RC22980 (W0311-059) November 7, 2003
Mathematics

IBM Research Report

From epsilon-Entropy to KL-Complexity: Analysis of Minimum Information Complexity Density Estimation

Tong Zhang
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

From ϵ-entropy to KL-complexity: analysis of minimum information complexity density estimation

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

Abstract

We extend the concept of ϵ-entropy to include randomized density estimation methods. Based on this extension, we develop a general information theoretical inequality that measures the statistical complexity of some deterministic and randomized density estimators. Consequences of the new inequality will be presented. In particular, we show that this technique can effortlessly lead to substantial improvements of some classical results concerning the convergence of minimum description length (MDL) and Bayesian posterior distributions. 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 has become a central technical tool in the traditional finite-sample convergence analysis. 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.

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, and we do not assume that q belongs 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 (which we shall specify later).

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

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 distribution*. In particular, a posterior distribution in our sense is not limited to the *Bayesian posterior distribution*, which has a very specific meaning. We are interested in the density estimation performance of randomized estimators that draw θ according to posterior distributions $\hat{w}_X(\theta)$ 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 distribution $\hat{w}_X(\theta)$.

The paper is organized as follows. In Section 2, we introduce the generalization of ϵ-entropy that we call *KL-complexity*. Then a fundamental information theoretical inequality, which forms the basis of our approach, will be obtained. Section 3 introduces the general information complexity minimization (ICM) density estimation formulation, where we derive various finite-sample convergence bounds using the fundamental information theoretical inequality obtained earlier. Section 4 and Section 5 apply the analysis to the case of MDL estimators and to the convergence of Bayesian posterior distributions. In particular, we are able to simplify and improve most results in [1] as well as various recent analysis on the consistency and concentration of Bayesian posterior distributions. Some concluding remarks will be presented in Section 6.

Throughout the paper, we ignore the measurability issue, and assume that all quantities appearing in the derivations to be measurable. Similar to empirical process theory [12], the analysis can also be written in the language of outer-expectations, so that the measurability requirement imposed in this paper can be relaxed.

2 The basic information theoretical inequality

In this section, we introduce an information theoretical complexity measure of randomized estimators represented as posterior densities. As we shall see, this quantity directly generalizes the concept of ϵ-entropy. We also develop a simple yet very general information theoretical inequality, which bounds the convergence behavior of an arbitrary randomized estimator using the newly introduced complexity measure. This inequality is the foundation of the approach introduced in this paper.

Definition 2.1 *Consider a probability density $w(\cdot)$ on Γ with respect to π. The KL-divergence $D_{KL}(wd\pi||d\pi)$ is defined as:*

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

Note that $D_{KL}(wd\pi||d\pi)$ may not always be finite. However, it is always non-negative.

KL-divergence is a rather standard information theoretical concept. In this section, we 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. We can immediately

see that the quantity directly generalizes the concept of ϵ-entropy on an ϵ-net: assume that we have N points in an ϵ-net, we may consider a prior that puts a mass of $1/N$ on every point. It is easy to see that any deterministic estimator in the ϵ-net can be regarded as a randomized estimator that is concentrated on one of the N points with posterior weight N (and weight of zero else-where). Clearly this estimator has a KL-complexity of $\ln N$, which is essentially the ϵ-entropy. In fact, it is also easy to verify that any randomized estimator on the ϵ-net has a KL-complexity upper-bounded by its ϵ-entropy $\ln N$. Therefore ϵ-entropy is the worst-case KL-complexity on an ϵ-net with uniform prior.

It is important to note that the KL-complexity can be applied to any prior π, instead of a discrete uniform prior as in the case of ϵ-entropy. We believe that this difference is significant since it is usually impossible (or very difficult) to perform computation based on a discrete ϵ-net, but it is certainly feasible to draw samples from a posterior distribution with respect to a continuous prior by using standard Monte Carlo techniques. Moreover, at the conceptual level, one may argue that an artificial discretization in the ϵ-net approach appears very unnatural. These concerns are addressed by our new complexity definition.

Another fundamental importance of the new definition is its ability to directly characterize local adaptivity of randomized estimators when we put more prior mass in certain regions of the model family. The issue of adaptivity (and related non-uniform prior) cannot be directly addressed with ϵ-entropy, and one has to use additional techniques (for example, peeling [11]) for this purpose. We can again argue that such an indirect route in the ϵ-entropy analysis is not as natural as simply allowing non-uniform priors, where a large prior mass in a certain region indicates that we want to achieve a more accurate estimate in that region, in exchange of slower convergence in a region with smaller prior mass. The prior structure reflects our belief that the true density is more likely to have a certain form than some alternatively forms. Therefore the theoretical analysis should also imply a more accurate estimate when we are lucky enough to guess the true density q correctly by putting a large prior mass around it. As we will see later, finite-sample convergence bounds derived in this paper using KL-complexity have this behavior.

