Statistica Sinica **10**(2000), 1313-1323

DECISION-MAKING WITH AUGMENTED ACTION SPACES

Andrew L. Rukhin

University of Maryland at Baltimore County

Abstract: In a multiple decision problem one has to choose the "correct" distribution out of a number of different distributions for an observation $\mathbf{x}$. When $\mathbf{x}$ is a random sample, it is known that the minimum Bayes risk decays at exponential rate, which coincides with that of the minimax risk, and is determined by an information–type divergence between these distributions.

There are situations when it is desirable to allow new possible decisions. For example, if the data $\mathbf{x}$ does not provide enough support to any of the models, one may want to allow a "no-decision" or "rejection" option. Another example of such a situation is the confidence estimation problem where the "correct" decisions correspond to one-point sets, and new non-standard actions are formed by subsets of the parameter space consisting of at least two elements.

In the version of the multiple decision problem with augmented action space, we derive the optimal exponential rate of the minimum Bayes risk, and show that it coincides with the mentioned information–type divergence in the classical multiple decision problem. However, the component of the Bayes risk corresponding to the error occurring when the decision belongs to the standard action space may decrease at a faster exponential rate. In a binomial example the accuracy of two asymptotic formulas for the risks containing oscillating (diverging) factors is compared.

Key words and phrases: Bayes risk, binomial distribution, Chernoff theorem, decision space, error probability, loss function, probabilities of large deviations.

1. Introduction

In a multiple decision problem different probability distributions $P_1, \ldots, P_g$ are given, and the goal is to classify an observation $\mathbf{x}$ as coming from one of these distributions. With the parameter space $\Theta = \{1, \ldots, g\}$, the decision (action) space in this statistical problem can be taken to be $\mathcal{D}_0 = \{d_1, \ldots . d_g\}$ where the decision d_θ has the meaning "distribution P_θ". Assume that $P_1, \ldots, P_g$ are mutually absolutely continuous measures (with respect to a σ-finite measure μ), so that the corresponding densities, $p_1, \ldots, p_g$, can be chosen to be positive on the same set $\{\mathbf{x} : p_\theta(\mathbf{x}) > 0\}$, which does not depend on θ.

Traditionally the performance of a multiple decision rule $\delta(\mathbf{x})$, taking values in the decision space $\mathcal{D}_0$ is measured by the error probabilities

$$P_\theta(\delta(\mathbf{x}) \neq d_\theta) = \sum_{\eta:\eta\neq\theta} P_\theta(\delta(\mathbf{x}) = d_\eta).$$

In some situations a priori probabilities $\pi_1, \dots, \pi_g$ can be assumed to be given. Then the global characteristic of δ is $\sum_\theta \pi_\theta P_\theta(\delta \neq d_\theta)$, which is minimized by the Bayes rule

$$\tilde{\delta}(\mathbf{x}) = \arg\max_\theta \left[\pi_\theta p_\theta(\mathbf{x})\right]. \tag{1}$$

For the "default" uniform prior distribution, $\tilde{\delta}$ is merely the maximum likelihood rule.

For fixed probabilities $\pi = (\pi_1, \dots, \pi_g)$ and $P_1, \dots, P_g$, denote by $e_\pi(\mathcal{D}_0)$ the minimum Bayes risk $\sum_\theta \pi_\theta P_\theta(\tilde{\delta} \neq d_\theta)$. Observe that the form of the rule (1) leads to the formula

$$e_\pi(\mathcal{D}_0) = \int [\min_\eta [\sum_1^g \pi_\theta p_\theta(\mathbf{x}) - \pi_\eta p_\eta(\mathbf{x})] \, d\mu(\mathbf{x}) = 1 - \int \max_\theta [\pi_\theta p_\theta(\mathbf{x})] \, d\mu(\mathbf{x}). \tag{2}$$

When the observation $\mathbf{x} = (x_1, \dots, x_n)$ is formed by a random sample from one of distributions $F_\theta, \theta = 1, \dots, g$, so that $P_\theta = F_\theta \otimes \cdots \otimes F_\theta$ and $p_\theta(\mathbf{x}) = \prod_{j=1}^n f_\theta(x_j)$, the limiting behavior of the minimum Bayes risk (or the minimax risk) is given by the following result which goes back to Chernoff (1952) when $g = 2$.

For real s and two probability distributions P and Q with densities p and q, denote by

$$H_s(P, Q) = \log \int \Big[\frac{d\,P}{d\,Q}\Big]^s d\,Q = \log \int \Big[p(x)\Big]^s \Big[q(x)\Big]^{1-s} d\mu(x) \tag{3}$$

the logarithm of the Hellinger type integral. This information-type divergence is known to be related to the *information number* $K(P, Q) = E^P \log \frac{d\,P}{d\,Q}(X)$ as $K(P, Q) = \frac{d}{ds} H_s(P, Q)|_{s=1}$. Intuitively both $K(P, Q)$ and $-\inf_{s>0} H_s(P, Q)$ are characteristics of the degree of separation (or dissimilarity) between P and Q.

