On Priors With a Kullback–Leibler Property

Stephen Walker, Paul Damien, and Peter Lenk

1. INTRODUCTION.

Recent Bayesian nonparametric literature has focused on consistency properties of Bayesian procedures (see, e.g., Wasserman 1998; Barron, Schervish, and Wasserman 1999; Ghosal, Ghosh, and Ramamoorthi 1999). Based on the results of these authors and on our own results, we argue that it is recommendable that Bayesians use priors that put positive mass on all Kullback–Leibler neighborhoods of all densities. A Kullback–Leibler neighborhood of size $\varepsilon > 0$ of the density $g$ is given by

$$A_g(\varepsilon) = \left\{ f : \int g(x) \log \frac{g(x)}{f(x)} dx < \varepsilon \right\}.$$ 

The property of the prior $\Pi$ with which we are concerned is that

$$\Pi(A_g(\varepsilon)) > 0$$

for all $\varepsilon > 0$ and all densities $g$. We call this the Kullback–Leibler property for $\Pi$. To achieve this, a nonparametric prior is required. (For specific examples of priors with the above foregoing property, see Barron et al. 1999; Ghosal et al. 1999; Petrone and Wasserman 2002.)

We consider only the case in which $f_0$ is a density function and $X^n = (X_1, \ldots, X_n)$ is available as an independent and identically distributed random sample from $f_0$, the first $n$ observations of a possibly infinite sequence of reals $X_1, X_2, \ldots$. Because $f_0$ is unknown, the Bayesian constructs a prior distribution on the relevant space of density functions, or distribution functions, reflecting available prior information about the location of $f_0$. Assuming that all the densities under consideration are dominated by some $\sigma$-finite measure, which we take to be the Lebesgue measure, Bayes’s theorem and the data $X^n$ combine to update the prior to the posterior.

In this article we demonstrate that the Kullback–Leibler property for a prior $\Pi$ provides good large sample properties for a number of Bayes procedures. Consequently, we argue that Bayesians should be constructing priors with the Kullback–Leibler property, at the very least when there is doubt about the underlying shape of the density function generating the data. The results are based on large samples, highly relevant these days when large datasets are becoming the norm. For example, financial datasets can be recorded in the tens of thousands.

In particular, from a Bayes factor perspective, we demonstrate that if a model has the Kullback–Leibler property, then the Bayes factor always supports this model eventually when compared with any other model without the Kullback–Leibler property. Practically speaking, a model with this property meets the requirements of an "asymptotically true" model.

We should point out that all Bayesian models $M = \{ f(x; \theta), \pi(\theta) \}$ define a prior probability $\Pi$ on the space of density functions. A random density function from $\Pi$ is chosen by first choosing a $\theta$ from $\pi$ and then putting $f(\cdot) \equiv f(\cdot; \theta)$. Hence, for us, a Bayesian model is precisely the prior $\Pi$ on the space of density functions. A parametric model of finite dimensions will not satisfy the Kullback–Leibler property, unless $f_0$ is known to belong to the particular parametric family.

The following reasons suggest that $\Pi$ should have the Kullback–Leibler property:

1. Many practicing statisticians would argue that parametric models are sufficient when combined with model checking and model comparison diagnostics (see, e.g., Bernardo and Smith 1994). However, Draper (1999), in an insightful discussion of the article by Walker et al. (1999), pointed out that allocating probability mass 1 to parametric subsets of densities should not be done lightly. The reason for this is that switching models when the original model(s) under consideration is (are) found to be deficient in some sense exposes the statistician to the very real possibility of poor calibration. Therefore, there is a very practical reason for assigning mass 1 to the set of all densities; the data can offer no surprises. Some authors go even further, pointing out that the allocation of probability 1 to a parametric model is a desire to check this allocation once the data has arrived represents an internal contradiction (see Lindsey 1999).

2. If $\Pi$ has the Kullback–Leibler property, then the Bayes factor comparing this model with any other model will always eventually support (under mild regularity conditions) the prior with the Kullback–Leibler property. The precise result is stated and proved in Section 2. The conclusion is that there is no motivation to put the prior $\Pi$ under the scrutiny of a Bayes factor.

