On the Convergence of MDL Density Estimation

Tong Zhang

IBM T.J. Watson Research Center
Yorktown Heights, NY, 10598, USA
tzhang@watson.ibm.com

Abstract. We present a general information exponential inequality that measures the statistical complexity of some deterministic and randomized density estimators. Using this inequality, we are able to improve classical results concerning the convergence of two-part code MDL in [1]. Moreover, we are able to derive clean finite-sample convergence bounds that are not obtainable using previous approaches.

1 Introduction

The purpose of this paper is to study a class of complexity minimization based density estimation methods using a generalization of ϵ-entropy which we call KL-complexity. Specifically, we derive a simple yet general information theoretical inequality that can be used to measure the convergence behavior of some randomized estimation methods. Consequences of this very basic inequality will then be explored. In particular, we apply this analysis to the two-part code MDL density estimator studied in [1], and refine their results.

We shall first introduce basic notations used in the paper. Consider a sample space $\mathcal{X}$ and a measure μ on $\mathcal{X}$ (with respect to some σ-field). In statistical inferencing, the nature picks a probability measure Q on $\mathcal{X}$ which is unknown. We assume that Q has a density q with respect to μ. In density estimation, the statistician considers a set of probability densities $p(\cdot|\theta)$ (with respect to μ on $\mathcal{X}$) indexed by $\theta \in \Gamma$.[1] Throughout this paper, we always denote the true underlying density by q, which may not belong to the model class Γ. Given Γ, the goal of the statistician is to select a density $p(\cdot|\theta) \in \Gamma$ based on the observed data $X = \{X_1, \ldots, X_n\} \in \mathcal{X}^n$, such that $p(\cdot|\theta)$ is as close to q as possible when measured by a certain distance function (to be specify later).

In the framework considered in this paper, we assume that there is a prior distribution $d\pi(\theta)$ on the parameter space Γ that is independent of the observed data. For notational simplicity, we shall call any observation X dependent probability density $\hat{w}_X(\theta)$ on Γ (measurable on $\mathcal{X}^n \times \Gamma$) with respect to $d\pi(\theta)$ a *posterior randomization measure*. In particular, a posterior randomization measure in our sense is not limited to the *Bayesian posterior distribution*, which has a specific meaning. We are interested in the density estimation performance of

[1] Without causing any confusion, we may also occasionally denote the model family $\{p(\cdot|\theta) : \theta \in \Gamma\}$ by the same symbol Γ.

randomized estimators that draw θ according to posterior randomization measure $\hat{w}_X(\theta)$ obtained from a class of density estimation schemes. We should note that in this framework, our density estimator is completely characterized by the associated posterior randomization density $\hat{w}_X(\theta)$.

2 Information Complexity Minimization Method

We introduce an information theoretical complexity measure of randomized estimators represented as posterior randomization densities.

Definition 1. *Consider a probability density $w(\cdot)$ on Γ with respect to π. The KL-divergence $D_{KL}(wd\pi||d\pi)$ is defined as:*

$$D_{KL}(wd\pi||d\pi) = \int_{\Gamma} w(\theta) \ln w(\theta) d\pi(\theta).$$

The definition becomes the differential entropy for measures on a real-line, when we choose the uniform prior. If we place the prior uniformly on an ϵ-net of the parameter space, then the KL-compleixty becomes ϵ-entropy. KL-divergence is a rather standard information theoretical concept. We will later show that it can be used to measure the complexity of a randomized estimator. We call such a measure the *KL-complexity* or *KL-entropy* of a randomized estimator.

For a real-valued function $f(\theta)$ on Γ, we denote by $\mathbf{E}_\pi f(\theta)$ the expectation of $f(\cdot)$ with respect to π. Similarly, for a real-valued function $\ell(x)$ on $\mathcal{X}$, we denote by $\mathbf{E}_q \ell(x)$ the expectation of $\ell(\cdot)$ with respect the true underlying distribution q. We also use $\mathbf{E}_X$ to denote the expectation with respect to the observation X (n independent samples from q).

The MDL method (7) which we will study in Section 5 can be regarded as a special case of a general class of estimation methods which we refer to as *Information Complexity Minimization*. The method produces a posterior randomization density. Let S be a pre-defined set of densities on Γ with respect to the prior π. We consider a general information complexity minimization estimator:

$$\hat{w}_X^S = \arg\min_{w \in S} \left[-\mathbf{E}_\pi \, w(\theta) \sum_{i=1}^{n} \ln p(X_i|\theta) + \lambda D_{KL}(wd\pi||d\pi) \right]. \tag{1}$$

If we let S be the set of all possible posterior randomization measures, then the estimator leads to the Bayesian posterior distribution with $\lambda = 1$ (see [11]). Therefore bounds obtained for (1) can also be applied to Bayesian posterior distributions. Instead of focusing on the more special MDL method presented later in (7), we shall develop our analysis for the general formulation in (1).

3 The Basic Information Theoretical Inequality

The key ingredient of our analysis using KL-complexity is a well-known convex duality, which has already been used in some recent machine learning papers to study sample complexity bounds [5, 7].

Proposition 1. *Assume that $f(\theta)$ is a measurable real-valued function on Γ, and $w(\theta)$ is a density with respect to π, we have $\mathbf{E}_\pi\, w(\theta)f(\theta) \leq D_{KL}(wd\pi||d\pi) + \ln \mathbf{E}_\pi \exp(f(\theta))$.*

The basis of the paper is the following lemma, where we assume that $\hat{w}_X(\theta)$ is a posterior randomization measure (density with respect to π that depends on X and measurable on $\mathcal{X}^n \times \Gamma$).

Lemma 1 (Information Exponential Inequality). *Consider any posterior randomization density $\hat{w}_X(\theta)$. Let α and β be two real numbers. The following inequality holds for all measurable real-valued functions $L_X(\theta)$ on $\mathcal{X}^n \times \Gamma$:*

$$\mathbf{E}_X \exp\left[\mathbf{E}_\pi \hat{w}_X(\theta)(L_X(\theta) - \alpha \ln \mathbf{E}_X e^{\beta L_X(\theta)}) - D_{KL}(\hat{w}_X d\pi||d\pi)\right] \leq \mathbf{E}_\pi \frac{\mathbf{E}_X e^{L_X(\theta)}}{\mathbf{E}_X^\alpha e^{\beta L_X(\theta)}}.$$

where $\mathbf{E}_X$ is the expectation with respect to the observation X.

