Improved MDL Estimators Using Local Exponential Family Bundles Applied to Mixture Families

Kohei Miyamoto
Kyushu University, Japan

Andrew R. Barron
Yale University, USA

Jun'ichi Takeuchi
Kyushu University, Japan

***Abstract*—The MDL estimators for density estimation, which are defined by two-part codes for universal coding, are analyzed. We give a two-part code for mixture families whose regret is close to the minimax regret, where regret of a code with respect to a target family $\mathcal{M}$ is the difference between the codelength of the code and the ideal codelength achieved by an element in $\mathcal{M}$. Our code is constructed using a probability density in an enlarged family of $\mathcal{M}$ (a bundle of local exponential families of $\mathcal{M}$) for data description. This result gives a tight upper bound on the risk of the MDL estimator defined by the two-part code, based on the theory introduced by Barron and Cover in 1991.**

I. INTRODUCTION

We consider estimators for a mixture family and give a new method to construct minimum description length (MDL) estimators [2], [4] for which the risk bound in terms of Rényi divergence is tight.

Given a data string $x^n = x_1x_2...x_n$ drawn from an element p_{θ^*} of a parametric family $\mathcal{M} = \{p_\theta(x)|\theta \in \Theta\}$, we call the following estimator as an α-MDL estimator in this paper.

$$\ddot{p} = \ddot{p}(\cdot; x^n) = \arg\min_{p\in\ddot{\mathcal{M}}}\Big(-\log p(x^n) + \alpha L_n(p)\Big), \quad (1)$$

where $\ddot{\mathcal{M}}$ is a countable set of densities and the function $L_n(p)$ is a codelength function defined over $\ddot{\mathcal{M}}$ satisfying the Kraft's inequality. Note that $\alpha L_n(p)$ is called parameter (model) description length and $-\log p(x^n)$ is called the data description length given by the model p. The right hand side of (1), denoted as $L_{\alpha,2\text{-}p}(x^n)$ is called the codelength of the α-two-part code associated with this α-MDL estimator. Note that $L_{\alpha,2\text{-}p}$ satisfies the Kraft's inequality over $\mathcal{X}^n$. Define

$$p_{\alpha,2\text{-}p}(x^n) = e^{-L_{\alpha,2\text{-}p}(x^n)} = \ddot{p}(x^n)e^{-\alpha L_n(\ddot{p})}.$$

Then, $p_{\alpha,2\text{-}p}$ is a sub-probability density over $\mathcal{X}^n$. Here, the extra factor α is a certain real number not less than 1. Since this codelength function is determined by α, $\ddot{\mathcal{M}}$ and L_n, we call the α-two-part code with this codelength function as α-two-part code $(\ddot{\mathcal{M}}, L_n)$. When $\alpha > 1$, the following inequality is known to hold [2], [4] for any $\lambda \in (0, 1 - 1/\alpha)$.

$$\mathrm{E}_{p_{\theta^*}} \bar{d}_\lambda(p_{\theta^*}||\ddot{p}) \leq \frac{1}{n}R_n(p_{\theta^*}, p_{\alpha,2\text{-}p}). \quad (2)$$

Here

$$R_n(p,q) = \mathrm{E}_p\big[-\log p(X^n) - \big(-\log q(X^n)\big)\big].$$

is the redundancy and $\bar{d}_\lambda$ is Rényi divergence [5], [10] of order λ. Note that p and q in the definition can be densities of non i.i.d. (sub) processes.

The fact means that the smaller the redundancy of a two-part code is, the smaller the risk of the MDL estimator induced by the two-part code is. It motivates us to pursue the two-part codes with as small redundancy as possible.

Though it is not well known compared to (2), the following holds provided X^n is drawn from p^* [3].

$$\Pr\Big\{\bar{d}_\lambda(p^*||\ddot{p}) - \frac{1}{n}\log\frac{p^*(X^n)}{p_{\alpha,2\text{-}p}(X^n)} \geq \epsilon\Big\} \leq e^{-\epsilon n/\alpha}. \quad (3)$$

If $p^* \in \mathcal{M}$, then $\log(p^*(X^n)/p_{\alpha,2\text{-}p}(X^n))$ is bounded upper by regret of $p_{\alpha,2\text{-}p}$ with respect to $(\mathcal{M}, x^n)$ defined as

$$\mathrm{REG}(p_{\alpha,2\text{-}p}, \mathcal{M}, x^n) = -\log p_{\alpha,2\text{-}p}(x^n) + \log\hat{p}(x^n)$$

where $\hat{p}$ is the maximum likelihood estimate in $\mathcal{M}$. Hence, (3) motivates us to pursue the minimax regret $\min_q \max_{x^n} \mathrm{REG}(q, \mathcal{M}, x^n)$ and the stochastic complexity. For various parametric families, the asymptotic evaluation of the minimax regret as below is known [6], [8], [9]:

$$\frac{K}{2}\log\frac{n}{2\pi} + \log\int_\Theta |J(\theta)|^{1/2}d\theta + o(1), \quad (4)$$

where K is the number of parameters, $J(\theta)$ is the Fisher information matrix of θ and $|J(\theta)|$ is its determinant.

