Dual roles of maximizing likelihood and Shannon entropy in Bayesian prediction

Toshio Ohnishi[1] and Takemi Yanagimoto[2]
[1] Kyushu University, Fukuoka, JAPAN
[2] Chuo University, Tokyo, JAPAN
Corresponding author: Toshio Ohnishi, e-mail: *ohnishi@econ.kyushu-u.ac.jp*

Abstract

The maximization of the likelihood and that of the Shannon entropy are the most famous principles in statistical inference. This paper reveals notable duality of these two important notions in the Bayesian prediction problems. We shed light on this duality through the dual Kullback-Leibler divergence losses. Under the e-divergence loss we find the following: 1) the minimization of the Bayesian risk is equivalent to the maximization of the Shannon entropy under a constraint, and 2) the maximization of the likelihood guarantees the minimum prediction in the sense that it derives the worst member of a class of nice predictors. An equality implying the balance of the log-likelihood ratio and the e-divergence plays an important role, which we call a saddlepoint equality. Dually, under the m-divergence loss the following findings are obtained: 1) the minimization of the Bayesian risk is equivalent to the maximization of the likelihood under a constraint, and 2) the maximization of the Shannon entropy guarantees the minimum prediction by deriving the worst member of a class of nice predictors. An equality showing a balance between the Shannon entropy difference and the m-divergence plays a key role.

Key Words: Kullback-Leibler divergence, Gateaux derivative, model averaging, saddlepoint equality.

1. Introduction.

The following situations are often encountered in actual Bayesian data analyses. 1) We have a sampling density $p(x;\mu,\tau)$ and assume a prior density $\pi(\mu|\tau)$ for a parameter μ of interest. We proceed to assume a prior density $\lambda(\tau)$ for an incidental parameter τ. 2) We have a sampling density $p(x;\mu)$ and assume a prior density $\pi(\mu;x_0,\delta)$ where x_0 and δ are hyper-parameters. We proceed to assume a hyper-prior density $\lambda(x_0,\delta)$.

This can be investigated from a viewpoint of Bayesian model averaging (Hoeting *et al.*, 1999) as follows. Suppose that we have Bayesian models $\{p_\xi(x;\theta)\pi_\xi(\theta)\}$ indexed by $\xi\in\Xi\subset\mathbb{R}^l$, where $x\in\mathcal{X}\subset\mathbb{R}^n$ be a random vector and $\theta\in\Theta\subset\mathbb{R}^m$ be a parameter vector. We assume a density $\lambda(\xi)$ that represents our prior belief for the ξth model, and will call it the *averaging prior density*. A standard Bayesian calculation gives the corresponding posterior belief for the ξth model as

$$\lambda(\xi|x)=\frac{\lambda(\xi)m_\xi(x)}{m(x)}, \tag{1.1}$$

where $m_\xi(x)$ is the marginal density in the ξth model and $m(x)=\int_\Xi\lambda(\xi)m_\xi(x)\,d\xi$. We will call $\lambda(\xi|x)$ the *averaging posterior density*.

We formulate two Bayesian prediction problems

$$\min_{q(y|x)}\mathrm{E}\Big[\mathrm{D}\big(q(y|x),\,p_\xi(y;\,\theta)\big)\;\Big|\;\pi_\xi(\theta|x)\lambda(\xi|x)\Big], \tag{1.2}$$

$$\min_{q(y|x)}\mathrm{E}\Big[\mathrm{D}\big(p_\xi(y;\theta),\,q(y|x)\big)\;\Big|\;\pi_\xi(\theta|x)\lambda(\xi|x)\Big], \tag{1.3}$$

where $\mathrm{D}(q,p)$ is the Kullback-Leibler divergence, $\mathrm{E}[f|p]$ is the expectation of f with respect to p, and $\pi_\xi(\theta|x)$ is the posterior density in the ξth model. The two losses $\mathrm{D}\big(q(y|x),\, p_\xi(y;\theta)\big)$ and $\mathrm{D}\big(p_\xi(y;\theta),\, q(y|x)\big)$ are said to be dual to each other and called the e-divergence and the m-divergence losses, respectively (Amari & Nagaoka, 2000). The aim of this paper is to reveal notable dual roles between the maximization of the likelihood and that of the Shannon entropy. We will shed light on this intrinsic dualistic structure through the dual divergence losses. This is a generalization of Ohnishi & Yanagimoto (2013) dealing with the discrete case of $\lambda(\xi)$. The two Bayesian prediction problems will be dealt with contrastingly in §2 and §3.

