# Convergence Rates for Posterior Distributions and Adaptive Estimation[1]

Tzee-Ming Huang

*Carnegie Mellon University*

**Abstract**

The main goal of this paper is to provide general theorems on convergence rates of posterior distributions that can be applied to many examples to obtain good convergence rates in the context of density estimation as well as regression. We show how to choose priors so that the posterior distributions converge at the optimal rate without prior knowledge of the amount of smoothness of the density function or the regression function to be estimated.

## 1 Intoduction

Bayesian methods have been used for non-parametric inference problems, and many theoretical results have been developed to investigate the asymptotic properties of non-parametric Bayesain methods. So far, the positive results are on consistency and convergence rates. For example, Doob (1948) proved the consistency of posterior distributions with respect to the joint distribution of the data and the prior under some weak conditions. For the frequentist version of consistency, see Diaconis and Freedman (1986) for a review on consistency results on tailfree and Dirichlet priors. Also, Barron,

[1]This paper is based on the author's thesis work under the supervision of Professor Larry Wasserman at Carnegie Mellon University.

AMS 1991 subject classifications. Primary 62A15; secondary 62G20, 62G07.
Key words and phrases. Convergence rate, non-parametric regression, density estimation, Bayesian, adaptive estimation, sieves.

Schervish and Wasserman (1999) gave some conditions to achieve the frequentist version of consistency in general. For convergence rates, although there have been some general results by Ghosal, Ghosh and van der Vaart (1998) and Shen and Wasserman (1998), adaptive estimation has not been considered. The goal of this paper is to develop theorems on convergence rates for posterior distributions which can be used for adaptive estimation. Also, we develop rates of convergence for non-parametric regression, which has not been done before.

In this paper, we have theorems on convergence rates in two contexts: density estimation and regression. In either case, we want to estimate a function $f$ using Bayesian methods based on the data $Z_1, ...., Z_n$ which are iid obsevations with density $p_f$, and we are interested in convergence rates for the posterior distributions for $f$. In other words, suppose that $f$ is some fixed function $f_0$, $\tilde{\pi}$ is a prior on some function space $\mathcal{F}$, and $\tilde{A}_n = \{f : d(f, f_0) \leq s_n\}$ is an $s_n$ neighborhood of $f_0$ with respect to metric $d$, where $d$ is the Hellinger distance in the density estimation case and is the $L^2$ distance in the regression case. We would like to show that the posterior probability

$$\pi(\tilde{A}_n^c | Z_1, ..., Z_n) = \frac{\int_{\tilde{A}_n^c} \prod_{i=1}^n p_f(Z_i) d\tilde{\pi}(f)}{\int_{\mathcal{F}} \prod_{i=1}^n p_f(Z_i) d\tilde{\pi}(f)} \tag{1}$$

converges to zero in $P_{f_0}^n$ probability and that this rate $s_n$ holds simultaneously for $f_0$ in function classes of different regularity. For the purpose of adaptive estimation, we consider a sieve prior, meaning a prior $\tilde{\pi}$ with the following form:

$$\tilde{\pi} = \sum_j a_j \tilde{\pi}_j,$$

where $a_j \geq 0$, $\sum_j a_j = 1$, and each $\tilde{\pi}_j$ is a prior supported on a space $\mathcal{F}_j \subset \mathcal{F} = \cup_j \mathcal{F}_j$. Also, we assume that each $\mathcal{F}_j$ is finite dimensional and can be represented as $\{f_\theta : \theta \in \Theta_j\}$ for some parameter space $\Theta_j$. Using this sieve prior, the posterior probability in (1) can be written as $U_n/V_n$, where

$$U_n = \sum_j a_j \int_{A_n^c} \prod_{i=1}^n \frac{p_{f_\theta}(Z_i)}{p_{f_0}(Z_i)} d\pi_j(\theta)$$

and

$$V_n = \sum_j a_j \int_{\Theta_j} \prod_{i=1}^n \frac{p_{f_\theta}(Z_i)}{p_{f_0}(Z_i)} d\pi_j(\theta),$$

where $A_n = \{\theta \in \cup_j \Theta_j : d(f_\theta, f_0) \leq s_n\}$ and $\pi_j$ and $\pi = \sum_j a_j \pi_j$ are the induced prior of $\tilde{\pi}_j$ and $\tilde{\pi}$ on $\Theta_j$ and $\cup_j \Theta_j$ respectively.

This paper is organized as follows. Section 2 gives a theorem on convergence rates in the density estimation case and some examples of applying the theorem. Section 3 contains the same things as in Section 2, but in the context of regression. Proofs are in Section 4.

# 2 Density Estimation

## 2.1 A Theorem

This section gives a general convergence rate theorem for Bayesian density estimation. In this case, we have iid observations $X_1,...,X_n$ from a distribution $P_0$ with density $p_0$, so the function $f_0$ in Section 1 is the density $p_0$, $(Z_1,...,Z_n)$ is $(X_1,...,X_n)$, and the metric $d$ is the Hellinger metric $d_H$, which is defined by $d_H^2(f,g) = \int(\sqrt{f} - \sqrt{g})^2$.

To bound $U_n$ and $V_n$, we need some conditions. To bound $U_n$, we assume a similar condition to Assumption 1 in Yang and Barron (1998) in order to apply their results. Let

$$B_{H,j}(\eta, r) = \{\theta \in \Theta_j : d_H^2(p_\eta, p_\theta) \le r^2\}$$

and $N(B, \delta, d')$ denote the $\delta$-covering number of a set $B$ with respect to a metric $d'$, which is defined as the smallest number of $\delta$-balls (with respect to $d'$) that are needed to cover the set $B$. Here is our assumption:

**Assumption 1** *For each $j$, there exist constants $A_j$ and $m_j$ such that $0 < A_j \le 0.0056$, $m_j \ge 1$ and for any $r > 0$, $\delta \le 0.0056r$, $\theta \in \Theta_j$,*

$$N(B_{H,j}(\theta, r), \delta, d_{j,\infty}) \le \left(\frac{A_j r}{\delta}\right)^{m_j},$$

*where $d_{j,\infty}(\theta, \eta) = \|\log p_\theta - \log p_\eta\|_\infty$ for all $\theta$, $\eta \in \Theta_j$.*

Also, suppose Assumption 1 holds, we specify the weights $a_j$ in the following way to give an upper bound for $U_n$:

$$a_j = \alpha \exp\left(-\left(1 + \frac{1-4\gamma}{8}\right)\eta_j\right), \tag{2}$$

where $\alpha$ is a normalizing constant so that $\sum_j a_j = 1$, $\gamma$ be the solution to $0.13\gamma/\sqrt{1-4\gamma} = 0.0056$, and

$$\eta_j = \frac{4m_j}{1-4\gamma}\log\left(\frac{46.2A_j\sqrt{1-4\gamma}}{\gamma}\right) + \frac{8C_j}{1-4\gamma} \tag{3}$$

for some $C_j$ such that $C_j \geq 0$ and $\sum_j e^{-C_j} \leq 1$.

To bound $V_n$, we will use Lemma 1 by Shen and Wasserman (1998), which gives a lower bound for $V_n$ by controlling the prior probability of a neighborhood of the true density $p_0$, as stated in Lemma 6 in this paper. To controll this prior probability, the idea is to consider a model $\Theta_{j_n}$ that is not too large, but large enough to contain some point $\beta_n$ so that $p_{\beta_n}$ is close to the true densisty $p_0$, and a prior $\pi_{j_n}$ on $\Theta_{j_n}$ that puts enough probability on a neighborhood of $\beta_n$. Then because $p_{\beta_n}$ is close to $p_0$, the prior probability near $p_0$ can't be too small.

Now we will put the above idea into assumptions. First, to measure both the approximation accuracy and the complexity for a model $\Theta_j$, for each $\theta \in \Theta_j$, define

$$R_{j,n}(\theta, p_0) = \max(D(p_0\|p_\theta), V(p_0\|p_\theta)) + \frac{\eta_j}{n},$$

where $D(f\|g) = \int f \log(f/g)$, $V(f\|g) = \int f(\log(f/g))^2$, and $\eta_j$ is as defined in (3) with $A_j$ and $m_j$ in Assumption 1, and let $\epsilon_n$ be such that

$$R_{j_n,n}(\beta_n, p_0) \leq \epsilon_n^2 \tag{4}$$

for some $j_n$ and some $\beta_n \in \Theta_{j_n}$. Second, to make our neighborhoods comparable to the neighborhoods in Shen and Wasserman's Lemma 1, we consider a metric $d_{j_n}$ such that the following assumption holds:

**Assumption 2** *$d_{j_n}$ is a metric on $\Theta_{j_n}$ satisfying*

$$\int p_0 \left(\log \frac{p_\eta}{p_\theta}\right)^2 \leq K_0' d_{j_n}^2(\eta, \theta) \tag{5}$$

*for all $\eta$, $\theta$ in $\Theta_{j_n}$, and*

$$D(p_0\|p_\theta) \leq K_0'' V(p_0\|p_\theta)$$

*for all $\theta \in \Theta_{j_n}$, where $K_0'$ and $K_0''$ are constants.*

Now with the $j_n$, $\beta_n$, $\epsilon_n$ and $d_{j_n}$ defined above, and the $A_j$, $m_j$ in Assumption 1, the condition that the size of model $\Theta_{j_n}$ can't be too large and the condition that the prior $\pi_{j_n}$ puts enough probability near $\beta_n$ can be expressed in the following two assumptions:

**Assumption 3** *There exists* $b_1 \geq 0$ *such that*

$$N(\Theta_{j_n}, \epsilon_n, d_{j_n}) \leq (A_{j_n}^{b_1} K_4)^{m_{j_n}},$$

*where* $N(\Theta_{j_n}, \epsilon_n, d_{j_n})$ *is the* $\epsilon_n$*-covering number of* $\Theta_{j_n}$ *with respect metric* $d_{j_n}$.

**Assumption 4** *There exist constants* $K_5$ *and* $b_2 \geq 0$ *such that for any* $\theta_1 \in \Theta_{j_n}$,

$$\frac{\pi_{j_n}(B_{d_{j_n}, j_n}(\theta_1, \epsilon_n))}{\pi_{j_n}(B_{d_{j_n}, j_n}(\beta_n, \epsilon_n))} \leq (A_{j_n}^{b_2} K_5)^{m_{j_n}}.$$

Finally, we assume

**Assumption 5**

$$n\epsilon_n^2 \to 0 \text{ as } n \to \infty.$$

Now we have the following theorem:

**Theorem 1** *Suppose that Assumption 1 holds and Assumptions 2 - 5 hold for some* $j_n$, $\beta_n$ *and* $\epsilon_n$ *satisfying (4) with* $\eta_j$ *defined in (3) and* $A_j$, $m_j$ *in Assumption 1 and for some metric* $d_{j_n}$, *then with* $a_j$ *defined in (2), we can choose* $s_n^2 = K_1 \epsilon_n^2$ *so that*

$$\pi(A_n^c | X_1, ..., X_n) \leq c \exp\left(-K_2 n \epsilon_n^2\right) \tag{6}$$

*except on a set of probability converging to zero, where the constants* $c$, $K_1$ *and* $K_2$ *do not depend on* $n$.

The proof of Theorem 1 is given in section 4.

Comparing to Theorem 2.1 by Ghosal et al (1998), Theorem 1 requires similar conditions. Assumption 3 in Theorem 1 is like their condition (2.1), and Assumption 4 plays the same role as their condition (2.3). However, Theorem 1 requires one more condition, which is Assumption 1, which gives us an exponential inequality in Lemma 4 that has two components set apart: one component $d_H(p_0, p_\theta)$ represents the approximation error of the model $\Theta_j$, and the other component $\xi_j/n$ represents the model complexity. Such an exponential inequality helps us specify a prior that adapts to the smoothness of the density function, as we will see later in examples.