Proof. From Proposition 1, we obtain

$$\begin{aligned}\hat{L}(X) =& \mathbf{E}_\pi \hat{w}_X(\theta)(L_X(\theta) - \alpha \ln \mathbf{E}_X e^{\beta L_X(\theta)}) - D_{KL}(\hat{w}_X d\pi||d\pi) \\ \leq& \ln \mathbf{E}_\pi \exp(L_X(\theta) - \alpha \ln \mathbf{E}_X e^{\beta L_X(\theta)}).\end{aligned}$$

Now applying Fubini's theorem to interchange the order of integration, we have:

$$\mathbf{E}_X e^{\hat{L}(X)} \leq \mathbf{E}_X \mathbf{E}_\pi e^{L_X(\theta) - \alpha \ln \mathbf{E}_X \exp(\beta L_X(\theta))} = \mathbf{E}_\pi \frac{\mathbf{E}_X e^{L_X(\theta)}}{\mathbf{E}_X^\alpha e^{\beta L_X(\theta)}}.$$

□

Remark 1. The main technical ingredients of the proof are motivated from techniques in the recent machine learning literature. The general idea for analyzing randomized estimators using Fubini's theorem and decoupling was already in [10]. The specific decoupling mechanism using Proposition 1 appeared in [5, 7] for related problems. A simplified form of Lemma 1 was used in [11] to analyze Bayesian posterior distributions.

The following bound is a straight-forward consequence of Lemma 1. Note that for density estimation, the loss $\ell_\theta(x)$ has a form of $\ell(p(x|\theta))$, where $\ell(\cdot)$ is a scaled log-loss.

Theorem 1 (Information Posterior Bounds). *Using the notation of Lemma 1. Let $X = \{X_1, \ldots, X_n\}$ be n-samples that are independently drawn from q. Consider a measurable function $\ell_\theta(x) : \Gamma \times \mathcal{X} \to R$. Consider real numbers α and β, and define*

$$c_n(\alpha, \beta) = \frac{1}{n} \ln \mathbf{E}_\pi \left(\frac{\mathbf{E}_q e^{-\ell_\theta(x)}}{\mathbf{E}_q^\alpha e^{-\beta \ell_\theta(x)}}\right)^n.$$

Then $\forall t$, the following event holds with probability at least $1 - \exp(-t)$:

$$-\alpha \mathbf{E}_\pi \hat{w}_X(\theta) \ln \mathbf{E}_q\, e^{-\beta \ell_\theta(x)} \leq \frac{\mathbf{E}_\pi\, \hat{w}_X(\theta) \sum_{i=1}^n \ell_\theta(X_i) + D_{KL}(\hat{w}_X d\pi||d\pi) + t}{n} + c_n(\alpha, \beta).$$

Moreover, we have the following expected risk bound:

$$-\alpha \mathbf{E}_X \mathbf{E}_\pi \hat{w}_X(\theta) \ln \mathbf{E}_q e^{-\beta \ell_\theta(x)} \leq \mathbf{E}_X \frac{\mathbf{E}_\pi \hat{w}_X(\theta) \sum_{i=1}^n \ell_\theta(X_i) + D_{KL}(\hat{w}_X d\pi || d\pi)}{n} + c_n(\alpha, \beta).$$

Proof. We use the notation of Lemma 1, with $L_X(\theta) = -\sum_{i=1}^n \ell_\theta(X_i)$. If we define $\hat{L}(X) = \mathbf{E}_\pi \hat{w}_X(\theta)(L_X(\theta) - \alpha \ln \mathbf{E}_X e^{\beta L_X(\theta)}) - D_{KL}(\hat{w}_X d\pi || d\pi)$, then by Lemma 1, we have $\mathbf{E}_X e^{\hat{L}(X)} \leq e^{nc_n(\alpha,\beta)}$. This implies $\forall \epsilon$: $e^\epsilon P(\hat{L}(X) > \epsilon) \leq e^{nc_n(\alpha,\beta)}$. Let $t = -\ln(1 - P(\hat{L}(X) \leq \epsilon))$, we obtain $\epsilon \leq nc_n(\alpha, \beta) + t$. Therefore with probability at least $1 - e^{-t}$, $\hat{L}(X) \leq \epsilon \leq nc_n(\alpha, \beta) + t$. Rearranging, we obtain the first inequality of the theorem.

To prove the second inequality, we still start with $\mathbf{E}_X e^{\hat{L}(X)} \leq e^{nc_n(\alpha,\beta)}$ from Lemma 1. From Jensen's inequality with the convex function e^x, we obtain $e^{\mathbf{E}_X \hat{L}(X)} \leq \mathbf{E}_X e^{\hat{L}(X)} \leq e^{nc_n(\alpha,\beta)}$. That is, $\mathbf{E}_X \hat{L}(X) \leq nc(\alpha, \beta)$. Rearranging, we obtain the desired bound. □

Remark 2. The special case of Theorem 1 with $\alpha = \beta = 1$ is very useful since in this case, the term $c_n(\alpha, \beta)$ vanishes. In fact, in order to obtain the correct rate of convergence for non-parametric problems, it is sufficient to choose $\alpha = \beta = 1$. The more complicated case with general α and β is only needed for parametric problems, where we would like to obtain a convergence rate of the order $O(1/n)$. In such cases, the choice of $\alpha = \beta = 1$ would lead to a rate of $O(\ln n/n)$, which is suboptimal.