We rewrite the Bayesian prediction problems (1.2) and (1.3). The corresponding Bayesian prediction problems in the ξth model are

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q(y|x),\, p_\xi(y;\theta)\big) \;\Big|\; \pi_\xi(\theta|x)\Big],$$
$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(p_\xi(y;\theta),\, q(y|x)\big) \;\Big|\; \pi_\xi(\theta|x)\Big].$$

According to Corcuera & Giummole (1999) and Aitchison (1975), the solutions are respectively given by

$$q^e_\xi(y|x) \propto \exp\Big\{\mathrm{E}\big[\log p_\xi(y;\theta) \;\big|\; \pi_\xi(\theta|x)\big]\Big\}, \tag{1.4}$$
$$q^m_\xi(y|x) = \mathrm{E}\big[\, p_\xi(y;\theta) \;\big|\; \pi_\xi(\theta|x)\,\big]. \tag{1.5}$$

Using the result in Yanagimoto & Ohnishi (2009), we can show that (1.2) and (1.3) are equivalent respectively to

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q(y|x),\, q^e_\xi(y|x)\big) \;\Big|\; \lambda(\xi|x)\Big], \tag{1.6}$$
$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q^m_\xi(y|x),\, q(y|x)\big) \;\Big|\; \lambda(\xi|x)\Big], \tag{1.7}$$

where $\lambda(\xi|x)$ is the posterior averaging density defined in (1.1).

2. Results in the e-divergence loss case

Letting $h(\xi)$ be an appropriate density, we investigate a minimization problem

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q(y|x),\, q^e_\xi(y|x)\big) \;\Big|\; h(\xi)\Big]. \tag{2.1}$$

Note that the Bayesian prediction problem (1.6) is its special case. We will call $h(\xi)$ the *canonical weight*. The following quantities will play vital roles.

Definition 2.1. (i) The *e-mixture density* of $q^e_\xi(y|x)$ in (1.4) with canonical weight $h(\xi)$ is defined by

$$f^e(y|x;h) = \frac{1}{K_x(h)} \exp\Big\{\mathrm{E}\big[\log q^e_\xi(y|x) \;\big|\; h(\xi)\big]\Big\}, \tag{2.2}$$

where $K_x(h)$ is the normalizing constant.
(ii) The *mean weight* corresponding to $h(\xi)$ is defined by

$$t_x(\xi;h) = \mathrm{E}\big[\log q^e_\xi(y|x) \;\big|\; f^e(y|x;h)\big]. \tag{2.3}$$

We may regard $\log q(x|x)$ as a Bayesian version of the log-likelihood and will call it the *Bayesian log-likelihood*. The optimal predictor balances the Bayesian log-likelihood ratio and the e-divergence loss.

Theorem 2.1. *The predictor* (2.2) *is the solution to the minimization problem* (2.1)*, and satisfies*

$$\mathrm{E}\left[\log \frac{f^e(x|x;h)}{q^e_\xi(x|x)} - \mathrm{D}\big(f^e(y|x;h),\, q^e_\xi(y|x)\big) \;\middle|\; h(\xi)\right] = 0 \quad \textit{for any } x \in \mathcal{X}. \tag{2.4}$$

The log-likelihood ratio is desired to be large according to the maximum likelihood principle while the e-divergence loss should be minimized. Therefore we will call (2.4) a *saddlepoint equality*.

The minimization of the Bayesian risk under the e-divergence loss is shown to be equivalent to the maximization of the Shannon entropy in the sense that they have the identical solution. Let $\mathrm{H}[p]$ be the Shannon entropy of the density p.

Theorem 2.2. *The following maximization problem of the Shannon entropy with a constraint has the solution $f^e(y|x;h)$ identical to that of* (2.1) *if and only if $s(\xi) = t_x(\xi;h)$ in* (2.3)*.*

$$\max_{q(y|x)} \mathrm{H}\big[q(y|x)\big],$$
$$\text{subject to } \mathrm{E}\big[\log q^e_\xi(y|x) \;\big|\; q(y|x)\big] = s(\xi) \quad \textit{for any } \xi \in \Xi.$$

If we 'maximizes' the Bayesian log-likelihood, another saddlepoint equality is derived. A functional $F(h)$ of h is said to have an extremum at $h^\dagger$ if its Gateaux derivative at $h^\dagger$ with increment $h - h^\dagger$ vanishes, where h is an arbitrary density.

