TWOFOLD STRUCTURE OF DUALITY IN BAYESIAN MODEL AVERAGING

Toshio Ohnishi* and Takemi Yanagimoto**

Two Bayesian prediction problems in the context of model averaging are investigated by adopting dual Kullback-Leibler divergence losses, the $e$-divergence and the $m$-divergence losses. We show that the optimal predictors under the two losses are shown to satisfy interesting saddlepoint-type equalities. Actually, the optimal predictor under the $e$-divergence loss balances the log-likelihood ratio and the loss, while the optimal predictor under the $m$-divergence loss balances the Shannon entropy difference and the loss. These equalities also hold for the predictors maximizing the log-likelihood and the Shannon entropy respectively under the $e$-divergence loss and the $m$-divergence loss, showing that enlarging the log-likelihood and the Shannon entropy moderately will lead to the optimal predictors. In each divergence loss case we derive a robust predictor in the sense that its posterior risk is constant by minimizing a certain convex function. The Legendre transformation induced by this convex function implies that there is inherent duality in each Bayesian prediction problem.

Key words and phrases: Bayesian prediction, convex function, deviance information criterion, duality, Kullback-Leibler divergence, Legendre transformation, model averaging, information theory, Shannon entropy, thermodynamic principles.

1. Introduction

Model averaging has attracted the interest of many researchers as an alternative to model selection since the choice of a single model among several competing models might be misleading. There is a large body of literature including Draper (1995), Raftery et al. (1997), Buckland et al. (1997), Hoeting et al. (1999), Brown et al. (2002), Hjort and Claeskens (2003) and Hansen (2007). The basic idea is to average estimators across the Bayesian models at hand. Bayesian model averaging averages estimators by posterior probabilities while frequentist model averaging uses the weights that are calculated from frequentists’ viewpoints. See Burnham and Anderson (2002) and Claeskens and Hjort (2008).

In this article we investigate the Bayesian prediction problem in the context of model averaging. Bayesian prediction has also been investigated by many authors such as Ng (1980), Harris (1989), Komaki (1996, 2006), Corcuera and Giannmole (1999) and Yanagimoto and Ohnishi (2009, 2011). See also the monograph by Aitchison and Dunsmore (1975) and Geisser (1993). When we have a Bayesian model $p(x; \theta)\pi(\theta)$, the prediction problem estimates the density $p(y; \theta)$

*Faculty of Economics, Kyushu University, 6-19-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. Email: ohnishi@econ.kyushu-u.ac.jp
**Department of Industrial and Systems Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan.
for future random variable $y$ by a predictor $q(y \mid x)$ with $x$ being the data. It is a generalization of the estimation problem where we estimate $p(y; \theta)$ by $p(y; \hat{\theta})$ where $\hat{\theta}$ is an estimator. We will adopt the two losses $D(q(y \mid x), p(y; \theta))$ and $D(p(y; \theta), q(y \mid x))$ where $D(q(y), p(y))$ stands for the Kullback-Leibler divergence from $q(y)$ to $p(y)$. They are said to be dual to each other, and are called the $e$-divergence and the $m$-divergence losses, respectively. See Nagaoka and Amari (2000, p. 55).

One of the key points of this article is the equality that a predictor satisfies. We present three examples. The first is the maximum likelihood estimator (MLE) in an exponential family density $p(x; \theta)$. As in Kullback (1959), it holds that

$$\log \frac{p(x; \hat{\theta}_M)}{p(x; \theta)} = D(p(y; \hat{\theta}_M), p(y; \theta)) \quad \text{for any } x \text{ and } \theta. \quad (1.1)$$

The left-hand side $\log \{p(x; \hat{\theta})/p(x; \theta)\}$ is the log-likelihood ratio that is desired to be as large as possible according to the maximum likelihood principle. The right-hand side $D(p(y; \hat{\theta}), p(y; \theta))$ is the loss for the estimator $\hat{\theta}$ and should be as small as possible. Since the MLE $\hat{\theta}_M$ balances both the quantities and reminds us of the saddlepoint of a multivariate function, we will call this type of equality a saddlepoint equality. The authors believe that (1.1) shows the merits of the exponential family as a statistical model since the equality does not hold in other models in general.

The second example is the James-Stein estimator $\hat{\theta}_S$ (Stein (1981)) satisfying

$$\mathbb{E} \left[ \log \frac{p(x; \hat{\theta}_S)}{p(x; \theta)} - D(p(y; \hat{\theta}_S), p(y; \theta)) \bigg| p(x; \theta) \right] = 0 \quad \text{for any } \theta, \quad (1.2)$$

where $p(x; \theta)$ is the $n$-variate normal density with mean $\theta$ and the identity covariance matrix. Here $n \geq 3$, and $\mathbb{E}[f(x) \mid p(x)]$ denotes the expectation of $f(x)$ under the density $p(x)$. We think that (1.2) is interesting, because it reveals the inferiority of the MLE. According to the maximum likelihood principle we should maximize the likelihood, but (1.1) and (1.2) suggest that we should enlarge it moderately.

The third example is found in the Bayesian prediction problem. Yanagimoto and Ohnishi (2009) derived

$$\mathbb{E} \left[ \log \frac{q^e(x \mid x)}{p(x; \theta)} - D(q^e(y \mid x), p(y; \theta)) \bigg| \pi(\theta \mid x) \right] = 0 \quad \text{for any } x, \quad (1.3)$$

where $q^e(y \mid x)$ is the optimal predictor under the $e$-divergence loss and $\pi(\theta \mid x)$ is the posterior density.

A statistical implication of (1.1)–(1.3) is that we can easily construct a Bayesian information criterion for the model selection. Let us consider the following condition for a predictor $q(y \mid x)$:

$$\mathbb{E} \left[ \log \frac{q(x \mid x)}{p(x; \theta)} - D(q(y \mid x), p(y; \theta)) \bigg| p(x; \theta)\pi(\theta) \right] = 0.$$
Note that this condition is satisfied if (1.1), (1.2) or (1.3) is satisfied. Yanagimoto and Ohnishi (2011) derived a Bayesian information criterion that is quite close to the deviance information criterion (DIC) by Spiegelhalter et al. (2002) as follows. A calculation verifies the following two equalities:

\[
E \left[ \log \frac{p(y; \theta)}{q(y \mid x)} - D(q(z \mid x), p(z; \theta)) \right] \bigg| p(y; \theta) = 0, \\
E \left[ \log \frac{p(x; \theta)}{p(y; \theta)} \right] \bigg| p(y; \theta)p(x; \theta) = 0.
\]

Therefore, if \(q(y \mid x)\) satisfies the saddlepoint condition, we obtain

\[
E \left[ \log \frac{q(x \mid x)}{q(y \mid x)} - D(q(z \mid x), p(z; \theta)) - D(q(z \mid x), p(z; \theta)) \right] \bigg| p(y; \theta)p(x; \theta)\pi(\theta) = 0.
\]

This provides us with a kind of unbiased estimator of \(\log q(y \mid x)\) as \(\log q(x \mid x) - p_D/2 - q_D/2\) where

\[
p_D = 2E[D(q(z \mid x), p(z; \theta)) \mid \pi(\theta \mid x)] \\
q_D = 2E[D(p(z; \theta), q(z \mid x)) \mid \pi(\theta \mid x)].
\]

Our criterion and the DIC are expressed as \(-2\log q(x \mid x) + p_D + q_D\) and \(-2\log q(x \mid x) + 2p_D\), respectively.

Another key point of this article is duality, and we mean that in the following two senses. One is the duality between the \(e\)-divergence and the \(m\)-divergence. A notable finding in this article is that the Shannon entropy appears as the counterpart of the log-likelihood. While the three saddlepoint equalities (1.1)–(1.3) concern the likelihood and the \(e\)-divergence loss, the pairing of the Shannon entropy and the \(m\)-divergence loss will be observed later. The duality between the log-likelihood and the Shannon entropy is interesting, since the entropy is one of the most fundamental quantities in physics and information theory and so is the log-likelihood in statistics. In statistics the maximization of the Shannon entropy has been applied to characterizing some probability densities. For example, the normal density is the one with the maximum entropy among the densities having the same first and second moments. See Kagan et al. (1973, Ch. 13).

The other is a well-known duality, which is observed between the canonical parameter \(\eta\) and the mean parameter \(\mu\) in the exponential family density \(p(x; \eta) = \exp\{\eta x - \psi(\eta)\}a(x)\). See Nagaoka and Amari (2000, p. 65). The Legendre transformation and Convex functions play a crucial role. Actually, the cumulant function \(\psi(\eta)\) is convex, and its Legendre transform is \(\phi(\mu) = \mu\eta - \psi(\eta)\) with \(\eta\) satisfying \(\mu = \psi'(\eta)\). We will show that dual parameterization casts a clear light on the predictors as the dual parameters \(\eta\) and \(\mu\) in the exponential family density. Duality is also closely related to conjugacy, which was discussed with regard to a generalized version of (1.1) by Ohnishi and Yanagimoto (2010).
The Legendre transformation is a useful mathematical tool and is often encountered in the theory of thermodynamics and analytical mechanics. A standard textbook of thermodynamics such as Callen (1985, Ch. 5) states that thermodynamic problems tend to be remarkably simple in the appropriate formulation while they tend to be remarkably complicated in an inappropriate formulation. In thermodynamics we apply equivalent principles such as the energy minimum principle and the Helmholtz potential minimum principle according to our situation. The Legendre transformation is incorporated as a fundamental aspect of thermodynamic theory. See also Chapter 6 of Callen (1985). We emphasize that the Bayesian prediction problem is another example where the Legendre transformation works well. Roughly speaking, we will verify the following two facts through the Legendre transformation: 1) Minimizing the risk in the $e$-divergence case is equivalent to maximizing the Shannon entropy; 2) minimizing the risk in the $m$-divergence case is equivalent to maximizing the log-likelihood.