4 Bounds for Information Complexity Minimization

Consider the Information Complexity Minimization (1). Given the true density q, if we define

$$\hat{R}_\lambda(w) = \frac{1}{n} \mathbf{E}_\pi w(\theta) \sum_{i=1}^n \ln \frac{q(X_i)}{p(X_i|\theta)} + \frac{\lambda}{n} D_{KL}(w d\pi || d\pi), \tag{2}$$

then it is clear that

$$\hat{w}_X^S = \arg\min_{w \in S} \hat{R}_\lambda(w).$$

The above estimation procedure finds a randomized estimator by minimizing the regularized empirical risk $\hat{R}_\lambda(w)$ among all possible densities with respect to the prior π in a pre-defined set S.

The purpose of this section is to study the performance of the estimator defined in (2) using Theorem 1. For simplicity, we shall only study the expected performance using the second inequality, although similar results can be obtained using the first inequality (which leads to exponential probability bounds).

One may define the true risk of w by replacing the empirical expectation in (1) with the true expectation with respect to q:

$$R_\lambda(w) = \mathbf{E}_\pi w(\theta) D_{KL}(q || p(\cdot|\theta)) + \frac{\lambda}{n} D_{KL}(w d\pi || d\pi), \tag{3}$$

where $D_{KL}(q||p) = \mathbf{E}_q \ln(q(x)/p(x))$ is the KL-divergence between q and p. The information complexity minimizer in (1) can be regarded as an approximate solution to (3) using empirical expectation.

Using empirical process techniques, one can typically expect to bound $R_\lambda(w)$ in terms of $\hat{R}_\lambda(w)$. Unfortunately, it does not work in our case since $D_{KL}(q||p)$ is not well-defined for all p. This implies that as long as w has non-zero concentration around a density p with $D_{KL}(q||p) = +\infty$, then $R_\lambda(w) = +\infty$. Therefore we may have $R_\lambda(\hat{w}_X^S) = +\infty$ with non-zero probability even when the sample size approaches infinity.

A remedy is to use a distance function that is always well-defined. In statistics, one often considers the ρ-divergence for $\rho \in (0, 1)$, which is defined as:

$$D_\rho(q||p) = \frac{1}{\rho(1-\rho)} \mathbf{E}_q \left[1 - \left(\frac{p(x)}{q(x)} \right)^\rho \right]. \tag{4}$$

This divergence is always well-defined and $D_{KL}(q||p) = \lim_{\rho \to 0} D_\rho(q||p)$. In the statistical literature, convergence results were often specified under the Hellinger distance ($\rho = 0.5$). In this paper, we specify convergence results with general ρ. We shall mention that bounds derived in this paper will become trivial when $\rho \to 0$. This is consistent with the above discussion since R_λ (corresponding to $\rho = 0$) may not converge at all. However, under additional assumptions, such as the boundedness of q/p, $D_{KL}(q||p)$ exists and can be bounded using the ρ-divergence $D_\rho(q||p)$.

The following bounds imply that up to a constant, the ρ-divergence with any $\rho \in (0, 1)$ is equivalent to the Hellinger distance. Therefore a convergence bound in any ρ-divergence implies a convergence bound of the same rate in the Hellinger distance. Since this result is not crucial in our analysis, we skip the proof due to the space limitation.

Proposition 2. *We have the following inequalities* $\forall \rho \in [0, 1]$*:*

$$\max(\rho, 1-\rho) D_\rho(q||p) \geq \frac{1}{2} D_{1/2}(q||p) \geq \min(\rho, 1-\rho) D_\rho(q||p).$$

4.1 A general convergence bound

The following general theorem is an immediate consequence of Theorem 1. Most of our later discussions can be considered as interpretations of this theorem under various different conditions.

Theorem 2. *Consider the estimator* $\hat{w}_X^S$ *defined in (1). Let* $\alpha > 0$*. Then* $\forall \rho \in (0, 1)$ *and* $\gamma \geq \rho$ *such that* $\lambda' = \frac{\lambda\gamma - 1}{\gamma - \rho} \geq 0$*, we have:*

$$\begin{aligned}
\mathbf{E}_X \mathbf{E}_\pi \hat{w}_X^S(\theta) D_\rho(q||p(\cdot|\theta)) \leq & \frac{-1}{\rho(1-\rho)} \mathbf{E}_X \mathbf{E}_\pi \hat{w}_X^S(\theta) \ln \mathbf{E}_q \left(\frac{p(x|\theta)}{q(x)} \right)^\rho \\
\leq & \frac{\gamma \inf_{w \in S} R_\lambda(w)}{\alpha\rho(1-\rho)} - \frac{\gamma - \rho}{\alpha\rho(1-\rho)} \mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S) + \frac{c_{\rho,n}(\alpha)}{\alpha\rho(1-\rho)},
\end{aligned}$$

where

$$c_{\rho,n}(\alpha) = \frac{1}{n} \ln \mathbf{E}_\pi \mathbf{E}_q^{(1-\alpha)n} \left(\frac{p(x|\theta)}{q(x)} \right)^\rho .$$

Proof Sketch. Consider an arbitrary data-independent density $w(\theta) \in S$ with respect to π, using (4), we can obtain from Theorem 1 the following chain of equations:

$$\begin{aligned}
&\alpha\rho(1-\rho)\mathbf{E}_X \mathbf{E}_\pi \hat{w}_X^S(\theta) D_\rho(q||p(\cdot|\theta)) \\
\leq &\alpha \mathbf{E}_X \mathbf{E}_\pi \hat{w}_X^S(\theta) \ln \frac{1}{1-\rho(1-\rho)D_\rho(q||p(\cdot|\theta))} \\
\leq &\mathbf{E}_X \left[\rho \mathbf{E}_\pi \, \hat{w}_X^S \sum_{i=1}^n \frac{1}{n} \ln \frac{q(X_i)}{p(X_i|\theta)} + \frac{D_{KL}(\hat{w}_X^S d\pi || d\pi)}{n} \right] + c_{\rho,n}(\alpha) \\
\leq &\mathbf{E}_X \left[\gamma \hat{R}_\lambda(w) + (\rho - \gamma) \hat{R}_{\lambda'}(\hat{w}_X^S) \right] + c_{\rho,n}(\alpha),
\end{aligned}$$

where $R_\lambda(w)$ is defined in (3). □

Remark 3. If $\gamma = \rho$ in Theorem 2, then we also require $\lambda\gamma = 1$, and let $\lambda' = 0$.

