MATHEMATICAL ENGINEERING TECHNICAL REPORTS

Relations Between the Conditional Normalized Maximum Likelihood Distributions and the Latent Information Priors

Mutsuki KOJIMA and Fumiyasu KOMAKI

METR 2014-40 December 2014

Relations Between the Conditional Normalized Maximum Likelihood Distributions and the Latent Information Priors

Mutsuki KOJIMA* and Fumiyasu KOMAKI*†

*Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo, Tokyo, Japan
†RIKEN Brain Science Institute, Wako-shi, Japan
{mutsuki_kojima,komaki}@mist.i.u-tokyo.ac.jp

December 2014

Abstract

We reveal the relations between the conditional normalized maximum likelihood (CNML) distributions and Bayesian predictive densities based on the latent information priors (LIPs). In particular, CNML3, which is one type of CNML distributions, is investigated. The Bayes projection of a predictive density, which is an information projection of the predictive density on a set of Bayesian predictive densities, is considered. We prove that the sum of the Bayes projection divergence of CNML3 and the conditional mutual information is asymptotically constant. This result implies that the Bayes projection of CNML3 (BPCNML3) is asymptotically identical to the Bayesian predictive density based on LIP. In addition, under some stronger assumptions, we show that BPCNML3 exactly coincides with the Bayesian predictive density based on LIP.

Keywords: Bayes projection, conditional mutual information, Kullback–Leibler divergence, least favorable prior, regret, Rényi divergence

1 Introduction

We construct predictive densities for future variables based on observed data. Let $(X, \mathcal{F})$ be a measurable space and let $\mathcal{M} = \{p(x|\theta)|x \in X, \theta \in \Theta \subset \mathbf{R}^d\}$ be a statistical model, where $p(x|\theta)$ is the probability density function with respect to a σ-finite measure μ on $(X, \mathcal{F})$. We assume that observations $x^N := (x_1, \ldots, x_N)^\top \in X^N$ and future variables $y^M := (y_1, \ldots, y_M)^\top \in X^M$ are independent and identically distributed random variables with probability distribution $\mathcal{M}$. Thus, the joint probability density function of x^N and y^M is

$$p(x^N, y^M|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \prod_{j=1}^{M} p(y_j|\theta).$$

A predictive density $q(y^M|x^N)$ is a conditional probability density, i.e., a function from $X^N \times X^M$ to $\mathbf{R}_+$ satisfying $\int_{X^M} d\mu(y^M) q(y^M|x^N) = 1$. The goodness of prediction fit of $q(y^M|x^N)$ is evaluated

by the average Kullback–Leibler divergence (simply referred to as *KL risk* in this paper) :

$$R_{\mathrm{KL}}^{N,M}(\theta, q) := \int_{X^N} \mathrm{d}\mu(x^N) p(x^N|\theta) \int_{X^M} \mathrm{d}\mu(y^M) p(y^M|\theta) \log \frac{p(y^M|\theta)}{q(y^M|x^N)}.$$

In information theory, the Bayes risk

$$R_{\mathrm{KL}}^{N,M}(\pi, p_\pi) := \int_\Theta \mathrm{d}\pi(\theta) R_{\mathrm{KL}}^{N,M}(\theta, p_\pi),$$

is called conditional mutual information when $N > 0$ (Cover and Thomas, 2006). Latent information priors (LIPs) are defined as prior distributions on Θ that maximize the conditional mutual information, see Komaki (2011). Bayesian predictive densities based on LIPs are minimax predictive densities under KL risk when $\mathcal{M}$ is a submodel of the multinomial distribution. The LIPs are different from Jeffreys priors in general. In addition, when $N > 0$ and the model is a joint location and scale model, we note that the minimax predictive densities under KL risk do not have to match the Bayesian predictive densities based on Jeffreys priors as shown by Liang and Barron (2004).

On the other hand, in the context of information-theoretic learning, the normalized maximum likelihood (NML) distributions, introduced by Shtarkov (1987), are important predictive densities with no observation ($N = 0$). The NML distribution is defined by

$$q_{\mathrm{NML}}(y^M) := \frac{p(y^M|\hat{\theta}(y^M))}{\int_{X^M} \mathrm{d}\mu(z^M) p(z^M|\hat{\theta}(z^M))},$$

where $\hat{\theta}(z^M) := \mathrm{argmax}_\theta\, p(z^M|\theta)$. Shtarkov (1987) showed that the NML distribution achieves the minimax regret:

$$q_{\mathrm{NML}} = \underset{q}{\mathrm{argmin}}\, \underset{y^M}{\max}\{-\log q(y^M) - (-\log p(y^M|\hat{\theta}(y^M)))\}.$$

However, NML distributions have a serious problem that the normalizing constants diverge to infinity even if $\mathcal{M}$ is a simple statistical model such as the normal, Poisson, or geometric distribution. To remedy the problem, Grünwald (2007) proposed three types of generalizations of NML distributions called conditional normalized maximum likelihood (CNML) distributions:

$$q_{\mathrm{CNML1}}(y^M|x^N) := \frac{p(x^N, y^M|\hat{\theta}(y^M))}{\int_{X^M} \mathrm{d}\mu(z^M) p(x^N, z^M|\hat{\theta}(z^M))},$$

$$q_{\mathrm{CNML2}}(y^M|x^N) := \frac{p(x^N, y^M|\hat{\theta}(x^N, y^M))}{\int_{X^M} \mathrm{d}\mu(z^M) p(x^N, z^M|\hat{\theta}(x^N, z^M))},$$

$$q_{\mathrm{CNML3}}(y^M|x^N) := \frac{p(y^M|x^N, \hat{\theta}(x^N, y^M))}{\int_{X^M} \mathrm{d}\mu(z^M) p(z^M|x^N, \hat{\theta}(x^N, z^M))},$$

where $\hat{\theta}(x^N, z^M) := \mathrm{argmax}_\theta\, p(x^N, z^M|\theta)$. By conditioning on observations x^N, the normalizing constants of CNML distributions do not diverge to infinity, and the distributions are defined as predictive densities with some observations ($N > 0$). As with the NML distribution, CNML-i ($i = 1, 2, 3$) achieves the minimax conditional regret-i ($i = 1, 2, 3$):

$$q_{\mathrm{CNML1}} = \underset{q}{\mathrm{argmin}}\, \underset{y^M}{\max}\{-\log q(y^M|x^N) - (-\log p(x^N, y^M|\hat{\theta}(y^M)))\},$$

$$q_{\mathrm{CNML2}} = \underset{q}{\mathrm{argmin}}\, \underset{y^M}{\max}\{-\log q(y^M|x^N) - (-\log p(x^N, y^M|\hat{\theta}(x^N, y^M)))\},$$

$$q_{\mathrm{CNML3}} = \underset{q}{\mathrm{argmin}}\, \underset{y^M}{\max}\{-\log q(y^M|x^N) - (-\log p(y^M|x^N, \hat{\theta}(x^N, y^M)))\}.$$

Our results are twofold. First, we show that the sum of the Bayes projection divergence of CNML3 and the conditional mutual information is asymptotically constant. The Bayes projection of a predictive density is an information projection, a generalization of the information projection studied by Csiszár (1975), of the predictive density on a set of Bayesian predictive densities (see Section 2). Throughout the paper, "asymptotic" means that the number of observations, N, is fixed, and the number of future variables, M, goes to infinity. Roughly speaking, the first result implies that the Bayes projection of CNML3 (BPCNML3) is asymptotically identical to the Bayesian predictive density based on LIP. Second, under some stronger assumptions, we show that the BPCNML3 exactly coincides with the Bayesian predictive density based on LIP. These results indicate that CNML3 is related to LIPs.

Among CNML distributions, CNML2 has received much attention (Kotłowski and Grünwald, 2011; Hedayati and Bartlett, 2012a,b; Bartlett et al., 2013; Harremoës, 2013), and it has been recognized as the only natural generalization of NML distributions (Grünwald, 2012). Grünwald (2007) showed that CNML1 and CNML2 are asymptotically equal to the Bayesian predictive density based on Jeffreys prior. Under some regularity conditions, Hedayati and Bartlett (2012a) showed that CNML2 is identical to the Bayesian predictive density based on Jeffreys prior even when M is finite. Because of the connection with Jeffreys prior, CNML2 is considered to be the most important predictive density among CNML distributions.

However, we argue that LIPs, not Jeffreys priors, are naturally related to minimax predictive densities under the conditional regret when $N > 0$. The reason is as follows. The regret and Kullback–Leibler divergence are widely known to be naturally related in the sense that they are special versions of the Rényi divergence (Rényi, 1961; van Erven and Harremoës, 2014). Notably, when $N = 0$ and statistical model $\mathcal{M}$ is the multinomial distribution, Xie and Barron (2000) showed that a Bayesian predictive density based on a modification of Jeffreys prior asymptotically achieves the minimax regret. When $N = 0$ and the model satisfies some regularity conditions, Clarke and Barron (1994) showed that Jeffreys prior is asymptotically least favorable under KL risk. Roughly speaking, when $N = 0$, Bayesian predictive densities based on Jeffreys priors are asymptotically minimax under both the regret and KL risk. In addition, the NML distribution is known to asymptotically coincide with the Bayesian predictive density based on Jeffreys prior (Grünwald, 2007). These studies imply that least favorable priors under KL risk are connected with minimax predictive densities under the regret when $N = 0$. Therefore, as is the case for $N = 0$, we insist that LIPs are naturally related to minimax predictive densities under the conditional regret because LIPs are least favorable priors under KL risk when $N > 0$.

Our results shed light on the connection between LIPs and CNML3. Although CNML2 has received the most attention among CNML distributions, we consider that CNML3, not CNML2, is more in line with the minimax KL risk approach and is the most important predictive density among CNML distributions. Notably, Grünwald (2007) also vaguely suggested that CNML3 is more in line with the minimax KL risk approach (called Liang and Barron's approach (Liang and Barron, 2004) in his book (Grünwald, 2007)) than CNML1 and CNML2.

The remainder of this paper is organized as follows. In Section 2, we define the Bayes projection of predictive densities and review the definition and properties of LIPs. In Section 3, we state the main results. In Section 4, we confirm that the main results hold for the binomial distributions through numerical experiments. In Section 5, we conclude our study.

2 Preliminaries

Let K be a compact set of Θ and $\mathcal{P}_K$ be the set of all probability measures on Θ whose support sets are contained in K. We assume that $\mathcal{P}_K$ is endowed with the weak convergence topology and the corresponding Borel sigma algebra. By the Prokhorov theorem, $\mathcal{P}_K$ is compact.