Theorem 1.1. *Assume that $\pi_\theta > 0$ for all $\theta \in \Theta$. Then for any classification rule $\delta = \delta(\mathbf{x})$ based on the random sample $\mathbf{x} = (x_1, \dots, x_n)$ from the family $\mathcal{P} = \{P_\theta = F_\theta \otimes \cdots \otimes F_\theta,\ \theta = 1, \dots, g\}$ one has*

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \log \max_\theta P_\theta(\delta(\mathbf{x}) \neq d_\theta) &\geq \liminf_{n\to\infty} \frac{1}{n} \log \sum_\theta \pi_\theta P_\theta(\delta(\mathbf{x}) \neq d_\theta) \\ \geq \liminf_{n\to\infty} \frac{1}{n} \log \sum_\theta \pi_\theta P_\theta(\tilde{\delta}(\mathbf{x}) \neq d_\theta) &= \max_{\eta\neq\theta} \inf_{s>0} H_s(F_\eta, F_\theta) = \rho. \end{aligned} \tag{4}$$

For an arbitrary g, Theorem 1.1 can be derived from Renyi (1970). This theorem shows that the Bayes risk and the minimax risk cannot tend to zero faster than at an exponential rate. The optimal exponential rate of the minimum

Bayes risk does not depend on positive prior probabilities, coincides with that of the minimax risk, and is determined by the information divergence ρ between probability distributions in the given family $P_1, \ldots, P_g$.

2. Augmented Decision Spaces

In this paper we look at multiple decision problems where the decision space $\mathcal{D}$ has the form $\mathcal{D}_a = \mathcal{D}_0 \bigcup \mathcal{D}_1$ with $\mathcal{D}_0 = \{d_1, \ldots, d_g\}$ having the interpretation of the classical decisions and a finite set $\mathcal{D}_1$ forming the "augmented" part of $\mathcal{D}$, which corresponds to, say $m = |\mathcal{D}_1|$, new alternative decisions.

For example, $\mathcal{D}_1$ may consist of one element d_0,"no-decision" or "rejection" of all distributions $P_\theta, \theta = 1, \ldots, g$ (see Rukhin (1998)). The advantage of this particular decision space is that one can find δ for which the individual error probabilities $P_\theta(\delta(\mathbf{x}) \neq d_\theta)$ are simultaneously small and which minimizes the probability of no-decision (see, for example, Nikulin (1989)). The importance of allowing a "no-decision" option has been illustrated recently by Berger, Boukai and Wang (1997) who demonstrated that in the classical hypothesis testing situation, i.e. when $g = 2$, with this option one can achieve the same error probabilities under frequentist and Bayesian approaches.

Another example is the set $\mathcal{D}_1$ formed by all subsets of Θ containing at least two elements. The space $\mathcal{D}_a$ corresponds to confidence estimation of the discrete parameter θ. While the classical decisions correspond to one-point sets, the additional "non-standard" decisions are formed by larger confidence regions. Such a decision problem is described as a *list scheme* in electrical engineering, Forney (1968). The practical situations where confidence sets are of interest include biometric readings like signatures, face images, fingerprints, which provide basis to a list of possible originators. Similar situations arise in the problems of subset selection (Gupta and Panchapakesan (1979), especially Sec 18.2).

We define the loss function $W(\theta, d)$ in the following way: $W(\theta, d_\theta) = 0$, $W(\theta, d_\eta) = 1$ when $\eta \neq \theta$, $d_\eta \in \mathcal{D}_0, \theta, \eta = 1, \ldots, g$; and to escape trivialities assume that for $d \in \mathcal{D}_1$, $0 < W(\theta, d) \leq 1$.

In our first example with $\mathcal{D}_1 = \{d_0\}$ one has to specify positive numbers $w_\theta = W(\theta, d_0) \leq 1$. In the second example a reasonable loss function can be derived from an array $\omega(\theta, t)$, $t = 0, 1, \ldots$ such that for any θ, $\omega(\theta, 0) = 0$ and $\omega(\theta, t)$ is an increasing sequence in t, representing the loss of including the θ-th model when choosing a subset of size $t + 1$. One can put, with a positive c,

$$W(\theta, d) = \begin{cases} \omega(\theta, |d| - 1), & \theta \in d, \\ \omega(\theta, |d| - 1) + c, & \theta \notin d. \end{cases}$$

Note that for a suitable choice of ω and c, $0 < W(\theta, d) < 1$.

When $g = 2$, in both of these examples, $m = 1$, i.e. there is just one additional decision d_0 which, however, has opposite meanings. In the latter example it is a decision consisting of both d_1 and d_2, while in the former it is the decision rejecting d_1 and d_2. In this situation the form of the Bayes rule is well known (see Problem 2.5 in Devroy, Gyorfi and Lugosi (1996)).

