad i = 1, \ldots, p.$$

Step 3 Calculate (12), (13), and (14) from $\Psi(\hat{\theta})$.
Step 4 Calculate (10) and (11) from the sample.
Step 5 Calculate the right-hand side of (15) and compare it with C_α.

Often, $\Psi(\theta)$ is not explicitly given, especially for a complex model. Then, $\hat{\theta}$ can be iteratively calculated using the Newton--Raphson method with the Jacobian matrix (12). Because $\ddot{\Psi}(\theta)$ is the variance-covariance matrix of the ξ_i terms under the $g(x; \theta)$ distribution, its value can be approximated from the generated sample. The alternative methods are as follows.

Step 2' Iteratively search for $\hat{\theta}$ with

$$\theta^{(n+1)} = \theta^{(n)} - \left(\eta(\theta^{(n)}) - \hat{\eta}\right)\left(\ddot{\Psi}(\theta^{(n)})\right)^{-1},$$

where $\eta(\theta^{(n)})$ and $\ddot{\Psi}(\theta^{(n)})$ are approximated by the sample mean and the sample covariance matrix of the ξ_i terms from the $g(x; \theta^{(n)})$ distribution.

Further, (12), (13), and (14) can also be approximated using the generated sample.

Step 3' Approximate (12), (13), and (14) using the sample moments and cumulants, where the sample is generated from $g(x; \hat{\theta})$.

The point here is that $\Psi(\theta)$ is not required for sample generation in Steps 2' and 3' if methods such as MCMC (requiring no normalizing constant) are used. Although Steps 2' and 3' are computationally heavy tasks, they enable construction of a complex model without calculation of Ψ.

3.4 Real data examples for p/n criterion

This section demonstrates use of the p/n criterion for a particular problem through two practical examples under the exponential family model.

Example 1 (*Red Wine*) The first example is a well-known dataset on wine quality, taken from the U.C.I. Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/wine+quality).

Only red wine data are used. The sample size is 1599, and the variables consist of 11 chemical substances (continuous variables) and "quality" indexes (integers from 3 to 8). The vector of the chemical substances and the "quality" variable are denoted by $x^{(1)} = (x_1^{(1)}, \ldots, x_{11}^{(1)})$ and $x^{(2)}$, respectively. We divided the sample into two halves randomly, one of which ("data_base") was used for the model formulation and the other ("data_est") was used for the estimation of the parameter.

For model formulation, we determined the following: normalization method of the original data, the reference (probability) measure $d\mu(x)$ and ξ elements. Using "data_base", we proceed as;

1. Each variable $x_i^{(1)} (i = 1, \ldots, 11)$ is divided by twice of its maximum such that its range is $[0,\ 1)$. Further, 2 is subtracted from each "quality" index to give a range of $\{1, 2, \ldots, 6\}$.
2. As $d\mu(x)$, 11 independent Beta distributions are applied to $x^{(1)}$ so that their means and variances are equal to those of the "data_base". The multinomial distribution of $x^{(2)}$ is adopted, using each category's sample relative frequency as the category probability parameter (say, $m_i,\ i = 1, \ldots, 6$). In addition, $x^{(1)}$ and $x^{(2)}$ are taken to be independent.

Consequently, $d\mu$ is selected as

$$x = (x^{(1)}, x^{(2)}), \quad d\mu(x) = \prod_{i=1}^{11} {x_i^{(1)}}^{(\beta_{1i}-1)} (1 - x_i^{(1)})^{(\beta_{2i}-1)} d(x^{(1)}) \times \prod_{i=1}^{6} m_i^{I(x^{(2)}=i)} d^*(x^{(2)}),$$

where $d(x^{(1)})$ is the Lebesgue measure on $[0,\ 1]^{11}$, $d^*(x^{(2)})$ is the counting measure on $\{1, 2, \ldots, 6\}$, and $I(\cdot)$ is the indicator function. Further, β_{1i}, β_{2i}, and m_i satisfy the relations