In particular for exponential families, the value (4) is achieved by slightly modified Jeffreys mixtures, whereas for non-exponential families, usual Bayes mixtures cannot achieve it. For this problem, it is known that mixtures of enlarged families of the target $\mathcal{M}$ using local exponential family bundle [1] achieve (4), in particular for mixture families [8], [9], which are typical examples of non-exponential families.

The minimum worst case regret of two-part codes is larger than the minimax value, but it is shown in [4] that a sub-probability density $p_{\alpha,2\text{-}p}(x^n)$ is shown, whose regret is asymptotically bounded upper by

$$\alpha\Big(\frac{K}{2}\log n + \log\int_\Theta |J(\theta)|^{1/2}d\theta - K\log a + c\Big) + \frac{Ka^2}{8}, \quad (5)$$

when $\mathcal{M}$ is an exponential family. Here a and c are arbitrary positive constants. Note that the quantization of $\mathcal{M}$ here is related to Fisher information of θ. In fact, Grünwald shows it only for $\alpha = 1$, but this generalization is straightforward. Whether we can generalize this proposition to non-exponential families or not, was an open problem.

In this paper, for mixture families, we establish a new two-part code which achieves the almost same value as (5) by encoding the data string by a density in a local exponential family bundle of the target $\mathcal{M}$. Here, a local exponential family bundle is a mathematical notion of enlargement of a parametric model of probability densities to higher dimensional spaces, by which the model $\mathcal{M}$ is enlarged to the direction of second order derivatives of $\log p_\theta$. Our result is obtained based on the fact that there is a density with higher likelihood in the enlarged model than the maximum likelihood in $\mathcal{M}$ when the empirical Fisher information is different from the Fisher information. These two-part codes yield the α-MDL estimators which enjoy tight risk and probabilistic loss bounds, owing to (2) and (3). In our method, the given data is encoded by a density in the enlarged family, which is outside of $\mathcal{M}$ for the cases that the empirical Fisher information differs from Fisher information. This means that the estimate by our MDL estimator may be outside $\mathcal{M}$.

II. PRELIMINARIES

Let $\mathcal{M} = \{p_\theta | \theta \in \Theta \subset \mathbb{R}^K\}$ be a parametric family of probability densities over a certain measurable set $\mathcal{X}$, and let $p_\theta(x^n) = \prod_{t=1}^n p_\theta(x_t)$. For given x^n, we define the empirical Fisher information $\hat{J}(\theta; x^n)$ by

$$\hat{J}_{ij}(\theta; x^n) = -\frac{1}{n}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(x^n)$$

Note that $\hat{J}(\theta; x^n) = \sum_{t=1}^n \hat{J}(\theta; x_t)/n$. The Fisher information is defined as $J(\theta) = \mathrm{E}_{p_\theta}\hat{J}(\theta, X)$.

Let $\ddot{\mathcal{M}}$ be a countable set of densities, which may be independent of $\mathcal{M}$. For $n = 1, 2, ...$, let $L_n : \ddot{\mathcal{M}} \to (0, \infty)$ be a codelength function over $\ddot{\mathcal{M}}$ satisfies the *Kraft's inequality*: $\sum_{p\in\ddot{\mathcal{M}}} e^{-L_n(p)} \le 1$.

Since $p_{\alpha,2\text{-}p}(x^n) = \ddot{p}(x^n)e^{-\alpha L_n(\ddot{p})}$, we have

$$\mathrm{REG}(p_{\alpha,2\text{-}p}, \mathcal{M}, x^n) = -\log\frac{\ddot{p}(x^n)}{p_{\hat{\theta}}(x^n)} + \alpha L_n(\ddot{p}).$$

In the chapter 10 of [4], for exponential families, Grünwald gives $\ddot{\Theta}$, which defines $\ddot{\mathcal{M}} = \{p_\theta \mid \theta \in \ddot{\Theta}\}$. He also gives the codelength function L_n such that for all $\theta \in \ddot{\Theta}$,

$$L_n(\theta) = \frac{K}{2}\log n + \log\int_\Theta |J(\theta)|^{1/2}d\theta - K\log a + o(1), \quad (6)$$

$$-\log\frac{p_{\ddot{\theta}}(x^n)}{p_{\hat{\theta}}(x^n)} \le \frac{Ka^2}{8} + o(1)$$

hold, where $a > 0$ is a certain constant. The quantization is done as follows, more detail is in [4]. We first construct a partition of Θ with hypercubes with side length $an^{-1/4}$. Let $\mathcal{S}$ represent this partition i.e. $\Theta \subseteq \cup_{S\in\mathcal{S}}S$, and for each $S \in \mathcal{S}$, θ_S be the centroid of $S \cap \Theta$. Then, we partition each $S \in \mathcal{S}$ into hyper-rectangles by using Fisher information $J(\theta_S)$. Then, we take quantized points based on these hyper-rectangles. For given $\theta \in \Theta$, we take the nearest point in $\ddot{\Theta}$ as a quantized point for it.