## 2.2 An Example: Spline Basis

In this section, we consider $\log p_0$ in the Sobolev space $W^s_\infty[0,1] = \{g : D^s g \in L_\infty[0,1]\}$ and approximate $\log p_0$ using a spline basis. We will see that using the sieve prior given below, the posterior distribution converges at the optimal rate $n^{-s/(1+2s)}$ in Hellinger distance:

**Lemma 1** *Suppose that* $\log p_0 \in W^s_\infty[0,1]$. *For integers* $k \geq 0$, $q \geq 1$, *define* $\ell = (k, q)$, $m_\ell = k + q$ *and for* $\theta \in R^{m_\ell}$, *define* $p_\theta$ *as in (21), where* $\{B_{\ell,i}\}$ *is a normalized B-spline basis of degree* $q$ *and* $k$ *equally spaced simple knots. For an integer* $L \geq 1$, *let* $j = (\ell, L)$ *and define*

$$\Theta_j = \{\theta \in R^{m_\ell} : \theta' 1_{m_\ell} = 0, \|D^r \log p_\theta\|_{L_\infty} \leq L, \forall r \in \{0, 1, ..., q-1\}\},$$

*where* $1_{m_\ell} = (1, ..., 1)' \in R^{m_\ell}$, *and*

$$\pi_j(A) = \frac{\int_{\{\theta' B \in A \cap \Theta_j\}} d\mu(\theta)}{\int_{\{\theta' B \in \Theta_j\}} d\mu(\theta)},$$

*where* $\mu$ *is Lebesgue measure on* $\Theta_j$. *Define* $\eta_j$ *according to (3) with*

$$A_j = 19.28\sqrt{q}(2q+1)9^{q-1}(L+1)e^{L/2} + 0.06, \quad m_j = m_\ell \text{ and } C_j = m_j + L, \tag{7}$$

*and define* $a_j$ *according to (2). Then for the prior* $\pi = \sum_j a_j \pi_j$, *the posterior probability* $\pi(\{\theta : d_H(p_\theta, p_0) > s_n\} | X_1, ..., X_n)$ *converges to zero in probability with* $s_n \propto n^{-s/(1+2s)}$.

Proof: The proof of Lemma 1 is given in Section 4.

This convergence rate of $n^{-s/(1+2s)}$ is optimal in the minimax sense. Note that the prior doesn't depend on $s$ here, so it adapts to the smoothness $s$.

## 2.3 Another Example: Haar Basis

In this section, we assume that $\log p_0$ is a continuous function on $[0,1]$ with $\|\log p_0\|_\infty \leq M_0$, and we approximate $\log p_0$ using the Haar basis $\{1_{[0,1]}(x), \psi_{j_1,k_1}(x) : 0 \leq j_1, 0 \leq k_1 \leq 2^{j_1}-1\}$, where $\psi_{j_1,k_1}(x) = 2^{j_1/2}\psi^*(2^{j_1}x - k_1)$ and $\psi^*(x) = 1_{[0,0.5]}(x) - 1_{[0.5,1]}(x)$. We also assume that the coefficients of the $L^2$ expansion of $\log p_0$ for the Haar basis, denoted by $d_{j_1,k_1}$, satisfy the following condition:

$$\sum_{j_1 \geq 0} (2^{j_1+1} - 1)^{2\alpha} \sum_{k_1=0}^{2^{j_1}-1} d^2_{j_1,k_1} \leq H_0^2 \tag{8}$$

for some $H_0 > 0$. According to Barron, Birge and Massart (1999) (p.330), the above condition on the Haar basis coefficients corresponds the Besov space $B^\alpha_{2,2}[0,1]$. The Besov space $B^\alpha_{2,2}[0,1]$ is indeed the Sobolev space $W^\alpha_2[0,1]$, so the optimal convergence rate is $n^{-2\alpha/(1+2\alpha)}$ using the loss function $\ell(\hat{p}, p_0) = \|\hat{p} - p_0\|^2$. We will see that using the sieve prior given below, the posterior distribution converges at the rate $n^{-\alpha/(1+2\alpha)}(\log n)^{1/2}$ in Hellinger distance, which is close to the optimal rate $n^{-\alpha/(1+2\alpha)}$ within a $(\log n)^{1/2}$ factor:

**Lemma 2** *Suppose that* $\log p_0$ *is in the space specified above. For an integer* $\ell \geq 0$*, define* $m_\ell = 2^{\ell+1} - 1$*. Re-index the Haar basis in the following way:*

$$\{\psi_{j_1,k_1} : 0 \leq j_1 \leq \ell, 0 \leq k_1 \leq 2^{j_1} - 1\} \stackrel{def}{=} \{B_{\ell,i} : 1 \leq i \leq m_\ell\}$$

*and then for* $\theta \in R^{m_\ell}$*, define* $p_\theta$ *as in (21). For an integer* $L \geq 1$*, let* $j = (\ell, L)$ *and define*

$$\Theta_j = \{\theta \in R^{m_\ell} : \|\theta' B\|_\infty \leq L\},$$

*and define*

$$\pi_j(A) = \frac{\int_{\{\theta' B \in A \cap \Theta_j\}} d\mu(\theta)}{\int_{\{\theta' B \in \Theta_j\}} d\mu(\theta)},$$

*where* $\mu$ *is Lebesgue measure on* $\Theta_j$*. Define* $\eta_j$ *according to (3) with*

$$A_j = 19.282^{(\ell+1)/2}(2L+1)e^L + 0.06, \;\; m_j = m_\ell + 1 \text{ and } C_j = m_j + L, \quad (9)$$

*and define* $a_j$ *according to (2). Then for the prior* $\pi = \sum_j a_j \pi_j$*, the posterior probability* $\pi(\{\theta : d_H(p_\theta, p_0) > s_n\}|X_1, ..., X_n)$ *converges to zero in probability with* $s_n \propto n^{-\alpha/(1+2\alpha)}(\log n)^{1/2}$*.*

Proof: The proof of Lemma 2 is given in Section 4.

# 3 Regression

## 3.1 A Theorem

In this section, a Bayesian convergence rate theorem is given in the context of regression: $\{(X_i, Y_i, \epsilon_i)\}_{i=1}^n$ are iid, $Y_i = f_0(X_i) + \epsilon_i$, $X_i$ and $\epsilon_i$ are independent, $X_i$ are distributed according to some probability measure $\mu$ and $\epsilon_i$ are normally distributed with mean zero and known variance $\sigma^2$, so the function $f_0$ in Section 1 is the regression function $f_0$ here, the $Z_i$s are $(X_i, Y_i)$s, and the metric $d$ is the $L^2(\mu)$ metric. Here we also assume that $\|f_0\|_\infty$ is bounded by a known constant $M$.

To bound $U_n$ and $V_n$, we modify the assumptions in Theorem 1 in the following way. Let

$$B_{\|\cdot\|,j}(\eta, r) = \{\theta \in \Theta_j : \|f_\eta - f_\theta\| \leq r^2\},$$

where $\|\|$ is the $L^2$ metric and Assumption 1 is replaced with:

**Assumption 6** *For each $j$, there exist constants $A_j$ and $m_j$ such that $0 < A_j \leq 0.0056$, $m_j \geq 1$ and for any $r > 0$, $\delta \leq 0.0056r$, $\theta \in \Theta_j$,*

$$N(B_{\|\cdot\|,j}(\theta, r), \delta, d_{j,\infty}) \leq \left(\frac{A_j r}{\delta}\right)^{m_j},$$

*where $d_{j,\infty}(\theta, \eta) = \|f_\theta - f_\eta\|_\infty$ for all $\theta$, $\eta \in \Theta_j$.*

Also, suppose Assumption 6 holds, we specify the weights $a_j$ in the following way to give an upper bound for $U_n$:

$$a_j = \alpha \exp\left(-\left(1 + \frac{1}{2\sigma^2} + \frac{0.0056}{\sigma}\right)\eta_j\right), \tag{10}$$

where $\alpha$ is a normalizing constant so that $\sum_j a_j = 1$ and

$$\eta_j = \frac{4m_j}{c_{1,M,\sigma}(1-4\gamma)} \log(1072.5A_j) + C_j \max\left(1, \frac{8}{c_{1,M,\sigma}(1-4\gamma)}\right) \tag{11}$$

for some $C_j$ such that $C_j \geq 0$ and $\sum_j e^{-C_j} \leq 1$. Assumption 2 is replaced with

**Assumption 7**

$$\|f_\theta - f_\eta\|^2 \leq K_0' d_{j_n}^2(\theta, \eta) \textit{ for all } \theta, \eta \in \Theta_{j_n} \tag{12}$$

Assumptions 3 - 5 remain unchanged.

Now we have the following theorem:

**Theorem 2** *Suppose that Assumption 6 holds and Assumption 7 and Assumptions 3 - 5 hold for some $j_n$, $\beta_n$ and $\epsilon_n$ satisfying (4) with $p_0 = p_{f_0}$, $p_\theta = p_{f_\theta}$, $\eta_j$ defined in (11) and $A_j$, $m_j$ in Assumption 6 and for some metric $d_{j_n}$, then with $a_j$ defined in (10), we can choose $s_n^2 = K_1\epsilon_n^2$ so that $\pi(A_n^c|X_1, ..., X_n)$ converges to zero in probability.*

Proof: The proof of Theorem 2 is given in section 4.

### 3.2 An Example

In this section, we consider $f_0 \in W_\infty^s[0,1] = \{g : D^s g \in L_\infty[0,1]\}$ and approximate $f_0$ using a spline basis. The minimax rate for this space, according to Stone (1982), is $n^{-2s/(1+2s)}$ using the loss function $\ell(\hat{f}, f_0) = \|\hat{f} - f_0\|^2$. We will see that using the sieve prior given below, the posterior distribution converges at the optimal rate $n^{-s/(1+2s)}$ in $L^2$ distance:

**Lemma 3** *Suppose that $f_0 \in W_\infty^s[0,1]$. For integers $k \geq 0$, $q \geq 1$, define $\ell = (k, q)$, $m_\ell = k + q$ and for $\theta \in R^{m_\ell}$, define $f_\theta$ as in (53), where $\{B_{\ell,i}\}$ is a normalized B-spline basis of degree $q$ and $k$ equally spaced simple knots. For an integer $L \geq 1$, let $j = (\ell, L)$ and define*

$$\Theta_j = \{\theta \in R^{m_\ell} : \|D^r f_\theta\|_{L_\infty} \leq L, \forall r \in \{0, 1, ..., q-1\} \text{ and } \|f_\theta\|_\infty \leq M\}$$

*and*

$$\pi_j(A) = \frac{\int_{\{\theta' B \in A \cap \Theta_j\}} d\mu(\theta)}{\int_{\{\theta' B \in \Theta_j\}} d\mu(\theta)},$$

*where $\mu$ is Lebesgue measure on $\Theta_j$. Define $\eta_j$ according to (11) with*

$$A_j = 9.64\sqrt{q}(2q+1)9^{q-1} + 0.06, \ \ m_j = m_\ell \text{ and } C_j = m_j + L, \tag{13}$$

*and define $a_j$ according to (10). Then for the prior $\pi = \sum_j a_j \pi_j$, the posterior probability $\pi(\{\theta : \|f_\theta - f_0\| > s_n\} | X_1, ..., X_n)$ converges to zero in probability with $s_n \propto n^{-s/(1+2s)}$.*

Proof: The proof for Lemma 3 is given in Section 4.

# 4 Proofs

### 4.1 A Proof for Theorem 1

We prove Theorem 1 by giving bounds for $U_n$ and $V_n$ respectively, and then combining the bounds to show that $U_n/V_n$ converges to zero. Let's start with finding an upper bound for $U_n$. We'd like to use Lemma 0 by Yang and Barron (1998). Since we condier a slightly different condition Assumption 1 instead of their Assumption 0, which is the condition for their Lemma 0, we have the following lemma which is a modified version of their Lemma 0:

**Lemma 4** *Suppose that Assumption 1 holds and*

$$\frac{\xi_j}{m_j} \geq \frac{4}{1-4\gamma} \log\left(\frac{46.2A_j\sqrt{1-4\gamma}}{\gamma}\right).$$

*Then*

$$P_0^*\left[\textit{for some } \theta \in \Theta_j, \frac{1}{n}\sum_{i=1}^{n} \log\frac{p_\theta(X_i)}{p_0(X_i)} \geq -\gamma d_H^2(p_0, p_\theta) + \frac{\xi_j}{n}\right]$$
$$\leq 15.1 \exp\left(-\frac{1-4\gamma}{8}\xi_j\right),$$

*where $P_0^*$ is the outer measure for $P_0^n$.*

Proof: The proof of this lemma follows from a part of the proof of Theorem 1 in Yang and Barron (1998), which shows that if Assumption 1 holds, then their Assumption 0 holds with $A = 3A_j$. So we can apply their Lemma 0 and obtain the above result.