Consequences of this theorem will later be applied to MDL methods. Although the bound in Theorem 2 looks complicated, the most important part on the right hand side is the first term. The second term is only needed to handle the situation $\lambda \leq 1$. The requirement that $\gamma \geq \rho$ is to ensure that the second term is non-positive. Therefore in order to apply the theorem, we only need to estimate a lower bound of $\hat{R}_{\lambda'}(\hat{w}_X^S)$, which (as we shall see later) is much easier than obtaining an upper bound. The third term is mainly included to get the correct convergence rate of $O(1/n)$ for parametric problems, and can be ignored for non-parametric problems. The effect of this term is quite similar to using localized ϵ-entropy in the empirical process approach for analyzing the maximum-likelihood method (for example, see [8]). As a comparison, the KL-entropy in the first term corresponds to the global ϵ-entropy.

Note that one can easily obtain a simplified bound from Theorem 2 by choosing specific parameters so that both the second term and the third term vanish:

Corollary 1. *Consider the estimator $\hat{w}_X^S$ defined in (1). Assume that $\lambda > 1$ and let $\rho = 1/\lambda$, we have*

$$\mathbf{E}_X \mathbf{E}_\pi \hat{w}_X^S(\theta) D_\rho(q||p(\cdot|\theta)) \leq \frac{1}{1-\rho} \inf_{w \in S} R_\lambda(w).$$

Proof. We simply let $\alpha = 1$ and $\gamma = \rho$ in Theorem 2. □.

An important observation is that for $\lambda > 1$, the convergence rate is solely determined by the quantity $\inf_{w \in S} R_\lambda(w)$, which we shall refer to as the *model resolvability* associated with S.

4.2 Some lower bounds on $\mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S)$

Lemma 2. *$\forall \lambda' \geq 1$: $\mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S) \geq -\frac{\lambda'}{n} \geq 0$.*

Proof. See Appendix A. □

By combining the above estimate with Theorem 2, we obtain the following refinement of Corollary 1.

Corollary 2. *Consider the estimator $\hat{w}_X^S$ defined in (1). Assume that $\lambda > 1$, then $\forall \rho \in (0, 1/\lambda]$:*

$$\mathbf{E}_X \mathbf{E}_\pi \hat{w}_X^S(\theta) D_\rho(q||p(\cdot|\theta)) \leq \frac{1}{\rho(\lambda - 1)} \inf_{w \in S} R_\lambda(w).$$

Proof. We simply let $\alpha = 1$ and $\gamma = (1-\rho)/(\lambda - 1)$ in Theorem 2. Note that in this case, $\lambda' = 1$, and hence by Lemma 2, $\mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S) \geq 0$. □

Note that Lemma 2 is only applicable for $\lambda' \geq 1$. If $\lambda' \leq 1$, then we need a discretization device, which generalizes the upper ϵ-covering number concept used in [2] for showing the consistency (or inconsistency) of Bayesian posterior distributions:

Definition 2. *The ϵ-upper bracketing number of Γ, denoted by $N(\Gamma, \epsilon)$, is the minimum number of non-negative functions $\{f_j\}$ on $\mathcal{X}$ with respect to μ such that $\mathbf{E}_q(f_j/q) = 1 + \epsilon$, and $\forall \theta \in \Gamma$, $\exists j$ such that $p(x|\theta) \leq f_j(x)$ a.e. $[\mu]$.*

The discretization device which we shall use in this paper is based on the following definition:

Definition 3. *An ϵ-upper discretization of Γ consists of a countable decomposition of Γ as measurable subsets $\{\Gamma_j\}$ such that $\cup_j \Gamma_j = \Gamma$ and $\mathbf{E}_q \sup_{\theta \in \Gamma_j}(p(x|\theta)/q(x)) \leq 1 + \epsilon$.*

Lemma 3. *Consider an ϵ-upper discretization $\{\Gamma_j\}$ of Γ. The following inequality is valid $\forall \lambda' \in [0, 1]$:*

$$\mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S) \geq -\left[\frac{\ln \sum_j \pi(\Gamma_j)^{\lambda'}}{n} + \ln(1+\epsilon)\right].$$

Proof. See Appendix B. □

Combine the above estimate with Theorem 2, we obtain the following simplified bound for $\lambda = 1$. Similar results can be obtained for $\lambda < 1$ but the case of $\lambda = 1$ is most interesting.

Corollary 3. *Consider the estimator defined in (1). Let $\lambda = 1$. Consider an ϵ-upper discretization $\{\Gamma_i\}$ of Γ. $\forall \rho \in (0, 1)$ and $\forall \gamma \geq 1$, we have:*

$$\mathbf{E}_X \mathbf{E}_\pi \hat{w}_X^S(\theta) D_\rho(q||p(\cdot|\theta)) \leq \frac{\gamma \inf_{w \in S} R_\lambda(w)}{\rho(1-\rho)} + \frac{\gamma - \rho}{\rho(1-\rho)} \left[\frac{\ln \sum_j \pi(\Gamma_j)^{\frac{\gamma-1}{\gamma-\rho}}}{n} + \ln(1+\epsilon)\right].$$

Proof. We let $\alpha = 1$ in Theorem 2, and apply Lemma 3. □

Note that the above results immediately imply the following bound using ϵ-upper entropy by letting $\gamma \to 1$ with a finite ϵ-upper bracketing cover of size $N(\Gamma, \epsilon)$ as the discretization:

$$\mathbf{E}_X \mathbf{E}_\pi \hat{w}_X^S(\theta) D_\rho(q||p(\cdot|\theta)) \leq \frac{\inf_{w \in S} R_\lambda(w)}{\rho(1-\rho)} + \frac{1}{\rho} \inf_{\epsilon > 0} \left[\frac{\ln N(\Gamma, \epsilon)}{n} + \ln(1+\epsilon) \right]. \tag{5}$$

It is clear that Corollary 3 is significantly more general than the covering number result (5). We are able to deal with an infinite cover as long as the decay of the prior π is fast enough on the discretization so that $\sum_j \pi(\Gamma_j)^{(\gamma-1)/(\gamma-\rho)} < +\infty$.

