Localized Upper and Lower Bounds for Some Estimation Problems

Tong Zhang

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

Abstract. We derive upper and lower bounds for some statistical estimation problems. The upper bounds are established for the Gibbs algorithm. The lower bounds, applicable for all statistical estimators, match the obtained upper bounds for various problems. Moreover, our framework can be regarded as a natural generalization of the standard minimax framework, in that we allow the performance of the estimator to vary for different possible underlying distributions according to a pre-defined prior.

1 Introduction

The purpose of this paper is to derive upper and lower bounds for some prediction problems in statistical learning. The upper bounds are obtained for the Gibbs algorithm. The lower bounds are obtained from some novel applications of well-known information theoretical inequalities (specifically, data-processing theorems). We show that the upper bounds and lower bounds have very similar forms, and match under various conditions.

In statistical prediction, we have input space $\mathcal{X}$ and output space $\mathcal{Y}$, and a space of predictors $\mathcal{G}$. For any $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, and any predictor $\theta \in \mathcal{G}$, we incur a loss $L_\theta(X,Y) = L_\theta(Z)$, where $Z = (X,Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Consider a probability measure D on $\mathcal{Z}$. Our goal is to find θ from a random sample $\hat{Z}$ from D, such that the loss $\mathbf{E}_Z L_\theta(Z)$ is small, where $\mathbf{E}_Z$ is the expectation with respect to D.

In the standard learning theory, we consider n random samples instead of one sample. The two formulations are in fact equivalent. To see this, consider $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_n\}$. Let $L_\theta(Z) = \sum_{i=1}^n L_{i,\theta}(X_i, Y_i)$. If $Z_i = (X_i, Y_i)$ are independent random variables, then it follows that $\mathbf{E}_Z L_\theta(Z) = \sum_{i=1}^n \mathbf{E}_{Z_i} L_{i,\theta}(X_i, Y_i)$. We shall thus focus on the one-sample case first without loss of generality.

In this paper, we consider randomized estimators. They are defined with respect to a prior π on $\mathcal{G}$, which is a probability measure on $\mathcal{G}$. For a randomized estimation method, given sample $\hat{Z}$ from D, we select θ from $\mathcal{G}$ based on a sample-dependent probability measure $d\hat{\pi}_{\hat{Z}}(\theta)$ on $\mathcal{G}$, In this paper, we shall call such a sample-dependent probability measure as a *posterior randomization*

measure (or simplified as posterior). The word posterior in this paper is not necessarily the Bayesian posterior distribution in the traditional sense. For notational simplicity, we also use the symbol $\hat{\pi}$ to denote $\hat{\pi}_{\hat{Z}}$. The randomized estimator associated with a posterior randomization measure is thus completely determined by its posterior $\hat{\pi}$. Its *posterior averaging risk* is the averaged risk of the randomized estimator drawn from this posterior randomization measure, which can be defined as

$$\mathbf{E}_{\theta\sim\hat{\pi}}\mathbf{E}_Z L_\theta(Z) = \mathbf{E}_Z \int L_\theta(Z) d\hat{\pi}(\theta).$$

In this paper, we are interested in estimating this average risk for an arbitrary posterior $\hat{\pi}$. The statistical complexity of this randomized estimator $\hat{\pi}$ will be measured by its *KL-entropy* respect to the prior, which is defined as:

$$D_{KL}(\hat{\pi}||\pi) = \int_{\mathcal{G}} \ln \frac{d\hat{\pi}(\theta)}{d\pi} d\hat{\pi}(\theta), \tag{1}$$

assuming it exists.

2 Analysis of the Gibbs Algorithm

Theoretical properties of the Gibbs algorithm have been studied by various researchers. In particular, some bounds obtained in this paper are related (but not identical) to independently obtained results in [3]. The main technical tool used here, based on the following lemma, is simpler and more general. See [8, 9] for its proof.

Lemma 1. *Consider randomized estimation, where we select posterior $\hat{\pi}$ on $\mathcal{G}$ based on $\hat{Z}$, with π a prior. Consider a real-valued function $L_\theta(Z)$ on $\mathcal{G}\times\mathcal{Z}$.*

$$c(\alpha) = \ln \mathbf{E}_{\theta\sim\pi}\mathbf{E}_Z^\alpha e^{-L_\theta(Z)},$$

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

$$-(1-\alpha)\mathbf{E}_{\theta\sim\hat{\pi}} \ln \mathbf{E}_Z\, e^{-L_\theta(Z)} \le \mathbf{E}_{\theta\sim\hat{\pi}} L_\theta(\hat{Z}) + D_{KL}(\hat{\pi}||\pi) + c(\alpha) + t.$$

Moreover, we have the following expected risk bound:

$$-(1-\alpha)\mathbf{E}_{\hat{Z}}\mathbf{E}_{\theta\sim\hat{\pi}} \ln \mathbf{E}_Z\, e^{-L_\theta(Z)} \le \mathbf{E}_{\hat{Z}}\left[\mathbf{E}_{\theta\sim\hat{\pi}} L_\theta(\hat{Z}) + D_{KL}(\hat{\pi}||\pi)\right] + c(\alpha).$$

If we choose $\alpha = 0$ in Lemma 1, then $c(\alpha) = 0$. However, choosing $\alpha \in (0,1)$ is useful for some parametric problems, where we would like to obtain a convergence rate of the order $O(1/n)$. In such cases, the choice of $\alpha = 0$ would lead to a rate of $O(\ln n/n)$, which is suboptimal.