The remainder of this article is organized as follows. Section 2 sets the model averaging framework for the two Bayesian prediction problems. Section 3 derives saddlepoint equalities in model averaging under the $e$-divergence loss. Dual parameters, which are essentially the same as the canonical and the mean parameters in the exponential family, are introduced to shed a clear light on the predictors. Section 4, which is ‘dual’ to Section 3, derives saddlepoint equalities implying the balance between the Shannon entropy difference and the $m$-divergence loss. Also, dual parameters are introduced which are linked with each other through a convex function. Section 5 reveals the background of such dual parameterization and pursues a similar structure between the Bayesian model averaging and thermodynamics. A numerical example is presented in Section 6. All the proofs of lemmas are given in the Appendix.

2. Formulation of two Bayesian prediction problems

Let $x \in \mathcal{X} \subset \mathbb{R}^n$ be a random vector and $\theta \in \Theta \subset \mathbb{R}^m$ be a parameter vector. Suppose that we have $K (\geq 2)$ Bayesian models $\{p_k(x; \theta)\pi_k(\theta)\}$ indexed by $k \in \{1, \ldots, K\}$. All the $K$ prior densities are assumed to be proper. The marginal and the posterior densities in the $k$th model are given respectively by

$$m_k(x) = \int_\Theta p_k(x; \theta)\pi_k(\theta)d\theta \quad \text{and} \quad \pi_k(\theta \mid x) = \frac{p_k(x; \theta)\pi_k(\theta)}{m_k(x)}.$$ 

Following the usual setting for Bayesian model averaging (Hoeting et al. (1999)), we aggregate $K$ Bayesian models into a Bayesian model in the following manner:

1) Let $k$ be a random variable taking $1, \ldots, K$ with probability $\lambda_1, \ldots, \lambda_K$, respectively, where $\sum_{k=1}^K \lambda_k = 1$. We will call $\{\lambda_k\}$ the prior probabilities.

2) Define the sampling and the prior densities as $p(x; \theta, k) = p_k(x; \theta)$ and $\pi(\theta, k) = \lambda_k\pi_k(\theta)$, respectively.

Here we assume that the true model is a linear combination of the $K$ Bayesian models with weights $\{\lambda_k\}$. In Sections 3 and 4 we discuss the case where we
mis-specify these weights. The corresponding marginal and posterior densities are calculated respectively as

\[
  m(x) = \sum_{k=1}^{K} \lambda_k m_k(x) \quad \text{and} \quad \pi(\theta, k \mid x) = \frac{\lambda_k m_k(x) \pi_k(\theta \mid x)}{m(x)}.
\]

The posterior probabilities are calculated as

\[
  \pi(k \mid x) = \int_{\Theta} \pi(\theta, k \mid x) d\theta = \frac{\lambda_k m_k(x)}{m(x)} \quad (k = 1, \ldots, K).
\]

Under this aggregated Bayesian model, we first investigate a Bayesian prediction problem

\[
  \min_{q(y \mid x)} \mathbb{E}[D(q(y \mid x), p(y; \theta, k)) \mid p(x; \theta, k) \pi(\theta, k)].
\]

In the above \(D(q(y), p(y))\) is the Kullback-Leibler divergence defined by \(\mathbb{E}[\log \{q(y)/p(y)\} \mid q(y)]\), where \(\mathbb{E}[f(x) \mid p(x)]\) denotes the expectation of \(f(x)\) under the density \(p(x)\).

Let us show that (2.3) is equivalent to

\[
  \min_{q(y \mid x)} \sum_{k=1}^{K} \pi(k \mid x) D(q(y \mid x), q_k^e(y \mid x)),
\]

where

\[
  q_k^e(y \mid x) = \frac{1}{c_k(x)} \exp\{\mathbb{E}[\log p_k(y; \theta) \mid \pi_k(\theta \mid x)]\}
\]

with \(c_k(x)\) being the normalizing constant. Corcuera and Giummole (1999) proved that (2.5) is the solution to the \(k\)th Bayesian prediction problem

\[
  \min_{q(y \mid x)} \mathbb{E}[D(q(y \mid x), p_k(y; \theta)) \mid p_k(x; \theta) \pi_k(\theta)],
\]

and called it the \(e\)-mixture predictor. Note that \(\pi(\theta, k \mid x) = \pi(k \mid x) \pi_k(\theta \mid x)\) and that \(p(x; \theta, k) \pi(\theta, k) = \pi(k \mid x) \pi_k(\theta \mid x) m(x)\). The posterior risk corresponding to the Bayes risk in (2.3) is expressed as

\[
  \mathbb{E}[D(q(y \mid x), p(y; \theta, k)) \mid \pi(k \mid x) \pi_k(\theta \mid x)]
\]

\[
  = \sum_{k=1}^{K} \pi(k \mid x) \mathbb{E}[D(q(y \mid x), p_k(y; \theta)) \mid \pi_k(\theta \mid x)].
\]

In order to show the equivalence, we only have to show that the quantity

\[
  \mathbb{E}[D(q(y \mid x), p_k(y; \theta)) \mid \pi_k(\theta \mid x)] - D(q(y \mid x), q_k^e(y \mid x))
\]
The equivalence is verified by almost the same calculation as in the $\varepsilon$-divergence case. We use the following Pythagorean relationship

$$E[PD(p_k(y; \theta), q_k^m(y \mid x), q(y \mid x)) \mid \pi_k(\theta \mid x)] = 0 \quad \text{for any} \quad x \in \mathcal{X},$$

which was presented in Yanagimoto and Ohnishi (2009). Note that $E[D(p_k(y; \theta), q_k^m(y \mid x)) \mid \pi_k(\theta \mid x)]$ does not depend on the choice of $q(y \mid x)$. 

Since $E[D(q_k^m(y \mid x), p_k(y; \theta)) \mid \pi_k(\theta \mid x)]$ is independent of the choice of the predictor $q(y \mid x)$, the desired equivalence is proved.

Next, we deal with model averaging in a Bayesian prediction problem under the $m$-divergence loss $D(p(y; \theta, k), q(y \mid x))$. The $m$-divergence loss has been adopted by many authors such as Aitchison (1975), Ng (1980), Harris (1989) and Komaki (1996, 2006). Let us show that the following Bayesian prediction problem

$$\min_{q(y \mid x)} E[D(p(y; \theta, k), q(y \mid x)) \mid p(x; \theta, k)\pi(\theta, k)]$$

is equivalent to

$$\min_{q(y \mid x)} \sum_{k=1}^{K} \pi(k \mid x)D(q_k^m(y \mid x), q(y \mid x)).$$

In the above $q_k^m(y \mid x)$ is the $m$-mixture predictor defined by

$$q_k^m(y \mid x) = E[p_k(y; \theta) \mid \pi_k(\theta \mid x)],$$

which is, as proved by Aitchison (1975), the optimal predictor for the $k$th Bayesian prediction problem

$$\min_{q(y \mid x)} E[D(p_k(y; \theta), q(y \mid x)) \mid p_k(x; \theta)\pi_k(\theta)].$$

The equivalence is verified by almost the same calculation as in the $\varepsilon$-divergence case. We use the following Pythagorean relationship

$$E[PD(p_k(y; \theta), q_k^m(y \mid x), q(y \mid x)) \mid \pi_k(\theta \mid x)] = 0 \quad \text{for any} \quad x \in \mathcal{X},$$

which was presented in Yanagimoto and Ohnishi (2009). Note that $E[D(p_k(y; \theta), q_k^m(y \mid x)) \mid \pi_k(\theta \mid x)]$ does not depend on the choice of $q(y \mid x)$.
3. Log-likelihood and \( e \)-divergence loss

This section derives a version of (1.3) in the context of model averaging. In order to make the discussion of this section clear, we first generalize the minimization problem (2.4). By allowing that \( \eta_k \) is negative, we tackle the minimization problem

\[
(3.1) \quad \min_{q(y)} \sum_{k=1}^{K} \eta_k D(q(y), p_k(y)),
\]

where \( \eta_1, \ldots, \eta_K \) are known parameters such that \( \sum_{k=1}^{K} \eta_k = 1 \), and \( p_1(y), \ldots, p_K(y) \) are known densities satisfying the following condition:

\[
(3.2) \quad \sum_{k=1}^{K} \eta_k \log \frac{p_k(y)}{p_K(y)} \text{ is constant if and only if } (\eta_1, \ldots, \eta_{K-1}) = (0, \ldots, 0).
\]

The density defined by

\[
(3.3) \quad f^e(y; \eta) = \frac{1}{c(\eta)} \prod_{k=1}^{K} \left\{ \frac{p_k(y)}{c(\eta)} \right\}^{\eta_k} \exp \left\{ \sum_{k=1}^{K-1} \eta_k \log \frac{p_k(y)}{p_K(y)} \right\},
\]

where \( \eta = (\eta_1, \ldots, \eta_{K-1}) \) and \( c(\eta) \) is the normalizing constant, will play a crucial role. We will call \( f^e(y; \eta) \) the \( e \)-mixture density of \( \{p_k(y)\} \) with weight \( \eta \). This is a \((K-1)\)-parameter exponential family density, the sufficient statistics and the canonical parameters of which are respectively \( \log \left\{ \frac{p_k(y)}{p_K(y)} \right\} \) and \( \eta_k \) \((k = 1, \ldots, K-1)\). We can define the corresponding mean parameter \( \mu = (\mu_1, \ldots, \mu_{K-1}) \). The standard theory of the exponential family density states that \( \eta \) and \( \mu \) are linked by

\[
(3.4) \quad \mu_k = \frac{\partial}{\partial \eta_k} \log c(\eta) \quad (k = 1, \ldots, K-1).
\]