4.3 Weak convergence bound

The case of $\lambda = 1$ is related to a number of important estimation methods in statistical applications such as the standard MDL and Bayesian methods. However, for an arbitrary prior π without any additional assumption such as the fast decay condition in Corollary 3, it is not possible to establish any convergence rate result in terms of Hellinger distance using the model resolvability quantity alone, as in the case of $\lambda > 1$ (Corollary 2). See Section 5.4 for an example demonstrating this claim. However, one can still obtain a weaker convergence result in this case. The following theorem essentially implies that the posterior randomization average $\mathbf{E}_\pi \hat{w}_X^S(\theta) p(\cdot|\theta)$ converges weakly to q as long as the model resolvability $\inf_{w \in S} R_\lambda(w) \to 0$ when $n \to \infty$.

Theorem 3. *Consider the estimator $\hat{w}_X^S$ defined in (1) with $\lambda = 1$. Then $\forall f : \mathcal{X} \to [-1, 1]$, we have:*

$$\mathbf{E}_X \left| \mathbf{E}_\pi \hat{w}_X^S(\theta) \mathbf{E}_{p(\cdot|\theta)} f(x) - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \leq 2A_n + \sqrt{2A_n},$$

where $A_n = \inf_{w \in S} \mathbf{E}_X R_\lambda(w) + \frac{\ln 2}{n}$.

Proof Sketch. Let $g_\epsilon(x) = 1 - \epsilon f(x)$, and $h_\epsilon(x) = \frac{q(x)}{p(x|\theta) g_\epsilon(x)}$, where $\epsilon \in (-1, 1)$ is a parameter to be determined later. Note that $g_\epsilon(x) > 0$. Let $\alpha = \beta = 1$ and $L_X(\theta) = -\sum_{i=1}^n \ln h_\epsilon(X_i)$ in Lemma 1, we have

$$\mathbf{E}_X \exp\left[\mathbf{E}_\pi \hat{w}_X^S(\theta) \left(-\sum_{i=1}^n \ln h_\epsilon(X_i) - \ln \mathbf{E}_X \prod_{i=1}^n \frac{1}{h_\epsilon(X_i)} \right) - D_{KL}(\hat{w}_X^S d\pi || d\pi) \right] \leq 1.$$

If we let

$$\Delta_\epsilon(X) = \mathbf{E}_\pi \hat{w}_X^S(\theta) \left(\sum_{i=1}^n \ln g_\epsilon(X_i) - n \ln \mathbf{E}_{p(\cdot|\theta)} g_\epsilon(x) \right),$$

then $\mathbf{E}_X e^{\Delta_\epsilon(X)-n\hat{R}_\lambda(\hat{w}_X^S)} \leq 1$. This implies that $\mathbf{E}_X[e^{\Delta_\epsilon(X)}+e^{\Delta_{-\epsilon}(X)}]e^{-n\hat{R}_\lambda(\hat{w}_X^S)} \leq 2$. Applying Jensen's inequality, we obtain

$$\mathbf{E}_X \ln[e^{\Delta_\epsilon(X)} + e^{\Delta_{-\epsilon}(X)}] \leq n\mathbf{E}_X \hat{R}_\lambda(\hat{w}_X^S) + \ln 2 \leq n \inf_{w\in S} R_\lambda(w) + \ln 2. \quad (6)$$

Consider $x \leq y < 1$. We have the following inequalities (which follow from Taylor expansion) $x \leq -\ln(1-x) \leq x+\frac{x^2}{2(1-y)^2}$. This implies $\ln g_\epsilon(x) \geq -\epsilon f(x) - \frac{\epsilon^2}{2(1-|\epsilon|)^2}$ and $-\ln \mathbf{E}_{p(\cdot|\theta)} g_\epsilon(x) \geq \epsilon \mathbf{E}_{p(\cdot|\theta)} f(x)$. Therefore

$$\Delta_\epsilon(X) \geq \epsilon \mathbf{E}_\pi \hat{w}_X^S(\theta) \left(-\sum_{i=1}^n f(X_i) + n\mathbf{E}_{p(\cdot|\theta)} f(x) \right) - \frac{n\epsilon^2}{2(1-|\epsilon|)^2}.$$

A similar bound can be obtained for $\Delta_{-\epsilon}(X)$. Now substitute them into (6) and observe that $|x| \leq \ln(e^x + e^{-x})$, we obtain

$$\mathbf{E}_X \left| \epsilon \mathbf{E}_\pi \hat{w}_X^S(\theta) \left(-\sum_{i=1}^n f(X_i) + n\mathbf{E}_{p(\cdot|\theta)} f(x) \right) \right| - \frac{n\epsilon^2}{2(1-|\epsilon|)^2} \leq n \inf_{w\in S} \mathbf{E}_X R_\lambda(w) + \ln 2.$$

Let $|\epsilon| = \sqrt{2A_n}/(\sqrt{2A_n}+1)$, we obtain the desired bound. □

5 MDL on Discrete Net

The minimum description length (MDL) method has been widely used in practice [6]. The version we consider here is the same as that of [1]. In fact, results in this section improve those of [1]. The MDL method considered in [1] can be regarded as a special case of information complexity minimization. The model space Γ is countable: $\theta \in \Gamma = \{1, 2, \ldots\}$. We denote the corresponding models $p(x|\theta = j)$ by $p_j(x)$. The prior π has a form $\pi = \{\pi_1, \pi_2, \ldots\}$ such that $\sum_j \pi_j = 1$, where we assume that $\pi_j > 0$ for each j. A randomized algorithm can be represented as a non-negative weight vector $w = [w_j]$ such that $\sum_j \pi_j w_j = 1$.