The important properties of this quantization are as follows. For all $S \in \mathcal{S}$, for all $\theta \in S$,

$$|\theta - \theta_S| \le \sqrt{K}an^{-1/4}. \quad (7)$$

For all $S \in \mathcal{S}$ and $\theta \in S$,

$$(\theta' - \theta)^T J(\theta_S)(\theta' - \theta) \le \frac{Ka^2}{4n}, \quad (8)$$

where θ' is an arbitrary point which is in the same hyper-rectangle as θ in quantization. This implies that, if θ and θ' belong to the identical hyper-rectangle, $|\theta - \theta'| \le \zeta^{-1/2}K^{1/2}n^{-1/2}a/2$. (Recall that ζ is a lower bound on eigenvalues of $J(\theta)$.)

Barron and Cover [2] gave a guarantee for the statistical risk of MDL estimators. In their theory, *Rényi divergence* of the order $\lambda \in (0, 1)$ [5], [10]:

$$\bar{d}_\lambda(p||q) = -\frac{1}{1-\lambda}\log\mathrm{E}_p\Big(\frac{q(X)}{p(X)}\Big)^{1-\lambda},$$

which can be regarded as a generalized version of the KL divergence, is employed to measure the performance of estimators. The following theorem is well known.

Theorem 1. *(Barron and Cover 1991, Grünwald 2007) Let $\alpha > 1$. Consider an arbitrary countable set of densities $\ddot{\mathcal{M}}$. Let L_n be an arbitrary codelength function of some prefix code on $\ddot{\mathcal{M}}$. Then, for all $\lambda \in (0, 1 - 1/\alpha]$, the MDL estimator $\ddot{p}$ of the α-two-part code $(\ddot{\mathcal{M}}, L_n)$ satisfies*

$$\mathrm{E}_{X^n\sim p^*}\bar{d}_\lambda(p^*||\ddot{p}) \le \frac{1}{n}R_n(p^*, p_{\alpha,2\text{-}p}).$$

This theorem is the version by Grünwald in the p. 478 of [4] with some modification for the notation. The original version is proved by Barron and Cover in [2]. Further, the following stochastic upper bound for the loss of MDL estimator [3] is known.

Theorem 2. *(Chatterjee and Barron 2014) Let $\alpha > 1$. Consider an arbitrary countable set of densities $\ddot{\mathcal{M}}$. Let L_n be an arbitrary codelength function of some prefix code on $\ddot{\mathcal{M}}$. Then, for all $\lambda \in (0, 1 - 1/\alpha]$ and $\epsilon > 0$, the MDL estimator $\ddot{p}$ of the α-two-part code $(\ddot{\mathcal{M}}, L_n)$ satisfies*

$$\Pr\big\{\bar{d}_\lambda(p^*||\ddot{p}) - \frac{1}{n}\log\frac{p^*(X^n)}{p_{\alpha,2\text{-}p}(X^n)} \ge \epsilon\big\} \le e^{-\epsilon n/\alpha}.$$

Finally, we introduce the definition of the mixture families and review some properties discussed in [9]. Let $q_0, q_1, \cdots, q_K$ be known densities on $\mathcal{X}$. The *mixture family* with components $q_0, q_1, \cdots, q_K$ is defined as follows.

$$\mathcal{M} = \Big\{p_\theta \;\Big|\; \theta \in \Theta,\ p_\theta(x) = \sum_{k=0}^K \theta_k q_k(x)\Big\},$$

$$\Theta = \Big\{\theta = (\theta_1, \cdots, \theta_K) \in [0, 1]^K \;\Big|\; \sum_{k=1}^K \theta_k \le 1\Big\},$$

where $\theta_0 = 1 - \sum_{k=1}^K \theta_k$. We assume that each q_i is different from one another in terms of KL-divergence. Then, the minimum eigenvalue of $J(\theta)$ for all $\theta \in \Theta$ is bounded below by a certain constant $\zeta > 0$.

Let $0 < \tau \leq 1/2$. To avoid parameters on the boundary of Θ, we introduce a subset $\Theta_\tau \in \Theta$. Θ_τ is a set of θ whose θ_k are in $[\tau, 1-\tau]$ for all $0 \leq k \leq K$. We also introduce a subset of the model $\mathcal{M}_\tau = \{p_\theta \mid \theta \in \Theta_\tau\}$.