Next we prove a simple information theoretical inequality using the KL-complexity of randomized estimators, which forms the basis of our analysis. 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 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 [7, 9].

Proposition 2.1 *Assume that $f(\theta)$ is a measurable real-valued function on Γ, and $w(\theta)$ is a density with respect to π, we have*

$$\mathbf{E}_\pi\, w(\theta) f(\theta) \leq D_{KL}(w d\pi || d\pi) + \ln \mathbf{E}_\pi \exp(f(\theta)).$$

Remark 2.1 *The above convex duality also has a straight-forward information theoretical interpretation: consider $v(\theta) = \exp(f(\theta))/\mathbf{E}_\pi \exp(f(\theta))$. Since $\mathbf{E}_\pi v(\theta) = 1$, we can regard it*

as a density with respect to π. Using this definition, it is easy to verify that the inequality in Proposition 2.1 can be rewritten as $D_{KL}(wd\pi||vd\pi) \geq 0$, which is a well-known information theoretical inequality, and follows easily from the Jensen's inequality.

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

Lemma 2.1 *Consider any posterior $\hat{w}_X(\theta)$. 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 2.1, 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 the 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 2.2 *The importance of the above inequality is that the left hand side is a quantity that involves an arbitrary posterior $\hat{w}_X$. The right hand side is a numerical constant independent of the estimator $\hat{w}_X$. Therefore the inequality gives a bound that can be applied to an arbitrary randomized estimator. The remaining issue is merely how to interpret the resulting bound, which we shall focus on later in the paper.*

Remark 2.3 *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 [14]. The specific decoupling mechanism using Proposition 2.1 appeared in [7, 9] for related problems. A simplified form of Lemma 2.1 was used in [15] to analyze Bayesian posterior distributions.*

The following bound is a straight-forward consequence of Lemma 2.1. 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 2.1 (Posterior Averaging Bounds) *Using the notation of Lemma 2.1. 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 2.1, with $L_X(\theta) = -\sum_{i=1}^n \ell_\theta(X_i)$. If we define

$$\begin{aligned}\hat{L}(X) =& \mathbf{E}_\pi \hat{w}_X(\theta)(L_X(\theta) - \alpha \ln \mathbf{E}_X e^{\beta L_X(\theta)}) - D_{KL}(\hat{w}_X d\pi || d\pi) \\ =& \mathbf{E}_\pi \hat{w}_X(\theta)(-\sum_{i=1}^n \ell_\theta(X_i) - n\alpha \ln \mathbf{E}_q e^{-\beta \ell_\theta(x)}) - D_{KL}(\hat{w}_X d\pi || d\pi),\end{aligned}$$

then by Lemma 2.1, 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 2.1. 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 2.4 *The special case of Theorem 2.1 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.*

3 Information complexity minimization

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:

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

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 this estimator using Theorem 2.1. 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.

Proposition 3.1 *We have the following inequalities* $\forall \rho \in [0,1]$*:*

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

Proof. We prove the first half of the two inequalities. Due to the symmetry $D_\rho(q||p) = D_{1-\rho}(p||q)$, we only need to consider the case $\rho \leq 1/2$. The proof of the second half (with $\rho \geq 1/2$) is identical except the sign in the Taylor expansion step is reversed.

We use Taylor expansion: let $x = \frac{p^{1/2}-q^{1/2}}{q^{1/2}}$, then $x \geq -1$, and there exists $\xi > -1$ such that

$$(1+x)^{2\rho} = 1 + 2\rho x + \rho(2\rho-1)(1+\xi)^{2\rho-2}x^2 \leq 1 + 2\rho x.$$

Now taking expectation with respect to q, we obtain:

$$\mathbf{E}_q \left(\frac{p}{q}\right)^\rho = \mathbf{E}_q \left(1 + \frac{p^{1/2}-q^{1/2}}{q^{1/2}}\right)^{2\rho} \leq 1 + 2\rho \mathbf{E}_q \frac{p^{1/2}-q^{1/2}}{q^{1/2}}.$$

Re-arranging, we obtain $2\rho(\frac{1}{4}D_{1/2}(q||p)) \leq \rho(1-\rho)D_\rho(q||p)$. $\square$

3.1 A general convergence bound

The following general theorem is an immediate consequence of Theorem 2.1. Most of our later discussions can be considered as interpretations of this theorem under various different conditions.

Theorem 3.1 *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. Consider an arbitrary data-independent density $w(\theta) \in S$ with respect to π, using (4), we can obtain from Theorem 2.1 the following chain of equations:

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

where $R_\lambda(w)$ is defined in (3). Note that the first inequality uses the fact $-\ln(1-x) \geq x$. The second inequality follows from Theorem 2.1 with the choice $\ell_\theta(x) = \rho \ln \frac{q(x)}{p(x|\theta)}$ and $\beta = 1$. The third inequality follows from the definition of $\hat{w}_X^S$ in (1). $\Box$

Remark 3.1 *If $\gamma = \rho$ in Theorem 3.1, then we also require $\lambda\gamma = 1$, and let $\lambda' = 0$.*

Although the bound in Theorem 3.1 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 [11]). 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 3.1 by choosing specific parameters so that both the second term and the third term vanish:

Corollary 3.1 *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 3.1. $\Box$.