This means that \( \log c(\eta) \) is the cumulant function in the exponential family density. Actually, we have

\[
(3.5) \quad \frac{\partial}{\partial \eta_k} \log c(\eta) = \frac{1}{c(\eta)} \int_X \left\{ \log \frac{p_k(y)}{p_K(y)} \right\} \frac{p_k(y)}{p_K(y)} \exp \left\{ \sum_{l=1}^{K-1} \eta_l \log \frac{p_l(y)}{p_K(y)} \right\} dy
\]

\[
= \mathbb{E} \left[ \log \frac{p_k(y)}{p_K(y)} \mid f^e(y; \eta) \right].
\]

As the relation (3.4) indicates, there is an underlying Legendre transformation, which will be discussed in Section 5. Another expression of \( \mu_k \) is given as

\[
(3.6) \quad \mu_k = D(f^e(y; \eta), p_K(y)) - D(f^e(y; \eta), p_k(y)) \quad (k = 1, \ldots, K-1).
\]

This is verified by noting that \( \log \left\{ \frac{f^e(y; \eta)}{p_K(y)} \right\} - \log \left\{ \frac{f^e(y; \eta)}{p_k(y)} \right\} = \log \left\{ \frac{p_k(y)}{p_K(y)} \right\} \).

We give three lemmas that are essential to prove the subsequent theorems.
Lemma 3.1.

(i) The Pythagorean relationship

\[ \sum_{k=1}^{K} \eta_k \text{PD}(q(y), f^e(y; \eta), p_k(y)) = 0 \]  

holds for any density \( q(y) \). Therefore, \( f^e(y; \eta) \) is the solution to (3.1).

(ii) The density \( f^e(y; \eta) \) satisfies

\[ \sum_{k=1}^{K} \eta_k \left\{ \log \frac{f^e(x; \eta)}{p_k(x)} - \text{D}(f^e(y; \eta), p_k(y)) \right\} = 0 \quad \text{for any} \quad x \in \mathcal{X}. \]

Lemma 3.2. Assume that for any \( x \in \mathcal{X} \) the maximizer \( \hat{\eta}_M(x) \) of \( \log f^e(x; \eta) \) exists and is given as the solution to

\[ \frac{\partial}{\partial \eta_k} \log f^e(x; \eta) = 0 \quad (k = 1, \ldots, K - 1). \]

Then, it holds that

\[ \log \frac{f^e(x; \hat{\eta}_M(x))}{p_k(x)} - \text{D}(f^e(y; \hat{\eta}_M(x)), p_k(y)) \]

\[ = 0 \quad \text{for any} \quad x \in \mathcal{X} \text{ and } k \in \{1, \ldots, K\}. \]

Recall that \( \log \{p_k(y)/p_K(y)\} \) and \( \mu_k \) are the sufficient statistic and the corresponding mean parameter respectively. Therefore, the \( k \)th component of \( \hat{\mu}_M(x) \) corresponding to \( \hat{\eta}_M(x) \) is given as

\[ \hat{\mu}_{Mk}(x) = \log \frac{p_k(x)}{p_K(x)}. \]

The minimum in (3.1) is expressed as a function of \( \eta \) as

\[ \sum_{k=1}^{K} \eta_k \text{D}(f^e(y; \eta), p_k(y)) = -\log c(\eta). \]

See the proof of Lemma 3.1.

Lemma 3.3. Assume that the maximizer \( \hat{\eta}_C \) of \( -\log c(\eta) \) exists and is given as the solution to

\[ \frac{\partial}{\partial \eta_k} \log c(\eta) = 0 \quad (k = 1, \ldots, K - 1). \]

Then, it holds for any set \( \{\tilde{\eta}_k\} \) of real numbers satisfying \( \sum_{k=1}^{K} \tilde{\eta}_k = 1 \) that

\[ \sum_{k=1}^{K} \tilde{\eta}_k \text{D}(f^e(y; \hat{\eta}_C), p_k(y)) = -\log c(\hat{\eta}_C). \]
The maximum in Lemma 3.3 is known as the \textit{Chernoff information} in information theory. See Cover and Thomas (2006, Ch. 11).

Let us return to the Bayesian prediction problem (2.4). As Corcuera and Giummole (1999) showed, its solution is given by the \(\varepsilon\)-mixture predictor
\[
q^\varepsilon(y \mid x) \propto E[\log p(y; \theta, k) \mid \pi(\theta, k \mid x)].
\]
To show that this solution is expressed as a type of average of \(\{q_k^\varepsilon(y \mid x)\}\), we define the \(\varepsilon\)-mixture density of \(\{q_k^\varepsilon(y \mid x)\}\) with weight \(\eta\) as
\[
f^\varepsilon(y \mid x; \eta) = q_K^\varepsilon(y \mid x) \exp \left\{ \sum_{k=1}^{K-1} \eta_k \log \frac{q_k^\varepsilon(y \mid x)}{q_K^\varepsilon(y \mid x)} \right\},
\]
where \(c(\eta; x)\) is the normalizing constant.

\textbf{Theorem 3.1.} Define \(\hat{\eta}_B(x) = (\pi(1 \mid x), \ldots, \pi(K-1 \mid x))\).

(i) The Pythagorean relationship
\[
\sum_{k=1}^{K} \pi(k \mid x) PD(q(y \mid x), f^\varepsilon(y \mid x; \hat{\eta}_B(x)), q_k^\varepsilon(y \mid x)) = 0 \quad \text{for any } x \in \mathcal{X}
\]
holds where \(q(y \mid x)\) is an arbitrary predictor. Therefore, \(f^\varepsilon(y \mid x; \hat{\eta}_B(x))\) is optimal.

(ii) The predictor \(f^\varepsilon(y \mid x; \hat{\eta}_B(x))\) satisfies
\[
(3.10) \sum_{k=1}^{K} \pi(k \mid x) \left\{ \log \frac{f^\varepsilon(x \mid x; \hat{\eta}_B(x))}{q_k^\varepsilon(x \mid x)} - D(f^\varepsilon(y \mid x; \hat{\eta}_B(x)), q_k^\varepsilon(y \mid x)) \right\} = 0 \quad \text{for any } x \in \mathcal{X}.
\]

\textbf{Proof.} (i) Apply Lemma 3.1(i) to (2.4).

(ii) Let \(z\) be an arbitrary value that is in general different from \(x\). An application of Lemma 3.1(ii) to the density \(f^\varepsilon(z \mid x; \eta)\) gives
\[
(3.11) \sum_{k=1}^{K} \pi(k \mid x) \left\{ \log \frac{f^\varepsilon(z \mid x; \hat{\eta}_B(x))}{q_k^\varepsilon(z \mid x)} - D(f^\varepsilon(y \mid x; \hat{\eta}_B(x)), q_k^\varepsilon(y \mid x)) \right\} = 0.
\]
The equality (3.10) is a special case. □

The equality (3.10) will be called the \textit{averaged saddlepoint equality}. By setting \(z = x\) in (3.11), we find that the first term in the bracket looks to be a Bayesian log-likelihood ratio.

We give a definition of a predictor that behaves like the MLE in an exponential family density. Recall that the saddlepoint equality (1.1) holds without being averaged. Replacing the variable \(y\) with the data \(x\) in \(\log f^\varepsilon(y \mid x; \eta)\), we will regard it as a Bayesian log-likelihood function. Then we define \(\hat{\eta}_M(x)\) by
\[
(3.12) \hat{\eta}_M(x) = \arg\max_{\eta} \log f^\varepsilon(x \mid x; \eta)
\]
when it makes sense.

**Theorem 3.2.** Under the assumption in Lemma 3.2 the predictor \( f^e(y \mid x; \hat{\eta}_M(x)) \) satisfies

\[
\log \frac{f^e(x \mid x; \hat{\eta}_M(x))}{q_k^e(x \mid x)} - D(f^e(y \mid x; \hat{\eta}_M(x)), q_k^e(y \mid x)) = 0 \quad \text{for any} \quad x \in \mathcal{X} \text{ and } k \in \{1, \ldots, K\}.
\]

**Proof.** Applying Lemma 3.2, we can prove this theorem in a slightly more general setting. Define \( \tilde{\eta}(z; x) \) by

\[
\tilde{\eta}(z; x) = \arg\max_{\eta} \log f^e(z \mid x; \eta).
\]

Here \( z \) is chosen as an arbitrary value as in the proof of Theorem 3.1(ii). It follows from (3.8) that the mean parameter value corresponding to \( \tilde{\eta}(z; x) \) has its \( k \)th components of the form

\[
\tilde{\mu}_k(z; x) = \log \frac{q_k^e(z \mid x)}{q_k^e(z \mid x)}.
\]

It follows from Lemma 3.2 that

\[
\log \frac{f^e(z \mid x; \tilde{\eta}(z; x))}{q_k^e(z \mid x)} - D(f^e(y \mid x; \tilde{\eta}(z; x)), q_k^e(y \mid x)) = 0 \quad \text{for any} \quad z \in \mathcal{X} \text{ and } k \in \{1, \ldots, K\}.
\]

Set \( z = x \) and note that \( \tilde{\eta}(x; x) = \hat{\eta}_M(x) \).

The equality in this theorem will be called the absolute saddlepoint equality.