Adapting Lemma 4, we have the following lemma which gives a bound for $U_n$ for a particular choice of the weights $a_j$:

**Lemma 5** *Suppose Assumption 1 holds. Let $a_j$ and $\eta_j$ be as specified in (2) and (3), then*

$$P_0^n\left[U_n > \alpha e^{-\gamma n s_n^2/2}\right] \leq 15.1 \exp\left(-\frac{(1-4\gamma)\gamma n s_n^2}{16}\right). \tag{14}$$

Proof: Take $\xi_j = \eta_j + \gamma n s_n^2/2$, then by Lemma 4,

$$\begin{aligned} U_n &\leq \left(\sum_j a_j e^{\xi_j}\right) e^{-\gamma n s_n^2} \\ &= \alpha e^{-\gamma n s_n^2/2} \sum_j \exp\left(-\frac{1-4\gamma}{8}\eta_j\right) \leq \alpha e^{-\gamma n s_n^2/2} \end{aligned}$$

except on a set of probability no greater than

$$\begin{aligned} &\sum_j 15.1 \exp\left(-\frac{1-4\gamma}{8}\xi_j\right) \\ &= 15.1 \exp\left(-\frac{(1-4\gamma)\gamma n s_n^2}{16}\right) \sum_j \exp\left(-\frac{1-4\gamma}{8}\eta_j\right) \\ &\leq 15.1 \exp\left(-\frac{(1-4\gamma)\gamma n s_n^2}{16}\right). \end{aligned}$$

Now we need to find a lower bound for $V_n$. To do this, we will use Lemma 1 by Shen and Wasserman (1998). Let

$$B_D(r) = \{p : D(p_0\|p) \le r^2, V'(p_0\|p) \le r^2\},$$

where $V'(f\|g) = \int f(\log(f/g) - D(f\|g))^2$. Here is their Lemma 1.

**Lemma 6** *For $t_n > 0$,*

$$P_0^n\left(V_n \le \frac{1}{2}\tilde{\pi}(B_D(t_n))e^{-2nt_n}\right) \le \frac{2}{nt_n}.$$

To use this lemma, we need to find a bound for $\tilde{\pi}(B_D(t_n))$. Note that

$$\tilde{\pi}(B_D(t_n)) = \sum_j a_j\tilde{\pi}_j(B_D(t_n)) \ge \sum_j a_j\pi_j(B_{D,j}(t_n)), \tag{15}$$

where

$$B_{D,j}(r) = \{\theta \in \Theta_j : D(p_0\|p_\theta) \le r^2, V(p_0\|p_\theta) \le r^2\}.$$

Therefore, it sufficies to give a lower bound for $\pi_j(B_{D,j}(t_n))$ for $j = j_n$. Now we will derive a lower bound for $\pi_{j_n}(B_{D,j_n}(t_n))$ assuming Assumptions 3 and 4 hold. By Assumption 3, there exist $\theta_1$,..., $\theta_{d^*} \in \Theta_j$ such that $d^* \le (A_{j_n}K_4)^{m_{j_n}}$ and $\cup_{i=1}^{d^*}B_{d_{j_n},j_n}(\theta_i,\epsilon_n) \supset \Theta_{j_n}$, where $B_{d_{j_n},j_n}(\theta,t_n)$ denotes the $d_{j_n}$-ball centered at $\theta$ with radius $t_n$ in $\Theta_{j_n}$ for $\theta \in \Theta_{j_n}$. Now, with the idea borrowed from a proof by Shen and Wasserman, it follows that

$$\pi_{j_n}(B_{D,j_n}(t_n)) \ge \frac{\pi_{j_n}(B_{D,j_n}(t_n))}{\sum_{i=1}^{d^*}\pi_{j_n}(B_{d_{j_n},j_n}(\theta_i,\epsilon_n))}. \tag{16}$$

To bound the numerator $\pi_{j_n}(B_{D,j_n}(t_n))$ in (16), we will show that

$$B_{d_{j_n},j_n}(\beta_n,\epsilon_n) \subset B_{D,j_n}(t_n), \tag{17}$$

where $\beta_n$ is given in (4). For $\theta \in B_{d_{j_n},j_n}(\beta_n,\epsilon_n)$, by Assumption 2,

$$V(p_0\|p_\theta) \le 2\epsilon_n^2 + 2K_0'\epsilon_n^2$$

and

$$D(p_0\|p_\theta) \le K_0''V(p_0\|p_\theta) \le 2K_0''(1+K_0')\epsilon_n^2,$$

so (17) holds for $t_n^2 = 2\max(1, K_0'')(1+K_0')\epsilon_n^2 \overset{\text{def}}{=} K'\epsilon_n^2$. Now by (17), (16) implies that

$$\pi_{j_n}(B_{D,j_n}(t_n)) \ge \frac{\pi_{j_n}(B_{d_{j_n},j_n}(\beta_n,\epsilon_n))}{\sum_{i=1}^{d^*}\pi_{j_n}(B_{d_{j_n},j_n}(\theta_i,\epsilon_n))},$$

which implies by Assumption 4 that

$$\pi_{j_n}(B_{D,j_n}(t_n)) \geq \left(\frac{1}{A_{j_n}^{b_1+b_2}K_4K_5}\right)^{m_{j_n}}.$$

Now we have obtained a lower bound for $\pi_{j_n}(B_{D,j_n}(t_n))$ with $t_n^2 = K'\epsilon_n^2$. By Lemma 6 and (15), except on a set of probability no greater than $2/(nt_n^2)$,

$$\begin{aligned}
V_n &\geq \frac{1}{2}e^{-2nt_n^2}a_{j_n}\pi_{j_n}(B_{D,j_n}(t_n)) \\
&\geq \frac{e^{-2nt_n^2}}{2}\alpha\exp\left(-\left(1+\frac{1-4\gamma}{8}\right)\eta_{j_n}\right)\left(\frac{1}{A_{j_n}^{b_1+b_2}K_4K_5}\right)^{m_{j_n}} \\
&\geq \frac{\alpha}{2}\exp\left(-2nt_n^2 - n\epsilon_n^2\left(1+\frac{1-4\gamma}{8}+b_1+b_2+(\log(K_4K_5))_+\right)\right) \\
&\geq \frac{\alpha}{2}e^{-Kn\epsilon_n^2}, \qquad (18)
\end{aligned}$$

where $K = 2K' + 1 + (1-4\gamma)/8 + b_1 + b_2 + (\log(K_4K_5))_+$. Here we have used the fact that

$$\frac{\eta_j}{m_j} \geq \frac{4}{1-4\gamma}\log\left(\frac{46.2A_j\sqrt{1-4\gamma}}{\gamma}\right) \geq \max(1, \log A_j)$$

for all $j$.

Now we will bound $U_n/V_n$ by combining (14) and (18). In (14) and (18), set

$$s_n^2 = \frac{4K\epsilon_n^2}{\gamma},$$

then

$$\pi(A_n^c|X_1,...,X_n) = \frac{U_n}{V_n} \leq 2\exp\left(-Kn\epsilon_n^2\right)$$

except on a set of probability no greater than

$$15.1\exp\left(-\frac{(1-4\gamma)Kn\epsilon_n^2}{4}\right) + \frac{2}{K'n\epsilon_n^2},$$

which converges to zero because $n\epsilon_n^2 \to 0$ by Assumption 5.

### 4.2 A Proof for Lemma 1

Here are two lemmas that we will use to prove Lemma 1. The first lemma follows from Lemma 1 by Yang and Barron (1998), which is useful for checking Assumption 1:

**Lemma 7** *Suppose that $\{S_\ell\}$ is a collection of linear function space on $[0,1]$. Suppose that for each $\ell$, there is a basis $B_{\ell,1}, ..., B_{\ell,m_\ell}$ for $S_\ell$ such that there exist constant $T_1$ and $T_2$ such that for $\theta = (\theta_1, ..., \theta_{m_\ell}) \in R^{m_\ell}$,*

$$\|\theta' B\|_\infty \overset{def}{=} \left\| \sum_{i=1}^{m_\ell} \theta_i B_{\ell,i} \right\|_\infty \le T_1 \max_i |\theta_i| \overset{def}{=} T_1 \|\theta\|_\infty \tag{19}$$

*and*

$$\|\theta' B\| \ge \frac{T_2}{\sqrt{m_\ell}} \sqrt{\sum_{i=1}^{m_\ell} \theta_i^2} \overset{def}{=} \frac{T_2}{\sqrt{m_\ell}} \|\theta\|, \tag{20}$$

*where $B = (B_{\ell,1}, ..., B_{\ell,m_\ell})$. Let*

$$\log p_\theta = -\psi(\theta) + \theta' B, \tag{21}$$

*where $\psi(\theta) = \log \int e^{\theta' B}$ is the normalizing constant. Suppose that $1 \in S_\ell$ and*

$$\Theta_j \subset \{\theta \in R^{m_\ell} : \|\log p_\theta\|_\infty \le L\}, \tag{22}$$

*where $j$ depends on $(\ell, L)$, then Assumption 1 holds with*

$$A_j = 19.28 \frac{T_1}{T_2} (L+1) e^{L/2} + 0.06 \tag{23}$$

*and $m_j = m_\ell$.*

The second lemma is a straightforward application of an equation in Lemma 1 by Barron and Sheu (1991), which says

$$D(p_0 \| p_\theta) \le \frac{1}{2} e^{\|\log p_0 - \log p_\theta\|_\infty} V(p_0 \| p_\theta), \tag{24}$$

and will be used for checking the second condition in Assumption 2.

**Lemma 8** *Suppose that*

$$\int p_0 \left( \log \frac{p_\eta}{p_\theta} \right)^2 \le K_0 d_{j_n}^2(\eta, \theta) \text{ for all } \eta, \theta \in \Theta_{j_n} \tag{25}$$

*for some constant* $K_0$ *and*

$$\sup_{\theta\in\Theta_{j_n}} \|\log p_0 - \log p_\theta\|_\infty \le \log K_3 \tag{26}$$

*for some constant* $K_3$*. Then Assumption 2 holds with* $K_0' = K_0$ *and* $K_0'' = K_3/2$*.*

Now let's verify Assumptions 1-5 for the spline example. To verify Assumption 1, we will apply Lemma 7. Note that (19) and (20) hold with $T_1 = 1$ and $T_2 = \gamma_q$ by the following fact taken from Schumaker (1981) (p.145, (4.86)):

$$\begin{cases} \|\theta' B\|_\infty \le \|\theta\|_\infty \\ \|\theta' B\| \ge \frac{\gamma_q}{\sqrt{m_\ell}}\|\theta\| \end{cases}$$

where $\gamma_q = 1/(\sqrt{q}(2q+1)9^{q-1})$, so by Lemma 7, Assumption 1 holds for $A_j$ and $m_j$ in (7).