In the general situation, the Bayes rule $\hat{\delta}$ has the following form. For $\theta = 1, \ldots, g$,

$$\{\hat{\delta}(\mathbf{x}) = d_\theta\} = \Big\{\pi_\theta p_\theta(\mathbf{x}) = \max_{\eta:\eta=1,\ldots,g} \pi_\eta p_\eta(\mathbf{x}) \vee \max_{d:d\in\mathcal{D}_1} \sum_{\eta=1}^{g} [1 - W(\eta, d)]\pi_\eta p_\eta(\mathbf{x})\Big\}$$

and for $d_0 \in \mathcal{D}_1$,

$$\begin{aligned}\{\hat{\delta}(\mathbf{x}) = d_0\} &= \Big\{\sum_{\eta=1}^{g} [1 - W(\eta, d_0)]\pi_\eta p_\eta(\mathbf{x}) \\ &= \max_{\eta:\eta=1,\ldots,g} \pi_\eta p_\eta(\mathbf{x}) \vee \max_{d:d\in\mathcal{D}_1} \sum_{\eta=1}^{g} [1 - W(\eta, d)]\pi_\eta p_\eta(\mathbf{x})\Big\}.\end{aligned}$$

Thus, this procedure coincides with the classical one for $g+m$ distributions family with densities $p_d(\mathbf{x}) = \sum_{\eta=1}^{g}[1 - W(\eta, d)]\pi_\eta p_\eta(\mathbf{x}) / \sum_{\eta=1}^{g}[1 - W(\eta, d)]\pi_\eta$, $d \in \mathcal{D}_1$, and appropriate prior probabilities.

Possible ties can be broken in any fashion without affecting the minimum Bayes risk, $e_\pi(\mathcal{D}_a)$, which has the form

$$\begin{aligned}e_\pi(\mathcal{D}_a) &= \sum_\theta \pi_\theta E_\theta W(\theta, \hat{\delta}(\mathbf{x})) = \sum_\theta \sum_{d\in\mathcal{D}_a} \pi_\theta W(\theta, d) P_\theta(\hat{\delta}(\mathbf{x}) = d) \\ &= \int \min_{d\in\mathcal{D}} \Big[\sum_\theta \pi_\theta W(\theta, d) p_i(\mathbf{x})\Big]\, d\mu(\mathbf{x}) \\ &= 1 - \int \Big[\max_\theta \pi_i p_\theta(\mathbf{x}) \vee \max_{d:d\in\mathcal{D}_1} \sum_{\eta=1}^{g} [1 - W(\eta, d)]\pi_\eta p_\eta(\mathbf{x})\Big]\, d\mu(\mathbf{x}). \qquad (5)\end{aligned}$$

It is obvious from the comparison of (5) and (2) that the minimum Bayes risk in the problem with the decision space $\mathcal{D}_a$ cannot exceed the minimum Bayes risk $e_\pi(\mathcal{D}_0)$ in the traditional multiple decision problem, i.e.

$$e_\pi(\mathcal{D}_a) \le 1 - \int \max_\eta [\pi_\eta p_\eta(\mathbf{x})]\, d\mu(\mathbf{x}) = e_\pi(\mathcal{D}_0). \qquad (6)$$

Also, according to (5),

$$\begin{aligned}1 - e_\pi(\mathcal{D}_a) &\le \int \Big[\max_\theta \pi_\theta p_\theta(\mathbf{x}) \vee \max_{d:d\in\mathcal{D}_1} \sum_{\eta=1}^{g} [1 - W(\eta, d)] \max_\theta \pi_\theta p_\theta(\mathbf{x})\Big]\, d\mu(\mathbf{x}) \\ &= \Big[\max_{d:d\in\mathcal{D}_1} \sum_{\eta=1}^{g} [1 - W(\eta, d)] \vee 1\Big]\Big[1 - e_\pi(\mathcal{D}_0)\Big], \qquad (7)\end{aligned}$$

so that a multiple decision problem with large $e_\pi(\mathcal{D}_0)$ necessarily leads to a large value of $e_\pi(\mathcal{D}_a)$. Clearly, if $\max_{d:d\in\mathcal{D}_1}\sum_{\eta=1}^{g}[1-W(\eta,d)]<1$, then $e_\pi(\mathcal{D}_a)=e_\pi(\mathcal{D}_0)$.

Because of (6) and Theorem 1.1, for i.i.d. observations $\mathbf{x}=(x_1,\ldots,x_n)$, the asymptotic decay of the Bayes probability is at least exponential with the rate ρ. We show now that the asymptotic behavior of the probability of non-standard decisions from $\mathcal{D}_1$, and of the minimum Bayes risk $e_\pi(\mathcal{D}_a)$ for any decision space $\mathcal{D}_a$, is the same as in the classical multiple decision problem. However, the component of the Bayes risk corresponding to the error occurring when the decision belongs to $\mathcal{D}_0$ may decrease at a faster exponential rate.