It should be noted that \( f^e(y \mid x; \hat{\eta}_B(x)) \) and \( f^e(y \mid x; \hat{\eta}_M(x)) \) satisfy the following condition:

\[
(3.13) \quad \mathbb{E} \left[ \log \frac{f^e(x \mid x; \hat{\eta}(x))}{q_k^e(x \mid x)} - D(f^e(y \mid x; \hat{\eta}(x)), q_k^e(y \mid x)) \middle| \pi(k \mid x)m(x) \right] = 0.
\]

Using an equality of this type, Yanagimoto and Ohnishi (2011) defined a saddlepoint predictor, and discussed its connection to DIC by Spiegelhalter et al. (2002). The following theorem explains the two extreme roles played by the two predictors.

**Theorem 3.3.** The predictors \( f^e(y \mid x; \hat{\eta}_B(x)) \) in Theorem 3.1 and \( f^e(y \mid x; \hat{\eta}_M(x)) \) in Theorem 3.2 are respectively the best and the worst among the predictors \( f^e(y \mid x; \hat{\eta}(x)) \) satisfying (3.13).

**Proof.** Since \( f^e(y \mid x; \hat{\eta}_B(x)) \) is the best among all the predictors, it is also the best among the predictors satisfying (3.13). From the definition of \( \hat{\eta}_M(x) \) in (3.12) we have

\[
\log f^e(x \mid x; \hat{\eta}_M(x)) \geq \log f^e(x \mid x; \hat{\eta}(x)) \quad \text{for any} \quad x \in \mathcal{X},
\]
where $f^e(y \mid x; \hat{\eta}(x))$ is any predictor satisfying (3.13). Applying Theorem 3.2, we have

$$E[D(f^e(y \mid x; \hat{\eta}(x)), q_k^e(y \mid x)) \mid \pi(k \mid x)m(x)]$$

$$= E \left[ \log \frac{f^e(x \mid x; \hat{\eta}(x))}{q_k^e(x \mid x)} \mid \pi(k \mid x)m(x) \right]$$

$$\geq E \left[ \log \frac{f^e(x \mid x; \hat{\eta}(x))}{q_k^e(x \mid x)} \mid \pi(k \mid x)m(x) \right]$$

$$= E[D(f^e(y \mid x; \hat{\eta}(x)), q_k^e(y \mid x)) \mid \pi(k \mid x)m(x)]$$

which completes the proof. □

From a Bayesian viewpoint we are supposed to use the weight $\hat{\eta}_B(x)$ that is essentially the posterior probabilities $\{\pi(k \mid x)\}$. However, note that the optimality of $f^e(y \mid x; \hat{\eta}_B(x))$ depends on the correct specification of the prior probabilities $\{\lambda_k\}$ and that we may not specify them correctly. Here we give such a robust predictor that has a constant posterior risk regardless of the choice of $\{\lambda_k\}$. We see from (3.9) that

$$-\log c(\eta; x) = \sum_{k=1}^{K} \eta_k D(f^e(y \mid x; \eta), q_k^e(y \mid x)),$$

which is the minimum value of $\sum_{k=1}^{K} \eta_k D(q(y \mid x), q_k^e(y \mid x))$. Define

$$(3.15) \quad \hat{\eta}_C(x) = \arg \max_{\eta} \{-\log c(\eta; x)\}$$

under the assumption in Lemma 3.3. Let $\{\hat{\lambda}_k\}$ denote the true prior probabilities. It follows from (2.1) and (2.2) that the corresponding posterior probabilities $\{\tilde{\pi}(k \mid x)\}$ are given by $\hat{\lambda}_km_k(x)/\sum_{l=1}^{K} \hat{\lambda}_lm_l(x)$. Applying Lemma 3.3 we obtain the following theorem.

**Theorem 3.4.** It holds for any prior probabilities $\{\hat{\lambda}_k\}$ that

$$\sum_{k=1}^{K} \tilde{\pi}(k \mid x)D(f^e(y \mid x; \hat{\eta}_C(x)), q_k^e(y \mid x)) = -\log c(\hat{\eta}_C(x); x).$$

We close this section by calculating the mean parameter values $\hat{\mu}_M(x)$ and $\hat{\mu}_C(x)$ corresponding respectively to $\hat{\eta}_M(x)$ in (3.12) and $\hat{\eta}_C(x)$ in (3.15). It follows from the proof of Theorem 3.2 that

$$(3.16) \quad \hat{\mu}_{Mk}(x) = \log \frac{q_k^e(x \mid x)}{q_K^e(x \mid x)} \quad (k = 1, \ldots, K - 1).$$

The right-hand side can be regarded as a Bayesian log-likelihood ratio of the $k$th model over $K$th model. The mean parameter value $\hat{\mu}_C(x)$ corresponding to $\hat{\eta}_C(x)$ is given by

$$(3.17) \quad \hat{\mu}_{Ck}(x) = 0 \quad (k = 1, \ldots, K - 1).$$
4. Shannon entropy and $m$-divergence loss

This section reveals an interesting aspect of the Shannon entropy of a density $p(x)$, which is defined by $H[p(x)] = E[- \log p(x) \mid p(x)]$. Indeed, duality between the Shannon entropy and the log-likelihood will be shown through saddlepoint equalities. To elucidate the heart of the subsequent discussion, we investigate the minimization problem

$$\min_{q(y)} \sum_{k=1}^{K} \eta_k D(p_k(y), q(y)),$$

which is a counterpart to (3.1). Here we assume that

$$p_1(y) - p_K(y), \ldots, p_{K-1}(y) - p_K(y) \text{ are linearly independent.}$$

The ordinary mixture density defined by

$$f^m(y; \eta) = \sum_{k=1}^{K} \eta_k p_k(y) = p_K(y) + \sum_{k=1}^{K-1} \eta_k \{p_k(y) - p_K(y)\},$$

with $\eta = (\eta_1, \ldots, \eta_{K-1})$, will play a vital role. In contrast to (3.3), we call this the $m$-mixture density of $\{p_k(y)\}$ with weight $\eta$. Motivated by (3.6), we define a new parameter vector $\mu = (\mu_1, \ldots, \mu_{K-1})$ by

$$\mu_k = D(p_K(y), f^m(y; \eta)) - D(p_k(y), f^m(y; \eta)) \quad (k = 1, \ldots, K-1).$$

We will call $\eta$ and $\mu$ the canonical and the entropy parameters, respectively. The reason why $\mu$ is called the entropy parameter is that another expression of $\mu_k$ is given by

$$\mu_k = H[p_k(y)] - H[p_K(y)] - \frac{\partial}{\partial \eta_k} H[f^m(y; \eta)].$$

To show this, we use the following formula for two densities $p(y)$ and $q(y)$

$$D(p(y), q(y)) + H[p(y)] + E[\log q(y) \mid p(y)] = 0,$$

which will be frequently used in the remainder of this article. The right-hand side of (4.4) is expressed as

$$H[p_k(y)] - H[p_K(y)] + E[\log f^m(y; \eta) \mid p_k(y) - p_K(y)].$$

We see from (4.3) that

$$\frac{\partial}{\partial \eta_k} f^m(y; \eta) = p_k(y) - p_K(y)$$

and therefore that

$$\frac{\partial}{\partial \eta_k} H[f^m(y; \eta)] = - \int_X \{1 + \log f^m(y; \eta)\} \frac{\partial}{\partial \eta_k} f^m(y; \eta) dy.$$
\[
= -\int_{\mathcal{X}} \{1 + \log f^m(y; \eta)\} \{p_k(y) - p_K(y)\} dy \\
= E[- \log f^m(y; \eta) \mid p_k(y) - p_K(y)].
\]

(4.8)

The expression (4.5) follows from (4.7) and (4.8).

We obtain three lemmas that are essential for the subsequent theorems.

**Lemma 4.1.**

(i) The Pythagorean relationship

\[
\sum_{k=1}^{K} \eta_k PD(p_k(y), f^m(y; \eta), q(y)) = 0
\]

holds for any \(q(y)\). Therefore, \(f^m(y; \eta)\) is the solution to (4.1).

(ii) The density \(f^m(y; \eta)\) satisfies

\[
\sum_{k=1}^{K} \eta_k \{H[f^m(y; \eta)] - H[p_k(y)] - D(p_k(y), f^m(y; \eta))\} = 0.
\]

**Lemma 4.2.** Assume that the maximizer \(\hat{\eta}_M\) of \(H[f^m(y; \eta)]\) exists and is obtained as the solution to

\[
\frac{\partial}{\partial \eta_k} H[f^m(y; \eta)] = 0 \quad (k = 1, \ldots, K - 1).
\]

Then, \(f^m(y; \hat{\eta}_M)\) satisfies

\[
H[f^m(y; \hat{\eta}_M)] - H[p_k(y)] - D(p_k(y), f^m(y; \hat{\eta}_M)) = 0 \quad \text{for any} \quad k \in \{1, \ldots, K\}.
\]

Since \((\partial/\partial \eta_k)H[f^m(y; \eta)]\) vanishes at \(\eta = \hat{\eta}_M\), the \(k\)th component of the corresponding entropy parameter value is given by

\[
\hat{\mu}_{MK} = H[p_k(y)] - H[p_K(y)].
\]

(4.9)

We define \(\psi(\eta)\) by the negative of the minimum of (4.1), i.e.,

\[
(4.10) \quad -\psi(\eta) = \sum_{k=1}^{K} \eta_k D(p_k(y), f^m(y; \eta)).
\]

A calculation using (4.3) and (4.6) gives

\[
-\psi(\eta) = - \sum_{k=1}^{K} \eta_k H[p_k(y)] - E[\log f^m(y; \eta) \mid f^m(y; \eta)]
\]

\[
= - \sum_{k=1}^{K} \eta_k H[p_k(y)] + H[f^m(y; \eta)]
\]
\[ (4.11) \quad = -H[p_K(y)] - \sum_{k=1}^{K-1} \eta_k \{ H[p_k(y)] - H[p_K(y)] \} + H[f^m(y; \eta)]. \]

It is seen from (4.5) that
\[ (4.12) \quad \mu_k = \frac{\partial}{\partial \eta_k} \psi(\eta) \quad (k = 1, \ldots, K - 1). \]

This is the same as the relationship between the canonical and the mean parameters in (3.4), showing duality between \( \eta \) and \( \mu \).