Now before verifying Assumptions 2 - 5, we need to find $j$ and $\beta$ so that $R_{j,n}(\beta, p_0)$ is small and then compute $\epsilon_n$ so that (4) holds for $j_n = j$ and $\beta_n = \beta$. Before we start to find such $j_n$ and $\beta_n$, note that for the $C_j$ specified in (7), $\sum_j e^{-C_j} = e^{-2}/(1-e^{-1})^3 < 1$ as required. Now we can start to find $j$ and $\beta$ so that $R_{j,n}(\beta, p_0)$ is small. To control the approximation error $\max(D(p_0\|p_\beta), V(p_0\|p_\beta))$ in $R_{j,n}(\beta, p_0)$, we use the following fact: for $q \ge s+1$, there exists $\beta \in R^{m_j}$ such that

$$\begin{cases} \|D^r(\log p_0 - \log p_\beta)\|_\infty \le \alpha_q \left(\frac{1}{k+1}\right)^{s-r} M_0, & \text{for } 0 \le r \le s-1 \\ \|D^s \log p_\beta\|_\infty \le \alpha_q M_0, \end{cases} \tag{27}$$

where $M_0 = \max_{0\le r\le s} \|D^r \log p_0\|_{L_\infty}$. This fact follows from (6.50) in Schumaker (1981) and the fact that

$$|\psi(\theta)| = \left|\log \int e^{\theta' B - \log p_0} p_0\right| \le \|\theta' B - \log p_0\|_\infty$$

for all $\theta$. Note that the inequality in (27) for the case $r = 0$ is sufficient to control the error $\max(D(p_0\|p_\beta), V(p_0\|p_\beta))$ as long as $\beta \in \Theta_j$, which means that we need to make sure $\beta' 1_{m_j} = 0$ and $\max_{0\le r\le q-1} \|D^r \log p_\beta\|_{L_\infty} \le L$. For the first condition $\beta' 1_{m_j} = 0$, we can assume it without loss of generality, because $\log p_\beta$ doesn't change when $\beta$ is shifted by a constant. For the second condition to hold, we need to make sure $L$ is large enough. Let

$$q = s + 1 \stackrel{\text{def}}{=} q^*$$

and $L = L^*$, the smallest integer such that

$$L \geq 2^s + \alpha_{q^*} M_0 + M_0, \tag{28}$$

then by (27), $\max_{0 \leq r \leq q-1} \|D^r \log p_\beta\|_{L_\infty} \leq L^*$, so $\beta \in \Theta_j$. Now by (27) and the definition of $\eta_j$, we can find two constants $c_1$ and $c_2$ which depend only on $s$ and $M_0$ so that

$$R_{j,n}(\beta, p_0) \leq c_1 \left(\frac{1}{k+1}\right)^{2s} + c_2 \frac{k}{n}.$$

To make $R_{j,n}$ small, take $k = k_n$ such that $c_3 n^{1/(1+2s)} \leq k_n \leq c_4 n^{1/(1+2s)}$, where $c_3$ and $c_4$ are positive constants, then

$$R_{j,n}(\beta, p_0) \leq \left(\frac{c_1}{c_3^{2s}} + c_2 c_4\right) n^{-2s/(1+2s)} \leq \epsilon_n^2$$

if we take

$$\epsilon_n^2 = \left(\frac{c_1}{c_3^{2s}} + c_2 c_4\right) n^{-2s/(1+2s)} \stackrel{\text{def}}{=} c_5 n^{-2s/(1+2s)}. \tag{29}$$

In other words, take $j_n = (k_n, q^*, L^*)$, $\beta_n = \beta$ and $\epsilon_n$ as in (29), then (4) holds.

Now let's verify Assumptions 2 and 3 with $d_{j_n} = d_{j_n,\infty}$ for the above $j_n$, $\beta_n$ and $\epsilon_n$. It's clear that (25) holds with $K_0 = 1$ and that (26) holds with $K_3 = e^{2L^*}$, so by Lemma 8, Assumption 2 holds. Also, by Theorem IV and Theorem XIV of Kolmogrov and Tihomirov (1961), there exists an $\epsilon_n$-net $F_{\epsilon_n}$ for $\Theta_{j_n}$ with respect to $d_{j_n}$ so that

$$\begin{aligned} \log \operatorname{card}(F_{\epsilon_n}) &\leq c_{q^*,L^*} \left(\frac{1}{\epsilon_n}\right)^{1/(q^*-1)} \\ &= c_{q^*,L^*} \left(\frac{1}{\epsilon_n}\right)^{1/s} \leq c_{q^*,L^*}(k_n + 1) \leq c_{q^*,L^*} m_{j_n}. \end{aligned}$$

Therefore, Assumption 3 holds with $K_4 = e^{c_{q^*,L^*}}$ and $b_1 = 0$.

Now let's verify Assumption 4 for $d_{j_n} = d_{j_n,\infty}$ and for the above $j_n$, $\beta_n$ and $\epsilon_n$. To bound $\pi_{j_n}(B_{d_{j_n},j_n}(\beta_n, \epsilon_n))$, we will now show that

$$\left\{\theta \in R^{m_{j_n}} : \theta' 1_{m_{j_n}} = 0, \|\theta - \beta_n\|_\infty \leq c_6 \left(\frac{1}{k_n+1}\right)^s\right\} \subset B_{d_{j_n},j_n}(\beta_n, \epsilon_n), \tag{30}$$

where $c_6 = \min(1, c_3^s\sqrt{c_5}/2)$. For $\|\theta - \beta_n\|_\infty \le c_6(1/(k_n+1))^s$,

$$\|\log p_\theta - \log p_{\beta_n}\|_\infty \le 2\|\theta - \beta_n\|_\infty \le 2\frac{c_3^s\sqrt{c_5}/2}{c_3^s} n^{-s/(1+2s)} = \epsilon_n$$

since $k_n \ge c_3 n^{1/(1+2s)}$. The second inequality holds because

$$|\psi(\theta) - \psi(\beta_n)| = \left|\log \int e^{(\theta-\beta_n)'B} e^{\beta' B - \psi(\beta_n)}\right| \le \|\theta - \beta_n\|_\infty.$$

Now it remains to show that $\theta \in \Theta_{k_n, q^*, L^*}$. To bound the derivatives of $\log p_\theta$, we need the following inequality:

$$\|D^r(\theta' B - \beta' B)\|_{L_\infty} \le 2^r (k+1)^r \|\theta - \beta\|_\infty, \tag{31}$$

which is deduced from (4.54) in Schumaker (1981). Now note that for $0 < r < s$,

$$\begin{aligned}
\|D^r \log p_\theta\|_\infty &= \|D^r \theta' B\|_\infty \\
&\le \|D^r(\theta' B - \beta_n' B)\|_\infty + \|D^r(\beta_n' B - \log p_0)\|_\infty \\
&\quad + \|D^r \log p_0\|_\infty \\
&\overset{(31),(27)}{\le} 2^r (k_n+1)^r \|\theta - \beta_n\|_\infty + \alpha_{q^*} M_0 \left(\frac{1}{k_n+1}\right)^{s-r} \\
&\quad + M_0 \\
&\le \left(\frac{1}{k_n+1}\right)^{s-r} (2^r + \alpha_{q^*} M_0) + M_0 \overset{(28)}{\le} L^*,
\end{aligned}$$

for $r = 0$,

$$\begin{aligned}
\|\log p_\theta\|_\infty &\le \|\log p_\theta - \log p_{\beta_n}\|_\infty + \|\log p_{\beta_n} - \log p_0\|_\infty + \|\log p_0\|_\infty \\
&\le 2\|\theta - \beta_n\|_\infty + \|\log p_{\beta_n} - \log p_0\|_\infty + M_0 \\
&\le \left(\frac{1}{k_n+1}\right)^s (2 + \alpha_{q^*} M_0) + M_0 \overset{(28)}{\le} L^*
\end{aligned}$$

and for $r = s$,

$$\begin{aligned}
\|D^s \log p_\theta\|_{L_\infty} &= \|D^s \theta' B\|_{L_\infty} \\
&\le \|D^s(\theta' B - \beta_n' B)\|_{L_\infty} + \|D^s \beta_n' B\|_{L_\infty} \\
&\overset{(31),(27)}{\le} 2^s + \alpha_{q^*} M_0 \overset{(28)}{\le} L^*.
\end{aligned}$$

Therefore, $\theta \in \Theta_{k_n,q^*,L^*}$ and (30) holds. To bound $\pi_{j_n}(B_{d_{j_n},j_n}(\theta_1,\epsilon_n))$ in Assumption 4, note that for all $\epsilon > 0$ and for all $j$,

$$\{\theta \in \Theta_j : \|\log p_\theta - \log p_{\theta_1}\|_\infty \leq \epsilon\} \subset \{\theta \in \Theta_j : \|\theta - \theta_1\|_\infty \leq 2\beta_q^* \epsilon\}, \quad (32)$$

which follows from Lemma 4.3 by Ghosal, Ghosh and Van Der Vaart:

$$\|\theta - \theta_1\|_\infty \leq \beta_q^* \|\log p_{\theta-\theta_1}\|_\infty,$$

and the fact that

$$\begin{aligned}
&\|\log p_{\theta-\theta_1} - (\log p_{\theta_1} - \log p_\theta)\|_\infty \\
&= \ |\psi(\theta-\theta_1) - (\psi(\theta) - \psi(\theta_1))| \\
&= \ \left|\log \int \exp\left(\theta' B - \psi(\theta) - (\theta_1' B - \psi(\theta_1))\right)\right| \\
&\leq \ \|\log p_{\theta_1} - \log p_\theta\|_\infty.
\end{aligned}$$

Then by (32) and by (30), we have

$$\begin{aligned}
\frac{\pi_{j_n}(B_{d_{j_n},j_n}(\beta_n,\epsilon_n))}{\pi_{j_n}(B_{d_{j_n},j_n}(\theta_1,\epsilon_n))} &\geq \frac{\left(c_6\left(\frac{1}{k_n+1}\right)^s\right)^{k_n+q^*-1}}{(2\beta_{q^*}^*\epsilon_n)^{k_n+q^*-1}} \\
&\geq \left(\frac{c_6}{2\beta_{q^*}^*\epsilon_n\left(1+\left(\frac{c_4\sqrt{c_5}/2}{\epsilon_n}\right)^{1/s}\right)^s}\right)^{k_n+q^*-1}.
\end{aligned}$$

For $n$ such that $0 < \epsilon_n \leq 1$,

$$\begin{aligned}
\frac{\pi_{j_n}(B_{d_{j_n},j_n}(\beta_n,\epsilon_n))}{\pi_{j_n}(B_{d_{j_n},j_n}(\theta_1,\epsilon_n))} &\geq \left(\frac{c_6}{2\beta_{q^*}^*\epsilon_n\left(\left(\frac{1}{\epsilon_n}\right)^{1/s}+\left(\frac{c_4\sqrt{c_5}/2}{\epsilon_n}\right)^{1/s}\right)^s}\right)^{k_n+q^*-1} \\
&= \left(\frac{c_6}{2\beta_{q^*}^*(1+(c_4\sqrt{c_5}/2)^{1/s})^s}\right)^{k_n+q^*-1}.
\end{aligned}$$

Without loss of generality, we can assume that $2\beta_{q^*}^* > 1$, so it is clear that Assumption 4 holds with $K_5 = 2\beta_{q^*}^*(1+(c_4\sqrt{c_5}/2)^{1/s})^s/c_6$ and $b_2 = 0$.

For Assumption 5, it should be clear that it holds with the $\epsilon_n$ specified in (29). Now by Theorem 1, the result in Lemma 1 holds.

### 4.3 A Proof for Lemma 2

Here we are going to use Lemma 8 and the following lemma to verify the conditions in Theorem 1 for the Haar basis example:

**Lemma 9** *Suppose that $\{S_\ell\}$ is a collection of linear function space on $[0,1]$ and that for each $\ell$, there exists a constant $K_\ell > 0$ such that for all $h \in S_\ell$,*

$$\|h\|_\infty \le K_\ell \|h\|. \tag{33}$$

*Suppose that each $S_\ell$ is spanned by a bounded and linearly independent (under $L_2$ norm) basis 1, $B_{\ell,1}, ..., B_{\ell,m_\ell}$. For $\theta \in R^{m_\ell}$, let $p_\theta$ be as defined in (21). Suppose that (22) holds, then Assumption 1 holds with*

$$A_j = 19.28 K_\ell (L+1) e^{L/2} + 0.06 \tag{34}$$

*and $m_j = m_\ell + 1$.*

Proof: This lemma follows directly from Lemma 2 by Yang and Barron (1998). Now let's verify Assumptions 1-5. To verify Assumption 1, we will apply Lemma 9. First, by equation (3.7) in Barron, Birge and Massart (1999), (33) holds for $K_\ell = 2^{(\ell+1)/2}$. Second, $|\phi(\theta)| = |\log \int e^{\theta' B}| \le \|\theta' B\|_\infty$, so (22) holds with $2L$ in the place of $L$ for our $\Theta_j$. Therefore, by Lemma 9, Assumption 1 holds for $A_j$ and $m_j$ in (9).

