Asymptotically Minimax Regret by Bayes Mixtures

Jun'ichi Takeuchi, *Member, IEEE,* and Andrew R. Barron, *Fellow, IEEE*

Abstract—We study the problems of data compression, gambling and prediction of a sequence $x^n = x_1 x_2 ... x_n$ from an alphabet $\mathcal{X}$, in terms of regret and expected regret (redundancy) with respect to various smooth families of probability distributions. We evaluate the regret of Bayes mixture distributions compared to maximum likelihood, under the condition that the maximum likelihood estimate is in the interior of the parameter space. For general exponential families (including the non-i.i.d. case) the asymptotically mimimax value is achieved when variants of the prior of Jeffreys are used. Interestingly, we also obtain a modification of Jeffreys prior which has measure outside the given family of densities, to achieve minimax regret with respect to non-exponential type families. This modification enlarges the family using local exponential tilting (a fiber bundle). Our conditions are confirmed for certain non-exponential families, including curved families and mixture families (where either the mixture components or their weights of combination are parameterized) as well as contamination models. Furthermore for mixture families we show how to deal with the full simplex of parameters. These results also provide characterization of Rissanen's stochastic complexity.

Index Terms—universal coding, universal prediction, regret, redundancy, exponential family, Bayes mixture, Jeffreys prior

I. Introduction

We study the problem of data compression, gambling and prediction of a string $x^n = x_1, x_2, ..., x_n$ from a given alphabet $\mathcal{X}$, in terms of regret and expected regret (redundancy) with respect to various families of probability distributions. We evaluate the regret and expected regret of Bayes mixture distributions and show that it asymptotically achieves the minimax value when variants of Jeffreys prior are used. Our results contain generalization of the results in [50], [51], [14]. Our results provide evaluation of stochastic complexity defined by Rissanen [37].

This paper's main concern is the regret of a coding or prediction strategy. This regret is defined as the difference of the loss incurred and the loss of an ideal coding or prediction strategy for each sequence. A coding scheme for strings of length n is equivalent to a probability mass function $q(x^n)$ on $\mathcal{X}^n$. We can also use q for prediction and gambling, that is, its conditionals $q(x_{i+1}|x^i)$ provide a distribution for the coding or prediction of the next symbol given the past. The minimax regret with respect to a target family of probability mass functions $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ and a given set of parameters $K \subset \Theta$ is defined as

$$\min_q \max_{x^n \in \mathcal{K}} \max_{\theta \in \Theta} \left(\log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\theta)} \right),$$

where $\mathcal{K}$ denotes the set $\{x^n : \hat{\theta}(x^n) \in K\}$, which is the set of strings for which the maximum likelihood estimate $\hat{\theta}$ is in K. Typically K is either all of Θ or a subset that excludes points near the boundary. The maximum is taken for all strings x^n in $\mathcal{K}$. The regret $\log(1/q(x^n)) - \log(1/p(x^n|\hat{\theta}))$ in the data compression context is also called the (pointwise) redundancy: the difference between the code length based on q and the minimum of the codelengths $\log(1/p(x^n|\theta))$ achieved by distributions in the family. Also, $\log(1/q(x^n)) - \log(1/p(x^n|\theta))$ is the sum of the incremental regrets of prediction $\log(1/q(x_{i+1}|x^i)) - \log(1/p(x_{i+1}|x^i, \theta))$, with what is sometimes called the log-loss. This regret is also called the pointwise regret, to emphasize the distinction from the expected regret also discussed below.

The heart of our analysis is the consideration of Bayes mixtures and the use of the Laplace method to approximate them. A Bayes mixture takes the form $q(x^n) = \int p(x^n|\theta)w(\theta)d\theta$, also called the Bayes factor or the marginal density of x^n obtained by integrating out θ from the joint distribution.

Though certain coding and prediction settings have a discrete alphabet $\mathcal{X}$, we are interested in minimax regret problems also for continuous spaces $\mathcal{X}$. Then the $p(x^n|\theta)$ as well as $q(x^n)$ are understood to be probability density functions with respect to a given reference measure.

For the case that S is the class of all discrete memoryless sources, it was proved [51] that the minimax regret asymptotically equals $(d/2) \log(n/2\pi) + \log C_J(K) + o(1)$, where d equals the size of alphabet minus 1 and $C_J(K)$ is the integral of the square root of the determinant of the Fisher information matrix over K. An important point in the above is that K is taken there to be Θ itself, i.e. we do not have to have any restriction for the sequence x^n. To obtain this asymptotically minimax regret, they use sequences of Bayes mixtures with certain prior distributions that weakly converge to the Jeffreys prior. The reason why one needs such variants of the Jeffreys prior is as follows: If we use the Jeffreys prior, the regret is asymptotically higher than the minimax value, for x^n such that $\hat{\theta}$ is near the boundary of Θ. Priors which have higher density near the boundaries than the Jeffreys prior give more prior attention to these boundary regions and thereby pull the regret down to not more than the asymptotically minimax level.

In this paper, we provide such regret results for more general parametric families of densities. For exponential families of arbitrary dimension, the Jeffreys mixture is shown to be asymptotically minimax, if K is a compact subset included in the interior of Θ. A boundary modification is shown to produce a variant of Jeffreys mixture asymptotically minimax for $K = \Theta$ in the one-dimensional exponential family case. For general smooth families that are not of exponential type, we find that any Bayes mixture that uses a prior restricted to the family is not asymptotically minimax, but a slight

modification (enlargement) of the family allows for priors for which the regret of Bayes mixtures does achieve the asymptotically minimax value.

The related problem of minimax expected regret (redundancy), which is defined as

$$\min_q \max_{\theta \in K} E_\theta \left(\log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\theta)} \right)$$

was studied by Clarke and Barron [14]. They considered fairly general classes of i.i.d. processes and showed that the minimax expected regret asymptotically equals $(d/2) \log(n/2\pi e) + \log C_J(K) + o(1)$, when K is a compact subset of the interior of Θ. Moreover the expected regret is related to the value of

$$\min_q \max_{\theta \in K} E_\theta \left(\log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\hat\theta)} \right)$$

using the target code length with $1/p(x^n|\hat\theta)$ instead of $1/p(x^n|\theta)$. It has a corresponding minimax value $(d/2) \log(n/2\pi) + \log C_J(K) + o(1)$. Preceding work [51], [50] evaluated the minimax expected regret for the class of discrete memoryless sources and showed that sequences of slightly modified Jeffreys mixtures achieve the minimax value asymptotically for the whole probability simplex Θ. As we shall see, the answer for the minimax regret and the minimax expected regret are similar.

To obtain the minimax regret results, we employ the Laplace integration method, which was used by [13], [14] to evaluate the expected regret of the Bayes procedures. Especially in [14], they succeeded to uniformly evaluate the expected regret by the Laplace integration for compact subsets K of Θ°, the interior of Θ. To handle all of Θ, careful modification of the Laplace method is required to handle behaviour near the boundary.

Since the Jeffreys mixture achieves asymptotically the minimax expected regret, one might expect that it also achieves asymptotically the minimax pointwise regret. However, it does not in general. Namely, we can see that for processes which are not exponential type, the worst case regret of the Jeffreys mixture is asymptotically higher than the minimax value, even if strings are restricted such that the maximum likelihood estimate (MLE) is in the interior of the parameter space.

The reason follows from examination of the Laplace approximation to the Bayes mixture density. This leads to an approximation to the regret of the Jeffreys mixture which converges to the minimax value, if and only if the difference between the determinant of the empirical Fisher information matrix and that of the Fisher information matrix at the MLE converges to 0. For exponential families, the empirical and expected Fisher informations are the same, hence the Jeffreys mixture is asymptotically minimax. The situation is different for families which are not exponential type, for then there exist sequence x^n for which the determinants of these matrices are asymptotically different.

Even though the Jeffreys mixture is not asymptotically minimax, we can obtain an asymptotic minimax regret, by a modified Jeffreys mixture obtained by adding a small contribution from a mixture of an enlarged set of densities, whose dimension is higher than the original set. The added

components deal with the strings for which the empirical Fisher information differs from the Fisher information. When the original set is a curved exponential family embedded in an exponential family, we have the option to use that family as the enlarged model. Furthermore, this method can be applied to non-i.i.d. families under certain assumptions.

Our result about minimax regret provides an alternative way to evaluate, more generally, the stochastic complexity in Rissanen [37], where he used the normalized maximum likelihood. Our results show that the codelength of the minimax strategy retains asymptotically the mixture codelength interpretation of earlier incarnations of his criterion. The work [6] shows that in some cases the exact minimax strategy has a signed mixture interpretation without resorting to asymptotics.

The codelength of a code based on a mixture $q(x^n) = \int_K p(x^n|\theta)w(\theta)d\theta$ has Laplace approximation (for $\hat\theta$ away from the boundary of K) given by

$$\log \frac{1}{q(x^n)} \sim \log \frac{1}{p(x^n|\hat\theta)} + \frac{d}{2} \log \frac{n}{2\pi} + \log \frac{|\hat{J}(\hat\theta, x^n)|^{1/2}}{w(\hat\theta)},$$

where $\hat{J}(\hat\theta, x^n)$ is the empirical Fisher information, defined as $(1/n)$ times the second derivative matrix of the minus log-likelihood, evaluated at the MLE $\hat\theta$. With Jeffreys prior in which $w(\theta)$ is proportional to $|J(\theta)|^{1/2}$, the Laplace approximation for $\log(1/q(x^n))$ is

$$\log \frac{1}{p(x^n|\hat\theta)} + \frac{d}{2} \log \frac{n}{2\pi} + \log C_J(K) + \frac{1}{2} \log \frac{|\hat{J}(\hat\theta)|}{|J(\hat\theta)|},$$

where

$$C_J(K) = \int_K |J(\theta)|^{1/2} d\theta.$$

Accordingly, the regret of the code based on Jeffreys prior takes the approximation form

$$\frac{d}{2} \log \frac{n}{2\pi} + \log C_J(K) + \frac{1}{2} \log \frac{|\hat{J}(\hat\theta)|}{|J(\hat\theta)|}.$$

Our modification of $q(x^n)$ by enlargement of the family with an associated modification to the prior shows that

$$\frac{d}{2} \log \frac{n}{2\pi} + \log C_J(K) \tag{1}$$

remains the asymptotically minimax regret even though in the non-exponential family cases there are sequences x^n for which $\log(|\hat{J}(\hat\theta, x^n)|/|J(\hat\theta)|)$ does not converge to 0.

This presents a challenge for the formulation and analysis of our asymptotic minimax procedures with the maximum of the regret taken over all x^n. In contrast the minimax expected regret is easier to achieve because the expectation washes out the effect of $\hat{J}$ different from J. In either formulation, with $\log(1/p(x^n|\hat\theta))$ as the target, the minimax value takes the asymptotic form (1).

In Section II, we formally introduce the notion of minimax and maximin regret. Section III gives keys to the bounds. In Section IV, we give the lower bound on the maximin regret for general smooth families in the i.i.d. setting (Theorems 1 and 2) and for families of non-i.i.d. densities (Theorems 3 and 4). These hold for any subset K of Θ with finite Jeffreys integral.

Likewise, in Section V, we give corresponding upper bound on the minimax regret for general smooth families with a compact subset K in the interior of Θ (Theorem 5), which incorporate the discussed innovations. In Section VI, we provide various types of families as concrete examples, including certain curved exponential families and certain mixture families that are not representable as exponential families.

II. PRELIMINARIES

Let $(\mathcal{X}, \mathcal{B}, \nu)$ be a measurable space with a reference measure ν, assumed to be sigma finite (such as counting measure or Lebesgue measure). Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ denote a parametric family of probability densities over $\mathcal{X}$ with respect to ν. Assume that $\Theta \subseteq \Re^d$. We let $p(x^n|\theta)$ denote $\prod_{i=1}^n p(x_i|\theta)$. Also, we let $\nu(dx^n)$ denote $\prod_{i=1}^n \nu(dx_i)$. Here, we are treating models for independently identically distributed (i.i.d.) random variables. (Section IV treats certain non i.i.d. cases.) We let P_θ denote the distribution function with density $p(\cdot|\theta)$ and E_θ denote expectation with respect to P_θ.

Define the Fisher information matrix by

$$J_{ij}(\theta) = E_\theta\Big(-\frac{\partial^2 \log p(x|\theta)}{\partial\theta^i\partial\theta^j}\Big).$$

(We let $\log$ denote the natural logarithm.) We assume that $J(\theta)$ exists and is strictly positive definite in the interior of Θ. Let K denote a subset of Θ. We let $C_J(K) \overset{\text{def}}{=} \int_K |J(\theta)|^{1/2} d\theta$, which we call the Jeffreys integral on K. Our interest is in cases in which $C_J(K)$ is finite. The Jeffreys prior over K is a prior distribution with density function

$$w_K(\theta) = \frac{|J(\theta)|^{1/2}}{C_J(K)} 1_K(\theta).$$

Define the Jeffreys mixture for K as

$$m_K(x^n) = \int_K p(x^n|\theta) w_K(\theta) d\theta.$$

We also introduce the empirical Fisher information as a function of x^n:

$$\hat{J}_{ij}(\theta) = \hat{J}_{ij,n}(\theta) = \hat{J}_{ij}(\theta, x^n) = \frac{-1}{n}\frac{\partial^2 \log p(x^n|\theta)}{\partial\theta^i\partial\theta^j}.$$

Note that $J_{ij}(\theta) = E_\theta[\hat{J}_{i,j}(\theta)]$. Moreover, in the present i.i.d. setting, $\hat{J}_{ij,n}(\theta)$ is near $J_{ij}(\theta)$ with high probability for large n, by the law of large numbers, when $J_{ij}(\theta)$ is finite, if we were to have the X_i distributed according to P_θ. We work with the arbitrary sequence perspective, so the empirical $\hat{J}_{ij,n}(\theta)$ need not be close to $J_{ij}(\theta)$. Nevertheless we will find a role for the expected value formulation of Fisher information in characterization of the minimax regret among arbitrary sequences.

Let $\hat{\theta}(x^n)$ be the maximum likelihood estimate (MLE), which is

$$\arg\max_{\theta\in\Theta} p(x^n|\theta).$$

More strictly, we define $\hat{\theta}$ to be an element of the set $\{\theta : p(x^n|\theta) = \max_\theta p(x^n|\theta)\}$. For the case where this set has more than one element, we presume there is an arbitrarily specified rule to choose one element.

For example, in the case of Bernoulli sources, we have $\hat{\theta} = \sum_{i=1}^n x_i/n$, where θ is the parameter denoting the probability that '1' occurs. The function $\hat{\theta}$ mapping each x^n in $\mathcal{X}^n$ to the value $\hat{\theta}(x^n)$ is the maximum likelihood estimator. In the above case of Bernoulli sources, $\hat{\theta}(x^n)$ is defined on the whole $\mathcal{X}^n$. However, in certain cases, there may exist strings for which $\max p(x^n|\theta)$ does not exists. In such cases, we restrict the domain of $\hat{\theta}$ to the set of x^n's such that the MLE does exist.

We introduce the notion of minimax regret and maximin regret. Let $\mathcal{K}_n$ denote a fixed subset of $\mathcal{X}^n$. Let $\mathcal{P}_n$ denote the set which consists of all probability densities over $\mathcal{X}^n$ with respect to the fixed sigma-finite reference measure.

For sample size n and set $\mathcal{K}_n$, define the maximum regret of a q in $\mathcal{P}_n$ (denoted by $\bar{r}_n(q, \mathcal{K}_n)$) as

$$\bar{r}_n(q, \mathcal{K}_n) = \sup_{x^n\in\mathcal{K}_n}\sup_{\theta\in\Theta} \log\frac{p(x^n|\theta)}{q(x^n)} = \sup_{x^n\in\mathcal{K}_n} \log\frac{p(x^n|\hat{\theta})}{q(x^n)}$$

and the *minimax* regret as

$$\bar{r}_n(\mathcal{K}_n) = \inf_{q\in\mathcal{P}_n} \bar{r}_n(q, \mathcal{K}_n).$$

For each $q \in \mathcal{P}_n$, the minimum average regret for sample size n denoted

$$\underline{r}_n(q) = \inf_{\tilde{q}\in\mathcal{P}_n} \int q(x^n)\Big(\log\frac{p(x^n|\hat{\theta})}{\tilde{q}(x^n)}\Big)\nu(dx^n)$$

is achieved by $\tilde{q} = q$. We define the *maximin* regret for the set $\mathcal{K}_n$ as

$$\underline{r}_n(\mathcal{K}_n) = \sup_{q\in\mathcal{P}_{\mathcal{K}_n}} \underline{r}_n(q),$$

where $\mathcal{P}_{\mathcal{K}_n}$ is the set of all probability densities supported on $\mathcal{K}_n$.

By the above definitions, $\bar{r}_n(\mathcal{K}_n) \geq \underline{r}_n(\mathcal{K}_n)$ holds. (In fact, $\bar{r}_n(\mathcal{K}_n) = \underline{r}_n(\mathcal{K}_n)$ holds [39], [51].) In this paper, we usually consider the minimax regret problem for the following form of the set $\mathcal{K}_n$: given $K \subseteq \Theta$,

$$\mathcal{K}_n = \mathcal{X}^n(K) = \{x^n : \hat{\theta} \in K\}.$$

Next, we remind the reader of the notions of the minimax and maximin value of expected regret (redundancy). Let a sample size n and a parameter set K be given. For each $q \in \mathcal{P}_n$, we define the maximum expected regret (denoted by $\bar{R}_n(q, K)$) as

$$\bar{R}_n(q, K) = \sup_{\theta\in K} E_\theta\Big(\log\frac{p(x^n|\theta)}{q(x^n)}\Big)$$

and the minimax expected regret as

$$\bar{R}_n(K) = \inf_{q\in\mathcal{P}_n} \bar{R}_n(q, K),$$

where E_θ denotes the expectation with respect to $p(x^n|\theta)$. Let $\Omega(K)$ denote the set which consists of all prior probability measures over K. For each prior $w \in \Omega(\Theta)$, the minimum Bayes expected regret for sample size n denoted

$$\underline{R}_n(w) = \inf_{q\in\mathcal{P}_n} \int E_\theta\Big(\log\frac{p(x^n|\theta)}{q(x^n)}\Big)w(d\theta)$$

is achieved by $q(x^n) = m(x^n) = \int p(x^n|\theta)w(d\theta)$. We define the maximin expected regret for the parameter set K as

$$\underline{R}_n(K) = \sup_{w \in \Omega(K)} \underline{R}_n(w).$$

We also have $\bar{R}_n(K) \geq \underline{R}_n(K)$. (In fact, $\bar{R}_n(K) = \underline{R}_n(K)$ holds [17], [25].)

Finally, we note some terminology. We let 1_A denote the characteristic function of a set A. We write $o(1)$ for an expression upper bounded by a positive quantity tending to zero. In the proofs, we denote certain positive constants by C_i ($i = 1, 2, ...$), where i is 'local' in each proof.

III. THE KEY TO HANDLING WORST CASE REGRET

The notion that the worst case regret is the same asymptotically as an average case regret should come as a surprise. Indeed, if one inspects the worst case regret of an average case optimal procedure, e.g. with the Jeffreys mixture, there will be strings for which the worst case performance of the procedure is quantitatively different in its asymptotics than the average case performance.

To succeed in bringing the worst case regret value down to the average case optimum requires a sequence of somewhat new specialized procedures. These specializations are modifications of Jeffreys mixtures that address two difficulties. One is that Laplace approximation is no longer valid for strings for which the MLE $\hat{\theta}$ is at or very close to the boundary of the parameter space. That is a familiar issue, addressed by modification of the prior to give small additional prior weights to parameter values near the boundary. Such boundary modification was needed in the special case of [50], [51]. Here we provide an improved near-boundary fix by using a multiplicative correction factor for the prior. This improved correction factor comes from the probability of K assigned by a normal distributions that arises in the Laplace approximation. This just-right probability correction factor from Laplace approximation is near 1 away from the boundary (by an amount of larger order than $1/\sqrt{n}$), and typically approaches $1/2$ as the parameter approaches the boundary (at a rate faster than order $1/\sqrt{n}$).

The more fundamental difficulty is the discrepancy between the worst case and average values of the regret of mixture procedures that can exist even when $\hat{\theta}$ is in the interior, which, as exhibited above, is due to individual strings having $(1/2)\log(|\hat{J}(\hat{\theta})|/|J(\hat{\theta})|)$ not near zero when not in an exponential family. We present and analyze ideas to overcome that difficulty which were initiated by the authors as discussed in [40], [7], [5].

For general smooth families we form a direct enlargement by a exponential tilting using linear combinations of the entries of the differences $\hat{J}(\theta) - J(\theta)$. It is formed as

$$p(x^n|\theta, \beta) = p(x^n|\theta)e^{n\beta \cdot (\hat{J}(\theta) - J(\theta)) - \psi_n(\theta, \beta)} \quad (2)$$

Where $\beta \cdot M$ for matrices β and M denotes the Frobenius inner product (the sum of products across all d^2 entries).

The idea for this enlargement in addressing minimax regret originate in preliminary form in [40], [7] as informally discussed in [5], [43]. Here $\psi_n(\theta, \beta)$ is the log of the required

normalization factor, so that $p(x^n|\theta, \beta)$ sums (integrates with respect to ν^n) to the value 1 for every $\theta \in K$ and every β in a neighborhood around 0. The prior assigns most of its weight to a Jeffreys prior on θ (with β set to 0) and small weight on a smooth prior on (θ, β) with β in a neighborhood of 0. Now demonstration of the success of this modification is based on demonstration of increased likelihood beyond that which is available at $\beta = 0$ when $\hat{J}(\hat{\theta}) - J(\hat{\theta})$ is not equal to 0. Indeed, as we show in Section V, the contribution to the mixture from β with $\beta \cdot (\hat{J}(\hat{\theta}) - J(\hat{\theta}))$ positive is sufficient to increase the value of $m(x^n)$ to again overcome the discrepancy in $(1/2)\log(|\hat{J}(\hat{\theta})|/|J(\hat{\theta})|)$ from Laplace approximation.

We have a somewhat different modification that can be used in the case of curved exponential families embedded in a full exponential family as developed in examples in Section VI. For both of the modifications we study, there is, in the analysis, the consideration of values of β in a neighborhood of a small multiple of $\hat{J}(\hat{\theta}) - J(\hat{\theta})$ which are sufficient to accomplish our objectives. We see the similarity of effect. In both cases there is opportunity to create a likelihood increase from a suitable linear combination of these statistics.

These ideas emanate from an underlying principle. Of importance in smooth statistical families is the parametric enlargement

$$p(x^n|\theta, \beta) = p(x^n|\theta)e^{\beta \cdot \nabla \log p(x^n|\theta) - \psi_n(\theta, \beta)}, \quad (3)$$

where $\nabla \log p(x^n|\theta)$ is the score function at θ and $\psi_n(\theta, \beta)$ is the log normalizing constant near $(1/2)\beta^t J(\theta)\beta$ for small β. Traditionally such a family arises in local asymptotic expansion of likelihood ratios, evaluated at a perturbation $\theta + \beta$ of a given θ, as used in demonstration of *local asymptotic normality* [30], [33], [36]. In Amari's information geometry it is a local exponential tangent to the family at θ. Again, in this setting, a prior can be put jointly on θ and β, where most of the prior weight is concentrated at $\beta = 0$ with Jeffreys prior on θ.

The use of the first derivative of $\log p(x^n|\theta)$ in the exponent would seem to produce a different sort of enlargement than occurs with equation (2) which uses the second derivative. Nevertheless, improvement in $m(x^n)$ still arises by suitable characterization of the increased likelihood available with the enlarged family. For each small β the maximum likelihood value $\hat{\theta}_\beta$ achieving $\max_\theta p(x^n|\theta, \beta)$ in (3) satisfies the approximate relationship

$$p(x^n|\hat{\theta}_\beta, \beta) = p(x^n|\hat{\theta}_0)e^{(n/2)\beta^T(\hat{J}(\hat{\theta}_0) - J(\hat{\theta}_0))\beta} \quad (4)$$

to within term of order $n|\beta|^3$ in the exponent. This is the desired effect, where $\hat{\theta}_0 = \hat{\theta}$ here is the MLE in the original family and $\hat{\theta}_\beta$ is the MLE in the tilted family (3). The likelihood is larger at some non-zero β than at $\beta = 0$ provided $\beta^T(\hat{J}(\hat{\theta}) - J(\hat{\theta}))\beta$ can be strongly positive, or equivalently provided $\hat{J}(\hat{\theta})(J(\hat{\theta}))^{-1}$ has some eigenvalue greater than 1. Consequently, when $(1/2)\log(|\hat{J}(\hat{\theta})|/|J(\hat{\theta})|)$ is positive (which is the only case of concern) optimization in this tangent family is sufficient to realize similar likelihood gain to optimization in the family tilted by second derivatives of log-likelihood.

This provides an enlargement of parametric families by exponential tilting using score functions suitable to endow the family with the property we need, comparable to the curved exponential family case. Analogous conditions are in Amari and Nagaoka [3], who use tilting by first and second derivatives. A distinction thought is that he is considering local families with a fixed true θ whereas we are considering the joint family parametrized by θ and β.

For general families in the present paper we use the enlargement of the form (2) in our analysis, now further motivated by the connections to locally asymptotically normal familiers we have discussed here.

IV. LOWER BOUNDS

Our targets are fairly general smooth parametric families under some conditions and some classes of non-i.i.d. stochastic processes. First we describe our results about lower bounds.

A. Lower Bounds for General i.i.d. Families

Let $S \stackrel{\text{def}}{=} \{p(\cdot|\theta) : \theta \in \Theta\}$ be a family of probability densities with respect to the reference measure ν, with a d-dimensional parameter θ. Assume that $\Theta \subseteq \Re^d$ and let Θ° denote its interior. We use K to denote subsets of Θ, with particular interest in sets for which $C_J(K)$ is finite and for which the boundary measure of K is zero. We handle both the case of bounded sets K (Lemma 1) and more general cases of possibly unbounded sets K (Theorem 1).

For each $\delta > 0$, define Mahalanobis neighborhoods of $\theta \in \Theta$ as

$$B_\delta(\theta) = \{\theta' \in \Theta : (\theta' - \theta)^t J_\theta(\theta' - \theta) \leq \delta^2\}.$$

We employ the assumptions described below.

Assumption 1: The density $p(x|\theta)$ is twice continuously differentiable for $\theta \in \Theta^\circ$ for every x. Moreover, for every $\theta \in \Theta^\circ$ there is a $r = r(\theta)$ such that, for every i, j,

$$E_\theta \left[\sup_{\theta' \in B_r(\theta)} |\hat{J}_{ij,1}(\theta')| \right] \quad (5)$$

is finite.

Assumption 2: The Fisher information $J(\theta)$ is continuous and positive definite in Θ°.

Consequences of continuity and positive definiteness are that the following two quantities tends to 0 as $\delta \to 0$ for any θ in Θ°

$$\sup_{\theta' \in B_\delta(\theta)} \frac{|J(\theta')|^{1/2}}{|J(\theta)|^{1/2}} - 1 \quad (6)$$

and

$$\sup_{\theta' \in B_\delta(\theta)} \max_{z \neq 0} \frac{z^t J(\theta) z}{z^t J(\theta') z} - 1. \quad (7)$$

Since $J(\theta)$ is symmetric and positive definite it has a real eigendecomposition with positive eigenvalues. Taking the positive square roots of the eigenvalues provides representation of the principle square root matrix $J(\theta)^{1/2}$ and taking their reciprocal provides representation of $J(\theta)^{-1}$. With $J(\theta)$ continuous

these inverses and square roots remain symmetric and continuous in θ ([28],p.411). Armed with these we can standardize the empirical Fisher information as $J(\theta)^{-1/2}\hat{J}(\theta')J(\theta)^{-1/2}$, which has expectation equal to the identity matrix I at $\theta' = \theta$.

From the continuity from Assumption 2 and the domination (finiteness of expected supremum) from Assumption 1, it follows by the monotone convergence theorem that for every i, j the following quantity also tends to 0 as $\delta \to 0$, for each θ in Θ°,

$$E_\theta \sup_{\theta' \in B_\delta(\theta)} \pm((J(\theta)^{-1/2}\hat{J}_1(\theta', x)J(\theta)^{-1/2})_{ij} - I_{ij}), \quad (8)$$

where the $\pm$ indicates that the statement is true with each choice of sign.

Assumption 3: The maximum likelihood estimator $\hat{\theta}$ is a consistent estimator of θ, for each θ in Θ°, so that

$$P_\theta\{||\hat{\theta}(x^n) - \theta||_{J(\theta)} > \delta\} = o(1),$$

for each $\delta > 0$.

Remark: We give a demonstration of consistency under suitable conditions in Lemma 18 in the Appendix. Assumptions 1, 2, and 3 are related to the conditions pioneered by Cramér (see p. 501 of [15]) for local asymptotic properties of maximum likelihood estimators building on earlier assumptions by Wald [47] for consistency of the maximum likelihood estimator.

Assumption 4: The set $K \subseteq \Theta$ has positive and finite

$$C_J(K) = \int_K |J(\theta)|^{1/2} d\theta$$

and the measure of $K \setminus K^\circ$ is zero.

We remark that the zero measure condition is to make $C_J(K^\circ) = C_J(K)$. It is clearly satisfied if K is open or if K is a closed set with boundary measure zero.

From the theory of finite measures, convergent functions are uniformly convergent except in sets of arbitrarily small measure. Consequently, as elaborated in the appendix, if Assumptions 1,2,4 hold, then, for any $\epsilon > 0$, there is a $\delta > 0$ and a good set G of parameters, such that for all θ in G the quantities in (6) and (7) are less than ϵ, the expected suprema in (8) are less than $\epsilon/(2d)$ and the Jeffreys measure of the complement of G, which is $\int_{K \setminus G} |J(\theta)|^{1/2} d\theta$, is less than $\epsilon C_J(K)$.

Moreover, using a law of large numbers, it is shown in the appendix, as a consequence of the expected suprema being less than $\epsilon/(2d)$, that for θ in G,

$$P_\theta \left(\inf_{z \neq 0} \inf_{\theta' \in B_\delta(\theta)} \frac{z^t \hat{J}(\theta', x^n) z}{z^t J(\theta) z} < 1 - \epsilon \right) = o(1) \quad (9)$$

as n goes to infinity.

To see the connection with the standardized empirical information, note that the infimum here among z is the same as

$$\inf_{\zeta : ||\zeta|| = 1} \inf_{\theta' \in B_\delta(\theta)} \zeta^t J(\theta)^{-1/2} \hat{J}(\theta') J(\theta)^{-1/2} \zeta$$

as can be seen by setting ζ to correspond to $J(\theta)^{1/2}z$ divided by its norm.

Two lower bounds related to maximin regret will be given. The first, using Assumption 1, is a general lower bound

showing the asymptotically validity of the $(d/2)\log(n/2\pi) + \log C_J(K)$ expression, but it somewhat less constructive. It uses the fact mentioned above that convergent functions are nearly uniformly convergent, so there exists a good set G with the indicated properties, while having a small Jeffreys measure for its complement. Also, as we said, it uses a law of large numbers for a sample average of such suprema, but, assuming only the finite expectation, there is, in general, no assurance of rates of approach in such a law of large numbers, and this translates to a lack of explicit rate of approach in the general lower bound. Moreover, for the purpose of obtaining the general lower bound on the minimax value it suffices to appeal to an approximation argument with Jeffreys mixtures that live on G.

The other lower bound we give is specific to the use of the Jeffreys prior on K without modification. In that case a stronger condition, refining Assumption 1, will be used with finite expected square (to permit application of Chebyshev's inequality for control of rate of convergence in the law of large numbers), as well as a continuity assumption of the expected suprema, to get uniform convergence (uniform closeness to zero of the expected suprema) within any compact subset of K. Armed with this more refined Assumption $1'$ (to be specified later below) we have more explicit control of rates of approach to the maximin value.

Now let $A \subseteq K$ be a parameter set to which Lemma 1 will appeal. This set A can be the good set G as discussed above, whose existence is a consequence of Assumptions 1, 2, and 4. Or if K is compact and additional assumptions are satisfied we may have $A = K$.

Define

$$\eta_{1,\delta}(A) = \sup_{\theta \in A} \sup_{\theta' \in B_\delta(\theta)} \max_{z \neq 0} \frac{z^t J(\theta)z}{z^t J(\theta')z} - 1, \quad (10)$$

$$\eta_{2,\delta}(A) = \sup_{\theta \in A} \sup_{\theta' \in B_\delta(\theta)} \frac{|J(\theta')|^{1/2}}{|J(\theta)|^{1/2}} - 1, \quad (11)$$

and, similarly, let $\eta_{0,\delta}(A)$ be $2d$ times the following supremum

$$\sup_{\theta \in A} E_\theta \sup_{\theta' \in B_\delta(\theta)} \pm (J(\theta)^{-1/2} \hat{J}_1(\theta', x) J(\theta)^{-1/2})_{ij} - I_{ij}).$$
$$(12)$$

As previously mentioned the $2d$ factor is so that when this $\eta_{0,\delta}(A) \leq \epsilon$ the conclusion of (9) holds.