To describe this rate let, for a fixed θ, η_θ denote any parametric value such that $K(F_\theta,F_{\eta_\theta})=\min_{\gamma:\gamma\neq\theta}K(F_\theta,F_\gamma)$ and let $h_\theta=\inf_{s>0}H_s\left(F_{\eta_\theta},F_\theta\right)$. (If η_θ is not defined uniquely, h_θ is the largest of all infima above.) Also put

$$\rho_a=\max_\theta h_\theta.$$

Clearly $\rho_a\le\rho$, and, as we will see, for $g\ge3$ strict inequality is possible.

Theorem 2.1. *Assume that $\pi_\theta>0$ and $0<W(\theta,d)<1$ for all $\theta=1,\ldots,g,d\in\mathcal{D}_1$. Then for the Bayes rule $\hat{\delta}$ taking values in $\mathcal{D}_a$*

$$\lim_{n\to\infty}\frac{1}{n}\log\Big[\sum_\theta\pi_\theta P_\theta(\hat{\delta}(\mathbf{x})\in\mathcal{D}_1)\Big]=\lim_{n\to\infty}\frac{1}{n}\log\Big[\sum_\theta\sum_{d\in\mathcal{D}_1}\pi_\theta W(\theta,d)P_\theta(\hat{\delta}(\mathbf{x})=d)\Big]$$
$$=\lim_{n\to\infty}\frac{1}{n}\log e_\pi(\mathcal{D}_a)=\max_{\eta\neq\theta}\inf_{s>0}H_s\left(F_\eta,F_\theta\right)=\rho. \tag{8}$$

However,

$$\lim_{n\to\infty}\frac{1}{n}\log\Big[\sum_\theta\pi_\theta P_\theta(\hat{\delta}(\mathbf{x})\neq d_\theta,\ \hat{\delta}(\mathbf{x})\in\mathcal{D}_0)\Big]=\rho_a. \tag{9}$$

For any procedure δ with values in $\mathcal{D}_a$,

$$\lim\frac{1}{n}\log\Big[\sum_\theta\pi_\theta P_\theta(\delta(\mathbf{x})\neq d_\theta,\ \delta(\mathbf{x})\in\mathcal{D}_0)+\sum_\theta\sum_{d\in\mathcal{D}_1}\pi_\theta W(\theta,d)P_\theta(\delta(\mathbf{x})=d)\Big]\ge\rho.$$

Proof. The proof of (8) follows closely that of Theorem 2.1 in Rukhin (1998) and is omitted.

To demonstrate (9) notice that for any fixed $\theta,\theta=1,\ldots,g$,

$$P_\theta(\hat{\delta}(\mathbf{x})\neq d_\theta,\ \hat{\delta}(\mathbf{x})\in\mathcal{D}_0)$$
$$\le(g-1)\max_{\eta:\eta\neq\theta}P_\theta\Big(\pi_\theta p_\theta(\mathbf{x})\le\pi_\eta p_\eta(\mathbf{x}),\max_{d\in\mathcal{D}_1}\sum_{\gamma=1}^{g}[1-W(\gamma,d)]\pi_\gamma p_\gamma(\mathbf{x})\le\pi_\eta p_\eta(\mathbf{x})\Big)$$
$$\le(g-1)\max_{\eta:\eta\neq\theta}\max_{d\in\mathcal{D}_1}P_\theta\Big(\pi_\eta p_\eta(\mathbf{x})\ge\max_{\gamma\neq\eta}[1-W(\gamma,d)]\pi_\gamma p_\gamma(\mathbf{x})\Big)$$

$$= (g-1)\max_{\eta:\eta\neq\theta}\max_{d\in\mathcal{D}_1} P_\theta\Big(\sum_{j=1}^n \log\frac{f_\eta}{f_\gamma}(x_j) \geq \log\ \frac{[1-W(\gamma,d)]\pi_\gamma}{\pi_\eta}\ \text{ for all }\ \gamma\neq\eta\Big).$$

By the multivariate Chernoff Theorem (Groeneboom, Oosterhoff and Ruymgaart (1979)) the limit of the logarithm of the probability in the right-hand side divided by n is $\max_{\eta:\eta\neq\theta}\inf_{s_\gamma\geq 0,\gamma\neq\eta}\ \log E_\theta\prod_{\gamma,\gamma\neq\eta}[f_\eta/f_\gamma]^{s_\gamma}$.