The matrix $\hat{J}$ is always positive-semidefinite. It implies that $-\log p_\theta(x)$ is convex as a function of θ. Further, for all $\theta \in \Theta_\tau$, x and $1 \leq i, j \leq K$,

$$|\hat{J}_{ij}(\theta; x)| \leq \frac{1}{\tau^2} \tag{9}$$

and hence, for all $\theta_1, \theta_2 \in \Theta_\tau$ and all x^n, the following two inequalities hold.

$$e^{-2\sqrt{K}|\theta_1-\theta_2|/\tau} \leq \frac{z^T \hat{J}(\theta_1; x^n) z}{z^t \hat{J}(\theta_2; x^n) z} \leq e^{2\sqrt{K}|\theta_1-\theta_2|/\tau}, \tag{10}$$

$$e^{-2\sqrt{K}|\theta_1-\theta_2|/\tau} \leq \frac{z^T J(\theta_1) z}{z^t J(\theta_2) z} \leq e^{2\sqrt{K}|\theta_1-\theta_2|/\tau}. \tag{11}$$

III. TWO-PART CODES FOR MIXTURE FAMILIES

In this section, we fix a mixture family $\mathcal{M}$ with its components $q_0, q_1, \cdots, q_K$. Let $\hat{\theta}_\tau(x^n)$ be the maximum likelihood estimate of θ in Θ_τ given x^n. Then, the regret related to $\mathcal{M}_\tau$ is given as $\mathrm{REG}(p_{\alpha,2\text{-}p}, \mathcal{M}_\tau, x^n) = L_{\alpha,2\text{-}p}(x^n) + \log p_{\hat{\theta}_\tau}(x^n)$. We will prove the following theorem.

Theorem 3. *Let $a > 0$, $\alpha > 1$, $\tau \in (0, 1/2)$ and $\delta \in (0,1)$. Then for the mixture family, there exists an α-two-part code, such that the following uniformly holds for all x^n,*

$$\begin{aligned}&\mathrm{REG}(p_{\alpha,2\text{-}p}, \mathcal{M}_\tau, x^n)\\ &\quad\leq \alpha\Big(\frac{K}{2}\log n + \log \int_{\Theta_\tau} |J(\theta)|^{1/2} d\theta - K \log a + c\Big)\\ &\qquad\qquad + f(n) + o(1),\end{aligned}$$

where c is an arbitrary small constant and

$$f(n) = \frac{Ka^2}{8}(1 + K\delta) e^{\frac{2}{\tau}\sqrt{K} O(n^{-1/4})}.$$

Remark. *If $p^* \in \mathcal{M}_\tau$, the following inequality holds for all x^n.*

$$-\log \frac{p_{\alpha,2\text{-}p}(x^n)}{p^*(x^n)} \leq \mathrm{REG}(p_{\alpha,2\text{-}p}, \mathcal{M}_\tau, x^n).$$

Therefore, Theorem 3 together with Theorems 1 and 2 implies the risk bound and the stochastic loss bound as follows.

$$\mathrm{E}_{X^n \sim p^*} \bar{d}_\lambda(p^* || p_{\ddot{\theta}(X^n)}) \leq \frac{1}{n} \overline{\mathrm{REG}},$$

$$\Pr\Big\{\bar{d}_\lambda(p^* || p_{\ddot{\theta}(X^n)}) - \frac{1}{n}\overline{\mathrm{REG}} \geq \epsilon\Big\} \leq e^{-\epsilon n/\alpha},$$

where $\overline{\mathrm{REG}}$ is the right side of the inequality in Theorem 3.

The code is designed by the quantization of Θ_τ we wrote in the previous section and using the idea of model enlargement via a local exponential family bundle. The idea of using a local exponential family bundle to design codes for non-exponential families in the MDL context was introduced in the literature for Bayes codes [9]. This idea allows us to achieve a small loss for the data description by encoding with a density in enlarged models. Let $\ddot{\Theta}_\tau$ be the quantization parameter space Θ_τ, L_n be the codelength function on $\ddot{\Theta}_\tau$ which satisfies (6). Define

$$V(\theta; x) = J^{-1/2}(\theta) \hat{J}(\theta; x) J^{-1/2}(\theta) - I$$

and the enlarged model $\bar{\mathcal{M}}_\tau$ as

$$\bar{\mathcal{M}}_\tau = \{\bar{p}_{\theta,\xi} \mid \theta \in \Theta_\tau,\ \xi \in R^{K\times K}\},$$

where

$$\bar{p}_{\theta,\xi}(x) = p_\theta(x) e^{\xi \cdot V(\theta;x) - \psi_\theta(\xi)},$$

$$\psi_\theta(\xi) = \log \int p_\theta(x) e^{\xi \cdot V(\theta;x)} dx.$$

Note that if $\xi = 0$, $\bar{p}_{\theta,\xi}$ corresponds to p_θ and that since $V(\theta; x^n) = \sum_{t=1}^n V(\theta; x_t)/n$ for a data string x^n,

$$\bar{p}_{\theta,\xi}(x^n) = p_\theta(x^n) e^{n\left(\xi \cdot V(\theta;x^n) - \psi_\theta(\xi)\right)}.$$

For fixed θ, the form of $\bar{p}_{\theta,\xi}$ can be regarded as an exponential family with the natural parameter ξ. This definition enlarges the original model $\mathcal{M}_\tau$ at each points $\theta \in \Theta_\tau$ with an exponential family $\bar{p}_{\theta,\xi}$. Since $V(\theta; x)$ consists of second derivatives of $\log p_\theta$, this enlarged model is an example of local exponential family bundles [1].