We shall consider the case of n iid samples $\hat{Z} = (\hat{Z}_1, \ldots, \hat{Z}_n) \in \mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$, where $\mathcal{Z}_1 = \cdots = \mathcal{Z}_n$. The loss function is $L_\theta(\hat{Z}) = \rho \sum_{i=1}^n \ell_\theta(\hat{Z}_i)$,

where ℓ is a function on $\mathcal{G} \times \mathcal{Z}_1$ and $\rho > 0$ is a constant. The Gibbs algorithm is a randomized estimator $\hat{\pi}_\rho$ defined as:

$$d\hat{\pi}_\rho = \frac{\exp(-\rho \sum_{i=1}^n \ell_\theta(\hat{Z}_i))}{\mathbf{E}_{\theta\sim\pi} \exp(-\rho \sum_{i=1}^n \ell_\theta(\hat{Z}_i))}\, d\pi. \tag{2}$$

It is not difficult to verify that it minimizes the right hand side of Lemma 1 among all probability distributions on $\mathcal{G}$:

$$\hat{\pi}_\rho = \arg\inf_{\hat{\pi}} \left[\rho \mathbf{E}_{\theta\sim\hat{\pi}} \sum_{i=1}^n \ell_\theta(\hat{Z}_i) + D_{KL}(\hat{\pi}||\pi) \right]. \tag{3}$$

Lemma 2. *Define resolvability*

$$r_\rho = -\frac{1}{\rho n} \ln \mathbf{E}_{\theta\sim\pi} e^{-\rho n \mathbf{E}_{Z_1} \ell_\theta(Z_1)}.$$

Then $\forall \alpha \in [0,1)$, the expected generalization performance of the Gibbs algorithm (2) can be bounded as

$$-\mathbf{E}_{\hat{Z}}\, \mathbf{E}_{\theta\sim\hat{\pi}_\rho} \ln \mathbf{E}_{Z_1} e^{-\rho \ell_\theta(Z_1)} \le \frac{\rho}{1-\alpha} \left[r_\rho + \frac{1}{\rho n} \ln \mathbf{E}_{\theta\sim\pi} \mathbf{E}_{Z_1}^{\alpha n} e^{-\rho \ell_\theta(Z_1)} \right].$$

Proof. We obtain from (3)

$$\begin{aligned}
&\mathbf{E}_{\hat{Z}} \left[\rho \mathbf{E}_{\theta\sim\hat{\pi}_\rho} \sum_{i=1}^n \ell_\theta(\hat{Z}_i) + D_{KL}(\hat{\pi}_\rho||\pi) \right] \\
\le &\inf_{\pi'} \mathbf{E}_{\hat{Z}} \left[\rho \mathbf{E}_{\theta\sim\pi'} \sum_{i=1}^n \ell_\theta(\hat{Z}_i) + D_{KL}(\pi'||\pi) \right] \\
\le &\inf_{\pi'} \left[\rho n \mathbf{E}_{\theta\sim\pi'} \mathbf{E}_{\hat{Z}_i} \ell_\theta(\hat{Z}_i) + D_{KL}(\pi'||\pi) \right] = \rho n r_\rho.
\end{aligned}$$

Let $L_\theta(\hat{Z}) = \rho \sum_{i=1}^n \ell_\theta(\hat{Z}_i)$. Substituting the above bound into the right hand side of the second inequality of Lemma 1, and using the fact that $\mathbf{E}_Z e^{-L_\theta(Z)} = \mathbf{E}_{Z_1}^n e^{-\rho \ell_\theta(Z_1)}$, we obtain the desired inequality.

Theorem 1. *Consider the Gibbs algorithm in (2). Assume there exist positive constants K such that $\forall \theta$:*

$$\mathbf{E}_{Z_1} \ell_\theta(Z_1)^2 \le K \mathbf{E}_{Z_1} \ell_\theta(Z_1).$$

Under either of the following conditions:

- *Bounded loss: $\exists M \ge 0$ s.t. $-\inf_{\theta,Z_1} \ell_\theta(Z_1) \le M$; let $\beta_\rho = 1 - K(e^{\rho M} - \rho M - 1)/(\rho M^2)$.*

– *Bernstein loss:* $\exists M, b > 0$ *s.t.* $\mathbf{E}_{Z_1}(-\ell_\theta(Z_1))^m \leq m! M^{m-2} K b \mathbf{E}_{Z_1}\ell_\theta(Z_1)$ *for all* $\theta \in \mathcal{G}$ *and integer* $m \geq 3$*; let* $\beta_\rho = 1 - K\rho(1-\rho M + 2b\rho M)/(2-2\rho M)$.

Assume we choose a sufficiently small ρ *such that* $\beta_\rho > 0$. *Let the (true) expected loss of* θ *be* $R(\theta) = \mathbf{E}_{Z_1}\ell_\theta(Z_1)$*, then the expected generalization performance of the Gibbs algorithm is bounded by*

$$\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \leq \frac{1}{(1-\alpha)\beta_\rho}\left[r_\rho + \frac{1}{\rho n}\ln \mathbf{E}_{\theta\sim\pi} e^{-\alpha\rho\beta_\rho n R(\theta)}\right], \tag{4}$$

where r_ρ *is the resolvability defined in Lemma 2.*

Proof. (Sketch) Under the bounded-loss condition, we can use the following moment generating function estimate:

$$\begin{aligned}\ln \mathbf{E}_{Z_1} e^{-\rho\ell_\theta(Z_1)} \leq & -\rho\mathbf{E}_{Z_1}\ell_\theta(Z_1) + \frac{e^{\rho M} - \rho M - 1}{M^2}\mathbf{E}_{Z_1}\ell_\theta(Z_1)^2 \\ \leq & -\left(\rho - \frac{K}{M^2}(e^{\rho M} - \rho M - 1)\right)\mathbf{E}_{Z_1}\ell_\theta(Z_1) = -\rho\beta_\rho \mathbf{E}_{Z_1}\ell_\theta(Z_1).\end{aligned}$$

Now substitute this bound into Lemma 2, and simplify, we obtain the desired result. The proof is similar for the Bernstein-loss condition (with appropriate logarithmic moment generating function estimate).

Remark 1. Since $(e^x - x - 1)/x \to 0$ as $x \to 0$, we know that the first condition $\beta_\rho > 0$ can be satisfied as long as we pick a sufficiently small ρ. In fact, using the inequality $(e^x - x - 1)/x \leq 0.5xe^x$ (when $x \geq 0$), we may also take $\beta_\rho = 1 - 0.5\rho K e^{\rho M}$ in the first condition of Theorem 1.

We shall now study consequences of (4) under some general conditions on the local prior structure $\pi(\epsilon) = \pi(\{\theta : R(\theta) \leq \epsilon\})$ around the best achievable parameter. For some specific forms of local prior conditions, convergence rates can be stated very explicitly.

Theorem 2. *If (4) holds with a non-negative function* $R(\theta)$*, then*

$$\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \leq \frac{\Delta(\alpha\beta_\rho, \rho n)}{(1-\alpha)\beta_\rho\rho n},$$

where

$$\Delta(a,b) = \ln \inf_{u,v}\left[\sup_{\epsilon\leq u}\frac{\max(0,\pi(\epsilon/a) - v)}{\pi(\epsilon)} + \inf_{\epsilon}\frac{v + (1-v)\exp(-bu)}{\pi(\epsilon)e^{-b\epsilon}}\right],$$

and $\pi(\epsilon) = \pi(\{\theta : R(\theta) \leq \epsilon\})$.