To show that this quantity (which will be shown to be equal to h_θ) also provides the lower bound, let again $z_\gamma = \pi_\gamma p_\gamma(\mathbf{x})/\sum_\eta \pi_\eta p_\eta(\mathbf{x})$ denote the posterior probabilities, and let us also fix $\eta, \eta\neq\theta$. With

$$\omega = \min_{d\in\mathcal{D}_1}\frac{W(\eta,d)}{W(\eta,d)+\max_{\gamma:\gamma\neq\eta}[1-W(\gamma,d)]},$$

the event $z_\eta = \max_\gamma z_\gamma > 1-\omega$ implies that $\max_{d\in\mathcal{D}_1}\sum_{\gamma=1}^g[1-W(\gamma,d)]z_\gamma < z_\eta$. Indeed as $\sum_{\gamma:\gamma\neq\eta} z_\gamma < \omega$, for any d, $\omega\max_{\gamma:\gamma\neq\eta}[1-W(\gamma,d)] \leq (1-\omega)W(\eta,d) < W(\eta,d)z_\eta$, so that by the definition of ω,

$$\sum_{\gamma=1}^g[1-W(\gamma,d)]z_\gamma < [1-W(\eta,d)]z_\eta + \omega\max_{\gamma:\gamma\neq\eta}[1-W(\gamma,d)] < z_\eta.$$

Therefore for $g\geq 3$,

$$\begin{aligned}
&P_\theta(\hat{\delta}(\mathbf{x})\neq d_\theta,\ \hat{\delta}(\mathbf{x})\in\mathcal{D}_0)\\
&\geq \max_{\eta:\eta\neq\theta}P_\theta\Big(\pi_\theta p_\theta(\mathbf{x}) < \pi_\eta p_\eta(\mathbf{x}),\ \max_{d\in\mathcal{D}_1}\sum_{\gamma=1}^g[1-W(\gamma,d)]\pi_\gamma p_\gamma(\mathbf{x}) < \pi_\eta p_\eta(\mathbf{x})\Big)\\
&\geq \max_{\eta:\eta\neq\theta}P_\theta\Big(\pi_\eta p_\eta(\mathbf{x}) > (1-\omega)\sum_\gamma \pi_\gamma p_\gamma(\mathbf{x})\Big)\\
&\geq \max_{\eta:\eta\neq\theta}P_\theta\Big(\omega\pi_\eta p_\eta(\mathbf{x}) > (1-\omega)(g-1)\max_{\gamma\neq\eta}\pi_\gamma p_\gamma(\mathbf{x})\Big)\\
&= \max_{\eta:\eta\neq\theta}P_\theta\Big(\sum_{j=1}^n\log\frac{f_\eta}{f_\gamma}(x_j) \geq \log\frac{(g-1)(1-\omega)\pi_\gamma}{\omega\pi_\eta}\ \text{ for all }\ \gamma\neq\eta\Big).
\end{aligned}$$

The multivariate Chernoff Theorem shows again that the logarithm of the latter probability divided by n has the same limit, which we prove now to be equal to h_θ. In other terms we prove that

$$\max_{\eta:\eta\neq\theta}\inf_{s_\gamma\geq 0,\gamma\neq\eta}\log E_\theta\prod_{\gamma:\gamma\neq\eta}\left[\frac{f_\eta}{f_\gamma}\right]^{s_\gamma} = h_\theta. \tag{10}$$

Indeed for each fixed η, the infimum in the left-hand side can be taken only with regard to $s_\gamma, \gamma\neq\eta$, such that the derivative of the convex function

$\log E_\theta \prod_{\gamma:\gamma\neq\eta} [f_\eta/f_\gamma]^{s_\gamma}$ evaluated at $s_\gamma = 0$ is negative. This condition means that $K(F_\theta, F_\eta) < K(F_\theta, F_\gamma)$, and by the definition of η_θ, for any η,

$$\inf_{s_\gamma\geq 0,\gamma\neq\eta_\theta} \log E_\theta \prod_{\gamma:\gamma\neq\eta} \left[\frac{f_{\eta_\theta}}{f_\gamma}\right]^{s_\gamma} = \inf_{s>0} H_s\left(F_{\eta_\theta}, F_\theta\right).$$

On the other hand for any η such that $E_\theta \log f_\eta/f_{\eta_\theta} < 0$,

$$\inf_{s_\gamma\geq 0,\gamma\neq\eta} \log E_\theta \prod_{\gamma,\gamma\neq\eta} \left[\frac{f_\eta}{f_\gamma}\right]^{s_\gamma} \leq \inf_{s\geq 0,t\geq 0} \log E_\theta \left[\frac{f_\eta}{f_{\eta_\theta}}\right]^t \left[\frac{f_\eta}{f_\theta}\right]^s$$

$$= \inf_{s,t} \log E_\theta \left[\frac{f_\eta}{f_{\eta_\theta}}\right]^t \left[\frac{f_\eta}{f_\theta}\right]^s \leq \inf_s \log E_\theta \left[\frac{f_\eta}{f_{\eta_\theta}}\right]^{-s} \left[\frac{f_\eta}{f_\theta}\right]^s = \inf_{s>0} H_s\left(F_{\eta_\theta}, F_\theta\right).$$

The penultimate equality here holds since the (global) infimum of $E_\theta\left[f_\eta/f_{\eta_\theta}\right]^t \left[f_\eta/f_\theta\right]^s$ is attained in the positive quadrant. Thus (10) is established, and $\lim_{n\to\infty} \frac{1}{n} \max_\theta \log P_\theta(\hat{\delta}(\mathbf{x}) \neq d_\theta,\ \hat{\delta}(\mathbf{x}) \in \mathcal{D}_0) = \rho_a$.