A. Proof of Theorem 3 for $\hat{\theta} \in \Theta_\tau$

We use $\bar{p}_{\theta,\xi}$ to encode x^n. This means that the estimate of our MDL estimator may be outside $\mathcal{M}$. We need to encode ξ similarly to θ. Since $\xi \in R^{K\times K}$ is a real matrix, it seems that we need large codelength to do this. However, indeed, we use only $\xi = 0$ or ξ with single non-zero value using some constant u. It gives a particular quantization for ξ, which we denote as $\ddot{\Xi}$. Using it, we can encode ξ with small codelength. Let $\bar{L}(\xi)$ be the codelength function for ξ and included in the parameter description length.

We first give a proof for the cases $\hat{\theta}(x^n) \in \Theta_\tau$. Let $\ddot{\hat{\theta}}(x^n) \in \ddot{\Theta}_\tau$ be the quantized point for $\hat{\theta}$. Let $\|A\|_s$ be the maximum absolute value of eigenvalues of a matrix A, similarly, $\|A\|_M$ be the maximum absolute value of elements of a matrix A. It is known that following inequality holds. For $A \in R^{K\times K}$,

$$\|A\|_M \leq \sqrt{K}\|A\|_s \leq K\sqrt{K}\|A\|_M. \tag{12}$$

Using $\delta \in (0,1)$, we define the following two sets of data strings G and G^c.

$$G = \{x^n \mid \hat{\theta}(x^n) \in \Theta_\tau,\ \|V(\hat{\theta}; x^n)\|_M \leq \delta\},$$

$$G^c = \{x^n \mid \hat{\theta}(x^n) \in \Theta_\tau,\ \|V(\hat{\theta}; x^n)\|_M > \delta\}.$$

Here, $x^n \in G$ implies that $\hat{J}(\hat{\theta}; x^n)$ is close to $J(\hat{\theta})$. For this case, we let $\ddot{\xi} = 0$, i.e. we use p_θ to encode the data. In the case $x^n \in G^c$, we let $\ddot{\xi} \neq 0$ as follows. First note that we can prove the following, similarly as in [9],

$$\|V(\ddot{\hat{\theta}}; x^n)\|_M > \frac{\delta}{2K\sqrt{K}}, \tag{13}$$

if $\hat{\theta}$ and $\ddot{\hat{\theta}}$ is sufficiently close. Precisely, (13) holds under the conditions; 1) the norm $\|V(\hat{\theta};x^n)\|_s$ is given by a positive eigenvalue of $V(\hat{\theta};x^n)$ and

$$|\hat{\theta}-\ddot{\hat{\theta}}| \leq \frac{\delta\tau}{8(K+\sqrt{K}\delta)}, \qquad (14)$$

or 2) the norm $\|V(\hat{\theta};x^n)\|_s$ is given by a negative eigenvalue of $V(\hat{\theta};x^n)$ and

$$|\hat{\theta}-\ddot{\hat{\theta}}| \leq \frac{\delta\tau}{8(K-\sqrt{K}\delta)}. \qquad (15)$$

Since $|\hat{\theta}-\ddot{\hat{\theta}}|$ is bounded by $O(n^{-1/2})$, these conditions are satisfied when n is sufficiently large. Let (i,j) be the index which satisfies $|V_{ij}(\ddot{\hat{\theta}};x^n)| = \|V(\ddot{\hat{\theta}};x^n)\|_M$. We let $\ddot{\xi}_{kl} = 0$ for $(k,l) \neq (i,j)$, and let $\ddot{\xi}_{ij} = u$ or $\ddot{\xi}_{ij} = -u$ so that $\ddot{\xi}\cdot V(\ddot{\hat{\theta}};x^n) = u\|V(\ddot{\hat{\theta}};x^n)\|_M$, where u is a certain positive constant. To encode $\ddot{\xi}$, we first indicate whether $x^n \in G$ or $x^n \in G^c$. When $x^n \in G$, we need only this codelength to encode $\ddot{\xi}$. When $x^n \in G^c$, we, next, encode the index of non-zero element (i,j) and the signature of $\ddot{\xi}_{ij}$. Since u is a constant independent of the data, we need not describe its value. Hence, we can define

$$\bar{L}(\xi) = \begin{cases} -\log r, & \text{when } \xi = 0, \\ -\log(1-r) + \log K^2 + \log 2, & \text{when } \xi \neq 0, \end{cases}$$

over $\ddot{\Xi}$, where r is an arbitrary value in $(0,1)$. We can denote the MDL estimator of the code as

$$(\ddot{\theta},\ddot{\xi}) = \arg\min_{(\theta,\xi)\in\ddot{\Theta}\times\ddot{\Xi}} \Big(-\log \bar{p}_{\theta,\xi}(x^n) + \alpha\big(L_n(\theta) + \bar{L}(\xi)\big)\Big).$$

Note that, even if the estimate does not achieve the minimum, Theorems 1 and 2 hold, and the risk and loss are bounded upper by the achieved redundancy.