2.1 Bayes Projection of Predictive Densities

We define the projection of predictive densities on a set of Bayesian predictive densities. Let $D_{K,q}^{N,M}(\pi)$ be a divergence from Bayesian predictive density based on π to predictive density q:

$$D_{K,q}^{N,M}(\pi) := \int_{X^N \times X^M} d\mu(x^N, y^M) p_\pi(x^N, y^M) \log \frac{p_\pi(x^N, y^M)}{q(y^M|x^N) p_\pi(x^N)}, \quad \pi \in \mathcal{P}_K,$$

where

$$p_\pi(x^k) := \int_\Theta d\pi(\theta) p(x^k|\theta).$$

Divergence $D_{K,q}^{N,M}$ is convex with respect to π. Let π_1 and π_2 in $\mathcal{P}_K$ and $w \in (0,1)$. We define $\pi_w := w\pi_1 + (1-w)\pi_2$. By the log sum inequality,

$$p_{\pi_w}(x^N, y^M) \log \frac{p_{\pi_w}(x^N, y^M)}{q(y^M|x^N) p_{\pi_w}(x^N)}$$
$$\leq w p_{\pi_1}(x^N, y^M) \log \frac{p_{\pi_1}(x^N, y^M)}{q(y^M|x^N) p_{\pi_1}(x^N)} + (1-w) p_{\pi_2}(x^N, y^M) \log \frac{p_{\pi_2}(x^N, y^M)}{q(y^M|x^N) p_{\pi_2}(x^N)}.$$

Therefore,

$$D_{K,q}^{N,M}(w\pi_1 + (1-w)\pi_2) \leq w D_{K,q}^{N,M}(\pi_1) + (1-w) D_{K,q}^{N,M}(\pi_2), \quad w \in (0,1).$$

Since $\mathcal{P}_K$ is compact, if map $\mathcal{P}_K \ni \pi \mapsto D_{K,q}^{N,M}(\pi) \in \mathbf{R}$ is strictly convex and lower semicontinuous, then there exists unique minimizer $\hat{\pi}_{K,q}^{N,M} \in \mathcal{P}_K$ such that

$$D_{K,q}^{N,M}(\hat{\pi}_{K,q}^{N,M}) = \inf_{\pi \in \mathcal{P}_K} D_{K,q}^{N,M}(\pi).$$

We refer to the Bayesian predictive density based on $\hat{\pi}_{K,q}^{N,M}$ as *Bayes projection* of q.

Komaki (2011) showed that KL risk of the Bayes projection of q is not larger than that of q if the statistical model is a submodel of the multinomial distribution.

2.2 Latent Information Priors

In information theory, the Bayes risk

$$R_{\mathrm{KL}}^{N,M}(\pi, p_\pi) := \int_\Theta d\pi(\theta) R_{\mathrm{KL}}^{N,M}(\theta, p_\pi),$$

is called mutual information when $N = 0$ and conditional mutual information when $N > 0$ (Cover and Thomas, 2006). The conditional mutual information is concave with respect to $\pi \in \mathcal{P}_K$. LIPs are defined as priors that maximize the conditional mutual information:

$$\hat{\pi}_{K,\mathrm{LIP}}^{N,M} := \underset{\pi \in \mathcal{P}_K}{\operatorname{argmax}} \, R_{\mathrm{KL}}^{N,M}(\pi, p_\pi).$$

Since $\mathcal{P}_K$ is compact, if map $\mathcal{P}_K \ni \pi \mapsto R_{\mathrm{KL}}^{N,M}(\pi, p_\pi) \in \mathbf{R}$ is strictly concave and upper semicontinuous, then $\hat{\pi}_{K,\mathrm{LIP}}^{N,M}$ is the unique maximizer.

Because LIPs are the least favorable priors (Ferguson, 1967), the Bayesian predictive densities based on LIPs are naturally related to minimax predictive densities under KL risk. Notably, Komaki (2011) showed that Bayesian predictive densities based on LIPs are minimax predictive densities under KL risk when $\mathcal{M}$ is a submodel of the multinomial distribution.

3   Main Results

Before showing the main results, we give basic assumptions and notations.

We assume that a maximum likelihood estimator (MLE) $\hat{\theta}(z^k) \in \Theta$ exists for all $k \in \mathbf{N}$ and $z^k \in X^k$. We take a compact set K contained in the interior of Θ such that $p(z|\theta)$ is strictly positive for all $z \in X$ and $\theta \in K$ and take a positive constant δ such that $K_\delta = \{\tilde{\theta} \in \Theta | \exists \theta \in K \text{ s.t. } |\theta - \tilde{\theta}| \leq \delta\}$ is also contained in the interior of Θ. Here, $|\theta|$ denotes the Euclidean norm. We denote probabilities of events and expectations of random variables by $P_\theta(\cdot)$ and $\mathrm{E}_\theta(\cdot)$, respectively.

We state conditions and lemmas required to prove the main results.

A1. For all $z \in X$, the log-likelihood function $\log p(z|\theta)$ is Lipschitz continuous on K_δ, i.e., there exists a measurable map $L_{K_\delta} : X \to \mathbf{R}_+$ and $1 \leq p \leq \infty$ such that for all $\theta_1, \theta_2 \in K_\delta$

$$\left| \log p(z|\theta_1) - \log p(z|\theta_2) \right| \leq L_{K_\delta}(z)|\theta_1 - \theta_2|,$$

where $L_{K_\delta}(\cdot)$ satisfies

$$\sup_{\theta \in K} \int_X \mathrm{d}\mu(z) p(z|\theta) \{L_{K_\delta}(z)\}^p < \infty.$$

We define $\{L_{K_\delta}\}^\infty := \operatorname{ess\,sup}_{z \in X} L_{K_\delta}(z)$.

A2.

$$\lim_{k \to \infty} \sup_{\theta \in K} \int_{X^k} \mathrm{d}\mu(z^k) p(z^k|\theta) \big| \hat{\theta}(z^k) - \theta \big|^q = 0,$$

where $q \geq 1$ satisfies $1/p + 1/q = 1$ ($q = \infty$ when $p = 1$ and $q = 1$ when $p = \infty$).

A3. There exists a measurable map $T_K : X \to \mathbf{R}_+$ and $1 < r \leq \infty$ such that

$$\sup_{\theta \in K} \{ \log p(z|\hat{\theta}(z)) - \log p(z|\theta) \} \leq T_K(z),$$

and T_K satisfies

$$\sup_{\theta \in K} \int_X \mathrm{d}\mu(z) p(z|\theta) \{T_K(z)\}^r < \infty.$$

A4.

$$\lim_{k \to \infty} \sup_{\theta \in K} \left| \int_{X^k} \mathrm{d}\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\hat{\theta}(z^k))}{p(z^k|\theta)} - \frac{d}{2} \right| = 0.$$

A5. There exist constants $C^{N,M}$ that do not depend on θ such that

$$\lim_{M \to \infty} \sup_{\theta \in K} \left| \int_{X^N} \mathrm{d}\mu(x^N) p(x^N|\theta) \log \left(\int_{X^M} \mathrm{d}\mu(y^M) p(y^M|\hat{\theta}(x^N, y^M)) \right) - C^{N,M} \right| = 0.$$

Remark 1. The integrand in condition A4 is known as the likelihood ratio statistic. The likelihood ratio statistic is widely known to converge in distribution to the chi-squared distribution with degrees of freedom $d/2$ under some mild conditions (Wilks, 1938). Because the mean of the chi-squared distribution is $d/2$, condition A4 is considered to be satisfied for many regular statistical models. However, except for Clarke and Barron (1989), we are not aware of studies about conditions on the L^1 convergence of the likelihood ratio statistic.

Lemma 1. *Assume that statistical model $\mathcal{M}$ satisfies condition A2. Then,*

$$\lim_{k \to \infty} \sup_{\theta \in K} P_\theta \left(\left\{ \hat{\theta}(z^k) \notin K_\delta \right\} \right) = 0.$$

Proof. By the Markov and Hölder inequalities, for all $\theta \in K$,

$$P_\theta \left(\left\{ \hat{\theta}(z^k) \notin K_\delta \right\} \right) \leq P_\theta(|\hat{\theta}(z^k) - \theta| > \delta) \leq \frac{1}{\delta} \left\{ \mathrm{E}_\theta(|\hat{\theta}(z^k) - \theta|^q) \right\}^{\frac{1}{q}}.$$

Since condition A2 is satisfied, the claim is verified. $\square$

Lemma 2. *Assume that conditions A1–A4 are satisfied. Then,*

$$\lim_{M \to \infty} \sup_{\theta \in K} \left| \int_{X^N \times X^M} \mathrm{d}\mu(x^N, y^M) \; p(x^N, y^M | \theta) \log \frac{p(y^M | \theta)}{p(y^M | \hat{\theta}(x^N, y^M))} + \frac{d}{2} \right| = 0.$$

Proof. See Appendix. $\square$

We state our first result.

Theorem 1. *Let K be a compact set that is contained in the interior of Θ and assume that $p(z|\theta)$ is strictly positive for all $z \in X$ and $\theta \in K$. Assume also that conditions A1–A5 are satisfied.*
Then,

$$\lim_{M \to \infty} \sup_{\pi \in \mathcal{P}_K} \left| D_{K, q_{\mathrm{CNML3}}}^{N,M}(\pi) + R_{\mathrm{KL}}^{N,M}(\pi, p_\pi) - \tilde{C}^{N,M} \right| = 0, \tag{1}$$

where $\tilde{C}^{N,M} = C^{N,M} - d/2$ that does not depend on the choice of π.

By deforming (1), we have

$$D_{K, q_{\mathrm{CNML3}}}^{N,M}(\pi) = -R_{\mathrm{KL}}^{N,M}(\pi, p_\pi) + \tilde{C}^{N,M} + o(1), \tag{2}$$

where term $o(1)$ satisfies $\lim_{M \to \infty} \sup_{\pi \in \mathcal{P}_K} |o(1)| = 0$.

Asymptotically, in the right-hand side of (2), only the first term $R_{\mathrm{KL}}^{N,M}(\pi, p_\pi)$ depends on the choice of π. Therefore, the LIP that maximizes $R_{\mathrm{KL}}^{N,M}(\pi, p_\pi)$ with respect to $\pi \in \mathcal{P}_K$ asymptotically coincides with the minimizer of the left-hand side of (2), i.e., $\hat{\pi}_{K, q_{\mathrm{CNML3}}}^{N,M}$. In other words, roughly speaking, BPCNML3 is asymptotically identical to the Bayesian predictive density based on the LIP. Notably, BPCNML3 is different from CNML3. Later, under some stronger conditions, we will show that BPCNML3 exactly coincides with the Bayesian predictive density based on the LIP even when M is finite (see Theorem 2).