Proof. We have

$$r_\rho = -\frac{1}{\rho n}\ln \mathbf{E}_{\theta\sim\pi} e^{-\rho n R(\theta)} = -\frac{1}{\rho n}\ln \underbrace{\int \pi(\epsilon/(\rho n))\, e^{-\epsilon} d\epsilon}_{A}.$$

Similarly, the second term on the right hand side of (4) is

$$\frac{1}{\rho n}\ln \mathbf{E}_{\theta\sim\pi} e^{-\alpha\rho\beta_\rho n R(\theta)} = \frac{1}{\rho n}\ln \int \pi(\epsilon/(\alpha\beta_\rho\rho n))\; e^{-\epsilon} d\epsilon$$

$$\leq \frac{1}{\rho n}\ln\left[\underbrace{\int_0^{\rho n u} (\pi(\epsilon/(\alpha\rho\beta_\rho n)) - v)\; e^{-\epsilon} d\epsilon}_{B} + \underbrace{\left(v + (1-v)e^{-\rho n u}\right)}_{C}\right].$$

To finish the proof, we only need to show that $(B+C)/A \leq e^{\Delta(\alpha\beta_\rho,\rho n)}$.

Consider arbitrary real numbers u and v. From the expressions, it is easy to see that $B/A \leq \sup_{\epsilon\leq u} \frac{\max(0,\pi(\epsilon/(\alpha\beta_\rho))-v)}{\pi(\epsilon)}$. Moreover, since $A \geq \sup_\epsilon(\pi(\epsilon)e^{-\rho n\epsilon})$, we have $C/A \leq C/\sup_\epsilon(\pi(\epsilon)e^{-\rho n\epsilon})$. Combining these inequalities, we have $(B+C)/A \leq e^{\Delta(\alpha\beta_\rho,\rho n)}$. The desired bound is now a direct consequence of (4).

In the following, we give two simplified bounds, one with global entropy, which gives correct rate of convergence for non-parametric problems. The other bound is a refinement that uses localized entropy, useful for parametric problems. They direct consequences of Theorem 2.

Corollary 1 (Global Entropy Bound). *If (4) holds, then*

$$\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \leq \frac{\inf_\epsilon[\rho n\epsilon - \ln\pi(\epsilon)]}{\beta_\rho\rho n} \leq \frac{2\bar{\epsilon}_{global}}{\beta_\rho},$$

where $\pi(\epsilon) = \pi(\{\theta : R(\theta)\leq\epsilon\})$ *and* $\bar{\epsilon}_{global} = \inf\left\{\epsilon : \epsilon \geq \frac{1}{\rho n}\ln\frac{1}{\pi(\epsilon)}\right\}$.

Proof. For the first inequality, we take $v = 1$ in Theorem 2, and let $\alpha \to 0$. For the second inequality, we simply note from the definition of $\bar{\epsilon}_{global}$ that $\inf_\epsilon[\rho n\epsilon - \ln\pi(\epsilon)] \leq 2\rho n\bar{\epsilon}_{global}$.

Corollary 2 (Local Entropy Bound). *If (4) holds, then*

$$\mathbf{E}_{\hat{Z}}\,\mathbf{E}_{\theta\sim\hat{\pi}_\rho} R(\theta) \leq \frac{\bar{\epsilon}_{local}}{(1-\alpha)\beta_\rho},$$

where $\pi(\epsilon) = \pi(\{\theta : R(\theta)\leq\epsilon\})$, *and*

$$\bar{\epsilon}_{local} = \frac{2}{\rho n} + \inf\left\{\frac{\epsilon}{\alpha\beta_\rho} : \epsilon \geq \sup_{\epsilon'\in[\epsilon,2u]} \frac{\alpha\beta_\rho}{\rho n}\ln\left[\frac{\pi(\epsilon'/(\alpha\beta_\rho))}{\pi(\epsilon')} + \frac{\exp(-\rho n u)}{\pi(u)}\right]\right\}.$$

Proof. For the first inequality, we simply take $u = u_2$ and $v = \pi(u_1/(\alpha\beta_\rho))$ in Theorem 2, and use the following bounds

$$\sup_{\epsilon\leq u}\frac{\max(0,\pi(\epsilon/(\alpha\beta_\rho))-v)}{\pi(\epsilon)} \leq \sup_{\epsilon\in[u_1,u_2]}\frac{\pi(\epsilon/(\alpha\beta_\rho))}{\pi(\epsilon)},$$

$$\frac{v}{\sup_\epsilon(\pi(\epsilon)e^{-\rho n\epsilon})} \leq \frac{v}{ve^{-\rho n u_1/(\alpha\beta_\rho)}} = \exp(\frac{\rho n u_1}{\alpha\beta_\rho}),$$

$$\frac{(1-v)\exp(-\rho n u)}{\sup_\epsilon(\pi(\epsilon)e^{-\rho n\epsilon})} \leq \frac{\exp(-\rho n u_2)}{\pi(u_2/2)e^{-\rho n u_2/2}} = \frac{\exp(-\rho n u_2/2)}{\pi(u_2/2)}.$$

For the second inequality, we let $u_2 = 2u$ and $u_1/(\alpha\beta_\rho) = \bar{\epsilon}_{local} - \ln 2/(\rho n)$. Then by the definition of $\bar{\epsilon}_{local}$, we have $\sup_{\epsilon\in[u_1,u_2]} \frac{\pi(\epsilon/(\alpha\beta_\rho))}{\pi(\epsilon)} + \exp(\frac{\rho n u_1}{\alpha\beta_\rho}) + \frac{\exp(-\rho n u_2/2)}{\pi(u_2/2)} \leq 2\exp(\frac{\rho n u_1}{\alpha\beta_\rho}) = \exp(\rho n \bar{\epsilon}_{local})$. This gives the second inequality.

Remark 2. By letting $u \to \infty$ in the definition of $\bar{\epsilon}_{local}$, we can see easily that $\bar{\epsilon}_{local} \leq \ln 2/(\rho n) + \bar{\epsilon}_{global}/(\alpha\beta_\rho)$. Therefore using the localized complexity $\bar{\epsilon}_{local}$ is always better (up to a constant) than using $\bar{\epsilon}_{global}$. If the ratio $\pi(\epsilon/(\alpha\beta_\rho))/\pi(\epsilon)$ is much smaller than $\pi(\epsilon)$, the localized complexity can be much better than the global complexity.