MDL gives a deterministic estimator, which corresponds to the set of weights concentrated on any one specific point k. That is, we can select S in (1) such that each weight w in S corresponds to an index $k \in \Gamma$ such that $w_k = 1/\pi_k$ and $w_j = 0$ when $j \neq k$. It is easy to check that $D_{KL}(wd\pi||d\pi) = \ln(1/\pi_k)$. The corresponding algorithm can thus be described as finding a probability density $p_{\hat{k}}$ with $\hat{k}$ obtained by

$$\hat{k} = \arg\min_k \left[\sum_{i=1}^n \ln \frac{1}{p_k(X_i)} + \lambda \ln \frac{1}{\pi_k} \right], \quad (7)$$

where $\lambda \geq 1$ is a regularization parameter. The first term corresponds to the description of the data, and the second term corresponds to the description of the model. The choice $\lambda = 1$ can be interpreted as minimizing the total description length, which corresponds to the standard MDL. The choice $\lambda > 1$ corresponds to

heavier penalty on the model description, which makes the estimation method more stable. This modified MDL method was considered in [1] for which the authors obtained results on the asymptotic rate of convergence. However, no simple finite sample bounds were obtained. For the case of $\lambda = 1$, only weak consistency was shown. In the following, we shall improve these results using the analysis presented in Section 4.

5.1 Modified MDL under global entropy condition

Consider the case $\lambda > 1$ in (7). We can obtain the following theorem from Corollary 2.

Theorem 4. *Consider the estimator $\hat{k}$ defined in (7). Assume that $\lambda > 1$, then $\forall \rho \in (0, 1/\lambda]$:*

$$\mathbf{E}_X D_\rho(q||p_{\hat{k}}) \leq \frac{1}{\rho(\lambda-1)} \inf_k \left[D_{KL}(q||p_k) + \frac{\lambda}{n} \ln \frac{1}{\pi_k} \right].$$

Note that in [1], the term $r_{\lambda,n}(q) = \inf_k \left[D_{KL}(q||p_k) + \frac{\lambda}{n} \ln \frac{1}{\pi_k} \right]$ is referred to as *index of resolvability*. They showed (Theorem 4) that $D_{1/2}(q||p_{\hat{k}}) = O_p(r_{\lambda,n}(q))$ when $\lambda > 1$. Theorem 4 is a slight generalization of a result developed by Andrew Barron and Jonathan Li, which gave the same inequality but only for the case of $\lambda = 2$ and $\rho = 1/2$. The result, with a proof quite similar to what we presented here, can be found in [4] (Theorem 5.5, page 78).

Examples of index of resolvabilities for various function classes can be found in [1], which we shall not repeat in this paper. In particular, it is known that for non-parametric problems, with appropriate discretization, the rate matches the minimax rate such as those in [9].

5.2 Local entropy analysis

Although the bound based on the index of resolvability in Theorem 4 is quite useful for non-parametric problems (see [1] for examples), it does not handle the parametric case satisfactorily. To see this, we consider a one-dimensional parameter family indexed by $\theta \in [0, 1]$, and we discretize the family using a uniform discrete net of size $N + 1$: $\theta_j = j/N$ $(j = 0, \ldots, N)$. If q is taken from the parametric family so that we can assume that $\inf_k D_{KL}(q||p_k) = O(N^{-2})$, then Theorem 4 with $\lambda = 2$, $\rho = 1/2$ and uniform prior on the net, becomes $\mathbf{E}_X D_{1/2}(q||p_{\hat{k}}) \leq O(N^{-2}) + \frac{4}{n} \ln \frac{1}{N}$. Now by choosing $N = O(n^{-1/2})$, we obtain a suboptimal convergence rate $\mathbf{E}_X D_{1/2}(q||p_{\hat{k}}) \leq O(\ln n/n)$. Note that convergence rates established in [1] for parametric examples are also of the order $O(\ln n/n)$.

The main reason for this sub-optimality is that the complexity measure $O(\ln N)$ or $O(-\ln \pi_k)$ corresponds to the globally defined entropy. However, readers who are familiar with the empirical process theory know that the rate of convergence of the maximum likelihood estimate is determined by local entropy

which appeared in [3]. For non-parametric problems, it was pointed out in [9] that the worst case local entropy is the same order of the global entropy. Therefore a theoretical analysis which relies on global entropy (such as Theorem 4) leads to the correct worst case rate at least in the minimax sense. For parametric problems, at the $O(1/n)$ approximation level, local entropy is constant but the global entropy is $\ln n$. This leads to a $\ln(n)$ difference in the resulting bound.

Although it may not be immediately obvious how to define a localized counterpart of the index of resolvability, we can make a correction term which has the same effect. As pointed out earlier, this is essentially the role of the $c_{\rho,n}(\alpha)$ term in Theorem 2. We include a simplified version below, which can be obtained by choosing $\alpha = 1/2$, and $\gamma = \rho = 1/\lambda$.

Theorem 5. *Consider the estimator $\hat{k}$ defined in (7). Assume that $\lambda > 1$, and let $\rho = 1/\lambda$:*

$$\mathbf{E}_X D_\rho(q||p_{\hat{k}}) \leq \frac{2}{1-\rho} \inf_k \left[D_{KL}(q||p_k) + \frac{\lambda}{n} \ln \frac{\sum_j \pi_j \mathbf{E}_q^{n/2} \left(\frac{p_j(x)}{q(x)}\right)^\rho}{\pi_k} \right].$$

The bound relies on a localized version of the index of resolvability, with the global entropy $-\ln \pi_k$ replaced by a localized entropy $\ln \sum_j \pi_j \mathbf{E}_q^{n/2} \left(\frac{p_j(x)}{q(x)}\right)^\rho - \ln \pi_k$. Since

$$\ln \sum_j \pi_j \mathbf{E}_q^{n/2} \left(\frac{p_j(x)}{q(x)}\right)^\rho \leq \ln \sum_j \pi_j = 0,$$

the localized entropy is always smaller than the global entropy. Intuitively, we can see that if $p_j(x)$ is far away from $q(x)$, then $\mathbf{E}_q^{n/2} \left(\frac{p_j(x)}{q(x)}\right)^\rho$ is very small as $n \to \infty$. It follows that the summation in $\sum_j \pi_j \mathbf{E}_q^{n/2} \left(\frac{p_j(x)}{q(x)}\right)^\rho$ is mainly contributed by terms such that $D_\rho(q||p_j)$ is small. This is equivalent to a re-weighting of prior π_k in such a way that we only count points that are localized within a small D_ρ ball of q.