Now before verifying Assumptions 2 - 5, we need to find $j$ and $\beta$ so that $R_{j,n}(\beta, p_0)$ is small and then compute $\epsilon_n$ so that (4) holds for $j_n = j$ and $\beta_n = \beta$. Before we start to find such $j_n$ and $\beta_n$, note that for the $C_j$ specified in (9), $\sum_j e^{-C_j} < 1$ as required. Now we can start to find $j$ and $\beta$ so that $R_{j,n}(\beta, p_0)$ is small. To control the approximation error $\max(D(p_0\|p_\beta), V(p_0\|p_\beta))$ in $R_{j,n}(\beta, p_0)$, we will use the coefficients of the $L^2$ expansion of $\log p_0$. Suppose that

$$\beta_0 + \sum_{i=1}^{m_\ell} \beta_{\ell,i} B_{\ell,i}$$

is the the $L^2$ projection of $\log p_0$ to the space spanned by 1 and $B_{\ell,i} : i = 1, ..., m_\ell$. Then by (8), $(\beta_0, \beta)$ satisfies

$$\|\log p_0 - \beta_0 - \beta' B\| \le H_0 (2^{k+2} - 1)^{-\alpha} \le H_0 \left(\frac{1}{m_j}\right)^\alpha,$$

and by the fact that $\beta_0 + \beta' B$ converges to $\log p_0$ uniformly, we have

$$\| \log p_0 - \beta_0 - \beta' B \|_\infty \le c_{2,p_0}.$$

To control the error $\log p_0 - \log p_\beta$ using $\log p_0 - \beta_0 - \beta' B$, we need to bound the difference between these two quantities, which is $\beta_0 + \psi(\beta)$, as follows:

$$\begin{aligned} |\beta_0 + \psi(\beta)| &= \left| \log \int e^{\beta_0 + \beta' B - \log p_0} p_0 \right| \\ &= |\log (1 + \Delta)| \\ &\le \max \left( \Delta, \frac{-\Delta}{1+\Delta} \right), \end{aligned}$$

where $\Delta = \int (e^{\beta_0 + \beta' B - \log p_0} - 1) p_0$. Let $b = \| \log p_0 - \beta_0 - \beta' B \|_\infty$, then it is clear that

$$e^{-b - M_0} \le 1 + \Delta \le e^{b + M_0},$$

so

$$\begin{aligned} |\beta_0 + \psi(\beta)| &\le |\Delta| e^{b + M_0} \\ &\le \left( 1 + \frac{1}{2} e^b \| \log p_0 - \beta_0 - \beta' B \| \right) \| \log p_0 - \beta_0 - \beta' B \| e^{b + 2M_0}, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality and equation (3.3) in Barron and Sheu (1991), which says that

$$\frac{z^2}{2} e^{-\max(-z,0)} \le e^z - 1 - z \le \frac{z^2}{2} e^{\max(z,0)} \text{ for all } z.$$

Therefore, we can now conclude that there exists $\beta \in R^{m_\ell}$ such that

$$\left\{ \begin{array}{l} \| \log p_0 - \log p_\beta \| \le c_{1,\alpha,M_0,H_0} \left( \frac{1}{m_j} \right)^\alpha \\ \| \log p_0 - \log p_\beta \|_\infty \le 2 c_{2,p_0} \end{array} \right. \tag{35}$$

Note that if $\beta \in \Theta_j$, then the $L^2$ bound in (35) gives a bound for the approximation error $\max(D(p_0 \| p_\beta), V(p_0 \| p_\beta))$ in $R_{j,n}(\beta, p_0)$ since

$$\begin{aligned} V(p_0 \| p_\theta) &= \int p_0 \left( \log \frac{p_0}{p_\beta} \right)^2 \\ &\le e^{\| \log p_0 \|_\infty} \| \log p_0 - \log p_\beta \|^2 \le e^{M_0} \| \log p_0 - \log p_\beta \|^2 \end{aligned} \tag{36}$$

and by (24) and (35),

$$D(p_0 \| p_\beta) \le \frac{1}{2} e^{2c_{2,p_0}} V(p_0 \| p_\beta). \tag{37}$$

To make sure that $\beta \in \Theta_j$, we have to make $L$ large enough so that $\| \log p_\beta \|_\infty \le L$. Take $L = L^*$, the smallest integer such that

$$L \ge 2c_{2,p_0} + 3M_0, \tag{38}$$

then (35) implies that $\| \log p_\beta \|_\infty \le L^*$, so $\beta \in \Theta_j$. Now by (35), (36), (37) and the definition of $\eta_j$, we can find two constants $k_1$ and $k_2$ which depend only on $\alpha$, $M_0$ and $H_0$ such that

$$R_{j,n}(\beta, p_0) \le k_1 \left( \frac{1}{m_j} \right)^{2\alpha} + k_2 \frac{m_j \log m_j}{n}.$$

Take $\ell = \ell_n$ such that $k_3 n^{1/(1+2\alpha)} \le m_j \le k_4 n^{1/(1+2\alpha)}$, where $k_3$ and $k_4$ are positive constants, then

$$R_{j,n}(\beta, p_0) \le \left( \frac{k_1}{k_3^{2\alpha}} + k_2 k_4 \log k_4 + \frac{k_2 k_4}{1+2\alpha} \right) n^{-2\alpha/(1+2\alpha)} \log n \le \epsilon_n^2$$

if we take

$$\epsilon_n^2 = \left( \frac{k_1}{k_3^{2\alpha}} + k_2 k_4 \log k_4 + \frac{k_2 k_4}{1+2\alpha} \right) n^{-2\alpha/(1+2\alpha)} \log n \stackrel{\text{def}}{=} k_5 n^{-2\alpha/(1+2\alpha)} \log n. \tag{39}$$

Now take $j_n = (\ell_n, L^*)$, $\beta_n = \beta$ and $\epsilon_n$ as in (39), then (4) holds.

Now let's verify Assumptions 2 and 3 with $d_{j_n}(\eta, \theta) = \|\eta - \theta\|$ for all $\eta$, $\theta \in \Theta_{j_n}$ for the above $j_n$, $\beta_n$ and $\epsilon_n$. First let's verify Assumption 2 using Lemma 8. For $\eta, \theta \in \Theta_{j_n}$, since

$$\begin{aligned}
\psi(\eta) - \psi(\theta) &= \log \int e^{(\theta-\eta)' B} p_\eta \\
&\le \log \int (1 + (\theta - \eta)' B e^{(\theta-\eta)' B}) p_\eta \\
&\le \log \left( 1 + \sqrt{\int ((\theta - \eta)' B)^2} \sqrt{\int e^{2(\theta-\eta)' B} p_\eta^2} \right) \\
&\le \log(1 + \|\theta - \eta\| e^{4L^*}) \\
&\le e^{4L^*} \|\theta - \eta\|,
\end{aligned}$$

$$\|\log p_\eta - \log p_\theta\|^2 = (\psi(\eta) - \psi(\theta))^2 + \|\eta - \theta\|^2$$

and

$$\int p_0 \left(\log \frac{p_\eta}{p_\theta}\right)^2 \quad \le \quad e^{\|\log p_0\|_\infty} \|\log p_\eta - \log p_\theta\|^2 \le e^{M_0} \|\log p_\eta - \log p_\theta\|^2,$$

(25) holds with $K_0 = e^{M_0}(1 + e^{8L^*})$ and clearly, (26) holds with $K_3 = e^{M_0 + 2L^*}$. Therefore, by Lemma 8, Assumption 2 holds. Also,

$$\Theta_{j_n} \subset \{\theta \in R^{\ell_n} : \|\theta\|_\infty \le L^*\},$$

which implies that for every $\epsilon > 0$, there exists an $\epsilon$-net $F_\epsilon$ for $\Theta_{j_n}$ with respect to $\|\cdot\|_\infty$ so that

$$\mathrm{card}(F_{\epsilon_n}) \le \left(1 + \frac{2L^*}{\epsilon}\right)^{m_{\ell_n}}.$$

By the fact that $\|\theta\| \le \sqrt{m_{\ell_n}} \|\theta\|_\infty$ for all $\theta \in \Theta_{j_n}$, there exist an $\epsilon_n$-net $F_{\epsilon_n}$ for $\Theta_{j_n}$ with respect to $d_{j_n}$ such that

$$\mathrm{card}(F_{\epsilon_n}) \le \left(1 + \frac{2L^* \sqrt{m_{\ell_n}}}{\epsilon_n}\right)^{m_{\ell_n}}.$$

Since

$$\frac{1 + \frac{2L^* \sqrt{m_{\ell_n}}}{\epsilon_n}}{A_{j_n}^{3\alpha}} \quad \le \quad \frac{\left(1 + \frac{2L^* \sqrt{k_4}}{k_5}\right) n^{1.5\alpha/(1+2\alpha)}}{k_3^{1.5\alpha} n^{1.5\alpha/(1+2\alpha)}},$$

Assumption 3 holds with $K_4 = (1 + 2L^* \sqrt{k_4}/k_5)/(k_3^{1.5\alpha})$ and $b_1 = 3\alpha$.

Now let's verify Assumption 4 for the above $d_{j_n}$, $j_n$, $\beta_n$ and $\epsilon_n$. To bound $\pi_{j_n}(B_{d_{j_n}, j_n}(\beta_n, \epsilon_n))$, we will now show that

$$\left\{\theta \in R^{m_{\ell_n}} : \|\theta - \beta_n\|_\infty \le \frac{\epsilon_n}{m_{j_n}\sqrt{m_{j_n} - 1}}\right\} \subset B_{d_{j_n}, j_n}(\beta_n, \epsilon_n) \tag{40}$$

for $n$ such that $\epsilon_n \le M_0$. For $\|\theta - \beta_n\|_\infty \le \epsilon_n/(m_{j_n}\sqrt{m_{j_n} - 1})$,

$$\|\theta - \beta\| \le \sqrt{m_{j_n} - 1} \|\theta - \beta\|_\infty \le \frac{\epsilon_n}{m_{j_n}} \le \epsilon_n,$$

so it sufficies to show that $\theta \in \Theta_{j_n}$. For $n$ such that $\epsilon_n \le M_0$,

$$\begin{aligned}\|\theta' B\|_\infty &\le \|\theta' B - \beta_n' B\|_\infty + \|\beta_0 + \beta_n' B - \log p_0\|_\infty + |\beta_0| + \|\log p_0\|_\infty \\ &\le m_{j_n}\|\theta - \beta_n\| + 2c_{2s}M_0 + 2M_0 \\ &\le \epsilon_n + 2c_{2s}M_0 + 2M_0 \\ &\le 2c_{2s}M_0 + 3M_0 \overset{(38)}{\le} L^*,\end{aligned}$$

so $\theta \in \Theta_{j_n}$ and (40) holds. To bound $\pi_{j_n}(B_{d_{j_n},j_n}(\theta_1, \epsilon_n))$ in Assumption 4, note that for all $\epsilon > 0$ and for all $j$,

$$\{\theta \in \Theta_j : \|\theta - \theta_1\| \le \epsilon\} \subset \{\theta \in \Theta_j : \|\theta - \theta_1\|_\infty \le \epsilon\}. \tag{41}$$

Now by (40) and (41) we have

$$\frac{\pi_{j_n}(B_{d_{j_n},j_n}(\theta_1, \epsilon_n))}{\pi_{j_n}(B_{d_{j_n},j_n}(\beta_n, \epsilon_n))} \le \left(\frac{\epsilon_n}{\epsilon_n/(m_{j_n}\sqrt{m_{j_n}-1})}\right)^{m_{j_n}-1} \le (m_{j_n}\sqrt{m_{j_n}-1})^{m_{j_n}}.$$

Since

$$\left(\frac{m_{j_n}\sqrt{m_{j_n}-1}}{A_{j_n}^3}\right)^{m_{j_n}} \le \left(\frac{m_{j_n}^{1.5}}{(\sqrt{m_{j_n}})^3}\right)^{m_{j_n}} = 1,$$

Assumption 4 holds with $b_2 = 3$ and $K_5 = 1$.