Proof of Theorem 1.

$$D_{K,q_{\mathrm{CNML3}}}^{N,M}(\pi) = \int_{\Theta \times X^N \times X^M} \mathrm{d}\pi(\theta)\mathrm{d}\mu(x^N, y^M)p(x^N, y^M|\theta) \log \frac{p_\pi(y^M|x^N)}{q_{\mathrm{CNML3}}(y^M|x^N)}$$

$$= -\int_{\Theta \times X^N \times X^M} \mathrm{d}\pi(\theta)\mathrm{d}\mu(x^N, y^M)p(x^N, y^M|\theta) \log \frac{p(y^M|\theta)}{p_\pi(y^M|x^N)}$$

$$+ \int_{\Theta \times X^N \times X^M} \mathrm{d}\pi(\theta)\mathrm{d}\mu(x^N, y^M)p(x^N, y^M|\theta) \log \frac{p(y^M|\theta)}{q_{\mathrm{CNML3}}(y^M|x^N)}.$$

The first term is $-R_{\mathrm{KL}}^{N,M}(\pi, p_\pi)$. The second term is decomposed as

$$\int_{\Theta \times X^N \times X^M} \mathrm{d}\pi(\theta)\mathrm{d}\mu(x^N, y^M)p(x^N, y^M|\theta) \log \frac{p(y^M|\theta)}{q_{\mathrm{CNML3}}(y^M|x^N)}$$

$$= \int_\Theta \mathrm{d}\pi(\theta)\left\{ \int_{X^N \times X^M} \mathrm{d}\mu(x^N, y^M)p(x^N, y^M|\theta) \log \frac{p(y^M|\theta)}{p(y^M|\hat{\theta}(x^N, y^M))} + \frac{d}{2} \right\}$$

$$+ \int_\Theta \mathrm{d}\pi(\theta)\left\{ \int_{X^N} \mathrm{d}\mu(x^N)\, p(x^N|\theta) \log \left(\int_{X^M} \mathrm{d}\mu(z^M)p(z^M|\hat{\theta}(x^N, z^M)) \right) - C^{N,M} \right\}$$

$$+ C^{N,M} - \frac{d}{2}.$$

By Lemma 2 and assumption A5, we have

$$\int_{\Theta \times X^N \times X^M} \mathrm{d}\pi(\theta)\mathrm{d}\mu(x^N, y^M)p(x^N, y^M|\theta) \log \frac{p(y^M|\theta)}{q_{\mathrm{CNML3}}(y^M|x^N)} = C^{N,M} - \frac{d}{2} + o(1),$$

where term $o(1)$ satisfies $\lim_{M\to\infty} \sup_{\pi \in \mathcal{P}_K} |o(1)| = 0$. Therefore, the claim is verified. $\qquad\square$

We give some examples that satisfy conditions A1–A5.

Example 1 (Multinomial Distributions). The first example is the multinomial distribution. Let $X = \{0, 1, \ldots, d\}$ and $\Theta = \{(p_1, \ldots, p_d)|0 \le p_i \le 1\ (i = 1, \ldots, d),\ \sum_{i=1}^d p_i \le 1\}$. We take a compact set K that is contained in the interior of Θ:

$$K \subset \left\{\theta = (p_1, \ldots, p_d)|0 < p_i < 1\ (i = 1, \ldots, d),\ \sum_{i=1}^d p_i < 1\right\}.$$

Since K is contained in the interior of Θ, we can find $\delta > 0$ such that compact set K_δ is also in the interior of Θ.

The probability function is

$$p(z|\theta) = \prod_{i=0}^d p_i^{z^{(i)}}, \quad z = (z^{(0)}, \ldots, z^{(d)})^\top \in \{0, 1\}^{d+1}, \quad p_0 := 1 - \sum_{i=1}^d p_i,$$

where we identify elements in X with $z = (z^{(0)}, \ldots, z^{(d)})^\top \in \{0, 1\}^{d+1}$ satisfying $\sum_{i=0}^d z^{(i)} = 1$. Since there exists a positive constant c_K such that $\inf_{\theta \in K} \min_{i=0,1,\ldots,d} p_i \ge c_K > 0$,

$$\sup_{\theta \in K}\{\log p(z|\hat{\theta}(z)) - \log p(z|\theta)\} \le \log 1 - \inf_{\theta \in K} \log p(z|\theta) \le -\log c_K.$$

Similarly, there exists a positive constant $c_{K_\delta} > 0$ such that $\inf_{\theta \in K_\delta} \min_{i=0,1,\dots,d} p_i \geq c_{K_\delta}$. By the mean value theorem, for all $\theta_1, \theta_2 \in K_\delta$ and $z \in X$,

$$\big| \log p(z|\theta_1) - \log p(z|\theta_2) \big| \leq \frac{1}{c_{K_\delta}} |\theta_1 - \theta_2|.$$

Therefore, condition A1 and A3 with any $p \in [1, \infty]$ and $r = \infty$ are satisfied. The MLE of the multinomial distribution is

$$\hat{\theta}(z^n) = \left(\frac{\sum_{i=1}^{n} z_i^{(1)}}{n}, \dots, \frac{\sum_{i=1}^{n} z_i^{(d)}}{n} \right), \quad z^n \in X^n,$$

and the variance of the MLE is

$$\mathrm{E}_\theta[|\hat{\theta}(z^n) - \theta|^2] = \frac{1}{n} \sum_{j=1}^{d} [p_j(1 - p_j)].$$

Hence, condition A2 with $q = 2$ is satisfied. Concerning conditions A4 and A5, we show two lemmas.

Lemma 3. *For the multinomial distributions, condition A4 is satisfied.*

Proof. Let G_n be the likelihood ratio statistic:

$$G_n(z^n; \theta) := \log \frac{p(z^n|\hat{\theta}(z^n))}{p(z^n|\theta)}.$$

Smith et al. (1981) showed that for θ in the interior of Θ,

$$\mathrm{E}_\theta(G_n(z^n; \theta)) = \frac{d}{2} + R_n(\theta),$$

where R_n satisfies

$$|R_n(\theta)| \leq \sum_{j=0}^{d} n p_j \frac{1}{6 n^3 p_j^3} \left| \mathrm{E}_\theta \left(\frac{\sum_{i=1}^{n} z_i^{(j)}}{n} - p_j \right)^3 \right| = \frac{1}{n} \sum_{j=0}^{d} \frac{|(1 - p_j)(1 - 2p_j)|}{6 p_j}.$$

Thus, $\lim_{n \to \infty} \sup_{\theta \in K} |R_n(\theta)| = 0$. Consequently, the claim is verified. $\square$

Lemma 4. *For the multinomial distributions, the normalizing constant of CNML3 is independent of x^N. Therefore, condition A5 is satisfied.*

Proof. See Appendix. $\square$

In conclusion, the multinomial distributions satisfy conditions A1-A5.

Example 2 (Normal Distributions with Restricted Mean)**.** We fix positive numbers a and b such that $a > b > 0$. Let $\Theta = [-a, a]$ and $K = [-b, b]$. Since a is strictly larger than b, we can take a positive constant δ satisfying $\delta < a - b$ and $K_\delta = [-b - \delta, b + \delta] \subset (-a, a)$.

We consider the normal distribution with mean $\theta \in \Theta$ and variance 1. The probability density function is

$$p(z|\theta) = \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{(z - \theta)^2}{2} \right), \quad z \in X.$$

For $\theta_1, \theta_2 \in K_\delta$, the log-likelihood function satisfies

$$|\log p(z|\theta_1) - \log p(z|\theta_2)| \le (|z| + a)|\theta_1 - \theta_2|.$$

Therefore, condition A1 is satisfied with $p = 2$.

The MLE is

$$\hat{\theta}(z^k) = \begin{cases} -a, & \text{if } \overline{z^k} < -a, \\ a, & \text{if } \overline{z^k} > a, \\ \overline{z^k}, & \text{otherwise,} \end{cases}$$

where $\overline{z^k} := \sum_{i=1}^k z_i / k$. We denote the probability density function of the one-dimensional normal distribution with mean μ and variance σ^2 by $\phi(z; \mu, \sigma^2)$. Since $\overline{z^k}$ is normally distributed with mean θ and variance $1/k$,

$$\mathrm{E}_\theta(\hat{\theta}(z^k) - \theta)^2$$
$$= \int_{-\infty}^{-a} \mathrm{d}z \, (-a - \theta)^2 \phi(z; \theta, 1/k) + \int_a^\infty \mathrm{d}z \, (a - \theta)^2 \phi(z; \theta, 1/k) + \int_{-a}^a \mathrm{d}z \, (z - \theta)^2 \phi(z; \theta, 1/k)$$
$$\le 4a^2 \int_{-\infty}^{\sqrt{k}(-a-\theta)} \mathrm{d}z \, \phi(z; 0, 1) + 4a^2 \int_{\sqrt{k}(a-\theta)}^\infty \mathrm{d}z \, \phi(z; 0, 1) + \int_{-\infty}^\infty \mathrm{d}z \, (z - \theta)^2 \phi(z; \theta, 1/k)$$
$$\le 8a^2 \int_{\sqrt{k}(a-b)}^\infty \mathrm{d}z \, \frac{z}{\sqrt{k}(a - b)} \phi(z; 0, 1) + \frac{1}{k}$$
$$= \frac{8a^2}{\sqrt{k}(a - b)} \exp\left(-\frac{k(a - b)^2}{2}\right) + \frac{1}{k}.$$

Consequently, we verify that condition A2 with $q = 2$ is fulfilled. Next, we verify that condition A3 holds. We have

$$\sup_{\theta \in K}\{\log p(z|\hat{\theta}(z)) - \log p(z|\theta)\} \le \sup_{\theta \in K} \frac{(z - \theta)^2}{2} \le \frac{(z - b)^2}{2} + \frac{(z + b)^2}{2} = z^2 + b^2.$$

Since moments of all orders exist and they are continuous in θ, condition A3 is satisfied with $r = 2$.

Conditions A4 and A5 are also fulfilled, and the proofs are described in Appendix.