$$\frac{\beta_{1i}}{\beta_{1i} + \beta_{2i}} = \text{Sample mean of } x_i^{(1)}, \quad i = 1, \ldots, 11$$

$$\frac{\beta_{1i}\beta_{2i}}{(\beta_{1i} + \beta_{2i})^2(\beta_{1i} + \beta_{2i} + 1)} = \text{Sample variance of } x_i^{(1)}, \quad i = 1, \ldots, 11$$

$$m_i = \text{Relative frequency of } i \text{ in } x^{(2)}$$

3. The candidate for the ξ_i terms are as follows;

$$\begin{aligned}
&\xi_1(x) = x_1^{(1)}x_2^{(1)}, \quad \xi_2(x) = x_1^{(1)}x_3^{(1)}, \quad \ldots \quad \xi_{10}(x) = x_1^{(1)}x_{11}^{(1)} \\
&\xi_{11}(x) = x_2^{(1)}x_3^{(1)}, \quad \ldots \quad \xi_{19}(x) = x_2^{(1)}x_{11}^{(1)} \\
&\qquad\qquad \ldots \\
&\xi_{55}(x) = x_{10}^{(1)}x_{11}^{(1)}
\end{aligned}$$

and

$$\xi_{56}(x) = x_1^{(1)}x^{(2)}, \quad \ldots \quad \xi_{66}(x) = x_{11}^{(1)}x^{(2)}.$$

Because some of these terms are highly correlated, we eliminate one of the pair with the correlation higher than 0.95. The following 20 ξ_i terms were removed from the full model:

$$\xi_i, \ i = 8, 17, 19, 24, 25, 27, 32, 34, 38, 40, 43, 45, 46, 47, 49, 53, 58, 62, 64.$$

Consequently, an exponential family model with $p = 47$ is formulated. As the probability distribution $g(x;\theta)d\mu$ equals $d\mu$ when the θ terms all equal zero, it is denoted by $g(x;0)$. Note that the $g(x;\theta_*)$ of this model is the closest to $g(x;0)$ in the sense that

$$D[g(x;\theta_*)|g(x;0)] = \min_{h\in\mathcal{H}} D[h(x)|g(x;0)],$$

where $\mathcal{H}$ is the p.d.f. set of $h(x)$ (w.r.t. $d\mu$) that satisfies

$$E_h[\xi_i(X)] = \int h(x)\xi_i(x)d\mu(x) = E[\xi_i(X)],$$

for each ξ_i in the model. This is the consequence of so-called "minimum relative entropy characterization" of an exponential family" (see Csiszár (1975)).

Under the formulated exponential family model, the algorithm in the previous section was implemented and the right-hand side of (15) was calculated using the "data_est", the size of which (n) equals 799. Because of the model complexity, the explicit form of $\Psi(\theta)$ could not be obtained; hence, Alternative Steps 2' and 3' were used. The R and RStan program codes for the whole risk calculation are presented in GitHub (https://github.com/YSheena/P-N_Criteria_Program.git). The first-and second-order terms and the estimation risk in the total of (15) were as follows;

First-order term: 2.95e-02, Second-order term: -1.30e-04, Estimation Risk: 2.93e-02

Note that the second-order term contributes little to the estimation risk; thus, the first-order approximation seems sufficient for this model and data. With the threshold (17), the equation 2.93e-02=$8\alpha^2$ gives the solution $\alpha \doteqdot 0.06$. Hence the Bayes error rate between $g(x;\hat{\theta})$ and $g(x;\theta_*)$ is higher than 0.44. If we set the threshold as

Table 1 Abalones by sex and rings

	1	2	3	4	5	6	7	8	9	10	11	12	13
F	*	*	*	*	4	16	44	122	238	248	200	128	88
I	1	1	12	51	100	216	267	274	173	92	62	21	24
M	*	*	3	6	11	27	80	172	278	294	225	118	91
	14	15	16	17	18	19	20	21	22	23	24	25 $\leq$	
F	56	41	30	26	19	15	12	7	3	6	4	*	
I	14	10	7	7	5	2	2	1	*	*	*	*	
M	56	52	30	25	18	15	12	6	3	3	3	*	