We extract nice sets within $\Theta \times \mathcal{X}^n$, which we will use in obtaining lower bounds on the maximin value. Define

$$\mathcal{B} = \{(\theta, x^n) : \theta, \hat{\theta}(x^n) \in K, ||\theta - \hat{\theta}(x^n)||_{J(\theta)} \leq \delta/4\}$$

$$\mathcal{N} = \left\{(\theta, x^n) : \theta \in A, \inf_{\theta' \in B_\delta(\theta)} \inf_{z \neq 0} \frac{z^t \hat{J}(\theta', x^n)z}{z^t J(\theta)z} \geq 1 - \epsilon\right\}.$$
$$(13)$$

Here, for any subset $\mathcal{G}$ of $\Theta \times \mathcal{X}^n$, we let $\mathcal{G}_{x^n}$ and $\mathcal{G}_\theta$ denote the section of $\mathcal{G}$ given x^n and θ, respectively, that is,

$$\mathcal{G}_{x^n} = \{\theta : (\theta, x^n) \in \mathcal{G}\},$$
$$\mathcal{G}_\theta = \{x^n : (\theta, x^n) \in \mathcal{G}\}.$$

Now fix a pair $(\theta, x^n) \in \mathcal{N}$. We show that the section $\mathcal{B}_{x^n}$ is included in $B_\delta(\theta)$. Assume $\theta' \in \mathcal{B}_{x^n}$. Then,

$$||\theta' - \hat{\theta}||_{J(\theta')} \leq \delta/4$$

holds. Likewise, since $(\theta, x^n) \in \mathcal{B}$,

$$||\theta - \hat{\theta}||_{J(\theta)} \leq \delta/4.$$

Since $\theta, \theta' \in \mathcal{B}_{x^n}$, and $\theta \in A$ we have, by the definition of $\eta_{1,\delta} = \eta_{1,\delta}(A)$,

$$||\theta' - \hat{\theta}||_{J(\theta)} \leq (1 + \eta_{1,\delta})^{1/2}||\theta' - \hat{\theta}||_{J(\theta')} \leq \delta/2,$$

assuming $\eta_{1,\delta} \leq 3$. Whence

$$||\theta' - \theta||_{J(\theta)} < \delta,$$

which means $\theta' \in B_\delta(\theta)$ and $\mathcal{B}_{x^n} \subset B_\delta(\theta)$. By this, if $x^n \in \mathcal{N}_\theta$ and $(\theta, x^n) \in \mathcal{B}$, the following holds,

$$\min_{\theta' \in \mathcal{B}_{x^n}} \min_{z \neq 0} \frac{z^t \hat{J}(\theta', x^n)z}{z^t J(\theta)z} \geq 1 - \epsilon. \quad (14)$$

We can prove the following Lemma, which is the heart of our main results for the lower bounds.

Lemma 1: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional family of probability densities. We suppose that Assumptions 1-4 hold for S. Let $0 < \epsilon < 1$ be given and assume for a specified $A \subset K$ that there is a $\delta > 0$ such that $\eta_{0,\delta}(A) \leq \epsilon$. Assume also that $\eta_{1,\delta}(A) \leq 3$ and let $\eta_{2,\delta} = \eta_{2,\delta}(A)$. Define a probability density m_δ supported on the set

$$\mathcal{A}_\delta = \{x^n : (\mathcal{N} \cap \mathcal{B})_{x^n} \neq \emptyset\} \subset \mathcal{K},$$

by

$$\bar{m}_\delta(x^n) = \int_{(\mathcal{N} \cap \mathcal{B})_{x^n}} p(x^n|\theta)w_A(\theta)d\theta, \quad (15)$$

$$m_\delta(x^n) = \frac{\bar{m}_\delta(x^n)}{\bar{C}_n}, \quad (16)$$

where

$$\bar{C}_n = \int_{\mathcal{K}} \bar{m}_\delta(x^n)\nu(dx^n) \quad (17)$$

Then

$$E_{m_\delta} \log \frac{p(x^n|\hat{\theta})}{m_\delta(x^n)} \geq \log \frac{(1-\epsilon)^{d/2}C_J(A)n^{d/2}}{(1+\eta_{2,\delta})(2\pi)^{d/2}} - o(1)$$

holds as n goes to infinity.

Remark: The support of the density m_δ is the set of x^n such that the section $(\mathcal{N} \cap \mathcal{B})_{x^n}$ is not empty.

Proof: We have

$$E_{m_\delta} \log \frac{p(x^n|\hat{\theta})}{m_\delta(x^n)} = E_{m_\delta} \log \frac{p(x^n|\hat{\theta})}{\bar{m}_\delta(x^n)} + \log \bar{C}_n. \quad (18)$$

Note that the log likelihood ratio in the first term of the right hand side is positive, while $\log(p(x^n|\hat{\theta})/m_\delta(x^n))$ can be negative.

For the first term. we have

$$E_{m_\delta} \log \frac{p(x^n|\hat\theta)}{\bar m_\delta(x^n)}$$

$$= \int_{\mathcal{K}} \frac{\bar m_\delta(x^n)}{\bar C_n} \log \frac{p(x^n|\hat\theta)}{\bar m_\delta(x^n)} \nu(dx^n)$$

$$= \frac{1}{\bar C_n} \int_{\mathcal{K}} \int_{(\mathcal{N} \cap \mathcal{B})_{x^n}} p(x^n|\theta) w_A(\theta) d\theta \log \frac{p(x^n|\hat\theta)}{\bar m_\delta(x^n)} \nu(dx^n)$$

$$= \frac{1}{\bar C_n} \int_{K} w_A(\theta) \left[\int_{(\mathcal{N} \cap \mathcal{B})_\theta} p(x^n|\theta) \log \frac{p(x^n|\hat\theta)}{\bar m_\delta(x^n)} \nu(dx^n) \right] d\theta.$$

Here, using the sections, reversed order integration is possible by the Fubini-Tonelli theorem because of positivity of the integrand.

Let $i(\theta)$ denote the quantity in the brackets [ ]. We will show a lower bound on $i(\theta)$.

Note that $x^n \in (\mathcal{N} \cap \mathcal{B})_\theta$ is equivalent to $(\theta, x^n) \in \mathcal{N} \cap \mathcal{B}$ and to $\theta \in (\mathcal{N} \cap \mathcal{B})_{x^n}$. By Taylor expansion of the log likelihood ratio around $\hat\theta$ under the condition that $x^n \in (\mathcal{N} \cap \mathcal{B})_\theta$, we have

$$\frac{\bar m_\delta(x^n)}{p(x^n|\hat\theta)}$$

$$= \int_{(\mathcal{N} \cap \mathcal{B})_{x^n}} \frac{p(x^n|\theta') w_A(\theta')}{p(x^n|\hat\theta)} d\theta'$$

$$= \int_{(\mathcal{N} \cap \mathcal{B})_{x^n}} e^{-n(\theta'-\hat\theta)^t \hat J(\theta'',x^n)(\theta'-\hat\theta)/2} w_A(\theta') d\theta',$$

where θ'' is a point between $\hat\theta$ and θ'. Since $x^n \in (\mathcal{N} \cap \mathcal{B})_\theta$, we have $\theta \in (\mathcal{N} \cap \mathcal{B})_{x^n}$. Now

$$\max_{\theta' \in B_\delta(\theta)} \frac{w_A(\theta')}{w_A(\theta)} \le (1 + \eta_{2,\delta})$$

and

$$\min_{\theta' \in B_\delta(\theta)} \min_{z \ne 0} \frac{z^t \hat J(\theta', x^n) z}{z^t J(\theta) z} \ge 1 - \epsilon$$

hold for $\theta \in A$, where $\eta_{2,\delta} = \eta_{2,\delta}(A)$. Hence we have for all $x^n \in (\mathcal{N} \cap \mathcal{B})_\theta$,

$$\frac{\bar m_\delta(x^n)}{p(x^n|\hat\theta)}$$

$$\le (1 + \eta_{J,\delta}) w_A(\theta) \int_{\mathcal{N}_{x^n}} e^{-n(1-\epsilon)(\theta'-\hat\theta)^t J(\theta)(\theta'-\hat\theta)/2} d\theta'$$

$$\le (1 + \eta_{2,\delta}) w_A(\theta) \int e^{-n(1-\epsilon)(\theta'-\hat\theta)^t J(\theta)(\theta'-\hat\theta)/2} d\theta'$$

$$= \frac{(1 + \eta_{2,\delta}) w_A(\theta)}{(1 - \epsilon)^{d/2}} \frac{(2\pi)^{d/2}}{n^{d/2} |J(\theta)|^{1/2}}$$

$$= \frac{(1 + \eta_{2,\delta})}{(1 - \epsilon)^{d/2}} \frac{(2\pi)^{d/2}}{n^{d/2} C_J(A)}.$$

Hence, for all $\theta \in A$, we have

$$i(\theta) \ge \log \frac{(1-\epsilon)^{d/2} C_J(A) n^{d/2}}{(1+\eta_{2,\delta})(2\pi)^{d/2}} \int_{(\mathcal{N} \cap \mathcal{B})_\theta} p(x^n|\theta)\nu(dx^n),$$

which implies

$$\int_K w_A(\theta) i(\theta) d\theta$$

$$\ge \left(\log \frac{(1-\epsilon)^{d/2} C_J(A) n^{d/2}}{(1+\eta_{2,\delta})(2\pi)^{d/2}} \right) \int_{\mathcal{K}} \bar m(x^n) \nu(dx^n)$$

$$= \left(\log \frac{(1-\epsilon)^{d/2} C_J(A) n^{d/2}}{(1+\eta_{2,\delta})(2\pi)^{d/2}} \right) \bar C_n.$$

Dividing both sides by $\bar C_n$, we have

$$E_{m_\delta} \log \frac{p(x^n|\hat\theta)}{\bar m_\delta(x^n)} \ge \log \frac{(1-\epsilon)^{d/2} C_J(A) n^{d/2}}{(1+\eta_{2,\delta})(2\pi)^{d/2}}. \tag{19}$$

Now we will proceed to a lower bound on $\bar C_n$. We have

$$\bar C_n = \int_{\mathcal{K}} \int_{(\mathcal{N} \cap \mathcal{B})_{x^n}} p(x^n|\theta) w_A(\theta) d\theta \nu(dx^n)$$

$$= \int_K \int_{(\mathcal{N} \cap \mathcal{B})_\theta} p(x^n|\theta) \nu(dx^n) w_A(\theta) d\theta$$

$$= \int_K P_\theta((\mathcal{N} \cap \mathcal{B})_\theta) w_A(\theta) d\theta$$

$$\ge \int_K (P_\theta(\mathcal{N}_\theta) + P_\theta(\mathcal{B}_\theta) - 1) w_A(\theta) d\theta \tag{20}$$

$$= \int_{K^\circ} (P_\theta(\mathcal{N}_\theta) + P_\theta(\mathcal{B}_\theta) - 1) w_A(\theta) d\theta, \tag{21}$$

where the last equality holds because the boundary of K is assumed to have zero measure. Here, because of (9),

$$P_\theta(\mathcal{N}_\theta) = 1 - o(1)$$

holds as n goes to infinity, for each $\theta \in A$. Now

$$\mathcal{B}_\theta = \{x^n : \hat\theta \in B_{\delta/4}(\theta) \cap K\}.$$

For θ on the boundary, the probability of this event generally converges to number less than one. But for θ in the interior, the consistency implies that $\hat\theta$ is in K with probability tending to 1. Hence for each $\theta \in K^\circ$, we have

$$P_\theta(\mathcal{B}_\theta) = 1 - o(1)$$

as n goes to infinity, because of Assumption 3. Obviously, $\bar C_n \le 1$. Together with (20) and the pointwise convergence of $P_\theta(\mathcal{N}_\theta)$ and $P_\theta(\mathcal{B}_\theta)$ to 1, by Fatou's Lemma (or the dominated convergence theorem) we have that $\liminf_n \bar C_n \ge 1$. Thus

$$\bar C_n = 1 - o(1) \tag{22}$$

as n goes to infinity. Hence, with (18) and (19), we have

$$E_{m_\delta} \log \frac{p(x^n|\hat\theta)}{m_\delta(x^n)} \ge \log \frac{(1-\epsilon)^{d/2} C_J(A) n^{d/2}}{(1+\eta_{2,\delta})(2\pi)^{d/2}} - o(1),$$

as n goes to infinity. *The proof is completed.*

Now we state the theorem about the lower bound on the maximin regret for general i.i.d. families.

Theorem 1: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional family of probability densities. Suppose that Assumptions 1-3 hold for S. Let K be a subset of Θ (possibly $K = \Theta$) which satisfies Assumption 4. Then the following holds

$$\liminf_{n \to \infty} \left(\underline r_n(\mathcal{K}) - \frac{d}{2} \log \frac{n}{2\pi} \right) \ge \log C_J(K). \tag{23}$$

Remark: This theorem allows for Θ and K to be unbounded as long as $C_J(K)$ is finite.

Proof of Theorem 1: We are given $K \subset \Theta$ satisfying the finite $C_J(K)$ assumption. Let $\epsilon > 0$ be arbitrarily small. We have the existence of the set $A = G$ for which $\int_{K-G} w_K(\theta)d\theta \leq \epsilon$ and within which the properties of convergence to zero of expressions (6), (7) and (8) hold uniformly within G. This is an appeal to Egorov's theorem for finite measures that pointwise convergence implies uniform convergence except in a set of negligible measure. Accordingly, we can arrange positive δ sufficiently small that $\eta_{1,\delta}(A) \leq \epsilon \leq 3$, $\eta_{2,\delta}(A) \leq \epsilon$, and $\eta_{0,\delta} \leq \epsilon$. As detailed in the appendix, the last inequality implies the in-probability behavior (9). By Lemma 1 we have a procedure providing a lower bound on the maximin regret

$$\underline{r}_n(K) \geq \log \frac{(1-\epsilon)^{d/2}C_J(A)n^{d/2}}{(1+\eta_{2,\delta})(2\pi)^{d/2}} - o(1)$$

Accordingly,

$$\liminf \left[\underline{r}_n(K) - (d/2)\log \frac{n}{2\pi} \right]$$

is at least

$$\log C_J(K) + (1+d/2)\log(1-\epsilon) - \log(1+\epsilon).$$

Now since ϵ is arbitrarily small it implies that

$$\liminf \left[\underline{r}_n(K) - (d/2)\log \frac{n}{2\pi} \right] \geq \log C_J(K).$$

This completes the proof of Theorem 1.

Under the following stronger assumptions than those for Theorem 1, we can obtain a lower bound on the average regret for mixtures using Jeffreys prior on all of K, when K is compact.

Assumption 1': The density $p(x|\theta)$ is twice continuously differentiable in θ for all x, and there is a $\bar{r} = \bar{r}(K) > 0$ so that for each i, j and every $\theta \in K$,

$$E_\theta \left[\sup_{\theta': |\theta'-\theta| \leq r} |\hat{J}_{ij,1}(\theta')|^2 \right] \quad (24)$$

is finite and continuous as a function of θ for $r \leq \bar{r}$.

Assumption 2': $J(\theta)$ is continuous and positive definite in Θ. Furthermore, the gradient $\nabla \log p(x|\theta)$ has expectation $E[\nabla \log p(X|\theta)] = \int \nabla p(x|\theta)dx$ equal to zero with covariance matrix

$$I(\theta) = E_\theta[\nabla \log p(X|\theta)\nabla^t \log p(X|\theta)]$$

also finite and continuous in Θ.

Here $I(\theta)$ and $J(\theta)$ are the two forms of Fisher Information. Usually $I(\theta)$ equals $J(\theta)$, though we do not need that here.

Assumption 3': The maximum likelihood estimator $\hat{\theta}$ is a consistent estimator of θ, indeed consistent uniformly in K, with tail probability tending to zero faster than $1/\log n$, that is,

$$\sup_{\theta \in K} P_\theta\{|\hat{\theta}(x^n) - \theta| > \delta\} = o(1/\log n),$$

for each $\delta > 0$.

Demonstration of consistency, indeed with tail probability tending to zero at rate $O(1/n)$ is given in the Appendix under sensible conditions.

Assumption 1' differs from Assumption 1 because of the presence of the square of the empirical Fisher information in (24). With Assumption 1' using Chebyshev's inequality we can obtain the following property which strengthens (9)

$$\max_{\theta \in K} P_\theta \left(\inf_{z \neq 0} \inf_{\theta': |\theta'-\theta| \leq \delta} \frac{z^t \hat{J}(\theta', x^n)z}{z^t J(\theta)z} \leq 1 - \epsilon \right) \quad (25)$$
$$= O(1/n).$$

Lemma 2: Under Assumptions 1', 2', 3', $\sqrt{n}J(\theta)^{1/2}(\hat{\theta}-\theta)$ converges in distribution to a mean zero normal random vector and there is a constant c such that the following inequality holds for any positive b,

$$\max_{\theta \in K} P_\theta \left(\sqrt{n}||J(\theta)^{1/2}(\hat{\theta}-\theta)|| > b\sqrt{\log n} \right)$$
$$\leq \frac{c}{b^2 \log n} + o\left(\frac{1}{\log n} \right).$$

The constant c equals $\max_{\theta \in K} \text{trace}(J(\theta)^{-1}I(\theta))$ which is the dimension d, the trace of an identity matrix, when $I(\theta) = J(\theta)$.

This Lemma 2 is proved in the Appendix.

Let $K_\delta = \{\theta \in \Theta : B(\theta, \delta) \subset K\}$ be the δ interior of K, where we are using the ordinary Euclidean ball $B(\theta, r) = \{\theta' \in R^d : ||\theta'-\theta|| \leq r\}$ and let $K \setminus K_\delta$ be the corresponding δ shell of K. (Here we could have used the ellipses based on $J(\theta)$ in place of Euclidean balls in defining an analogous shell.)

Assumption 4': The set $K \subseteq \Theta$ has positive and finite

$$C_J(K) = \int_K |J(\theta)|^{1/2}d\theta$$

and, with δ_n of order $\sqrt{\log n}/\sqrt{n}$, the contribution of the shell $\int_{K \setminus K_{\delta_n}} |J(\theta)|^{1/2}d\theta$ is of order $o(1/\log n)$.

Note that, if the boundary of K has finite surface measure, then the measure of the δ shell is of order δ, which is of order $\sqrt{\log n}/\sqrt{n}$ with the chosen δ_n, so the shell measure vanishes at rate much faster than the required $o(1/\log n)$.

The asymptotic lower bound of Theorem 1 was established by using a code distribution (16) analogous to the Jeffreys mixture code, but with modification by restriction for each x^n to θ in $(\mathscr{B} \cap \mathscr{N})_{x^n}$. It is of interest to know whether a Bayes mixture using Jeffreys prior on all of K has the same asymptotic lower bound. The next result (Theorem 2) shows this to be true for compact K with slightly more stringent assumptions. Indeed with these assumptions the Jeffreys mixture provides a lower bound which is as good as the asymptotically maximin value.

Theorem 2: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional family of probability densities. Suppose that K is a compact subset of Θ for which Assumptions 1', 2', 3', and 4' hold. Define a density m_K over $\mathcal{K} = \mathcal{X}^n(K)$ as

$$m_K(x^n) = \frac{\bar{m}_K(x^n)}{C_n}$$
$$\bar{m}_K(x^n) = 1_\mathcal{K}(x^n) \int_K p(x^n|\theta)w_K(\theta)d\theta,$$

where C_n is the normalization constant and $w_K(\theta)$ is the Jeffreys prior on K. Then the following holds

$$\liminf_{n\to\infty}\left(\underline{r}_n(m_K) - \frac{d}{2}\log\frac{n}{2\pi}\right) \geq \log C_J(K).$$

Remark: The code distribution $m_K(x^n)$ here differs from code distribution m_δ used in Theorem 1 because here we do not make the restriction to $\mathscr{B}\cap\mathscr{N}$.

Proof: We examine the value $\underline{r}_n(m_K)$. We have

$$E_{m_K}\log\frac{p(x^n|\hat\theta)}{m_K(x^n)}$$

$$= \int_{\mathcal{K}} m_K(x^n)\log\frac{p(x^n|\hat\theta)}{m_K(x^n)}\nu(dx^n)$$

$$= \int_{\mathcal{K}} m_K(x^n)\log\frac{p(x^n|\hat\theta)}{\bar m_K(x^n)}\nu(dx^n) + \log C_n$$

$$= \frac{1}{C_n}\int_{\mathcal{K}}\int_K p(x^n|\theta)w_K(\theta)d\theta\log\frac{p(x^n|\hat\theta)}{\bar m_K(x^n)}\nu(dx^n)$$
$$+ \log C_n.$$

We proceed to lower bound it as follows, where m_δ is the same as in Lemma 1 and $1_{\mathscr{G}\cap\mathscr{N}} = 1_{\mathscr{G}\cap\mathscr{N}}(\theta,x^n)$,

$$\geq \frac{1}{C_n}\int_{\mathcal{K}}\int_K 1_{\mathscr{G}\cap\mathscr{N}}\, p(x^n|\theta)w_K(\theta)d\theta\log\frac{p(x^n|\hat\theta)}{\bar m_K(x^n)}\nu(dx^n)$$
$$+ \log C_n$$

$$= \frac{\bar C_n}{C_n}\int_{\mathcal{K}} m_\delta(x^n)\log\frac{p(x^n|\hat\theta)}{\bar m_K(x^n)}\nu(dx^n) + \log C_n$$

$$= \frac{\bar C_n}{C_n}\int_{\mathcal{K}} m_\delta(x^n)\log\frac{p(x^n|\hat\theta)}{m_K(x^n)}\nu(dx^n) + (1-\frac{\bar C_n}{C_n})\log C_n$$

$$\geq \frac{\bar C_n}{C_n} E_{m_\delta}\log\frac{p(x^n|\hat\theta)}{m_\delta(x^n)} + (1-\frac{\bar C_n}{C_n})\log C_n$$

$$\geq \bar C_n E_{m_\delta}\log\frac{p(x^n|\hat\theta)}{m_\delta(x^n)} + \log C_n,$$

where the first inequality in the above manipulation follows from $p(x^n|\hat\theta)/\bar m_K(x^n) \geq 1$, the second inequality follows from the positivity of Kullback divergence of m_δ from m_K, and the last inequality follows from $C_n \leq 1$, which implies $\log C_n \leq 0$.

We appeal to Lemma 1 using $A = K$. By Assumption $1'$, the expected supremum in (8) is continuous in θ and by the monotone convergence theorem it converges monotonically to 0 as $\delta \to 0$, so by Dini's Theorem (see e.g. ([38]), p.150) the convergence is uniform on the compact K. Accordingly, $\eta_{0,\delta}(K)$ tends to zero as $\delta \to 0$. Likewise, by compactness, $\eta_{1,\delta}(K)$ and $\eta_{2,\delta} = \eta_{2,\delta}(K)$ tend to zero as $\delta \to 0$. Given $\epsilon > 0$, pick δ such that $\eta_{0,\delta}(K) \leq \epsilon$ and $\eta_{1,\delta}(K) \leq 3$.

By Lemma 1, the expectation in the last line above is lower bounded as

$$E_{m_\delta}\log\frac{p(x^n|\hat\theta)}{m_\delta(x^n)} \geq \log\frac{(1-\epsilon)^{d/2}C_J(K)n^{d/2}}{(1+\eta_{2,\delta})(2\pi)^{d/2}} - o(1).$$

Picking $\delta > 0$ and $\epsilon > 0$ sufficiently small, the above lower bound is sufficient for our purpose, provided $\bar C_n$ and C_n converges to 1 at a rate sufficiently fast. In fact, noting that the

lower bound contains the term $(d/2)\log n$, we need to prove $\bar C_n = 1 - o(1/\log n)$ and $C_n = 1 - o(1)$, which imply the claim of Theorem 2.

First examine $\bar C_n$. Here, we already have $\bar C_n = 1 - o(1)$ in the proof of Lemma 1, but it is not sufficient for our purpose. However, using Assumptions $1'$-$3'$, we can show $\bar C_n = 1 - o(1/\log n)$. Indeed, by (20), we have

$$\bar C_n \geq \int_{K^\circ}(P_\theta(\mathscr{N}_\theta) + P_\theta(\mathscr{B}_\theta) - 1)w_K(\theta)d\theta. \qquad (26)$$

Here, by (25), we have

$$\min_{\theta\in K}P_\theta(\mathscr{N}_\theta) \geq 1 - o(1/\log n).$$

Hence

$$\int_K P_\theta(\mathscr{N}_\theta)w_K(\theta)d\theta \geq 1 - o(1/\log n). \qquad (27)$$

For evaluation of $P_\theta(\mathscr{B}_\theta)$, we employ Lemma 2. Let $\delta_n = \sqrt{b\log n}/\sqrt{n}$. Then by Lemma 2,

$$\min_{\theta\in K}P_\theta(\hat\theta \in B(\theta,\delta_n/2)) \geq 1 - \frac{4c}{b^2\log n} - o\left(\frac{1}{\log n}\right).$$

Provided $\theta \in K_{\delta_n}$, $\hat\theta \in B(\theta,\delta_n/2)$ implies $\hat\theta \in K$ and $\hat\theta \in B_\delta(\theta)$ for sufficiently large n, that is, $x^n \in \mathscr{B}_\theta$. Then, we have

$$\min_{\theta\in K_{\delta_n}}P_\theta(\mathscr{B}_\theta) \geq 1 - \frac{4c}{b^2\log n} - o\left(\frac{1}{\log n}\right).$$

Hence we have

$$\int_K P_\theta(\mathscr{B}_\theta)w_K(\theta)d\theta$$
$$\geq \int_{K_{\delta_n}} P_\theta(\mathscr{B}_\theta)w_K(\theta)d\theta$$
$$\geq \frac{C_J(K_{\delta_n})}{C_J(K)}\left(1 - \frac{4c}{b^2\log n} - o\left(\frac{1}{\log n}\right)\right).$$

Here, by Assumption $4'$, $C_J(K_{\delta_n})/C_J(K) = 1 - o(1/\log n)$ holds. Hence, we have

$$\int_K P_\theta(\mathscr{B}_\theta)w_K(\theta)d\theta \geq 1 - \frac{4c}{b^2\log n} - o\left(\frac{1}{\log n}\right).$$

Together with (26) and (27), we have

$$\bar C_n \geq 1 - \frac{4c}{b^2\log n} - o\left(\frac{1}{\log n}\right).$$

Noting that b can be arbitrarily large, we have $\bar C_n \geq 1 - o(1/\log n)$.

Next, we will examine C_n, for which we have

$$C_n = \int_{\mathcal{K}}\int_K 1_K(x^n)p(x^n|\theta)w_K(\theta)d\theta \qquad (28)$$

$$= \int_K P_\theta(\hat\theta \in K)w_K(\theta)d\theta. \qquad (29)$$

By the same way as for $\bar C_n$, we can prove this is $1 - o(1/\log n)$. *The proof is completed.*

B. Lower Bounds for Stochastic Processes

Here, we consider the case in which the model consists of non i.i.d. stochastic processes.

Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a parameterized family of joint densities of a stochastic process, i.e. $p(x^n|\theta)$ satisfies

$$p(x^n|\theta) = \int_{\mathcal{X}} p(x^{n+1}|\theta)\nu(dx_{n+1})$$

and

$$\int_{\mathcal{X}} p(x_1|\theta)\nu(dx_1) = 1.$$

We introduce the following notation

$$J_{n,\theta} = J_n(\theta) \stackrel{\text{def}}{=} E_\theta \hat{J}(\theta, x^n).$$

In this subsection, define the Jeffreys prior for each n over K as

$$w_K^{(n)}(\theta) = \frac{|J_{n,\theta}|^{1/2}}{C_{J,n}(K)}$$

where $C_{J,n}(K)$ is the Jeffreys integral

$$C_{J,n}(K) = \int_K |J_{n,\theta}|^{1/2} d\theta$$

based on the Fisher information indexed by n.

With the use of $J_{n,\theta}$ we modify the definition of $B_\delta(\theta)$ as

$$B_\delta(\theta) = B_{n,\delta}(\theta) = \{\theta' \in \Theta : (\theta' - \theta)^t J_{n,\theta}(\theta' - \theta) \leq \delta^2\}.$$

We use the following assumptions.

Assumption 1″: For each i, j, for each $\theta \in K$, and for each $\tilde{\epsilon} > 0$, there is a $\delta = \delta(\theta, \tilde{\epsilon}) > 0$, such that

$$\sup_{\theta' \in B_\delta(\theta)} \left(\pm(J_{n,ij}(\theta) - \hat{J}_{ij}(\theta', x^n))\right) \leq \tilde{\epsilon} \quad (30)$$

holds, with P_θ probability converging to 1 as $n \to \infty$.

To define $B_\delta(\theta)$ properly, the smallest eigenvalue of $J_n(\theta)$ must be lower bounded by a positive constant. Hence we use the following assumption, which is used in the upper bound results, too.

Assumption 2″: The matrix $J_n(\theta)$ is positive definite, and the collection of functions $J_n(\theta)$, $n \in \mathbb{N}$ is equicontinuous for $\theta \in \Theta$, with a continuous $\lambda(\theta) > 0$ serving as a lower bound on its smallest eigenvalue, uniformly for $n \in \mathbb{N}$.

It follows from these two assumption that the assertion of Assumption 1″ can be refined. Indeed, it follows that for for each θ in K and for any $\epsilon > 0$, there is an $\tilde{\epsilon} = \tilde{\epsilon}(\epsilon, \theta) > 0$ such that at $\delta = \delta(\tilde{\epsilon}, \theta)$,

$$\sup_{\theta' \in B_\delta(\theta)} \left(\pm\left(J_n(\theta)^{-1/2}(J_n(\theta) - \hat{J}(\theta', x^n))J_n(\theta)^{-1/2}\right)_{ij}\right) \leq \frac{\epsilon}{d}$$

with P_θ probability converging to 1 as $n \to \infty$, and consequently,

$$P_\theta \left(\inf_{z \neq 0} \inf_{\theta' \in B_\delta(\theta)} \frac{z^t \hat{J}(\theta', x^n)z}{z^t J_n(\theta)z} < 1 - \epsilon\right) = o(1) \quad (31)$$

as n goes to infinity. This corresponds to (9) for the i.i.d. case.

From Assumption 2″, we also have

$$\min_{n \in \mathbb{N}} \min_{\theta \in K} \min_{z:|z|=1} z^t J_n(\theta)z > 0. \quad (32)$$

for compact subsets K of Θ. This is because $\lambda(\theta)$, being positive and continuous, is therefore strictly positive on any compact set.

Sufficient condition for Assumptions 1″ and 2″ are given in the appendix for Markov processes.

Define two nice sets as

$$\mathscr{B} = \{(\theta, x^n) : \theta, \hat{\theta}(x^n) \in K, ||\theta - \hat{\theta}(x^n)||_{J_{n,\theta}} \leq \delta/4\}$$
$$(33)$$

$$\mathscr{N} = \left\{(\theta, x^n) : \inf_{\theta' \in B_\delta(\theta)} \inf_{z \neq 0} \frac{z^t \hat{J}(\theta', x^n)z}{z^t J_n(\theta)z} \geq 1 - \epsilon\right\}. \quad (34)$$

Further define

$$\eta_{1,J,\delta} = \sup_{\theta \in K} \sup_{\theta' \in B_\delta(\theta)} \max_{z \neq 0} \frac{z^t J_n(\theta)z}{z^t J_n(\theta')z} - 1, \quad (35)$$

$$\eta_{2,J,\delta} = \sup_{\theta \in K} \sup_{\theta' \in B_\delta(\theta)} \max_{z \neq 0} \frac{z^t J_n(\theta')z}{z^t J_n(\theta)z} - 1, \quad (36)$$

and $\eta_{J,\delta} = \max\{\eta_{1,J,\delta}, \eta_{2,J,\delta}\}$. By equicontinuity of $J_n(\theta)$ and compactness of K_δ, this $\eta_{J,\delta}$ converges to 0 as δ tends to zero, uniformly for $n \in \mathcal{N}$.

For $\epsilon > 0$, let $\delta = \delta_\epsilon > 0$, with $(1 + \eta_{J,\delta})^{1/2} < 2$, be sufficiently small that (31) holds for each $\theta \in K$, and $\delta_\epsilon \to 0$ as $\epsilon \to 0$.

As before, when (30) holds we have

$$\min_{\theta' \in \mathscr{B}_{x^n}} \min_{z \neq 0} \frac{z^t \hat{J}(\theta', x^n)z}{z^t J_n(\theta)z} \geq 1 - \epsilon. \quad (37)$$

Assumption 3″: For each θ in Θ°, the maximum likelihood estimator $\hat{\theta}(x^n)$ is consistent. Namely, $\lim_n |\hat{\theta}(x^n) - \theta| = 0$ in P_θ probability, where P_θ is the distribution determined by the x^n with densities $p(x^n)$ for $n \geq 1$.

Assumption 4″: The Jeffreys integral $C_{J,n}(K)$ is finite.

We can prove the following.

Lemma 3: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional family of stochastic processes. We suppose that Assumptions 1″, 2″, 3″, and 4″ hold for S. Define a density m_δ supported on the set

$$\{x^n : (\mathscr{N} \cap \mathscr{B})_{x^n} \neq \emptyset\} \subset \mathcal{K},$$

by

$$\bar{m}_\delta(x^n) = \int_{(\mathscr{N} \cap \mathscr{B})_{x^n}} p(x^n|\theta)w_K^{(n)}(\theta)d\theta,$$

$$m_\delta(x^n) = \frac{\bar{m}_\delta(x^n)}{\bar{C}_n},$$

where

$$\bar{C}_n = \int_{\mathcal{K}} \bar{m}_\delta(x^n)\nu(dx^n). \quad (38)$$

Then

$$E_{m_\delta} \log \frac{p(x^n|\hat{\theta})}{m_\delta(x^n)} \geq \log \frac{(1-\epsilon)^{d/2}C_{J,n}(K)n^{d/2}}{(1+\eta_{J,\delta})(2\pi)^{d/2}} - o(1)$$

holds as n goes to infinity.