This localization leads to the correct rate of convergence for parametric problems. The effect is similar to using localized entropy in the empirical process analysis. We consider the maximum likelihood estimate with a general one dimensional problem discussed at the beginning of the section with a uniform discretization consisted of $N+1$ points. For one-dimensional parametric problems, it is natural to assume that the number of k such that $\rho(1-\rho)D_\rho(q||p_k) \leq 1 - \exp(-m^2/N^2)$ is $O(m)$ for $m \geq 1$. This implies that $\forall N = O(n^{1/2})$,

$$\ln \sum_j \mathbf{E}_q^{n/2} \left(\frac{p_j(x)}{q(x)}\right)^\rho \leq \ln \sum_m O(m)(e^{-m^2/N^2})^{n/2} = O(1).$$

Since $\pi_j = 1/N$, the localized entropy

$$\ln \frac{\sum_j \pi_j \mathbf{E}_q^{n/2} \left(\frac{p_j(x)}{q(x)}\right)^\rho}{\pi_k} = O(1)$$

is a constant when $N = O(n^{1/2})$. Therefore with a discretization size $N = O(n^{1/2})$, Theorem 5 implies a convergence rate of the correct order $O(1/n)$.

5.3 The standard MDL ($\lambda = 1$)

The standard MDL with $\lambda = 1$ in (7) is more complicated to analyze. It is not possible to give a bound similar to Theorem 4 that only depends on the index of resolvability. As a matter of fact, no bound was established in [1]. As we will show later, the method can converge very slowly even if the index of resolvability is well-behaved.

However, it is possible to obtain bounds in this case under additional assumptions on the rate of decay of the prior π. The following theorem is a straightforward interpretation of Corollary 3, where we consider the family itself as an 0-upper discretization: $\Gamma_i = \{p_i\}$:

Theorem 6. *Consider the estimator defined in (7) with $\lambda = 1$. $\forall \rho \in (0,1)$ and $\forall \gamma \geq 1$, we have:*

$$\mathbf{E}_X D_\rho(q||p_{\hat{k}}) \leq \frac{\gamma \inf_k \left[D_{KL}(q||p_k) + \frac{1}{n} \ln \frac{1}{\pi_k} \right]}{\rho(1-\rho)} + \frac{\gamma - \rho}{\rho(1-\rho)n} \ln \sum_j \pi_j^{(\gamma-1)/(\gamma-\rho)}.$$

The above theorem only depends on the index of resolvability and decay of the prior π. If π has a fast decay in the sense of $\sum_j \pi_j^{(\gamma-1)/(\gamma-\rho)} < +\infty$ and does not change with respect to n, then the second term on the right hand side of Theorem 6 is $O(1/n)$. In this case the convergence rate is determined by the index of resolvability. The prior decay condition specified here is rather mild. This implies that the standard MDL is usually Hellinger consistent when used with care.

5.4 Slow convergence of the standard MDL

The purpose of this section is to illustrate that the index of resolvability cannot by itself determine the rate of convergence for the standard MDL. We consider a simple example related to the Bayesian inconsistency counter-example given in [2], with an additional randomization argument. Note that due to the randomization, we shall allow two densities in our model class to be identical. It is clear from the construction that this requirement is for convenience only, rather than anything essential.

Given a sample size n, and consider an integer m such that $m \gg n$. Let the space $\mathcal{X}$ consist of $2m$ points $\{1, \ldots, 2m\}$. Assume that the truth q is the uniform distribution: $q(u) = 1/2m$ for $u = 1, \ldots 2m$.

Consider a density class Γ' consisted of all densities p such that either $p(u) = 0$ or $p(u) = 1/m$. That is, a density p in Γ' takes value $1/m$ at m of the $2m$ points, and 0 elsewhere. Now let our model class Γ be consisted of the true

density q with prior $1/4$, and 2^n densities p_j $(j = 1, \ldots, 2^n)$ that are randomly (and uniformly) drawn from Γ', each with the same prior $3/2^{n+2}$.

We shall show that for a sufficiently large integer m, with large probability we will estimate one of the 2^n densities from Γ' with probability of at least $1 - e^{-1/2}$. Since the index of resolvability is $\ln 4/n$, which is small when n is large, the example implies that the convergence of the standard MDL method cannot be characterized by the index of resolvability alone.

Let $X = \{X_1, \ldots, X_n\}$ be a set of n-samples from q and $\hat{p}$ be the estimator from (7) with $\lambda = 1$ and Γ randomly generated above. We would like to estimate $P(\hat{p} = q)$. By construction, $\hat{p} = q$ only when $\prod_{i=1}^n p_j(X_i) = 0$ for all $p_j \in \Gamma' \cap \Gamma$. Now pick m large enough such that $(m-n)^n/m^n \geq 0.5$, we have

$$\begin{aligned} P(\hat{p} = q) =& P\left(\forall p_j \in \Gamma' \cap \Gamma : \prod_{i=1}^n p_j(X_i) = 0\right) \\ =& \mathbf{E}_X \, P\left(\prod_{i=1}^n p_1(X_i) = 0 | X\right)^{2^n} \leq \mathbf{E}_X \left(1 - \left(\frac{m-n}{2m}\right)^n\right)^{2^n} \leq e^{-0.5}, \end{aligned}$$

where $|X|$ denotes the number of distinct elements in X. Therefore with a constant probability, we have $\hat{p} \neq q$ no matter how large n is.

This example shows that it is not possible to obtain any rate of convergence result using index of resolvability alone. In order to estimate convergence, it is thus necessary to make additional assumptions, such as the prior decay condition of Theorem 6. We shall also mention that from this example together with a construction scheme similar to that of the Bayesian inconsistency counter example in [2], it is not difficult to show that the standard MDL is not Hellinger consistent even when the index of resolvability approaches zero as $n \to \infty$. For simplicity, we skip the detailed construction in this paper.