Theorem 2.3. *Assume that there exists a canonical weight $h^\dagger_x$ at which $\log f^e(x|x;h)$ has an extremum. The predictor $f^e(y|x;h^\dagger_x)$ satisfies*

$$\log \frac{f^e(x|x;h^\dagger_x)}{q^e_\xi(x|x)} - \mathrm{D}\big(f^e(y|x;h^\dagger_x),\, q^e_\xi(y|x)\big) = 0 \quad \textit{for any } x \in \mathcal{X} \textit{ and } \xi \in \Xi.$$

The maximization of the Bayesian log-likelihood guarantees a minimum prediction in the following sense.

Theorem 2.4. *Define h^*_x by $h^*_x(\xi) = \lambda(\xi|x)$ in* (1.1)*, and suppose that $h^\dagger_x$ in Theorem 2.3 maximizes the Bayesian log-likelihood $\log f^e(x|x;h)$. Then, the predictors $f^e(y|x;h^*_x)$ and $f^e(y|x;h^\dagger_x)$ are respectively the best and the worst among the predictors satisfying*

$$\mathrm{E}\left[\log \frac{f^e(x|x;h)}{q^e_\xi(x|x)} - \mathrm{D}\big(f^e(y|x;h),\, q^e_\xi(y|x)\big) \;\middle|\; \lambda(\xi|x)m(x)\right] = 0, \tag{2.5}$$

where $m(x)$ is the grand marginal density in (1.1)*.*

The condition (2.5) establishes such a class of predictors that the optimal predictor $f^e(y|x;h^*_x)$ is its best member and $f^e(y|x;h^\dagger_x)$ is its worst one. Any predictor

in this class has a risk equal to or better than that of $f^e(y|x;h_x^\dagger)$. Yanagimoto & Ohnishi (2011) examined the condition (2.5), and discussed its implication to the information criterion.

We give such a robust predictor that has a constant posterior risk regardless of the choice of the prior averaging density $\lambda(\xi)$. It follows from Theorem 2.1 that $-\log K_x(h)$ is the minimum of the minimization problem (2.1). The 'maximization' of $-\log K_x(h)$ with respect to h leads to this predictor.

Theorem 2.5. *Assume that there exists a canonical weight $h_x^{\rm c}$ at which $\log K_x(h)$ has an extremum. Then, $f^e(y|x;h_x^{\rm c})$ satisfies*

$$\mathrm{D}\big(f^e(y|x;h_x^{\rm c}),\, q_\xi^e(y|x)\big) = -\log K_x(h_x^{\rm c}) \quad \textit{for any } \xi\in\Xi.$$

3. Results in the m-divergence loss case

This section, together with §2, reveals a notable duality between the maximization of the log-likelihood and that of the Shannon entropy. Theorems 3.1 – 3.5 below correspond to Theorems 2.1 – 2.5, respectively. Equalities balancing the Shannon entropy difference and the m-divergence loss play a key role.

We investigate the minimization problem

$$\min_{q(y|x)} \mathrm{E}\Big[\mathrm{D}\big(q_\xi^m(y|x),\, q(y|x)\big) \;\Big|\; h(\xi)\Big], \tag{3.1}$$

which include (1.7) as a special case. The density $h(\xi)$ is called the canonical weight as in §2.

Definition 3.1. (i) The *m-mixture density* of $q_\xi^m(y|x)$ in (1.5) with canonical weight $h(\xi)$ is defined by

$$f^m(y|x;h) = \mathrm{E}\big[q_\xi^m(y|x) \;\big|\; h(\xi)\big]. \tag{3.2}$$

(ii) The *entropy weight* corresponding to $h(\xi)$ is defined by

$$t_x(\xi;h) = -\log f^m(x|x;h) - \mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h)\big). \tag{3.3}$$

The optimal predictor satisfies an interesting equality balancing the Shannon entropy difference and the m-divergence loss.

Theorem 3.1. *The predictor* (3.2) *is the solution to the minimization problem* (3.1)*, and satisfies*

$$\mathrm{E}\Big[\mathrm{H}\big[f^m(y|x;h)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] - \mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h)\big) \;\Big|\; h(\xi)\Big] = 0.$$

The Shannon entropy difference is desired to be large according to the Shannon entropy maximization principle while the m-divergence loss should be minimized. Thus, this is also called a *saddlepoint equality*.