The last formula of the theorem is true because the Bayes risk of any procedure δ cannot be smaller than the minimum Bayes risk corresponding to $\tilde{\delta}$.

Theorem 2.1 shows that the optimal rate of the exponential decay of the minimum Bayes risk in the decision problem with a no-decision option coincides with that in the traditional setting. In particular, it is independent of positive prior probabilities and of the values $W(\theta, d)$, $0 < W(\theta, d) < 1$, of the loss function.

As an example consider the situation with three normal distributions on the real line $F_1 = N(0.7, 0.07)$, $F_2 = N(0.15, 0.06)$ and $F_3 = N(0, 1)$. In this case F^θ is a two-parameter exponential family for the vector $(x, -x^2/2)^T$. When η is the mean of a normal distribution and κ is its variance, the natural parameter vector of this exponential family has the form $\theta = (v, w)^T = (\eta/\kappa, 1/\kappa)^T$. Thus with $\theta = (v, w)^T$, the logarithm of the moment generating function is $\chi(\theta) = (v^2/w - \log w)/2$. One has, with $t = (t_1, t_2)^T$, $2K(F^\theta, F^t) = w(\frac{t_1}{t_2} - \frac{u}{w})^2 + \frac{w}{t_2} - \log\frac{w}{t_2} - 1$ and $\inf_{s>0} H_s(F^\theta, F^t) = \chi(s\theta + (1-s)t) - s\chi(\theta) - (1-s)\chi(t)$, where $s, 0 < s < 1$, is found from the condition $(\theta - t)^T\chi'(s\theta + (1-s)t) = \chi(\theta) - \chi(t)$.

Calculation after these formulas shows that $\inf_{s>0} H_s\left(F_1, F_2\right) = -0.2928..$, $\inf_{s>0} H_s\left(F_1, F_3\right) = -0.4785..$, $\inf_{s>0} H_s\left(F_2, F_3\right) = -0.4518..$, while $K(F_1, F_2) = 1.2667.. > K(F_1, F_3) = 0.9871..$, and $K(F_2, F_1) = 1.0860.. > K(F_2, F_3) = 0.9423...$ Thus in this case $\eta_1 = \eta_2 = 3$, but $\rho_a = -0.4518.. < \rho = -0.2928....$

Note that for one-parameter exponential families one has $\rho_a = \rho$.

3. Example: Exact Asymptotics of the Error Probability

According to Theorem 2.1, $n^{-1}\log e_\pi(\mathcal{D}_a)$ and $n^{-1}\log P_\theta(\hat{\delta} \in \mathcal{D}_1)$, have a common limit. However convergence can be fairly slow. When $m = 1$ and the distribution of the likelihood ratios has an absolutely continuous component, the asymptotic expansions for the minimum Bayes risk and the minimax risk are derived in Rukhin (1998). These expansions give much more accurate approximations to the corresponding error probabilities than $\exp\{n\rho\}$.

Here we look at the example of two binomial distributions. Let $g = 2, \pi_1 = 1/2$, $F_1 = Bin(1, p_1), F_2 = Bin(1, p_2)$ with $p_1 < p_2$. It is easy to see that in the classical two-action problem $\tilde{\delta}(x_1, \ldots, x_n) = 1$ if and only if

$$X = x_1 + \cdots + x_n < n \frac{\log \frac{1-p_1}{1-p_2}}{\log \frac{(1-p_1)p_2}{(1-p_2)p_1}} = nq.$$

In this situation with X denoting a binomial random variable, $e_\pi(\mathcal{D}_0) = \frac{1}{2}\left[P_1\left(X \geq nq\right) + P_2\left(X < nq\right)\right]$. Let $c(x)$ denote the ceiling function so that for an integer X, $X \geq nq$ if and only if $X \geq c(nq)$. The asymptotic representation of the binomial distribution function by Fu and Wong (1980) (see also Fu, Leu and Peng (1990)) shows that

$$P_1\left(X \geq nq\right) = P_1\left(X \geq c(nq)\right) = \frac{(1-p_1)\Gamma(n)}{(q-p_1)\Gamma(c(nq))\Gamma(n+1-c(nq))}$$
$$\times \exp\{c(nq)\log p_1 + (n - c(nq))\log(1-p_1)\}\left[1 + O\left(\frac{1}{n}\right)\right]$$

and

$$P_2\left(X < nq\right) = P_2\left(X < c(nq)\right) = P_2\left(X \leq c(nq) - 1\right)$$
$$= \frac{p_2\Gamma(n)}{(q-p_2)\Gamma([nq]+1)\Gamma(n-[nq])}\exp\{c(nq)\log p_2 + (n - c(nq))\log(1-p_2)\}$$
$$\times\left[1 + O\left(\frac{1}{n}\right)\right].$$