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.

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

Lemma 3.1 $\forall \lambda' \geq 1$: $\mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S) \geq -\frac{\lambda'}{n} \geq 0$.

Proof. Applying the convex duality in Proposition 2.1 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

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

which proves the lemma. □

By combining the above estimate with Theorem 3.1, we obtain the following refinement of Corollary 3.1.

Corollary 3.2 *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 3.1. Note that in this case, $\lambda' = 1$, and hence by Lemma 3.1, $\mathbf{E}_X \hat{R}_{\lambda'}(\hat{w}_X^S) \geq 0$. □

Note that Lemma 3.1 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 3.1 *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 3.2 *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 3.2 *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. The proof is similar to that of Lemma 3.1, 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 $\{a_j\}$: $(\sum_j a_j)^{\lambda'} \leq \sum_j a_j^{\lambda'}$. $\square$

Combine the above estimate with Theorem 3.1, 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 3.3 *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 3.1, and apply Lemma 3.2. $\square$

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 3.3 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 prior π is fast enough on the discretization so that $\sum_j \pi(\Gamma_j)^{(\gamma-1)/(\gamma-\rho)} < +\infty$.

3.3 Weak convergence bound

The case of $\lambda = 1$ is related to a number of important estimation methods in statistical applications. However, for an arbitrary prior π without any additional assumption such as the fast decay condition in Corollary 3.3, 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 3.2). See Section 4.4 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 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 3.2 *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. Although we can work with Theorem 2.1, slightly better and simpler bounds can be obtained by working with the basic inequality in Lemma 2.1. The first half of the proof, leading to (6), is very similar to that of Theorem 2.1. The second half is very similar to that of Theorem 3.1.

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 2.1, we have

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

That is, 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

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

where the second inequality follows from the definition of $\hat{w}_X^S$ in (1). This inequality plays the same role of Theorem 2.1 in the proof of Theorem 3.1.

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. □

4 MDL on discrete net

The minimum description length (MDL) method has been widely used in practice [8]. 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

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

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 3.

4.1 Modified MDL under global entropy condition

Consider the case $\lambda > 1$ in (7). We can obtain the following theorem from Corollary 3.2.

Theorem 4.1 *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 $\lambda > 1$, which is a direct consequence of Theorem 4.1.

To the author's knowledge, the only previous result relatively close to the simple form presented in Theorem 4.1 is a bound in [6] on maximum-likelihood estimate over a finite-net. The bound obtained there applies to the case where the size of Γ is N with uniform prior $\pi_j = 1/N$:

$$\mathbf{E}_X D_{1/2}(q||p_{\hat{k}}) \leq 2 \inf_k \left[D_{KL}(q||p_k) + \frac{2}{n} \ln \frac{1}{N} \right]. \tag{8}$$

Clearly this result is identical to that inferred by Theorem 4.1 with $\lambda = 2$ and $\rho = 1/2$.

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 resulted from (8) matches the minimax rate such as those in [13].

4.2 Local entropy analysis

Although the bound based on the index of resolvability in Theorem 4.1 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 the bound in (8), which relies on the index of resolvability, 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 appeared in [3]. For non-parametric problems, it was pointed out in [13] 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 4.1) 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 3.1. We include a simplified version below, which can be obtained by choosing $\alpha = 1/2$, and $\gamma = \rho = 1/\lambda$.

Theorem 4.2 *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 4.2 implies a convergence rate of the correct order $O(1/n)$.

4.3 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 4.1 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 straight-forward interpretation of Corollary 3.3, where we consider the family itself as an 0-upper discretization: $\Gamma_i = \{p_i\}$:

Theorem 4.3 *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 4.3 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.

4.4 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 $\mathcal{X}$ be consisted 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

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

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 4.3. 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.

4.5 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 3.2, which we shall restate here:

Theorem 4.4 *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.

5 Bayesian posterior distributions

Assume we observe n-samples $X = \{X_1, \ldots, X_n\} \in \mathcal{X}^n$, independently drawn from the true underlying distribution Q with density q. As mentioned earlier, we call any probability density $\hat{w}_X(\theta)$ with respect to π that depends on the observation X (and measurable on $\mathcal{X}^n \times \Gamma$) a posterior distribution. $\forall \gamma > 0$, we define a generalized Bayesian posterior $\pi_\gamma(\cdot|X)$ with respect to π as:

$$\pi_\gamma(\theta|X) = \frac{\prod_{i=1}^n p^\gamma(X_i|\theta)}{\int_\Gamma \prod_{i=1}^n p^\gamma(X_i|\theta) d\pi(\theta)}. \tag{9}$$

We call π_γ the γ-Bayesian posterior. The standard Bayesian posterior is denoted as $\pi(\cdot|X) = \pi_1(\cdot|X)$.