Proof: Similar to the proof of Lemma 1.

The following result, under the stated assumptions in the non i.i.d. setting, states that $\log C_{J,n}(K)$ provides control on the minimax regret from below. It corresponds to Theorem 1 in the i.i.d. setting.

Theorem 3: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional family of stochastic processes. Suppose that Assumptions $1''$, $2''$, $3''$, and $4''$ hold for S. Let K be an arbitrary measurable subset of Θ. Then the following holds

$$\liminf_{n\to\infty}\Big(r_n(\mathcal{K}) - \frac{d}{2}\log\frac{n}{2\pi} - \log C_{J,n}(K)\Big) \geq 0.$$

Proof: The conclusion is a consequence of Lemma 3 and is shown in the same fashion as Theorem 1.

V. UPPER BOUNDS

For the upper bound we handle both i.i.d. and stochastic process cases at once. We need strength Assumption $2''$, which is a certain equi-continuity of Fisher information. Similarly as for the lower bound results, we use a compact subset K of Θ°. In this section, K denotes a convex set which is the closure of an open set.

We employ the following stronger assumption than Assumption $2''$.

Assumption $2'''$: The matrix $J_n(\theta)$ is positive definite, and the collection of functions $J_{n,ij}(\theta)$ is Lipschitz continuous in K, uniformly for all $n \in \mathbb{N}$ and for all i, j,

$$|J_{n,ij}(\theta') - J_{n,ij}(\theta)| \leq M|\theta' - \theta|$$

with a continuous $\lambda(\theta) > 0$ serving as a lower bound on its smallest eigenvalue, uniformly for $n \in \mathbb{N}$.

The main theorem gives the upper bound which matches the lower bound we have obtained.

To obtain the upper bound, we will construct an enlarged model of probability densities from the model S.

Define a matrix-valued random variable $V \in \mathbb{R}^{d \times d}$ as

$$V_n(\theta) = V(x^n|\theta) = J_{n,\theta}^{-1/2}\hat{J}(\theta, x^n)J_{n,\theta}^{-1/2} - I.$$

We also use the notation $V_{ij,n}(\theta)$ to denote the (i,j)-entry of $V_n(\theta)$. Let $\mathcal{B} = (-b/2, b/2)^{d \times d}$ for some small $b > 0$. We consider the case where the following assumption holds.

Assumption 5: For every compact set K included in Θ°, there exists a $b > 0$ and a $C_1 = C_{1,K} > 0$, such that the following holds.

$$\forall n \in \mathbb{N}, \ \forall \theta \in K, \ \forall \beta \in \mathcal{B},$$

$$\left(\int p(x^n|\theta)\exp(nV_n(\theta)\cdot\beta)\nu_n(dx^n)\right)^{1/n} < C_1, \quad (39)$$

where $V(x^n|\theta)\cdot\beta$ denotes the matrix inner product of Frobenius

$$\text{trace}(V_n(\theta)\beta^t) = \sum_{ij} V_{ij,n}(\theta)\beta_{ij}.$$

Define a function $\psi_n(\theta, \beta)$ as

$$\psi_n(\theta, \beta) = \frac{1}{n}\log\int p(x^n|\theta)\exp(nV_n(\theta)\cdot\beta)\nu_n(dx^n).$$

Note that Assumption 5 is the boundedness of $\psi_n(\theta, \beta)$.

We define the enlarged family $\bar{S}$ by

$$\bar{S} = \{p_e(x^n|\theta, \beta) : \theta \in \Theta, \beta \in \mathcal{B}\},$$

$$p_e(x^n|\theta, \beta) = p(x^n|\theta)\exp\Big(n(V_n(\theta)\cdot\beta - \psi_n(\theta, \beta))\Big).$$

Let $u = (\theta, \beta)$ and $\mathcal{U} = \{u = (\theta, \beta) : \theta \in \Theta, \beta \in \mathcal{B}\}$. We have

$$\frac{\partial\psi_n(\theta, \beta)}{\partial\beta_{ij}} = E_{\theta,\beta}V_{ij,n}(\theta)$$

and

$$\frac{\partial^2\psi_n(\theta, \beta)}{\partial\beta_{kl}\partial\beta_{ij}}$$
$$= n\Big(E_{\theta,\beta}V_{kl,n}(\theta)V_{ij,n}(\theta) - E_{\theta,\beta}V_{kl,n}(\theta)E_{\theta,\beta}V_{ij,n}(\theta)\Big),$$

where $E_{\theta,\beta}$ denotes expectation with respect to $p_e(x^n|\theta, \beta)$. The latter is the covariance of $\sqrt{n}V_{ij,n}(\theta)$ and $\sqrt{n}V_{kl,n}(\theta)$ at $p_e(x^n|\theta, \beta)$. Correspondingly, let $\text{Cov}_n(\theta, \beta)$ denote the matrix whose (ij, kl)-entry is

$$[\text{Cov}_n(\theta, \beta)]_{ij,kl} = \frac{\partial^2\psi_n(\theta, \beta)}{\partial\beta_{ij}\beta_{kl}}.$$

We let λ_n^* denote the maximum of the largest eigenvalue of $\text{Cov}_n(\theta, \beta)$ among $(\theta, \beta) \in \Theta \times \mathcal{B}$.

When the model is i.i.d., the enlarged family can be single-letterized as

$$\bar{S} = \{p_e(x|\theta, \beta) : \theta \in \Theta, \beta \in \mathcal{B}\},$$

$$p_e(x|\theta, \beta) = p(x|\theta)\exp\Big(V_1(\theta)\cdot\beta - \psi(\theta, \beta)\Big),$$

where (39) is reduced to

$$\int p(x|\theta)\exp(V_1(\theta)\cdot\beta)\nu(dx) < C_1,$$

and

$$\psi(\theta, \beta) = \log\int p(x|\theta)\exp(V_1(\theta)\cdot\beta)\nu(dx).$$

Note that the enlarged family $p_e(x^n|\theta, \beta)$ is still i.i.d., that is, $p_e(x^n|\theta, \beta) = \prod_{t=1}^n p_e(x_t|\theta, \beta)$. Further we have

$$p_e(x^n|\theta, \beta) = p(x^n|\theta)\exp\Big(n(V_n(\theta)\cdot\beta - \psi(\theta, \beta))\Big).$$

where the following holds.

$$V(x^n|\theta) = \frac{1}{n}\sum_{t=1}^n V(x_t|\theta).$$

Then, $\partial^2\psi(\theta, \beta)/\partial\beta_{kl}\partial\beta_{ij}$ is the covariance of $V_{ij}(x|\theta)$ and $V_{kl}(x|\theta)$ at $p_e(x|\theta, \beta)$, and we have

$$[\text{Cov}_n(\theta, \beta)]_{ij,kl} = \frac{\partial^2\psi(\theta, \beta)}{\partial\beta_{kl}\partial\beta_{ij}}.$$

In the i.i.d. case, $[\text{Cov}_n(\theta, \beta)]_{ij,kl} = \partial^2\psi(\theta, \beta)/\partial\beta_{kl}\partial\beta_{ij}$ dose not depend on n, λ_n^* also does not depend on n. Hence, we use the symbols $\text{Cov}(\theta, \beta)$ and λ^* for i.i.d. cases.

For a given compact K included in Θ°, define subsets $G_{n,\delta}$ and $G_{n,\delta}^c$ of $\mathcal{K}$ as follows.

$$G_{n,\delta} = \Big\{x^n : \hat{\theta} \in K, ||V(x^n|\hat{\theta})||_s \leq \delta\Big\},$$

$$G_{n,\delta}^c = \Big\{x^n : \hat{\theta} \in K, ||V(x^n|\hat{\theta})||_s > \delta\Big\},$$

where $||A||_s$ for a symmetric matrix $A \in \Re^{d \times d}$ denotes the spectral norm of A defined as

$$||A||_s = \max_{z:|z|=1} ||Az|| = \max_{z:|z|=1} |z^t A z|.$$

Note the following relation between the spectral norm $||A||_s$ and the Frobenius norm $||A||$ (which is the Euclidean norm of the associated vector with d^2 elements).

$$\frac{||A||}{\sqrt{d}} \le ||A||_s \le ||A||.$$

We call elements of $G_{n,\delta}$ and $G_{n,\delta}^c$ good strings and not good strings, respectively.

We also define the two kinds of neighborhood of θ as

$$B_\epsilon(\theta) = \{\theta' : (\theta' - \theta)^t J_{n,\theta}(\theta' - \theta) \le \epsilon^2\}, \qquad (40)$$

$$\hat{B}_\epsilon(\theta) = \{\theta' : (\theta' - \theta)^t \hat{J}(\hat{\theta})(\theta' - \theta) \le \epsilon^2\}, \qquad (41)$$

where the latter depends on x^n.

We further put the following assumptions.

Assumption 6: Let $w_n(\theta)$ be the prior density function we use for $p(x^n|\theta)$. Assume the following equi-semicontinuity, that is,

$$w_n(\theta') \ge (1 - \eta_r) w_n(\theta)$$

holds for all large n, for all small $r > 0$, for all $\theta \in K$, and for all θ' in $B_r(\theta)$, where $\eta_r = \eta(r)$ is a positive valued function defined on $(0, \infty)$ with $\lim_{r \to 0} \eta_r = 0$.

Assumption 7: There exist a positive number $\bar{h}_K \ge 1$ and a small positive number $\epsilon = \epsilon_K$ such that the following inequality holds for all large n, for all $x^n \in \mathcal{K}$, for all $\tilde{\theta} \in B_\epsilon(\hat{\theta}) \cap K$, and for all z.

$$z^t \hat{J}(\tilde{\theta}, x^n) z \le \bar{h}_K z^t \hat{J}(\hat{\theta}, x^n) z. \qquad (42)$$

This assumption does hold for the mixture family with compact $K \subset \Theta^\circ$. However, there are other families for which we cannot show it holds. To treat those cases too, we use the following weakened assumption. In fact, we prove Theorem 4 under this assumption.

Assumption 7': There exist a positive number $\bar{h}_K \ge 1$ and a small positive number $\epsilon = \epsilon_K$ such that the following inequality holds for all large n, for all $x^n \in \mathcal{K}$, for all $\tilde{\theta} \in B_\epsilon(\hat{\theta}) \cap K$, and for all z.

$$z^t \hat{J}(\tilde{\theta}, x^n) z \le \bar{h}_K \max\{z^t \hat{J}(\hat{\theta}, x^n) z, z^t J_{n,\hat{\theta}} z\}. \qquad (43)$$

Assumption 8: There exist a number $\kappa_J = \kappa_J(K) > 0$ and a small number $\delta_0 = \delta_0(K) > 0$ such that the following inequality holds for all large n, for all $x^n \in G_{n,\delta_0}$, for all small $\epsilon > 0$, for all $\tilde{\theta} \in B_\epsilon(\hat{\theta}) \cap K$, and for all z.

$$z^t \hat{J}(\tilde{\theta}, x^n) z \le (1 + \kappa_J \epsilon) z^t \hat{J}(\hat{\theta}, x^n) z. \qquad (44)$$

Assumption 9: There exists an $\epsilon > 0$ and a $\zeta \in (0, 1)$, such that,

$$\forall n \in \mathbb{N}, \ \forall x^n \in \mathcal{K}, \ \inf_{\tilde{\theta} \in \hat{B}_\epsilon(\hat{\theta}) \cap K} ||V(x^n|\tilde{\theta})||_s \ge \zeta ||V(x^n|\hat{\theta})||_s.$$

If the collection of empirical Fisher information matrices $\hat{J}(\theta) = \hat{J}(\theta, x^n)$ for all x^n and all n, is equicontinuous as functions of θ in K, then Assumptions 9', 10, 11 hold. For Assumption 9, if also the minimum of the smallest eigenvalue

of $\hat{J}_n(\hat{\theta})$ for all n and for all $x^n \in \mathcal{K}$ is lower bounded by a positive constant, then it holds.

The following is used to control the behavior of the ideal prior for stochastic processes. Note that it holds automatically for i.i.d. cases.

Assumption 10: For each $\theta \in K$, $\sup_{n \in \mathbb{N}} ||J_n(\theta)||_s$ is finite.

Under this assumption and Assumption 2''', we can show that $\sup_{n \in \mathbb{N}} \sup_{\theta \in K} ||J_n(\theta)||_s$ is finite, similarly as (32).

Hereafter we use the symbol Φ to denote the probability measure of the d-dimensional standard normal distribution.

Lemma 4: Let m_{w_n} be the mixture of S with respect to a prior density $w_n(\theta)$ over a convex and compact set K included in Θ°. Suppose that Assumptions 2''', 6, and 8 hold. Then, for an arbitrary $\epsilon > 0$, and for all $x^n \in G_{n,\delta}$ with $\delta \le \delta_0$, the following holds

$$\log \frac{p(x^n|\hat{\theta})}{m_{w_n}(x^n)} \le \frac{d}{2} \log \frac{n}{2\pi} + \log \frac{|J_n(\hat{\theta})|^{1/2}}{w_n(\hat{\theta})\Phi(U_{K,\epsilon,n}(\hat{\theta}))}$$
$$+ \log \frac{(1 + \kappa_J \epsilon)^{d/2}(1 + \delta)^{d/2}}{1 - \eta_\epsilon},$$

where

$$U_{K,\epsilon,n}(\theta) = N_{\sqrt{n}\epsilon}(0) \cap (nJ_n(\theta))^{1/2}(K - \theta),$$

and

$$N_r(0) = \{z : |z| \le r\}.$$

Ignoring the restriction to $N_{\sqrt{n}\epsilon}(0)$ the $\Phi(U_{K,\epsilon,n}(\hat{\theta}))$ may be interpreted as the probability that a Gaussian variable of mean $\hat{\theta}$ and covariance $(nJ_n(\hat{\theta}))^{-1}$ belongs to the parameter set K. When $\hat{\theta}$ is sufficiently in the interior this $\Phi(U_{K,\epsilon,n}(\hat{\theta}))$ is near 1 for large n. Whereas, for $\hat{\theta}$ at the boundary or within order $n^{-1/2}$ of the boundary, it provides the modification for the classical Laplace approximation to account for the fact that the near boundary case makes the normal integral approximation be an incomplete normal integral.

Proof of Lemma 4: By a Laplace integration for the mixture m_w restricted in $B_\epsilon(\hat{\theta})$, we have

$$\frac{m_{w_n}(x^n)}{p(x^n|\hat{\theta})} \ge \int_{B_\epsilon(\hat{\theta}) \cap K} e^{-n(\theta - \hat{\theta})^t \hat{J}(\tilde{\theta})(\theta - \hat{\theta})/2} w_n(\theta) d\theta$$
$$\ge (1 - \eta_\epsilon) w_n(\hat{\theta})$$
$$\cdot \int_{B_\epsilon(\hat{\theta}) \cap K} e^{-n(1 + \kappa_J \epsilon)(\theta - \hat{\theta})^t \hat{J}(\hat{\theta})(\theta - \hat{\theta})/2} d\theta,$$

by Assumption 6. . Since $x^n \in G_{n,\delta}$, by Assumption 8,

$$\frac{m_{w_n}(x^n)}{p(x^n|\hat{\theta})} \ge (1 - \eta_\epsilon) w_n(\hat{\theta})$$
$$\cdot \int_{B_\epsilon(\hat{\theta}) \cap K} e^{-n(1 + \kappa_J \epsilon)(1 + \delta)(\theta - \hat{\theta})^t J_n(\hat{\theta})(\theta - \hat{\theta})/2} d\theta.$$

Noting

$$(nJ_n(\hat{\theta}))^{1/2}(B_\epsilon(\hat{\theta}) - \hat{\theta}) = N_{\sqrt{n}\epsilon}(0),$$

by change of variables

$$z = ((1 + \kappa_J \epsilon)(1 + \delta)nJ_n(\hat{\theta}))^{1/2}(\theta - \hat{\theta}),$$

the integration in the last line is transformed as

$$\int_{B_\epsilon(\hat\theta)\cap K} e^{-n(1+\kappa_J\epsilon)(1+\delta)(\theta-\hat\theta)^t J_n(\hat\theta)(\theta-\hat\theta))/2}d\theta$$

$$=\frac{\int_{(1+\kappa_J\epsilon)^{1/2}(1+\delta)^{1/2}U_{K,\epsilon,n}(\theta)} e^{-|z|^2/2}dz}{(1+\kappa_J\epsilon)^{d/2}(1+\delta)^{d/2}n^{d/2}|J_n(\hat\theta)|^{1/2}},$$

which is larger than

$$\frac{(2\pi)^{d/2}\Phi(U_{K,\epsilon,n}(\theta))}{(1+\kappa_J\epsilon)^{d/2}(1+\delta)^{d/2}n^{d/2}|J_n(\hat\theta)|^{1/2}},$$

since $U_{K,\epsilon,n}(\theta)$ is included in

$$(1+\kappa_J\epsilon)^{1/2}(1+\delta)^{1/2}\big(N_{\sqrt{n}\epsilon}(0)\cap(nJ_n(\theta))^{1/2}(K-\theta)\big)$$

because K is a convex set. Hence we have

$$\frac{m_{w_n}(x^n)}{p(x^n|\hat\theta)}\geq\frac{(1-\eta_\epsilon)w_n(\hat\theta)(2\pi)^{d/2}\Phi(U_{K,\epsilon,n}(\theta))}{(1+\kappa_J\epsilon)^{d/2}(1+\delta)^{d/2}n^{d/2}|J_n(\hat\theta)|^{1/2}},$$

which yields the claim of the lemma. *This completes the proof of Lemma 4.*

Lemma 4 suggests the following choice of a prior density:

$$w_{K,\epsilon,n}(\theta)=\frac{|J_n(\theta)|^{1/2}}{C_{K,\epsilon,n}\Phi(U_{K,\epsilon,n}(\theta))},$$

where $C_{K,\epsilon,n}$ is the normalization constant. This idea is that the prior proportional to $|J_n(\theta)|^{1/2}/\Phi(U_{K,\epsilon,n}(\theta))$ provides the approximate modification to Jeffreys prior to equalize the regret in accordance with Laplace approximation, not only in the interior, but also for $\hat\theta$ near the boundary of K.

However, unfortunately this prior $w_{K,\epsilon,n}$ does not satisfy Assumption 6, as seen below. Recall that the assumption requires that the prior density $w_n(\theta')$ keeps its value over the range $B_\epsilon(\theta)$. We can make a counter example as follows. Assume that θ is at the boundary of K and that $\Phi(N_{\sqrt{n}\epsilon}(0))$ almost equals 1. Then, $\Phi(U_{K,\epsilon,n}(\theta))$ is smaller than a half, which means that $w_{K,\epsilon,n}(\theta)$ is larger than twice the value of Jeffreys prior. Note that there exists θ' in $B_\epsilon(\theta)$ such that $B_\epsilon(\theta')\subset K$, provided ϵ is sufficiently small. For that θ', $\Phi(U_{K,\epsilon,n}(\theta'))$ almost equals 1. Hence, $w_{K,\epsilon,n}(\theta')\lessapprox w_{K,\epsilon,n}(\theta)/2$ holds.

To overcome this problem, we employ the slightly modified prior than $w_{K,\epsilon,n}$, which we call the ideal priors for the good strings. To this end, we define a modified region of $U_{K,\epsilon,n}(\theta)$ as below.

$$U_{K,\epsilon,n}^{(\alpha)}(\theta)=N_{\sqrt{n}\epsilon}(0)\cap((n/\alpha^2)J_n(\theta))^{1/2}(K-\theta)$$

where α is a positive real value.

Remark: If $J_n(\theta)$ does not depend on n, i.e. $J_n(\theta)=J(\theta)$, then

$$U_{K,\epsilon,n}^{(\alpha)}(\theta)=U_{K,\alpha\epsilon,n/\alpha^2}(\theta)$$

holds.

Note that

$$U_{K,\epsilon,n}^{(\alpha)}(\theta)\subset U_{K,\epsilon,n}(\theta)$$

holds for $\alpha\geq 1$, since K is a convex set, which means

$$\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))\leq\Phi(U_{K,\epsilon,n}(\theta)). \qquad (45)$$

Definition 1 (The ideal prior for the good strings): Define the prior density $w_{K,\epsilon,n}^{(\alpha)}$ over K as

$$w_{K,\epsilon,n}^{(\alpha)}(\theta)=\frac{|J_n(\theta)|^{1/2}}{C_{K,\epsilon,n}^{(\alpha)}\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))},$$

where

$$C_{K,\epsilon,n}^{(\alpha)}=\int_K\frac{|J_n(\theta)|^{1/2}}{\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))}d\theta,$$

where we assume $n\epsilon^2\to\infty$, $\alpha/\sqrt{n}\epsilon\to\infty$ and $\epsilon\alpha\to 0$ as n goes to infinity.

Remark: As for setting of ϵ and α, we may let $\epsilon=\sqrt{\log n/n}$ and $\alpha=\log n$ for example.

For this prior we can show that Assumption 6 holds under Assumption 2''' and that $C_{J,n}(K)/C_{K,\epsilon,n}^{(\alpha)}\to 1$ as n goes to infinity. First, we will show the latter.

Note that the following holds.

Proposition 1: For all $n>0$ and $\epsilon>0$,

$$\Phi(N_{\sqrt{n}\epsilon}(0))\geq 1-\exp\Big(-\frac{n\epsilon^2}{2}\Big(1-\frac{d}{n\epsilon^2}\log\frac{n\epsilon^2}{d}\Big)+\frac{d}{2}\Big),$$

which is larger than

$$1-\exp\Big(-\frac{n\epsilon^2}{4}+\frac{d}{2}\Big),$$

when $n\epsilon^2/d\geq 2$.

See Appendix H for the proof.

Using this proposition, we have the following lemma, where

$$\rho_n^{(\alpha)}(\epsilon,\theta)=\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))$$

and

$$\rho_n^{(\alpha)}(\epsilon)=\inf_{\theta\in K}\rho_n^{(\alpha)}(\epsilon,\theta).$$

Lemma 5: Let K be an arbitrary compact set in Θ°. Suppose Assumption 2''' holds and let λ denote a lower bound on the smallest eigenvalue of $J_{n,\theta}$ among $n\in\mathbb{N}$ and $\theta\in K$. Let

$$K_\epsilon=\{\theta:B_\epsilon(\theta)\subset K\},$$

Then for ϵ such that $n\epsilon^2/d\geq 2$ and $\epsilon^2\alpha^2\leq\lambda$, we have

$$C_{J,n}(K)\leq C_{K,\epsilon,n}^{(\alpha)}\leq\frac{C_{J,n}(K)}{1-e^{-n\epsilon^2/4+d/2}}+\frac{C_{J,n}(K\setminus K_{\epsilon\alpha})}{\rho_n^{(\alpha)}(\epsilon)},$$

where

$$C_{J,n}(A)=\int_A|J_{n,\theta}|^{1/2}d\theta.$$

The proof is in Appendix I.

We have a lower bound for $\rho_n^{(\alpha)}(\epsilon)$ as follows, where $\text{vol}(A)$ denotes the volume of $A\in\mathbb{R}^d$, $\text{diam}(A)$ the diameter of $A\in\mathbb{R}^d$, and V_k the volume of the k-dimensional sphere of radius 1, that is,

$$V_k=\frac{\pi^{k/2}}{\Gamma(k/2+1)}.$$

Lemma 6: Assume that K is compact, convex, and the closure of an open set in $\Re^d$. Then, for all ϵ such that $\text{vol}(N_\epsilon(0))\leq\text{vol}(J_{n,\theta}^{1/2}K)/2$, the following holds.

$$\rho_n^{(\alpha)}(\epsilon,\theta)\geq\frac{\text{vol}(J_{n,\theta}^{1/2}K)}{2(\text{diam}(J_{n,\theta}^{1/2}K))^dV_d}\Phi(N_{\sqrt{n}\epsilon}(0)).$$

Since Assumptions 2‴ and 10, the minimum eigenvalue of $J_{n,\theta}$ has a positive lower bound and the maximum eigenvalue has an upper bound, which are uniform for $\theta \in K$. Hence, by Lemma 6, $\rho_n^{(\alpha)}(\epsilon)$ is lower bounded by a positive constant when $\sqrt{n}\epsilon/\alpha$ is not small.

Noting this fact, we can show the following lemma.

Lemma 7: Suppose that Assumptions 2‴ and 10 hold. If $\epsilon/\alpha \to 0$ and $n\epsilon^2/\alpha^2 \to \infty$ as n goes to infinity, then

$$\lim_{n\to\infty}\left(\log C_{K,\epsilon,n}^{(\alpha)} - \log C_{J,n}(K)\right) = 0.$$

Proof: Since $\sup_{\theta \in K}\sup_n |J_n(\theta)|^{1/2}$ is finite under Assumption 10, $C_{J,n}(K \setminus K_{\epsilon\alpha})$ converges to zero uniformly for all n, as ϵ tends to zero. Noting that $\rho_n^{(\alpha)}$ is lower bounded by a certain positive constant, since Lemma 6. Hence by Lemma 5, which is shown under Assumption 2‴, we have the claim of the lemma. *This completes the proof.*

To show that Assumption 6 holds for the ideal priors, we will show the following lemma.

Lemma 8: Let λ_n denote the maximum of the largest eigenvalue of $J_n(\theta)$ among $\theta \in K$. For a certain $\epsilon > 0$, for all $r \leq \epsilon$, for all $\theta' \in B_r(\theta) \cap K$, and for all $\theta \in K$,

$$\frac{\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta'))}{\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))}$$
$$\leq 1 + \frac{\sqrt{n\lambda_n}r}{\rho_n(\epsilon,\theta)\alpha} + \frac{C_d \operatorname{diam}(K)\sqrt{n\lambda_n}\max\{g_{\theta,r},0\}}{\rho_n(\epsilon,\theta)\alpha}$$

holds, where $C_d = 2^{1-d/2}\Gamma(d/2)$ and

$$g_{\theta,r} = \max_{\theta' \in B_r(\theta)\cap K}||J_n(\theta)^{-1/2}J_n(\theta')^{1/2}||_s - 1.$$

The proof is in Appendix K.

Note that $\rho_n^{(\alpha)}(\epsilon,\theta)$ is lower bounded by a positive constant, since $\sqrt{n}\epsilon \to \infty$ when n goes to infinity because of Lemma 6. Since $g_{\theta,r}$ is of order r uniformly over K by Assumption 2‴, we have

$$\sup_{\theta' \in K\cap B_r(\theta)}\sup_{\theta \in K}\frac{\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))}{\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta'))} \geq 1 - O\left(\frac{\sqrt{n}\epsilon}{\alpha}\right), \quad (46)$$

when $\sqrt{n}\epsilon/\alpha$ tends to 0. From this, we have

$$\inf_{\theta \in K}\inf_{\theta' \in B_r(\theta)}\frac{w_{K,\epsilon,n}^{(\alpha)}(\theta')}{w_{K,\epsilon,n}^{(\alpha)}(\theta)} \geq 1 - O\left(\frac{\sqrt{n}\epsilon}{\alpha}\right),$$

Since we assume $\sqrt{n}\epsilon/\alpha$ tends to 0 as n goes to infinity, this means that $w_{K,\epsilon,n}^{(\alpha)}$ satisfies Assumption 6. Hence, we can immediately show the following lemma from Lemma 4 and (45). In fact, for the ideal priors, Assumption 6 holds under Assumption 2‴ on Fisher information.

Lemma 9: Let $m_{K,\epsilon,n}^{(\alpha)}$ denote the mixture of S with respect to the prior density $w_{K,\epsilon,n}^{(\alpha)}$. Suppose that Assumptions 2‴, 7, 8, and 10 hold. Then, for all $x^n \in G_{n,\delta}$ and for all $\epsilon > 0$, we have

$$\log\frac{p(x^n|\hat{\theta})}{m_{K,\epsilon,n}^{(\alpha)}(x^n)} \leq \frac{d}{2}\log\frac{n}{2\pi} + \log C_{K,\epsilon,n}^{(\alpha)}$$
$$+ \log\frac{(1+\kappa_J\epsilon)^{d/2}(1+\delta)^{d/2}}{1-\eta_\epsilon}.$$

Here, we treat the strings $x^n \in G_{n,\delta}^c$. We use the mixture of the extended model $\bar{S}$. Define a prior density $\bar{w}(\theta,\beta)$ over $\bar{S}$ as

$$\bar{w}(\theta,\beta) = w_K^{(n)}(\theta)b^{-d^2}.$$

Note that this is the direct product of the Jeffreys prior $w_K^{(n)}(\theta)$ and the uniform probability density over $\mathcal{B}$.

Recall that λ_n^* denotes the maximum of the largest eigenvalue of $\operatorname{Cov}_n(\theta,\beta)$ among $(\theta,\beta) \in \Theta \times \mathcal{B}$. We have the following Lemma.

Lemma 10: Let $\bar{m}$ denote the mixture of $\bar{S}$ with respect to the prior $\bar{w}$. Suppose Assumptions 6, 7′, 9, and 10. Set $\tilde{a} = \min\{1/2\lambda_n^*, 1\}$ and $\tilde{\delta} = \zeta\delta/\sqrt{d}$. Then for all $x^n \in G_{n,\delta}^c$, we have

$$\frac{\bar{m}(x^n)}{p(x^n|\hat{\theta})} \geq \frac{(1-\eta_\epsilon)\pi^{d/2}\Phi(U_{K,\epsilon,n})}{\left(\bar{h}_K(1+||V(x^n|\hat{\theta})||_s)n\right)^{d/2}C_{J,n}(K)}$$
$$\cdot \frac{V_{d^2}(\tilde{a}\tilde{\delta})^{d^2}}{2^{d^2}}\exp\left(\frac{\zeta n||V(x^n|\hat{\theta})||_s\tilde{a}\tilde{\delta}}{16}\right). \quad (47)$$

Remark: The key factor in this expression is the exponential in $n||V(x^n|\hat{\theta})||\tilde{\delta}$, which makes the coding distribution $\bar{m}(x^n)$ large when the empirical Fisher information deviation $V(x^n|\hat{\theta})$ is not small.

Proof: Note that

$$\frac{\bar{m}(x^n)}{p(x^n|\hat{\theta})}$$
$$= \int\frac{p_e(x^n|\theta,\beta)\bar{w}(\theta,\beta)}{p(x^n|\hat{\theta})}d\theta d\beta$$
$$= b^{-d^2}\int\int e^{n(V_n(\theta)\cdot\beta - \psi_n(\theta,\beta))}d\beta\frac{p(x^n|\theta)}{p(x^n|\hat{\theta})}w_K^{(n)}(\theta)d\theta.$$

First, we will evaluate the integral with respect to β assuming $\theta \in B_\epsilon(\hat{\theta})$, which implies $||V_n(\theta)||_s \geq \zeta||V_n(\hat{\theta})||_s > \zeta\delta$, by Assumption 9 and $x^n \in G_{n,\delta}^c$. Hence, $||V_n(\theta)|| \geq \zeta\delta/\sqrt{d} = \tilde{\delta}$ holds. Below, We will find a certain region where the integrand $e^{n(V_n(\theta)\cdot\beta - \psi_n(\theta,\beta))}$ is exponentially large in the polynomial of n.