By combining these two formulas one obtains, with $\mathbf{I}_p(q) = q\log p + (1-q)\log(1-p)$,

$$e_\pi(\mathcal{D}_0) = \frac{1}{2}\Big[\frac{\Gamma(n)(1-p_1)}{\Gamma(c(nq))\Gamma(n+1-c(nq))(q-p_1)}\exp\left\{n\mathbf{I}_{p_1}\left(\frac{c(nq)}{n}\right)\right\}$$
$$+\frac{\Gamma(n)p_2}{\Gamma(c(nq))\Gamma(n+1-c(nq))(q-p_2)}\exp\left\{n\mathbf{I}_{p_2}\left(\frac{c(nq)}{n}\right)\right\}\Big]\left[1+O\left(\frac{1}{n}\right)\right]. \tag{11}$$

One can derive from (11) the approximation obtained from the asymptotic representation of the binomial distribution function by Bahadur (1960, Corollary 2, p.50). Indeed let

$$\mathbf{H}_p(q) = -\left[q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}\right] = \mathbf{I}_p(q) - \mathbf{I}_q(q)$$

denote the classical relative entropy function. According to this representation

$$e_\pi(\mathcal{D}_0) = \frac{e^{-n\mathbf{H}_{p_1}(q)}\Delta_n}{2\sqrt{2\pi n}}\left[1 + O\left(\frac{1}{n}\right)\right],$$

where

$$\Delta_n = \frac{(1-p_1)\sqrt{q}}{\sqrt{1-q}(q-p_1)}\left[\frac{q(1-p_1)}{p_1(1-q)}\right]^{c(nq)-nq} + \frac{p_2\sqrt{1-q}}{\sqrt{q}(p_2-q)}\left[\frac{q(1-p_2)}{p_2(1-q)}\right]^{nq+1-c(nq)}.$$

This formula immediately shows that $\rho = \mathbf{H}_{p_1}(q) = \mathbf{H}_{p_2}(q)$, which also can be checked directly. More precisely,

$$\frac{\log e_\pi(\mathcal{D}_0)}{n} = \rho - \frac{\log n}{2n} + \frac{1}{n}\log\left(\frac{\Delta_n}{2\sqrt{2\pi}}\right) + O\left(\frac{1}{n^2}\right).$$

However this formula is less accurate numerically than the one obtained from (11). Also observe that the sequence $c(nq) - nq$ does not converge, so that the probabilities of large deviations in general do not have the form $e^{-n\rho}n^{-1/2}b_n$ with a *convergent* sequence b_n.

With only one additional decision d_0 (like 'no-decision' or a confidence set consisting of both d_1 and d_2) when $W(1,d_0)=W(2,d_0)=w < 1/2$, $\hat\delta(x_1,\ldots,x_n)=d_1$ if and only if

$$X < nq - \frac{\log\frac{1-w}{w}}{\log\frac{(1-p_1)p_2}{(1-p_2)p_1}} = nq - r.$$

Also $\hat\delta(x_1,\ldots,x_n) = d_2$ when $X > nq + r$, with the decision d_0 taken if $nq - r \le X \le nq + r$. Therefore,

$$\begin{aligned}
e_\pi(\mathcal{D}_a) &= \frac{1}{2}\Big[P_1(X \ge nq + r) + P_2(X < nq - r)\\
&+w\Big[P_1(nq - r \le X < nq + r) + P_2(nq - r \le X < nq + r)\Big]\Big]\\
&= \frac{1}{2}\Big[(1-w)P_1(X \ge nq + r) + wP_1(X \ge nq - r) + (1-w)P_2(X < nq - r)\\
&+wP_2(X < nq + r)\Big].
\end{aligned}$$

As above,

$$\begin{aligned}
2e_\pi(\mathcal{D}_a) \sim{}& \frac{1-p_1}{q-p_1}\Big[\frac{(1-w)\Gamma(n)}{\Gamma(c(nq+r))\Gamma(n+1-c(nq+r))}\exp\Big\{n\mathbf{H}_{p_1}\Big(\frac{c(nq+r)}{n}\Big)\Big\}\\
&+\frac{w\Gamma(n)}{\Gamma(c(nq-r))\Gamma(n+1-c(nq-r))}\exp\Big\{n\mathbf{H}_{p_1}\Big(\frac{c(nq-r)}{n}\Big)\Big\}\Big]\\
&+\frac{p_2}{q-p_2}\Big[\frac{w\Gamma(n)}{\Gamma(c(nq+r))\Gamma(n+1-c(nq+r))}\exp\Big\{n\mathbf{H}_{p_2}\Big(\frac{c(nq+r)}{n}\Big)\Big\}\\
&+\frac{(1-w)\Gamma(n)}{\Gamma(c(nq-r))\Gamma(n+1-c(nq-r))}\exp\Big\{n\mathbf{H}_{p_2}\Big(\frac{c(nq-r)}{n}\Big)\Big\}\Big]. \qquad (12)
\end{aligned}$$