The key starting point of our analysis is the following simple observation that relates the Bayesian posterior to an instance of information complexity minimization which we have already analyzed in this paper. The proof is rather straight-forward, which we shall skip. In fact, the inequality is essentially the same as that of Proposition 2.1 with $f(\theta) = \frac{1}{\lambda}\sum_{i=1}^n \ln p(X_i|\theta)$.

Proposition 5.1 *Consider prior π and $\lambda > 0$. Then*

$$\hat{R}_\lambda(\pi_{1/\lambda}(\cdot|X)) = -\lambda \ln \mathbf{E}_\pi \exp\left(\frac{1}{\lambda}\sum_{i=1}^n \ln \frac{p(X_i|\theta)}{q(X_i)}\right) = \inf_w \hat{R}_\lambda(w),$$

where $\hat{R}_\lambda(w)$ is defined in (2), and the inf *on the right hand side is over all possible densities w with respect to the prior π.*

The above Proposition indicates that the generalized Bayesian posterior can be regarded as a minimum information complexity estimator (1) with S consisted of all possible densities. Therefore results parallel to those of MDL can be obtained.

5.1 Generalized Bayesian methods

Instead of the index of resolvability complexity measure for MDL, the complexity of Bayesian-like methods is controlled by the *Bayesian resolvability* defined as:

$$r_{\lambda,n}(q) = \inf_w \left[\mathbf{E}_\pi w(\theta) D_{KL}(q||p(\cdot|\theta)) + \frac{\lambda}{n} D_{KL}(wd\pi||d\pi)\right] = -\frac{\lambda}{n}\ln \mathbf{E}_\pi e^{-\frac{n}{\lambda}D_{KL}(q||p(\cdot|\theta))}. \tag{10}$$

The following proposition gives a simple and intuitive estimate of the Bayesian index of resolvability. This bound implies that the Bayesian resolvability can be estimated using local properties of the prior π around the true density q. The quantity is small as long as there is a positive prior mass in a small KL-ball around the truth q.

Proposition 5.2 *The Bayesian resolvability defined in (10) can be upper bounded as:*

$$r_{\lambda,n}(q) \leq \inf_{\epsilon>0}\left[\epsilon - \frac{\lambda}{n}\ln \pi(\{p \in \Gamma : D_{KL}(q||p) \leq \epsilon\})\right].$$

Proof. $\forall \epsilon > 0$, we simply note that $\mathbf{E}_\pi e^{-\frac{n}{\lambda} D_{KL}(q||p(\cdot|\theta))} \geq e^{-\frac{n}{\lambda}\epsilon} \pi(\{p \in \Gamma : D_{KL}(q||p) \leq \epsilon\})$. Now taking logarithm and using (10), we obtain the desired inequality. □

The following bound is a direct consequence of Corollary 3.2.

Theorem 5.1 *Consider the generalized Bayesian posterior $\pi_{1/\lambda}(\theta|X)$ defined in (9) with $\lambda > 1$. Then $\forall \rho \in (0, 1/\lambda]$:*

$$\mathbf{E}_X \pi_{1/\lambda}(\theta|X) D_\rho(q||p(\cdot|\theta)) \leq -\frac{\lambda}{\rho(\lambda-1)n} \ln \mathbf{E}_\pi \exp\left(-\frac{n}{\lambda} D_{KL}(q||p(\cdot|\theta))\right).$$

The above theorem gives a general convergence bound on the γ-Bayesian method with $\gamma < 1$, depending only on the globally defined Bayesian resolvability. Note that similar to Theorem 4.2 for the MDL case, a bound using a localized Bayesian resolvability can also be obtained.

Theorem 5.1 immediately implies the concentration of generalized Bayesian posterior. To see this, we denote the posterior mass outside an ϵ D_ρ-ball around q as

$$\pi_{1/\lambda}(\{p \in \Gamma : D_\rho(q||p) \geq \epsilon\}|X).$$

Using the bound in Theorem 5.1 and Proposition 5.2, we obtain the following bound, which indicates that with large probability, the generalized Bayesian posterior concentrate in a D_ρ-ball of size $O(\epsilon_{\pi,n})$ around the truth.

Corollary 5.1 *Let $\lambda > 1$ and $\rho \in (0, 1/\lambda]$. Then $\forall \gamma > 0$:*

$$\mathbf{E}_X \pi_{1/\lambda}(\{p \in \Gamma : D_\rho(q||p) \geq \gamma\epsilon_{\pi,n}\}|X) \leq \frac{2}{\gamma\rho(\lambda-1)},$$

where the critical prior-mass radius $\epsilon_{\pi,n} = \inf\{\epsilon : \epsilon \geq -\frac{\lambda}{n} \ln \pi(\{p \in \Gamma : D_{KL}(q||p) \leq \epsilon\})\}$.