3. Decisions made via the maximization of expected utility are consistent when the prior has the Kullback–Leibler property. This is proved in Section 3. The utility function and $f_0$ define the correct action, which is unknown, just as $f_0$ is unknown. Decisions are consistent if the decision rule eventually converges to this correct action.
4. For those interested in density estimation, there exists a Kullback–Leibler consistent sequence of predictive densities based on a prior with the Kullback–Leibler property. This is stated precisely in Section 4.

Before proceeding, we introduce the notation used throughout the article. Let \( \Pi^n \) denote the posterior distribution given \( X^n \). Then define \( I_n = \int R_n(f) \Pi^n(df) \), \( n \geq 1 \), and \( l_0 = 1 \), where \( R_n(f) = \prod_{i=1}^n f(X_i) / f_0(X_i) \). Define

\[
 f_n(x) = \int f(x) \Pi^n(df)
\]

to be the predictive density, and also define \( D(f, f_0) = \int \log f_0 f / f \) to be the Kullback–Leibler divergence between \( f_0 \) and \( f \). In the following, a.s. is with respect to the infinite product measure \( F_0^{\infty} \).

2. BAYES FACTORS

Bayes factors are widely used in Bayesian model selection problems (see, e.g., Bernardo and Smith 1994 for a review). To date, asymptotic studies of Bayes factors have been formulated only when one of the models is “correct” (see, e.g., Gelfand and Dey 1994). The Bayes factor for comparing model 1 with model 2 is given by

\[
 B_n = I_{1n} / I_{2n},
\]

where \( I_{jn} = \int R_n(f) \Pi_j(df) \), and the Bayesian models \( M_1 \) and \( M_2 \) are fully characterized by the priors \( \Pi_1 \) and \( \Pi_2 \).

Bayesian models, characterized by \( \Pi \), will be associated with a \( \delta_\Pi \geq 0 \). This \( \delta_\Pi \) is such that \( \Pi\{ f : D(f, f_0) < \delta_1 \} > 0 \) only for, and all, \( d > \delta_\Pi \). For such priors, it is expected that \( \liminf_n D(f_n, f_0) \geq \delta_\Pi \) a.s. Let \( \delta_j \geq 0 \) be the value associated with \( \Pi_j \).

**Theorem 1.** If \( \Pi_j \{ f : D(f, f_0) < \delta_1 \} > 0 \) only for, and all, \( d > \delta_j \), and \( \liminf_n D(f_{jn}, f_0) \geq \delta_j \) a.s., then

\[
 n^{-1} \log B_n \to \delta_2 - \delta_1 \quad \text{a.s.,}
\]

provided that \( \sum_n n^{-2} \text{var} [\log(I_{jn}/I_{jn-1})] < \infty \).

**Proof.** Dropping the subscript \( j \), consider the martingale

\[
 S_N = \sum_{n=1}^N [\log(I_n/I_{n-1}) + D(f_{n-1}, f_0)]
\]

which is a martingale by virtue of \( E[\log(I_n/I_{n-1}) | \mathcal{F}_{n-1}] = -D(f_{n-1}, f_0) \). Here \( \mathcal{F}_n = \sigma(X_1, \ldots, X_n) \). For such martingales, it is known that if

\[
 \sum_n n^{-2} \text{var} [\log(I_n/I_{n-1})] < \infty,
\]

then \( S_N / N \to 0 \) a.s. Therefore,

\[
 N^{-1} \log I_N + N^{-1} \sum_{n=1}^N D(f_{n-1}, f_0) \to 0
\]

a.s. Now, by assumption,

\[
 \liminf_N N^{-1} \sum_{n=1}^N D(f_{n-1}, f_0) \geq \delta
\]

a.s., and hence \( \limsup_N N^{-1} \log I_N \leq -\delta \) a.s. With the Kullback–Leibler (\( \delta \)) property for \( \Pi \), we have \( \liminf_N N^{-1} \times \log I_N \geq -\delta \) a.s. (see Barron et al. 1999), and hence

\[
 N^{-1} \log I_N \to -\delta
\]

a.s.

Consequently, \( B_n \to \infty \) a.s. (preferring model 1) if and only if \( \delta_1 < \delta_2 \). This makes sense. Note that the rate will be exponential; that is, \( B_n \sim \exp(n(\delta_2 - \delta_1)) \). Obviously, if \( \delta_1 = 0 \), then \( B_n \to \infty \) a.s. for all \( \delta_2 > 0 \). We present illustrations of Theorem 1 in Section 5.