For Assumption 5, it should be clear that Assumption 5 holds with the $\epsilon_n$ specified in (39). Now by Theorem 1, the result in Lemma 2 holds.

### 4.4 A Proof for Theorem 2

We prove Theorem 2 by giving bounds for $U_n$ and $V_n$ respectively, and then combining the bounds to show that $U_n/V_n$ converges to zero.

Let's start with finding an upper bound for $U_n$. To bound $U_n$, we will derive an exponential inequality similar to the one in Lemma 0 by Yang and Barron (1998), and then apply that inequality to derive a bound for $U_n$.

**Lemma 10** *Suppose that Assumption 6 holds and $\gamma \in (0, 0.25)$ is defined so that*

$$0.0056 = \frac{0.13}{c_{2,c_0,M}\sqrt{c_{1,M,\sigma}}}\frac{\gamma}{\sqrt{1-4\gamma}}.$$

*Then for $\xi_j$ such that*

$$\frac{\xi_j}{m_j} \ge \frac{4}{c_{1,M,\sigma}(1-4\gamma)}\log(1072.5A_j),$$

$$
\begin{aligned}
P_{f_0}^* &\Bigg[ \frac{1}{n}\sum_{i=1}^n (Y_i - f_0(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2 \geq -\gamma\|f_0 - f_\theta\|^2 \\
&\quad + \frac{\xi_j}{n} + 0.0224 \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \right| \sqrt{\frac{\xi_j}{n}} \text{ for some } \theta \in \Theta_j \text{ and} \\
&\quad \frac{1}{n}\sum_{i=1}^n |\epsilon_i| \leq c_0, \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \leq c_0^2 \Bigg] \\
&\leq \; 15.1 \exp\left( -\frac{c_{1,M,\sigma}(1-4\gamma)\xi_j}{8} \right),
\end{aligned}
$$

*where*

$$
c_{1,M,\sigma} = \min\left( \frac{1 - \exp\left(-\frac{M^2}{2\sigma^2}\right)}{2M^2}, \frac{1}{2\sigma^2} \right) \text{ and } c_{2,c_0,M} = 2(c_0 + 2M).
$$

Proof: The proof for Lemma 10 is long and is deferred to Section 4.4.1.

Apply Lemma 10, then we have the following lemma which gives a bound for $U_n$:

**Lemma 11** *Suppose that Assumption 6 holds. Take* $c_0 = 2\sigma$ *and define* $\gamma$ *as in Lemma 10. Let* $C_j \geq 0$ *be such that* $\sum_j e^{-C_j} \leq 1$ *and define* $\eta_j$ *and* $a_j$ *as in (11) and (10) respectively. Then*

$$
P_{f_0}\left[ U_n \leq \alpha \exp\left( \frac{0.0056 Z_n^2}{\sigma} - \frac{\gamma n s_n^2}{4\sigma^2} \right) \right] \geq 1 - (p_1 + p_2 + p_3), \tag{42}
$$

*where*

$$
Z_n = \frac{1}{\sqrt{n}\sigma}\sum_{i=1}^n \epsilon_i \sim N(0,1),
$$

$$
p_1 = P\left[ \frac{1}{n}\sum_{i=1}^n |\epsilon_i| > c_0 \right],
$$

$$
p_2 = P\left[ \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 > c_0^2 \right]
$$

*and*

$$
p_3 = 15.1 \exp\left( -\frac{c_{1,M,\sigma}(1-4\gamma)\gamma n s_n^2}{32(0.5 + 0.0056\sigma)} \right).
$$

Proof: Take

$$\xi_j = \eta_j + \frac{\gamma n s_n^2}{4(0.5 + 0.0056\sigma)}.$$

Since $U_n$ is

$$\sum_j a_j \int_{(B_{d,\Theta_j}(s_n))^c} \frac{\exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - f_0(X_i))^2\right)}{\exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2\right)} d\pi_j(\theta),$$

Lemma 10 gives

$$\begin{aligned}
U_n &\leq \sum_j a_j \exp\left(\frac{1}{2\sigma^2}\left(-\gamma n s_n^2 + \xi_j + 0.0224\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\right|\sqrt{\xi_j}\right)\right) \\
&= \sum_j a_j \exp\left(-\frac{\gamma n s_n^2}{2\sigma^2} + \frac{\xi_j}{2\sigma^2} + \frac{0.0112}{\sigma}|Z_n|\sqrt{\xi_j}\right) \\
&\leq \sum_j a_j \exp\left(-\frac{\gamma n s_n^2}{2\sigma^2} + \frac{\xi_j}{2\sigma^2} + \frac{0.0056}{\sigma}(Z_n^2 + \xi_j)\right) \\
&= \exp\left(\frac{0.0056 Z_n^2}{\sigma} - \frac{\gamma n s_n^2}{4\sigma^2}\right)\sum_j a_j \exp\left(\frac{0.5 + 0.0056\sigma}{\sigma^2}\eta_j\right) \\
&= \alpha \exp\left(\frac{0.0056 Z_n^2}{\sigma} - \frac{\gamma n s_n^2}{4\sigma^2}\right)\sum_j e^{-\eta_j} \\
&\leq \alpha \exp\left(\frac{0.0056 Z_n^2}{\sigma} - \frac{\gamma n s_n^2}{4\sigma^2}\right)
\end{aligned}$$

except on a set of probability no greater than

$$P\left[\frac{1}{n}\sum_{i=1}^n |\epsilon_i| > c_0\right] + P\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 > c_0^2\right] + 15.1\sum_j \exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\xi_j}{8}\right).$$

Note that

$$\begin{aligned}
&\sum_j \exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\xi_j}{8}\right) \\
&= \exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\gamma n s_n^2}{32(0.5+0.0056\sigma)}\right)\sum_j \exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\eta_j}{8}\right) \\
&\leq \exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\gamma n s_n^2}{32(0.5+0.0056\sigma)}\right),
\end{aligned}$$

so now we have the following bound for $U_n$:

$$P_{f_0}\left[U_n \le \alpha \exp\left(\frac{0.0056 Z_n^2}{\sigma} - \frac{\gamma n s_n^2}{4\sigma^2}\right)\right] \ge 1 - (p_1 + p_2 + p_3).$$

The process of deriving a bound for $V_n$ is the same as in Section 4.1 except the following changes:

- Set $p_\theta = p_{f_\theta}$ and $p_0 = p_{f_0}$.
- $\eta_j$ is as defined in (11) instead of (3).
- (17) follows from Assumption 7 instead of Assumption 2. The arguments for obtaining (17) are modified in the following way: First, note that in our regression setting, for all $\theta \in \Theta_j$ and for all $j$,

$$D(p_0 \| p_\theta) = \frac{\|f_0 - f_\theta\|^2}{2\sigma^2} \tag{43}$$

and

$$V(p_0\|p_\theta) = \frac{\|f_0 - f_\theta\|^2}{\sigma^2} + \frac{1}{4\sigma^4}\int (f_0 - f_\theta)^4 \le \left(\frac{1}{\sigma^2} + \frac{M^2}{\sigma^4}\right)\|f_0 - f_\theta\|^2. \tag{44}$$

Then for $\theta \in B_{d_{j_n}, j_n}(\beta, \epsilon_n)$ with $\beta = \beta_n$ satisfying (4), by (43) and (12),

$$D(p_0\|p_\theta) \le D(p_0\|p_\beta) + \frac{\|f_\beta - f_\theta\|^2}{2\sigma^2} \le \epsilon_n^2 + \frac{K_0'\epsilon_n^2}{2\sigma^2}$$

and by (44),

$$V(p_0\|p_\theta) \le \left(2 + \frac{2M^2}{\sigma^2}\right) D(p_0\|p_\theta),$$

so (17) holds for

$$t_n^2 = \left(2 + \frac{2M^2}{\sigma^2}\right)\left(1 + \frac{K_0'}{2\sigma^2}\right)\epsilon_n^2 \stackrel{\text{def}}{=} K'\epsilon_n^2.$$

- The process of deriving a lower bound for $V_n$ in (18) is replaced by the following:

$$V_n \ \ge \ \frac{1}{2} e^{-2nt_n^2} a_{j_n} \pi_{j_n}(B_{D,j_n}(t_n))$$

$$\begin{aligned}
&\geq \frac{\alpha e^{-2nt_n^2}}{2} \exp\left(-\left(1+\frac{1}{2\sigma^2}+\frac{0.0056}{\sigma}\right)\eta_{j_n}\right)\left(\frac{1}{A_{j_n}^{b_1+b_2}K_4K_5}\right)^{m_{j_n}} \\
&\geq \frac{\alpha}{2}\exp\left(-2nt_n^2 - n\epsilon_n^2\left(1+\frac{1}{2\sigma^2}+\frac{0.0056}{\sigma}+c_1(b_1+b_2\right.\right. \\
&\quad \left.\left. +(\log(K_4K_5))_+)\right)\right) \\
&\geq \frac{\alpha}{2}e^{-Kn\epsilon_n^2}, \qquad (45)
\end{aligned}$$

where $c_1 = c_{1,M,\sigma}$ and

$$K = 2K' + 1 + \frac{1}{2\sigma^2} + \frac{0.0056}{\sigma} + c_1(b_1 + b_2 + (\log(K_4K_5))_+).$$

Here we have used the fact that

$$\frac{c_1\eta_j}{m_j} \geq \frac{4}{1-4\gamma}\log\left(\frac{1072.5A_j\sqrt{1-4\gamma}}{\gamma}\right) \geq \max(1, \log A_j)$$

for all $j$.

Now we will bound $U_n/V_n$ by combining (42) and (45). In (42) and (45), set

$$s_n^2 = \frac{8\sigma^2 K\epsilon_n^2}{\gamma},$$

then

$$\pi(A_n^c|X_1, ..., X_n) = \frac{U_n}{V_n} \leq 2\exp\left(\frac{0.0056Z_n^2}{\sigma}\right)\exp\left(-Kn\epsilon_n^2\right)$$

except on a set of probability no greater than

$$p_1 + p_2 + 15.1\exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)8\sigma^2Kn\epsilon_n^2}{32(0.5+0.0056\sigma)}\right) + \frac{2}{K'n\epsilon_n^2},$$

where

$$Z_n = \frac{1}{\sqrt{n}\sigma}\sum_{i=1}^n \epsilon_i \sim N(0,1),$$

$$p_1 = P\left[\frac{1}{n}\sum_{i=1}^n |\epsilon_i| > c_0\right],$$

$$p_2 = P\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 > c_0^2\right].$$

Note that $c_0 = 2\sigma > \max(E|\epsilon_i|, E\epsilon_i^2)$, so $p_1 + p_2 \to 0$ as $n \to \infty$. Now suppose that Assumption 5 holds. Since $2e^{0.0056 Z_n^2/\sigma}$ converges in distribution and $e^{-Kn\epsilon_n^2}$ converges to zero, we have $2e^{0.0056 Z_n^2/\sigma} e^{-Kn\epsilon_n^2}$ converges to zero in probability, so $\pi(A_n^c|X_1, ..., X_n)$ converges to zero in probability. Now we have proved Theorem 2.