Proof. Apply the Markov inequality:

$$\mathbf{E}_X \pi_{1/\lambda}(\theta|X) D_\rho(q||p) \geq \mathbf{E}_X \pi_{1/\lambda}(\{p \in \Gamma : D_\rho(q||p) \geq \epsilon\}|X)\epsilon.$$

Now using Theorem 5.1 and Proposition 5.2, we obtain

$$\begin{aligned}
&\pi_{1/\lambda}(\{p \in \Gamma : D_\rho(q||p) \geq \gamma\epsilon_{\pi,n}\}|X)\gamma\epsilon_{\pi,n} \\
\leq &\frac{1}{\rho(\lambda-1)} \left[\epsilon_{\pi,n} - \frac{\lambda}{n} \ln \pi(\{p \in \Gamma : D_{KL}(q||p) \leq \epsilon_{\pi,n}\})\right] \leq \frac{2\epsilon_{\pi,n}}{\rho(\lambda-1)}.
\end{aligned}$$

Now divide both sides by $\gamma\epsilon_{\pi,n}$. □

5.2 The standard Bayesian method

For the standard Bayesian posterior distribution, it is not possible to bound its convergence using only the Bayesian resolvability. The reason is the same as in the MDL case. In fact, it is immediately obvious that the example for MDL can also be applied here (also see [2]).

Therefore in order to obtain a rate of convergence (and concentration) for the standard Bayesian method, additional assumptions are necessary. Clearly a bound similar to Theorem 4.3 using ϵ-upper discretization can be easily obtained from Corollary 3.3. However, we shall only list a simplified version using ϵ-upper bracketing cover of Γ since its interpretation is more intuitive. We shall then use the interpretation in a corresponding result concerning the concentration of Bayesian posterior distributions.

Theorem 5.2 *Consider the Bayesian posterior $\pi(\cdot|X) = \pi_1(\cdot|X)$ defined in (9). For all $\epsilon > 0$, let $N(\Gamma, \epsilon)$ be its ϵ-upper bracketing covering number. Then $\forall \rho \in (0,1)$, we have:*

$$\mathbf{E}_X \mathbf{E}_\pi \pi(\theta|X) D_\rho(q||p(\cdot|\theta)) \leq \frac{\ln \mathbf{E}_\pi^{-1/n} e^{-n D_{KL}(q||p(\cdot|\theta))}}{\rho(1-\rho)} + \frac{1}{\rho} \inf_{\epsilon>0} \left[\frac{\ln N(\Gamma, \epsilon)}{n} + \ln(1+\epsilon) \right].$$

Similar to Corollary 5.1, we obtain the following concentration result for the standard Bayesian posterior distribution.

Corollary 5.2 *Let $\rho \in (0,1)$, then*

$$\mathbf{E}_X \pi_{1/\lambda}(\{p \in \Gamma : D_\rho(q||p) \geq \gamma(\epsilon_{\pi,n} + (1-\rho)\epsilon_{upper,n})\}|X) \leq \frac{2}{\gamma\rho(1-\rho)},$$

where the critical prior-mass radius $\epsilon_{\pi,n} = \inf\{\epsilon : \epsilon \geq -\frac{1}{n} \ln \pi(\{p \in \Gamma : D_{KL}(q||p) \leq \epsilon\})\}$, and the critical upper-bracketing radius $\epsilon_{upper,n} = \inf\{\epsilon : \epsilon \geq \frac{1}{n} \ln N(\Gamma, \epsilon))\}$.

Note that the result implies that if the critical upper-bracketing radius $\epsilon_{upper,n}$ is at the same (or smaller) order of the critical prior-mass radius $\epsilon_{\pi,n}$, then with large probability, the standard Bayesian posterior distribution will concentrate in a D_ρ ball of size $\epsilon_{\pi,n}$. In this case, the standard Bayesian posterior has the same rate of convergence when compared with the generalized Bayesian posterior with $\lambda > 1$. However, if $\epsilon_{upper,n}$ is large, then the standard Bayesian method may fail to concentrate in a small D_ρ ball around the truth q, even when the critical prior radius $\epsilon_{\pi,n}$ is small. This can be easily seen from the same counter-example used to illustrate the slow convergence of the standard MDL.

Although the standard Bayesian posterior distribution may not concentrate even when $\epsilon_{\pi,n}$ is small, Theorem 3.2 implies that the Bayesian density estimator $\mathbf{E}_\pi \pi(\theta|X) p(\cdot|X)$ is close to q in the sense of weak convergence.