Denoting $\xi\cdot V(x^n;\theta) - \psi_\theta(\xi)$ as $g_\theta(\xi)$, the regret of our code is given as follows.

$$\begin{aligned}&\mathrm{REG}(p_{\alpha,2\text{-}p},\mathcal{M},x^n)\\ &\quad = -\log\frac{p_{\ddot{\theta}}(x^n)}{p_{\hat{\theta}}(x^n)} - ng_{\ddot{\theta}}(\ddot{\xi}) + \alpha\big(L_n(\ddot{\theta}) + \bar{L}(\ddot{\xi})\big). \qquad (16)\end{aligned}$$

Note that the regret related to $\mathcal{M}_\tau$ is not larger than the above. From (6), we have already obtained the L_n term with the form in Theorem 3. We will give an upper bound for the other terms. Since $(\ddot{\theta},\ddot{\xi})$ minimizes the codelength, it is sufficient to evaluate an upper bound for $-\log(p_{\ddot{\hat{\theta}}}(x^n)/p_{\hat{\theta}}(x^n)) - ng_{\ddot{\hat{\theta}}}(\xi) + \alpha\bar{L}(\xi)$.

First, we give the proof for the case $x^n \in G$. From the Taylor expansion of $-\log p_{\ddot{\hat{\theta}}}(x^n)$ around $\hat{\theta}$, there exists a dividing point θ' between $\ddot{\hat{\theta}}$ and $\hat{\theta}$ such that

$$-\log\frac{p_{\ddot{\hat{\theta}}}(x^n)}{p_{\hat{\theta}}(x^n)} = \frac{n}{2}(\ddot{\hat{\theta}}-\hat{\theta})^T \hat{J}(\theta';x^n)(\ddot{\hat{\theta}}-\hat{\theta}). \qquad (17)$$

We want to bound the right hand side. For $z \in R^K \setminus \{0\}$, let $z' = J^{1/2}z$. By definitions of $V(\theta;x^n)$ and $\|V\|_s$, we have

$$\begin{aligned} z^T\hat{J}(\theta;x^n)z &= z^TJ(\theta)z\big(1 + \frac{z'^TV(\theta;x^n)z'}{z'^Tz'}\big)\\ &\leq z^TJ(\theta)z\big(1 + \|V(\theta;x^n)\|_s\big).\end{aligned}$$

Therefore, from (17), (10) and (11), denoting $\ddot{\hat{\theta}}-\hat{\theta}$ as z, we have

$$\begin{aligned} &-\log\frac{p_{\ddot{\hat{\theta}}}(x^n)}{p_{\hat{\theta}}(x^n)}\\ &\quad\leq \frac{n}{2}e^{\frac{2}{\tau}\sqrt{K}|z|}z^T\hat{J}(\hat{\theta};x^n)z\\ &\quad\leq \frac{n}{2}e^{\frac{2}{\tau}\sqrt{K}|z|}z^TJ(\hat{\theta})z\big(1+\|V(\hat{\theta})\|_s\big)\\ &\quad\leq \frac{n}{2}e^{\frac{2}{\tau}\sqrt{K}(|z|+|\hat{\theta}-\theta_S|)}z^TJ(\theta_S)z\big(1+\|V(\hat{\theta})\|_s\big),\end{aligned}$$

where S is the hypercube used in the quantization with $\hat{\theta} \in S$ and $V(\hat{\theta})$ is an abbreviation for $V(\hat{\theta};x^n)$.

Therefore, from (7), (8) and (12), for $x^n \in G$, we have

$$-\log\frac{p_{\ddot{\hat{\theta}}}(x^n)}{p_{\hat{\theta}}(x^n)} \leq \frac{Ka^2}{8}e^{\frac{2}{\tau}\sqrt{K}O(n^{-1/4})}(1+K\delta). \qquad (18)$$

This is the form in Theorem 3. Since $\ddot{\xi} = 0$, i.e. $g(\ddot{\xi}) = 0$ for $x^n \in G$, we have for $x^n \in G$,

$$\begin{aligned}&\mathrm{REG}(p_{\alpha,2\text{-}p},\mathcal{M},x^n)\\ &\quad\leq \alpha\Big(\frac{K}{2}\log n + \log\int_\Theta |J(\theta)|^{1/2}d\theta - K\log a - \log r\Big)\\ &\qquad + \frac{Ka^2}{8}(1+K\delta)e^{\frac{2}{\tau}\sqrt{K}O(n^{-1/4})} + o(1). \qquad (19)\end{aligned}$$

Next, we give the proof for $x^n \in G^c$. In this case, since $\|V\|_s$ is larger than δ, we can not use the same method as the previous case. However, similarly to (18), we have

$$-\log\frac{p_{\ddot{\hat{\theta}}}(x^n)}{p_{\hat{\theta}}(x^n)} \leq \frac{Ka^2}{8}e^{\frac{2}{\tau}\sqrt{K}O(n^{-1/4})}\big(1+K\|V(\ddot{\hat{\theta}};x^n)\|_M\big). \qquad (20)$$