Lemma 5. *For this model, condition A4 is satisfied.*

Proof. See Appendix. $\qquad\square$

Lemma 6. *For this model, condition A5 is satisfied.*

Proof. See Appendix. $\qquad\square$

In summary, the one-dimensional normal distributions with restricted mean satisfy conditions A1–A5.

Remark 2. As we will see later, numerous statistical models, including normal and Weibull distributions, satisfy a stronger condition than A5, i.e., the normalizing constant of CNML3 does not depend on the value of observations x^N (see condition B2 and Theorem 2). In Example 2, we verify that the one-dimensional normal model with restricted mean satisfies condition A5. However,

this model does not satisfy the stronger condition (condition B2) and the normalizing constant of CNML3 does depend on x^N.

The quantity

$$\log \left(\int_{X^M} \mathrm{d}\mu(y^M) p(y^M|\hat{\theta}(x^N, y^M)) \right)$$

is not only the logarithm of the normalizing constant of CNML3 but also the minimax conditional regret-3 when we observe x^N and predict M future variables. Intuitively speaking, if the statistical model has "uniformity" such as group structure (for example location-scale models), the conditional regret-3 is equal irrespective of the observations. Even when the uniformity is not equipped with the model such as Example 2, condition A5 is considered to hold because the information of future variables y^M increases as M goes to infinity and therefore the effect of x^N on the conditional regret decreases.

Example 3 (Normal Distributions with Unknown Means)**.** The third example is the normal distribution with unknown means. Let $X = \mathbf{R}^d$ and $\Theta = \mathbf{R}^d$. We take a compact subset K of Θ and fix a positive number $\delta > 0$.

We consider a normal distribution with mean $\theta = (\theta^{(1)}, \ldots, \theta^{(d)}) \in \Theta$ and covariance matrix $\sigma^2 I_d$. Here, $\sigma^2 > 0$ is a known parameter, and I_d is the $d \times d$ identity matrix. The probability density function is

$$p(z|\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \exp \left(-\frac{\sum_{i=1}^{d}(z^{(i)} - \theta^{(i)})^2}{2\sigma^2} \right), \quad z = (z^{(1)}, \ldots, z^{(d)}) \in \mathbf{R}^d.$$

For any compact set $\tilde{K} \subset \mathbf{R}^d$, there exist $\theta_{\min,\tilde{K}}^{(i)} := \min_{\theta \in \tilde{K}} \theta^{(i)}$ and $\theta_{\max,\tilde{K}}^{(i)} := \max_{\theta \in \tilde{K}}, \theta^{(i)}$. For $\theta_1, \theta_2 \in K_\delta$, the log-likelihood function satisfies

$$|\log p(z|\theta_1) - \log p(z|\theta_2)| \le \sum_{i=1}^{d} \left(\frac{1}{\sigma^2}|z^{(i)}| + \frac{|\theta_{\max,K_\delta}^{(i)}| + |\theta_{\min,K_\delta}^{(i)}|}{2\sigma^2} \right) |\theta_1 - \theta_2|.$$

Therefore, condition A1 is satisfied with $p = 2$. The MLE is the sample mean and its variance is $\mathrm{E}_\theta[|\hat{\theta}(z^k) - \theta|^2] = d\sigma^2/k$. Thus, condition A2 is satisfied with $q = 2$. We have

$$\sup_{\theta \in K}\{\log p(z|\hat{\theta}(z)) - \log p(z|\theta)\} = \sup_{\theta \in K} \sum_{i=1}^{d} \frac{(z^{(i)} - \theta^{(i)})^2}{2\sigma^2}$$

$$\le \sum_{i=1}^{d} \left\{ \frac{(z^{(i)} - \theta_{\min,K}^{(i)})^2}{2\sigma^2} + \frac{(z^{(i)} - \theta_{\max,K}^{(i)})^2}{2\sigma^2} \right\}.$$

Because moments of all orders exist and are continuous in θ, condition A3 is satisfied with $r = 2$.

Since for any $\theta \in \Theta$ and for all $j = 1, \ldots, d$,

$$\mathrm{E}_\theta \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{k} \left(z_i^{(j)} - \frac{1}{k} \sum_{l=1}^{k} z_l^{(j)} \right)^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{k} (z_i^{(j)} - \theta)^2 \right\} = \frac{1}{2},$$

condition A4 is satisfied.

Finally, we show that condition A5 holds. Let $\overline{x^{(i)}} := \sum_{j=1}^N x_j^{(i)}/N$ and $\overline{y^{(i)}} := \sum_{j=1}^M y_j^{(i)}/M$. By the translation invariance of the Lebesgue measure,

$$
\int_{\mathbf{R}^{dM}} \mathrm{d}y^M \, p(y^M|\hat{\theta}(x^N, y^M))
$$

$$
= \int_{\mathbf{R}^{dM}} \mathrm{d}y^M \frac{1}{(2\pi\sigma^2)^{\frac{dM}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^d \sum_{j=1}^M \left(y_j^{(i)} - \frac{N\overline{x^{(i)}} + M\overline{y^{(i)}}}{N+M} \right)^2 \right)
$$

$$
= \int_{\mathbf{R}^{dM}} \mathrm{d}z^M \frac{1}{(2\pi\sigma^2)^{\frac{dM}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^d \sum_{j=1}^M \left(z_j^{(i)} - \frac{\sum_{k=1}^M z_k^{(i)}}{N+M} \right)^2 \right),
$$

where $z_j^{(i)} := y_j^{(i)} - \sum_{k=1}^N x_k^{(i)}/N$. Therefore the normalizing constant of CNML3 does not depend on x^N, and thus condition A5 is satisfied. In summary, the normal distributions satisfy conditions A1–A5.

Example 4 (Exponential Distributions)**.** The fourth example is the exponential distribution. Let $X = (0, \infty)$ and $\Theta = (0, \infty)$. We take a compact set K that is contained in Θ. We fix a positive constant δ such that $\inf_{\theta \in K_\delta} \theta > 0$. We define $\theta_{\min,K} := \min_{\theta \in K} \theta > 0$, $\theta_{\max,K} := \max_{\theta \in K} \theta < \infty$ and $\theta_{\min,K_\delta} := \min_{\theta \in K_\delta} \theta > 0$.

The probability density function is

$$
p(z|\theta) = \theta \exp(-\theta z), \quad z \in X, \quad \theta \in \Theta,
$$

and by the mean value theorem, for all $\theta_1, \theta_2 \in K_\delta$,

$$
|\log p(z|\theta_1) - \log p(z|\theta_2)| \leq \left(\frac{1}{\theta_{\min,K_\delta}} + z \right)|\theta_1 - \theta_2|.
$$

Therefore, condition A1 with $p = 2$ is satisfied. Condition A3 with $r = 2$ is also satisfied because

$$
\sup_{\theta \in K}\{\log p(z|\hat{\theta}(z)) - \log p(z|\theta)\} \leq -\log z + |\log \theta_{\min,K}| + |\log \theta_{\max,K}| + |\theta_{\max,K}|z,
$$

and

$$
\sup_{\theta \in K} \mathrm{E}_\theta[z^2] < \infty, \quad \sup_{\theta \in K} \mathrm{E}_\theta[(\log z)^2] < \infty.
$$

The MLE is $\hat{\theta}(z^k) = k/\sum_{i=1}^k z_i$ and $\sum_{i=1}^k z_i$ follows the gamma distribution with mean k/θ and variance k/θ^2. Therefore,

$$
\mathrm{E}_\theta[|\hat{\theta}(z^k) - \theta|^2] = \theta^k \int_0^\infty \mathrm{d}u \left(\frac{k}{u} - \theta \right)^2 \frac{u^{k-1}e^{-\theta u}}{\Gamma(k)} = \frac{2(k+1)}{(k-1)(k-2)}\theta^2,
$$

and condition A2 is satisfied with $q = 2$ because $0 < \theta_{\min,K} \leq \theta \leq \theta_{\max,K} < \infty$ for all $\theta \in K$. Next, we verify that condition A4 holds.

$$
\int_{X^k} \mathrm{d}\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\hat{\theta}(z^k))}{p(z^k|\theta)} = k\log k - k - k\log\theta - k\mathrm{E}_\theta\left[\log\sum_{i=1}^k z_i \right] + \theta\mathrm{E}_\theta\left[\sum_{i=1}^k z_i \right]
$$

$$
= k\log k - k\log\theta - k\int_0^\infty \mathrm{d}u \log u \frac{\theta^k u^{k-1}}{\Gamma(k)} e^{-\theta u}
$$

$$
= k\log k - k\log\theta - k(\psi(k) - \log\theta)
$$

$$
= k(\log k - \psi(k)),
$$

11

where ψ is the digamma function (Gradshteyn and Ryzhik, 2007). The digamma function is represented as

$$\psi(k) = \log k - \frac{1}{2k} - 2 \int_0^\infty \mathrm{d}u \frac{u}{(u^2 + k^2)(\exp(2\pi u) - 1)}.$$

Since $k^2 \leq u^2 + k^2$,

$$0 \leq \int_0^\infty \mathrm{d}u \frac{u}{(u^2 + k^2)(\exp(2\pi u) - 1)} \leq \frac{1}{k^2} \int_0^\infty \mathrm{d}u \frac{u}{\exp(2\pi u) - 1} = \frac{1}{24k^2}.$$

Therefore,

$$\lim_{k \to \infty} k(\log k - \psi(k)) = \frac{1}{2},$$

and thus, condition A4 is satisfied. Finally, we show that condition A5 holds. Let $\bar{x} = \sum_{i=1}^N x_i / N$ and $\bar{y} = \sum_{i=1}^M y_i / M$. The normalizing constant of CNML3 is

$$\int_{X^M} \mathrm{d}\mu(y^M) p(y^M | \hat{\theta}(x^N, y^M)) = \int_{X^M} \mathrm{d}y^M \left(\frac{N+M}{N\bar{x} + M\bar{y}} \right)^M \exp\left(-\frac{N+M}{N\bar{x} + M\bar{y}} \sum_{i=1}^M y_i \right).$$

Let $z_i = My_i / (N\bar{x})$ and $\bar{z} = \sum_{i=1}^M z_i / M$. Then,

$$\int_{X^M} \mathrm{d}\mu(y^M) p(y^M | \hat{\theta}(x^N, y^M)) = \int_{X^M} \mathrm{d}z^M \left(\frac{N+M}{M + M\bar{z}} \right)^M \exp\left(-\frac{N+M}{M + M\bar{z}} \sum_{i=1}^M z_i \right).$$

This is independent of x^N. In conclusion, the exponential distributions satisfy conditions A1–A5.