$\alpha = 0.05$, we must trim the model further. For example, if we eliminate one of the ξ elements from the pair with correlation higher than 0.9, then p becomes as small as 37. For this model, the estimation risk is lower than the target value $8 * (0.05)^2 = 0.02$ as follows;

First-order term: 1.60e-02, Second-order term: 2.04e-04, Estimation Risk: 1.62e-02

Example 2 (*Abalone Data*) The next example also features a well-known dataset, in this case, for the physical measurement of abalones (U.C.I. Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Abalone). This data comprise eight properties (sex, length, diameter, etc.) of 4177 abalones. Here, only two discrete variables were considered: "sex" and "ring," where "sex" had three values "Female," "Infant," and "Male"; and "rings" had integer values from 1 to 29. The frequency of each classified group by "sex" and "rings" is given in Table 1. The original frequencies were aggregated at both ends. In the table, if a cell with a star mark is located to the immediate left or right, the number in the cell is aggregated. For example, of the female abalones, cells with 24 or more rings were aggregated to frequency 4. The total number of cells was 63.

A multinomial distribution over 63 cells was considered; hence, $p = 62$. First the simple criterion (19) is adopted, then

$$p/n = 62/4177 \doteqdot 0.015 < 1/25,$$

but $p/n > 1/625$. Consequently, the model distribution is close to the information projection (this case, the true distribution) to the extent that the Bayes error rate is more than 0.45 but less than 0.49.

In order to use the second order term, M needs to be estimated. From the sample relative frequency of each cell $\hat{m}_i$, where $i = 0, \ldots, 62$,

$$\hat{M} = \sum_{i=0}^{62} \hat{m}_i^{-1} = 36128.33,$$

Use of the n formula (20) yielded

$$n \geq 1642,$$

which indicates the actual sample size 4177 is large enough for Bayes error rate 0.45. However, to attain Bayes error rate of 0.49, the required sample size equals 38847, which is far beyond 1642.

Acknowledgements The author greatly appreciates the reviewers' constructive comments on the previous version of the manuscript, which made the present version more concise and readable. This research was partially supported by Grant-in-Aid for Scientific Research (20K11706).

Appendix

Proof of Theorem 3 A suitably fine partition $S_i,\ i = 1, \ldots, m$ of the domain of $d\mu$ and the associated step functions of $g_j(x) = \sum_{i=1}^{m} c_{ji} I(x \in S_i),\ j = 1, 2$ are taken such that the two integrations

$$\begin{aligned} Er[g_1(x) \mid g_2(x)] &= \frac{1}{2} \int \min\Big(g_1(x), g_2(x)\Big) d\mu \\ &= \frac{1}{2} \int g_1(x) \min\Big(1, g_2(x)/g_1(x)\Big) d\mu, \\ D[g_1(x) \mid g_2(x)] &= \int g_1(x) \log\Big(g_1(x)/g_2(x)\Big) d\mu, \end{aligned}$$

are sufficiently well approximated by

$$\frac{1}{2} \sum_{i=1}^{m} \min(1, c_{2i}/c_{1i}) \int_{S_i} c_{1i} d\mu \tag{21}$$

$$\sum_{i=1}^{m} \log(c_{1i}/c_{2i}) \int_{S_i} c_{1i} d\mu, \tag{22}$$

respectively. Furthermore, we can choose the partition such that

$$\int_{S_i} c_{1i} d\mu = 1/m, \quad i = 1, \ldots, m.$$

Then, (21) and (22) equal

$$\frac{1}{2m} \sum_{i=1}^{m} \min(1, \Delta_i)\ (= t(\Delta))$$

$$\frac{1}{m}\sum_{i=1}^{m} -\log \Delta_i,$$

where $\Delta_i = c_{2i}/c_{1i},\ i = 1, \ldots, m$. Suppose that $D[g(X;\theta_1) \mid g(x;\theta_2)] \leq \delta$. Then, with sufficiently finer $S_i,\ i = 1, \ldots, m$, we have

$$f(\Delta) = \frac{1}{m}\sum_{i=1}^{m} \log \Delta_i \geq -\delta. \tag{23}$$

The lower bound of $t(\Delta)$ is searched for, under the condition of (23). Let

$$\tilde{m} = \sum_{i=1}^{m} \Delta_i, \qquad \tilde{1} = \frac{\tilde{m}}{m}. \tag{24}$$