The consistency theorem given in [2] also relies on the upper entropy number $N(\Gamma, \epsilon)$. However, no convergence rates were established. Therefore Corollary 5.1 in some sense can be regarded as a refinement of their analysis using their covering definition. Other kinds of covering numbers (e.g. Hellinger covering) can also be used in convergence analysis of non-parametric Bayesian methods [5, 10].

The convergence rate result in [10] employed the chaining technique from empirical process, which can possibly lead to suboptimal convergence rates when the covering number grows relatively fast as the scale $\epsilon \to 0$. We shall focus on [5], which employed techniques from hypothesis testing in [4]. The resulting convergence theorem from their analysis cannot be as simply stated as those in this paper. Moreover, some of their conditions can be relaxed. In the following, we shall prove an improvement of their result based on the following lemma.

Lemma 5.1 *Consider a partition of Γ as the union of countably many disjoint measurable sets Γ_j $(j = 1, \ldots)$. We have $\forall \rho \in (0,1)$ and $\gamma \geq 1$:*

$$-\mathbf{E}_X \sum_j \pi(\Gamma_j|X) \ln \mathbf{E}_{X'} \left(\frac{p_j(X')}{q(X')} \right)^\rho \leq (\gamma - \rho) \ln \sum_j \pi(\Gamma_j)^{\frac{\gamma-1}{\gamma-\rho}} - \gamma \ln \sum_j \pi(\Gamma_j) e^{-D_{KL}(q(X)||p_j(X))},$$

where $\pi(\Gamma_j) = \int_{\Gamma_j} d\pi(\theta)$ is the prior probability of Γ_j, $q(X) = \prod_{i=1}^n q(X_i)$ is the true density of X, $p_j(X) = \frac{1}{\pi(\Gamma_j)} \int_{\Gamma_j} \prod_{i=1}^n p(X_i|\theta) d\pi(\theta)$ is the mixture density over Γ_j under π, and $\pi(\Gamma_j|X) = \frac{\pi(\Gamma_j) p_j(X)}{\sum_\ell \pi(\Gamma_\ell) p_\ell(X)}$ is the Bayesian posterior probability of Γ_j.

Proof. We shall apply Corollary 3.3 with a slightly different interpretation. Instead of considering X as n independent samples X_i as before, we simply regard it as one random variable by itself. Consider family Γ' which is consisted of discrete densities $p_j(X)$, with prior $\pi_j = \pi(\Gamma_j)$. This discretization itself can be regarded as a 0-upper discretization of Γ'. Also given X, it is easy to see that the Bayesian posterior on Γ' with respect to $\{\pi_j\}$ is $\hat{\pi}_j = \pi(\Gamma_j|X)$. We can thus apply a slightly strengthened version of Corollary 3.3 on Γ' (that is, we shall replace the left-hand side by the second inequality in Theorem 3.1), which leads to the stated bound (with the help of Equation 10). □

Now to obtain the posterior concentration result from this bound, we require an additional estimate:

Lemma 5.2 *Let $q(X) = \prod_{i=1}^n q(X_i)$ and $p_j(X) = \frac{1}{\pi(\Gamma_j)} \int_{\Gamma_j} \prod_{i=1}^n p(X_i|\theta) d\pi(\theta)$, then*

$$\inf_{p \in \mathrm{co}(\Gamma_j)} \mathbf{E}_q \left(\frac{p(X_1)}{q(X_1)} \right)^\rho \leq \mathbf{E}_q^{1/n} \left(\frac{p_j(X)}{q(X)} \right)^\rho \leq \sup_{p \in \mathrm{co}(\Gamma_j)} \mathbf{E}_q \left(\frac{p(X_1)}{q(X_1)} \right)^\rho,$$

and

$$\inf_{p \in \mathrm{co}(\Gamma_j)} D_{KL}(q(X_1)||p(X_1)) \leq \frac{1}{n} D_{KL}(q(X)||p_j(X)) \leq \sup_{p \in \mathrm{co}(\Gamma_j)} D_{KL}(q(X_1)||p(X_1)),$$

where $\mathrm{co}(\Gamma_j)$ is the convex hull of densities in Γ_j.

Proof. The first two inequalities can be proved in the same as Lemma 4 on page 478 of [4], which dealt with the existence of tests under the Hellinger distance. The proof is relatively simple, and in fact similar to that of the last two inequalities. We shall thus only write down

the proof of the last two inequalities.

$$\begin{aligned} D_{KL}(q(X)||p_j(X)) =& \mathbf{E}_X \sum_{i=1}^{n} \mathbf{E}_{X_i} \ln \frac{q(X_i)}{\frac{\int_{\Gamma_j} \prod_{k=1}^{i} p(X_k|\theta) d\pi(\theta)}{\int_{\Gamma_j} \prod_{k=1}^{i-1} p(X_k|\theta) d\pi(\theta)}} \\ =& \mathbf{E}_X \sum_{i=1}^{n} D_{KL}\left(q(X_i)||\int_{\Gamma_j} w_i(\theta) p(X_i|\theta) d\pi(\theta)\right), \end{aligned}$$

where $w_i(\theta) = \frac{\prod_{k=1}^{i-1} p(X_k|\theta)}{\int_{\Gamma_j} \prod_{k=1}^{i-1} p(X_k|\theta) d\pi(\theta)}$. Since each $\int_{\Gamma_j} w_i(\theta) p(X_i|\theta) d\pi(\theta) \in \mathrm{co}(\Gamma_j)$, we obtain the last two inequalities. □