In the following, we consider three cases of local prior structures, and derive the corresponding rates of convergence. Comparable lower-bounds are given in Section 4.

2.1 Non-parametric type local prior

It is well known that for standard nonparametric families such as smoothing splines, etc, the ϵ-entropy often grows at the order of $O(\epsilon^{-r})$ for some $r > 0$. We shall not list detailed examples here, and simply refer the readers to [5–7] and references there-in. Similarly, we assume that there exists constants C and r such that the prior $\pi(\epsilon)$ satisfies the condition:

$$C_1\epsilon^{-r} \leq \ln\frac{1}{\pi(\epsilon)} \leq C_2\epsilon^{-r}.$$

This implies that $\bar{\epsilon}_{global} \leq (C_2/(\rho n))^{1/(1+r)}$. It is easy to check that $\bar{\epsilon}_{local}$ is the same order of $\bar{\epsilon}_{global}$ when $C_1 > 0$. Therefore, for prior that behaves non-parametrically around the truth, it does not matter whether we use global complexity or local complexity.

2.2 Parametric type local prior

For standard parametric families, the prior π has a density with an underlying dimensionality d: $\pi(\epsilon) = O(\epsilon^{-d})$. We may assume that the following condition holds:

$$C_1 + d\ln\frac{1}{\epsilon} \leq \ln\frac{1}{\pi(\epsilon)} \leq C_2 + d\ln\frac{1}{\epsilon}.$$

This implies that $\bar{\epsilon}_{global}$ is of the order $d\ln n/n$. However, we have

$$\bar{\epsilon}_{local} \leq \frac{\ln 2 + C_2 - C_1 - d\ln(\alpha\beta_\rho)}{\rho n},$$

which is of the order $O(d/n)$ for large d. In this case, we obtain a better rate of convergence using localized complexity measure.

2.3 Singular local prior

It is possible to obtain a rate of convergence faster than $O(1/n)$. This cannot be obtained with either $\bar{\epsilon}_{global}$ or $\bar{\epsilon}_{local}$, which are of the order no better than n^{-1}. The phenomenon of faster than $O(1/n)$ convergence rate is related to super-efficiency and hence can only appear at countably many isolated points.

To see that it is possible to obtain faster than $1/n$ convergence rate (super efficiency) in our framework, we only consider the simple case where

$$\sup_{\epsilon \leq 2u} \frac{\pi(\epsilon/(\alpha\beta_\rho))}{\pi(\epsilon)} = 1.$$

That is, we have a point-like prior mass at the truth with zero density around it (up to a distance of $2u$). In this case, we can apply Corollary 2 with $u_1 = -\infty$ and $u_2 = 2u$, and obtain

$$\mathbf{E}_{\hat{Z}} \mathbf{E}_{\theta \sim \hat{\pi}_\rho} R(\theta) \leq \frac{\ln\left[1 + \frac{\exp(-\rho n u)}{\pi(u)}\right]}{(1-\alpha)\beta_\rho \rho n}.$$

This gives an exponential rate of convergence. Clearly this example can be generalized to the case that a point is not completely isolated from its neighbor.

3 Some Examples

We focus on consequences and applications of Theorem 1. Specifically, we give two important examples for which (4) holds with some positive constants α, ρ, and β_ρ.

3.1 Conditional density estimation

Conditional density estimation is very useful in practical applications. It includes the standard density estimation problem widely studied in statistics as a special case. Moreover, many classification algorithms (such as decision trees or logistic regression) can be considered as conditional density estimators.

Let $Z_1 = (X_1, Y_1)$, where X_1 is the input variable, and Y_1 is the output variable. We are interested in estimating the conditional density $p(Y_1|X_1)$. In this framework, we assume (with a slight abuse of notation) that each parameter θ corresponds to a conditional density function: $\theta(Z_1) = p(Y_1|\theta, X_1)$. In density estimation, we consider negative log loss function $-\ln\theta(Z_1)$. Our goal is to find a randomized conditional density estimator θ from the data, such that the expected log-loss $-\mathbf{E}_{Z_1} \ln\theta(Z_1)$ is as small as possible.

In this case, the Gibbs estimator in (2) becomes

$$d\hat{\pi}_\rho \propto \prod_{i=1}^{n} \theta(\hat{Z}_i)^\rho d\pi, \tag{5}$$

which corresponds to the Bayesian posterior distribution when $\rho = 1$. Lemma 2 can be directly applied since the left hand side can be interpreted as a (Hellinger-like) distance between distributions. This approach has been taken in [8]. However, in this section, we are interested in using the log-loss on the left-hand side.

We further assume that θ is defined on a domain $\mathcal{G}$ which is a closed convex density class. However, we do not assume that $\mathcal{G}$ contains the true conditional density. We also let $\theta_{\mathcal{G}}$ be the optimal density in $\mathcal{G}$ with respect to the log loss:

$$\mathbf{E}_{Z_1} \ln \frac{1}{\theta_{\mathcal{G}}(Z_1)} = \inf_{\theta \in \mathcal{G}} \mathbf{E}_{Z_1} \ln \frac{1}{\theta(Z_1)}.$$

In the following, we are interested in a bound which compare the performance of the randomized estimator (5) to the best possible predictor $\theta_{\mathcal{G}} \in \mathcal{G}$, and thus define

$$\ell_\theta(Z_1) = \ln \frac{\theta_{\mathcal{G}}(Z_1)}{\theta(Z_1)}.$$

In order to apply Theorem 1, we need the following variance bound. We skip the proof due to the space limitation.

Proposition 1. *If there exists a constant $M_{\mathcal{G}} \geq 0$ such that $-M_{\mathcal{G}} \mathbf{E}_{Z_1} \ell_\theta(Z_1)^2 \leq \mathbf{E}_{Z_1} \ell_\theta(Z_1)^3$. Then $\mathbf{E}_{Z_1} \ell_\theta(Z_1)^2 \leq \frac{8M_{\mathcal{G}}}{3} \mathbf{E}_{Z_1} \ell_\theta(Z_1)$.*

Using this result, we obtain the following theorem from Theorem 1.