By Taylor expansion of $\psi_n(\theta,\beta)$ with respect to β at $(\theta,0)$, we have

$$V_n(\theta)\cdot\beta - \psi_n(\theta,\beta)$$
$$= V_n(\theta)\cdot\beta - \beta^t n\operatorname{Cov}_n(\theta,\beta')\beta/2$$
$$\geq V_n(\theta)\cdot\beta - \lambda_n^*||\beta||^2/2, \quad (48)$$

where $\beta' = t\beta$ with a certain $t \in [0,1]$, and

$$\beta^t n\operatorname{Cov}_n(\theta,\beta')\beta = \sum_{ij,kl}\beta_{ij}[n\operatorname{Cov}_n(\theta,\beta')]_{ij,kl}\beta_{kl}.$$

Let $f(\beta)$ denote the last side of (48). Introduce new parameters $(v,\beta_\perp) \in \Re \times \mathcal{B}$ by

$$\beta = v\frac{V_n(\theta)}{||V_n(\theta)||} + \beta_\perp,$$

where $\beta_\perp \cdot V_n(\theta) = 0$ is assumed. Then, we have

$$f(\beta) = ||V_n(\theta)||v - \frac{\lambda_n^*}{2}(v^2 + ||\beta_\perp||^2).$$

Now consider the region $N_{\tilde{a}\tilde{\delta}/2}(\tilde{\beta})$ with $\tilde{\beta} = (\tilde{v}, \tilde{\beta}_\perp) = (\tilde{a}\tilde{\delta}, 0)$. Note that $\tilde{a}\lambda_n^* \leq 1/2$ holds since $\tilde{a} = \min\{1/2\lambda_n^*, 1\}$. For all β in $N_{\tilde{a}\tilde{\delta}/2}(\tilde{\beta})$, we have

$$
\begin{aligned}
&f(\beta)\\
&= \|V_n(\theta)\|v - \frac{\lambda_n^*}{2}\big((\tilde{a}\tilde{\delta} + v - \tilde{a}\tilde{\delta})^2 + \|\beta_\perp\|^2\big)\\
&= \|V_n(\theta)\|v\\
&\quad - \frac{\lambda_n^*}{2}\big(\tilde{a}^2 2\tilde{\delta}^2 + 2\tilde{a}\tilde{\delta}(v - \tilde{a}\tilde{\delta}) + (v - \tilde{a}\tilde{\delta})^2 + \|\beta_\perp\|^2\big)\\
&\geq \|V_n(\theta)\|v - \frac{\lambda_n^*}{2}\Big(\tilde{a}^2 2\tilde{\delta}^2 + 2\tilde{a}\tilde{\delta}(v - \tilde{a}\tilde{\delta}) + \frac{\tilde{a}^2 2\tilde{\delta}^2}{4}\Big)\\
&= \big(\|V_n(\theta)\| - \lambda_n^* \tilde{a}\tilde{\delta}\big)v - \frac{3\lambda_n^*\tilde{a}^2 2\tilde{\delta}^2}{8}\\
&\geq \frac{\|V_n(\theta)\|}{2}v - \frac{3\lambda_n^*\tilde{a}^2 2\tilde{\delta}^2}{8}\\
&\geq \frac{\|V_n(\theta)\|\tilde{a}\tilde{\delta}}{4} - \frac{3\lambda_n^*\tilde{a}^2 2\tilde{\delta}^2}{8}\\
&= \frac{\tilde{a}\tilde{\delta}}{4}\Big(\|V_n(\theta)\| - \frac{3\lambda_n^*\tilde{a}\tilde{\delta}}{2}\Big)\\
&\geq \frac{\tilde{a}\tilde{\delta}}{4}\Big(\|V_n(\theta)\| - \frac{3\tilde{\delta}}{4}\Big)\\
&\geq \frac{\tilde{a}\tilde{\delta}\|V_n(\theta)\|}{16}.
\end{aligned}
$$

To derive the first inequality above, we have used $(v - \tilde{a}\tilde{\delta})^2 + \|\beta_\perp\|^2 \leq (\tilde{a}\tilde{\delta}/2)^2$ for $\beta \in N_{\tilde{a}\tilde{\delta}/2}(\tilde{\beta})$. The second follows from $\lambda_n^*\tilde{a} \leq 1/2$ and $\tilde{\delta} \leq \|V_n(\theta)\|$. The others also follow from them and $v \geq \tilde{a}\tilde{\delta}/2$. In this evaluation, we used the condition $a\lambda_n^* \leq 1/2$, but we actually defined $\tilde{a} = \min\{1/2\lambda_n^*, 1\}$, since too large $\tilde{a}$ may cause the problem that $N_{\tilde{a}\tilde{\delta}/2}(\tilde{\beta})$ is beyond $\mathcal{B}$.

From this, the following lower bound provided $\theta \in B_\epsilon(\hat{\theta})$ is obtained.

$$
\begin{aligned}
&\int e^{n(V(x^n|\theta)\cdot\beta - \psi_n(\theta,\beta))}d\beta\\
&\geq \int_{N_{\tilde{a}\tilde{\delta}/2}(\tilde{\beta})} e^{n(V(x^n|\theta)\cdot\beta - \psi_n(\theta,\beta))}d\beta\\
&\geq \frac{V_{d^2}(\tilde{a}\tilde{\delta})^{d^2}}{2^{d^2}}\exp\Big(\frac{n\|V_n(\theta)\|\tilde{a}\tilde{\delta}}{16}\Big)\\
&\geq \frac{V_{d^2}(\tilde{a}\tilde{\delta})^{d^2}}{2^{d^2}}\exp\Big(\frac{n\zeta\|V_n(\hat{\theta})\|_s\tilde{a}\tilde{\delta}}{16}\Big), \qquad (49)
\end{aligned}
$$

where we have used Assumption 11 and the inequality $\|A\|_s \leq \|A\|$.

Next we will evaluate the integral with respect to θ. Since the bound for the integral with β is uniform for $\theta \in B_\epsilon(\hat{\theta})$, it suffices to evaluate

$$
\int_{B_\epsilon(\hat{\theta})} \frac{p(x^n|\theta)w_K^{(n)}(\theta)}{p(x^n|\hat{\theta})}d\theta.
$$

For that purpose, we make use of (42) in Assumption 7'. We need a case argument depending on which one is the maximum in the right side of (42). For $\theta \in B_\epsilon(\hat{\theta})$, first assume $(\theta -$

$\hat{\theta})^t J_{n,\hat{\theta}}(\theta - \hat{\theta}) \leq (\theta - \hat{\theta})^t \hat{J}(\hat{\theta})(\theta - \hat{\theta})$, then we have

$$
\begin{aligned}
\frac{p(x^n|\theta)}{p(x^n|\hat{\theta})} &\geq \exp\Big(\frac{-n(\theta - \hat{\theta})^t \hat{J}(\theta')(\theta - \hat{\theta})}{2}\Big)\\
&\geq \exp\Big(\frac{-n\bar{h}_K(\theta - \hat{\theta})^t \hat{J}(\hat{\theta})(\theta - \hat{\theta})}{2}\Big),
\end{aligned}
$$

where $\theta' = t\theta + (1 - t)\hat{\theta}$ with a certain $t \in [0, 1]$. The last inequality follows from Assumption 7'. Since

$$
\|(J_{n,\hat{\theta}})^{-1/2}\hat{J}(\hat{\theta})(J_{n,\hat{\theta}})^{-1/2}\|_s \leq \|V(x^n|\hat{\theta})\|_s + 1
$$

holds, we have

$$
\frac{(\theta - \hat{\theta})^t \hat{J}(\hat{\theta})(\theta - \hat{\theta})}{(\theta - \hat{\theta})^t J_{n,\hat{\theta}}(\theta - \hat{\theta})} \leq \|\hat{V}\|_s + 1,
$$

where $\hat{V}$ denotes $V(x^n|\hat{\theta})$. Hence for any $\theta \in B_\epsilon(\hat{\theta})$ with $(\theta - \hat{\theta})^t J_{n,\hat{\theta}}(\theta - \hat{\theta}) \leq (\theta - \hat{\theta})^t \hat{J}(\hat{\theta})(\theta - \hat{\theta})$,

$$
\frac{p(x^n|\theta)}{p(x^n|\hat{\theta})} \geq e^{-n\bar{h}_K(1+\|\hat{V}\|_s)(\theta - \hat{\theta})^t J_{n,\hat{\theta}}(\theta - \hat{\theta})/2}. \quad (50)
$$

Next assume $(\theta - \hat{\theta})^t J_{n,\hat{\theta}}(\theta - \hat{\theta}) > (\theta - \hat{\theta})^t \hat{J}(\hat{\theta})(\theta - \hat{\theta})$ for $\theta \in B_\epsilon(\hat{\theta})$. Then, directly from Assumption 7', we have

$$
\frac{p(x^n|\theta)}{p(x^n|\hat{\theta})} \geq e^{-n\bar{h}_K(\theta - \hat{\theta})^t J_{n,\hat{\theta}}(\theta - \hat{\theta})/2}.
$$

Hence, (50) holds for any $\theta \in B_\epsilon(\hat{\theta})$.

Then, in a manner similar to the proof of Lemma 4, by Laplace approximation over $B_\epsilon(\hat{\theta})$, we have

$$
\begin{aligned}
&\int_{B_\epsilon(\hat{\theta})} \frac{p(x^n|\theta)w_{K,n}(\theta)}{p(x^n|\hat{\theta})}d\theta\\
&\geq \frac{(1 - \eta_\epsilon)(2\pi)^{d/2}\Phi(U_{K,\epsilon,n})w_K^{(n)}(\hat{\theta})}{\bar{h}_K^{d/2}n^{d/2}(1 + \|\hat{V}\|_s)^{d/2}|J_{n,\hat{\theta}}|^{1/2}}\\
&= \frac{(1 - \eta_\epsilon)(2\pi)^{d/2}\Phi(U_{K,\epsilon,n})}{\bar{h}_K^{d/2}n^{d/2}(1 + \|\hat{V}\|_s)^{d/2}C_{J,n}(K)}
\end{aligned}
$$

Together with (49), we have (47). *This completes the proof of Lemma 10.*

Theorem 4: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional family of stochastic processes. Suppose that Assumptions 7, 8, 9', 10, and 11 hold. Further, we assume that $J_n(\theta)$ is equicontinuous for all n and that

$$
\lambda^* = \overline{\lim_{n\to\infty}} \lambda_n^*
$$

is finite. Fix a sequence ϵ_n which converges to 0 slower than $1/\sqrt{n}$. Define a mixture m_n over $\bar{S}$ as

$$
m_n(x^n) = (1 - n^{-r})m_{K,\epsilon_n,n}(x^n) + n^{-r}\bar{m}(x^n).
$$

Then, the following holds.

$$
\begin{aligned}
\overline{\lim_{n\to\infty}}\Big(&\sup_{x^n\in\mathcal{K}}\log\frac{p(x^n|\hat{\theta})}{m_n(x^n)} - \frac{d}{2}\log\frac{n}{2\pi} - \log C_{J,n}(K)\Big)\\
&\leq 0.
\end{aligned}
\tag{51}
$$

Proof: In this proof, we use Lemmas 5 and 8, plugging in $\delta_n = n^{-1/2+\gamma}$ in place of δ, which means that $\tilde{\delta}$ is replaced by

$$\tilde{\delta}_n = \frac{\zeta\delta_n}{\sqrt{d}}.$$

When $x^n \in G_{n,\delta_n}$, by Lemma 9, we have

$$\log\frac{p(x^n|\hat{\theta})}{m_n(x^n)} \le \frac{d}{2}\log\frac{n}{2\pi} + \log C_{K,\epsilon_n,n}$$
$$+ \log\frac{(1+\kappa_J\epsilon_n)^{d/2}(1+\delta_n)^{d/2}}{1-\eta(\epsilon_n)} - \log(1-n^{-r}).$$

Here, δ_n, n^{-r} and $|\log C_{K,\epsilon_n,n} - \log C_{J,n}(K)|$ converge to 0 as n goes to infinity (the last one follows from Lemma 7), hence we have

$$\varlimsup_{n\to\infty} \sup_{x^n \in G_{n,\delta_n}} \left(\log\frac{p(x^n|\hat{\theta})}{m_n(x^n)} - \frac{d}{2}\log\frac{n}{2\pi} - \log C_{J,n}(K)\right)$$
$$\le 0.$$

Next, we move to the case that $x^n \in G_{n,\delta_n}^c$. Since

$$\frac{m_n(x^n)}{p(x^n|\hat{\theta})} \ge \frac{n^{-r}\bar{m}(x^n)}{p(x^n|\hat{\theta})},$$

we evaluate the lower bound of $\bar{m}(x^n)/p(x^n|\hat{\theta})$ using Lemma 10. When n is large and ϵ is small, the right side of (47) times $C_{J,n}(K)$ is not less than $g(||\hat{V}||_s)$ with

$$g(\xi) = \frac{A(\tilde{a}\tilde{\delta}_n)^{d^2}}{(1+\xi)^{d/2}n^{d/2}} \exp\left(\frac{\zeta n\xi\tilde{a}\tilde{\delta}_n}{16}\right),$$

where A is a constant which does not depend on $||\hat{V}||_s$, n, and $\tilde{\delta}_n$. The first derivative of $\log g(\xi)$ is

$$-\frac{d}{2(1+\xi)} + \frac{\zeta n\tilde{a}\tilde{\delta}_n}{16} \ge -\frac{d}{2} + \frac{\zeta n\tilde{a}\tilde{\delta}_n}{16},$$

which is positive when n is large, since $n\tilde{\delta}_n$ diverges to infinity as n goes to infinity. Hence, $g(||\hat{V}||_s)$ is not less than

$$g(\tilde{\delta}_n) = \frac{A(\tilde{a}\tilde{\delta}_n)^{d^2}}{(1+\delta_n)^{d/2}n^{d/2}} \exp\left(\frac{\zeta n\delta_n\tilde{a}\tilde{\delta}_n}{16}\right)$$
$$= \frac{A(\tilde{a}\tilde{\delta}_n)^{d^2}}{(1+\delta_n)^{d/2}n^{d/2}d^{d^2/2}} \exp\left(\frac{\zeta^2 n\tilde{a}\delta_n^2}{16\sqrt{d}}\right)$$

for large n. Therefore, we have

$$\frac{C_{J,n}(K)m_n(x^n)}{p(x^n|\hat{\theta})} \tag{52}$$
$$\ge \frac{n^{-r}A\tilde{a}^{d^2}n^{d^2(-1/2+\gamma)}}{(1+\delta_n)^{d/2}n^{d/2}d^{d^2/2}} \exp\left(\frac{\zeta^2\tilde{a}n^{2\gamma}}{16\sqrt{d}}\right),$$

which diverges to infinity as n goes to infinity. It means that the regret for not bad sequences is negative and can be ignored. *This completes the proof.*

VI. REGRET BOUNDS FOR SPECIAL CASES

In this section, employing the main theorems for general cases, we will show asymptotic minimax bounds on regret of some special models: exponential families, curved exponential families, models with hidden variables including mixture families (with fixed components), and contaminated Gaussian location families. Concerning the mixture families, we show a stronger upper bound than that obtained by Theorem 5, that is, we show that the same form of the asymptotic upper bound is valid without restricting the strings to $\mathcal{K}$.

In the first subsection below, we describe the formal definition of exponential families and review their basic properties. In the remaining subsections, we discuss regret bounds for each example.

A. Definition of Exponential Families

Define the exponential family, following [12], [2].

Definition 2 (Exponential Family): Given Borel measurable functions $T : \mathcal{X} \to \Re^d$ and $U : \mathcal{X} \to \Re$, define the natural parameter space

$$\Theta = \left\{\theta \in \Re^d : \int_{\mathcal{X}} \exp\big(\theta \cdot T(x) + U(x)\big)\nu(dx) < \infty\right\}$$

a subset of $\Re^d$, assumed to have a non-empty interior Θ°. Define $\psi(\theta) = \log\int_{\mathcal{X}} \exp(\theta \cdot T(x) + U(x))\nu(dx)$ and a probability density on $\mathcal{X}$ with respect to ν by

$$p(x|\theta) = \exp\big(\theta \cdot T(x) - \psi(\theta) + U(x)\big).$$

Let K denote a subset of Θ such that $\bar{K} = \bar{K}^\circ$. We refer to the set $S(K) = \{p(x|\theta) : \theta \in K\}$ as an *exponential family* of densities on $\mathcal{X}$.

The class of exponential families includes many common statistical models such as Gaussian, Poisson, Bernoulli sources and so on. The following familiar case expresses how the Gaussian family is an exponential family.

Example 1 (Gaussian distributions): Let $\mathcal{X} = \Re$ and $\nu(dx)$ be the Lebesgue measure dx. The density of $N(\mu, \sigma^2)$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$= \exp\left(\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{\log(2\pi\sigma^2)}{2}\right).$$

Let $\theta = (\mu/\sigma^2, -1/(2\sigma^2))$, $T(x) = (x, x^2)$, $U(x) = 0$ and $\psi(\theta) = \mu^2/(2\sigma^2) + (1/2)\log(2\pi\sigma^2)$. Then, we see that the Gaussian is an exponential family, where $\Theta = \Re \times (-\infty, 0)$.

Hereafter, we absorb the factor $\exp(U(x))$ into the reference measure $\nu(dx)$ (that is, we change $\exp(U(x))\nu(dx)$ to $\nu(dx)$).

It is known that Θ is a convex set. Let $\mathcal{T}$ denote the range of $T(x)$ for x in the support of ν and let $\overline{\mathrm{conv}}(\mathcal{T})$ be the closure of its convex hull. We can assume that the model is arranged to be full rank such that $\dim(\overline{\mathrm{hull}}(\mathcal{T})) = \dim\Theta = d$ holds without loss of generality [12]. An exponential family which satisfies this condition is said to be minimal. We assume $S(\Theta)$ is minimal in this paper.

It is known that $\psi(\theta)$ is of class C^∞ and strictly convex on Θ°. We refer to θ as the natural parameter (or θ-coordinates). We define the expectation parameter (or η-coordinates) as $\eta_i =$

$E_\theta(T_i)$. It is known that the function on Θ° mapping $\theta \mapsto \eta$ is an injection and of class C^∞. Let $\mathcal{H} = \{\eta(\theta) : \theta \in \Theta^\circ\}$ be the range of this map. We also assume that $S(\Theta)$ is steep: that is $E_\theta(|T(x)|) = \infty$ holds for any $\theta \in \Theta - \Theta^\circ$. It is known that $\mathcal{H}$ is the interior of $\mathrm{conv}(\mathcal{T})$ for steep exponential families. Further, if Θ is open, then $S(\Theta)$ is said to be regular. Note that any regular family is steep automatically (vacuously) since $\Theta - \Theta^\circ$ is empty.

Note that $\partial\psi/\partial\theta^i = E_\theta(T_i(x)) = \eta_i$ and $\partial^2\psi/\partial\theta^j\partial\theta^i = E_\theta((T_i(x) - \eta_i)(T_j(x) - \eta_j))$ hold on Θ°. Here, this $E_\theta((T_i(x) - \eta_i)(T_j(x) - \eta_j))$ is the Fisher information matrix with respect to θ denoted $J(\theta)$. We let $I(\eta)$ denote the Fisher information matrix of η. This $I(\eta)$ is known to be the inverse matrix of $J(\theta)$. It is known that $J(\theta)$ is strictly positive definite in Θ°.

Given the string x^n, we have $p(x^n|\theta) = \prod_{\tau=1}^n p(x_\tau|\theta) = \prod_{\tau=1}^n \exp(\theta \cdot T(x_\tau) - \psi(\theta)) = \exp(n(\theta \cdot \bar{T} - \psi(\theta)))$, where

$$\bar{T} = \bar{T}(x^n) = \frac{\sum_{\tau=1}^n T(x_\tau)}{n}.$$

These yield

$$\frac{\partial \log p(x^n|\theta)}{\partial\theta_i} = n(\bar{T}_i - \eta_i),$$

$$\frac{\partial^2 \log p(x^n|\theta)}{\partial\theta_i\partial\theta_j} = -nJ_{ij}(\theta). \tag{53}$$

The latter implies that $\log p(x^n|\theta)$ is strictly concave in Θ°, hence, the former implies that $\log p(x^n|\theta)$ takes the unique maximum at $\eta = \bar{T}$ under the condition that $\bar{T} \in \mathcal{H}$. Let $\hat{\eta} = \hat{\eta}(x^n)$ denote the maximum likelihood estimate of η given x^n. Then, we have $\hat{\eta}(x^n) = \bar{T}(x^n)$ for x^n with $\bar{T}(x^n) \in \mathcal{H}^\circ$. The identity (53) also implies

$$\hat{J}(\theta, x^n) = J(\theta)$$

for all $\theta \in \Theta$ and for all x^n. Therefore, Assumptions 1, 1′, 2, and 2′ hold for the natural parameter θ of exponential familes.

For the expectation parameters $(\tilde{\eta}, \eta) \in \mathcal{H}^2$, corresponding to $p(\cdot|\theta)$ and $p(\cdot|\tilde{\theta})$, let $D(\tilde{\eta}||\eta)$ denote the Kullback divergence from $p(\cdot|\tilde{\theta})$ to $p(\cdot|\theta)$. Then, we have

$$D(\tilde{\eta}||\eta) = E_{\tilde{\theta}}\left(\log\frac{p(x|\tilde{\theta})}{p(x|\theta)}\right) = \tilde{\theta}\cdot\tilde{\eta} - \psi(\tilde{\theta}) - \theta\cdot\tilde{\eta} + \psi(\theta),$$

We extend the domain $\mathcal{H}^2$ to $\Re^d \times \mathcal{H}$, following [12]:

$$D(\tilde{\eta}||\eta)$$
$$= \begin{cases} \lim_{\epsilon\to+0}\inf\{D(\eta'||\eta) : \eta' \in \mathcal{H}, |\tilde{\eta} - \eta'| < \epsilon\}, \\ \qquad\qquad \text{for } (\tilde{\eta}, \eta) \in \partial\mathcal{H} \times \mathcal{H}, \\ \infty, \qquad\qquad \text{for } (\tilde{\eta}, \eta) \in \bar{\mathcal{H}}^c \times \mathcal{H}, \end{cases}$$

where $\partial\mathcal{H}$ is the boundary of $\mathcal{H}$ and $\bar{\mathcal{H}}^c$ is the complement of the closure.

We have the following relation between differentiation with respect to η and differentiation with respect to θ:

$$\frac{\partial}{\partial\eta_i} = \sum_j \frac{\partial\theta_j}{\partial\eta_i}\frac{\partial}{\partial\theta_j} = \sum_j I_{ij}(\eta)\frac{\partial}{\partial\theta_j}.$$

This holds for $\theta \in \Theta^\circ$ (equivalently for $\eta \in \mathcal{H}$). Thus θ and $\psi(\theta)$ are infinitely differentiable with respect to η as well as with respect to θ. Hence, $D(\tilde{\eta}||\eta)$ is of class C^∞ on $\mathcal{H}^2$.

The following is a large deviation inequality.

Lemma 11: Suppose that $\eta \in \mathcal{H}$. Let Λ be a closed half space of $\Re^d$ (i.e. $\Lambda = \{x \in \Re^d : x \cdot \xi \geq \gamma\}$ for any specified $\xi \in \Re^d$ and $\gamma \in \Re$). Then, the following inequality holds

$$P_\theta(\bar{T}(x^n) \in \Lambda) \leq \exp(-n\inf_{\tilde{\eta}\in\Lambda} D(\tilde{\eta}||\eta)).$$

See [16], [12] for the proof.

B. Lower and Upper Bounds for Exponential Families

From Lemma 11, Assumptions 3 and 3′ hold. We continue to assume finiteness of $C_J(K)$ (Assumption 4′). For the regular exponential families, all the remaining assumptions for Lemmas 1 and 2 automatically hold. Hence, Lemmas 1 and 2 and Theorems 1 and 3 can be applied to show the desired lower bound indeed holds for the maximin regret for regular exponential families.

Next, we consider the upper bound. For the natural parameter θ of exponential families, $\hat{J}(\theta, x^n) = J(\theta)$ holds for any $\theta \in \Theta^\circ$ and for any x^n. Hence, if we use the natural parameter, we do not need the fiber bundle of local exponential families to obtain the asymptotic minimax regret. In fact, we can prove the following theorem, where we suppose only Assumptions about K. (K is compact, convex, and included in Θ°.)

Theorem 5: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional exponential family. Let K be a compact and convex subset of Θ°. Let $m_{K,\epsilon_n,n}$ be the same one as in Theorem 5. Then, the following holds.

$$\overline{\lim_{n\to\infty}}\left(\sup_{x^n\in\mathcal{K}}\log\frac{p(x^n|\hat{\theta})}{m_{K,\epsilon_n,n}(x^n)} - \frac{d}{2}\log\frac{n}{2\pi}\right) \tag{54}$$
$$\leq \log C_J(K).$$

This theorem immediately follows from Lemmas 5 and 7, which hold in much simpler forms then the original ones, since any string x^n is in the good set $G_{n,\delta}$ by the fact $\hat{J}(\theta, x^n) = J(\theta)$,

C. Lower and Upper Bounds for Curved Exponential Families

As a more general case than exponential families, we consider curved exponential families. It is defined as a surface in the natural parameter space of an exponential family. If the surface is a hyperplane, the surface forms an exponential family again. Hence, we are interested in cases that the surface is curved. Below, we give a formal definition of curved exponential families.

Definition 3 (Curved Exponential Families): Let $S = \{\bar{p}(\cdot|u) : u \in \mathcal{U}\}$ be a $\bar{d}$-dimensional steep exponential family, where $\bar{p}(x|u) = \exp(u \cdot T(x) - \psi(u))$ denotes the density with respect to a measure $\nu(dx)$ and u its natural parameter. Let Θ be an open subset of $\Re^d$ ($d < \bar{d}$) and $\phi = (\phi_1, \ldots, \phi_{\bar{d}})$ be a function from Θ to $\mathcal{U}$. Assume that ϕ is twice continuously differentiable in Θ and that the Jacobian of ϕ is of rank d for every θ in Θ. Further, we assume that ϕ is an injection. Let $p(\cdot|\theta) = \bar{p}(\cdot|\phi(\theta))$. Then, the family $M = \{p(\cdot|\theta) : \theta \in K\}$ with densities

$$p(x|\theta) = \exp(\phi(\theta) \cdot T(x) - \psi(\phi(\theta)))$$

is referred to as a d-dimensional curved exponential family of densities embedded in the exponential family S.

In the above definition, let η denote the function which maps u to the corresponding expectation parameter of S, that is, $\eta(u) = \int T(x)\bar{p}(x|u)\nu(dx)$.

For the curved exponential family M, the score function is

$$
\begin{aligned}
\frac{\partial \log p(x|\theta)}{\partial \theta_i} &= \frac{\partial \phi(\theta) \cdot T(x)}{\partial \theta_i} - \sum_k \frac{\partial \psi(\phi(\theta))}{\partial u_k} \frac{\partial \phi_k(\theta)}{\partial \theta_i} \\
&= \frac{\partial \phi(\theta) \cdot T(x)}{\partial \theta_i} - \eta \cdot \frac{\partial \phi(\theta)}{\partial \theta_i} \\
&= \frac{\partial \phi(\theta) \cdot (T(x) - \eta)}{\partial \theta_i}.
\end{aligned} \tag{55}
$$

Then, by the chain rule of differentiation, we have

$$
\begin{aligned}
&\frac{\partial^2 \log p(x|\theta)}{\partial \theta_j \partial \theta_i} \\
&= \frac{\partial^2 \phi(\theta) \cdot (T(x) - \eta)}{\partial \theta_j \partial \theta_i} - \sum_{k,l} \frac{\partial \phi_l(\theta)}{\partial \theta_i} \frac{\partial \eta_l}{\partial u_k} \frac{\partial \phi_k(\theta)}{\partial \theta_j} \\
&= \frac{\partial^2 \phi(\theta) \cdot (T(x) - \eta)}{\partial \theta_j \partial \theta_i} - \sum_{k,l} \frac{\partial \phi_l(\theta)}{\partial \theta_i} \bar{J}_{kl}(u) \frac{\partial \phi_k(\theta)}{\partial \theta_j},
\end{aligned} \tag{56}
$$

where $\bar{J}_{kl}(u)$ is the Fisher information matrix of the exponential family S. Taking the expected value, we have

$$
-J_{ij}(\theta) = -\sum_{k,l} \frac{\partial \phi_l(\theta)}{\partial \theta_i} \bar{J}_{kl}(u) \frac{\partial \phi_k(\theta)}{\partial \theta_j}.
$$

By this we have confirmed that the Fisher information exists and is finite. Further, since ϕ is twice continuously differentiable and its Jacobian is of rank d for every $\theta \in \Theta$, the Fisher information is continuous and is positive definite for all $\theta \in \Theta$. (That is, Assumption 2 holds.) Further (56) yields

$$
\hat{J}_{ij}(\theta, x) = J_{ij}(\theta) - \frac{\partial^2 \phi(\theta)}{\partial \theta_i \partial \theta_j} \cdot (T(x) - \eta(\phi(\theta))). \tag{57}
$$

This shows how the empirical Fisher information differs from the Fisher information in curved exponential families. Since $T(x)$ and its covariance matrix with respect to $\bar{p}(\cdot|u)$ for all $u \in \mathcal{U}$ is finite, Assumptions 1 and 1' hold. Assumption 2' also holds because of (55). Further, note that the maximum likelihood estimator for curved exponential families is consistent uniformly for $\theta \in K$, when K is compact and interior to Θ. This can be easily confirmed by Lemma 9 (the large deviation theorem). Hence, Lemmas 1 and 2 and Theorems 1 and 3 work for curved exponential families and we have the validity of the desired lower bound.

Next, we consider the upper bound. By (57), we have

$$
\hat{J}_{ij}(\theta, x^n) = J_{ij}(\theta) - \frac{\partial^2 \phi(\theta)}{\partial \theta_i \partial \theta_j} \cdot (\bar{T} - \eta(\phi(\theta))). \tag{58}
$$

Let H^k denote the Hessian of $\phi_k(\theta)$ (the Hessian of the kth component of $\phi(\theta)$). Then we can write

$$
\hat{J}(\theta, x^n) = J(\theta) - \sum_k (\bar{T}_k - \eta_k(\phi(\theta))) H^k(\theta). \tag{59}
$$

By this, we have

$$
V(x^n|\theta) = -\sum_k (\bar{T}_k - \eta_k(\phi(\theta))) J_\theta^{-1/2} H^k(\theta) J_\theta^{-1/2}. \tag{60}
$$

Since the exponential moments of $T(x)$ exist for $u \in \mathcal{U}^\circ$ by definition of the natural parameter space of exponential families, Assumption 7 holds for a certain $b > 0$. Further, note that, if $T(x)$ is bounded, then $\bar{T}(x^n)$ is uniformly bounded for all n. Hence, if $T(x)$ is bounded, the collection of functions $V(x^n|\theta)$, $x^n \in \mathcal{X}^n$, $n = 1, 2, \ldots$ is equicontinuous uniformly for $\theta \in K$. Then, Assumptions 8 to 11 hold. As for the ideal prior, Assumption 9 does hold. That is, we can use Theorem 5, and the general minimax strategy works. On the contrary, if $T(x)$ is not bounded, it is not clear whether the assumptions hold or not.

However, to deal with such curved exponential families with unbounded $T(x)$, we can establish another minimax strategy. In this alternative strategy, to deal with the string x^n whose empirical Fisher information differs from the Fisher information, we mix in small measure component living in the full $\bar{d}$-dimensional exponential family S instead of the fiber bundle of local exponential families.

Let K be a compact set included in Θ° such that $\bar{K} = \bar{K}^\circ$. Let $\mathcal{U}_c$ be a compact set included in $\mathcal{U}^\circ$ such that $\mathcal{U}_c^\circ \supset \{\phi(\theta) : \theta \in \bar{K}\}$. Our asymptotic minimax strategy is established in the following theorem. Note that we do not assume the boundness of $T(x)$ here.

Theorem 6: Let $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ be a d-dimensional curved exponential family of densities defined in Definition 3 and let

$$
m_n(x^n) = (1 - n^{-r}) m_{K,\epsilon,n}(x^n) + n^{-r} \int \bar{p}(x^n|u) w(u) du,
$$

where $w(u)$ is the uniform prior over $\mathcal{U}_c$. Then, the following holds.

$$
\limsup_{n \to \infty} \left(\sup_{x^n \in \mathcal{K}} \log \frac{p(x^n|\hat{\theta})}{m_n(x^n)} - \frac{d}{2} \log \frac{n}{2\pi} \right) \tag{61}
$$
$$
\leq \log C_J(K).
$$

Proof: Define a set of good strings:

$$
G_n = \{x^n : \hat{\theta} \in K \text{ and } |\hat{\eta} - \eta(\phi(\hat{\theta}))| \leq \delta_n\}
$$

(recall that η denotes the expectation parameter and that $\hat{\eta} = \bar{T}$), where we let $\delta_n = n^{-1/2+\gamma}$ $(0 < \gamma < 1/2)$. Define also

$$
G_n^c = \{x^n : \hat{\theta} \in K \text{ and } |\hat{\eta} - \eta((\phi(\hat{\theta}))| > \delta_n\}
$$

We can prove the usual asymptotic bound for regret when $x^n \in G_n$, while for $x^n \in G_n^c$, it turns out that the regret of our strategy becomes negative, that is, the code length becomes shorter than $-\log p(x^n|\hat{\theta})$.

For $x^n \in G_n$, from (59) and the Cauchy-Schwartz inequality, we have for each (i, j),

$$
\begin{aligned}
(\hat{J}_{ij}(\theta, x^n) - J_{ij}(\theta))^2 &\leq |\bar{T} - \eta(\phi(\theta))|^2 \sum_k (H_{ij}^k(\theta))^2 \\
&\leq \delta_n^2 \sum_k (H_{ij}^k(\theta))^2.
\end{aligned}
$$

Since $H^k(\theta)$ is continuous and K is compact, $\sum_k (H_{ij}^k(\theta))^2$ is bounded by a constant $C_{H,K}^2$. Accordingly, for $x^n \in G_n$,

$$\max_{ij} \max_{\theta \in K} |\hat{J}_{ij}(\theta, x^n) - J_{ij}(\theta)| \le C_{H,K}\, \delta_n$$

holds, which implies

$$\max_{\theta \in K} ||\hat{J}(\theta, x^n) - J(\theta)|| \le d\, C_{H,K}\delta_n.$$

Therefore, uniformly over $x^n \in G_n$,

$$\max_{\theta \in K} ||\hat{J}(\theta, x^n) - J(\theta)||_s \le d\, C_{H,K}\delta_n$$

which means, for all unit vector z, we have

$$\max_{\theta \in K} |z^t(\hat{J}(\theta, x^n) - J(\theta))z| \le d\, C_{H,K}\delta_n.$$

Since $J(\theta)$ is continuously differentiable, for $\epsilon > 0$,

$$\max_{\theta' \in B_\epsilon(\theta) \cap K} \max_{\theta \in K} |z^t(\hat{J}_n(\theta')z - z^t J(\theta))z| \le d\, C_{H,K}\delta_n + O(\epsilon).$$

Now since $\delta_n = n^{-1/2+\gamma}$, the right side is $O(\epsilon)$ for $n \ge (1/\epsilon)^{2/(1-2\gamma)}$. Therefore, for such n, we have the following two inequalities holding uniformly for $x^n \in G_n$ and uniformly for all unit vectors z,

$$\max_{\tilde{\theta} \in B_\epsilon(\hat{\theta}) \cap K} |z^t \hat{J}(\tilde{\theta}, x^n)z - z^t J(\hat{\theta})z| \le O(\epsilon)$$

and

$$|z^t \hat{J}(\hat{\theta}, x^n)z - z^t J(\hat{\theta})z| \le O(\epsilon).$$

Since $z^t J(\hat{\theta}))z$ is lower bounded by a positive constant (even when we take a minimum over all unit vectors z), for sufficiently small ϵ, the $z^t(\hat{J}(\hat{\theta}, x^n)z$ is also lower bounded by a positive constant, uniformly over unit vectors z and $x^n \in G_n$. Consequently, for all $n \ge (1/\epsilon)^{2/(1-2\gamma)}$ and all $x^n \in G_n$, we have

$$\max_{\tilde{\theta} \in B_\epsilon(\hat{\theta}) \cap K} \frac{z^t \hat{J}(\tilde{\theta}, x^n)z}{z^t \hat{J}(\hat{\theta}, x^n)z} \le 1 + O(\epsilon).$$

This is the last inequality in Assumption 9 and is sufficient to obtain the same bound for good strings as in Theorem 5. To show that, we have the condition that ϵ must converge to 0 slower than $1/\sqrt{n}$. Now, we face the new condition $\epsilon \ge \delta_n = n^{-1/2+\gamma}$, which does not contradict the previous condition.