Thus far, we have considered asymptotic situations, but next, we provide a non-asymptotic result. We state conditions for the result.

B1. For all $\theta \in K$, and for all N and M,

$$\int_{X^N \times X^M} \mathrm{d}\mu(x^N, y^M) p(x^N, y^M | \theta) \log \frac{p(y^M | \theta)}{p(y^M | \hat{\theta}(x^N, y^M))}$$

does not depend on θ.

B2. For all $\theta \in K$, and for all N and M,

$$\log \left(\int_{X^M} \mathrm{d}\mu(y^M) p(y^M | \hat{\theta}(x^N, y^M)) \right)$$

does not depend on x^N.

Theorem 2. *Let K be a compact set that is contained in the interior of Θ and assume that $p(z|\theta)$ is strictly positive for all $z \in X$ and $\theta \in K$. Assume also that conditions B1 and B2 are satisfied.*
Then, for any $\pi \in \mathcal{P}_K$ and for all N and M,

$$D_{K, q_{\mathrm{CNML3}}}^{N,M}(\pi) + R_{\mathrm{KL}}^{N,M}(\pi, p_\pi) = C_*^{N,M}, \tag{3}$$

where $C_*^{N,M}$ is a constant that is independent of π. Therefore, BPCNML3 exactly coincides with the Bayesian predictive density based on the LIP.

Proof. The left-hand side of (3) is

$$D^{N,M}_{K,q_{\mathrm{CNML3}}}(\pi) + R^{N,M}_{\mathrm{KL}}(\pi, p_\pi)$$

$$= \int_\Theta \pi(\mathrm{d}\theta) \left\{ \int_{X^N \times X^M} \mathrm{d}\mu(x^N, y^M) p(x^N, y^M | \theta) \log \frac{p(y^M | \theta)}{p(y^M | \hat{\theta}(x^N, y^M))} \right\}$$

$$+ \int_\Theta \pi(\mathrm{d}\theta) \left\{ \int_{X^N} \mathrm{d}\mu(x^N) p(x^N | \theta) \log \left(\int_{X^M} \mathrm{d}\mu(y^M) p(y^M | \hat{\theta}(x^N, y^M)) \right) \right\}.$$

By assumptions B1 and B2, the claim is verified. $\qquad\square$

Example 5 (One-Dimensional Normal Distribution with Unknown Mean)**.** In Example 3, we show that the normal distribution satisfies condition B2. Here, we verify that condition B1 holds. Assume that x^N and y^M are independent and identically normally distributed with unknown mean θ and variance 1. Let $\bar{x} := \sum_{i=1}^N x_i / N$ and let $\bar{y} := \sum_{i=1}^M y_i / M$. Then,

$$\int_{X^N \times X^M} \mathrm{d}\mu(x^N, y^M) p(x^N, y^M | \theta) \log \frac{p(y^M | \theta)}{p(y^M | \hat{\theta}(x^N, y^M))}$$

$$= \mathrm{E}_\theta \sum_{i=1}^M \frac{1}{2} \left\{ -(y_i - \theta)^2 + \left(y_i - \frac{N\bar{x} + M\bar{y}}{N + M} \right)^2 \right\}$$

$$= -\frac{M}{2} + \frac{M(1 + \theta^2)}{2} - \frac{NM\theta^2 + M(1 + M\theta^2)}{N + M} + \frac{M\theta^2}{2} + \frac{M}{2(N + M)}$$

$$= -\frac{M}{2(N + M)}.$$

Condition B1 is satisfied, and thus, Theorem 2 holds.

Example 6 (Weibull Distribution with Unknown Scale Parameter)**.** Let $X = (0, \infty)$ and $\Theta = (0, \infty)$. We consider the Weibull distribution with unknown scale parameter $\theta \in \Theta$ and known shape parameter $k \in (0, \infty)$. The Weibull distributions are widely known to include numerous other probability distributions, such as the exponential distributions ($k = 1$) and the Rayleigh distributions ($k = 2$).

The probability density function is

$$p(z | \theta) = \frac{k}{\theta} \left(\frac{z}{\theta} \right)^{k-1} \exp \left\{ -\left(\frac{x}{\theta} \right)^k \right\}, \quad z \in X.$$

The MLE is

$$\hat{\theta}(z^n) = \left(\frac{1}{n} \sum_{i=1}^n z_i^k \right)^{\frac{1}{k}}.$$

First we show that condition B1 is satisfied. We have

$$\int_{X^N \times X^M} \mathrm{d}\mu(x^N, y^M) p(x^N, y^M | \theta) \log \frac{p(y^M | \theta)}{p(y^M | \hat{\theta}(x^N, y^M))}$$

$$= \mathrm{E}_\theta \sum_{i=1}^M \left\{ -k \log \theta + k \log \hat{\theta}(x^N, y^M) - \frac{y_i^k}{\theta^k} + \frac{y_i^k}{(\hat{\theta}(x^N, y^M))^k} \right\}$$

$$= \mathrm{E}_\theta \left\{ M \log \left(\sum_{i=1}^N \frac{x_i^k}{\theta^k} + \sum_{i=1}^M \frac{y_i^k}{\theta^k} \right) - \sum_{i=1}^M \frac{y_i^k}{\theta^k} + \frac{(N + M) \sum_{i=1}^M \frac{y_i^k}{\theta^k}}{\sum_{i=1}^N \frac{x_i^k}{\theta^k} + \sum_{i=1}^M \frac{y_i^k}{\theta^k}} \right\} - M \log(N + M).$$

If a random variable Z follows the Weibull distribution with scale parameter θ and shape parameter k, then $(Z/\theta)^k$ follows the exponential distribution with mean 1. In addition, if two random variables Z_1 and Z_2 follow the gamma distributions with common scale parameter ξ and shape parameters α and β, respectively, then $Z_1/(Z_1 + Z_2)$ follows the beta distribution with shape parameters α and β. From these facts and the reproductive property of the gamma distribution,

$$
\int_{X^N \times X^M} d\mu(x^N, y^M) p(x^N, y^M|\theta) \log \frac{p(y^M|\theta)}{p(y^M|\hat{\theta}(x^N, y^M))}
$$
$$
= M\psi(N+M) - M + (N+M) \times \frac{M}{N+M} - M\log(N+M)
$$
$$
= M(\psi(N+M) - \log(N+M)),
$$

where ψ is the digamma function. Hence, condition B1 is fulfilled. Because we can verify that condition B2 holds in the same manner as Example 4, we omit the proof.

4 Numerical Experiments

In Example 1, we verify that the multinomial distribution satisfies condition A1–A5 and thus, Theorem 1 holds. In this section, we confirm the validity of Theorem 1 for the binomial distribution through numerical experiments.

We explain the settings of the numerical experiments. Let $\Theta = [0, 1]$ and $K = [0.1, 0.9]$. Since $\mathcal{P}_K$ is infinite-dimensional space, we approximate $\mathcal{P}_K$ by the set of discrete distributions $\tilde{P}_K^{100}$:

$$
\tilde{P}_K^{100} := \left\{ \sum_{i=0}^{100} \pi_i \delta_{0.1+0.08i}(d\theta) \middle| 0 \le \pi_i \le 1 \text{ for all } i, \ \sum_{i=0}^{100} \pi_i = 1. \right\},
$$

where $\delta_a(d\theta)$ denotes the Dirac measure with support $a \in \Theta$. By numerical optimization, we calculate the approximation of the LIP

$$
\tilde{\pi}_{K,\text{LIP}}^{N,M} := \underset{\pi \in \tilde{\mathcal{P}}_K}{\operatorname{argmax}} R_{\text{KL}}^{N,M}(\pi, p_\pi)
$$
$$
= \underset{\pi \in \tilde{\mathcal{P}}_K}{\operatorname{argmax}} \sum_{i,j,k} \pi_i \binom{N}{j} \binom{M}{k} \theta_i^{j+k}(1-\theta_i)^{N+M-j-k} \log \frac{\theta_i^k (1-\theta_i)^{M-k} p_\pi(j)}{p_\pi(j,k)},
$$

and BPCNML3

$$
\tilde{\pi}_{K,q_{\text{CNML3}}}^{N,M} := \underset{\pi \in \tilde{\mathcal{P}}_K}{\operatorname{argmin}} D_{K,q_{\text{CNML3}}}^{N,M}(\pi)
$$
$$
= \underset{\pi \in \tilde{\mathcal{P}}_K}{\operatorname{argmin}} \sum_{i,j,k} \pi_i \binom{N}{j} \binom{M}{k} \theta_i^{j+k}(1-\theta_i)^{N+M-j-k} \log \frac{p_\pi(j,k)}{(\hat{\theta}_{j,k})^k (1-\hat{\theta}_{j,k})^{M-k} p_\pi(j)},
$$

where $\theta_i := 0.1 + 0.08i$, $\hat{\theta}_{j,k} := (j+k)/(N+M)$, $p_\pi(j) := \sum_{i=0}^{100} \pi_i \theta_i^j (1-\theta_i)^{N-j}$ and $p_\pi(j,k) := \sum_{i=0}^{100} \pi_i \theta_i^{j+k} (1-\theta_i)^{N+M-j-k}$. We used the free software R (R Development Core Team, 2009) and constrOptim function for the optimization.