Theorem 3. *Consider the estimator (5) for conditional density estimation (under log-loss). Then $\forall \alpha \in [0, 1)$, inequality (4) holds with $R(\theta) = \mathbf{E}_{Z_1} \ln \frac{\theta_{\mathcal{G}}(Z_1)}{\theta(Z_1)}$ under either of the following two conditions:*

- $\sup_{\theta_1, \theta_2 \in \mathcal{G}, Z_1} \ln \frac{\theta_1(Z_1)}{\theta_2(Z_1)} \leq M_{\mathcal{G}}$*: we pick ρ such that $\beta_\rho = (11\rho M_{\mathcal{G}} + 8 - 8e^{\rho M_{\mathcal{G}}})/(3\rho M_{\mathcal{G}}) > 0$.*
- $\forall \theta \in \mathcal{G}$ *and* $m \geq 3$, $\mathbf{E}_{Z_1}(\ln \frac{\theta(Z_1)}{\theta_{\mathcal{G}}(Z_1)})^m \leq m! M_{\mathcal{G}}^{m-2} b \mathbf{E}_{Z_1}(\ln \frac{\theta(Z_1)}{\theta_{\mathcal{G}}(Z_1)})^2$*: we pick ρ such that $\beta_\rho = 1 - 8b\rho M_{\mathcal{G}}(1 - \rho M_{\mathcal{G}} + 2b\rho M_{\mathcal{G}})/(1 - \rho M_{\mathcal{G}}) > 0$.*

Proof. Under the first condition, using Proposition 1, we may take $K = 8/3M_{\mathcal{G}}$ and $M = M_{\mathcal{G}}$ in Theorem 1 (bounded loss case). Under the second condition, using Proposition 1, we may take $K = 16M_{\mathcal{G}} b$ and $M = M_{\mathcal{G}}$ in Theorem 1 (Bernstein loss case).

Similar to the remark after Theorem 1, we may also let $\beta_\rho = (3 - 4\rho M_{\mathcal{G}} e^{\rho M_{\mathcal{G}}})/3$ under the first condition of Theorem 3. The second condition involves moment inequalities that needs to be verified for specific problems. It applies to certain unbounded conditional density families such as conditional Gaussian models with bounded variance. We shall discuss a related scenario in the least squares regression case. Note that under Gaussian noise with identical variance, the conditional density estimation using the log-loss is equivalent to the estimation of conditional mean using least squares regression.

Since for log-loss, (4) holds under appropriate boundedness or moment assumptions on the density family, consequences in Section 2 applies. As we shall show in Section 4, similar lower bounds can be derived.

3.2 Least squares regression

Let $Z_1 = (X_1, Y_1)$, where X_1 is the input variable, and Y_1 is the output variable. We are interested in predicting Y_1 based on X_1. We assume that each parameter θ corresponds to a predictor: $\theta(X_1)$. The quality of the predictor is measured by the mean squared error $\mathbf{E}_{Z_1}(\theta(X_1)-Y_1)^2$. In this framework, the Gibbs estimator in (2) becomes

$$\hat{\pi}_\rho \propto \exp\left[-\rho \sum_{i=1}^{n} (\theta(X_i) - Y_i)^2\right]. \tag{6}$$

We further assume that θ is defined on a domain $\mathcal{G}$, which is a closed convex function class. Let $\theta_{\mathcal{G}}$ be the optimal predictor in $\mathcal{G}$ with respect to the least squares loss:

$$\mathbf{E}_{Z_1}(\theta_{\mathcal{G}}(X_1) - Y_1)^2 = \min_{\theta \in \mathcal{G}} \mathbf{E}_{Z_1}(\theta(X_1) - Y_1)^2.$$

In the following, we are interested in comparing the performance of the randomized estimator (5) to the best possible predictor $\theta_{\mathcal{G}} \in \mathcal{G}$. Define

$$\ell_\theta(Z_1) = (\theta(X_1) - Y_1)^2 - (\theta_{\mathcal{G}}(X_1) - Y_1)^2.$$

We have the following proposition. Again, we skip the proof due to the limitation of space.

Proposition 2. *Let* $A_{\mathcal{G}} = \sup_{X_1, \theta \in \mathcal{G}} |\theta(X_1) - \theta_{\mathcal{G}}(X_1)|$ *and* $\sup_{X_1, \theta \in \mathcal{G}} \mathbf{E}_{Y_1|X_1} |\theta(X_1) - Y_1|^m \le m! B_{\mathcal{G}}^{m-2} M_{\mathcal{G}}$ *for* $m \ge 2$. *Then we have:*

$$\mathbf{E}_{Z_1}(-\ell_\theta(Z_1))^m \le m!(2A_{\mathcal{G}}B_{\mathcal{G}})^{m-2} 4 M_{\mathcal{G}} \mathbf{E}_{Z_1} \ell_\theta(Z_1).$$

The moment estimates can be combined with Theorem 1, and we obtain the following theorem.

Theorem 4. *Consider the estimator (6) for least squares regression. Then* $\forall \alpha \in [0, 1)$, *inequality (4) holds with* $R(\theta) = \mathbf{E}_{Z_1}(\theta(X_1) - Y_1)^2 - \mathbf{E}_{Z_1}(\theta_{\mathcal{G}}(X_1) - Y_1 Z)^2$, *under either of the following conditions:*

- $\sup_{\theta \in \mathcal{G}, Z_1} (\theta(X_1) - Y_1)^2 \le M_{\mathcal{G}}$*: we pick* ρ *such that* $\beta_\rho = (5\rho M_{\mathcal{G}} + 4 - 4e^{\rho M_{\mathcal{G}}})/(\rho M_{\mathcal{G}}) > 0$.
- *Proposition 2 holds for all integer* $m \ge 2$*: we pick small* ρ *such that* $\beta_\rho = 1 - 4M_{\mathcal{G}}\rho/(1 - 2A_{\mathcal{G}}B_{\mathcal{G}}\rho) > 0$.

Proof. Under the first condition, using Proposition 2, we have $M_{\mathcal{G}} \le \sup_{\theta \in \mathcal{G}, Z_1} (\theta(X_1) - Y_1)^2$. We may thus take $K = 4M_{\mathcal{G}}$ and $M = M_{\mathcal{G}}$ in Theorem 1 (bounded loss case). Under the second condition, using Proposition 2, we can let $K = 8M_{\mathcal{G}}$, $M = 2A_{\mathcal{G}}B_{\mathcal{G}}$ and $b = 1/2$ in Theorem 1 (Bernstein loss case).