If \( \Pi \) has the Kullback–Leibler property and the competing model does not, then the Bayes factor will eventually prefer the model that does have the Kullback–Leibler property.

3. BAYESIAN DECISION THEORY

Here we provide further support for the notion that a prior \( \Pi \) should be constructed with the Kullback–Leibler property. Taking the notation from Hirschleifer and Riley (1992), the elements of a decision problem are as follows:

1. A finite set of actions indexed by \( a \); for practical purposes, we assume \( a \in \{1, \ldots, N\} \), for some integer \( N \). Although much of decision theory is written up with the notion of a continuous set of actions, in practice the number of decisions that can be made is finite (see Lindley 1985 for a discussion).

2. A set of states of nature, which we take to be the appropriate space of distribution functions \( F \), say \( \mathcal{F} \). (We assume that the relevant unknown state of nature is the distribution generating the data. This gives us a general framework to work with. Certainly, knowing the true distribution will solve all decision problems associated with the data.)

3. A consequence function \( v(a, F) \) showing outcomes under all combinations of actions and states of nature.

4. A preference scaling function \( v(c) \) measuring the desirability of the consequence \( c \).

5. A probability distribution on \( \mathcal{F} \) representing beliefs in the true state of nature. In a Bayesian context, this probability is the prior \( \Pi \) in the no sample problem and is \( \Pi^n \) once the data \( X^n \) have been observed.

The Von Neumann–Morgenstern (Von Neumann and Morgenstern 1947) expected utility rule then asserts that the best decision is to take the action \( a \) that maximizes

\[
 U_n(a) = \int v(c(a, F)) \Pi^n(dF).
\]

This expected utility rule is applicable if and only if the \( v(\cdot) \) function has been determined in a particular way that leads to \( v(c) \) being bounded, specifically \( 0 \leq v(c) \leq 1 \); that is, the \( v(c) \) has a probabilistic interpretation (see Hirschleifer and Riley 1992). There are differing opinions on the point of a bounded elementary utility function [see, e.g., De Groot (1970) who relaxed the axioms of Von Neumann–Morgenstern]. We point out that with unbounded \( v(\cdot) \), it is not guaranteed that \( U_n(a) \) even exists, and because this depends on \( f_0 \), which is unknown, the bounded \( v(\cdot) \) makes the most sense.

It is not our intention to discuss the expected utility rule further. Our aim is to show that if \( \Pi \) has the Kullback–Leibler
property, then the decision rule eventually converges to the action that maximizes \( U_0(a) = v(c(a, F_0)) \), which can be classified as the correct action, obviously unknown because \( F_0 \) is unknown.

**Theorem 2.** If \( \Pi \) has the Kullback–Leibler property, then \( U_0(a) \rightarrow U_0(\theta) \) a.s. for all \( a \).

**Proof.** If \( \Pi \) has the Kullback–Leibler property, then \( \Pi_n \) converges weakly to \( \Pi_0 \) a.s., where \( \Pi_0 \) is the probability measure with point mass one at \( F_0 \) (see Schwartz 1965). The Portmanteau theorem (see, e.g., Billingsley 1968, thm. 2.1) then gives the desired convergence result for \( U_n(a) \), assuming that \( v \) is suitably smooth. Essentially, we need \( v[c(a, F)] \) to be a continuous function with respect to \( F \); that is, if \( F \) converges weakly to \( F^* \), then \( v[c(a, F)] \rightarrow v[c(a, F^*)] \).

Clearly, if \( U_n(a) \rightarrow U_0(a) \) a.s. for all \( a \), then the maximizer over \( U_n(a) \), say \( a_n \), will eventually converge to \( a_0 \), which maximizes \( U_0(a) \). To see this, if \( a_n \neq a_0 \) infinitely often, then there exists \( a^* \neq a_0 \) such that \( U_n(a^*) > U_n(a_0) \) infinitely often. Such an \( a^* \) exists because of the finiteness of the set of \( a \)'s, and so there will be an accumulation point of the set of \( a_n \) for which \( U_n(a) > U_n(a_0) \). Consequently, to contradict the fact that \( a_0 \) maximizes \( U_0 \) and \( U_n(a) \rightarrow U_0(a) \) a.s. for all \( a \), \( U_0(a^*) = U_0(a_0) \) must hold.