Next we consider $x^n \in G_n^c$. Let $N_\delta(u)$ denote the δ neighborhood of u and let $\mathcal{U}_c'$ a compact set included in $\mathcal{U}_c^\circ$. Now we will prove that the following holds for a certain $\kappa > 0$.

$$\forall n \in \mathbb{N}, \forall x^n \in G_n^c, \exists \tilde{u} \in \mathcal{U}_c',$$
$$\inf_{u \in N_{\kappa\delta_n}(\tilde{u})} \frac{1}{n} \log \frac{\bar{p}(x^n|u)}{p(x^n|\hat{\theta})} \ge C_2 \delta_n^2. \qquad (62)$$

Indeed, by Taylor expansion with respect to u at $\phi(\hat{\theta})$, we have

$$\frac{1}{n} \log \frac{\bar{p}(x^n|u)}{p(x^n|\hat{\theta})}$$
$$= \frac{1}{n} \log \frac{\bar{p}(x^n|u)}{\bar{p}(x^n|\phi(\hat{\theta}))}$$
$$= (u - \phi(\hat{\theta})) \cdot \bar{T} - \psi(u) + \psi(\phi(\hat{\theta}))$$
$$= (\bar{T} - \eta(\phi(\hat{\theta}))) \cdot (u - \phi(\hat{\theta}))$$
$$\quad - (1/2)(u - \phi(\hat{\theta}))^t \bar{J}(u')(u - \phi(\hat{\theta})),$$

where $u' = \xi u + (1 - \xi)\phi(\hat{\theta})$ for some $\xi \in [0, 1]$. Let

$$\tilde{u} = \phi(\hat{\theta}) + 2\kappa\delta_n \frac{\bar{T} - \phi(\hat{\theta}))}{|\bar{T} - \phi(\hat{\theta})|}.$$

Then, for all $u \in N_{\kappa\delta_n}(\tilde{u})$, we have $\kappa\delta_n \le |u - \phi(\hat{\theta})| \le 3\kappa\delta_n$. Hence, we have

$$\inf_{u \in N_{\kappa\delta_n}(\tilde{u})} \frac{1}{n} \log \frac{\bar{p}(x^n|u)}{p(x^n|\hat{\theta})} \ge \kappa\delta_n^2 - 9\kappa^2\lambda^*\delta_n^2/2$$
$$= \kappa(\delta_n^2 - 9\kappa\lambda^*\delta_n^2/2),$$

where λ^* denotes the maximum value among the maximum eigenvalues of $\bar{J}(u)$ for $u \in \mathcal{U}_c$. Let $\kappa = 1/9\lambda^*$, then the above implies (62). Hence we have for all $n \in \mathbb{N}$ and for all $x^n \in G_n^c$,

$$\frac{n^{-r} \int \bar{p}(x^n|u)w(u)du}{p(x^n|\hat{\theta})}$$
$$\ge \int_{N_{\kappa\delta_n}(\tilde{u})} \frac{n^{-r}\bar{p}(x^n|u)w(u)}{p(x^n|\hat{\theta})}du$$
$$\ge C_3 n^{-r} \int_{N_{\kappa\delta_n}(\tilde{u})} \exp(C_2 n\delta_n^2)du$$
$$\ge C_3 n^{-r}(\kappa\delta_n)^{\bar{d}} \exp(C_2 n\delta_n^2)$$
$$\ge C_4 n^{-r} n^{-\bar{d}/4} \exp(C_2 n^{2\gamma}).$$

This means that the regret of m_n is negative for all large n and can be ignored. *This completes the proof.*

Lastly we give a comment about the case of stochastic processes. Though Theorem 6 above is for the i.i.d. case only, it can be extended to the case of Markov sources. It is known that some parametric model of Markov sources approaches an exponential family as n goes to infinity, in the sense that the difference between Fisher and empirical Fisher information matrices converges to zero as n goes to infinity. We call the model of stochastic processes with that condition as an asymptotic exponential family. In particular, the Markov model defined by a context tree [48] (a tree model) is interesting. When the context tree satisfies a certain condition, the model is called a finite state machine X (FSMX) model [49]. It is shown that the tree model is an asymptotic exponential family if and only if the model is an FSMX model [45].

As a special case, consider the model of kth order Markov chains, which is an example of an FSMX model. For this case, it has been shown that the Jeffreys mixture asymptotically achieves the minimax regret [44]. Note that any tree model can be regarded as a sub family (a surface) of the model of a certain order Markov chains. It means that a tree model can be regarded as a curved exponential family embedded in an asymptotic exponential family. It suggests that we can establish a similar theorem as Theorem 6 concerning tree models. In fact, such a theorem is in [42].

D. Lower and Upper Bounds for Models with Hidden Variables

The contents of this section was presented in part in 2014 IEEE International Symposium on Information Theory [41].

First we introduce the notion of models with hidden variables below, where $(\mathcal{Y}, \mathcal{B}_y, \nu_y)$ is a measurable space.

Definition 4 (Model with Hidden Variables): Let $q(y|\theta)$ be a density of a d-dimensional exponential family over $\mathcal{Y}$ with natural parameter θ. Define a class S of probability density functions over $\mathcal{X}$ by

$$S = \left\{ p(x|\theta) = \int \kappa(x|y)q(y|\theta)\nu_y(dy) \; : \; \theta \in \Theta \right\}, \quad (63)$$

where $\kappa(x|y)$ is a fixed conditional probability density function of x given y. Then, S is called a model with a hidden variable y.

For the above definition, we will assume that the Fisher information is positive definite over Θ°. (Assumption 2.) Whether this holds or not depends on the property of $\kappa(x|y)$. Later we will see that if x and y are not independent random variables with respect to the joint density $\kappa(y|x)q(y|\theta)$, then $J(\theta)$ is positive definite. This is reasonable because we can draw information about θ from x, if x and y are not independent.

If $q(y|\theta)$ is the multinomial model over the finite set $\mathcal{Y} = \{0, 1, \ldots, d\}$, then $p(x|\theta)$ is in the following form:

$$p(x|\theta) = \sum_{y=1}^{d} \theta_y \kappa(x|y) + \left(1 - \sum_{y=1}^{d} \theta_y\right)\kappa(x|0). \quad (64)$$

Here, θ denotes a d-dimensional vector $(\theta_1, \ldots, \theta_d) \in \Re^d$ satisfying $0 \leq \sum_{i=1}^{d} \theta_i \leq 1$ and $\theta_i \geq 0$ for all i. This model is called a mixture family in information geometry [2], [3]. The mixture family is the dual object of the exponential family in information geometry [2], [3]. Note that this mixture family is a simpler object than the more general class of mixture models such as Gaussian mixtures, binomial mixtures, etc., which are used in many settings.

Typically, models with a hidden variable are examples of non exponential families. But note that if it is the mixture family and if $\mathcal{X}$ is a finite subset of an Euclidean space and each $\kappa(\cdot|y)$ is a point mass at an $x = x_y$ in $\mathcal{X}$, then it is also an exponential family. That is the case of the multinomial Bernoulli model.

In this section we first argue the general case of models with hidden variables, assuming $T(y)$ is bounded, for which Theorem 5 gives the minimax solution for the set of strings $\mathcal{K} = \mathcal{X}^n(K)$ for a compact K interior to Θ.

After that, for the mixture family case, we show a minimax strategy for the problem without restriction to the set of strings $\mathcal{X}^n$. The shown strategy is different from the one used in Theorem 1.

1) Lower Bound for Models with Hidden Variables: Here, we consider the lower bound. We examine Assumptions for Theorems 1 and 2. Here, we assume θ in $q(y|\theta)$ is the natural parameter of exponential family $q(y|\theta)$:

$$q(y|\theta) = \exp(\theta \cdot T(y) - \psi(\theta)).$$

Below, let η denote the expectation parameter: $\eta = \int T(y)q(y|\theta)\nu_y(dy)$ and let $G(\theta)$ denote the Fisher information matrix for $q(y|\theta)$ with respect to θ.

We will show some nice properties of the score function, the Fisher information, and the empirical Fisher information. Let $\mathrm{Cov}_\theta[\bar{T}(Y^n)|x^n]$ denote the covariance matrix of $\sqrt{n}\bar{T}(y^n)$ with respect to the conditional probability distribution $p(y^n|x^n, \theta)$ with a fixed x^n, where $\bar{T} = \sum_{t=1}^{n} T(y_t)/n$. Further, define

$$\mathrm{Cov}_\theta[T(Y)|X] = \int \mathrm{Cov}_\theta[\bar{T}(Y)|x]p(x|\theta)\nu(dx).$$

Then, we have the following lemma.

Lemma 12: For the model with hidden variable (63), the following equalities hold.

$$\frac{1}{n}\frac{\partial \log p(x^n|\theta)}{\partial \theta_i} = \tilde{t}_i - \eta_i, \quad (65)$$

$$\hat{J}(\theta, x^n) = G(\theta) - \mathrm{Cov}_\theta[\bar{T}(Y^n)|x^n], \quad (66)$$

$$J(\theta) = G(\theta) - \mathrm{Cov}_\theta[T(Y)|X] \quad (67)$$

and

$$\frac{\partial \hat{J}_{ij}(\theta, x^n)}{\partial \theta_k} = \frac{\partial G_{ij}(\theta)}{\partial \theta_k} - \sqrt{n}\tilde{w}_{ijk}(\theta, x^n),$$

where

$$\tilde{t}_i = \tilde{t}_i(\theta, x^n) = \int \bar{T}(y^n)p(y^n|x^n, \theta)\nu_y(dy^n)$$

and

$$\tilde{w}_{ijk} = \tilde{w}_{ijk}(\theta, x^n)$$
$$= n^{3/2}\int(\bar{T}_i - \tilde{t}_i)(\bar{T}_j - \tilde{t}_j)(\bar{T}_k - \tilde{t}_k)p(y^n|x^n, \theta)\nu_y(dy^n).$$

Proof: Note that

$$\frac{\partial p(x^n|\theta)}{\partial \theta_i} = \int \kappa(x^n|y^n)n(\bar{T}_i - \eta_i)q(y^n|\theta)\nu(dy^n). \quad (68)$$

Hence, we have

$$\frac{\partial \log p(x^n|\theta)}{\partial \theta_i} = \frac{\int \kappa(x^n|y^n)n(\bar{T}_i - \eta_i)q(y^n|\theta)\nu(dy^n)}{p(x^n|\theta)} \quad (69)$$
$$= \int n(\bar{T}_i - \eta_i)p(y^n|x^n, \theta)\nu(dy^n)$$
$$= n(\tilde{t}_i - \eta_i),$$

where $p(y^n|x^n, \theta) = \kappa(x^n|y^n)q(y^n|\theta)/p(x^n|\theta)$. This yields (65). To further differentiate, we need $(\partial/\partial\theta_i)p(y^n|x^n, \theta)$. By the product rule of differentiation

$$\frac{\partial p(y^n|x^n, \theta)}{\partial \theta_i}$$
$$= \frac{\kappa(x^n|y^n)n(\bar{T}_i - \eta_i)q(y^n|\theta)}{p(x^n|\theta)}$$
$$\quad - \frac{\kappa(x^n|y^n)q(y^n|\theta)}{(p(x^n|\theta))^2}\frac{\partial p(x^n|\theta)}{\partial \theta_i}$$
$$= n(\bar{T}_i - \eta_i)p(y^n|x^n, \theta)$$
$$\quad - \frac{\kappa(x^n|y^n)q(y^n|\theta)}{p(x^n|\theta)}\frac{\partial \log p(x^n|\theta)}{\partial \theta_i}$$
$$= n(\bar{T}_i - \eta_i)p(y^n|x^n, \theta)$$
$$\quad - p(y^n|x^n, \theta)\int n(\bar{T}_i - \eta_i)p(y^n|x^n, \theta)\nu(dy^n)$$
$$= n(\bar{T}_i - \eta_i)p(y^n|x^n, \theta) - n(\tilde{t}_i - \eta_i)p(y^n|x^n, \theta)$$
$$= n(\bar{T}_i - \tilde{t}_i)p(y^n|x^n, \theta).$$

Using this formula, we have

$$\frac{\partial^2 \log p(x^n|\theta)}{\partial\theta_j\partial\theta_i}$$

$$= \int \frac{\partial n(\bar{T}_i - \eta_i)}{\partial\theta_j} p(y^n|x^n,\theta)\nu(dy^n)$$

$$+ \int n(\bar{T}_i - \eta_i)\frac{\partial p(y^n|x^n,\theta)}{\partial\theta_j}\nu(dy^n)$$

$$= \int (-nG_{ij})(\theta)p(y^n|x^n,\theta)\nu(dy^n)$$

$$+ \int n(\bar{T}_i - \eta_i)n(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu(dy^n)$$

$$= -nG_{ij}(\theta)$$

$$+ \int n(\bar{T}_i - \eta_i)n(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu(dy^n)$$

$$= -nG_{ij}(\theta)$$

$$+ \int n(\bar{T}_i - \tilde{t}_i)n(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu(dy^n)$$

$$+ \int n(\tilde{t}_i - \eta_i)n(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu(dy^n)$$

$$= -nG_{ij}(\theta)$$

$$+ \int n(\bar{T}_i - \tilde{t}_i)n(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu(dy^n)$$

$$= -nG_{ij}(\theta) + n\tilde{v}_{ij},$$

where $\tilde{v}_{ij} = \tilde{v}_{ij}(\theta, x^n)$ denotes

$$\int \sqrt{n}(\bar{T}_i - \tilde{t}_i)\sqrt{n}(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu_y(dy^n).$$

That is, we have

$$\frac{\partial^2 \log p(x^n|\theta)}{\partial\theta_j\partial\theta_i} = -nG_{ij}(\theta) + n\tilde{v}_{ij}, \tag{70}$$

Since $\tilde{v}_{ij}$ is ij-entry of $\mathrm{Cov}_\theta[\bar{T}(Y^n)|x^n]$, this is (66), which directly yields (67).

To differentiate (70) again, note the following

$$\frac{\partial\tilde{t}_i}{\partial\theta_j} = \frac{\partial}{\partial\theta_j}\int \bar{T}_i p(y^n|x^n,\theta)\nu(dy^n)$$

$$= \int \bar{T}_i n(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu(dy^n)$$

$$= n\int (\bar{T}_i - \tilde{t}_i)(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu(dy^n)$$

$$= \tilde{v}_{ij}.$$

Then,

$$\frac{\partial^3 \log p(x^n|\theta)}{\partial\theta_k\partial\theta_j\partial\theta_i}$$

$$= -n\frac{\partial G_{ij}(\theta)}{\partial\theta_k}$$

$$+ \int \frac{\partial n(\bar{T}_i - \tilde{t}_i)}{\partial\theta_k}n(\bar{T}_j - \tilde{t}_j)p(y^n|x^n,\theta)\nu(dy^n)$$

$$+ \int n(\bar{T}_i - \tilde{t}_i)\frac{\partial n(\bar{T}_j - \tilde{t}_j)}{\partial\theta_k}p(y^n|x^n,\theta)\nu(dy^n)$$

$$+ \int n(\bar{T}_i - \tilde{t}_i)n(\bar{T}_j - \tilde{t}_j)\frac{\partial p(y^n|x^n,\theta)}{\partial\theta_k}\nu(dy^n).$$

Since $\partial\bar{T}/\partial\theta_k = 0$ holds, and since $\partial\tilde{t}_i/\partial\theta_k = \tilde{v}_{ik}$ dose not depend on y^n, the second and third terms in the last expression in the integral yield 0. Hence,

$$\frac{\partial^3 \log p(x^n|\theta)}{\partial\theta_k\partial\theta_j\partial\theta_i}$$

$$= -n\frac{\partial G_{ij}(\theta)}{\partial\theta_k}$$

$$+ \int n(\bar{T}_i - \tilde{t}_i)n(\bar{T}_j - \tilde{t}_j)\frac{\partial p(y^n|x^n,\theta)}{\partial\theta_k}\nu(dy^n).$$

The second term is

$$\int n(\bar{T}_i - \tilde{t}_i)n(\bar{T}_j - \tilde{t}_j)n(\bar{T}_k - \tilde{t}_k)p(y^n|x^n,\theta)\nu(dy^n),$$

which equals $n^{3/2}\tilde{w}_{ijk}$. Then, we have

$$\frac{\partial^3 \log p(x^n|\theta)}{\partial\theta_k\partial\theta_j\partial\theta_i} = -n\frac{\partial G_{ij}(\theta)}{\partial\theta_k} + n^{3/2}\tilde{w}_{ijk}. \tag{71}$$

This completes the proof of the lemma.

Recall that $G(\theta)$ equals the covariance of $T(y)$ with respect to $q(y|\theta)$ and note that the expectation of $\mathrm{Cov}_\theta[T(y)|x]$ with respect to $p(x|\theta)$, which we denote by $\mathrm{Cov}_\theta[T(y)|X]$, is not more than $\mathrm{Cov}_\theta[T(y)] = G(\theta)$ in the sense of positive definiteness. Hence, $J(\theta)$ is positive semi-definite. Here, the equality holds only if x and $T(y)$ are independent of each other. To be more precisely, if and only if x and $z^t T(y)$ (the linear combination of each entry of $T(y)$) is independent, $z^t G(\theta)z = z^t\mathrm{Cov}_\theta[T(y)|X]z$ holds. Therefore, we assume that x and $z^t T(y)$ with any non-zero z is not independent for $\theta \in \Theta^\circ$.

Since we assume that $T(y)$ is bounded,

$$E_\theta[\max_{\theta' \in B_\epsilon(\theta)} |\hat{J}_{ij}(\theta, x)|]$$

and

$$E_\theta[\max_{\theta' \in B_\epsilon(\theta)} (\hat{J}_{ij}(\theta, x)|)^2]$$

are finite. This means that Assumptions 1 and 1' hold. By (65),

$$\sup_{n\in\mathbb{N}} \sup_{x^n\in\mathcal{X}^n} \sup_{\theta\in K} \left|\frac{1}{n}\frac{\partial \log p(x^n|\theta)}{\partial\theta_i}\right|$$

is finite for any compact set $K \subset \Theta^\circ$. That is, the log likelihood function is equicontinuous for all $\theta \in K$, n, and x^n. Hence, noting positive definiteness of $J(\theta)$, we can prove the maximum likelihood estimator is uniformly consistent for all $\theta \in K$. In conclusion, all the assumptions for Theorems 1 and 2 are satisfied and we have demanded lower bound results.

2) Upper Bounds for Models with Hidden Variables: Note that any order's moments of $\bar{T}(Y^n)$ exist at $p(y^n|x^n,\theta)$. (This is true even if $T(x)$ is not bounded.) Hence, by (66) of Lemma 12, Assumption 7 holds for any $b > 0$. Further, we can show that $|\hat{J}(x^n,\theta)|$ and $|\partial\hat{J}(\theta,x^n)/\partial\theta_k|$ is uniformly bounded for all x^n such that $\hat{\theta} \in K$, which means all the assumptions for the upper bound results hold. In fact, by (71), we have

$$\frac{\partial\hat{J}_{ij}(\theta, x^n)}{\partial\theta_k} = \frac{\partial G_{ij}(\theta)}{\partial\theta_k} - n^2\tilde{w}_{ijk}(\theta, x^n).$$

Since the posterior density $p(y^n|x^n, \theta)$ is i.i.d., $n^2\tilde{w}(\theta, x^n)$ equals the third order moment of $T(y)$, which does not depend on n. Recalling we assume $T(y)$ is bounded, $|\partial\hat{J}_{ij}(\theta, x^n)/\partial\theta_k|$ is bounded for $\theta \in K$.

Hence, we see that our general minimax strategy works for models with hidden variables assuming that $T(y)$ is bounded and $z^t T(y)$ for any nonzero z is not independent of x for all $\theta \in \Theta$.

Note that $\hat{J}(\theta, x^n)$ is positive definite for all x^n and for all $\theta \in \Theta$, which implies the uniqueness of MLE.

3) Extension to the Whole Strings Set Case for Mixture Families: Next, we consider the mixture family case defined as (64). About the lower bound, we should examine the consistency of maximum likelihood estimator

$$\hat{\theta}(x^n) = \arg\max_{\theta \in \Theta} p(x^n|\theta),$$

which is defined relative to the whole parameter space Θ. For the mixture family, as we see later, log likelihood function is concave. Using that property, we can prove $\hat{\theta}$ is uniformly consistent for $\theta \in \Theta$. Therefore, we can apply Theorem 1 for $K = \Theta$ case. (J.T.: Is this true?)

Now we consider the upper bound for the case that $\mathcal{K} = \mathcal{X}^n$. Define a subset Θ_τ of Θ as

$$\Theta_\tau = \{\theta \in \Theta : \theta_i \geq \tau, \, i = 0, 1, \ldots, d\}.$$

Since Θ_τ for $\tau > 0$ is a compact subset of Θ°, it may be possible to design a minimax strategy for the model $\{p(x|\theta) : \theta \in \Theta_\tau\}$ utilizing the result for the general case. In fact, we can do it as we show later.

Then we argue the situation that τ converges to zero as n goes to infinity. Under that situation we utilize Lemmas 2 and 3. Then we have to control the behavior of $\kappa_J(K)$ and λ_K^* (the maximum of the largest eigenvalue of $J(\theta)$ among $\theta \in K$), since K changes as n increases.

To treat the strings with the maximum likelihood estimate being near the boundary, we employ the technique introduced in [51], which utilizes the Dirichlet(α) prior $w_{(\alpha)}(\theta) \propto \prod_{i=0}^d \theta_i^{-(1-\alpha)}$ with $\alpha < 1/2$. Note that this prior with $\alpha = 1/2$ has the same form as the Jeffreys prior for the multinomial model. When $\alpha < 1/2$ it has higher density than the Jeffreys prior as θ approaches the boundary of Θ.

Our asymptotic minimax strategy is the mixture

$$
\begin{aligned}
m_n(x^n) &= (1 - 2n^{-r})m_\Theta(x^n) \\
&+ n^{-r}\int p_e(x^n|\theta, \beta)\bar{w}(\theta, \beta)d\theta d\beta \\
&+ n^{-r}\int p(x^n|\theta)w_{(\alpha)}(\theta)d\theta,
\end{aligned}
\tag{72}
$$

where m_Θ is the Jeffreys mixture over Θ. Here, the first and second terms are for the strings with the MLE being away from the boundary, while the third term works when the MLE approaches the boundary.

We can show the following theorem.

Theorem 7: The strategy m_n defined as (72) with $r < (1/2 - \alpha)(1 - p)$ asymptotically achieves the minimax regret for the mixture family, i.e.

$$\sup_{x^n \in \mathcal{X}^n} \log\frac{p(x^n|\hat{\theta})}{m_n(x^n)} \leq \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(\Theta) + o(1).$$

To prove Theorem 7, we use the following useful inequalities for the model with hidden variables. Here $A \leq B$ for two matrices A and B means that $B - A$ is positive semidefinite.

Lemma 13: Given a data string x^n, let $\hat{\theta}$ denote the MLE for a model with hidden variables $p(x^n|\theta)$ defined by (63). Then, the following holds for all $x^n \in \mathcal{X}^n$.

$$\forall\theta \in \Theta, \quad \frac{1}{n}\log\frac{p(x^n|\hat{\theta})}{p(x^n|\theta)} \leq D(q(\cdot|\hat{\theta})\|q(\cdot|\theta)), \tag{73}$$

$$\forall\theta \in \Theta, \quad \hat{J}(\theta, x^n) \leq G(\theta). \tag{74}$$

In particular, when $q(y|\theta)$ is the multinomial model, the following holds

$$\frac{p(x^n|\hat{\theta})}{p(x^n|\theta)} \leq \exp(nD(q(\cdot|\hat{\theta})\|q(\cdot|\theta))) = \prod_{y\in\mathcal{Y}}\frac{\hat{\eta}_y^{n\hat{\eta}_y}}{\eta_y^{n\hat{\eta}_y}}, \tag{75}$$

where $\eta_y = q(y|\theta)$ and $\hat{\eta}_y = q(y|\hat{\theta})$.

The proof is in Appdendix L.

Note that

$$
\begin{aligned}
\hat{J}_{ij}(\theta, x) &= -\frac{\partial^2 \log p(x|\theta)}{\partial\theta_i\partial\theta_j} \\
&= \frac{(p_i(x) - p_0(x))(p_j(x) - p_0(x))}{(p(x|\theta))^2}
\end{aligned}
$$

holds, where we have

$$\left|\frac{p_i(x) - p_0(x)}{p(x|\theta)}\right| \leq \frac{p_i(x) + p_0(x)}{\sum_{i=0}^d \theta_i p_i(x)} \leq \frac{1}{\theta_i} + \frac{1}{\theta_0}.$$

The last expression is not more than $2\max_{\theta\in K}\max_i \theta_i^{-1}$, which is finite. Further, $\partial V(x^n|\theta)/\partial\theta_i$ is similarly bounded, so $V(x^n|\theta)$ is equicontinuous for all $x^n : \hat{\theta} \in K$. This implies the general mixture strategy works for mixture families with a compact K interior to Θ.

Next, we give a discussion on the case with the set of strings $\{x^n : \hat{\theta}(x^n) \in \Theta^\circ\}$. We are to examin Assumptions 7 and 8 for $\theta \in \Theta_\tau$. Note that, for $z \in \Re^d$,

$$
\begin{aligned}
z^t\hat{J}(\theta, x)z &= \frac{\sum_{ij} z_i z_j (p_i(x) - p_0(x))(p_j(x) - p_0(x))}{p(x|\theta)^2} \\
&= \frac{(\sum_i z_i(p_i(x) - p_0(x)))^2}{p(x|\theta)^2} \geq 0,
\end{aligned}
$$

Hence we have

$$z^t\hat{J}(\theta, x^n)z \geq 0$$

for all $z \in \Re^d$ and for all $\theta \in \Theta$. Since

$$\frac{\partial\hat{J}_{ij}(\theta, x)}{\partial\theta_k} = \frac{-2(p_k(x) - p_0(x))}{(p(x|\theta))}\hat{J}_{ij}(\theta, x),$$

we have

$$\frac{\partial z^t\hat{J}(\theta, x)z}{\partial\theta_k} = \frac{-2(p_k(x) - p_0(x))}{(p(x|\theta))}z^t\hat{J}(\theta)z, \tag{76}$$

which yields

$$
\begin{aligned}
&\frac{\partial z^t\hat{J}(\theta, x^n)z}{\partial\theta_k} \\
&= \frac{1}{n}\sum_t \frac{-2(p_k(x_t) - p_0(x_t))}{(p(x_t|\theta))}z^t\hat{J}(\theta, x_t)z.
\end{aligned}
\tag{77}
$$

Hence we have for $\theta \in \Theta_\tau$,

$$\frac{\partial z^t \hat{J}(\theta, x^n)z}{\partial \theta_k} \leq \frac{1}{n}\sum_t \frac{2z^t \hat{J}(\theta, x_t)z}{\tau} = \frac{2z^t \hat{J}(\theta, x^n)z}{\tau}$$

By this inequality and the fact $z^t \hat{J}(\tilde{\theta}, x^n)z \geq 0$, we have the following inequalities.

$$e^{-2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} z^t \hat{J}(\hat{\theta}, x^n)z \leq z^t \hat{J}(\tilde{\theta}, x^n)z \qquad (78)$$
$$\leq e^{2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} z^t \hat{J}(\hat{\theta}, x^n)z.$$

By this, we can see that $\bar{h}_K$ for $K = \Theta_\tau$ is not more than $e^{2\sqrt{d}}$, as long as $\epsilon \leq \tau$. Further, if $|\tilde{\theta} - \hat{\theta}| \leq \tau/2\sqrt{d}$,

$$z^t \hat{J}(\tilde{\theta}, x^n)z \leq z^t \hat{J}(\hat{\theta}, x^n)z\left(1 + \frac{e\tau}{2\sqrt{d}}\right)$$

holds for all $z \in \Re^d \setminus \{0\}$. Note that

$$\max_{\tilde{\theta} \in B_\epsilon(\hat{\theta})} |\tilde{\theta} - \hat{\theta}| = \frac{\epsilon}{\lambda_{\min}},$$

where $\lambda_{\min}$ is the minimum value among $\theta \in \Theta_\tau$ of the smallest eigenvalue of $J(\theta)$. Hereafter, we assume $\epsilon \leq \tau\lambda_{\min}/2\sqrt{d}$.

When $\epsilon \leq \tau\lambda_{\min}/2\sqrt{d}$ is satisfied,

$$z^t \hat{J}(\tilde{\theta}, x^n)z \leq z^t \hat{J}(\hat{\theta}, x^n)z\left(1 + \frac{e\epsilon}{\lambda_{\min}}\right)$$

holds for all $\tilde{\theta} \in B_\epsilon(\hat{\theta})$. This implies Assumption 8 holds with $\kappa_J = e/\lambda_{\min}$ for $\epsilon \leq \tau\lambda_{\min}/2\sqrt{d}$. Note that the argument here works even if τ and ϵ depend on n.

Next we examine Assumption 9. Concerning this matter, we have the following lemma.

Lemma 14: Assume that $||V(x^n|\hat{\theta})||_s \geq \delta$, $\delta \geq 8\sqrt{d}\epsilon/\tau$, $4\sqrt{d}\epsilon/\tau \leq 1/2$, and $\delta \leq 1/2$. Then, we have for $\tilde{\theta} \in B_\epsilon(\hat{\theta})$

$$||T(x^n|\tilde{\theta})||_s \geq ||T(x^n|\hat{\theta})||_s/4.$$

Proof: Note that there exists a unit vector $\bar{z} \in \Re^d$ such that $|\bar{z}^t V(x^n|\hat{\theta})\bar{z}| = ||V(x^n|\hat{\theta})||_s$. Here we have two cases; i) $\bar{z}^t V(x^n|\hat{\theta})\bar{z} = ||V(x^n|\hat{\theta})||_s$ and ii) $-\bar{z}^t V(x^n|\hat{\theta})\bar{z} = ||V(x^n|\hat{\theta})||_s$.

First consider the case i), for which we have

$$||V(x^n|\hat{\theta})||_s = \bar{z}^t J(\hat{\theta})^{-1/2} \hat{J}(\hat{\theta}) J(\hat{\theta})^{-1/2}\bar{z} - 1$$
$$= \frac{\bar{z}^t J(\hat{\theta})^{-1/2} \hat{J}(\hat{\theta}) J(\hat{\theta})^{-1/2}\bar{z}}{\bar{z}^t\bar{z}} - 1$$
$$= \frac{\tilde{z}^t \hat{J}(\hat{\theta})\tilde{z}}{\tilde{z}^t J(\hat{\theta})\tilde{z}} - 1,$$

where $\tilde{z}$ denotes $J(\hat{\theta})^{-1/2}\bar{z}$. Hence,

$$\frac{\tilde{z}^t \hat{J}(\hat{\theta})\tilde{z}}{\tilde{z}^t J(\hat{\theta})\tilde{z}} = 1 + ||V(x^n|\hat{\theta})||_s \qquad (79)$$

holds. For the numerator in the left side, from (78) we have

$$\tilde{z}^t \hat{J}(\tilde{\theta})\tilde{z} \geq e^{-2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau}\tilde{z}^t \hat{J}(\hat{\theta})\tilde{z}.$$

As for the denominator, similarly as (78) we can show by (76)

$$e^{-2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} \leq \frac{\tilde{z}^t J(\tilde{\theta})\tilde{z}}{\tilde{z}^t J(\hat{\theta})\tilde{z}} \leq e^{2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} \qquad (80)$$

for $\tilde{\theta} \in \Theta_\tau$. Then, we have for all $\tilde{\theta} \in B_\epsilon(\hat{\theta}) \cap \Theta_\tau$,

$$\frac{\tilde{z}^t \hat{J}(\tilde{\theta})\tilde{z}}{\tilde{z}^t J(\tilde{\theta})\tilde{z}} \geq (1 + ||V(x^n|\hat{\theta})||_s)e^{-4\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau}$$
$$\geq (1 + ||V(x^n|\hat{\theta})||_s)e^{-\delta/2}$$
$$\geq (1 + ||V(x^n|\hat{\theta})||_s)(1 - \delta/2)$$
$$= 1 + ||V(x^n|\hat{\theta})||_s - \delta/2 - \delta||V(x^n|\hat{\theta})||_s/2$$
$$\geq 1 + ||V(x^n|\hat{\theta})||_s/2 - \delta||V(x^n|\hat{\theta})||_s/2$$
$$= 1 + (1 - \delta)||V(x^n|\hat{\theta})||_s/2$$
$$= 1 + ||V(x^n|\hat{\theta})||_s/4.$$

Letting $\xi = J(\tilde{\theta})^{1/2}\tilde{z}/|J(\tilde{\theta})^{1/2}\tilde{z}|$, that is, $\tilde{z} = |J(\tilde{\theta})^{1/2}\tilde{z}|J(\tilde{\theta})^{-1/2}\xi$, we have

$$\frac{\tilde{z}^t \hat{J}(\tilde{\theta})\tilde{z}}{\tilde{z}^t J(\tilde{\theta})\tilde{z}} = \frac{\xi^t J(\tilde{\theta})^{-1/2} \hat{J}(\tilde{\theta}) J(\tilde{\theta})^{-1/2}\xi}{\xi^t\xi}$$
$$= \xi^t J(\tilde{\theta})^{-1/2} \hat{J}(\tilde{\theta}) J(\tilde{\theta})^{-1/2}\xi$$
$$\leq 1 + ||V(x^n|\tilde{\theta})||_s$$

Hence, we have $||T(x^n|\tilde{\theta})||_s \geq ||T(x^n|\hat{\theta})||_s/4$.