The theorem applies to unbounded regression problems with exponentially decaying noise such as Gaussian noise. For example, the following result holds.

Corollary 3. *Assume that there exists function $y_0(X)$ such that*

- *For all X_1, the random variable $|Y_1 - y_0(X_1)|$, conditioned on X_1, is dominated by the absolute value of a zero-mean Gaussian random variable*[1] *with standard deviation σ.*
- *$\exists$ constant $b > 0$ such that $\sup_{X_1} |y_0(X) - \theta(X_1)| \leq b$.*

If we also choose A such that $A \geq \sup_{X_1, \theta \in \mathcal{G}} |\theta(X_1) - \theta_{\mathcal{G}}(X_1)|$, then (4) holds with $\beta_\rho = 1 - 4\rho(b+\sigma)^2/(1 - 2A(b+\sigma)\rho) > 0$.

4 Lower Bounds

The purpose of this section is to prove some lower bounds which hold for arbitrary statistical estimators. Our goal is to match these lower bounds to the upper bounds proved earlier (at least for certain problems), which implies that the Gibbs algorithm is near optimal.

Upper bounds we obtained in previous sections are for every possible realization of the underlying distribution. It is not possible to obtain a lower bound for any specific realization since we can always design an estimator that picks a parameter that achieves the best possible performance under this particular distribution. However, such an estimator will not work well for a different distribution. Therefore as far as lower bounds are concerned, we are interested in the performance averaged over a set of underlying distributions.

In order to obtain lower bounds, we associate each parameter θ with a probability distribution $q_\theta(x, y)$ so that we can take samples $Z_i = (X_i, Y_i)$ from this distribution. In addition, we shall design the map in such a way that the optimal parameter under this distribution is θ. For (conditional) density estimation, the map is the density itself. For regression, we associate each predictor θ with a conditional Gaussian distribution with constant variance and the conditional mean given by the prediction $\theta(X_1)$ of each input X_1.

We consider the following scenario: we put a prior π on θ, which becomes a prior on the distribution $q_\theta(x, y)$. Assume that we are interested in estimating θ, under a loss function $\ell_\theta(Z_1)$, then the quantity

$$R_\theta(\theta') = \mathbf{E}_{Z_1 \sim q_\theta} \ell_{\theta'}(Z_1)$$

is the true risk between an estimated parameter θ' and the true distribution parameter θ. The average performance of an arbitrary randomized estimator $\hat{\theta}(Z)$ can thus be expressed as

$$\mathbf{E}_{\theta \sim \pi} \mathbf{E}_{Z \sim q_\theta(Z)} R_\theta(\hat{\theta}(Z)), \tag{7}$$

where Z consists of n independent samples $Z = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from the underlying density. In this section, we are mainly interested in obtaining a

[1] That is, conditioned on X_1, the moments of $|Y_1 - y_0(X_1)|$ with respect to Y_1 are no larger than the corresponding moments of the dominating Gaussian random variable.

lower bound for any possible estimator, so that we can compare this lower bound to the upper bound for the Gibbs algorithm developed earlier.

Note that (7) only gives one performance measure, while the upper bound for the Gibbs method is specific for every possible truth q_θ. It is thus useful to study the best local performance around any possible θ with respect to the underlying prior π. To address this issue, we observe that for every partition of the θ space into the union of disjoint small balls B_k, we may rewrite (7) as

$$\sum_j \pi(B_k)\mathbf{E}_{\theta\sim\pi_{B_k}}\mathbf{E}_{Z\sim q_\theta(Z)}R_\theta(\hat{\theta}(Z)),$$

where for each small ball B_k, the localized prior is defined as:

$$\pi_{B_k}(A) = \frac{\pi(A\cap B_k)}{\pi(B_k)}.$$

Therefore, instead of bounding the optimal Bayes risk with respect to the global prior π in (7), we shall bound the optimal risk with respect to a local prior π_B for a small ball B around any specific parameter θ, which gives a more refined performance measure. In this framework, if for some small local ball π_B, the Gibbs algorithm has performance not much worse than the best possible estimator, then we can say that it is *locally near optimal*.

The main theorem in our lower bound analysis is presented below. Related techniques appeared in [2, 4, 7].

Theorem 5. *Consider an arbitrary randomized estimator $\hat{\theta}(Z)$ that takes value in $B'\subset\mathcal{G}$, where Z consists of n independent samples $Z=\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ from some underlying density q_θ, then for all non-negative functions $R_\theta(\theta')$, we have*

$$\mathbf{E}_{\theta\sim\pi_B}\mathbf{E}_{Z\sim q_\theta(Z)}R_\theta(\hat{\theta}(Z)) \geq 0.5\sup\left\{\epsilon : \inf_{\theta'\in B'}\ln\frac{1}{\pi_B(\{\theta : R_\theta(\theta')<\epsilon\})} \geq 2n\Delta_B+4\right\},$$

where $\Delta_B = \mathbf{E}_{\theta\sim\pi_B}\mathbf{E}_{\theta'\sim\pi_B}D_{KL}(q_\theta(Z_1)||q_{\theta'}(Z_1))$.

Proof. The joint distribution of (θ,Z) is given by $\prod_{i=1}^n q_\theta(Z_i)d\pi_B(\theta)$. Denote by $I(\theta,Z)$ the mutual information between θ and Z. Now let Z' be a random variable independent of θ and with the same marginal of Z, then by definition, the mutual information can be regarded as the KL-divergence between the joint distributions of (θ,Z) and (θ,Z'), which we write (with a slight abuse of notation) as:

$$I(\theta,Z) = D_{KL}((\theta,Z)||(\theta,Z')).$$

Now consider an arbitrary estimator $\hat{\theta}:\mathcal{Z}\to B'$. By the data processing theorem for KL-divergence (that is, processing does not increase KL-divergence), with

input $(\theta, Z) \in \mathcal{G} \times \mathcal{Z}$ and binary output $\mathbf{1}(R_\theta(\hat{\theta}(Z)) \leq \epsilon)$, we obtain

$$\begin{aligned}
&D_{KL}(\mathbf{1}(R_\theta(\hat{\theta}(Z)) \leq \epsilon)||\mathbf{1}(R_\theta(\hat{\theta}(Z')) \leq \epsilon)) \\
\leq& D_{KL}((\theta, Z)||(\theta, Z')) = I(\theta, Z) \\
=& \mathbf{E}_{\theta_1 \sim \pi_B} \mathbf{E}_{Z \sim q_{\theta_1}(Z)} \ln \frac{q_{\theta_1}(Z)}{\mathbf{E}_{\theta_2 \sim \pi_B} q_{\theta_2}(Z)} \\
\leq& \mathbf{E}_{\theta_1 \sim \pi_B} \mathbf{E}_{Z \sim q_{\theta_1}(Z)} \mathbf{E}_{\theta_2 \sim \pi_B} \ln \frac{q_{\theta_1}(Z)}{q_{\theta_2}(Z)} \\
=& n \mathbf{E}_{\theta_1 \sim \pi_B} \mathbf{E}_{\theta_2 \sim \pi_B} D_{KL}(q_{\theta_1}||q_{\theta_2}) = n\Delta_B.
\end{aligned}$$

The second inequality is a consequence of Jensen's inequality and the concavity of logarithm.