4. **PREDICTIVE DENSITY**

Here we point out that if \( \Pi \) has the Kullback–Leibler property, then there exists a Kullback–Leibler consistent sequence of densities.

**Theorem 3.** Suppose that \( \Pi \) has the Kullback–Leibler property and

\[
\sum_n n^{-2} \text{var}[\log(I_n/I_{n-1})] < \infty.
\]

If

\[
f^N = \frac{1}{N} \sum_{n=1}^{N} f_n,
\]

then \( D(f^N, f_0) \rightarrow 0 \) a.s.

**Proof.** This follows from Theorem 1.

Hence, for those who see density estimation as an important statistical procedure, \( f^N \) is an easily available Kullback–Leibler consistent sequence of densities. Nonparametric predictable densities are often hard to construct, but are not hard to sample from. So if it is possible to sample from \( f_n \), then obviously it is also possible to sample from \( f^N \).

If \( \Pi \) has the Kullback–Leibler property, then the condition

\[
\sum_n n^{-2} \text{var}[\log(I_n/I_{n-1})] < \infty
\]

is an extremely mild one. For if \( \Pi \) does have the Kullback–Leibler property, then it is possible to write \( I_n = \exp(-nt_n) \) and \( t_n \rightarrow 0 \) a.s. (see Barron et al. 1999 for this result). For the foregoing sum to be finite, therefore, it is sufficient that

\[
\sum_n \text{E}(t_n - I_{n-1})^2 < \infty.
\]

Note that if \( \Pi \) has the Kullback–Leibler property, then

\[
d_H(f^N, f_0) \rightarrow 0 \text{ a.s.,}
\]

where \( d_H(f, f_0) \) is the Hellinger distance between \( f \) and \( f_0 \) (see Walker 2003).

5. **ILLUSTRATIONS**

Here we present five examples illustrating the ideas developed in this article.

**Example 1.** In the first example, we take the true density function to be \( f_0(x) = \exp(-x) \). We take model 1 to be \( f_1(x; \theta) = \theta \exp(-\theta x) \) with prior \( \pi_1(\theta) = \exp(-\theta) \) and take model 2 to be fixed at \( f_2(x) = .5 \exp(-.5x) \). It is easy to see that \( \delta_1 = 0 \) and \( \delta_2 = \log 2 - .5 = .193 \). We calculate that

\[
\int \prod_{i=1}^{n} f(x_i) \Pi_1(df) = \frac{n!}{(1 + s_n)^{1+n}},
\]

where \( s_n = \sum_{i=1}^{n} x_i \), and

\[
\int \prod_{i=1}^{n} f(x_i) \Pi_2(df) = (1/2)^n \exp(-s_n/2).
\]

Following a simulation of data from \( f_0 \), Figure 1 plots \( n^{-1} \times \log B_n \) for \( n = 1, \ldots, 3,000 \). The convergence of \( n^{-1} \log B_n \) to the correct value of .193 is evident.

**Example 2.** In this example, we consider the case where both models are wrong, in the sense that neither prior has the Kullback–Leibler property. We now take \( f_0(x) = x \exp(-x) \) and keep model 1 as in the first example. Model 2 is Weibull, \( f_2(x) = x \exp(-x^2/2) \) with \( \pi_2(\theta) = \exp(-\theta) \). Then \( \delta_1 = .116 \) and \( \delta_2 = .099 \), and so \( \delta_2 - \delta_1 = .017 \). Again, we performed a simulation of \( n^{-1} \log B_n \); Figure 2 plots the convergence to the correct value. It should be noted in this case that convergence is slow, and we took 1,000,000 samples. Figure 2 shows every 350th value of \( n^{-1} \log B_n \).