**Lemma 4.3.** Assume that the maximizer \( \hat{\eta}_C \) of \( -\psi(\eta) \) exists and is obtained as the solution to
\[ \frac{\partial}{\partial \eta_k} \psi(\eta) = 0 \quad (k = 1, \ldots, K - 1). \]

It holds for any set \( \{ \tilde{\eta}_k \} \) of real numbers satisfying \( \sum_{k=1}^{K} \tilde{\eta}_k = 1 \) that
\[ \sum_{k=1}^{K} \tilde{\eta}_k D(p_k(y), f^m(y; \hat{\eta}_C)) = -\psi(\hat{\eta}_C). \]

Note that the corresponding entropy parameter value is \( \hat{\mu}_C = 0 \), which is derived by (4.12).

Let us resume the investigation of the minimization problem (2.8). We define the \( m \)-mixture density of \( \{ p^m_k(y \mid x) \} \) with weight \( \eta \) as
\[ f^m(y \mid x; \eta) = q^m_K(y \mid x) + \sum_{k=1}^{K-1} \eta_k \{ q^m_k(y \mid x) - q^m_K(y \mid x) \}. \]

It follows from (4.4) that the entropy parameter \( \mu \) corresponding to the canonical parameter \( \eta \) has its \( k \)th component of the form
\[ \mu_k = D(q^m_K(y \mid x), f^m(y \mid x; \eta)) - D(q^m_K(y \mid x), f^m(y \mid x; \hat{\eta}_B(x))). \]

Applying Lemma 4.1 to (2.8), we obtain the following theorem.

**Theorem 4.1.** Define \( \hat{\eta}_B(x) = (\pi(1 \mid x), \ldots, \pi(K - 1 \mid x)) \).

(i) The Pythagorean relationship
\[ \sum_{k=1}^{K} \pi(k \mid x) PD(q^m_k(y \mid x), f^m(y \mid x; \hat{\eta}_B(x)), q(y \mid x)) = 0 \quad \text{for any} \quad x \in \mathcal{X} \]
holds where \( q(y \mid x) \) is an arbitrary predictor. Therefore, \( f^m(y \mid x; \hat{\eta}_B(x)) \) is optimal.

(ii) The predictor \( f^m(y \mid x; \hat{\eta}_B(x)) \) satisfies
\[ \sum_{k=1}^{K} \pi(k \mid x) \{ H[f^m(y \mid x; \hat{\eta}_B(x))] - H[q^m_k(y \mid x)] \}
- D(q^m_K(y \mid x), f^m(y \mid x; \hat{\eta}_B(x))) = 0. \]
The maximization of the Shannon entropy leads us to a dual version of Theorem 3.2. Define \( \hat{\eta}_M(x) \) by

\[
\hat{\eta}_M(x) = \arg\max_{\eta} H[f^m(y \mid x; \eta)]
\]

when it makes sense. The following theorem is a direct consequence of Lemma 4.2.

**Theorem 4.2.** Under the assumption in Lemma 4.2 the predictor \( f^m(y \mid x; \hat{\eta}_M(x)) \) satisfies

\[
H[f^m(y \mid x; \hat{\eta}_M(x))] - H[q^m_k(y \mid x)] - D(q^m_k(y \mid x), f^m(y \mid x; \hat{\eta}_M(x))) = 0 \quad \text{for any } x \in \mathcal{X} \text{ and } k \in \{1, \ldots, K\}.
\]

The equalities in Theorem 4.1(ii) and 4.2 will be called the *averaged* and *absolute saddlepoint equalities*, respectively. Note that the Shannon entropy difference and the \( m \)-divergence loss appear instead of the log-likelihood ratio and the \( e \)-divergence loss. From (4.9) the \( k \)th component of \( \hat{\mu}_M(x) \) corresponding to \( \hat{\eta}_M(x) \) is given by

\[
\hat{\mu}_{Mk}(x) = H[q^m_k(y \mid x)] - H[q^m_K(y \mid x)].
\]

**Remark 4.1.** We observe two interesting differences between Sections 3 and 4. One is that while we employed a discussion using an arbitrary value \( z \) in the proofs of Theorems 3.1 and 3.2, such an arbitrary value does not appear in the proofs of Theorems 4.1 and 4.2. Another slight difference is as follows: The averaged saddlepoint equality in Theorem 3.1(ii) is obtained by formally setting \( q(y \mid x) \) as the Dirac mass at \( x \) in the Pythagorean relationship in Theorem 3.1(i). On the other hand, Theorem 4.1(ii) is obtained by setting \( q(y \mid x) \) as the uniform density in Theorem 4.1(i) in the case where \( \mathcal{X} \) is finite. Note that the uniform density is defined \( q(y \mid x) = L^{-1} \) where \( L = \int_{\mathcal{X}} dx \).

The two predictors in Theorems 4.1 and 4.2 play the two extreme roles among the predictors satisfying

\[
E[H[f^m(y \mid x; \hat{\eta}(x))] - H[q^m_k(y \mid x)] - D(q^m_k(y \mid x), f^m(y \mid x; \hat{\eta}(x))) \mid \pi(k \mid x)m(x)] = 0.
\]

**Theorem 4.3.** The predictors \( f^m(y \mid x; \hat{\eta}_B(x)) \) in Theorem 4.1 and \( f^m(y \mid x; \hat{\eta}_M(x)) \) in Theorem 4.2 are respectively the best and the worst among the predictors \( f^m(y \mid x; \hat{\eta}(x)) \) satisfying (4.14).

**Proof.** The proof is parallel to that of Theorem 3.3. A key idea is the following inequality

\[
H[f^m(y \mid x; \hat{\eta}_M(x))] \geq H[f^m(y \mid x; \hat{\eta}(x))],
\]
where \( f^m(y \mid x; \hat{\eta}(x)) \) is any predictor satisfying (4.14). This is verified by the definition of \( \hat{\eta}_M(x) \). \( \square \)

We give a robust predictor that has a constant posterior risk. Set

\[
-\psi(\eta; x) = \sum_{k=1}^K \eta_k D(q^m_k(y \mid x), f^m(y \mid x; \eta)),
\]

which is the minimum of \( \sum_{k=1}^K \eta_k D(q^m_k(y \mid x), q(y \mid x)) \). We define

\[
\hat{\eta}_C(x) = \arg\max_{\eta} \{-\psi(\eta; x)\}
\]

under the assumption in Lemma 4.3. The following theorem is obtained by applying Lemma 4.3.

**Theorem 4.4.** Let \( \{\tilde{\pi}(k \mid x)\} \) be the posterior probabilities corresponding to the prior probabilities \( \{\tilde{\lambda}_k\} \). It holds for any \( \{\tilde{\lambda}_k\} \) that

\[
\sum_{k=1}^K \tilde{\pi}(k \mid x) D(q^m_k(y \mid x), f^m(y \mid x; \hat{\eta}_C(x))) = -\psi(\hat{\eta}_C(x); x).
\]

We can calculate the entropy parameter value \( \hat{\mu}_C(x) \) that corresponds to \( \hat{\eta}_C(x) \) by using (4.12). The \( k \)th component of \( \hat{\mu}_C(x) \) is given by

\[
\hat{\mu}_{Ck}(x) = 0.
\]

5. Duality in Bayesian prediction

In Sections 3 and 4 we introduced the dual parameters for the predictors. Here we reveal the background of such parameterization by giving another perspective on the minimization problems (3.1) and (4.1).

First, we give a minimization problem that is equivalent to (3.1) in the sense that both have the identical solution. We will use the term “equivalence” in this sense in this section. Define the dual parameter spaces \( H \) and \( M \) by

\[
H = \{\eta = (\eta_1, \ldots, \eta_{K-1}) \mid c(\eta) < \infty\},
\]

\[
M = \{\mu = (\mu_1, \ldots, \mu_{K-1}) \mid \eta \in H\},
\]

where \( c(\eta) \) is the normalizing constant in (3.3), and \( \mu \) is linked with \( \eta \) through (3.4). We obtain the following theorem, the proof of which is given in the Appendix.

**Theorem 5.1.**

(i) Under assumption (3.2) \( \log c(\eta) \) is convex.

(ii) The Legendre transform of \( \log c(\eta) \) is expressed as \( \phi(\mu) = D(f^c(y; \eta), p_K(y)) \).
Suppose \( s = (s_1, \ldots, s_{K-1}) \in M \). The minimization problem with \( K - 1 \) constraints

\[
\min_{q(y)} D(q(y), p_K(y)),
\]

subject to \( D(q(y), p_K(y)) - D(q(y), p_k(y)) = s_k (k = 1, \ldots, K - 1) \)

has the solution \( f^*(y; \eta) \), identical to that of (3.1) if and only if \( s = \mu \).

We may call (3.1) and (5.1) respectively the canonical representation and the mean representation of the minimization problem since the minimization problem (5.1) specifies the mean parameter value while (3.1) specifies the canonical parameter value. Recall (3.6), and note that \( s_k \) in (5.1) is the mean of \( \log \{ p_k(y)/p_K(y) \} \) under \( q(y) \). When specifying the weight for the predictor in Theorem 3.1, it is easy to employ the canonical representation or the canonical parameter. In contrast, we may employ the mean representation or the mean parameter when specifying the weight for the predictors in Theorems 3.2 and 3.4.