Figure 1–3 show the result of comparison of KL risk among CNML3, BPCNML3, and Bayesian predictive densities based on LIP (simply abbreviated to BPDLIP) when $N = 1$ and $M =$
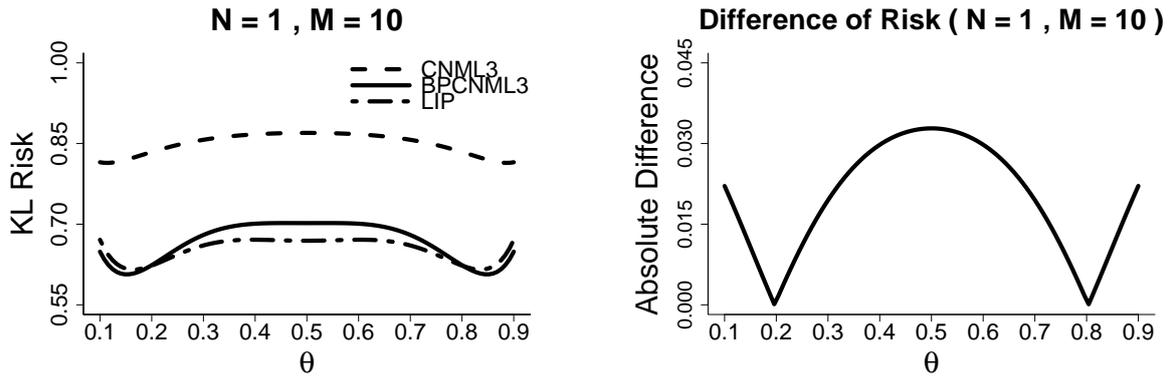
Figure 1: Comparison of KL risk when $N = 1$, $M = 10$. The right panel shows the absolute difference of KL risk between BPCNML3 and BPDLIP.
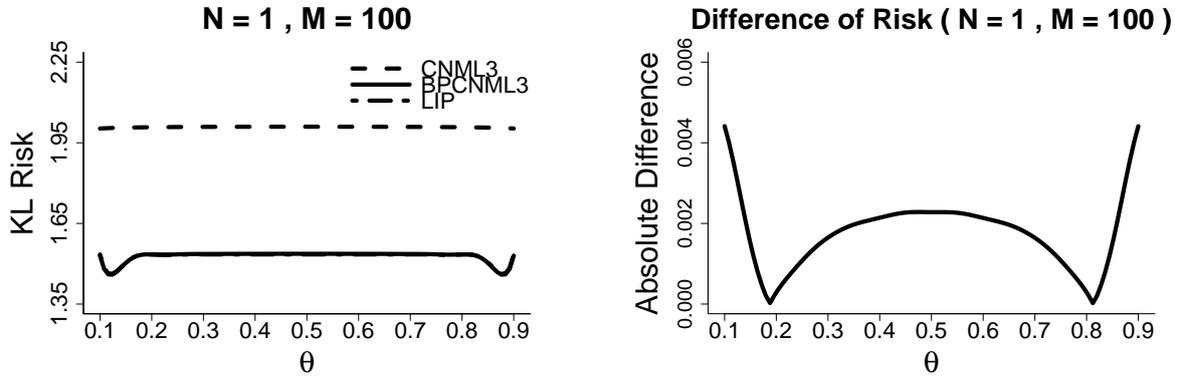
Figure 2: Comparison of KL risk when $N = 1$, $M = 100$. Since the KL risk of BPCNML3 is almost the same as that of BPDLIP, we plot the absolute difference of KL risk between BPCNML3 and BPDLIP in the right panel.
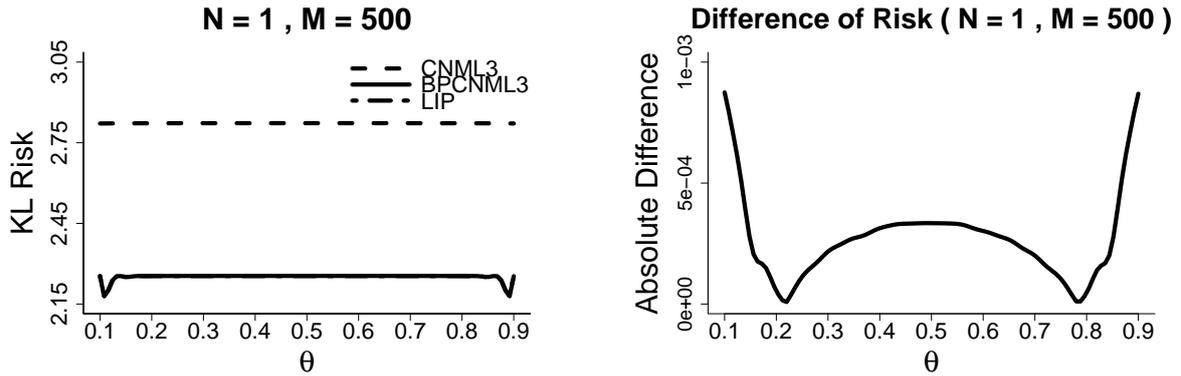
Figure 3: Comparison of KL risk when $N = 1$, $M = 500$. Since the KL risk of BPCNML3 is almost the same as that of BPDLIP, we plot the absolute difference of KL risk between BPCNML3 and BPDLIP in the right panel.

$10, 100, 500$. When $N = 1$ and $M = 100, 500$ (Figure 2 and 3), KL risk of BPCNML3 is almost the same as that of BPDLIP. Therefore, we plot the absolute difference of KL risk between BPCNML3 and BPDLIP.

Implications from the figures are twofold. First, KL risk of BPCNML3 is much lower than that of CNML3. Notably, for submodels of the multinomial distributions, Komaki (2011) showed that KL risk of the Bayes projection of predictive density q is not larger than that of q. In addition, the amount of reduction increases as M increases. Second, we find that the difference of KL risk between BPCNML3 and BPDLIP goes to zero as M increases. This finding implies that BPCNML3 is asymptotically identical to BPDLIP.

5   Conclusion

In this study, we discussed the relations between the Bayes projection of CNML3 (BPCNML3) and the Bayesian predictive density based on the LIP (BPDLIP). In Theorem 1, we proved that the sum of the Bayes projection divergence of CNML3 and the conditional mutual information is asymptotically constant. Roughly speaking, this result implies that the BPCNML3 is asymptotically identical to the BPDLIP. The numerical results in Section 4 confirmed that the BPCNML3 is asymptotically identical to the BPDLIP for the binomial model. Under stronger conditions B1 and B2, we showed that the BPCNML3 exactly coincides with the BPDLIP in Theorem 2.

Our results shed light on the connection between CNML3 and LIPs. Although CNML2 has received the most attention among CNML distributions, we argue that CNML3, not CNML2, is more in line with the minimax KL risk approach and is the most important predictive density among CNML distributions.

Finally, we provide our future plans for this study. The plans are threefold. First, we will study the sufficient conditions for A5 and B2. These conditions are concerned with the conditional minimax regret-3. As reported in Remark 2, we believe that numerous regular statistical models satisfy these conditions. Second, we will address the boundary of the parameter space. In the same manner as Clarke and Barron (1994), we restricted the support set of the prior distributions that should be contained in the fixed compact set. Using the methods such as in Xie and Barron (2000) or Komaki (2012), we may treat the boundary of the parameter space. Finally, we plan to study the predictive performance of the BPDLIP under the conditional regret-3. It is an interesting study because it parallels to the study of Xie and Barron (2000).

A   Appendix: Proofs of Lemmas

A.1   Proof of Lemma 2

Proof. We define several notations as follows:

$$I_{N,M}(\theta) := \int_{X^N \times X^M} \mathrm{d}\mu(x^N, y^M) \, p(x^N, y^M | \theta) \log \frac{p(y^M | \theta)}{p(y^M | \hat{\theta}(x^N, y^M))} + \frac{d}{2},$$

$$R_k(\theta) := \int_{X^k} \mathrm{d}\mu(z^k) \, p(z^k | \theta) \log \frac{p(z^k | \theta)}{p(z^k | \hat{\theta}(z^k))} + \frac{d}{2}.$$

Note that since $p(x^N, y^M | \hat{\theta}(x^N, y^M)) = p(x^N | \hat{\theta}(x^N, y^M)) p(y^M | \hat{\theta}(x^N, y^M)) > 0$ for all x^N and y^M,

$$p(x^N | \hat{\theta}(x^N, y^M)) > 0, \quad p(y^M | \hat{\theta}(x^N, y^M)) > 0.$$

Since $p(y^M|\hat{\theta}(y^M)) \geq p(y^M|\hat{\theta}(x^N, y^M))$

$$R_M(\theta) \leq I_{N,M}(\theta), \quad \forall \theta \in K. \tag{4}$$

For $\theta \in K$, the integrand in the claim of Lemma 2 is decomposed as

$$\frac{p(y^M|\theta)}{p(y^M|\hat{\theta}(x^N, y^M))} = \frac{p(x^N, y^M|\theta)}{p(x^N, y^M|\hat{\theta}(x^N, y^M))} \frac{p(x^N|\hat{\theta}(x^N, y^M))}{p(x^N|\theta)}. \tag{5}$$

By condition A1, for $(x^N, y^M) \in \{\hat{\theta}(x^N, y^M) \in K_\delta\}$

$$\log \frac{p(x^N|\hat{\theta}(x^N, y^M))}{p(x^N|\theta)} \leq |\hat{\theta}(x^N, y^M) - \theta| \sum_{i=1}^{N} L_{K_\delta}(x_i).$$

In addition, by condition A3, for $(x^N, y^M) \in \{\hat{\theta}(x^N, y^M) \notin K_\delta\}$

$$\log \frac{p(x^N|\hat{\theta}(x^N, y^M))}{p(x^N|\theta)} = \sum_{i=1}^{N} \{\log p(x_i|\hat{\theta}(x^N, y^M)) - \log p(x_i|\theta)\}$$

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

$$\leq \sum_{i=1}^{N} T_K(x_i).$$

By the Hölder inequality, for all $\theta \in K$

$$\int_{X^N \times X^M} d\mu(x^N, y^M) p(x^N, y^M|\theta) \log \frac{p(x^N|\hat{\theta}(x^N, y^M))}{p(x^N|\theta)}$$

$$\leq \sup_{\theta \in K} \int_{\{\hat{\theta}(x^N, y^M) \in K_\delta\}} d\mu(x^N, y^M) p(x^N, y^M|\theta)|\hat{\theta}(x^N, y^M) - \theta| \sum_{i=1}^{N} L_{K_\delta}(x_i)$$

$$+ \sup_{\theta \in K} \int_{\{\hat{\theta}(x^N, y^M) \notin K_\delta\}} d\mu(x^N, y^M) p(x^N, y^M|\theta) \sum_{i=1}^{N} T_K(x_i)$$

$$\leq \sup_{\theta \in K} \left\{ \int_{X^N} d\mu(x^N) p(x^N|\theta) \left(\sum_{i=1}^{N} L_{K_\delta}(x_i) \right)^p \right\}^{\frac{1}{p}}$$

$$\times \sup_{\theta \in K} \left\{ \int_{X^N \times X^M} d\mu(x^N, y^M) p(x^N, y^M|\theta)|\hat{\theta}(x^N, y^M) - \theta|^q \right\}^{\frac{1}{q}}$$

$$+ \sup_{\theta \in K} \left\{ P_\theta^{N+M}\left(\hat{\theta}(x^N, y^M) \notin K_\delta \right) \right\}^{\frac{1}{s}} \sup_{\theta \in K} \left\{ \int_{X^N} d\mu(x^N) p(x^N|\theta) \left(\sum_{i=1}^{N} T_K(x_i) \right)^r \right\}^{\frac{1}{r}},$$

where s satisfies $1/r + 1/s = 1$. We denote the upper bound by $U_{M,N}$. Note that $U_{M,N}$ is nonnegative and does not depend on θ. By conditions A1–A3 and Lemma 1, we have

$$\lim_{M \to \infty} U_{M,N} = 0. \tag{6}$$

17

From (4) and (5),

$$R_M(\theta) \le I_{M,N}(\theta) \le R_{M+N}(\theta) + U_{M,N}.$$

Therefore, since $|I_{M,N}(\theta)| \le \max\{|R_M(\theta)|, |R_{M+N}(\theta)| + U_{M,N}\}$

$$\sup_{\theta \in K} |I_{M,N}(\theta)| \le \max\left\{ \sup_{\theta \in K} |R_M(\theta)|, \ \sup_{\theta \in K} |R_{M+N}(\theta)| + U_{M,N} \right\}$$

$$\le \sup_{\theta \in K} |R_M(\theta)| + \sup_{\theta \in K} |R_{M+N}(\theta)| + U_{M,N}.$$