5.5 Weak convergence of the standard MDL

Although Hellinger consistency cannot be obtained for standard MDL based on index of resolvability alone, it was shown in [1] that as $n \to \infty$, if the index of resolvability approaches zero, then $p_{\hat{k}}$ converges weakly to q. Therefore MDL is effectively weakly consistent as long as q belongs to the information closure of Γ. This result is a direct consequence of Theorem 3, which we shall restate here:

Theorem 7. *Consider the estimator defined in (7) with $\lambda = 1$. Then $\forall f : \mathcal{X} \to [-1, 1]$, we have:*

$$\mathbf{E}_X \left| \mathbf{E}_{p_{\hat{k}}} f(x) - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \leq 2A_n + \sqrt{2A_n},$$

where $A_n = \inf_k \left[D_{KL}(q||p_k) + \frac{1}{n} \ln \frac{1}{\pi_k} \right] + \frac{\ln 2}{n}$.

Note that this theorem essentially implies that the standard MDL estimator is weakly consistent as long as the index of resolvability approaches zero when $n \to 0$. Moreover, it establishes a rate of convergence result which only depends on the index of resolvability. This theorem improves the consistency result in [1], where no rate of convergence results were established, and f was assumed to be an indicator function.

6 Discussions

This paper studies certain randomized (and deterministic) density estimation methods which we call information complexity minimization. We introduced a general KL-complexity based convergence analysis, and demonstrated that the new approach can lead to simplified and improved convergence results for two-part code MDL, which improves the classifical results in [1].

An important observation from our study is that generalized information complexity minimization methods with regularization parameter $\lambda > 1$ are more robust than the corresponding standard methods with $\lambda = 1$. That is, their convergence behavior is completely determined by the local prior density around the true distribution measured by the model resolvability $\inf_{w \in S} R_\lambda(w)$. For MDL, this quantity (index of resolvability) is well-behaved if we put a not too small prior mass at a density that is close to the truth q. We have also demonstrated through an example that the standard MDL does not have this desirable property in that even we can guess the true density by putting a relatively large prior mass at the true density q, we may not estimate q very well as long as there exits a bad (random) prior structure even at places very far from the truth q.

A Proof of Lemma 2

Applying the convex duality in Proposition 1 with $f(x) = -\frac{1}{\lambda'} \sum_{i=1}^n \ln \frac{q(X_i)}{p(X_i|\theta)}$, we obtain

$$\hat{R}_{\lambda'}(\hat{w}_X^S) \geq -\frac{\lambda'}{n} \ln \mathbf{E}_\pi \exp\left(-\frac{1}{\lambda'} \sum_{i=1}^n \ln \frac{q(X_i)}{p(X_i|\theta)}\right).$$

Taking expectation and using Jensen's inequality with the convex function $\psi(x) = -\ln(x)$, we obtain

$$\mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S) \geq -\frac{\lambda'}{n} \ln \mathbf{E}_X \mathbf{E}_\pi \exp\left(-\frac{1}{\lambda'} \sum_{i=1}^n \ln \frac{q(X_i)}{p(X_i|\theta)}\right) \geq 0.$$

B Proof of Lemma 3

The proof is similar to that of Lemma 2, but with a slightly different estimate. We again start with the inequality

$$\hat{R}_{\lambda'}(\hat{w}_X^S) \geq -\frac{\lambda'}{n} \ln \mathbf{E}_\pi \exp\left(-\frac{1}{\lambda'} \sum_{i=1}^n \ln \frac{q(X_i)}{p(X_i|\theta)}\right).$$

Taking expectation and using Jensen's inequality with the convex function $\psi(x) = -\ln(x)$, we obtain

$$\begin{aligned}
-\mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S) \leq & \frac{1}{n} \ln \mathbf{E}_X \mathbf{E}_\pi^{\lambda'} \exp\left(-\frac{1}{\lambda'} \sum_{i=1}^n \ln \frac{q(X_i)}{p(X_i|\theta)}\right) \\
\leq & \frac{1}{n} \ln \mathbf{E}_X \left[\sum_j \pi(\Gamma_j) \exp\left(-\frac{1}{\lambda'} \sum_{i=1}^n \ln \frac{q(X_i)}{\sup_{\theta\in\Gamma_j} p(X_i|\theta)}\right)\right]^{\lambda'} \\
\leq & \frac{1}{n} \ln \mathbf{E}_X \left[\sum_j \pi(\Gamma_j)^{\lambda'} \exp\left(-\sum_{i=1}^n \ln \frac{q(X_i)}{\sup_{\theta\in\Gamma_j} p(X_i|\theta)}\right)\right] \\
= & \frac{1}{n} \ln \left[\sum_j \pi(\Gamma_j)^{\lambda'} \mathbf{E}_X \prod_{i=1}^n \frac{\sup_{\theta\in\Gamma_j} p(X_i|\theta)}{q(X_i)}\right] \\
\leq & \frac{1}{n} \ln \left[\sum_j \pi(\Gamma_j)^{\lambda'} (1+\epsilon)^n\right].
\end{aligned}$$

The third inequality follows from the fact that $\forall \lambda' \in [0,1]$ and positive numbers $\{a_j\}$: $(\sum_j a_j)^{\lambda'} \leq \sum_j a_j^{\lambda'}$.

References

1. Andrew Barron and Thomas Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034–1054, 1991.
2. Andrew Barron, Mark J. Schervish, and Larry Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561, 1999.
3. Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53, 1973.
4. J.Q. Li. *Estimation of Mixture Models*. PhD thesis, The Department of Statistics. Yale University, 1999.
5. Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.
6. J. Rissanen. *Stochastic complexity and statistical inquiry*. World Scientific, 1989.
7. M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *JMLR*, 3:233–269, 2002.
8. S.A. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.
9. Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27:1564–1599, 1999.
10. Tong Zhang. Theoretical analysis of a class of randomized regularization methods. In *COLT 99*, pages 156–163, 1999.
11. Tong Zhang. Learning bounds for a generalized family of Bayesian posterior distributions. In *NIPS 03*, 2004. to appear.