Now let $p_1 = P(R_\theta(\hat{\theta}(Z)) \leq \epsilon)$ and $p_2 = P(R_\theta(\hat{\theta}(Z')) \leq \epsilon)$, then the above inequality can be rewritten as:

$$D_{KL}(p_1||p_2) = p_1 \ln \frac{p_1}{p_2} + (1 - p_1) \ln \frac{1 - p_1}{1 - p_2} \leq n\Delta_B.$$

Since $\hat{\theta}(Z')$ is independent of θ, we have

$$p_2 \leq \sup_{\theta' \in B'} \pi_B(\{\theta : R_\theta(\theta') \leq \epsilon\}).$$

Now we consider any ϵ such that $\sup_{\theta' \in B'} \pi_B(\{\theta : R_\theta(\theta') < \epsilon\}) \leq 0.25 e^{-2n\Delta_B}$. This implies that $p_2 \leq 0.25 e^{-2n\Delta_B} \leq 0.25$.

We now show that in this case, $p_1 \leq 1/2$. Since $D_{KL}(p_1||p_2)$ is increasing in $[p_2, 1]$, we only need to show that $D_{KL}(0.5||p_2) \geq n\Delta_B$. This easily follows from the inequality

$$D_{KL}(0.5||p_2) \geq 0.5 \ln \frac{0.5}{p_2} + 0.5 \ln \frac{0.5}{1} \geq n\Delta_B.$$

Now, we have shown that $p_1 \leq 0.5$, which implies that $P(R_\theta(\hat{\theta}(Z)) \geq \epsilon) \geq 0.5$. Therefore we have $\mathbf{E}_{\theta \sim \pi_B} \mathbf{E}_{Z \sim q_\theta(Z)} R_\theta(\hat{\theta}(Z)) \geq 0.5\epsilon$.

Theorem 5 has a form that resembles Corollary 2. In the following, we state a result which shows the relationship more explicitly.

Corollary 4 (Local Entropy Lower Bound). *Under the notations of Theorem 5. Consider a reference point $\theta_0 \in \mathcal{G}$, and balls $B(\theta_0, \epsilon) \subset \mathcal{G}$ which contains θ_0 and indexed by $\epsilon > 0$, such that*

$$\sup_{\theta_1, \theta_2 \in B(\theta_0, \epsilon)} D_{KL}(q_\theta(Z_1)||q_{\theta'}(Z_1)) \leq \epsilon.$$

Given $u > 0$, consider $\underline{\epsilon}(\theta_0, u)$ which satisfies:

$$\underline{\epsilon}(\theta_0, u) = \sup_{\epsilon > 0} \left\{ \epsilon : \inf_{\theta' \in B'} \ln \frac{\pi(B(\theta_0, u\epsilon))}{\pi(\{\theta : R_\theta(\theta') < \epsilon\} \cap B(\theta_0, u\epsilon))} \geq 2nu\epsilon + 4 \right\},$$

then locally around θ_0, we have

$$\mathbf{E}_{\theta\sim\pi_{B(\theta_0,u\underline{\epsilon}(\theta_0,u))}}\mathbf{E}_{Z\sim q_\theta(Z)}R_\theta(\hat{\theta}(Z)) \geq 0.5\underline{\epsilon}(\theta_0,u).$$

The definition of $B(\theta_0,\epsilon)$ requires that within the $B(\theta_0,\epsilon)$ ball, the distributions q_θ are nearly indistinguishable up to a scale of ϵ, when measured by their KL-divergence. Corollary 4 implies that the local performance of an arbitrary statistical estimator cannot be better than $\underline{\epsilon}(\theta_0,u)/2$. The bound in Corollary 4 will be good if the ball $\pi(B(\theta_0,\epsilon))$ is relatively large. That is, there are many distributions that are statistical nearly indistinguishable (in KL-distance). Therefore the bound of Corollary 4 is similar to Corollary 2, but the localization is within a small ball which is statistically nearly indistinguishable (rather than the $R(\cdot)$ localization for the Gibbs estimator). From an information theoretical point of view, this difference is rather intuitive and clearly also necessary since we allow arbitrary statistical estimators (which can simply estimate the specific underlying distribution q_θ if they are distinguishable).

It follows that if we want the upper bound in Corollary 2 to match the lower bound in Corollary 4, we need to design a map $\theta \to q_\theta$ such that locally around θ_0, a ball with small $R(\cdot)$ risk is also small information theoretically in terms of the KL-distance between q_θ and q_{θ_0}. Consider the following two types of small R balls:

$$B_1(\theta,\epsilon) = \{\theta' : R_\theta(\theta') < \epsilon\}, \quad B_2(\theta,\epsilon) = \{\theta' : R_{\theta'}(\theta) < \epsilon\}.$$

Now assume that we can find a map $\theta \to q_\theta$ such that locally around θ_0, q_θ within a small B_1-ball is also small in the information theoretical sense (small KL-distance). That is, we have for some $c > 0$ that

$$\sup\{D_{KL}(q_\theta(Z_1)||q_{\theta'}(Z_1)) : \theta,\theta' \in B_1(\theta_0,c\epsilon)\} \leq \epsilon. \tag{8}$$

For problems such as density estimation and regression studied in this paper, it is easy to design such a map (under mild conditions such as the boundedness of the loss). We shall not go into the details for verifying specific examples of (8). Now, under this condition, we can take

$$\underline{\epsilon}(\theta_0,u) = \sup_{\epsilon>0}\left\{\epsilon : \inf_{\theta'\in B'} \ln\frac{\pi(B_1(\theta_0,cu\epsilon))}{\pi(B_2(\theta',\epsilon)\cap B_1(\theta_0,cu\epsilon))} \geq 2nu\epsilon + 4\right\}.$$