For the case ii), we can show the same conclusion. *This completes the proof.*

Remark: This means that Assumption 10 holds with $\zeta = 1/4$.

To see Assumption 7 holds is easy because of (78).

Finally, we evaluate $a = \min\{1/2\lambda^*, 1\}$ in Lemma 10. Since $|\hat{J}_{ij}(\theta)|$ is bounded by $4/\tau^2$ for $\theta \in \Theta_\tau$ and since the smallest eigenvalue of $J(\theta)$ is lower bounded by a certain positive constant, λ^* is of order τ^{-4}. Hence, a is lower bounded by τ^4 times a certain constant.

Now, we will give the proof of Theorem 7.

Proof of Theorem 7: We divide the strings to the following three categories:

$$A_n = \{x^n : \hat{\theta} \in \Theta_\tau \text{ and } ||V(x^n|\hat{\theta})||_s \leq \delta\},$$
$$B_n = \{x^n : \hat{\theta} \in \Theta_\tau \text{ and } ||V(x^n|\hat{\theta})||_s > \delta\},$$
$$C_n = \{x^n : \hat{\theta} \notin \Theta_\tau\}.$$

Set $\tau = n^{-(1-p)}$, $\delta = n^{-1/2+\gamma}$, and $\epsilon = n^{-1/2+\iota}$, assuming $1/2 + \iota < p < 1$, $0 < \gamma < 1/2$, and $1 - p < \gamma/2$, that is, $2(1-p) < \gamma < 1/2$. Note that $\epsilon/\tau = n^{-1/2+\iota+1-p} = n^{-(p-1/2-\iota)} \to 0$ as n goes to infinity under this setting. Let $\tilde{\epsilon} = \epsilon/\tau$ and assume n is large enough so that $\tilde{\epsilon} = \epsilon/\tau \leq \lambda_{\min}/2\sqrt{d}$ holds. Let $C_1 = 2\sqrt{d}/\lambda_{\min}$. Then, $\tilde{\epsilon} \leq 1/C_1$.

When x^n belongs to A_n, the Laplace approximation for the integral with respect to θ is correct enough. In fact by (78),

$$z^t \hat{J}(\tilde{\theta}, x^n)z \leq e^{2\sqrt{d}|\tilde{\theta}-\hat{\theta}|/\tau} z^t \hat{J}(\hat{\theta}, x^n)z$$
$$\leq e^{2\sqrt{d}\epsilon/\lambda_{\min}\tau} z^t \hat{J}(\hat{\theta}, x^n)z$$
$$= e^{C_1\tilde{\epsilon}} z^t \hat{J}(\hat{\theta}, x^n)z$$

holds for all $\tilde{\theta} \in B_\epsilon(\hat{\theta})$. Hence for $x^n \in A_n$, similarly as the

proof of Lemma 4, we have

$$\log \frac{p(x^n|\hat{\theta})}{m_\Theta(x^n)} \le \frac{d}{2} \log \frac{n}{2\pi} + \log \frac{|J(\hat{\theta})|^{1/2}}{w_J(\hat{\theta})\Phi(U_{\Theta,\epsilon,n}(\hat{\theta}))}$$
$$+ \log \frac{e^{dC_1\bar{\epsilon}/2}(1+\delta)^{d/2}}{1-\eta_\epsilon}.$$

Here, since $B_\epsilon(\hat{\theta}) \subset \Theta$ holds when n is large, $\Phi(U_{\Theta,\epsilon,n}(\hat{\theta}))$ converges to 1 as n goes to infinity, the above means

$$\log \frac{p(x^n|\hat{\theta})}{(1-2n^{-r})m_\Theta(x^n)} \le \frac{d}{2} \log \frac{n}{2\pi} + \log C_J(\Theta) + o(1).$$

Next, we consider the case that $x^n \in B_n$. By Lemma 14, Assumption 9 holds with $\zeta = 1/4$. Then, similarly as the proof of Lemma 10, we can show (47) with $\bar{m}(x^n) = \int p_e(x^n|\theta,\beta)\bar{w}(\theta,\beta)d\theta d\beta$, $\zeta = 1/4$ and $\bar{h}_K = \bar{h}_\Theta < \infty$. Hence, we can show the same inequality as (52) for $x^n \in B_n$. In this case, with a certain $C > 0$, we have

$$an^{2\gamma} \ge C\tau^4 n^{2\gamma} = Cn^{2\gamma - 4(1-p)},$$

which is polynomially large, since we set $1 - p < \gamma/2$. This makes the regret of m_n go to negative infinity as n goes to infinity.

Finally, we consider the case that $x^n \in C_n$, in which there exists y such that $\hat{\theta}_y \le \tau = n^{-(1-p)}$. Then, we can use the help from the third term in (72) as follows. Note that the following holds from (75).

$$\frac{\int p(x^n|\theta)w_{(\alpha)}(\theta)d\theta}{p(x^n|\hat{\theta})} \ge \frac{\int \prod_{i=0}^d \theta_i^{n\hat{\theta}_i} w_{(\alpha)}(\theta)d\theta}{\prod_{i=0}^d \hat{\theta}_i^{n\hat{\theta}_i}}.$$

Then, by Lemma 4 of [51], the right side is not less than

$$\frac{1}{Rn^{d/2 - (1/2-\alpha)(1-p)}}$$

for $x^n \in C_n$, where R is a constant depending only on d. Let r be smaller than $(1/2-\alpha)(1-p)$, then for large n the third term of the right side of (72) is larger than $(2\pi)^{d/2}/(C_J(\Theta)n^{d/2})$. This provides an upper bound on regret of m_n smaller than the maximin value for all $x^n \in C_n$. *This completes the proof.*

E. Mixture with Fixed Weights

We consider the following model.

$$p(x|\theta) = 0.5g_0(x) + 0.5g_1(x|\theta),$$

where $q(x|\theta)$ is an exponential family with the natural parameter θ: $g_1(x|\theta) = \exp(\theta \cdot T(x)) - \psi(\theta)$, and $g_0(x)$ is a certain fixed probability density function. This is another example of non-exponential families. Let η denote the expectation parameter for $g_1(x|\theta)$. For this case, we assume $T(x)$ is bounded again.

We have

$$\frac{\partial \log p(x|\theta)}{\partial \theta_i} = \frac{0.5(T_i - \eta_i)g_1(x|\theta)}{p(x|\theta)}$$
$$= (T_i - \eta_i)r(x|\theta),$$

where we have defined

$$r(x|\theta) = \frac{0.5g_1(x|\theta)}{p(x|\theta)}.$$

Note that $0 \le r(x|\theta) \le 1$ and

$$\int r(x|\theta)p(x|\theta)\nu(dx) = 0.5$$

hold. The score function for x^n is

$$\frac{\partial \log p(x^n|\theta)}{\partial \theta_i} = \sum_{t=1}^n (T_i(x_t) - \eta_i)r(x_t|\theta),$$

We have

$$\frac{\partial r(x|\theta)}{\partial \theta_i} = \frac{0.5(T_i - \eta_i)g_1(x|\theta)}{p(x|\theta)}$$
$$- \frac{0.5g_1(x|\theta)0.5(T_i(x) - \eta_i)g_1(x|\theta)}{(p(x|\theta))^2}$$
$$= (T_i - \eta_i)(r(x|\theta) - r(x|\theta)^2)$$
$$= (T_i - \eta_i)r(x|\theta)(1 - r(x|\theta))$$
$$= (T_i - \eta_i)q(x|\theta).$$

where $q(x|\theta) = r(x|\theta)(1 - r(x|\theta))$. Hence,

$$\frac{\partial^2 \log p(x|\theta)}{\partial \theta_j \partial \theta_i} = -G_{ij}(\theta)r(x|\theta)$$
$$+ (T_i - \eta_i)(T_j - \eta_j)q(x|\theta),$$

by which, we have

$$\hat{J}(\theta, x^n) = G_{ij}(\theta)\bar{r} \tag{81}$$
$$- \frac{1}{n}\sum_{t=1}^n (T_i(x_t) - \eta_i)(T_j(x_t) - \eta_j)q(x_t|\theta), \tag{82}$$

where $\bar{r} = (1/n)\sum_{t=1}^n r(x_t|\theta)$. We also have

$$J_{ij}(\theta) = \frac{G_{ij}(\theta)}{2} - \int (T_i - \eta_i)(T_j - \eta_j)q(x|\theta)p(x|\theta)\nu(dx).$$

Hence, $\hat{J}(\theta, x^n) - J(\theta)$ is bounded for any $\theta \in \Theta^\circ$, which means that Assumption 7 holds for any $b > 0$. Further, since $\int (T_i - \eta_i)(T_j - \eta_j)p(x|\theta)\nu(dx) = G_{ij}(\theta)$ and $0 \le q(x|\theta) \le 1/4$, $J(\theta)$ is positive definite for all θ.

Now we have

$$\frac{\partial q(x|\theta)}{\partial \theta_i} = (T_i - \eta_i)q(x|\theta)(1 - r(x|\theta))$$
$$- (T_i - \eta_i)r(x|\theta)q(x|\theta)$$
$$= (T_i - \eta_i)q(x|\theta)(1 - 2r(x|\theta)).$$

Using the above, we have

$$\frac{\partial^3 \log p(x|\theta)}{\partial \theta_k \partial \theta_j \partial \theta_i}$$
$$= -\frac{\partial G_{ij}(\theta)}{\partial \theta_k}r(x|\theta) - G_{ij}(\theta)(T_k - \eta_k)q(x|\theta)$$
$$- G_{ik}(\theta)(T_j - \eta_j)q(x|\theta)$$
$$- (T_i - \eta_i)G_{jk}(\theta)q(x|\theta)$$
$$+ (T_i - \eta_i)(T_j - \eta_j)(T_k - \eta_k)q(x|\theta)(1 - 2r(x|\theta)).$$

Hence, noting $T(x)$ is assumed to be bounded, $\hat{J}(\theta, x^n)$ is equicontinuous in $\theta \in \Theta$ for all n and for all x^n. Therefore, we can apply Theorem 5 to show the asymptotic minimax strategy.

F. Contaminated Gaussian Location Families

In this subsection, we treat a simple case of contamination Gaussian location families, which is a basic model for robust estimation.

Let us define the model formally below. Let

$$p(x|\theta) = (1-\nu)g_0(x|\theta) + \nu g_1(x|\theta), \tag{83}$$

where

$$g_0(x|\theta) = N(\theta, I),$$
$$g_1(x|\theta) = N(\theta, s^2 I)$$

are Gaussian densities over $\mathbb{R}^d$, I the unit matrix of order d, ν a small positive number, and s a number larger than 1. Here, $(1-\nu)g_0(x|\theta)$ is the main target for estimation and $\nu g_1(x|\theta)$ represents a noise distribution. This $p(x|\theta)$ is an example of non-exponential families, In [34], MDL estimation based on two-stage codes for this model is discussed.

Note that our setting that ν and the forms of g_0 and g_1 are known is the simplest one in robust statistics. See [29] for more general settings.

Now we evaluate the score function and the empirical Fisher information. Let ∂_i denote $\partial/\partial\theta_i$ and $r = r(x|\theta) = (1-\nu)g_0(x|\theta)/p(x|\theta)$. We have

$$
\begin{aligned}
\partial_i \log p(x|\theta) &= \frac{\partial_i(1-\nu)g_0(x|\theta) + \partial_i \nu g_1(x|\theta)}{p(x|\theta)} \\
&= -r(\theta_i - x_i) - \frac{(1-r)(\theta_i - x_i)}{s^2} \\
&= -(\theta_i - x_i)w,
\end{aligned}
$$

where

$$w = w(x|\theta) = r(x|\theta) + \frac{1 - r(x|\theta)}{s^2} = \left(1 - \frac{1}{s^2}\right)r(x|\theta) + \frac{1}{s^2}.$$

Hence,

$$
\begin{aligned}
\partial_j \partial_i \log p(x|\theta) &= -I_{ji}w - (\theta_i - x_i)\partial_j w \\
&= -I_{ij}w - (\theta_i - x_i)\left(1 - \frac{1}{s^2}\right)\partial_j r.
\end{aligned}
$$

For the last term, we have

$$
\begin{aligned}
\partial_j r &= -\frac{\partial_j p(x|\theta)}{p(x|\theta)^2}(1-\nu)g_0(x|\theta) + \frac{\partial_j(1-\nu)g_0(x|\theta)}{p(x|\theta)} \\
&= -r(x|\theta)\partial_j \log p(x|\theta) - r(x|\theta)(\theta_j - x_j) \\
&= r(x|\theta)((\theta_j - x_j)w - (\theta_j - x_j)) \\
&= r(x|\theta)(\theta_j - x_j)(w - 1) \\
&= r(x|\theta)(\theta_j - x_j)(1 - r(x|\theta))\left(\frac{1}{s^2} - 1\right) \\
&= (\theta_j - x_j)q\left(\frac{1}{s^2} - 1\right), \tag{84}
\end{aligned}
$$

where $q = q(x|\theta) = r(x|\theta)(1 - r(x|\theta))$. Hence, we have

$$
\begin{aligned}
&\partial_j \partial_i \log p(x|\theta) \\
&= -wI_{ij} + (\theta_i - x_i)(\theta_j - x_j)\left(1 - \frac{1}{s^2}\right)^2 q,
\end{aligned}
$$

which yields

$$
\begin{aligned}
&\hat{J}_{ij}(\theta, x) \tag{85} \\
&= wI_{ij} - \left(1 - \frac{1}{s^2}\right)^2 (\theta_i - x_i)(\theta_j - x_j)q.
\end{aligned}
$$

By this, we will evaluate the Fisher information. First note that $J(\theta)$ dose not depend on θ. Since $Er = 1$, we have $Ew = 1$. Hence, our task is to evaluate the expectation of the second term. Since

$$r(x|\theta)(1 - r(x|\theta))p(x|\theta) \le \min\{(1-\nu)g_0(x|\theta), \nu g_1(x|\theta)\}$$

holds, we have

$$\int (\theta_i - x_i)(\theta_j - x_j)q\, p(x|\theta)dx \le \min\{1 - \nu, \nu s^2\}I_{ij}.$$

Hence, the following holds.

$$I \ge J(\theta) \ge \max\{\nu, 1 - \nu s^2\}I.$$

The third part is positive definite and equals I if and only if $\nu = 0$. Note that I equals the Fisher information of $g_0(x|\theta) = N(x|\theta, I)$.

The empirical Fisher information $\hat{J}_1(\theta) = \hat{J}(\theta, x)$ has an interesting property. To see it, consider the $d = 1$ case. Assume $x = 0$ without loss of generality. Then, we have

$$\hat{J}_1(\theta) = r + \frac{1 - r}{s^2} - \left(1 - \frac{1}{s^2}\right)^2 \theta^2 q. \tag{86}$$

Assume s is very large and ν is very small. Then, if $|\theta| < 1$ and $r(0|\theta) \doteqdot 1$, then $\hat{J}_1(\theta)$ nearly equals the empirical Fisher information of $N(0|\theta, 1)$, which is 1. On the other hand, let us consider the case that $r(0|\theta) \doteqdot 0$, which is realized when θ^2 is large. In particular, assume that θ^2 is much larger than $\log(s/\nu)$, then

$$
\begin{aligned}
\theta^2 r(0|\theta) &\le \frac{\theta^2(1-\nu)g_0(0|\theta)}{\nu g_1(0|\theta)} \\
&\le \frac{s\theta^2}{\nu}\exp\left(-\frac{\theta^2}{2} + \frac{\theta^2}{2s^2}\right) \doteqdot 0.
\end{aligned}
$$

Hence, $\hat{J}(\theta, 0) \doteqdot 1/s^2$, when θ^2 is much larger than $\log(s/\nu)$.

The above is reasonable observations and not surprising. More interestingly, when θ takes an intermediate value, $\hat{J}_1(\theta)$ can be a large negative value. Let us see it below.

Let c be a solution to the equation $r(0|\theta) = 1/2$, which is

$$\frac{1-\nu}{\sqrt{2\pi}}\exp\left(-\frac{c^2}{2}\right) = \frac{\nu}{\sqrt{2\pi s^2}}\exp\left(-\frac{c^2}{2s^2}\right).$$

This is modified to

$$s^2 = \frac{\nu^2}{(1-\nu)^2}\exp\left(c^2 - \frac{c^2}{s^2}\right).$$

Hence, we have

$$c^2 = \left(1 - \frac{1}{s^2}\right)^{-1}\left(\log s^2 + \log\frac{(1-\nu)^2}{\nu^2}\right),$$

which implies $c^2 \ge \log s^2$, because we assume ν is small. Therefore, the following holds.

$$\hat{J}_1(\pm c) \le \frac{1}{2}\left(1 + \frac{1}{s^2}\right) - \frac{\log s^2}{4}\left(1 - \frac{1}{s^2}\right)^2.$$

The right side is negative when $\log s^2 \ge 4$. Further, its absolute value can be arbitrarily large by letting s be large.

Now, we can show that the MLE is not unique for the string x^2 with $x_1 = c$ and $x_2 = -c$. When s^2 is large, $\hat{J}(0, x^2)$ is negative, which means that $\theta = 0$ is not the maximizer of the

log likelihood, because $-\hat{J}(0) > 0$ is the second derivative of the log likelihood at $\theta = 0$. Hence, the maximum is taken at a certain $\tilde{\theta} \neq 0$. Then by symmetry, $-\tilde{\theta}$ also achieves the maximum.

Next, we consider $\hat{J}(\theta, x^n)$. From (85) we can obtain the following representation.

$$\hat{J}(\theta, x^n) = \left(\left(1 - \frac{1}{s^2}\right)\bar{r} + \frac{1}{s^2}\right)I - \left(1 - \frac{1}{s^2}\right)^2 \tilde{\Sigma}, \quad (87)$$

where we have defined

$$\bar{r} = \frac{1}{n}\sum_{t=1}^{n} r(x_t|\theta),$$

$$\tilde{\Sigma} = \frac{1}{n}\sum_{t=1}^{n} q(x_t|\theta)(x_t - \theta)(x_t - \theta)^t.$$

Now, we concern whether (87) is equicontinuous or not for $x^n \in \mathcal{K}$. By (84), the derivative of the first term with respect to θ_i is proportional to $\theta_i - x_i$, which is unbounded. It may seem to cause a problem, but it is not, since $\theta_i - x_i$ is with a factor $q(x|\theta)$ which is of order $\exp(-C|\theta - x|^2)$. As a result, the derivative of the first term is bounded. This holds for all the terms which contain the factor $\theta_i - x_i$. Hence, the derivative of the empirical Fisher information is bounded, which implies it is equicontinuous for all x^n and for all n as functions of θ.

APPENDIX A
ON ASSUMPTIONS FOR LOWER BOUNDS OF THEOREM 1

Suppose that Assumption 1 holds. This means that $E_\theta \sup_{\theta' \in B_r(\theta)} |\hat{J}_{ij}(\theta', x)|$ is finite for r not more than $r(\theta)$, for all i, j in $\{1, \ldots, d\}$. Now $J(\theta)^{-1/2}\hat{J}_1(\theta', x)J(\theta)^{-1/2}$ is a linear combination of the $\hat{J}_{ij}(\theta', x)$. Consequently,

$$E_\theta \sup_{\theta' \in B_\delta(\theta)} \pm((J(\theta)^{-1/2}\hat{J}(\theta', x)J(\theta)^{-1/2})_{ij} - I_{ij}) \quad (88)$$

is finite for δ not more than $r(\theta)$. Also by Assumption 1, $\hat{J}(\theta', x)$ is continuous in θ' for each x. Accordingly, as δ decreases to 0, the supremum in the above expression convergences monotonically to $\pm((J(\theta)^{-1/2}\hat{J}(\theta, x)J(\theta)^{-1/2})_{ij} - I_{ij})$ which has zero expectation. Then by the monotone convergence theorem, the expected supremum in (88) converges to 0 as $\delta \to 0$ for each θ in Θ°.

The plus-minus $\pm$ means that the indicated property holds with each sign choice. It is the flipped sign case $\sup_{\theta' \in B_\delta(\theta)}(I_{ij} - ((J(\theta)^{-1/2}\hat{J}(\theta', x)J(\theta)^{-1/2})_{ij})$ that especially matters for our development. Nevertheless, it is appropriate to see that it can be handled with either sign. One might be tempted to take an absolute value, but $|(J(\theta)^{-1/2}\hat{J}(\theta, x)J(\theta)^{-1/2})_{ij} - I_{ij}|$ does not have zero expectation.

On K, the assumption that $C_J(K)$ is finite means that the Jeffreys measure is a finite measure on K. Accordingly, by Egorov's Theorem [8], [9], the convergent sequences (6), (7), and (88) are uniformly convergent to 0 as $\delta \to 0$, except for θ in sets of arbitrarily small measure. In particular, for any $\epsilon > 0$, there is a $\delta > 0$ and a good set $G \subset K$, such that for all θ in G the quantities in (6) and (7) are less than ϵ, the expected suprema in (88) are less than $\epsilon/(2d)$ and the Jeffreys

measure of the complement of G, which is $\int_{K \setminus G} |J(\theta)|^{1/2}d\theta$, is less than $\epsilon C_J(K)$.

Next consider $(J(\theta)^{-1/2}\hat{J}(\theta', x^n)J(\theta)^{-1/2})_{ij}$ for a sample of size n. By the definition of the empirical Fisher information, this is

$$\frac{1}{n}\sum_{t=1}^{n}\left(J(\theta)^{-1/2}\hat{J}(\theta', x_t)J(\theta)^{-1/2}\right)_{ij}.$$

The law of large numbers informs us that it is near its expectation with high probability. We use the expected supremum property to get uniformity over $B_\delta(\theta)$. Indeed, the supremum of an average is less than the average of the supremum. In particular, $\sup_{\theta' \in B_\delta(\theta)} \pm(I_{ij} - (J(\theta)^{-1/2}\hat{J}(\theta', x^n)J(\theta)^{-1/2})_{ij}$ is not more than

$$\frac{1}{n}\sum_{t=1}^{n}\sup_{\theta' \in B_\delta(\theta)} \pm\left(I_{ij} - (J(\theta)^{-1/2}\hat{J}(\theta', x_t)J(\theta)^{-1/2})_{ij}\right).$$

When the x_t are independent from the distribution P_θ, this is a sample average having expectation bounded by $\epsilon/(2d)$. Accordingly, by the law of large numbers, for each i, j, and for each sign choice, it is within ϵ/d of 0 except in an event of (P_θ) probability tending to 0 as $n \to \infty$. Accordingly, taking the exception event as the finite union of these tail events for $1 \leq i, j \leq d$ and for each sign, we have that $\max_{i,j} \sup_{\theta' \in B_\delta(\theta)} |(I_{ij} - (J(\theta)^{-1/2}\hat{J}(\theta', x^n)J(\theta)^{-1/2})_{ij}|$ is not more than ϵ/d except in an event having a (P_θ) probability, say δ_n, which tends to zero as $n \to \infty$.

Now the spectral norm of a d by d matrix is within a factor d of the maximum absolute value of its entries. Accordingly,

$$\sup_{|\zeta|=1}\sup_{\theta' \in B_\delta(\theta)} \zeta^t\left(I - (J(\theta)^{-1/2}\hat{J}(\theta', x^n)J(\theta)^{-1/2})\right)\zeta \quad (89)$$

is less than ϵ, except in the event of probability δ_n. Any unit vector ζ may be obtained as $J(\theta)^{1/2}z/\|J(\theta)^{1/2}z\|$ for some non-zero vector z. Accordingly, switching from ζ to z, this gives

$$\sup_{\theta' \in B_\delta(\theta)}\inf_{z \neq 0}\frac{z^t\hat{J}(\theta', x^n)z}{z^tJ(\theta)z} \geq 1 - \epsilon$$

except in the event of vanishing P_θ probability, asymptotically.

APPENDIX B
ON ASSUMPTIONS FOR LOWER BOUNDS OF THEOREM 2

The analysis in the proof of Theorem 2 is similar. The extra assumptions of continuity of the expected supremum function and continuity of the Fisher information are used, along with Dini's Theorem, to deduce that the convergence to 0 of the expected supremum as well as the convergences to 0 of the expressions in (6) and (7), are in fact uniform convergences within the whole presumed compact K (with no need for an Egorov style exception set $K \setminus G$). In particular there is a $\delta > 0$, sufficiently small, such that the expected value of the supremum in expression (89) is less than $\epsilon/2$, uniformly over θ in K.

Another distinction is the use of the assumption of the continuous and finite expected square, to permit an appeal to Chebyshev's inequality. The variance of the supremum in expression (89) is not more than v_θ/n, where v_θ is the

variance of the supremum when $n = 1$. From Assumption 1′ this variance is finite and continuous so it has a finite bound v_K on the compact domain K. Accordingly, by Chebyshev's inequality, the probability that expression (89) exceeds ϵ is not more than $4v_K/(n\epsilon^2)$, uniformly over θ in K. This converges to zero faster than the required $1/\log n$. [A similar sufficiently fast tail probability conclusion can be arranged if using a $1+\alpha$ moment of the empirical Fisher information for any $\alpha > 0$].

APPENDIX C
ASYMPTOTIC NORMALITY AND UNIFORM TIGHTNESS OF THE DISTRIBUTION OF THE MLE

First we restate and then prove Lemma 2 from Section IV, which is for general i.i.d. families.

Lemma 2: Under Assumptions 1′, 2′, 3′, $\sqrt{n}J(\theta)^{1/2}(\hat\theta - \theta)$ converges in distribution to a mean zero normal random vector and there is a constant c such that the following inequality holds for any positive b,

$$\max_{\theta \in K} P_\theta\left(\sqrt{n}||J(\theta)^{1/2}(\hat\theta - \theta)|| > b\sqrt{\log n}\right)$$
$$\leq \frac{c}{b^2 \log n} + o\left(\frac{1}{\log n}\right).$$

The constant c equals $\max_{\theta \in K} \text{trace}(J(\theta)^{-1}I(\theta))$ which is the dimension d, the trace of an identity matrix, when $I(\theta) = J(\theta)$.

Proof: Assume that x^n is drawn from $p(\cdot|\theta)$. Let $l_{x^n}(\theta)$ be the log likelihood function $\log p(x^n|\theta)$. Then, the first order Taylor expansion of the score function $\nabla l_{x^n}(\theta)$ at $\hat\theta$ shows

$$\nabla l_{x^n}(\theta) = \nabla\nabla^t l_{x^n}(\tilde\theta)(\theta - \hat\theta),$$

where $\nabla f(\theta)$ and $\nabla\nabla^t f(\theta)$ denote the gradient and the Hessian of f, respectively, and $\tilde\theta$ is a point on the line between θ and $\hat\theta$. From the definition of the empirical Fisher information, this equation is

$$\nabla l_{x^n}(\theta) = n\hat J(\tilde\theta, x^n)(\hat\theta - \theta), \tag{90}$$

which is equivalent to

$$J(\theta)^{-1/2}\frac{1}{\sqrt{n}}\nabla l_{x^n}(\theta)$$
$$= [J(\theta)^{-1/2}\hat J(\tilde\theta, x^n)J(\theta)^{-1/2}][\sqrt{n}J(\theta)^{1/2}(\hat\theta - \theta)].$$

Let's call the random vector on the left side Z_n. It is asymptotic normal by the central limit theorem (for sums of i.i.d. random vectors with finite covariance) and it has covariance $EZ_nZ_n^t = J(\theta)^{-1/2}I(\theta)J(\theta)^{-1/2}$.

By Assumption 3′ the estimator $\hat\theta$ is in a δ neighborhood of θ with probability greater than $1 - o(1/\log n)$ uniformly for $\theta \in K$. Then by (25), which is a consequence of Assumption 1', $J(\theta)^{-1/2}\hat J(\tilde\theta, x^n)J(\theta)^{-1/2}$ is close to I, indeed not less than $(1 - \epsilon)I$, for all $\theta \in K$ with probability at least $1 - o(1/\log n)$. (The inequality of matrices is in the sense that the difference is a non-negative definite matrix.) Accordingly, the random vector of interest defined by $\xi_n = \xi_n(\theta) = \sqrt{n}J(\theta)^{1/2}(\hat\theta - \theta)$ satisfies $\xi_n = A_nZ_n$ where $A_n \leq (1 - \epsilon)^{-1}I$. It follows that $E\xi_n\xi_n^t \leq (1 - \epsilon)^{-2}EZ_nZ_n^t$ and that $E||\xi_n||^2 \leq (1 - \epsilon)^{-2}E||Z_n||^2$ which is

$(1 - \epsilon)^{-2}\text{trace}(J(\theta)^{-1}I(\theta))$, not more than $(1 - \epsilon)^{-2}c$. Thus by Chebyshev's inequality

$$\max_{\theta \in K} P_\theta\left(\sqrt{n}||J(\theta)^{1/2}(\hat\theta - \theta)|| > b\sqrt{\log n}\right)$$
$$\leq \frac{(1 - \epsilon)^{-2}c}{b^2 \log n} + o\left(\frac{1}{\log n}\right).$$

This means that $\limsup_n(\log n)\max_{\theta \in K} P_\theta(||\xi_n(\theta)|| \geq b\sqrt{\log n}) \leq (1 - \epsilon)^{-2}c/b^2$. Since this bound holds for all positive ϵ, it follows that this limsup is less than c/b^2, which is the desired result. *This completes the proof of Lemma 2*

Remark on Uniform Tightness: Consider the family of approximately standardized random variables $\xi_n(\theta) = \sqrt{n}J(\theta)^{1/2}(\hat\theta - \theta)$. Convergence in distribution provides a form of stochastic tightness, that is, $P_\theta(||\xi_n(\theta)|| > a)$ tends to zero as $a \to \infty$, uniformly in n. Here we used slightly stronger assumptions to get that the convergence is also uniformly valid for $\theta \in K$ and that moderately sized tails with $a_n = b\sqrt{\log n}$ have tail probability bounded by a small multiple of $1/\log n$.

APPENDIX D
CONSISTENCY OF THE MLE

The present setting is for a model in which the $x_1, \ldots, x_n$ given θ are independent and identically distributed.

Here we state conditions that are sufficient for the consistency of the MLE for each θ. The subsequent section addresses uniform consistency for θ in compacta.

For consistency, we use Assumptions 11-13 described below. The first assumption concerns continuity.

Assumption 11: Continuity. For almost every $x \in \mathcal{X}$ and for all $\theta \in \Theta$, $p(x|\theta)$ is upper half continuous at θ:

$$\lim_{\delta \to 0} \sup_{\theta':|\theta - \theta'|<\delta} p(x|\theta') = p(x|\theta).$$

Moreover, the relative entropy $D(\theta||\theta')$ is a continuous function of θ and θ' in Θ.

Assumption 12: For each θ and θ' in the interior of Θ, there is a $\delta > 0$ such that the following holds

$$E_\theta\left(\sup_{\theta''\in\Theta:|\theta''-\theta'|<\delta} \log\frac{p(x|\theta'')}{p(x|\theta)}\right) < \infty.$$

The above two assumptions are sufficient for consistency when the parameter space Θ is compact. Accordingly, for bounded parameter spaces, it is natural to include in the parameter space any boundary points that do correspond to probability densities.

For non-compact parameter spaces, the following additional assumption is used to handle those cases in which the densities lose mass as the parameter approaches boundary points not in Θ. For $B \geq 1$, we let Θ_B be an increasing sequence of compact subsets whose union is Θ.

For one example, consider the family of exponential densities $p(x|\theta) = \theta e^{-\theta x}$ for $x > 0$. It has the natural parameter space $\Theta = \{\theta : \theta > 0\}$ and the sets

$$\Theta_B = \{\theta : 1/B \leq \theta \leq B\}$$

are increasing compact sets whose union is Θ. Moreover, as $\theta' \to 0$ or as $\theta' \to \infty$ it has $\lim p(x|\theta') = 0$ for each $x > 0$.

When the parameter space is $\Theta = R^d$ we may take

$$\Theta_B = \{\theta : |\theta| \leq B\}.$$

Assumption 13: Let an increasing sequence of compact subsets Θ_B be given whose union is Θ. For each $\theta \in \Theta^\circ$, there is a B sufficiently large such that

$$E_\theta \Big(\sup_{\theta' \in \Theta \backslash \Theta_B} \log \frac{p(x|\theta')}{p(x|\theta)} \Big) < \infty.$$

Moreover, there is a value $a > 0$, such that, for any sequence θ_B, $B \geq 1$, with $\theta_B \in \Theta \backslash \Theta_B$, in the limit the likelihood ratio satisfies $\limsup_B p(x|\theta_B)/p(x|\theta) \leq \exp\{-a\}$, for each $x \in \mathcal{X}$.