**Example 3.** In this example we take a nonparametric prior, not infinite-dimensional but with a large number of parameters. With samples from \([0, 1]\), we have model 1 as \( f_1(x; \theta) = \theta x^{\theta-1} \) with prior \( \pi_1(\theta) = \exp(-\theta) \). We take model 2 to be a histogram on \( m = 1,000 \) bins, with each bin of length \( 1/m \). The density function is \( f_2(x) = m_{2k} \mathbb{I}(k-1/m < x < k/m) \), and we take \( (q_1, \ldots, q_m) \) to have a Dirichlet prior with parameters all equal to 1. Then

\[
\int \prod_{i=1}^{n} f(x_i) \Pi_1(df) = \frac{n! \prod_{i=1}^{n} x_i^{-1}}{(1 + t_n)^{1+n}}.
\]

![Figure 1](image_url)  \( \log B_n \) \( \rightarrow \) .193.
where \( t_n = -\sum_{i=1}^{n} \log x_i \), and
\[
\int \prod_{i=1}^{n} f(x_i) \prod_{k=1}^{m} \Gamma(n_k + 1),
\]
where \( n_k = \sum_{i=1}^{n} \mathbb{1}(k-1)/m < x_i < k/m \). If \( f_0 \) is uniform on \([0, 1]\), then both \( \delta_1 = 0 \) and \( \delta_2 = 0 \). As can be seen from the simulation of \( n^{-1} \log B_n \) in Figure 3, the Bayes factor always prefers the parametric model, although asymptotically it prefers neither model, because \( n^{-1} \log B_n \to 0 \) a.s.

**Example 4.** In this slight variation of Example 3, we retain model 2 and \( f_0 \) as in Example 3 and take \( f_1(x) = 2x \) to be fixed. Then \( \delta_1 = .306 \), and Figure 4 shows the convergence of \( n^{-1} \log B_n \) to \(-.306\). Note that in this case the Bayes factor always prefers the nonparametric model.

**Example 5.** Here we compare a parametric model with a semiparametric extension. The parametric model for the density function \( f_1 \) with support \( \mathcal{Y} \) is given by
\[
f_1(y|\beta) = \frac{\exp[\sum_{j=1}^{p} h_j(y)\beta_j]}{\int_{\mathcal{Y}} \exp[\sum_{j=1}^{p} h_j(x)\beta_j]dG(x)},
\]
where the \( \{h_j\} \) are known functions on \( \mathcal{Y} \), \( G \) is the known dominating measure, and the \( \{\beta_j\} \) are unknown parameters. One choice of prior for \( \beta_j \) is a multivariate normal distribution. The semiparametric model for the density function \( f_2 \) extends the exponential model via
\[
f_2(y|\beta, Z) = \frac{\exp[\sum_{j=1}^{p} h_j(y)\beta_j + Z(y)]}{\int_{\mathcal{Y}} \exp[\sum_{j=1}^{p} h_j(x)\beta_j + Z(x)]dG(x)},
\]
where \( Z \) is a mean-0, second-order Gaussian process with bounded covariance function on \( \mathcal{Y} \), \( E[Z(x), Z(y)] = \sigma(x, y) \), and \( f_2 ZdG = 0 \), to identify the model. The variance of \( Z \) determines the amount that \( f_2 \) varies from \( f_1 \) in a Kullback–Leibler sense, and its covariance determines the smoothness of the sample paths for \( f_2 \) and the corresponding smoothness of the density estimator (see Whittle 1958).

Lenk (2003) used the Karhunen–Loève expansion (Grenander 1981) to express \( Z \) as an infinite series with random coefficients,
\[
Z(y) = \sum_{k=1}^{\infty} \theta_k \phi_k(y),
\]
where the \( \{\phi_k\} \)’s form an orthogonal basis for the square-integrable functions on \( \mathcal{Y} \) with respect to \( G \). We assume that the \( \{\phi_k\} \)’s are different from the basis functions. The random Fourier coefficients of the expansion are given by
\[
\theta_k = \int_{\mathcal{Y}} Z(y)\phi_k(y)dG(y), \quad k = 1, 2, \ldots.
\]
The basis functions satisfy \( \int_{\mathcal{Y}} \phi_k dG = 0 \) for the sample paths of \( Z \) to integrate to 0. That is, the basis elements are orthogonal to the constant function, which is excluded from the expansion. Because \( Z \) is a mean-0 Gaussian process, the \( \{\theta_k\} \)’s are mutually independent and normally distributed with mean 0. Their standard deviations \( \{\sigma_k\} \) are determined by the expansion of the covariance function,
\[
\sigma(x, y) = \sum_{k=1}^{\infty} \sigma_k^2 \phi_k(x)\phi_k(y),
\]
assuming that \( \sum \sigma_k^2 < \infty \). One choice for the basis functions is the cosine functions,
\[
\phi_k(y) = \sqrt{2} \cos[k\pi G(y)], \quad k = 1, 2, \ldots.
\]
The cosine basis has a natural ordering in terms of the smoothness of the sample paths of \( f_2 \); the Fourier coefficients for high-frequency cosine terms will be relatively small for smooth functions. A smoothing prior for these variances is
\[
\sigma_k^2 = \tau^2 \exp(-k^2y).
\]
where \( \tau^2 \) has the inverted gamma distribution and \( \gamma \) has an exponential distribution.