By condition A4 and (6), we have

$$\lim_{M \to \infty} \sup_{\theta \in K} |I_{M,N}(\theta)| = 0.$$

$\square$

A.2   Proof of Lemma 4

Proof. We define a family of polynomials with one variable t as follows:

$$f_{M,a}^{(2)}(t) := \sum_{i=0}^{M} \binom{M}{i} (t+i)^i (M + a - t - i)^{M-i},$$

where M is a positive integer and a is a real number. We also define $f_{0,a}^{(2)}(t) \equiv 1$. If we set $a = N$ and $x \in \{0, 1, \ldots, N\}$, then $f_{M,N}(x)/(M+N)^M$ is the normalizing constant of CNML3 for the binomial distributions with observations x^N satisfying $\sum_{i=1}^{N} x_i = x$. For any nonnegative integer M and any real number a, we first prove that $f_{M,a}^{(2)}$ does not depend on the value of t, i.e., $f_{M,a}^{(2)}$ is a constant function.

It suffices to show that for any real number a,

$$\frac{\mathrm{d}}{\mathrm{d}t} f_{M,a}^{(2)}(t) = 0, \quad \forall t \in \mathbf{R}, \tag{7}$$

since $f_{M,a}^{(2)}$ is a polynomial in t. We prove this by mathematical induction with respect to M. For $M = 0$, (7) is evident by the definition of $f_{0,a}^{(2)}$. Assume that (7) holds for $M = m$ and any $a \in \mathbf{R}$. From this assumption, $f_{m,a}^{(2)}$ is a constant function. Then,

$$\frac{\mathrm{d}}{\mathrm{d}t} f_{m+1,a}^{(2)}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left\{ (m + 1 + a - t)^{m+1} + (t + m + 1)^{m+1} \right\}$$

$$+ \sum_{i=1}^{m} \binom{m+1}{i} \frac{\mathrm{d}}{\mathrm{d}t} (t+i)^i (m + 1 + a - t - i)^{m+1-i}$$

$$= -(m+1)(m + 1 + a - t)^m + (m+1)(t + m + 1)^m$$

$$+ \sum_{i=1}^{m} \binom{m+1}{i} i(t+i)^{i-1} (m + 1 + a - t - i)^{m+1-i}$$

$$- \sum_{i=1}^{m} \binom{m+1}{i} (m + 1 - i)(t+i)^i (m + 1 + a - t - i)^{m-i}.$$

18

Since
$$\binom{m+1}{i} i = (m+1)\binom{m}{i-1}, \quad \binom{m+1}{i}(m+1-i) = (m+1)\binom{m}{i},$$
hold,

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} f_{m+1,a}^{(2)}(t) = & -(m+1)(m+1+a-t)^m + (m+1)(t+m+1)^m \\
& + (m+1)\sum_{i=1}^{m}\binom{m}{i-1}(t+i)^{i-1}(m+1+a-t-i)^{m+1-i} \\
& - (m+1)\sum_{i=1}^{m}\binom{m}{i}(t+i)^{i}(m+1+a-t-i)^{m-i} \\
= & \ (m+1)(t+m+1)^m \\
& + (m+1)\sum_{i=0}^{m-1}\binom{m}{i}(t+i+1)^{i}(m+a-t-i)^{m-i} \\
& - (m+1)\sum_{i=0}^{m}\binom{m}{i}(t+i)^{i}(m+1+a-t-i)^{m-i} \\
= & \ (m+1)\sum_{i=0}^{m}\binom{m}{i}(t+i+1)^{i}(m+a-t-i)^{m-i} \\
& - (m+1)\sum_{i=0}^{m}\binom{m}{i}(t+i)^{i}(m+1+a-t-i)^{m-i} \\
= & \ (m+1)\{f_{m,a+1}(t+1) - f_{m,a+1}(t)\}.
\end{aligned}
$$

By the assumption of the induction, $f_{m,a+1}(t+1) - f_{m,a+1}(t) = 0$. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t} f_{m+1,a}^{(2)}(t) = 0,$$

and (7) is verified for any M and a. In addition, from this result, the claim of Lemma 4 is verified for the binomial distributions.

Next, we show that the claim holds for $(d+1)$-nomial distributions $(d \geq 2)$. We define a family of polynomials with d variables $(t_1, \ldots, t_d)$ as follows:

$$f_{M,a}^{(d+1)}(t_1, \ldots, t_d) := \sum_{\substack{0 \leq i_1,\ldots,i_d \leq M, \\ \sum_{l=1}^{d} i_l = M}} \frac{M! \prod_{l=1}^{d}(t_l+i_l)^{i_l}}{i_1! \ldots i_k! (M - \sum_{l=1}^{k} i_l)!} \left(M+a-\sum_{l=1}^{d}(t_l+i_l)\right)^{M-\sum_{l=1}^{d} i_l}.$$

where M is a positive integer and a is a real number. For the same reason discussed in the case of the binomial distributions, it suffices to show that $f_{M,a}^{(d+1)}$ is a constant function to verify that the normalizing constant of CNML3 for $(d+1)$-nomial distribution is independent of x^N.

Note that $f_{M,a}^{(d+1)}$ is symmetric with respect to any permutation of variables, i.e., for any permutation σ,

$$f_{M,a}^{(d+1)}(t_{\sigma(1)}, \ldots, t_{\sigma(d)}) = f_{M,a}^{(d+1)}(t_1, \ldots, t_d).$$

19

Therefore, it is sufficient to show that

$$\frac{\partial}{\partial t_1} f_{M,a}^{(d+1)}(t_1,\dots,t_d) = 0,$$

since $f_{M,a}^{(d+1)}$ is a symmetric polynomial in $(t_1,\dots,t_d)$.

$$\frac{\partial}{\partial t_1} f_{M,a}^{(d+1)}(t_1,\dots,t_d)$$

$$= \frac{\partial}{\partial t_1} \sum_{\substack{0 \le i_2,\dots,i_d \le M, \\ \sum_{l=2}^{d} i_l \le M}} \sum_{i_1=0}^{M-\sum_{l=2}^{d} i_l} \frac{M! \prod_{l=1}^{d}(t_l + i_l)^{i_l}}{i_1! \dots i_d! (M - \sum_{l=1}^{d} i_l)!} \left(M + a - \sum_{l=1}^{d}(t_l + i_l) \right)^{M - \sum_{l=1}^{d} i_l}$$

$$= \frac{\partial}{\partial t_1} \sum_{\substack{0 \le i_2,\dots,i_d \le M, \\ \sum_{l=2}^{d} i_l \le M}} \frac{M! \prod_{l=2}^{d}(t_l + i_l)^{i_l}}{i_2! \dots i_d! (M - \sum_{l=2}^{d} i_l)!}$$

$$\times \sum_{i_1=0}^{M-\sum_{l=2}^{d} i_l} \frac{(M - \sum_{l=2}^{d} i_l)! (t_1 + i_1)^{i_1}}{i_1! (M - \sum_{l=2}^{d} i_l - i_1)!} \left(M - \sum_{l=2}^{d} t_l + a - \sum_{l=2}^{d} i_l - t_1 - i_1 \right)^{M - \sum_{l=2}^{d} i_l - i_1}$$

$$= \sum_{\substack{0 \le i_2,\dots,i_d \le M, \\ \sum_{l=2}^{d} i_l \le M}} \frac{M! \prod_{l=2}^{d}(t_l + i_l)^{i_l}}{i_2! \dots i_d! (M - \sum_{l=2}^{d} i_l)!} \frac{\partial}{\partial t_1} f_{(M-\sum_{l=2}^{d} i_l),(a-\sum_{l=2}^{d} t_l)}^{(2)}(t_1)$$

$$= 0,$$

since (7) holds. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

A.3   Proof of Lemma 5

Proof. Note that it is easy to verify that

$$\int_{X^k} d\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\overline{z^k})}{p(z^k|\theta)} = \frac{1}{2}, \tag{8}$$

and thus, we omit the calculation. Let $A_k := \{\hat{\theta}(z^k) \in (-a, a)\}$. Since (8) holds and $\hat{\theta}(z^k) = \overline{z^k}$ for $z^k \in A_k$

$$\frac{1}{2} - \int_{X^k} d\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\hat{\theta}(z^k))}{p(z^k|\theta)}$$

$$= \int_{X^k} d\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\overline{z^k})}{p(z^k|\theta)} - \int_{X^k} d\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\hat{\theta}(z^k))}{p(z^k|\theta)}$$

$$= \int_{A_k^c} d\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\overline{z^k})}{p(z^k|\hat{\theta}(z^k))}. \tag{9}$$

Since $p(z^k|\overline{z^k}) \ge p(z^k|\hat{\theta}(z^k))$, (9) is positive. Hence,

$$\left| \int_{X^k} d\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\hat{\theta}(z^k))}{p(z^k|\theta)} - \frac{1}{2} \right| = \int_{A_k^c} d\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\overline{z^k})}{p(z^k|\hat{\theta}(z^k))}. \tag{10}$$

The integrand in the right-hand side of (10) is

$$\log \frac{p(z^k|\overline{z^k})}{p(z^k|\hat{\theta}(z^k))} = k(\overline{z^k} - \hat{\theta}(z^k))\overline{z^k} - \frac{k}{2}((\overline{z^k})^2 - (\hat{\theta}(z^k))^2).$$