Combining the above two inequalities, we obtain the following result:

Theorem 5.3 *Using notations of Lemma 5.1 and Lemma 5.2. Consider a partition of* Γ *as the union of countably many disjoint measurable sets* Γ_j *(*$j = 1, \ldots$*). Then* $\forall \rho \in (0,1)$ *and* $\gamma \geq 1$*:*

$$\mathbf{E}_X \sum_j \pi(\Gamma_j|X) \inf_{p \in \mathrm{co}(\Gamma_j)} D_\rho(q||p) \leq \frac{(\gamma-\rho) \ln \sum_j \pi(\Gamma_j)^{\frac{\gamma-1}{\gamma-\rho}} - \gamma \ln \sum_j \pi(\Gamma_j) e^{-n \sup_{p \in \mathrm{co}(\Gamma_j)} D_{KL}(q||p)}}{\rho(1-\rho)n}.$$

Proof. We simply substitute the estimates of Lemma 5.2 into Lemma 5.1, and use the inequality $\rho(1-\rho) D_\rho(q||p) \leq -\ln \mathbf{E}_q (p(x)/q(x))^\rho$ (see Theorem 3.1). □

An immediate consequence of the above theorem is a result on the concentration of Bayesian posterior distributions that improves that of [5] and complements the ϵ-upper discretization based bound in Corollary 5.2. For simplicity, we only state a simplified version with a finite convex cover since it is more intuitive to interpret.

Corollary 5.3 *Let* $\epsilon_{\pi,n} = \inf\{\epsilon : \epsilon \geq -\frac{1}{n} \ln \pi(\{p \in \Gamma : D_{KL}(q||p) \leq \epsilon\})\}$*. Assume that* $\{p \in \Gamma : D_\rho(q||p) \geq \epsilon\}$ *can be covered by the union of* N_ϵ *measurable convex sets* Γ_j *(*$j = 1, \ldots$*) such that* $\inf_{p \in \cup_j \Gamma_j} D_\rho(q||p) \geq \epsilon/2$*. Then* $\forall \epsilon \geq \epsilon_{\pi,n} / \min(\rho, 1-\rho)$*, we have*

$$\mathbf{E}_X \pi(\{p \in \Gamma : D_\rho(q||p) \geq \epsilon\}|X) \leq \frac{(2-2\rho)\frac{\ln(N_\epsilon+2)}{n} + 4\epsilon_{\pi,n}}{\rho(1-\rho)\epsilon}.$$

Proof. Let $\Gamma_1 = \{p \in \Gamma : D_{KL}(q||p) \leq \epsilon_{\pi,n}\}$. Since $D_{KL}(q||p) = D_0(q||p)$, using Proposition 3.1, we know that $\forall \epsilon \geq \epsilon_{\pi,n} / \min(\rho, 1-\rho)$, $\Gamma_1 \subset \Gamma' = \{p \in \Gamma : D_\rho(q||p) < \epsilon\}$. Let $\Gamma_2 = \Gamma' - \Gamma_1$, and by assumption, it is clear that $\Gamma - \Gamma'$ can be partitioned into the union of N_ϵ disjoint measurable sets $\Gamma_3, \ldots, \Gamma_{N_\epsilon+2}$ such that $\inf_{p \in \cup_{j \geq 3} \mathrm{co}(\Gamma_j)} D_\rho(q||p) \geq \epsilon/2$. Therefore for this partition, we have

$$\mathbf{E}_X \sum_j \pi(\Gamma_j|X) \inf_{p \in \mathrm{co}(\Gamma_j)} D_\rho(q||p) \geq \mathbf{E}_X \pi(\Gamma - \Gamma'|X)\epsilon/2.$$

Let $\gamma = 1$, then $\ln \sum_j \pi(\Gamma_j)^{\frac{\gamma-1}{\gamma-\rho}} \leq \ln(N_\epsilon + 2)$. In addition, we have

$$-\ln \sum_j \pi(\Gamma_j) e^{-n \sup_{p \in \mathrm{co}(\Gamma_j)} D_{KL}(q||p)} \leq -\ln \pi(\Gamma_1) + n \sup_{p \in \mathrm{co}(\Gamma_1)} D_{KL}(q||p) \leq 2n\epsilon_{\pi,n}.$$

Combining the above estimates, and plug them into Theorem 5.3, we obtain the desired result. $\Box$

Clearly if $\frac{1}{n} \ln N_\epsilon = O(\epsilon_{\pi,n})$ for some $\epsilon = O(\epsilon_{\pi,n})$, then with large probability, Bayesian posterior distributions concentrate on a D_ρ ball of size $O(\epsilon_{\pi,n})$ around q. Note that this result relaxes a condition of [5], where our definition of $\epsilon_{\pi,n}$ was replaced by the smaller ball $\{p \in \Gamma : D_{KL}(q||p) \leq \epsilon, \mathbf{E}_q \ln(\frac{q}{p})^2 \leq \epsilon\}$. Moreover, their covering definition N_ϵ does not apply to arbitrary convex sets.