Let us show that minimizing the risk in the \( \varepsilon \)-divergence case can be regarded as equivalent to maximizing the Shannon entropy. The quantity to be minimized in (5.1) is expressed as

\[
D(q(y), p_K(y)) = \mathbb{E} \left[ \frac{q(y)}{p_K(y)} \log \frac{q(y)}{p_K(y)} \middle| p_K(y) \right],
\]

which is the negative of the Shannon entropy of the density \( q(y)/p_K(y) \) when the supporting measure is \( p_K(y) \). Another way of showing that \( D(q(y), p_K(y)) \) is the negative of the Shannon entropy of \( q(y) \) is as follows: Assuming that \( \mathcal{X} \) is finite, we set \( p_K(y) = L^{-1} \) where \( L = \int_{\mathcal{X}} dx \). We have

\[
D(q(y), p_K(y)) = -H[q(y)] + \log L.
\]

We give an explanation to the weight \( \hat{\mu}_{Mk}(x) \) in (3.16) in the mean representation. An implication is the weighting of the \( k \)th model by a Bayesian log-likelihood \( \log q_k^e(x \mid x) \). Let us examine an implication of this weight. For simplicity we consider the case where \( p_k(y; \theta) \) is an exponential family density. As stated in (1.3), Yanagimoto and Ohnishi (2009) showed

\[
E \left[ \log \frac{q_k^e(x \mid x)}{p_k(x; \theta)} - D(q_k^e(y \mid x), p_k(y; \theta)) \middle| \pi_k(\theta \mid x) \right] = 0 \quad \text{for any} \quad x \in \mathcal{X}.
\]

They also showed the Pythagorean relationship (2.7). Since (1.1) holds for \( p_k(x; \theta) \), we have

\[
\log \frac{p_k(x; \hat{\theta}_{Mk})}{p_k(x; \theta)} - D(p_k(y; \hat{\theta}_{Mk}), p_k(y; \theta)) = 0 \quad \text{for any} \quad x \in \mathcal{X} \text{ and } \theta \in \Theta,
\]
where $\hat{\theta}_{Mk}$ is the MLE. Setting $q(y \mid x) = p_k(y; \hat{\theta}_{Mk})$ in (2.7) and combining this with (5.2) and (5.3), we get
\begin{equation}
\log q_k^e(x \mid x) = \log p_k(x; \hat{\theta}_{Mk}) - D(p_k(y; \hat{\theta}_{Mk}), q_k^e(y \mid x)).
\end{equation}
This expression is slightly similar to the weight that was proposed by Buckland et al. (1997) in the context of frequentist model averaging. They proposed to use Akaike’s information criterion (Akaike (1973)) as the weight by which the estimators are averaged. There is a difference in the multiplication constant in the penalty term. A reason for this seems to lie in whether we are dealing with the predictors or estimators.

The weight $\hat{\mu}_{Ck}(x)$ in (3.17) implies that the predictors are averaged with equal weight in the mean representation. Recall that this is obtained by maximizing $-\log c(\eta; x)$ in (3.14), which is the minimized risk, with respect to $\eta$. This equal weight is to be compared with the equal weight $\lambda_1 = \cdots = \lambda_K = 1/K$, which is sometimes suggested for the prior probabilities $\{\lambda_k\}$ when we lack any information. See for example Burnham and Anderson (2002, Ch. 2). It is the weight in the canonical representation, and is obtained by maximizing the Shannon entropy $-\sum_{k=1}^K \lambda_k \log \lambda_k$ of the prior probabilities.

Next, we give an equivalent problem to the minimization problem (4.1). Define the dual parameter spaces by
\begin{align*}
H &= \{\eta = (\eta_1, \ldots, \eta_{K-1}) \mid H[f^m(y; \eta)] \text{ is finite}\},
M &= \{\mu = (\mu_1, \ldots, \mu_{K-1}) \mid \eta \in H\},
\end{align*}
where $\mu$ are linked with $\eta$ through (4.12). The following theorem is obtained, whose proof is given in the Appendix.

**Theorem 5.2.**
(i) Under assumption (4.2) $\psi(\eta)$ defined in (4.10) is convex.
(ii) The Legendre transform of $\psi(\eta)$ is expressed as $\phi(\mu) = D(p_K(y), f^m(y; \eta))$.
(iii) Suppose that $s = (s_1, \ldots, s_{K-1}) \in M$. The minimization problem with $K-1$ constraints
\begin{equation}
\min_{q(y)} D(p_K(y), q(y)),
\end{equation}
subject to $D(p_K(y), q(y)) - D(p_k(y), q(y)) = s_k(k = 1, \ldots, K - 1)$
has the solution $f^m(y; \eta)$, identical to that of (4.1) if and only if $s = \mu$.

We will call (4.1) and (5.5) respectively the **canonical representation** and the **entropy representation** of the minimization problem. Let us show that the quantity to be minimized in the entropy representation can be regarded as the log-likelihood. Setting $p_K(y)$ as the Dirac mass at $x$ and noting the Shannon entropy of the Dirac mass should be zero, we have
\begin{equation}
D(\delta(y - x), q(y)) = -\log q(x).
\end{equation}
This, together with the discussion after Theorem 5.1, suggests the duality between the log-likelihood and the Shannon entropy. It is interesting to note that the Dirac mass and the uniform density appear here and in Remark 4.1. While the Dirac mass is employed in deriving saddlepoint equality in the ε-divergence case and in rewriting the minimization problem in the m-divergence case, the uniform density is used in deriving saddlepoint equality in the m-divergence case and in rewriting the minimization problem in the ε-divergence case.

The entropy representation enables us to have a clear understanding of \( f^m(y \mid x; \hat{\eta}_M(x)) \) and \( f^m(y \mid x; \hat{\eta}_C(x)) \). We see from \( \hat{\mu}_{Mk}(x) \) in (4.13) that the weight of the \( k \)th model in \( f^m(y \mid x; \hat{\eta}_M(x)) \) is the Shannon entropy \( H[q_k^m(y \mid x)] \) in the entropy representation. Let us examine an implication of this weight. For simplicity we consider the case where \( p_k(y; \theta) \) is the ordinary mixture model.

First, we show the averaged saddlepoint equality for the optimal predictor \( q_k^m(y \mid x) \) in the \( k \)th model

\[
E[H[q_k^m(y \mid x)] - H[p_k(y; \theta)] - D(p_k(y; \theta), q_k^m(y \mid x) \mid \pi_k(\theta \mid x)] = 0 \quad \text{for any} \quad x \in \mathcal{X}.
\]

We use the following Pythagorean relationship given by Yanagimoto and Ohnishi (2009)

\[
E[PD(p_k(y; \theta), q_k^m(y \mid x), q(y \mid x) \mid \pi_k(\theta \mid x)] = 0 \quad \text{for any} \quad x \in \mathcal{X},
\]

where \( q(y \mid x) \) is an arbitrary predictor. A calculation using (4.6) gives

\[
D(p_k(y; \theta), q(y \mid x)) - D(q_k^m(y \mid x), q(y \mid x)) = H[q_k^m(y \mid x)] - H[p_k(y; \theta)] - E[\log q(y \mid x) \mid p_k(y; \theta) - q_k^m(y \mid x)].
\]

Combining (2.9), (5.7) and (5.8), we get (5.6). It holds for the ordinary mixture model that

\[
H[p_k(y; \hat{\theta}_{MK})] - H[p_k(y; \theta)] - D(p_k(y; \theta), p_k(y; \hat{\theta}_{MK})) = 0 \quad \text{for any} \quad \theta \in \Theta,
\]

where \( \hat{\theta}_{MK} \) is the maximizer of the Shannon entropy \( H[p_k(y; \theta)] \). This is verified by the proof of Lemma 4.2. Setting \( q(y \mid x) = p_k(y; \hat{\theta}_{MK}) \) in (5.7) and using (5.6) and (5.9), we obtain

\[
H[q_k^m(y \mid x)] = H[p_k(y; \hat{\theta}_{MK})] - D(p_k(y; \hat{\theta}_{MK}), q_k^m(y \mid x)),
\]

which is parallel to (5.4).

An implication of the weight \( \hat{\mu}_{CK}(x) \) in (4.16) is that the predictors have equal weight in the entropy representation. Recall that this weight is obtained by the maximization of \(-\psi(\eta; x)\) in (4.15), which is the minimized risk, with respect to \( \eta \).

We can point out the following two similarities between our discussion and statistical physics or thermodynamics. One is that the normalizing constant \( c(\eta) \)
is a fundamental quantity and is called a **canonical partition sum** in statistical
physics. See a textbook such as Callen (1985, Ch. 16). The **Helmholtz free energy** \( F \) is given by
\[
F = -k_B T \log Z
\]
where \( Z \) is the canonical partition sum, \( k_B \) is Boltzmann’s constant and \( T \) is the absolute temperature. The thermodynamic quantities are calculated from the canonical partition sum. The other similarity is as follows: What Theorems 5.1 and 5.2 tell us can be compared with the equivalence of the two principles in thermodynamics, the energy minimum principle and the Helmholtz potential minimum principle. See Chapters 5 and 6 of Callen (1985). The Helmholtz potential is a Legendre transform of the (internal) energy, and the minimizer is the **equilibrium state**.