Figure 1 shows the behavior of the sequences $n^{-1}\log e_\pi(\mathcal{D}_0)$ and $n^{-1}\log e_\pi(\mathcal{D}_a)$ along with the approximations from (11), (12) and Bahadur's approximation when $p_1 = 0.3, p_2 = 0.5, \pi_1 = 1/2$. The constant line there represents the value of $\rho = -0.0213..$; the lowest curve corresponds to $n^{-1}\log e_\pi(\mathcal{D}_a)$ calculated from the exact formula for the binomial distribution function; the upper curve is Bahadur's approximation; the middle curve depicts (12). This Figure shows that (12) provides a much better approximation to the exact value of $e_\pi(\mathcal{D}_a)$ than does Bahadur's approximation.

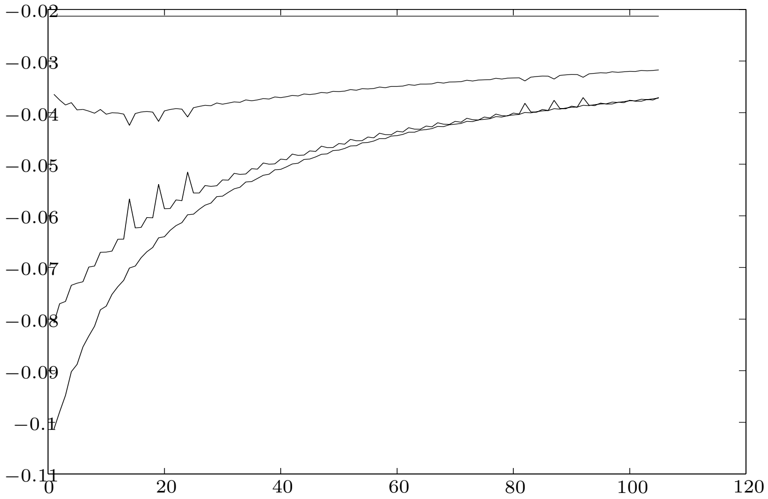

Figure 1. Plots of the sequence $n^{-1}\log e_\pi(\mathcal{D}_a)$ and the approximations from (12) and Bahadur's approximation for $n = 15, \ldots, 120$.

Acknowledgement

This research has been supported by the von Humboldt Stiftung DMS-9000999 award. The author is grateful to the editor and two referees for their patience and help.

References

Bahadur, R. R. (1960). Some approximations to the binomial distribution function. *Ann. Math. Statist.* **31**, 43-54.

Berger, J., Boukai, B. and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypotheses. *Statist. Sci.* **12**, 133-160.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. *Ann. Math. Statist.* **23**, 493-507.

Devroye, L., Gyorfi, L. and Lugosi, G. (1996). *A Probabilistic Theorey of Pattern Recognition.* Springer, New York.

Forney, J. C. D. (1968). Exponential error bounds for erasure, list and decision feedback schemes. *IEEE Trans. Information Theory* IT-14, 206-220.

Fu, J. C., Leu, C. M. and Peng, C. Y. (1990). A numerical comparison of normal and large deviation approximation for tail probabilities. *J. Japan Statist. Soc.* **20**, 61-67.

Fu, J. C. and Wong, R. (1980). An asymptotic expansion of a beta-type integral and its application to probabilities of large deviations. *Proc. Amer. Math. Soc.* **79**, 410-414.

Groeneboom, P., Oosterhoff, J. and Ruymgaart, F. H. (1979). Large deviation theorems for empirical probability measures. *Ann. Probab.* **7**, 553-586.

Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations.* John Wiley, New York.

Nikulin, M. S. (1989). A result of L. N. Bol'shev from the theory of statistical testing of hypotheses. *J. Soviet Math.* **44**, 522-529.

Renyi, A. (1970). On some problems of statistics from the point of view of information theory. In *Proceedings of the Colloquium on Information Theory, Vol 2.* Budapest, Bolyai Math. Soc.

Rukhin, A. L. (1998). Decision making with a no-decision option. *Statist. Decisions* **16**, 259-272.

Department of Mathematics and Statistics, University of Maryland at Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, U.S.A.

E-mail: rukhin@math.umbc.edu

(Received October 1998; accepted December 1999)