Finally we shall mention that concentration bounds of the forms in Corollary 5.2 and Corollary 5.3 are known to produce optimal convergence rates for many non-parametric problems in the minimax sense. See [5, 10] for illustrations. It is also useful to note that Corollary 5.1 requires less assumptions to achieve good convergence rates, implying that generalized Bayesian methods are more stable than the standard Bayesian method.

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 MDL and Bayesian posterior distributions.

The KL-complexity measure used in this paper generalizes the concept of ϵ-entropy that has become an essential tool in many traditional statistical convergence analysis. Using the new concept, we introduced an information theoretical inequality that is not only very general but also leads to simple convergence bounds not obtainable using previous approaches. In many cases, the algebraic approach based on the new information theoretical inequality can replace convergence analysis based on ϵ-entropy and hypothesis testing. We note that although the theory of hypothesis testing is very powerful, it is not very useful for producing simple and clean convergence bounds such as those obtained in this paper.

In this regard, our approach provides a useful technical tool from a novel perspective. Moreover since the information theoretical inequality introduced in this paper gives more precise algebraic information than many previous approaches (such as those based on the theory of hypothesis testing), one can obtain substantial improvements over previous results in many interesting cases (as demonstrated by this paper). Such improvements suggest that the new machinery introduced in this paper not only can reproduce old results, but also can lead to new bounds that are not obtainable from old approaches.

It is not difficult to see that the newly introduced information theoretical inequality can be used to study general loss minimization estimation methods such as regression and classification. However, since the problem of density estimation is by itself rich enough, we shall leave the general case to another article. Moreover, since the fundamental inequality in Lemma 2.1 is an exponential bound, our analysis can also be applied to model selection

methods where the risk minimization formula under consideration contains tuning parameters to be adjusted for optimal estimation accuracy. In fact, one can already see this idea from the proof of Theorem 3.2, where we added two exponential bounds in Lemma 2.1 with different parameters. The general application of this idea isn't explored carefully in this paper since we have focused on more basic issues.

For density estimation, we considered a general randomized estimation procedure based on log-loss minimization with KL-complexity regularization. We call this family of methods, which include MDL and Bayesian methods, information complexity minimization. Using a posterior averaging bound derived from the basic information theoretical inequality, we obtained a number of general convergence results for information complexity minimization, and applied them to MDL and Bayesian methods.

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. For Bayesian posterior, this quantity (Bayesian resolvability) is well-behaved if we put a not too small prior mass in a small KL-ball around q. We have also demonstrated through an example that the standard MDL (and Bayesian posterior) 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.

Therefore although the standard Bayesian method is "optimal" in a certain averaging sense, its behavior is heavily dependent on the regularity of the prior distribution globally. Intuitively, the standard Bayesian method can put too much emphasis on the difficult part of the prior distribution, which degrades the estimation quality in the easier parts where we are actually more interested in. Therefore even if one is able to guess the true distribution by putting a large prior mass around its neighborhood, the Bayesian method can still ill-behave if one accidentally makes bad choices elsewhere. It is thus arguably more difficult to design good Bayesian priors. The new theoretical insights obtained here imply that unless one completely understands the impact of the prior, it is much safer to use a generalized Bayesian method with $\lambda > 1$. This interesting observation may potentially lead to useful robust Bayesian statistical procedures.

Acknowledgments

The author would like to thank Andrew Barron for helpful discussions that motivated some ideas presented in this paper, and Matthias Seeger for useful comments on an earlier version of the paper.

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] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[5] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[6] J.Q. Li. *Estimation of Mixture Models*. PhD thesis, The Department of Statistics. Yale University, 1999.

[7] Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

[8] J. Rissanen. *Stochastic complexity and statistical inquiry*. World Scientific, 1989.

[9] M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *JMLR*, 3:233–269, 2002.

[10] Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001.

[11] S.A. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

[12] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

[13] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27:1564–1599, 1999.

[14] Tong Zhang. Theoretical analysis of a class of randomized regularization methods. In *COLT 99*, pages 156–163, 1999.

[15] Tong Zhang. Learning bounds for a generalized family of Bayesian posterior distributions. In *NIPS 03*, 2004. to appear.