### 6. A numerical example

In order to present quantitative differences among Bayes risks of predictors in study, we consider the model in which \( K = 2 \), the sampling density is the Bernoulli density
\[
p_1(x; \theta) = p_2(x; \theta) = \theta^x (1 - \theta)^{1-x},
\]
and the prior density is the beta density of the form
\[
\pi_k(\theta) = \frac{1}{B(\alpha_k, \beta_k)} \theta^{\alpha_k-1} (1 - \theta)^{\beta_k-1}
\]
with \((\alpha_1, \beta_1) = (0.4, 0.6)\) and \((\alpha_2, \beta_2) = (0.7, 0.3)\). This model is easily tractable and is useful in practice applications. As shown by Yanagimoto and Ohnishi (2009), an exponential family density and an ordinary mixture density are closed respectively under \( e \)-mixture and \( m \)-mixture. Therefore, the Bernoulli density is closed under both \( e \)-mixture and \( m \)-mixture. Bayes risks induced from the \( e \)-divergence and the \( m \)-divergence losses are employed to compare predictors of the forms, \( f^e(y \mid x; \hat{\eta}(x)) \) and \( f^m(y \mid x; \hat{\eta}(x)) \), respectively.

We consider the case where our specified prior probabilities \( \{\lambda_1, 1 - \lambda_1\} \) are not always true. Suppose that the true prior probabilities take the values \( \{\tilde{\lambda}_1, 1 - \tilde{\lambda}_1\} \). In this setting the following five weights are examined. The two weights \( \hat{\eta}_C(x) \) and \( \hat{\eta}_M(x) \) do not depend on the prior probabilities. On the other hand, the weight
\[
(6.1) \quad \hat{\eta}_B(x) = \frac{\lambda_1 m_1(x)}{\lambda_1 m_1(x) + (1 - \lambda_1) m_2(x)}
\]
is dependent on \( \lambda_1 \) and therefore impractical. In order to introduce practical weights, it is necessary to choose a suitable value of \( \lambda_1 \) by the rule of thumb. In the present studies we choose two values, 0.2 and 0.8; the two weights \( \hat{\eta}_{B(0.2)}(x) \) and \( \hat{\eta}_{B(0.8)}(x) \) are given by inserting the values 0.2 and 0.8 into \( \lambda_1 \) in (6.1), respectively. For reference we deal with the weight \( \hat{\eta}_{Bt}(x) \) that is obtained by setting \( \lambda_1 = \tilde{\lambda}_1 \) in (6.1). We compare the Bayesian risks of the predictors corresponding to \( \hat{\eta}_{Bt}(x), \hat{\eta}_{B(0.2)}(x), \hat{\eta}_{B(0.8)}(x), \hat{\eta}_C(x) \) and \( \hat{\eta}_M(x) \) as \( \lambda_1 \) varies. For notational convenience we will identify a weight \( \hat{\eta}(x) \) and the corresponding predictor in the later description on numerical results.
First, we treat the case of the $e$-divergence loss, and compare the predictors of the form $f^e(y \mid x; \hat{\eta}(x))$. Theorem 3.1(i) implies that $\hat{\eta}_{Bt}(x)$ has a minimum Bayes risk for any $\tilde{\lambda}_1$. Figure 1 indicates that the curve of the Bayes risk of $\hat{\eta}_{Bt}(x)$ is concave and has fairly large variation. The two lines depicting the Bayes risks of $\hat{\eta}_{B(0.2)}(x)$ and $\hat{\eta}_{B(0.8)}(x)$ are tangent to the Bayes risk curve for $\hat{\eta}_{Bt}(x)$. Large variation of the Bayes risk of $\hat{\eta}_{Bt}(x)$ yields that their Bayes risks depend largely on $\tilde{\lambda}_1$. Among the four practical predictors, $\hat{\eta}_C(x)$ seems to be promising in this setting. It is quite close to $\hat{\eta}_{Bt}(x)$ for all values of $\tilde{\lambda}_1$ and is superior to $\hat{\eta}_{B(0.2)}(x)$ and $\hat{\eta}_{B(0.8)}(x)$ for large portions of the region for $\tilde{\lambda}_1$. The predictor $\hat{\eta}_M(x)$ behaves poorly compared with the other predictors. But, its risk is close to those of $\hat{\eta}_{B(0.2)}(x)$ and $\hat{\eta}_{B(0.8)}(x)$, when $\tilde{\lambda}_1$ is near 1 and 0, respectively. Actually, omitted figures under different settings for $(\alpha_1, \beta_1)$ and $(\alpha_2, \beta_2)$ show that $\hat{\eta}_M(x)$ has less risk than those of $\hat{\eta}_{B(0.2)}(x)$ and $\hat{\eta}_{B(0.8)}(x)$ for some portion of the region of $\tilde{\lambda}_1$.

Secondly, we compare the Bayes risks for the five predictors of the form $f^m(y \mid x; \hat{\eta}(x))$ under the $m$-divergence loss. The predictors in study are chosen corresponding to the above case. The result is described in Fig. 2. The predictor $\hat{\eta}_C(x)$ again performs favorably by showing almost flat risk in $\tilde{\lambda}_1$. A notable difference from Fig. 1 is that the Bayes risk of $\hat{\eta}_{Bt}(x)$ has small variation in $\tilde{\lambda}_1$. Accordingly, the predictors $\hat{\eta}_{B(0.2)}(x)$ and $\hat{\eta}_{B(0.8)}(x)$ behave satisfactorily. As a result $\hat{\eta}_M(x)$ is much worse than the other predictors.

![Figure 1](image.png)

Figure 1. Numerical Bayes risk comparison of the five predictors of the form $f^e(y \mid x; \hat{\eta}(x))$ under the $e$-divergence loss for various values of $\tilde{\lambda}_1$: $\hat{\eta}_{Bt}(x)$ (thick solid curve), $\hat{\eta}_C(x)$ (thick dashed line), $\hat{\eta}_M(x)$ (thick dot-dashed line), $\hat{\eta}_{B(0.2)}(x)$ (thin solid line) and $\hat{\eta}_{B(0.8)}(x)$ (thin dashed line).
Appendix A

Proof of Lemma 3.1. (i) It follows from (2.6) that

\[
PD(q(y), f^e(y; \eta), p_k(y)) = \mathbb{E}\left[\log\frac{f^e(y; \eta)}{p_k(y)} \mid q(y) - f^e(y; \eta)\right].
\]

This, together with (3.3), gives

(A.1) \[ \sum_{k=1}^{K} \eta_k \log \frac{f^e(y; \eta)}{p_k(y)} = -\log c(\eta), \]

which is constant in \( y \). Therefore, the left-hand side of (3.7) vanishes.

(ii) A calculation using (A.1) gives

\[
\sum_{k=1}^{K} \eta_k D(f^e(y; \eta), p_k(y)) = \sum_{k=1}^{K} \eta_k \mathbb{E}\left[\log\frac{f^e(y; \eta)}{p_k(y)} \mid f^e(y; \eta)\right] = \mathbb{E}\left[\sum_{k=1}^{K} \eta_k \log\frac{f^e(y; \eta)}{p_k(y)} \mid f^e(y; \eta)\right] = -\log c(\eta).
\]

The equality (A.1) with \( y \) being replaced by \( x \) also holds. \( \square \)

Proof of Lemma 3.2. We give a proof based on the saddlepoint equality (1.1). Since \( \hat{\eta}_M(x) \) can be regarded as the MLE in the exponential family density
\[ f^e(y; \eta), \text{ we have} \]
\[
\log \frac{f^e(x; \hat{\eta}_M(x))}{f^e(x; \eta)} - D(f^e(y; \hat{\eta}_M(x)), f^e(y; \eta)) = 0 \quad \text{for any } x \in X \text{ and } \eta \in H,
\]
where \( H \) is the canonical parameter space defined in Section 5. When \( \eta_l = \delta_{lk} \) where \( \delta_{lk} \) is the Kronecker delta, we have \( f^e(x; \eta) = p_k(x) \). Therefore, the equality in Lemma 3.2 is obtained. \( \square \)

**Proof of Lemma 3.3.** It follows from the definition of \( \hat{\eta}_C \) and (3.4) that the corresponding mean parameter value \( \hat{\mu}_C \) is equal to 0. Recalling (3.6), we see that \( D(f^e(y; \hat{\eta}_C), p_k(y)) \) is constant in \( k \in \{1, \ldots, K\} \). Note that (3.9) is valid for \( \eta = \hat{\eta}_C \). Therefore, we have \( D(f^e(y; \hat{\eta}_C), p_k(y)) = -\log c(\hat{\eta}_C) \) for \( k = 1, \ldots, K \). \( \square \)

**Proof of Lemma 4.1.** (i) From (2.6) we have
\[
PD(p_k(y), f^m(y; \eta), q(y)) = E \left[ \log \frac{f^m(y; \eta)}{q(y)} \middle| p_k(y) - f^m(y; \eta) \right].
\]
The weighted sum of the right-hand side with \( \{\eta_k\} \) being the weights vanishes due to (4.3).