As a comparison, according to Corollary 2, the Gibbs method at θ_0 gives an upper bound of the following form (which we simplify to focus on the main term) with some constant $u' \in (0,1)$:

$$\bar{\epsilon}_{local} \leq \frac{2}{\rho n} + \inf\left\{\epsilon : \rho n\epsilon \geq \sup_{\epsilon'\geq\epsilon} \ln\frac{\pi(B_1(\theta_0,\epsilon'))}{\pi(B_1(\theta_0,u'\epsilon'))}\right\}.$$

Essentially, the local upper bound for the Gibbs algorithm is achieved at $\bar{\epsilon}_{local}$ such that

$$n\bar{\epsilon}_{local} \sim \sup_{\epsilon'\geq\bar{\epsilon}_{local}} \ln\frac{\pi(B_1(\theta_0,\epsilon'))}{\pi(B_1(\theta_0,u'\epsilon'))},$$

where we use $\sim$ to denote approximately the same order, while the lower bound in Corollary 4 implies that (let $u' = 1/(cu)$):

$$n\underline{\epsilon} \sim \inf_{\theta' \in B'} \ln \frac{\pi(B_1(\theta_0, \underline{\epsilon}))}{\pi(B_2(\theta', u'\underline{\epsilon}) \cap B_1(\theta_0, \underline{\epsilon}))}.$$

From this, we see that our upper and lower bounds are very similar. There are two main differences which we outline below.

- In the lower bound, for technical reasons, B_2 appears in the definition of the local entropy. In order to argue that the difference does not matter, we need to assume that the prior probabilities of B_1 and B_2 are of the same order.
- In the lower bound, we use the smallest local entropy in a neighbor of θ_0, While in the upper bound, we use the largest local entropy at θ_0 across different scales. This difference is not surprising since the lower bound is with respect to the average in a small neighborhood of θ_0.

Both differences are relatively mild and somewhat expected. We consider two situations which parallel Section 2.1 and Section 2.2.

4.1 Non-parametric type local prior

Similar to Section 2.1, we assume that for some sufficiently large constant v: there exist $0 < C_1 < C_2$ such that

$$C_2\epsilon^{-r} \leq \inf_{\theta' \in B'} \ln \frac{1}{\pi(B_2(\theta', \epsilon) \cap B_1(\theta_0, v\epsilon))}, \qquad \ln \frac{1}{\pi(B_1(\theta_0, v\epsilon))} \leq C_1\epsilon^{-r},$$

which measures the order of global entropy around a small neighborhood of θ_0. Now under the condition (8) and let $u = v/c$, Corollary 4 implies that

$$\underline{\epsilon} \geq \sup\left\{\epsilon : 2un\epsilon + 4 \leq (C_2 - C_1)\epsilon^{-r}\right\}.$$

This implies that $\underline{\epsilon}$ is of the order $n^{-1/(1+r)}$, which matches the order of the Gibbs upper bound $\bar{\epsilon}_{global}$ in Section 2.1.

4.2 Parametric type local prior

Similar to Section 2.2, we assume that for some sufficiently large constant v: there exist $0 < C_1 < C_2$ such that

$$C_2 + d\ln\frac{1}{\epsilon} \leq \inf_{\theta' \in B'} \ln \frac{1}{\pi(B_2(\theta', \epsilon) \cap B_1(\theta_0, v\epsilon))}, \qquad \ln \frac{1}{\pi(B_1(\theta_0, v\epsilon))} \leq C_1 + d\ln\frac{1}{\epsilon}.$$

which measures the order of global entropy around a small neighborhood of θ_0. Now under the condition (8) and let $u = v/c$, Corollary 4 implies that

$$\underline{\epsilon} \geq \sup\left\{\epsilon : 2c^{-1}n\epsilon \leq C_2 - C_1 - 4\right\}.$$

That is, we have a convergence rate of the order $1/n$, which matches the parametric upper bound $\bar{\epsilon}_{local}$ for the Gibbs algorithm in Section 2.2.

5 Discussions

In this paper, we established upper and lower bounds for some statistical estimation problems. Our upper bound analysis is based on a simple information theoretical inequality, which can be used to analyze randomized estimation methods such as Gibbs algorithms. The resulting upper bounds rely on the local prior decaying rate in some small ball around the truth. Moreover, we are able to obtain lower bounds that have similar forms as the upper bounds. For some problems (such as density estimation and regression), the upper and lower bounds match under mild conditions. This suggests that both of our upper bound and lower bound analysis are relatively tight.

This work can be regarded as an extension of the standard minimax framework since we allow the performance of the estimator to vary for different underlying distributions, according to the pre-defined prior. The framework we study here is closely related to the concept of adaption in the statistical literature. At the conceptual level, both seek to find locally near optimal estimators around any possible true underlying distribution within the class.

This paper also shows that in theory, the Gibbs algorithm is better behaved than (possibly penalized) empirical risk minimization that picks an estimator to minimize the (penalized) empirical risk. In particular, for certain problems such as density estimation and regression, the Gibbs algorithm can achieve the best possible convergence rate under relatively mild assumptions on the prior structure. However, it is known that for non-parametric problems, empirical risk minimization can lead to sub-optimal convergence rate if the covering number grows too rapidly (or in our case, prior decays too rapidly) when ϵ (the size of the covering ball) decreases [1].

References

1. Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, 97(1-2):113–150, 1993.
2. R.E. Blahut. Information bounds of the Fano-Kullback type. *IEEE Transactions on Information Theory*, 22:410–421, 1976.
3. Olivier Catoni. A PAC-Bayesian approach to adaptive classification. Available online at http://www.proba.jussieu.fr/users/catoni/homepage/classif.pdf.
4. Te Sun Han and Sergio Verdú. Generalizing the Fano inequality. *IEEE Transactions on Information Theory*, 40:1247–1251, 1994.
5. S.A. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.
6. Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.
7. Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27:1564–1599, 1999.
8. Tong Zhang. Learning bounds for a generalized family of Bayesian posterior distributions. In *NIPS 03*, 2004.
9. Tong Zhang. On the convergence of MDL density estimation. In *COLT 2004*, pages 315–330, 2004.