We call this latter property the loss of mass for any sequences diverging from the set. Often, as in the example above, the indicated limit property holds trivially, with full loss of mass $\lim p(x|\theta_B) = 0$, if there be sequences θ_B heading out of compacta in Θ.

For compact Θ, taking $\Theta_B = \Theta$, the set $\Theta \backslash \Theta_B$ is empty and Assumption 13 is regarded as holding vacuously, with the supremum of an empty set taken to be $-\infty$.

Under 11, 12, 13, the maximum likelihood estimate is a consistent estimator of θ. We prove the following.

Lemma 15: For a family of densities $S = \{p(\cdot|\theta) : \theta \in \Theta\}$, suppose Assumptions 11, 12, and 13 hold. Then, for each θ in Θ°, we have convergence to zero in probability of the relative entropy between the densities at the maximum likelihood estimate $\hat{\theta}(x^n)$ and at θ. That is,

$$\lim_{n \to \infty} D(\theta||\hat{\theta}) = 0, \quad \text{in } P_\theta \text{ probability.}$$

Moreover, if $D(\theta||\theta')$ is continuous in θ', and if $P_{\theta'} = P_\theta$ only at $\theta' = \theta$, then also $\hat{\theta}(x^n)$ converges to θ in probability. That is, for every $\epsilon > 0$.

$$P_\theta(|\hat{\theta}(x^n) - \theta| > \epsilon) = o(1).$$

Proof: For positive ξ, let U be a subset of Θ defined as $U \overset{\text{def}}{=} \{\theta' \in \Theta : D(\theta||\theta') \geq \xi\}$. By the continuity of $D(\theta||\theta')$ as a function of θ', when Θ is compact, so also is U. When Θ is not compact, in place of U we use $U \cap \Theta_B$ for which will discuss the size of B later below. By Assumption 11, we have, for each θ' in U,

$$\lim_{\delta \to 0} \sup_{\theta'':|\theta'' - \theta'| < \delta} \log \frac{p(x|\theta'')}{p(x|\theta)} = \log \frac{p(x|\theta')}{p(x|\theta)}.$$

Therefore, we have

$$\lim_{\delta \to 0} E_\theta \Big(\sup_{\theta'':|\theta'' - \theta'| < \delta} \log \frac{p(x|\theta'')}{p(x|\theta)} \Big) = -D(\theta||\theta'),$$

by the monotone convergence theorem, since $\sup_{\theta'':|\theta'' - \theta'| < \delta} \log(p(x|\theta'')/p(x|\theta))$ decreases as δ decreases for each $x \in \mathcal{X}$ and since it has finite expectation by Assumption 12.

For each $\xi > 0$ and each $\theta' \in U$, let $\delta(\theta') = \delta(\theta', \xi) > 0$ be so small that

$$E_\theta \Big(\sup_{\theta'':|\theta'' - \theta'| < \delta(\theta')} \log \frac{p(x|\theta'')}{p(x|\theta)} \Big) \leq -D(\theta||\theta') + \xi/2$$

holds. Then, since $D(\theta||\theta') \geq \xi$ in U, we have

$$E_\theta \Big(\sup_{\theta'':|\theta'' - \theta'| < \delta(\theta')} \log \frac{p(x|\theta'')}{p(x|\theta)} \Big) \leq -\xi/2.$$

Now consider $\frac{1}{n} \sup_{\theta'':|\theta'' - \theta'| < \delta(\theta')} \log \frac{p(x^n|\theta'')}{p(x^n|\theta)}$ for samples of size n. Since we are presently treating models in which the x_1 through x_n are independent. It is

$$\sup_{\theta'':|\theta'' - \theta'| < \delta(\theta')} \frac{1}{n} \sum_{i=1}^n \log \frac{p(x_i|\theta'')}{p(x_i|\theta)}.$$

Then since the supremum of the average is less than the average of the supremum, we can apply the law of large numbers and deduce that it has a limit in (P_θ) probability not greater than $-\xi/2$, for each θ' in U. We will need to arrange an analogous property holding uniformly over θ' in U, and we will come back to that matter momentarily.

For non-compact domains Θ, consider the sequence $\sup_{\theta' \in \Theta \backslash \Theta_B} \log \frac{p(x|\theta')}{p(x|\theta)}$. It is decreasing in B and it has limit less than or equal to $-a$ as $B \to \infty$, by Assumption 11. Then with $0 \leq \xi \leq a$, by the monotone convergence theorem, we can arrange that B is sufficiently large that

$$E_\theta \Big(\sup_{\theta' \in \Theta \backslash \Theta_B} \log \frac{p(x|\theta')}{p(x|\theta)} \Big) < -\xi/2.$$

Accordingly, again using that a supremum of an average is less than the average of a supremum, we can deduce that $\frac{1}{n} \sup_{\theta' \in \Theta \backslash \Theta_B} \log \frac{p(x^n|\theta'')}{p(x^n|\theta)}$ has a limsup in probability not greater than $-\xi/2$.

To obtain that $D(\theta||\hat{\theta})$ is less than ξ it is enough to confirm that the likelihood at θ is higher than the supremum of the likelihoods for all θ' with $D(\theta||\theta') \geq \xi$. Indeed, this gives an instance of a parameter value in the Kullback ball (namely at the center point θ) that has higher likelihood than at all points outside the Kullback ball. Accordingly the maximum must be in that ball. So if suffices for Kullback consistency to show that with high probability

$$\sup_{\theta':D(\theta||\theta') \geq \xi} \log \frac{p(x^n|\theta')}{p(x^n|\theta)} < 0.$$

In the compact Θ case, this supremum is over all θ' in U. Now the compact U is covered by the (infinite) union of all the variable radius balls $\{\theta'' : |\theta'' - \theta'| \leq \delta(\theta')\}$ for θ' in U. For a compact set every cover has a finite subcover. Thus, there exist a finite number of points $\theta_1, \theta_2, ..., \theta_N$ such that the family of sets $\mathcal{U} = \{U_i\}$ ($i = 1, 2, ..., N$) covers $\bar{U}$, where $U_i \overset{\text{def}}{=} \{\theta : |\theta - \theta_i| < \delta(\theta_i)\}$.

Similarly, in the non-compact Θ case, we obtain such a cover of the compact $U \cap \Theta_B$ and append the set $U_0 = \{\theta' \in \Theta \backslash \Theta_B : D(\theta||\theta') \geq \xi\}$ to obtain thereby $U_0 \cup U_1 \cup ... \cup U_N$ as a cover of $\{\theta' \in \Theta : D(\theta||\theta') \geq \xi\}$.

Now, we have

$$\sup_{\theta':D(\theta||\theta)>\xi} \left(\frac{1}{n} \log \frac{p(x^n|\theta')}{p(x^n|\theta)} \right)$$
$$\leq \max_{0\leq k\leq N} \left(\frac{1}{n} \sup_{\theta'\in U_k} \log \frac{p(x^n|\theta')}{p(x^n|\theta)} \right)$$
$$\leq \max_{0\leq k\leq N} \left(\frac{1}{n} \sum_{i=1}^{n} \sup_{\theta'\in U_k} \log \frac{p(x_j|\theta')}{p(x_j|\theta)} \right).$$

Now as we showed each of these finitely many averages $\frac{1}{n} \sum_{i=1}^{n} \sup_{\theta'\in U_k} \log \frac{p(x_j|\theta')}{p(x_j|\theta)}$ is less than 0 except in an event $E_{k,n}$ of probability $P_\theta(E_{k,n})$ which tends to zero as $n \to \infty$. Accordingly, the finite union of these exception sets $E_n = \cup_{k=0}^{N} U_{k,n}$ has probability $P_\theta(E_n)$ tending to zero as well. Outside E_n, we have $\sup_{\theta':D(\theta||\theta)>\xi} \log \frac{p(x^n|\theta')}{p(x^n|\theta)} < 0$.

So for each $\xi > 0$, the maximum likelihood estimate is in the Kullback ball of size ξ except in an event of probability tending to zero as $n \to \infty$. This means that $D(\theta||\hat{\theta})$ converges to 0 in probability.

Now for any $\epsilon > 0$, with $D(\theta||\theta')$ continuous in θ' and zero only at $\theta' = \theta$, it follows that in any compact $K \subseteq \Theta$ (such as $K = \Theta_B$)

$$\inf_{\theta'\in K:|\theta'-\theta|\geq\epsilon} D(\theta||\theta') > 0$$

Moreover, for non-compact Θ the condition on the limit of $p(x|\theta')/p(x|\theta)$ as θ' diverges from compacta, prevents $D(\theta||\theta')$ from tending to zero for such sequences. Accordingly, the convergence of $D(\theta||\hat{\theta})$ to zero (in probability) implies $|\hat{\theta}(x^n) - \theta| \to 0$ (in probability). *This completes the proof of the Lemma.*

*R*emarks: The first consistency proof of this type is in Wald [47]. Our conditions are similar, but slightly weaker in that we use suprema of log density ratios (demonstrated close to relative entropies). In contrast Wald use suprema of log densities together with finite entropy conditions. Secondly, here we organized the proof to exhibit the primacy of the conclusion of information consistency ($D(\theta||\hat{\theta})$ tending to zero in probability) with the parameter consistency conclusion $\hat{\theta} \to \theta$ of Wald as a consequence.

Extension to parameter spaces that are separable metric spaces is straightforward, provided the finite balls of radius B are compact (so that covers have finite subcovers). Extension to almost sure convergence is also straightforward. Also, the conclusion extends to the case that the governing distribution P is not in the family $\{P_\theta : \theta \in \Theta\}$, but is an information limit of a sequence of members of the family (that is, there exists P_{θ_k} with $D(P||P_{\theta_k})$ tending to 0 as k grows).

APPENDIX E
CONSISTENCY CONDITIONS AND THE SHTARKOV VALUE

Recall that the minimax regret value is defined by $r_n(K) = \min_q \sup_{x^n\in\mathcal{X}^n} \sup_{\theta\in K} \log p(x^n|\theta)/q(x^n)$, where the minimum is over q which are non-negative and integrate to not more than 1 with respect to ν. Per the theory of Shtarkov [39], the exact minimax value is $r_n(K) = \log c_{n,K}$ achieved

by the normalized maximum likelihood distribution $q(x^n) = \sup_{\theta\in K} p(x^n|\theta)/c_{n,K}$ where $c_{n,K} = \int \sup_{\theta\in K} p(x^n|\theta)\eta(d\theta)$ is the Shtarkov constant. As explained in the introduction, it is the properties of $r_n(K)$ that are of interest to us here. In particular, the use of Bayes mixture approximations to the minimax distribution arises from aim of approximate implementation as well as the aim of determining approximations of the Shtarkov value suitable for use in Rissanen's stochastic complexity formulation of the minimum description length (MDL) principle.

In this appendix we show a relationship between the consistency conditions and finiteness of the Shtarkov constant.

Let's focus attention on the case that the parameter space is taken to be a set K (possibly a subset of the whole parameter space) for which $c_{n,K}$ is finite.

For independent random variables, if the total sample size $N = nm$ is a multiple of a block size m, one has the option to organize the sequence $x_1, x_2, \ldots, x_N$ as the sequence of independent blocks $\underline{x}_1^m, \underline{x}_2^m, \ldots \underline{x}_n^m$, each of size m, where

$$\underline{x}_i^m = (x_{(i-1)m+1}, x_{(i-1)m+2}, \ldots, x_{im}).$$

Accordingly, for $m \geq 1$, the Assumptions 12 and 13 generalize as follows.

Assumption 12[*m*]: For the specified blocksize m and for each θ and θ' in the interior of K, there is a $\delta > 0$ such that the following holds

$$E_\theta \left(\sup_{\theta''\in\Theta:|\theta''-\theta'|<\delta} \log \frac{p(x^m|\theta'')}{p(x^m|\theta)} \right) < \infty.$$

Assumption 13[*m*]: Let an increasing sequence of compact subsets K_B be given whose union is K. For the specified m and for each $\theta \in K^\circ$, there is a B sufficiently large such that

$$E_\theta \left(\sup_{\theta'\in K\setminus K_B} \log \frac{p(x^m|\theta')}{p(x^m|\theta)} \right) < \infty.$$

Moreover, there is the loss of mass property for sequences diverging from K. That is, there is a value $a > 0$, such that, for any sequence θ_B, $B \geq 1$, with $\theta_B \in K \setminus K_B$, in the limit, the likelihood ratio satisfies $\limsup_B p(x^m|\theta_B)/p(x^m|\theta) < \exp\{-a\}$, for each $x^m \in \mathcal{X}$.

With these assumptions we have the following consistency conclusion.

Lemma 16: For a family of densities $S = \{p(\cdot|\theta) : \theta \in \Theta\}$, and a blocksize m, suppose Assumptions 11, 12[*m*] and 13[*m*]. Consider the sequence of maximum likelihood estimates $\hat{\theta}(x^N)$ with sample sizes $N = nm$ a multiple of m. Then, for each θ in Θ°, we have convergence to zero in probability of the relative entropy between the densities at the maximum likelihood estimate $\hat{\theta}$ and at θ. That is,

$$\lim_{n\to\infty} D(\theta||\hat{\theta}) = 0, \quad \text{in } P_\theta \text{ probability}.$$

Moreover, if $D(\theta||\theta')$ is continuous in θ', and if $P_{\theta'} = P_\theta$ only at $\theta' = \theta$, then also $\hat{\theta}(x^n)$ converges to θ in probability. That is, for every $\epsilon > 0$.

$$P_\theta(|\hat{\theta}(x^n) - \theta| > \epsilon) = o(1).$$

Proof: Taking the sequence of sample sizes nm which are a multiple of m, the proof is essentially the same as in the $m = 1$ case.

Lemma 17: If the Shtarkov value $c_{1,K}$ at $m = 1$ is finite, then it implies the satisfaction of assumptions 12 and 13, and hence (together with the continuity assumption 11 and the loss of mass assumption if K is not compact), a finite Shtarkov value implies the full consistency of the MLE, per the conclusions of Lemma 15. Likewise, for $m \geq 1$, finiteness of the Shtarkov value $c_{m,K}$ implies the satisfaction of assumptions 12[m] and 13[m] and, hence (along with the continuity assumption 11 and the loss of mass assumption if K is not compact), it implies the consistency of the MLE for the sequence of sample sizes that are multiples of m, per the conclusions of Lemma 16.

Proof: The expected suprema in 12[m] and 13[m] are

$$E_\theta\left(\log \frac{\sup_{\theta'' \in K \cap B_{\theta',\delta}} p(x^m|\theta'')}{p(x^m|\theta)}\right)$$

and

$$E_\theta\left(\log \frac{\sup_{\theta' \in K \setminus K_B} p(x^m|\theta')}{p(x^m|\theta)}\right)$$

which by Jensen's inequality have the upper bounds

$$\log \int \sup_{\theta'' \in K \cap B_{\theta',\delta}} p(x^m|\theta'')\nu(dx^m)$$

and

$$\log \int \sup_{\theta' \in K \setminus K_B} p(x^m|\theta')\nu(dx^m),$$

respectively. Thus (local) domination of the joint densities is sufficient for (local) domination of the log likelihood ratios.

Next, we can further upperbound these by a global dominating quantity (when finite)

$$\log \int \sup_{\theta' \in K} p(x^m|\theta')\nu(dx^m)$$

which is $\log c_{m,K}$, the Shtarkov minimax regret. The indicated consistency properties then follow when it is finite.

APPENDIX F
UNIFORM CONSISTENCY OF THE MLE

We turn our attention to assumptions for uniform consistency, on compact subsets of Θ, as used in the development of Theorem 2.

Recall that we assumed continuity of the mean of the log density ratio $D = D(\theta||\theta')$ in assumption 11. Here we will also assume continuity of its variance

$$V(\theta||\theta') = E_\theta[(\log p(x|\theta)/p(x|\theta') - D)^2].$$

Assumption 11': $V(\theta||\theta')$ is a continuous function of θ' and θ in Θ.

Assumption 12': Within each compact subset of Θ, there exists a $\bar{\delta} > 0$ such that for $0 < \delta \leq \bar{\delta}$ the following expressions are finite and continuous as a function of θ and θ'

$$E_\theta\left[\sup_{\theta'' \in \Theta: |\theta'' - \theta'| < \delta} \log \frac{p(x|\theta'')}{p(x|\theta)}\right]$$

and

$$E_\theta\left[\left(\sup_{\theta'' \in \Theta: |\theta'' - \theta'| < \delta} \log \frac{p(x|\theta'')}{p(x|\theta)}\right)^2\right].$$

The above assumption 12', along with 11 and 11', is sufficient to show uniform consistency for θ within any compact subset when the parameter space on which the likelihood is maximized is also compact.

There can also be interest to know whether uniform consistency within compact subsets K can hold when the parameter space Θ is not compact. For that purpose we have the following assumption.

Assumption 13': Let K be a compact subset of Θ. There exists a certain positive number B such that

$$\sup_{\theta \in K} E_\theta\left[\left(\sup_{\theta' \in \Theta \setminus \Theta_B} \log \frac{p(x|\theta')}{p(x|\theta)}\right)^2\right] < \infty$$

and

$$\sup_{\theta \in K} E_\theta \sup_{\theta' \in \Theta \setminus \Theta_B} \log \frac{p(x|\theta')}{p(x|\theta)} < 0$$

hold.

We prove the following uniform consistency of maximum likelihood estimators.

Lemma 18: Consider a family of densities $S = \{p(\cdot|\theta) : \theta \in \Theta\}$. Consider any global maximizer $\hat{\theta}(x^n)$ of the likelihood over θ in Θ. When Θ is compact, suppose Assumptions 11, 11' and 12'. Then for each $\xi > 0$

$$\sup_{\theta \in \Theta} P_\theta(D(\theta||\hat{\theta}) > \xi) = O(1/n).$$

and likewise, for each $\epsilon > 0$,

$$\sup_{\theta \in \Theta} P_\theta(|\hat{\theta}(x^n) - \theta| > \epsilon) = O(1/n).$$

If Θ is not compact, suppose Assumptions 11, 11', 12' and 13'. Then, for any compact subset K, we have

$$\sup_{\theta \in K} P_\theta(D(\theta||\hat{\theta}) > \xi) = O(1/n)$$

and

$$\sup_{\theta \in K} P_\theta(|\hat{\theta}(x^n) - \theta| > \epsilon) = O(1/n).$$

Proof: In the compact Θ case, let $\bar{\delta} > 0$ be a choice such that the square of the expected supremum in Assumption 12' is finite and and continuous in θ' and θ. We could proceed as in the proof of the previous lemma, arranging, for each $\xi > 0$, positive $\delta(\theta, \theta') \leq \bar{\delta}$ sufficiently small to achieve the negative expected supremum as explained there. But we can arrange matters a little better, because we have convergence as δ goes to 0 of functions, here assumed to be continuous on the compact set, which are converging to the continuous function $-D(\theta||\theta')$. So by Dini's theorem it is a uniform convergence. Consequently, we can deduce that for each $\xi > 0$, there is a common choice of positive $\delta < \bar{\delta}$ with which the expected supremum is less than $-D(\theta||\theta') + \xi/2$ (and moreover, the variance of the expected supremum is arbitrarily close to its continuous limit $V(\theta||\theta')$). From the common choice δ of the radii of the balls, when we arrange the finite cover $U_1, \ldots, U_N$,

if desired, we have control on the number of them. Then, as in the proof of the previous lemma,

$$\sup_{\theta':D(\theta||\theta)>\xi} \left(\frac{1}{n} \log \frac{p(x^n|\theta')}{p(x^n|\theta)} \right)$$

$$\leq \max_{1 \leq k \leq N} \left(\frac{1}{n} \sum_{i=1}^{n} \sup_{\theta' \in U_k} \log \frac{p(x_j|\theta')}{p(x_j|\theta)} \right),$$

where each of these suprema have expectation less than $-\xi/2$. Then by Chebyshev's inequality each of these averages is less than 0 except in an event of probability not more than $4v_\delta/(n\xi^2)$, where v_δ is the maximum over θ' and θ of the variance of the expected suprema (this maximum being finite since it is a maximum over a compact set of a function assumed continuous). So we have, crudely, the union bound on the exception event $4Nv_\delta/(n\xi^2)$ of order $1/n$. This proves the conclusion in the compact Θ case.

Maximization of the likelihood over non-compact Θ is handled with an analogous incorporation of the Assumption $13'$. This completes the proof.

Remarks: Note that in the non-compact Θ setting, with Assumption $13'$, the domain for maximization of the likelihood is allowed to be larger than the set K of θ values for which we bound $P_\theta\{|\hat{\theta}(x^n)-\theta|| \geq \epsilon\}$. That Assumption may be dropped if the likelihood is maximized only within the compact K.

APPENDIX G
ON ASSUMPTIONS IN THE NON-I.I.D. SETTING

Stochastic processes for the joint densities $p(x^n|\theta)$, $n \geq 1$, provide considerable generality, beyond the i.i.d. model setting, in which the regret could potentially be analyzed. Two natural settings include Markov models with time homogeneous transitions and stationary ergodic processes. The core ingredients are the conditional densities $p(x_{t+1}|x_t, \ldots x_1, \theta)$.

Such conditional densities provide the heart of what is needed for arithmetic coding and predictive gambling procedures. Indeed, to facilitate implementation, the rational for Bayes procedures that approximate the mimimax regret is that these yield predictive densities $q(x_{t+1}|x_t, \ldots, x_1)$ which can be computed (e.g. via posterior Monte Carlo) as the posterior means of the $p(x_{t+1}|x_t, \ldots x_1, \theta)$ using the posterior $w(\theta|x_t, \ldots, x_1)$ built from the (suitably modified) Jeffreys prior.

For stationary processes, there is, for each $k \geq 1$, a function $p(s_{k+1}|s_k, \ldots, s_1, \theta)$ on $\mathcal{X}^{k+1}$ such that the conditional probability density for X_{t+1}, evaluated at $X_{t+1} = s_k$, given $X_t = s_k, \ldots, X_{t+1-k} = s_1$ and given θ is the same for all $t \geq k$, equal to this $p(s_{k+1}|s_k, \ldots, s_1, \theta)$.

These stationary conditionals exist also for Markov processes that do not necessarily start in the invariant distribution. More generally, for asymptotically mean stationary processes (in the sense of Gray and Kieffer [22]), associated to the process with densities $p(x^n|\theta)$ there is a stationary process with densities $\bar{p}(x^n|\theta)$ with corresponding stationary conditionals.

For simplicity, here we present results for irreducible Markov processes (of arbitrary order k). And leave to briefer

mention settings in which for large k these approximate more general asymptotically mean stationary processes which have a stationary ergodic $\bar{p}(x^n|\theta)$, $n \geq 1$.

We consider the following setting for the Markov model of order k. The model of joint densities takes the form

$$p(x^n|\theta) = p_{init}(x^k) \prod_{t=k+1}^{n} p(x_t|x_{t-1}, \ldots x_{t-k}, \theta),$$

where the conditional (transition) density function is time-homogeneous as mentioned above. It depends on the parameter θ. It is assumed that there is an invariant probability distribution $P^*_{X^k|\theta}$, which, in general, would depend on the parameter. It may have a density $p^*(x^k|\theta)$. It will assumed that the Markov model is irreducible, so the invariant distribution is unique.

As for the initial density $p_{init}(x^k)$, with additional complication, we could consider the case that the initial is the invariant density $p^*(x^k|\theta)$. But for simplicity here, let's simply take the case that $p_{init}(x^k)$ is fixed (not depending on the parameter) and known.

Now the empirical Fisher information matrix takes the form

$$\hat{J}_{ij}(\theta, x^n) = -\frac{1}{n} \sum_{t=k+1}^{n} \frac{\partial^2}{\partial \theta^i \partial \theta^j} \log p(x_t|x_{t-1}, \ldots x_{t-k}, \theta),$$

where it will be assumed that $\log p(s_{k+1}|s^k, \theta)$ is twice continuously differentiable for every s^{k+1} in $\mathcal{X}^{k+1}$. The second derivatives in the expression above will also be denoted $\partial_i \partial_j \log p(x_t|x_{t-1}, \ldots x_{t-k}, \theta)$.

In investigating the expectation of the empirical Fisher information, let $J_{ij}(\theta|s^k)$ denote the conditional expectation of minus the second derivatives of the log of transition density, given an arbitrarily specified vector of previous values s^k. Then the Fisher information matrix $J_{ij,n}(\theta) = E_\theta[\hat{J}_{ij}(\theta, x^n)]$ takes the form

$$J_{ij,n}(\theta) = E_{\bar{P}_{n,\theta}}[J_{ij}(\theta|\cdot)] = \int J_{ij}(\theta|s^k)\bar{p}_n(s^k|\theta)\nu(s^k)$$

where $\bar{P}_{n,\theta} = (1/n)\sum_{t=k+1}^{n} P_{X^{t-1}_{t-k}|\theta}$ is the average across t of the distributions for $X^{t-1}_{t-k} = (x_{t-k}, \ldots, x_{t-1})$ for $t \geq k+1$ having joint density

$$\bar{p}_n(s^k|\theta) = \frac{1}{n} \sum_{t=k+1}^{n} p_{X^{t-1}_{t-k}}(s^k|\theta).$$

Continuity of $J_{ij,n}(\theta)$ as a function of θ is inherited from the continuity of $J_{ij}(\theta|s^k)$, and of the distributions $P_{X^{t-1}_{t-k}|\theta}$. These distributions $P_{X^{t-1}_{t-k}|\theta}$ with density functions $p_{X^{t-1}_{t-k}}(s^k|\theta)$ are induced by the transitions starting from the specified initial distribution P_{init} for the first k.

In general the distribution $\bar{P}_{n,\theta}$ converges (weakly) to the invariant P^*_θ distribution on $\mathcal{X}^k$.

Let's assume that $J_{ij}(\theta|s^k)$ is a bounded function of s^k in $\mathcal{X}^k$ for each θ (which allows that the expectation with respect to the any initial distribution is finite). With $J_{ij}(\theta|s^k)$ is bounded and continuous for s^k in $\mathcal{X}^k$ for each θ, from the convergence of $\bar{P}_{n,\theta}$ to the invariant P^*_θ, it follows that the

Fisher information $J_{ij,n}(\theta)$ converges to the limit $J_{ij}(\theta)$ as $n \to \infty$, where the limiting Fisher information matrix

$$J_{ij}(\theta) = E_{P_\theta^*}[J_{ij}(\theta|\cdot)]$$

is the expectation using the invariant distribution.

Moreover, by the ergodic theorem (for asymptotically mean stationary processes [22]), if $\partial_i \partial_j \log p(x_t|x_{t-1}, \dots x_{t-k}, \theta)$ has finite expected absolute value, with respect to the stationary P_θ^*, it follows that the empirical Fisher information $\hat{J}_{ij}(\theta, x^n)$ converges to $J_{ij}(\theta)$ in P_θ probability. The following assumption allows the extension of that conclusion via local domination.

Assumption 14: The stochastic process is kth order Markov with a time homogeneous one-step-ahead transition density that is twice continuously differentiable for θ in Θ° and has a unique invariant distribution. The conditional Fisher information $J_{ij}(\theta|\cdot)$ is assumed bounded as a function $\mathcal{X}^k$ for each θ in Θ, and continuous as a function of θ for each s^k. Moreover, for every $\theta \in \Theta^\circ$ there is an $r = r(\theta)$ such that, for every i, j,

$$E_\theta^* \left[\sup_{\theta' \in B_r(\theta)} |\partial_i \partial_j \log p(x_{k+1}|x_k, \dots x_1, \theta')| \right] \qquad (91)$$

is finite, where the expectation is taken with respect to the stationary distribution. The Fisher information $J_n(\theta)$ and its limit $J(\theta)$ are assumed to be continuous as a function of θ in Θ. Moreover, $J(\theta)$ is strictly positive definite.

Accordingly, as in Appendix A, given any ϵ, there exists $\delta(\theta)$ sufficiently small that the expected supremum in Assumption 14 is less than any prescribed positive value, and likewise for the expected supremum when the matrix of second derivatives of $\log p(x_{k+1}|x_k, \dots x_1, \theta')$ is hit on the left and right by $J(\theta)^{-1/2}$. Moreover, using that a supremum of averages is less than the average of suprema, and appealing to the ergodic Theorem, we find again that, for each i, j, and for each sign choice,

$$\sup_{\theta' \in B_\delta(\theta)} \pm (I_{ij} - (J(\theta)^{-1/2} \hat{J}(\theta', x^n) J(\theta)^{-1/2})_{ij})$$

is within ϵ/d of 0 except in an event of (P_θ) probability tending to 0 as $n \to \infty$.

In compact sets, as we have previously indicated (via Dini's Theorem), convergence of continuous functions to a continuous limit is uniformly convergent. Accordingly, the convergence of $J_n(\theta)$ to $J(\theta)$ is a uniform convergence in K. Accordingly, the collection of $J_n(\theta)$ for large n is equicontinuous, uniformly in K, and they share a positive lower bound on their minimum eigenvalue. Moreover $\int_K |J_n(\theta)|^{1/2} d\theta$ converges to $\int_K |J(\theta)|^{1/2}$.

For consistency, one proceeds similarly, using the relative entropy rate. The relative entropy rate $D(\theta||\theta')$ defined by $\lim_n (1/n) D(P_{X^n|\theta}||P_{X^n|\theta'})$, in the Markov case, is equal to

$$D(\theta||\theta') = E_{P^*_{X^k|\theta}} E_{X_{k+1}|X^k,\theta} \left[\log \frac{p(X_{k+1}|X^k, \theta)}{p(X_{k+1}|X^k, \theta')} \right].$$

Assumption 15: For the Markov model, the relative entropy rate $D(\theta||\theta')$ is assumed to be continuous as a function of θ' and θ, and to satisfy the identifiability property that it is zero only at $\theta' = \theta$. Moreover, local domination of log-likelihood ratios is assumed for each θ and θ' in a compact K. Namely for each θ and θ' there is an $r = r(\theta, \theta')$ such that

$$E_\theta^* \left[\sup_{\theta'' \in K \cap B_r(\theta')} \log \frac{p(X_{k+1}|X^k, \theta'')}{p(X_{k+1}|X^k, \theta)} \right]$$

is finite and continuous as a function of θ and θ' in K.

These Assumptions 14 and 15 for Markov models are sufficient for the conditions of Theorem 3 for the lower bound conclusion for the regret for stochastic processes, for compact parameter sets.

Further details are omitted since they closely parallel the previous analysis from the i.i.d. setting.

One may wonder about further generalization beyond the Markov model setting. For consistency, the core matter is the convergence of $(1/n)$ log density ratios to a relative entropy (an asymptotic equipartion property (AEP)). One setting beyond Markov, is when target stationary ergodic processes P^* are approximated by possibly variable order Markov models P_θ. For this setting one can appeal to the moderately general AEP of [11], [21], and [1]. Presummably, shifting back to a time 0 reference, the domination condition would be of the form of an assumption of finiteness of

$$E_\theta^* \left[\sup_{\theta'' \in K \cap B_r(\theta')} \log \frac{p(X_0|\theta'', X_{-1}, X_{-2}, \dots, X_{-k})}{p^*(X_0|X_{-1}, X_{-2}, \dots)} \right].$$

where perhaps the Markov order $k = k(\theta')$ would depend on θ. To deal more generally for families of non-Markov stationary ergodic processes, there were early attempts in [35], and counterexamples for pairs of stationary (but not ergodic) processes in [32]. Nevertheless the question of formulation of suitably general conditions for an AEP for pairs of stationary ergodic processes P_θ and P_θ' is largely open. Until such is obtained, the matter of consistency of maximum likelihood in non-Markov models requires that it be addressed on a case-by-case basis.

APPENDIX H
LOWER BOUND ON GAUSSIAN MEASURE (PROPOSITION 1)

Recall that Φ denotes the measure of the d-dimensional standard normal distribution and $N_r(0) = \{z : |z| \le r\}$. The following holds.

Proposition 1: For all $n > 0$ and $\epsilon > 0$,

$$\Phi(N_{\sqrt{n}\epsilon}(0)) \ge 1 - \exp\left(-\frac{n\epsilon^2}{2}\left(1 - \frac{d}{n\epsilon^2} \log \frac{n\epsilon^2}{d}\right) + \frac{d}{2}\right),$$

which is larger than

$$1 - \exp\left(-\frac{n\epsilon^2}{4} + \frac{d}{2}\right),$$

when $n\epsilon^2/d \ge 2$.