In practice, the finite-series expansion of \( Z \) is truncated to a finite \( K \), resulting in a model similar to the models of Efron and Tibshirani (1996) and Stone, Hansen, Kooperberg, and Truong (1997). Stone (1990) detailed large-sample properties of the maximum likelihood estimator for log-spline models as the sample size and \( K \) go to infinity at the appropriate rates. The behavior of the maximum likelihood estimator is sensitive to the choice of \( K \). If \( K \) is too large, then sampling variation in high-frequency Fourier coefficients dominates and results in estimators that are too wavy. If \( K \) is too small, then important features of the density are missed. In contrast, the Bayesian analysis can include as many basis functions as possible, up to aliasing, and
shrinks the Fourier coefficients to 0. The amount of shrinkage depends on the smoothing prior for the variances and the sampling variation in the Fourier coefficients.

We performed simulation study to compare the two models $f_1$ and $f_2$. Briefly, the parametric model is a truncated gamma distribution on $(0, 10)$. When the data are generated from the truncated gamma distribution, for sample sizes of 500, the Bayes factor supported the parametric model in 98% of the simulations. Again, this further illustrates the findings of Example 3. If $\delta_1 = \delta_2 = 0$, then the Bayes factor tends to support the reduced model.

6. DISCUSSION

For those interested in subjective issues, consider the following. Walker, Damien, Laud, and Smith (1999) showed that it is possible to take subjective information from a parametric model and incorporate it into a nonparametric model. Then, for those who would acknowledge the existence of $f_0$, we have demonstrated the practical relevance of a nonparametric model. For those who would not accept that an object such as $f_0$ exists, the nonparametric approach using the Kullback–Leibler property offers a surprise-free approach (see point 1 in Sec. 1), and at a minimum avoids the poor calibration that may confront the statistician who is happy to assign probability 1 to a host of possible models (see Draper 1999 for a detailed discussion of this point).

For those concerned with working in high-dimensional spaces, the message from the collection of applied articles edited by Dey, Sinha, and Müller (1998) is that it is no more difficult to routinely implement Bayesian nonparametric procedures than parametric ones, following the advent and rapid growth of user-friendly Markov chain Monte Carlo methods.

Other ideas for avoiding poor calibration include Bayesian model averaging (Draper 1995) and model selection (Draper 2003), both of which are based on a fixed set of models with associated probabilities of plausibility, rather than probabilities of correctness. Practically speaking, it may not be difficult to assign probabilities to models; if there is a finite set, then assigning equal probability is one option. Several researchers have recently pointed out that model averaging usually outperforms model selection, and intuitively it is easy to see why this might be the case. We see model averaging as an attempt to construct a prior with large support (the idea being that at least one of the models may be close to $f_0$) using a collection of parametric models, and this could be seen as being equivalent to a Bayesian nonparametric statistician who makes finite the infinite-dimensional nonparametric model. This often happens, such as in the case of Pólya trees (Lavine 1992) and the infinite-dimensional exponential family; indeed, it is necessary in these cases.

This last comment has implications for the Kullback–Leibler property. For example, for Pólya trees one would take the levels of the tree to be finite, and hence the Kullback–Leibler property would no longer hold. Recalling Section 2, if $\Pi$ is the finite-dimensional approximation to the infinite-dimensional $\Pi$, then there could be a positive $\delta_\Pi$, where

$$\delta_\Pi = \inf \left\{ \int f_0(x) \log \frac{f_0(x)}{f(x)} \, dx : f \text{ in support of } \Pi \right\}.$$

However, this could be made as small as possible with no serious effect on any of the three procedures outlined in the article. For example, we would have

$$\frac{1}{N} \sum_{n=1}^{N} D(f_n, f_0) \rightarrow \delta_\Pi \quad \text{a.s.}$$

under the condition of Theorem 3.

{Received February, 2002. Revised November 2003.}

REFERENCES