Here, $\|V(\ddot{\hat{\theta}};x^n)\|_M$ may be large, but in that case $-ng_{\ddot{\theta}}(\xi)$ in (16) can help. We evaluate the lower bound on $g_{\ddot{\hat{\theta}}}(\ddot{\xi}) = \ddot{\xi}\cdot V(x^n;\ddot{\hat{\theta}}) - \psi_{\ddot{\hat{\theta}}}(\ddot{\xi})$, and prove that the term $-ng_{\ddot{\hat{\theta}}}(\ddot{\xi})$ defeats the $\|V\|_M$ term in (20).

Recall that (i,j) is the index of the non-zero element of $\ddot{\xi}$. Then $\ddot{\xi}\cdot V(\theta;x)$ equals $\ddot{\xi}_{ij}V_{ij}(\theta;x)$. By definition of $\psi_\theta(\xi)$, we have

$$\begin{aligned}\left.\frac{\partial\psi_\theta(\xi)}{\partial\xi_{ij}}\right|_{\xi_{ij}=0} &= \mathrm{E}_{\bar{p}_{\theta,0}}V_{ij}(\theta;X)\\ &= \big(\mathrm{E}_{p_\theta}V(\theta;X)\big)_{ij} = 0,\\ \frac{\partial^2\psi_\theta(\xi)}{\partial\xi_{ij}^2} &= \mathrm{E}_{\bar{p}_{\theta,\xi}}V_{ij}(\theta;X)^2 - \big(\mathrm{E}_{\bar{p}_{\theta,\xi}}V_{ij}(\theta;X)\big)^2\\ &\leq \mathrm{E}_{\bar{p}_{\theta,\xi}}V_{ij}(\theta;X)^2.\end{aligned}$$

Therefore, by Taylor expansion around $\xi_{ij} = 0$, there exists $\xi' \in R^{K\times K}$ such that

$$\psi_\theta(\xi) \le \frac{u^2}{2}\mathrm{E}_{\bar{p}_{\theta,\xi'}} V_{ij}(X;\theta)^2.$$

Recall that the minimum eigenvalue of $J(\theta)$ is not less than ζ for all $\theta \in \Theta$. Then, from (9) and (12), we can prove that

$$\|J^{-1/2}(\theta)\hat{J}(\theta;x)J^{-1/2}(\theta)\|_M \le \frac{K\sqrt{K}}{\zeta\tau^2}$$

for all x and $\theta \in \Theta_\tau$. This implies that

$$\|V(\theta;x)\|_M \le \frac{K\sqrt{K}}{\zeta\tau^2} + 1.$$

Let $B = (K\sqrt{K}/\zeta\tau^2 + 1)^2$. Then, we have $\psi_\theta(\xi) \le u^2 B/2$. Therefore,

$$\begin{aligned} g_{\ddot{\hat{\theta}}}(\ddot{\xi}) &\ge u\|V(\ddot{\hat{\theta}};x^n)\|_M - \frac{B}{2}u^2 \\ &\ge u\|V(\ddot{\hat{\theta}};x^n)\|_M\left(1 - \frac{B}{2\|V(\ddot{\hat{\theta}};x^n)\|_M}u\right) \\ &> u\|V(\ddot{\hat{\theta}};x^n)\|_M\left(1 - K\sqrt{K}\frac{B}{\delta}u\right) \end{aligned}$$

holds for sufficiently large n, where the last inequality follows from (13). Now assume $0 < u < \delta/K\sqrt{K}B$. In particular, let $u = \delta/2K\sqrt{K}B$. Then we have,

$$g_{\ddot{\hat{\theta}}}(\ddot{\xi}) > \frac{\delta}{4K\sqrt{K}B}\|V(\ddot{\hat{\theta}};x^n)\|_M. \qquad (21)$$

Therefore, we have the following inequality for sufficiently large n and all $x^n \in G^c$.

$$\begin{aligned} -\log\frac{\bar{p}_{\ddot{\hat{\theta}},\ddot{\xi}}(x^n)}{p_{\hat{\theta}}(x^n)} &\le \frac{Ka^2}{8}e^{\frac{2}{\tau}\sqrt{K}O(n^{-1/4})} \\ &- n\|V(\ddot{\hat{\theta}};x^n)\|_M\left(\frac{\delta}{4K\sqrt{K}B} - \frac{Ka^2}{8n}e^{\frac{2}{\tau}\sqrt{K}O(n^{-1/4})}\right). \end{aligned}$$

Since the right hand side of this diverges to negative infinity as $n \to \infty$, the codelength for $\ddot{\xi} \neq 0$, which equals $\bar{L}(\ddot{\xi}) = -\log(1-r) + \log K^2 + \log 2$, is negligible for an arbitrary r, when n is large. It implies that the term $\bar{L}(0) = -\log r$ in (19) can be set to an arbitrarily small constant c.