Proof: Note that

$$\Phi(N_{\sqrt{n}\epsilon}(0)) = \Pr\left\{ \sum_{i=1}^d X_i^2 \le n\epsilon^2 \right\},$$

where each X_i is an independent standard normal random variable. The density function of normal distribution with mean 0 and variance σ^2 is

$$\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\Big(\frac{-x^2}{2\sigma^2}\Big) = \exp\Big(\frac{-x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\Big)$$
$$= \exp(ux^2 - \psi(u)),$$

where $u = -1/(2\sigma^2)$ and $\psi(u) = (1/2)\log(2\pi\sigma^2) = (1/2)\log(-\pi/u)$. Let $q(x|u) = \exp(ux^2 - \psi(u))$, which forms an exponential family, for which the natural parameter is $u = -1/(2\sigma^2)$. Let $v = E_u(x^2) = \sigma^2$. Then v is the expectation parameter coresponding to u. We have

$$q(x^d|u) = \exp(d(u\|x\|^2/d - \psi(u))),$$

where $x^d = (x_1, \ldots, x_d)$ and $\|x\|^2$ denotes the square of x^d's Euclidian norm. Note that the maximum likelihood estimate of $v = \sigma^2$ given x^d is $\hat{\sigma^2} = \|x\|^2/d$. Then by the large deviation inequality (Lemma 11), we have

$$\Pr\Big\{\sum_{i=1}^d X_i^2 \geq n\epsilon^2\Big\} \leq \exp(-dD(n\epsilon^2/d\|1)),$$

where $D(v\|v')$ denotes the Kullback-Leibler divergence from $q(x|u)$ to $q(x|u')$ (v' is the expectation parameter corresponding to u'). Here, we have

$$D(v\|v') = (u - u')v - \psi(u) + \psi(u')$$
$$= -\frac{1}{2} + \frac{1}{2}\frac{v}{v'} + \frac{1}{2}\log\frac{v'}{v}$$
$$= \frac{1}{2}\Big(-1 + \frac{v}{v'} - \log\frac{v}{v'}\Big).$$

Hence we have

$$dD(n\epsilon^2/d\|1) = \frac{d}{2}\Big(-1 + \frac{n\epsilon^2}{d} - \log\frac{n\epsilon^2}{d}\Big)$$
$$= -\frac{d}{2} + \frac{n\epsilon^2}{2}\Big(1 - \frac{d}{n\epsilon^2}\log\frac{n\epsilon^2}{d}\Big),$$

which is not less than

$$-\frac{d}{2} + \frac{n\epsilon^2}{2}\Big(1 - \frac{\log 2}{2}\Big) \geq -\frac{d}{2} + \frac{n\epsilon^2}{4},$$

when $n\epsilon^2/d \geq 2$. *This completes the proof.*

APPENDIX I
NORMALIZATION CONSTANT OF THE IDEAL PRIORS

Lemma 5: Let K be an arbitrary compact set in Θ°. Suppose Assumption $2'''$ holds and let $\underline{\lambda}$ denote a lower bound on the smallest eigenvalue of $J_{n,\theta}$ among $n \in \mathbb{N}$ and $\theta \in K$. Let

$$K_\epsilon = \{\theta : B_\epsilon(\theta) \subset K\},$$

Then for ϵ such that $n\epsilon^2/d \geq 2$ and $\epsilon^2\alpha^2 \leq \underline{\lambda}$, we have

$$C_{J,n}(K) \leq C_{K,\epsilon,n}^{(\alpha)} \leq \frac{C_{J,n}(K)}{1 - e^{-n\epsilon^2/4+d/2}} + \frac{C_{J,n}(K \setminus K_{\epsilon\alpha})}{\rho_n^{(\alpha)}(\epsilon)},$$

where

$$C_{J,n}(A) = \int_A |J_{n,\theta}|^{1/2}d\theta.$$

Proof: Assume $\theta \in K_{\epsilon\alpha}$. Then, by $\epsilon^2\alpha^2 \leq \underline{\lambda}$, $N_{\sqrt{n}\epsilon}(0) \subset (\sqrt{n}/\alpha)(J_{n,\theta})^{1/2}(K - \theta)$ is satisfied, which means

$$\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta)) = \Phi(N_{\sqrt{n}\epsilon}(0))$$

holds for $\theta \in K_{\epsilon\alpha}$. Hence by Proposition 1 and the assumption $n\epsilon^2/d \geq 2$, the following holds.

$$\inf_{\theta \in K_{\epsilon\alpha}} \Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta)) \geq 1 - e^{-n\epsilon^2/4+d/2}.$$

Using this inequality, we have

$$\int_K \frac{|J_{n,\theta}|^{1/2}}{\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))}d\theta$$
$$\leq \int_{K_{\epsilon/\alpha}} \frac{|J_{n,\theta}|^{1/2}}{1 - e^{-n\epsilon^2/4+d/2}}d\theta + \int_{K \setminus K_{\epsilon\alpha}} \frac{|J_{n,\theta}|^{1/2}}{\rho_n^{(\alpha)}(\epsilon)}d\theta$$
$$\leq \frac{C_{J,n}(K)}{1 - e^{-n\epsilon^2/4+d/2}} + \frac{C_{J,n}(K \setminus K_{\epsilon\alpha})}{\rho_n^{(\alpha)}(\epsilon)}.$$

This completes the proof of the Lemma.

APPENDIX J
LOWER BOUNDS ON $\rho_n^{(\alpha)}(\epsilon)$

For a subset A of $\mathbb{R}^d$, let $\mathrm{diam}(A)$ and $\mathrm{vol}(A)$ denote the diameter of A and the volume of A, respectively.

Lemma 6: Assume that K is compact, convex, and the closure of an open set in $\Re^d$. Then, for all ϵ such that $\mathrm{vol}(N_\epsilon(0)) \leq \mathrm{vol}(J_{n,\theta}^{1/2}K)/2$, the following holds.

$$\rho_n^{(\alpha)}(\epsilon, \theta) \geq \frac{\mathrm{vol}(J_{n,\theta}^{1/2}K)}{2(\mathrm{diam}(J_{n,\theta}^{1/2}K))^d V_d}\Phi(N_{\sqrt{n}\epsilon}(0)).$$

Proof: Let $L = J_{n,\theta}^{1/2}K$. Then, we have

$$\rho_n^{(\alpha)}(\epsilon, \theta) = \Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))$$
$$= \Phi(N_{\sqrt{n}\epsilon}(0) \cap (\sqrt{n}/\alpha)(L - J_{n,\theta}^{1/2}\theta))$$
$$= \Phi\Big((\sqrt{n}/\alpha)\big(N_{\epsilon\alpha}(0) \cap (L - J_{n,\theta}^{1/2}\theta)\big)\Big).$$

Define for $\eta \in L$ and for $\epsilon > 0$,

$$A(\eta) = \{u \in \Re^d : |u| = 1, \exists \kappa > 0, \ \eta + \kappa u \in L^\circ\},$$
$$C(\eta, \epsilon) = \{u \in \Re^d : |u| = 1, \ \eta + \epsilon u \in L^\circ\},$$

and

$$\tilde{C}(\eta, \epsilon) = \{v \in \Re^d : |v| = \epsilon, \ \eta + v \in L^\circ\}.$$

By definition, we have

$$A(\eta) \subset \bigcup_{\epsilon > 0} C(\eta, \epsilon).$$

By the convexity of L,

$$A(\eta) \supset C(\eta, \epsilon_1) \supset C(\eta, \epsilon_2)$$

holds for $0 < \epsilon_1 \leq \epsilon_2$. Hence, $A(\eta) = \bigcup_{\epsilon > 0} C(\eta, \epsilon)$ holds.

Note that $\tilde{C}(\eta, \epsilon)$ is an open subset of the surface of $N_\epsilon(0)$, since it is the intersection of $L^\circ - \eta$ (an open set) and the surface of $N_\epsilon(0)$. Hence it is measurable in the surface of $N_\epsilon(0)$, which means $C(\eta, \epsilon)$ is also measurable.

Let $r(\eta, \epsilon)$ denote the ratio of the $d-1$ dimensional volume of $C(\eta, \epsilon)$ to that of the surface of $N_1(0)$ ($d\pi^{d/2}/\Gamma(d/2+1)$). Since $C(\eta, \epsilon)$ does not decrease as ϵ decreases, $r(\eta, \epsilon)$ does not decrease as ϵ decreases.

For $\eta \in L$, define

$$\text{cone}(\eta, \kappa) = \{\eta + \xi u : u \in C(\eta, \kappa), \xi \geq 0\}.$$

If $C(\eta, \kappa)$ is connected, this is an unbounded cone with the vertex η. If not, it is a union of unbounded cones. Note that

$$N_\epsilon(\eta) \cap L \supset N_\kappa(\eta) \cap \text{cone}(\eta, \kappa) \tag{92}$$

holds for all $\kappa \in (0, \epsilon]$. This is shown as follows. Let $\eta + v$ be an element of $\text{cone}(\eta, \kappa)$ with $|v| \leq \kappa$. Then, $v/|v| \in C(\eta, \kappa)$ holds by the definition of $C(\eta, \kappa)$. This means that $\eta + \kappa v/|v| \in L^\circ$. Since $\kappa/|v| \geq 1$, $\eta + v$ is on the line segment between $\eta \in L$ and $\eta + \kappa v/|v| \in L^\circ$. Hence $\eta + v \in L^\circ$ holds. Since $|v| \leq \kappa \leq \epsilon$, (92) holds.

Because of (92), for any nonnegative integrand, the integral over $(\sqrt{n}/\alpha)(N_{\epsilon\alpha}(0) \cap (L - \eta))$ is lower bounded by that over $(\sqrt{n}/\alpha)(N_{\epsilon\alpha}(0) \cap (\text{cone}(\eta, \epsilon\alpha) - \eta))$. Then due to the symmetry of Φ, we have for each $\theta \in K$,

$$\rho_n^{(\alpha)}(\epsilon, \theta) \geq r(\eta, \epsilon\alpha)\Phi(N_{\sqrt{n}\epsilon}(0)),$$

where $\eta = J_{n,\theta}^{1/2}\theta$.

Finally, we will show that $\inf_{\eta \in L} r(\eta, \epsilon\alpha)$ is lower bounded by the positive constant determined by $\text{vol}(L)$ and $\text{diam}(L)$. Let ϵ_0 so small that $\text{vol}(N_{\epsilon_0}(0)) \leq \text{vol}(L)/2$ and assume that $\epsilon\alpha \leq \epsilon_0$. Let Δ denote $\text{diam}(L)$. Recall that $C(\eta, \epsilon\alpha) \supset C(\eta, \epsilon_0) \supset C(\eta, \kappa)$ holds for all $\kappa \geq \epsilon_0$ and for all ϵ such that $\epsilon\alpha \leq \epsilon_0$. Since

$$\frac{\eta' - \eta}{|\eta' - \eta|} \in C(\eta, \kappa)$$

holds for all $\eta' \in L^\circ$ with $|\eta' - \eta| = \kappa$, the set $L^\circ \setminus N_{\epsilon_0}(\eta)$ is included in $\text{cone}(\eta, \epsilon_0)$. Note that L is included in $N_\Delta(\eta)$. Then, we see that $L^\circ \setminus N_{\epsilon_0}(\eta)$ is included in $\text{cone}(\eta, \epsilon_0) \cap N_\Delta(\eta)$. Comparing the volumes of both sets, we have

$$\text{vol}(L) - \text{vol}(N_{\epsilon_0}(\eta)) \leq V_d \Delta^d r(\eta, \epsilon_0).$$

Since the left side is not less than $\text{vol}(L) - \text{vol}(L)/2 = \text{vol}(L)/2$, we have

$$r(\eta, \epsilon\alpha) \geq \frac{\text{vol}(L)}{2(\text{diam}(L))^d V_d}$$

holds for all ϵ such that $\epsilon\alpha \in (0, \epsilon_0)$. *The proof is completed.*

APPENDIX K
CONTINUITY OF $\Phi(U_{K,\epsilon,n}^{(\alpha)})$

Lemma 8: Let λ_n denote the maximum of the largest eigenvalue of $J_n(\theta)$ among $\theta \in K$. For a certain $\epsilon > 0$, for all $r \leq \epsilon$, for all $\theta' \in B_r(\theta) \cap K$, and for all $\theta \in K$,

$$\frac{\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta'))}{\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta))}$$
$$\leq 1 + \frac{\sqrt{n\lambda_n}r}{\rho_n(\epsilon, \theta)\alpha} + \frac{C_d \text{diam}(K)\sqrt{n\lambda_n}\max\{g_{\theta,r}, 0\}}{\rho_n(\epsilon, \theta)\alpha}$$

holds, where $C_d = 2^{1-d/2}\Gamma(d/2)$ and

$$g_{\theta,r} = \max_{\theta' \in B_r(\theta) \cap K} ||J_n(\theta)^{-1/2}J_n(\theta')^{1/2}||_s - 1.$$

Proof: First assume that $J_n(\theta)$ does not depend on θ. Let J_n denote the value. Note that

$$U_{K,\epsilon,n}^{(\alpha)}(\theta') \setminus U_{K,\epsilon,n}^{(\alpha)}(\theta)$$
$$= N_{\sqrt{n}\epsilon}(0) \cap (\sqrt{n}/\alpha)J_n^{1/2}((K - \theta') \setminus (K - \theta))$$
$$= N_{\sqrt{n}\epsilon}(0) \cap (\sqrt{n}/\alpha)((L - \eta') \setminus (L - \eta)),$$

holds, where $L = J_n^{1/2}K$, $\eta = J_n^{1/2}\theta$, and $\eta' = J_n^{1/2}\theta'$. Then, we have

$$\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta') \setminus U_{K,\epsilon,n}^{(\alpha)}(\theta)) \leq \Phi((\sqrt{n}/\alpha)((L - \eta') \setminus (L - \eta))).$$

Recall that Φ is the d variate standard Gaussian measure. Note that $L - \eta'$ is given by displacing $L - \eta$ by $\eta - \eta'$. Hence, a point $\tilde{\eta}$ in $L - \eta$ has a representation as a scalar amount in the direction $\eta' - \eta$ together with $d - 1$ amounts in directions orthogonal to $\eta' - \eta$. The displacement from $L - \eta$ to $L - \eta'$ only affects that scalar translation by the amount $|\eta' - \eta|$. Hence, we have

$$\Phi\Big((\sqrt{n}/\alpha)\big((L - \eta') \setminus (L - \eta)\big)\Big)$$
$$\leq \Pr\{|y| \leq (\sqrt{n}/\alpha)|\eta' - \eta|/2\}$$
$$\leq (\sqrt{n}/\alpha)|\eta' - \eta|,$$

where y denotes the (univariate) standard Gaussian variable. Since $|\theta' - \theta| \leq r \leq \epsilon$ and since the eigenvalues of $J_n(\theta)$ are bounded by λ_n,

$$\Phi\Big((\sqrt{n}/\alpha)\big((L - \eta') \setminus (L - \eta)\big)\Big) \leq \frac{\sqrt{n\lambda_n}r}{\alpha}.$$

Recalling that $\Phi(U_{K,\epsilon,n}^{(\alpha)}(\theta)) \geq \rho_n(\epsilon, \theta)$, we have the claim of Lemma.

Next, consider the general $J_n(\theta)$ case. We have

$$J_n(\theta')^{1/2}(K - \theta') \setminus J_n(\theta)^{1/2}(K - \theta)$$
$$\subset J_n(\theta)^{1/2}(K - \theta') \setminus J_n(\theta)^{1/2}(K - \theta)$$
$$\cup J_n(\theta')^{1/2}(K - \theta') \setminus J_n(\theta)^{1/2}(K - \theta').$$

The quantity produced by the first component of the union in the last expression is bounded as in the same way as the constant J_n case. As for the second component, we have

$$J_n(\theta')^{1/2}(K - \theta') \setminus J_n(\theta)^{1/2}(K - \theta')$$
$$= J_n(\theta)^{1/2}\Big(\big(J_n(\theta)^{-1/2}J_n(\theta')^{1/2}(K - \theta')\big) \setminus (K - \theta')\Big)$$
$$\subset J_n(\theta)^{1/2}\Big(\big((1 + g_{\theta,r})(K - \theta')\big) \setminus (K - \theta')\Big)$$
$$= (1 + g_{\theta,r})J_n(\theta)^{1/2}(K - \theta') \setminus J_n(\theta)^{1/2}(K - \theta').$$

Let $L' = J_n(\theta)(K - \theta')$, then we have

$$J_n(\theta')^{1/2}(K - \theta') \setminus J_n(\theta)^{1/2}(K - \theta')$$
$$\subset (1 + g_{\theta,r})L' \setminus L'.$$

Hereafter, assume $g_{\theta,r} > 1$. Otherwise, the analysis below is not necessary, because $(1 + g_{\theta,r})L' \setminus L'$ is the empty set in that case.

Note that L', which is compact, includes the origin. Assume now that the origin is included in L'°, that is, $\theta' \in K^\circ$. Since L' is convex, the set $(1 + g_{\theta,r})L' \setminus L'$ is a shell outside L' as follows.

$$
\begin{aligned}
&(1 + g_{\theta,r})L' \setminus L' \\
&= \{yu : u \in \partial L', 1 < y \le 1 + g_{\theta,r}\} \\
&= \{y_u u/|u| : u \in \partial L', |u| < y_u \le |u|(1 + g_{\theta,r})\},
\end{aligned} \tag{93}
$$

where $\partial L' = L' - L'^\circ$. Note that $\{u/|u| : u \in \partial L'\}$ is the unit sphere, when the origin is in the interior of L'. The equality (93) itself holds even if the origin is not included in the interior of L'. Hence, we do not need the assumption $\theta' \in K^\circ$ for the argument below.

Now, we can evaluate

$$
\begin{aligned}
&\Phi\Big((\sqrt{n}/\alpha)\big(J_n(\theta')^{1/2}(K - \theta') \setminus J_n(\theta)^{1/2}(K - \theta')\big)\Big) \\
&\le \Phi\Big((\sqrt{n}/\alpha)\big((1 + g_{\theta,r})L' \setminus L'\big)\Big)
\end{aligned}
$$

Let ξ be a d-variate standard Gaussian variable. Then, we transform it to $(v, y) = (\xi/|\xi|, |\xi|)$. Here, v and y are independent of each other, the marginal density of v is uniform on the surface of the unit sphere, and that of y is

$$
C_d y^{d-1} \exp\Big(-\frac{y^2}{2}\Big) \le C_d(d-1)^{(d-1)/2} e^{-(d-1)/2},
$$

where $C_d = 2^{(1-d)/2}/\Gamma(d/2)$ is the normalization constant and the equality of the above holds if and only if $y = \sqrt{d-1}$. Since $|u|$ $(u \in \partial L')$ is upper bounded by $\mathrm{diam}(L')$, we have the following, where $y_0 = \sqrt{n}|u|/\alpha$ and $y_1 = \sqrt{n}|u|(1 + g_{\theta,r})/\alpha$.

$$
\begin{aligned}
&\Phi\Big((\sqrt{n}/\alpha)\big((1 + g_{\theta,r})L' \setminus L'\big)\Big) \\
&\le \max_{u \in \partial L'} \Pr\{y_0 < y \le y_1\} \\
&\le \max_{u \in \partial L'} C_d \int_{y_0}^{y_1} y^{d-1} \exp\Big(-\frac{y^2}{2}\Big) dy \\
&\le \max_{u \in \partial L'} C_d \int_{y_0}^{y_1} (d-1)^{(d-1)/2} e^{-(d-1)/2} dy \\
&\le C_d \,\mathrm{diam}(L')\sqrt{n} g_{\theta,r}/\alpha \\
&= C_d \mathrm{diam}(K)\sqrt{n\lambda_n} g_{\theta,r}/\alpha.
\end{aligned}
$$

This completes the proof.

APPENDIX L
SOME INEQUALITY FOR MODELS WITH HIDDEN VARIABLES (LEMMA 13)

Recall that we defined the model with hidden variable in (63) as

$$
p(x|\theta) = \int \kappa(x|y)q(y|\theta)\nu_y(dy),
$$

where $q(y|\theta)$ is the density of an exponential family. We have the following, where $G(\theta)$ is the Fisher information of $q(y|\theta)$.

Lemma 13: Given a data string x^n, let $\hat\theta$ denote the MLE for a model with hidden variables $p(x^n|\theta)$ defined by (63).

Then, the following holds for all $x^n \in \mathcal{X}^n$.

$$
\forall \theta \in \Theta, \ \frac{1}{n}\log\frac{p(x^n|\hat\theta)}{p(x^n|\theta)} \le D(q(\cdot|\hat\theta)\|q(\cdot|\theta)), \tag{94}
$$

$$
\forall \theta \in \Theta, \ \hat J(\theta, x^n) \le G(\theta). \tag{95}
$$

In particular, when $q(y|\theta)$ is the multinomial model, the following holds

$$
\frac{p(x^n|\hat\theta)}{p(x^n|\theta)} \le \exp(nD(q(\cdot|\hat\theta)\|q(\cdot|\theta))) = \prod_{y\in\mathcal{Y}} \frac{\hat\eta_y^{n\hat\eta_y}}{\eta_y^{n\hat\eta_y}}, \tag{96}
$$

where $\eta_y = q(y|\theta)$ and $\hat\eta_y = q(y|\hat\theta)$.

Proof: Note that

$$
\begin{aligned}
q(x^n|\theta) &= \prod_{t=1}^{n} \int \kappa(x_t|y_t)p(y_t|\theta)\nu_y(dy_t) \\
&= \int \prod_t \kappa(x_t|y_t)p(y_t|\theta)\nu_y(dy^n) \\
&= \int \kappa(x^n|y^n)p(y^n|\theta)\nu_y(dy^n).
\end{aligned}
$$

We have

$$
\begin{aligned}
\frac{q(x^n|\theta)}{q(x^n|\theta')} &= \frac{\int \kappa(x^n|y^n)p(y^n|\theta)\nu_y(dy^n)}{\int \kappa(x^n|y^n)p(y^n|\theta')\nu_y(dy^n)} \\
&= \int \frac{p(y^n|\theta)}{p(y^n|\theta')} \frac{\kappa(x^n|y^n)p(y^n|\theta')\nu_y(dy^n)}{\int \kappa(x^n|z^n)p(z^n|\theta')\nu_y(dz^n)}.
\end{aligned}
$$

Define $q(y^n|x^n, \theta')$ by

$$
q(y^n|x^n, \theta') = \frac{\kappa(x^n|y^n)p(y^n|\theta')}{\int \kappa(x^n|z^n)p(z^n|\theta')\nu_y(dz^n)},
$$

which is the posterior distribution of y^n given x^n provided x^n is drawn from $q(x^n|\theta')$.

Using it, we can write

$$
\frac{q(x^n|\theta)}{q(x^n|\theta')} = \int q(y^n|x^n, \theta')\frac{p(y^n|\theta)}{p(y^n|\theta')}\nu_y(dy^n).
$$

Then by Jensen's inequality, we have

$$
\begin{aligned}
&\frac{1}{n}\log\frac{q(x^n|\theta)}{q(x^n|\theta')} \\
&\ge \frac{1}{n}\int q(y^n|x^n, \theta')\log\frac{p(y^n|\theta)}{p(y^n|\theta')}\nu_y(dy^n).
\end{aligned} \tag{97}
$$

Let $f(\theta, \theta')$ denote the left side, and $g(\theta, \theta')$ the right side. Then, we have

$$
\forall \theta, \theta' \in \Theta, \ f(\theta, \theta') - g(\theta, \theta') \ge 0, \tag{98}
$$

where equality holds when $\theta = \theta'$. Hence, Hessian of the left side is semi positive-definite. That is, the matrix whose ij entry is

$$
\frac{\partial^2 \log f(\theta, \theta')}{\partial\theta_i\partial\theta_j} - \frac{\partial^2 \log g(\theta, \theta')}{\partial\theta_i\partial\theta_j} \tag{99}
$$

is semi positive definite. Note that

$$
g(\theta, \theta') = \theta\bar\eta^t - \psi(\theta) - (\theta'\bar\eta^t - \psi(\theta')), \tag{100}
$$

where

$$\bar{\eta} = \frac{1}{n} \int q(y^n|x^n, \theta') \sum_{t=1}^{n} y_t \nu_y(dy^n).$$

From (100), we have

$$-\frac{\partial^2 \log g(\theta, \theta')}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 \psi(\theta)}{\partial \theta_i \partial \theta_j} = G_{ij}(\theta).$$

Hence, semi positive-definiteness of (99) implies

$$\forall \theta \in \Theta, \ \ \hat{J}(\theta, x^n) \le G(\theta),$$

Plugging in $\hat{\theta}$ to θ' in (98) and noting $f(\theta, \hat{\theta}) \le 0$, we have

$$\forall \theta \in \Theta, \ 0 \ge f(\theta, \hat{\theta}) \ge g(\theta, \hat{\theta}), \tag{101}$$

where both inequality hold as equality, when $\theta = \hat{\theta}$. That is,

$$g(\theta, \hat{\theta}) \le g(\hat{\theta}, \hat{\theta}) = 0.$$

Together with (100) the following holds

$$g(\theta, \hat{\theta}) = \theta \bar{\eta}^t - \psi(\theta) - (\hat{\theta} \bar{\eta}^t - \psi(\hat{\theta})) \tag{102}$$

$$\le \hat{\theta} \bar{\eta}^t - \psi(\hat{\theta}) - (\hat{\theta} \bar{\eta}^t - \psi(\hat{\theta})) = 0, \tag{103}$$

which implies $\bar{\eta} = \hat{\eta}$. Here $\hat{\eta}$ denotes the coo responding value of expectation parameter η to $\hat{\theta}$.

Note that

$$g(\theta, \hat{\theta}) = -D(p(\cdot|\hat{\theta})||p(\cdot|\theta)),$$

where $D(p(\cdot|\hat{\theta})||p(\cdot|\theta))$ is the Kullback divergence from $p(y|\hat{\theta})$ to $p(y|\theta)$ defined as

$$D(p(\cdot|\hat{\theta})||p(\cdot|\theta)) = \int p(y|\hat{\theta}) \log \frac{p(y|\hat{\theta})}{p(y|\theta)} \nu_y(dy).$$

Hence from (97), we have

$$\frac{1}{n} \log \frac{q(x^n|\hat{\theta})}{q(x^n|\theta)} \le D(\hat{\theta}||\theta). \tag{104}$$

This completes the proof.

REFERENCES

[1] P. H. Algoet and T. M. Cover, "Asymptotic Optimality and Asymptotic Equipartition Properties of Log-Optimum Investment," *The Annals of Probability,* Vol. 16, No. 2, pp. 876-898, Apr. 1988.

[2] S. Amari, *Differential-geometrical methods in statistics* (*2nd pr.*), Lecture Notes in Statistics, Vol.28, Springer-Verlag, 1990.

[3] S. Amari and H. Nagaoka, *Methods of Information Geometry*, AMS & Oxford University Press, 2000.

[4] A. R. Barron, *Logically Smooth Density Estimation,* Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., 1985.

[5] A. R. Barron, J. Rissanen and B. Yu, "The minimum description length principle in coding and modeling," *IEEE trans. Inform. Theory,* Vol. 44 No. 6, pp. 2743 - 2760, 1998.

[6] A. R. Barron, T. Roos, and K. Watanabe. "Bayesian Properties of Normalized Maximum Likelihood and its Fast Computation," *Proc. IEEE International Symposium on Information Theory.* Honolulu, HI, June 2014. pp.1667-1671.

[7] A. R. Barron and J. Takeuchi, "Mixture models achieving optimal coding regret," *Proc. of 1998 Inform. Theory Workshop,* 1998.

[8] R. G. Bartle, *Elements of Integration,* Publisher: John Wiley & Sons Inc, 1966.

[9] R. G. Bartle, "An Extension of Egorov's Theorem," *The American Mathematical Monthly,* 87:8, 628-633, 1980.

[10] J. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. B.* 41. pp. 113-147, 1979.

[11] A. R. Barron, "The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem," *Annals of Probability*, Vol. 13, pp. 1292-1303, 1985.

[12] L. Brown, *Fundamentals of statistical exponential families*, Institute of Mathematical Statistics, 1986.

[13] B. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE trans. on IT,* vol. 36. no. 3, pp. 453-471, 1990.

[14] B. Clarke and A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *J. Statistical Planning and Inference,* 41:37-60, 1994.

[15] H. Cramèr, *Mathematical methods of statistics*, Princeton University Press, 1946.

[16] I. Csiszàr, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. of probability*, vol. 12, No. 3, pp. 768-793, 1984.

[17] L. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. 26, pp. 166-174, March 1980.

[18] M. Feder, N. Merhav and M. Gutman, "Universal prediction of individual sequences," *IEEE trans. Inform. Theory,* vol. 38, pp. 1258-1268, 1992.

[19] D. P. Foster, "Prediction in the worst case," *Ann. of Statist.,* vol. 19, pp. 1084-1090, 1991.

[20] Y. Freund, "Predicting a binary sequence almost as well as the optimal biased coin," *Proc. of the 9th Annual Workshp on Computational Learning Theory,* pp. 89-98, Morgan Kaufmann, 1996.

[21] R. M. Gray, *Probability, random processes, and ergodic properties,* Springer, 1988.

[22] R. M. Gray and J. C. Kieffer, "Asymptotically Mean Stationary Measures." *Ann. Probab.* 8 (5) 962 - 973, October, 1980.

[23] J. A. Hartigan, "Invariant prior distributions," *Ann. Math. Statist.,* vol. 35, pp. 836-845, 1964.

[24] J. A. Hartigan, "The asymptotically unbiased density," *Ann. Math. Statist.,* vol. 36, pp. 1137-1152, 1965.

[25] D. Haussler, "A general minimax result for relative entropy," *IEEE trans. Inform. Theory,* vol. 43, no. 4, pp. 1276-1280, 1997.

[26] D. Haussler and A. R. Barron, "How well do Bayes methods work for on-line prediction of $\{\pm 1\}$ values," *Proceedings 1992 NEC Symposium on Computation and Cognition,* Chapter 4.

[27] D. Haussler, J. Kivinen and M. K. Warmuth, "Tight worst-case loss bounds for prediction with expert advice," *IEEE trans. Inform. Theory,* vol. 44, 1998.

[28] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge: Cambridge Univ. Press. 1990.

[29] P. J. Huber, *Robust Statistics*, Wiley & Sons, Hoboken, NJ, USA, 2004.

[30] I. Ibragimov and R. Hasminski, *Statistical estimation; Asymptotic theory,* Springer, New York, 1981.

[31] H. Jeffreys, *Theory of probability, 3rd ed.* , Univ. of California Press, Berkeley, Cal, 1961.

[32] J. C. Kieffer, "A Counterexample to Perez's Generalization of the Shannon-McMillan Theorem," *Ann. Probab.* 1(2): 362-364, April, 1973.

[33] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory* (Springer Series in Statistics), Springer New York, 1986.

[34] K. Miyamoto and J. Takeuchi, "On MDL Estimation for Simple Contaminated Gaussian Location Families," in *Proc. of the 2020 International Symposium on Information Theory and Its Applications (ISITA 2020),* Octeber 2020.

[35] A. Perez. "Extensions of Shannon-McMillan's limit theorem to more general stochastic processes processes," In *Third Prague Conf. on Inform. Theory, Decision Functions,and Random Processes,* pp. 545.574, Prague and New York, 1964. Publishing House Czech. Akad. Sci. and Academic Press.

[36] D. Pollard, Online notes, http://web.archive.org/web/20160524113206/ http://www.stat.yale.edu/~pollard/Books/Asymptopia/, viewed on March 1, 2024.

[37] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. Inform. Theory,* vol. 40, pp. 40-47, 1996.

[38] W. R. Rudin, *Principles of Mathematical Analysis*, Third Edition, McGraw–Hill, 1976.

[39] Yu M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3-17, July 1988.

[40] J. Takeuchi and A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *Proc. of 1998 IEEE ISIT,* p. 1998.

[41] J. Takeuchi and A. R. Barron, "Asymptotically minimax regret for models with hidden variables," *Proc. of 2014 IEEE International Symposium on Information Theory,* pp. 3037-3041, Honolulu, HI, USA, June 29 - July 4, 2014.

[42] J. Takeuchi and A. R. Barron, "Stochastic Complexity for Tree Models," *Proc. of 2014 IEEE Information Theory Workshop,* pp. 223-227, Hobart, Tasmania, Australia, November 2-5, 2014.

[43] J. Takeuchi and A. R. Barron. Device, method, and medium for predicting a probability of an occurrence of data. US Patent Publication 6892164 B2. Also US6766280. Filed August 26, 2002. Granted. Published July 20, 2004 and May 10, 2005.

[44] J. Takeuchi, T. Kawabata, and A. R. Barron, "Properties of Jeffreys Mixture for Markov Sources," *of IEEE Transactions on Information Theory,* Vol. 59, No. 1, pp. 438-457, January 2013.

[45] J. Takeuchi and H. Nagaoka, "Information Geometry of The Family of Markov Kernels Defined by A Context Tree," in *Proc. of 2017 IEEE Information Theory Workshop,* Kaohsiung, Taiwan, November 6-10, 2017.

[46] V. Vovk, "Aggregating starategies," *Proc. 3rd Annual Workshop on Computational Learning Theory,* Morgan, Kaufmann, pp. 371-383, 1990.

[47] A. Wald, "Note on the Consistency of the Maximum Likelihood Estimate," *The Annals of Mathematical Statistics*, Vol. 20, No. 4, pp. 595-601, Dec., 1949.

[48] F. Willems, Y. Shtarkov and T. Tjalkens, "The context tree weighting method: basic properties," *IEEE trans. Inform. Theory,* vol. 41. no. 3, pp. 653-664, 1995.

[49] M. J. Weinberger, J. Rissanen, and M. Feder, "A universal finite memory source," *IEEE trans. Inform. Theory,* Vol. 41. No. 3, pp. 643-652, 1995.

[50] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE trans. Inform. Theory,* vol. 43, no. 2, pp. 646-657, 1997.

[51] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," *IEEE trans. Inform. Theory,* vol. 46, no. 2, pp. 431-445, 2000.

[52] J. Takeuchi, *Computational learning theory of probabilistic knowledge representations* (*in Japanese*)*,* Doctoral dissertation, Faculty of Engineering, Univ. of Tokyo, 1995.

[53] J. Takeuchi and S. Amari, "The α-parallel prior and its properties (in Japanese)," *Technical Report of IEICE,* 1996.