Since the sample mean $\overline{z^k}$ is normally distributed with mean θ and variance $1/k$,

$$
\begin{aligned}
&\int_{A_k^c} d\mu(z^k) p(z^k|\theta) \log \frac{p(z^k|\overline{z^k})}{p(z^k|\hat{\theta}(z^k))} \\
&= \int_a^\infty du \; \phi(u;\theta,1/k)\frac{k(u-a)^2}{2} + \int_{-\infty}^{-a} du \; \phi(u;\theta,1/k)\frac{k(u+a)^2}{2} \\
&= \int_{\sqrt{k}(a-\theta)}^\infty du \; \phi(u;0,1)\frac{(u-\sqrt{k}(\theta+a))^2}{2} + \int_{-\infty}^{-\sqrt{k}(a+\theta)} du \; \phi(u;0,1)\frac{(u-\sqrt{k}(\theta-a))^2}{2} \\
&\leq \int_{\sqrt{k}\delta}^\infty du \; \phi(u;0,1)\frac{(u-\sqrt{k}(\theta+a))^2}{2} + \int_{-\infty}^{-\sqrt{k}\delta} du \; \phi(u;0,1)\frac{(u-\sqrt{k}(\theta-a))^2}{2} \\
&= \int_{\sqrt{k}\delta}^\infty du \; \phi(u;0,1)\frac{(u-\sqrt{k}\theta-\sqrt{k}a)^2}{2} + \int_{\sqrt{k}\delta}^\infty du \; \phi(u;0,1)\frac{(u+\sqrt{k}\theta-\sqrt{k}a)^2}{2} \\
&\leq \int_{\sqrt{k}\delta}^\infty du \; \phi(u;0,1)(u^2 + k(a^2 + \theta^2)).
\end{aligned}
$$

By the Lebesgue convergence theorem

$$\lim_{k\to\infty} \int_{\sqrt{k}\delta}^\infty du \; \phi(u;0,1)u^2 = 0.$$

In addition, since $u/(\sqrt{k}\delta) \geq 1$ for $u \geq \sqrt{k}\delta$,

$$\int_{\sqrt{k}\delta}^\infty du \; \phi(u;0,1)k(a^2+\theta^2) \leq \int_{\sqrt{k}\delta}^\infty du \; \phi(u;0,1)\frac{2uka^2}{\sqrt{k}\delta} = \frac{2\sqrt{k}a^2}{\delta}\exp(-k\delta^2/2).$$

Therefore,

$$\lim_{k\to\infty} \sup_{\theta\in K} \int_{\sqrt{k}\delta}^\infty du \; \phi(u;0,1)k(a^2+\theta^2) = 0.$$

By (10), the claim is verified. □

A.4   Proof of Lemma 6

Proof. First, we derive

$$\int_{\mathbf{R}^M} dy^M p(y^M|\hat{\theta}(x^N, y^M)) = 1 + \frac{M}{N}\int_{-\frac{aN}{\sqrt{M}}-\frac{1}{\sqrt{M}}\sum_{i=1}^N x_i}^{\frac{aN}{\sqrt{M}}-\frac{1}{\sqrt{M}}\sum_{i=1}^N x_i} dv \; \phi(v;0,1).$$

From this equation, we find that the normalizing constant of CNML3 does depend on the value $\sum_{i=1}^N x_i$ (see Remark 2).

Let $u := \sum_{i=1}^N x_i$ and let $v^M := (v_1, \ldots, v_M)^\top$ satisfying

$$v^M = Hy^M,$$

21

where H is the $M \times M$ orthogonal matrix of the Helmert transformation. From the definition of H, we have

$$\frac{1}{\sqrt{M}} \sum_{i=1}^{M} y_i = v_1, \quad \sum_{i=1}^{M} y_i^2 = \sum_{i=1}^{M} v_i^2.$$

MLE $\hat{\theta}(x^N, y^M)$ is represented in terms of u and v_1:

$$\hat{\theta}(u, v_1) := \begin{cases} -a, & \text{if } \frac{u + \sqrt{M} v_1}{N + M} < -a, \\ a, & \text{if } \frac{u + \sqrt{M} v_1}{N + M} > a, \\ \frac{u + \sqrt{M} v_1}{N + M}, & \text{otherwise.} \end{cases}$$

Because H is orthogonal,

$$\int_{\mathbf{R}^M} \mathrm{d}y^M p(y^M | \hat{\theta}(x^N, y^M))$$

$$= \int_{\mathbf{R}^M} \mathrm{d}v^M \frac{1}{(2\pi)^{\frac{M}{2}}} \exp\left(-\frac{1}{2} \sum_{i=2}^{M} v_i^2 - \frac{1}{2}(v_1 - \sqrt{M}\hat{\theta}(u, v_1))^2 \right)$$

$$= \int_{\mathbf{R}} \mathrm{d}v_1 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(v_1 - \sqrt{M}\hat{\theta}(u, v_1))^2 \right)$$

$$= \int_{-\infty}^{-\frac{a(N+M)}{\sqrt{M}} - \frac{u}{\sqrt{M}}} \mathrm{d}v_1 \frac{\exp\left(-\frac{(v_1 + \sqrt{M}a)^2}{2} \right)}{\sqrt{2\pi}} + \int_{\frac{a(N+M)}{\sqrt{M}} - \frac{u}{\sqrt{M}}}^{\infty} \mathrm{d}v_1 \frac{\exp\left(-\frac{(v_1 - \sqrt{M}a)^2}{2} \right)}{\sqrt{2\pi}}$$

$$+ \int_{-\frac{a(N+M)}{\sqrt{M}} - \frac{u}{\sqrt{M}}}^{\frac{a(N+M)}{\sqrt{M}} - \frac{u}{\sqrt{M}}} \mathrm{d}v_1 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(N v_1 - \sqrt{M} u)^2}{2(N+M)^2} \right)$$

$$= 1 + \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}} - \frac{u}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}} - \frac{u}{\sqrt{M}}} \mathrm{d}v_1 \phi(v_1; 0, 1).$$

Next, we verify that condition A5 is satisfied.

$$\frac{M}{N} \int_{-\frac{aN}{\sqrt{M}} - \frac{u}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}} - \frac{u}{\sqrt{M}}} \mathrm{d}v_1 \, \phi(v_1; 0, 1) = \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}}} \mathrm{d}v_1 \, \phi(v_1; 0, 1) \exp\left(\frac{u}{\sqrt{M}} v_1 - \frac{u^2}{2M} \right)$$

$$\leq \exp\left(\frac{aN|u|}{M} \right) \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}}} \mathrm{d}v_1 \, \phi(v_1; 0, 1).$$

Since $\exp(aN|u|/M) \geq 1$,

$$1 + \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}} - \frac{u}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}} - \frac{u}{\sqrt{M}}} \mathrm{d}v_1 \, \phi(v_1; 0, 1) \leq \exp\left(\frac{aN|u|}{M} \right) \left(1 + \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}}} \mathrm{d}v_1 \, \phi(v_1; 0, 1) \right).$$

Similarly, we find the lower bound as follows:

$$1 + \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}} - \frac{u}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}} - \frac{u}{\sqrt{M}}} \mathrm{d}v_1 \, \phi(v_1; 0, 1) \geq \exp\left(-\frac{aN|u|}{M} - \frac{u^2}{2M} \right) \left(1 + \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}}} \mathrm{d}v_1 \, \phi(v_1; 0, 1) \right).$$

Therefore,

$$\left| \log \int_{\mathbf{R}^M} \mathrm{d}y^M p(y^M | \hat{\theta}(x^N, y^M)) - \log\left(1 + \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}}} \mathrm{d}v_1 \ \phi(v_1; 0, 1) \right) \right| \leq \frac{aN|u|}{M} + \frac{u^2}{2M}.$$

From this inequality, if we set

$$C^{N,M} = \log\left(1 + \frac{M}{N} \int_{-\frac{aN}{\sqrt{M}}}^{\frac{aN}{\sqrt{M}}} \mathrm{d}v_1 \ \phi(v_1; 0, 1) \right),$$

then the claim is verified because $\mathrm{E}_\theta[|u|]$ and $\mathrm{E}_\theta[u^2]$ is uniformly bounded in $\theta \in K$ and do not depend on M. $\qquad\square$

References

Bartlett, P., Grünwald, P. D., Harremoës, P., Hedayati, F., and Kotłowski, W. (2013). Horizon-independent optimal prediction with log-loss in exponential families. In *Proceedings of the Twenty Sixth Annual Conference on Learning Theory*.

Clarke, B. and Barron, A. R. (1989). Information theoretic asymptotics of Bayes methods. Technical Report 26, University of Illinois, Department of Statistics.

Clarke, B. and Barron, A. R. (1994). Jeffreys prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 40:37–60.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience, second edition.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3:146–158.

Ferguson, T. S. (1967). *Mathematical statistics: a decison theoretic approach*. Academic Press, New York.

Gradshteyn, I. S. and Ryzhik, I. M. (2007). *Table of integrals, series, and products*. Academic Press, Amsterdam, seventh edition.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT Press, Cambridge.

Grünwald, P. D. (2012). Commentary on "The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators". *JMLR Workshop and Conference Proceedings*, 23:7.14–7.17.

Harremoës, P. (2013). Extendable MDL. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 1516–1520.

Hedayati, F. and Bartlett, P. (2012a). Exchangeability characterizes optimality of sequential normalized maximum likelihood and Bayesian prediction with Jeffreys prior. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*.

Hedayati, F. and Bartlett, P. (2012b). The optimality of Jeffreys prior for online density estimation and the asymptotic normality of maximum likelihood estimators. In *Proceedings of the Twenty Fifth Annual Conference on Learning Theory.*

Komaki, F. (2011). Bayesian predictive densities based on latent information priors. *Journal of Statistical Planning and Inference*, 141:3705–3715.

Komaki, F. (2012). Asymptotically minimax Bayesian predictive densities for multinomial models. *Electronic Journal of Statistics*, 6:934–957.

Kotłowski, W. and Grünwald, P. D. (2011). Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the Twenty Fourth Annual Conference on Learning Theory.*

Liang, F. and Barron, A. R. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50:2708–2726.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rényi, A. (1961). On measures of information and entropy. In *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, volume 1, pages 547–561.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*, 23:175–186.

Smith, P. J., Rae, D. S., Manderscheid, R. W., and Silbergeld, S. (1981). Approximating the moments and distribution of the likelihood ratio statistic for multinomial goodness of fit. *Journal of the American Statistical Association*, 76:737–740.

van Erven, T. and Harremoës, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*, 60:3797–3820.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62.

Xie, Q. and Barron, A. R. (2000). Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46:431–445.