B. Proof for the case of $\hat{\theta} \notin \Theta_\tau$

In this case, we can not use the technique using (10) and (11) at $\hat{\theta}$. However, from the fact that $-\log p_\theta(x^n)$ is convex, when $\hat{\theta} \notin \Theta_\tau$, $\hat{\theta}_\tau$ is always on the boundary of Θ_τ. In this paper, we prove only for the simplest case $K = 1$. When $K = 1$ i.e. $\Theta_\tau = [\tau, 1-\tau]$, $\hat{\theta}_\tau$ is τ or $1-\tau$. Therefore, in this case, we first indicate the fact that $\hat{\theta}(x^n) \notin \Theta_\tau$, then describe either $\hat{\theta}_\tau$ is τ or $1-\tau$, and finally encode the data using $-\log p_{\hat{\theta}_\tau}(x^n)$ nats. The codelength to indicate the fact $\hat{\theta} \notin \Theta_\tau$ can be an arbitrary small constant c' similarly to $\bar{L}(0)$ for the case $\hat{\theta} \in \Theta_\tau$, and the codelength to describe $\hat{\theta}_\tau$ can be designed to be $\log 2$ nats. Therefore, the regret of this case is a constant $c' + \log 2$. Comparing to the regret of the case $x^n \in G$, this is negligible for large n. We have proved Theorem 3 for $K = 1$.

For general cases $K \ge 2$, we give only the outline. Similarly to the case $K = 1$, we fix to τ each elements of $\hat{\theta}$ and $\hat{\theta}_0$, which are smaller than τ. We can indicate these fixed elements with $(K+1)\log 2$ nats. Let K' be the number of elements which are not fixed. Then, the rest part of $\hat{\theta}$ is on a K'-dimensional subspace of Θ_τ. By quantizing this subspace with the same way for Θ_τ, we can encode the data with the regret related to $\mathcal{M}_\tau$, whose main term is $(\alpha K'/2)\log n$. Since $K' < K$, the regret for x^n such that $\hat{\theta} \in \Theta_\tau$, whose main term is $(\alpha K/2)\log n$, is dominant for large n. Therefore, we have Theorem 3.

C. Parameters used for the code

We use parameters τ and δ to design a code. We want to make these parameters small, since large δ makes the regret bound loose and since large τ makes the range of the true density small.

These parameters can be designed to decrease as the data size n increases. Let $r_1, r_2 > 0$ and $\tau_n = n^{-r_1}/2,\ \delta_n = n^{-r_2}$. In (19), consider $O(n^{-1/4})/\tau_n$ in the exponent, this have to converge to 0 as $n \to \infty$. Hence, we should design τ_n to be larger than $n^{-1/4}$, i.e. $0 < r_1 < 1/4$. Further, in (14) and (15), $\delta_n\tau_n$ have to decrease slower than $|\hat{\theta} - \ddot{\hat{\theta}}| = O(n^{-1/2})$. Therefore, we should let $r_2 < 1/2 - r_1$. Moreover, from (21), since $ng_{\ddot{\hat{\theta}}}(\xi)$ have to diverge, we should design τ_n and δ_n so that $n\delta/B$ diverges. Therefore, we should let $r_2 < 1 - 4r_1$. In conclusion, we should design τ_n and δ_n as $0 < r_1 < 1/4$ and $0 < r_2 < \min\{1/2 - r_1, 1 - 4r_1\}$.

IV. Concluding Remark

There still remains a problem of proving the upper bound of the regret related to the original model $\mathcal{M}$ for all x^n.

Acknowledgment

This research was partially supported by JSPS KAKENHI Grant Number 18H03291.

References

[1] S. Amari and H. Nagaoka, *Methods of information geometry*. American Mathematical Soc., 2007, vol. 191.

[2] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. on Inf. Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.

[3] S. Chatterjee and A. Barron, "Information theory of penalized likelihoods and its statistical implications," *arXiv:1401.6714v2*, 2014.

[4] P. D. Grünwald, *The minimum description length principle*. MIT press, 2007.

[5] A. Rényi, "On measures of entropy and information," in *Proc. of the Fourth Berkeley Symp. on Math. Stat. and Prob., vol. 1*, Berkeley, Calif., 1961, pp. 547–561.

[6] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. on Inf. Theory*, vol. 42, no. 1, pp. 40–47, 1996.

[7] Y. M. Shtar'kov, "Universal sequential coding of single messages," *Problemy Peredachi Informatsii*, vol. 23, no. 3, pp. 3–17, 1987.

[8] J. Takeuchi and A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," in *Proc. 1998 IEEE Intl. Symp. on Inf. Theory*, 1998, p. 318.

[9] ——, "Asymptotically minimax regret for models with hidden variables," in *Proc. 2014 IEEE Intl. Symp. on Inf. Theory*, 2014, pp. 3037–3041.

[10] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. on Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.