Note that, as the partition $S_i,\ i = 1, \ldots, m$ becomes finer,

$$\sum_{i=1}^{m} \int_{S_i} c_{2i} d\mu = \sum_{i=1}^{m} \Delta_i/m = \tilde{1} \to \int g_2(x) d\mu = 1.$$

Without loss of generality, the following can be assumed:

$$\Delta_1 \geq \cdots \geq \Delta_s > 1 > \Delta_{s+1} \geq \cdots \geq \Delta_m > 0, \quad \exists s (\geq 1).$$

Let $t = m - s$ and

$$\Delta^+ = \frac{1}{s}\sum_{i=1}^{s} \Delta_i, \qquad \Delta^- = \frac{1}{t}\sum_{i=s+1}^{m} \Delta_i.$$

Note that

$$t(\underbrace{\Delta^+, \cdots, \Delta^+}_{s}, \underbrace{\Delta^-, \cdots, \Delta^-}_{t}) = t(\Delta)$$

and, because of the concavity of $f(\Delta)$,

$$f(\underbrace{\Delta^+, \cdots, \Delta^+}_{s}, \underbrace{\Delta^-, \cdots, \Delta^-}_{t}) \geq f(\Delta) \geq -\delta.$$

Therefore, in search of the lower bound of $t(\Delta)$, we must only consider the case where

$$\begin{aligned} &\Delta_1 = \Delta_2 = \cdots = \Delta_s = \Delta^+ > 1, \\ &0 < \Delta_{s+1} = \Delta_{s+t} = \cdots = \Delta_m = \Delta^- < 1, \end{aligned} \tag{25}$$

Under condition (25), the relations (23) and (24) are

$$\frac{1}{m}(s\log\Delta^+ + t\log\Delta^-) \geq -\delta,$$
$$s\Delta^+ + t\Delta^- = \tilde{m},$$

respectively, or equivalently,

$$x\log\Delta^+ + (1-x)\log\Delta^- \geq -\delta, \tag{26}$$
$$x\Delta^+ + (1-x)\Delta^- = \tilde{1}, \tag{27}$$

where

$$0 < x = s/m < 1. \tag{28}$$

Substituting the relation from (27), i.e.,

$$\Delta^- = \frac{\tilde{1} - x\Delta^+}{1-x}$$

into $\Delta^- > 0$ and (26) gives

$$1 < \Delta^+ < \frac{\tilde{1}}{x} \tag{29}$$
$$h(x;\Delta^+) = x\log\Delta^+ + (1-x)\log\left(\frac{\tilde{1} - x\Delta^+}{1-x}\right) \geq -\delta. \tag{30}$$

Furthermore, under condition (25),

$$\begin{aligned} t(\Delta) &= \frac{1}{2m}\sum_{i=1}^{m}\min(1,\Delta_i) \\ &= \frac{1}{2m}(s + t\Delta^-) \\ &= \frac{1}{2}(x + (1-x)\Delta^-) \\ &= \frac{1}{2}\left(\tilde{1} + x(1-\Delta^+)\right) \left(= t(x;\Delta^+)\right) \end{aligned}$$