### 4.4.1 An Exponentional Inequality

First, we claim that to prove Lemma 10, it sufficies to prove the following lemma with a slightly different assumption:

**Assumption 8** *$\Theta$ is a parameter space such that for $\theta \in \Theta$, $\|f_\theta\|_\infty \le M$, and there exist constants $A > 0$, $m \ge 1$ and $0 < \rho \le A$ such that for any $r > 0$, $\delta \le \rho r$, $\theta \in \Theta$, the $\delta$-covering number of set $B_{\|\cdot\|,\Theta}(r)$ with respect to metric $d_\infty$*

$$N(B_{\|\cdot\|,\Theta}(r), \delta, d_\infty) \le \left(\frac{Ar}{\delta}\right)^m,$$

*where $B_{\|\cdot\|,\Theta}(r) = \{\theta \in \Theta : \|f_0 - f_\theta\| \le r\}$ and $d_\infty(\theta, \eta) = \|f_\theta - f_\eta\|_\infty$ for all $\theta$, $\eta \in \Theta$.*

**Lemma 12** *Suppose that Assumption 8 holds with*

$$\rho \ge \frac{0.13}{c_{2,c_0,M}\sqrt{c_{1,M,\sigma}}} \frac{\gamma}{\sqrt{1-4\gamma}}.$$

*Then for $\xi$ such that*

$$\frac{\xi}{m} \ge \frac{4}{c_{1,M,\sigma}(1-4\gamma)} \log\left(15.4 c_{2,c_0,M}\sqrt{c_{1,M,\sigma}} A \frac{\sqrt{1-4\gamma}}{\gamma}\right),$$

$$\begin{aligned} P^* \Bigg[ & \frac{1}{n}\sum_{i=1}^n (Y_i - f_0(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2 \ge -\gamma\|f_0 - f_\theta\|^2 + \frac{\xi}{n} \\ & +4\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right| \delta \textit{ for some } \theta \in \Theta \textit{ and } \frac{1}{n}\sum_{i=1}^n |\epsilon_i| \le c_0, \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \le c_0^2 \Bigg] \\ \le \; & 15.1 \exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right), \end{aligned}$$

*where*

$$\delta = \frac{2\gamma}{15.4 c_{2,c_0,M}\sqrt{c_{1,M,\sigma}(1-4\gamma)}}\sqrt{\frac{\xi}{n}},$$

$$c_{1,M,\sigma} = \min\left(\frac{1-\exp\left(-\frac{M^2}{2\sigma^2}\right)}{2M^2}, \frac{1}{2\sigma^2}\right) \quad \textit{and } c_{2,c_0,M} = 2(c_0+2M).$$

Our claim follows from a part of the proof of Theorem 1 in Yang and Barron (1998), which shows that if Assumption 6 holds, then Assumption 8 holds with $A = 3A_j$ and $\rho = 0.0056$. Suppose that Lemma 12 is true, then Lemma 10 follows by setting $\rho = 0.0056$ and choosing $\gamma$ such that

$$\rho = \frac{0.13}{c_{2,c_0,M}\sqrt{c_{1,M,\sigma}}}\frac{\gamma}{\sqrt{1-4\gamma}}.$$

So let's prove Lemma 12. The proof for Lemma 12 is very similar to the proof of Lemma 0 in Yang and Barron (1998). First, divide the space $\Theta$ into rings

$$\Theta_i = \{\theta \in \Theta : r_{i-1} \le \|f_0 - f_\theta\| \le r_i\}, i = 0, 1, ...,$$

where $r_i = 2^{i/2}\sqrt{\xi/n}$ for $i \ge 0$ and $r_{-1} = 0$. For each ring $\Theta_i$, we will use a chaining argument to bound

$$q_i \stackrel{\text{def}}{=} P^*\left[\frac{1}{n}\sum_{i=1}^n (Y_i - f_0(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2 \ge -\gamma\|f_0 - f_\theta\|^2 + \frac{\xi}{n}\right.$$

$$\left. +4\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right|\delta \text{ for some } \theta \in \Theta_i \text{ and } \frac{1}{n}\sum_{i=1}^n |\epsilon_i| \le c_0, \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \le c_0^2\right].$$

Then we will put all the bounds for $q_i$ together to complete the proof. So let's focus on one $\Theta_i$ first. Let $\{\delta_k\}_{k=0}^\infty$ be a sequence decreasing to zero with $\delta_0 \le \min(\rho r_0, \delta)$ and define $\tilde{\delta}_k = \delta_k$ for $k \ge 1$ and $\tilde{\delta}_0 = \delta_0/2$. Then by assumption we can find a sequence of nets $\tilde{F}_0$, $\tilde{F}_1$,..., where each $\tilde{F}_k$ is a $\tilde{\delta}_k$ net in $\Theta_i$ satisfying the cardinal number constraint in Assumption 8. In other words, for each $k$, there exists a mapping $\tilde{\tau}_k : \Theta_i \to \tilde{F}_k$ such that $\|f_{\tilde{\tau}_k(\theta)} - f_\theta\|_\infty \le \tilde{\delta}_k$ for all $\theta \in \Theta_i$, and

$$\text{Card}(\tilde{F}_k) \le \left(\frac{Ar_k}{\tilde{\delta}_k}\right)^m.$$

Instead of applying the chaining argument using the nets $\tilde{F}_k$, we will modify the net $\tilde{F}_0$ first and then apply the chaining argument using the nets $F_k$, where $F_k = \tilde{F}_k$ for $k \ge 1$ and $F_0$ is the modified $\tilde{F}_0$. Now modify the net $\tilde{F}_0$

in the following way: consider a positive number $\epsilon$. For each $\tilde{\theta}_0$ in $\tilde{F}_0$, find $\theta_0$ in

$$\tilde{\tau}_0^{-1}(\tilde{\theta}_0) = \{\theta \in \Theta_i : \tilde{\tau}_0(\theta) = \tilde{\theta}_0\}$$

such that

$$\|f_0 - f_{\theta_0}\|^2 \leq \inf_{\theta \in \tilde{\tau}_0^{-1}(\tilde{\theta}_0)} \|f_0 - f_\theta\|^2 + \epsilon.$$

Define $\tau(\tilde{\theta}_0) = \theta_0$, and $F_0 = \{\tau(\tilde{\theta}_0) : \tilde{\theta}_0 \in \tilde{F}_0\}$. Define $\tau_0 = \tau(\tilde{\tau}_0)$ and $\tau_k = \tilde{\tau}_k$ for $k \geq 1$. Then by triangle inequality, $\|f_{\tau_0(\theta)} - f_\theta\|_\infty \leq \delta_0$, so $F_0$ is a $\delta_0$ net and for each $k$, $F_k$ is a $\delta_k$ net. Now we can start the chaining argument. For each $\theta \in \Theta_i$, define

$$\ell_0 = \frac{1}{n}\sum_{i=1}^n (Y_i - f_0(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f_{\tau_0(\theta)}(X_i))^2$$

and

$$\ell_k = \frac{1}{n}\sum_{i=1}^n (Y_i - f_{\tau_{k-1}(\theta)}(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f_{\tau_k(\theta)}(X_i))^2$$

for $k \geq 1$, then

$$\frac{1}{n}\sum_{i=1}^n (Y_i - f_0(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2 = \ell_0 + \sum_{k=1}^\infty \ell_k.$$

Now, instead of giving bounds for $\ell_k - E\ell_k$ as in Yang and Barron (1998), we will give bounds for $\ell_k - E_\epsilon \ell_k$, where

$$E_\epsilon \ell_k = \frac{2}{n}\sum_{i=1}^n \epsilon_i \int (f_{\tau_k(\theta)} - f_{\tau_{k-1}(\theta)}) + \|f_0 - f_{\tau_{k-1}(\theta)}\|^2 - \|f_0 - f_{\tau_k(\theta)}\|^2$$

is the conditional expectation of $\ell_k$ given $\epsilon_1$,...,$\epsilon_n$ for $k \geq 1$. Note that

$$\begin{aligned}
\sum_{k=1}^\infty E_\epsilon \ell_k &= 2\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i\right)\int (f_\theta - f_{\tau_0(\theta)}) + \|f_0 - f_{\tau_0(\theta)}\|^2 - \|f_0 - f_\theta\|^2 \\
&\leq 2\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right| \int (f_\theta - f_{\tau_0(\theta)}) + \epsilon \\
&\leq 4\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right| \delta_0 + \epsilon \leq 4\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right| \delta + \epsilon,
\end{aligned}$$

so

$$\begin{aligned} q_i &\le P^*(B_0 \cap B) \\ &\le P^*\left(\left\{\ell_0 \ge -2\gamma r_i^2 + \frac{\xi}{n} - \epsilon \text{ for some } \theta \in \Theta_i\right\} \cap B\right) \\ &\quad + \sum_{k=1}^{\infty} P^*\left(\{\ell_k - E_\epsilon \ell_k \ge \eta_k \text{ for some } \theta \in \Theta_i\} \cap B\right) \\ &\stackrel{\text{def}}{=} q_i^{(1)} + q_{i,k}^{(2)} \end{aligned}$$

if $\sum_{k=1}^{\infty} \eta_k \le \gamma r_i^2$, where

$$B_0 = \left\{\ell_0 + \sum_{k=1}^{\infty}(\ell_k - E_\epsilon \ell_k) \ge -\epsilon - \gamma \|f_0 - f_\theta\|^2 + \frac{\xi}{n} \text{ for some } \theta \in \Theta_i\right\}$$

and

$$B = \left\{\frac{1}{n}\sum_{i=1}^{n}|\epsilon_i| \le c_0, \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \le c_0^2\right\}.$$

To bound $q_i^{(1)}$, we will use the following inequality by Chernoff (1952):

**Fact 1** *Suppose that $X_i$ are iid from a distribution with density $g_2$ with respect to measure $\mu$ and $g_1$ is a density with respect to the same measure. Then*

$$P\left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{g_1(X_i)}{g_2(X_i)} \ge t\right] \le \exp\left(-\frac{n}{2}\left(d_H^2(g_1, g_2) + t\right)\right).$$

Since

$$\ell_0 = \frac{2\sigma^2}{n}\sum_{i=1}^{n}\log\frac{p_{f_{\tau_0(\theta)}}(X_i)}{p_{f_0}(X_i)},$$

Fact 1 implies that for a $\tau_0(\theta)$,

$$P\left[\ell_0 \ge t\right] \le \exp\left(-\frac{n}{2}\left(d_H^2(p_{f_{\tau_0(\theta)}}, p_{f_0}) + t/(2\sigma^2)\right)\right). \tag{46}$$

To replace the Hellinger distance $d_H^2(p_{f_{\tau_0(\theta)}}, p_{f_0})$ with the $L^2$ distance $\|f_{\tau_0(\theta)} - f_0\|$ in (46), note that

$$\begin{aligned} d_H^2(p_{f_{\tau_0(\theta)}}, p_{f_0}) &= 2\int\left(1 - \exp\left(-\frac{(f_{\tau_0(\theta)}(x) - f_0(x))^2}{8\sigma^2}\right)\right) d\mu(x) \\ &\ge \frac{1 - \exp\left(-\frac{M^2}{2\sigma^2}\right)}{2M^2}\int (f_{\tau_0(\theta)}(x) - f_0(x))^2 d\mu(x) \\ &\stackrel{\text{def}}{=} c_{0,M,\sigma}\|f_{\tau_0(\theta)} - f_0\|^2. \end{aligned} \tag{47}$$

Here the equality follows from direct calculation and the inequality follows from the fact that $(1-e^{-x})/x$ is decreasing with $x$ on $(0,\infty)$ and that $\|f_{\tau_0(\theta)}\|_\infty, \|f_0\|_\infty \le M$. Now by (46) and (47), we have

$$\begin{aligned} P[\ell_0 \ge t] &\le \exp\left(-\frac{n}{2}\left(c_{0,M,\sigma}\|f_{\tau_0(\theta)} - f_0\|^2 + t/(2\sigma^2)\right)\right) \\ &\le \exp\left(-\frac{c_{1,M,\sigma}n}{2}(\|f_{\tau_0(\theta)} - f_0\|^2 + t)\right), \end{aligned}$$

where $c_{1,M,\sigma} = \min(c_{0,M,\sigma}, 1/(2\sigma^2))$. Set $t = -2\gamma r_i^2 + \frac{\xi}{n} - \epsilon$, then for a $\tau_0(\theta)$,

$$P\left[\ell_0 \ge -2\gamma r_i^2 + \frac{\xi}{n} - \epsilon\right] \le \exp\left(-\frac{c_{1,M,\sigma}n}{2}(r_{i-1}^2 - 2\gamma r_i^2 + \frac{\xi}{n} - \epsilon)\right).$$

Therefore,

$$\begin{aligned} q_i^{(1)} &\le \mathrm{Card}(F_0)\exp\left(-\frac{c_{1,M,\sigma}n}{2}(r_{i-1}^2 - 2\gamma r_i^2 + \frac{\xi}{n} - \epsilon)\right) \\ &\le \mathrm{Card}(F_0)\exp\left(-\frac{c_{1,M,\sigma}n}{2}((i+1)(1-4\gamma)\frac{\xi}{n} - \epsilon)\right), \end{aligned} \tag{48}$$

where the last inequality was verified in Yang and Barron (1998) (from the end of p.111 to the beginning of p.112).