(ii) Using (4.6), we have
\[
(A.2) \quad H[f^m(y; \eta)] - H[p_k(y)] - D(p_k(y), f^m(y; \eta))
= E[\log f^m(y; \eta) \mid p_k(y) - f^m(y; \eta)].
\]
Similarly, the weighted sum of the right-hand side vanishes. \( \square \)

**Proof of Lemma 4.2.** We first prove the following equality
\[
(A.3) \quad H[f^m(y; \hat{\eta}_M)] - H[f^m(y; \eta)] - D(f^m(y; \eta), f^m(y; \hat{\eta}_M))
= 0 \quad \text{for any } \eta \in H,
\]
where \( H \) is the canonical parameter space defined in Section 5. A calculation similar to the derivation of (A.2) gives
\[
(A.4) \quad H[f^m(y; \hat{\eta}_M)] - H[f^m(y; \eta)] - D(f^m(y; \eta), f^m(y; \hat{\eta}_M))
= E[\log f^m(y; \hat{\eta}_M) \mid f^m(y; \eta) - f^m(y; \hat{\eta}_M)].
\]
It follows from (4.3) that
\[
(A.5) \quad f^m(y; \eta) - f^m(y; \hat{\eta}_M) = \sum_{k=1}^{K-1} \{\eta_k - \hat{\eta}_{Mk}\} \{p_k(y) - p_K(y)\}.
\]
The definition of \( \hat{\eta}_M \), together with (4.8), gives
\[
(A.6) \quad E[\log f^m(y; \hat{\eta}_M) \mid p_k(y) - p_K(y)] = 0 \quad (k = 1, \ldots, K - 1).
\]
Combining (A.4), (A.5) and (A.6), we obtain (A.3). By considering the case where \( \eta_l = \delta_{lk} \) as in Proof of Lemma 3.2, we obtain the required result. It should be noted that (A.3) is valid for the ordinary mixture model, which is used in Section 5. \( \square \)

**Proof of Lemma 4.3.** The proof is almost the same as that of Lemma 3.3. We see from \( \hat{\mu}_C = 0 \) and (4.4) that all the \( K \) divergences \( D(p_k(y), f^m(y; \hat{\eta}_C)) \) are the same. Setting \( \eta = \hat{\eta}_C \) in (4.10), we have \( D(p_k(y), f^m(y; \hat{\eta}_C)) = -\psi(\hat{\eta}_C) \) for \( k = 1, \ldots, K \). \( \square \)

**Proof of Theorem 5.1.** (i) Differentiating (3.5) with respect to \( \eta_l \) and employing the usual calculation in the exponential family density, we have

\[
\frac{\partial^2}{\partial \eta_k \partial \eta_l} \log c(\eta) = \text{Cov} \left[ \log \frac{p_k(y)}{p_K(y)}, \log \frac{p_l(y)}{p_K(y)} \mid f^e(y; \eta) \right],
\]

where \( \text{Cov}[f(y), g(y) \mid p(y)] \) denotes the covariance of \( f(y) \) and \( g(y) \) under the density \( p(y) \). Let \( a_1, \ldots, a_{K-1} \) be arbitrary real numbers. The variance of \( \sum_{k=1}^{K-1} a_k \log \{p_k(y)/p_K(y)\} \) is the quadratic form

\[
\sum_{k=1}^{K-1} \sum_{l=1}^{K-1} a_k a_l \frac{\partial^2}{\partial \eta_k \partial \eta_l} \log c(\eta),
\]

which should be positive due to assumption (3.2). This implies the convexity of \( \log c(\eta) \).

(ii) Noting (3.3) and (3.5), we calculate the Legendre transform of \( \log c(\eta) \) as

\[
\sum_{k=1}^{K-1} \eta_k \frac{\partial}{\partial \eta_k} \log c(\eta) - \log c(\eta) = \sum_{k=1}^{K-1} \eta_k \left[ \log \frac{p_k(y)}{p_K(y)} \mid f^e(y; \eta) \right] - \log c(\eta)
\]

\[
= \text{E} \left[ \sum_{k=1}^{K-1} \eta_k \log \frac{p_k(y)}{p_K(y)} \mid f^e(y; \eta) \right] - \log c(\eta)
\]

\[
= \text{E} \left[ \log \frac{c(\eta)f^e(y; \eta)}{p_K(y)} \mid f^e(y; \eta) \right] - \log c(\eta)
\]

\[
= D(f^e(y; \eta), p_K(y)).
\]

(iii) Let \( \tilde{\eta} \in H \) be the canonical parameter value corresponding to \( s \). We see from (5.1) that

\[
\text{E} \left[ \log \frac{p_k(y)}{p_K(y)} \mid q(y) \right] = s_k.
\]

This, together with (3.3), gives

\[
\text{E} \left[ \log \frac{f^e(y; \tilde{\eta})}{p_K(y)} \mid q(y) \right] = \text{E} \left[ \sum_{k=1}^{K-1} \tilde{\eta}_k \log \frac{p_k(y)}{p_K(y)} - \log c(\tilde{\eta}) \mid q(y) \right]
\]
\[
\begin{align*}
  &= \sum_{k=1}^{K-1} s_k \bar{\eta}_k - \log c(\bar{\eta}) \\
  &= \phi(s).
\end{align*}
\]

Note the equality
\[
D(q(y), p_K(y)) = E \left[ \log \frac{q(y)}{f^e(y; \bar{\eta})} + \log \frac{f^e(y; \bar{\eta})}{p_K(y)} \right] q(y).
\]

Combining this with (A.7), we have
\[
D(q(y), p_K(y)) = D(q(y), f^e(y; \bar{\eta})) + \phi(s).
\]

Clearly, the solution is \( q(y) = f^e(y; \bar{\eta}) \). This is equal to \( f^e(y; \eta) \) if and only if \( \bar{\eta} = \eta \), i.e., \( s = \mu \). □

**Proof of Theorem 5.2.** (i) It follows from (4.11) that
\[
\frac{\partial^2}{\partial \eta_k \partial \eta_l} \psi(\eta) = -\frac{\partial^2}{\partial \eta_k \partial \eta_l} H[f^m(y; \eta)].
\]

Differentiating both the sides of (4.8) with respect to \( \eta_l \), we have
\[
\frac{\partial^2}{\partial \eta_k \partial \eta_l} \psi(\eta) = E \left[ \frac{p_l(y) - p_K(y)}{f^m(y; \eta)} \left| \frac{p_k(y) - p_K(y)}{f^m(y; \eta)} \right| f^m(y; \eta) \right].
\]

Let \( a_1, \ldots, a_{K-1} \) be arbitrary real numbers. The quadratic form
\[
\sum_{k=1}^{K-1} \sum_{l=1}^{K-1} a_k a_l E \left[ \frac{p_k(y) - p_K(y)}{f^m(y; \eta)} \frac{p_l(y) - p_K(y)}{f^m(y; \eta)} \left| f^m(y; \eta) \right| \right]
\]
\[
= E \left[ \sum_{k=1}^{K-1} \sum_{l=1}^{K-1} a_k a_l \frac{p_k(y) - p_K(y)}{f^m(y; \eta)} \frac{p_l(y) - p_K(y)}{f^m(y; \eta)} \left| f^m(y; \eta) \right| \right]
\]
\[
= E \left[ \left\{ \sum_{k=1}^{K-1} a_k \frac{p_k(y) - p_K(y)}{f^m(y; \eta)} \right\}^2 \left| f^m(y; \eta) \right| \right]
\]

is positive due to the assumption (4.2), which shows the desired convexity.

(ii) A calculation using (4.5) and (4.11) gives
\[
\phi(\mu) = \sum_{k=1}^{K-1} \mu_k \eta_k - \psi(\eta)
\]
\[ \phi(\mu) = \sum_{k=1}^{K-1} \eta_k \frac{\partial}{\partial \eta_k} H[\log f_m(y; \eta)] - H[p_K(y)] + H[f_m(y; \eta)]. \]

Applying (4.8) to the first term, we have

\[ \phi(\mu) = \sum_{k=1}^{K-1} \eta_k E[\log f_m(y; \eta) | p_k(y) - p_K(y)] - H[p_K(y)] + H[f_m(y; \eta)] \]
\[ = E[\log f_m(y; \eta) | f_m(y; \eta) - p_K(y)] - H[p_K(y)] + H[f_m(y; \eta)] \]
\[ = -E[\log f_m(y; \eta) | p_K(y)] - H[p_K(y)] \]
\[ = D(p_K(y), f_m(y; \eta)). \]

The formula (4.6) was used in the last equality.

(iii) The proof is similar to that of Theorem 5.1(iii). We have only to show that

(A.8) \[ D(p_K(y), q(y)) = D(f_m(y; \tilde{\eta}), q(y)) + \phi(s), \]

where \( q(y) \) is the density satisfying the constraints in (5.5) and \( \tilde{\eta} \) be the canonical parameter value corresponding to \( s \). It follows from Part (ii) and (2.6) that

\[ D(p_K(y), q(y)) - D(f_m(y; \tilde{\eta}), q(y)) - \phi(s) \]
\[ = PD(p_K(y), f_m(y; \tilde{\eta}), q(y)) \]
\[ = E \left[ \log \frac{f_m(y; \tilde{\eta})}{q(y)} \mid p_K(y) - f_m(y; \tilde{\eta}) \right]. \]

We use the technique of applying (4.6) to the two divergences twice. First, we see that the \( k \)th constraint in (5.5) is rewritten as

\[ E[\log q(y) | p_k(y) - p_K(y)] = s_k - H[p_k(y)] + H[p_K(y)]. \]

Secondly, we see from (4.4) that

\[ E[\log f_m(y; \tilde{\eta}) | p_k(y) - p_K(y)] = s_k - H[p_k(y)] + H[p_K(y)]. \]

These two, together with (4.3), gives

\[ E \left[ \log \frac{f_m(y; \tilde{\eta})}{q(y)} \mid p_K(y) - f_m(y; \tilde{\eta}) \right] = -\sum_{k=1}^{K-1} \tilde{\eta}_k E \left[ \log \frac{f_m(y; \tilde{\eta})}{q(y)} \mid p_k(y) - p_K(y) \right] \]
\[ = 0, \]

which means (A.8). □

Acknowledgements

The authors are grateful to anonymous referees for their helpful comments, which improved the earlier version of this paper. This research work was supported by Grants-in-Aid for Scientific Research (C) 23500354 and 23500357, and partly by Grants-in-Aid for Scientific Research (B) 21300107 and 23310104.
References