The minimization of the Bayesian risk under the m-divergence loss is proved to equivalent to the maximization of the log-likelihood in the sense that they have the identical solution.

Theorem 3.2. *The following maximization problem of the Bayesian log-likelihood with a constraint has the solution $f^m(y|x;h)$ identical to that of* (3.1)

if and only if $s(\xi) = t_x(\xi; h)$ *in* (3.3).

$$\max_{q(y|x)} \log q(x|x),$$
$$\text{subject to } -\log q(x|x) - \mathrm{D}\big(q_\xi^m(y|x),\, q(y|x)\big) = s(\xi) \quad \textit{for any } \xi \in \Xi.$$

Such a predictor that 'maximizes' the Shannon entropy satisfies another saddlepoint equality.

Theorem 3.3. *Assume that there exists a canonical weight* $h_x^\dagger$ *at which* $\mathrm{H}\big[f^m(y|x;h)\big]$ *has an extremum. Then, the predictor* $f^m(y|x;h_x^\dagger)$ *satisfies*

$$\mathrm{H}\big[f^m(y|x;h_x^\dagger)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] - \mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h_x^\dagger)\big) = 0$$
$$\textit{for any } x \in \mathcal{X} \textit{ and } \xi \in \Xi.$$

Maximizing the Shannon entropy guarantees the minimum prediction in the following sense.

Theorem 3.4. *Define* h_x^* *by* $h_x^*(\xi) = \lambda(\xi|x)$ *as in Theorem 2.4, and suppose that* $h_x^\dagger$ *maximizes* $\mathrm{H}\big[f^m(y|x;h)\big]$. *Then, the predictors* $f^m(y|x;h_x^*)$ *and* $f^m(y|x;h_x^\dagger)$ *are respectively the best and the worst among the predictors satisfying*

$$\mathrm{E}\Big[\mathrm{H}\big[f^m(y|x;h)\big] - \mathrm{H}\big[q_\xi^m(y|x)\big] - \mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h)\big) \;\Big|\; \lambda(\xi|x)m(x)\Big] = 0. \tag{3.4}$$

We learn that the condition (3.4) specifies a class of predictors whose Bayesian risks are equal to or smaller than that of $f^m(y|x;h_x^\dagger)$. Note that the optimal predictor $f^m(y|x;h_x^*)$ is a member of this class.

The scheme in Theorem 2.5 derives a predictor that has a constant posterior risk also in the m-divergence loss case. Let $-\psi_x(h)$ be the minimum of the minimization problem (3.1). It follows from Theorem 3.1 that $\psi_x(h)$ is given by

$$\psi_x(h) = \mathrm{E}\Big[\mathrm{H}\big[q_\xi^m(y|x)\big] \;\Big|\; h(\xi)\Big] - \mathrm{H}\big[f^m(y|x;h)\big].$$

The 'maximization' of $-\psi_x(h)$ yields such a robust predictor.

Theorem 3.5. *Assume that there exists a canonical weight* h_x^{c} *at which* $\psi_x(h)$ *has an extremum. The predictor* $f^m(y|x;h_x^{\mathrm{c}})$ *satisfies*

$$\mathrm{D}\big(q_\xi^m(y|x),\, f^m(y|x;h_x^{\mathrm{c}})\big) = -\psi_x(h_x^{\mathrm{c}}) \quad \textit{for any } \xi \in \Xi.$$

References

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62**, 547-554.

Amari, S-I. and Nagaoka, H. (2000). *Methods of Information Geometry.* American Mathematical Society, Load Island.

Corcuera, J.M. and Giummole F. (1999). A generalized Bayes rule for prediction. *Scandinavian Journal of Statistics.* **26**, 265-279.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382-417.

Ohnishi, T. and Yanagimoto, T. (2013). Twofold structure of duality in Bayesian model averaging. *Journal of the Japan Statistical Society*, *to appear*.

Yanagimoto, T. and Ohnishi, T. (2009). Bayesian prediction of a density function in terms of e-mixture. *Journal of Statistical Planning and Inference*, **139**, 3064-3075.

Yanagimoto, T. and Ohnishi, T. (2011). Saddlepoint condition on a predictor to reconfirm the need for the assumption of a prior distribution. *Journal of Statistical Planning and Inference*, **141**, 1990-2000.