Consider the minimization of $t(x;\Delta^+)$ under conditions (28), (29), and (30). Notice that

$$\frac{d}{dx}h(x;\Delta^+) = h'(x;\Delta^+) = \log\Delta^+ - \log\left(\frac{\tilde{1} - x\Delta^+}{1-x}\right) + (1-x)\left\{\frac{-\Delta^+}{\tilde{1} - x\Delta^+} + \frac{1}{1-x}\right\}$$

$$= \log\left(\frac{\Delta^+(1-x)}{\tilde{\mathrm{I}} - x\Delta^+}\right) + \frac{\tilde{\mathrm{I}} - \Delta^+}{\tilde{\mathrm{I}} - x\Delta^+}$$
$$\leq \frac{\Delta^+ - \tilde{\mathrm{I}}}{\tilde{\mathrm{I}} - x\Delta^+} + \frac{\tilde{\mathrm{I}} - \Delta^+}{\tilde{\mathrm{I}} - x\Delta^+} = 0 \qquad (\because \log(1+x) \leq x).$$

Since

$$x < \frac{\tilde{\mathrm{I}}}{\Delta^+} = x + (1-x)\frac{\Delta^-}{\Delta^+} < 1,$$

and

$$\lim_{x \to \tilde{\mathrm{I}}/\Delta^+} h(x; \Delta^+) = -\infty,$$

the minimum value of $t(x; \Delta^+)$(say, t^*) is attained when (30) holds with the equation. Let (x^*, Δ^+_*) denote the point that attains t^*; then,

$$\Delta^+_* = (\tilde{\mathrm{I}} - 2t^*)/x^* + 1. \tag{31}$$

Inserting (31) into the left-hand side of (30) and equating it with $-\delta$ gives

$$x^* \log\left(\frac{\tilde{\mathrm{I}} - 2t^*}{x^*} + 1\right) + (1-x^*)\log\left(\frac{2t^* - 1}{1-x^*} + 1\right) = -\delta,$$

while, from (28), (29), and (31),

$$0 < x^* < 2t^* < \tilde{\mathrm{I}}.$$

Let us define the region $\tilde{A}(\delta)$ by

$$\tilde{A}(\delta) = \Big\{(x^*, t^*)$$
$$\Big| x^* \log\left(\frac{\tilde{\mathrm{I}} - 2t^*}{x^*} + 1\right) + (1-x^*)\log\left(\frac{2t^* - 1}{1-x^*} + 1\right) = -\delta, \quad 0 < x^* < 2t^* < \tilde{\mathrm{I}}.\Big\}$$

Then,

$$\frac{1}{2m}\sum_{i=1}^{m} \min(1, \Delta_i) = t(x; \Delta^+) \geq \min\{t^* \mid (x^*, t^*) \in \tilde{A}(\delta)\}.$$

Taking the limit operation for both sides as the partition becomes finer gives the result.

References

Barndorff-Nielsen OE (2014) Information and exponential families in statistical theory. Wiley, New York
Barron AR, Sheu C (1991) Approximation of density functions by sequences of exponential families. Ann Stat 19(3):1347–1369
Brown LD (1986) Fundamentals of statistical exponential families. IMS
Csiszár I (1975) I-divergence geometry of probability distributions and minimization problems. Ann Probab 3:146–158
Hartigan JA (1998) The maximum likelihood prior. Ann Stat 26(6):2083–2103
Komaki F (1996) On asymptotic properties of predictive distributions. Biometrika 83(2):299–313
Komaki F (2015) Asymptotic properties of Bayesian predictive densities when the distributions of data and target variables are different. Bayesian Anal 10(1):31–51
Scott DW (2015) Multivariate density estimation. Wiley, New York
Sheena Y (2018) Asymptotic expansion of the risk of maximum likelihood estimator with respect to α-divergence as a measure of the difficulty of specifying a parametric model. Commun Stat Theory Methods 47(16):4059–4087
Sheena Y (2021) Mle convergence speed to information projection of exponential family: criterion for model dimension and sample size—complete proof version–. arXiv, 2105.08947
Sundberg R (2019) Statistical modeling for exponential families. Cambridge University Press, Cambridge
van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Wainwright MJ, Jordan MI (2008) Graphical models, exponential families, and variational inference. Now Publishers
Zhang F, Shi Y, Ng HKT, Wang R (2018) Information geometry of generalized Bayesian prediction using α-divergence as loss functions. IEEE Trans Inf Theory 64(3):1812–1824

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.