To bound $q_{i,k}^{(2)}$, we will use Hoeffding's inequality:

**Fact 2** *Suppose that $\{Y_i\}_{i=1}^n$ are independent with mean zero and that $a_i \le Y_i \le b_i$ for all $i$. Then for $\eta > 0$,*

$$P\left[\sum_{i=1}^n Y_i \ge \eta\right] \le \exp\left(\frac{-2\eta^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

For a pair $(\tau_{k-1}(\theta), \tau_k(\theta))$,

$$\begin{aligned} &|(Y_i - f_{\tau_{k-1}(\theta)}(X_i))^2 - (Y_i - f_{\tau_k(\theta)}(X_i))^2| \\ \le\ & 2|f_{\tau_{k-1}(\theta)}(X_i) - f_{\tau_k(\theta)}(X_i)|\left|\epsilon_i + f_0(X_i) - \frac{f_{\tau_{k-1}(\theta)}(X_i) + f_{\tau_k(\theta)}(X_i)}{2}\right| \\ \le\ & 2(\delta_{k-1} + \delta_k)(|\epsilon_i| + 2M) \le 4(|\epsilon_i| + 2M)\delta_{k-1}. \end{aligned}$$

By Hoeffding's inequality, the conditional probability

$$\begin{aligned} P\left[\ell_k - E_\epsilon \ell_k \ge \eta | \epsilon_1, ..., \epsilon_n\right] &\le \exp\left(\frac{-2n^2\eta^2}{\sum_{i=1}^n 64(|\epsilon_i| + 2M)^2\delta_{k-1}^2}\right) \\ &\le \exp\left(\frac{-2n\eta^2}{64(c_0 + 2M)^2\delta_{k-1}^2}\right) \end{aligned}$$

if $\sum_{i=1}^n |\epsilon_i|/n \le c_0$ and $\sum_{i=1}^n \epsilon_i^2/n \le c_0^2$. Integrate the conditional probability over set $B$, then we have

$$P\left(\{\ell_k - E_\epsilon \ell_k \ge \eta\} \cap B\right) \le \exp\left(\frac{-2n\eta^2}{64(c_0+2M)^2\delta_{k-1}^2}\right).$$

Therefore,

$$q_{i,k}^{(2)} \le \mathrm{Card}(F_{k-1})\mathrm{Card}(F_k) \le \exp\left(\frac{-2n\eta_k^2}{64(c_0+2M)^2\delta_{k-1}^2}\right). \tag{49}$$

Now combine (48) and (49) and let $\epsilon \to 0$, then we have

$$\begin{aligned}
q_i &\le \mathrm{Card}(F_0)\exp\left(-\frac{nc_{1,M,\sigma}}{2}(i+1)(1-4\gamma)\frac{\xi}{n}\right)\\
&\quad + \sum_{k=1}^{\infty}\mathrm{Card}(F_{k-1})\mathrm{Card}(F_k)\exp\left(\frac{-2n\eta_k^2}{64(c_0+2M)^2\delta_{k-1}^2}\right)\\
&\le \left(\frac{Ar_i}{\tilde{\delta}_0}\right)^m \exp\left(-\frac{c_{1,M,\sigma}}{2}(i+1)(1-4\gamma)\xi\right)\\
&\quad + \sum_{k=1}^{\infty}\left(\frac{Ar_i}{\tilde{\delta}_{k-1}}\right)^m\left(\frac{Ar_i}{\tilde{\delta}_k}\right)^m \exp\left(\frac{-2n\eta_k^2}{64(c_0+2M)^2\delta_{k-1}^2}\right).
\end{aligned}$$

Now choose $\delta_0$, $\delta_j$ so that

$$\log\left(\frac{Ar_0}{\tilde{\delta}_k}\right)^m = \frac{c_{1,M,\sigma}(k+1)(1-4\gamma)\xi}{4}$$

and $\eta_k$ such that

$$\begin{aligned}
\frac{2n\eta_k^2}{64(c_0+2M)^2\delta_{k-1}^2} &= im\log 2 + \frac{(2k+1)c_{1,M,\sigma}(1-4\gamma)\xi}{4}\\
&\quad + \frac{(i+1)kc_{1,M,\sigma}(1-4\gamma)\xi}{8}.
\end{aligned}$$

Now the bound for $q_i$ becomes

$$\begin{aligned}
q_i &\le 2^{im/2}\exp\left(\frac{c_{1,M,\sigma}(1-4\gamma)\xi}{4}\right)\exp\left(-\frac{c_{1,M,\sigma}}{2}(i+1)(1-4\gamma)\xi\right)\\
&\quad + \sum_{k=1}^{\infty}\exp\left(-\frac{(i+1)c_{1,M,\sigma}k(1-4\gamma)\xi}{8}\right)
\end{aligned}$$

$$\begin{aligned}
&\leq \exp\left(\frac{im}{2}\log 2 - \frac{(i+1)c_{1,M,\sigma}(1-4\gamma)\xi}{4}\right) \\
&\quad + \exp\left(-\frac{(i+1)c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right) \\
&\quad \cdot \left(1 - \exp\left(-\frac{(i+1)c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right)\right)^{-1}.
\end{aligned}$$

Note that by assumption,

$$\frac{m}{2}\log\frac{2A}{\rho_0} \leq \frac{c_{1,M,\sigma}(1-4\gamma)\xi}{8},$$

where

$$\rho_0 = \frac{2\gamma}{15.4c_{2,c_0,M}\sqrt{c_{1,M,\sigma}(1-4\gamma)}}.$$

Since $\rho_0 \leq \rho \leq A$, we have

$$\frac{\log 2}{2} \leq \frac{m}{2}\log 2 \leq \frac{m}{2}\log\frac{2A}{\rho_0} \leq \frac{c_{1,M,\sigma}(1-4\gamma)\xi}{8}, \tag{50}$$

so

$$\begin{aligned}
q_i &\leq \exp\left(-\frac{(i+1)c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right) \\
&\quad \cdot \left(1 + \left(1 - \exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right)\right)^{-1}\right) \\
&\leq \left(1 + \frac{\sqrt{2}}{\sqrt{2}-1}\right)\exp\left(-\frac{(i+1)c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right),
\end{aligned}$$

and

$$\begin{aligned}
&P^*\Bigg[\frac{1}{n}\sum_{i=1}^n (Y_i - f_0(X_i))^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - f_\theta(X_i))^2 \geq -\gamma\|f_0 - f_\theta\|^2 + \frac{\xi}{n} \\
&\qquad +4\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right|\delta \text{ for some } \theta \in \Theta \text{ and } \frac{1}{n}\sum_{i=1}^n |\epsilon_i| \leq c_0, \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \leq c_0^2\Bigg] \\
&\leq \sum_{i=0}^\infty q_i \\
&\leq \left(1 + \frac{\sqrt{2}}{\sqrt{2}-1}\right)\exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right)
\end{aligned}$$

$$\cdot\left(1-\exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right)\right)^{-1}$$
$$\overset{(50)}{\leq} 15.1\exp\left(-\frac{c_{1,M,\sigma}(1-4\gamma)\xi}{8}\right).$$

It remains to check that $\{\delta_k\}_{k=0}^{\infty}$ is a decreasing sequence,

$$\sum_{k=1}^{\infty}\eta_k \leq \gamma r_i^2 \tag{51}$$

and

$$\delta_0 \leq \min(r_0\rho, \delta) \tag{52}$$

as claimed in the beginning of the proof. Let's start with verifying that $\{\delta_k\}_{k=0}^{\infty}$ is a decreasing sequence. It is clear that the sequence $\{\delta_k\}_{k=1}^{\infty}$ is decreasing by definition and that

$$\frac{\delta_0}{\delta_1} = 2\exp\left(\frac{c_{1,M,\sigma}(1-4\gamma)\xi}{4}\right) \geq 1$$

by (50), so $\{\delta_k\}_{k=0}^{\infty}$ is decreasing. To verify (51), let $c_2 = 2(c_0+2M)$ and $c_1 = c_{1,M,\sigma}$, and then

$$\begin{aligned}
\eta_1 &= 2c_2A\sqrt{\frac{\xi}{n}}\exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right)\sqrt{\frac{im8\log 2}{n}+\frac{(i+7)c_1(1-4\gamma)\xi}{n}}\\
&\overset{(50)}{\leq} 2c_2A\frac{\xi}{n}\sqrt{c_1(1-4\gamma)}\exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right)\sqrt{3i+9}
\end{aligned}$$

and for $k \geq 2$,

$$\begin{aligned}
\eta_k &= c_2A\sqrt{\frac{\xi}{n}}\exp\left(-\frac{c_1k(1-4\gamma)\xi}{4m}\right)\\
&\quad\cdot\sqrt{\frac{im8\log 2}{n}+\frac{2(2k+1)c_1(1-4\gamma)\xi}{n}+\frac{(i+1)kc_1(1-4\gamma)\xi}{n}}\\
&\overset{(50)}{\leq} c_2A\frac{\xi}{n}\sqrt{c_1(1-4\gamma)}\exp\left(-\frac{c_1k(1-4\gamma)\xi}{4m}\right)\\
&\quad\cdot\sqrt{2(2k+1)+(i+1)(k+2)}\\
&\leq c_2A\frac{\xi}{n}\sqrt{c_1(1-4\gamma)}\exp\left(-\frac{c_1k(1-4\gamma)\xi}{4m}\right)\sqrt{(i+5)(k+2)}\\
&\leq c_2A\frac{\xi}{n}\sqrt{c_1(1-4\gamma)}\exp\left(-\frac{c_1k(1-4\gamma)\xi}{8m}\right)\sqrt{i+5}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
\sum_{k=1}^{\infty} \eta_k &\leq c_2 A \frac{\xi}{n} \sqrt{c_1(1-4\gamma)} \exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right) \sqrt{i+5} \\
&\quad \cdot \left(2\sqrt{3} + \frac{1}{1-\exp\left(-\frac{c_1(1-4\gamma)\xi}{8m}\right)}\right) \\
&\leq c_2 A \frac{\xi}{n} \sqrt{c_1(1-4\gamma)} \exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right) \sqrt{5} 2^i \\
&\quad \cdot \left(2\sqrt{3} + \frac{1}{1-\exp\left(-\frac{c_1(1-4\gamma)\xi}{8m}\right)}\right) \\
&\stackrel{(50)}{\leq} c_2 A \frac{\xi}{n} \sqrt{c_1(1-4\gamma)} \exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right) \sqrt{5} 2^i \left(2\sqrt{3} + \frac{\sqrt{2}}{\sqrt{2}-1}\right) \\
&\leq 15.4 c_2 \sqrt{c_1} A \frac{\sqrt{1-4\gamma}}{\gamma} \exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right) \gamma 2^i \frac{\xi}{n} \\
&= 15.4 c_2 \sqrt{c_1} A \frac{\sqrt{1-4\gamma}}{\gamma} \exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right) \gamma r_i^2.
\end{aligned}
$$

To make (51) holds, it is sufficient to require that

$$
\frac{\xi}{m} \geq \frac{4}{c_1(1-4\gamma)} \log\left(15.4 c_2 \sqrt{c_1} A \frac{\sqrt{1-4\gamma}}{\gamma}\right)
$$

as in the assumption. Now it remains to verify (52). (52) follows from the fact that

$$
\delta_0 = 2A\sqrt{\frac{\xi}{n}} \exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right) \stackrel{(50)}{\leq} \rho_0 \sqrt{\frac{\xi}{n}} = \delta
$$

and that

$$
\frac{\delta_0}{r_0} = 2A \exp\left(-\frac{c_1(1-4\gamma)\xi}{4m}\right) \stackrel{(50)}{\leq} 2A\frac{\rho_0}{2A} \leq \rho.
$$

The proof for Lemma 12 is complete.

## 4.5 A Proof for Lemma 3

Here is a lemma that we will use to verify Assumption 6 to prove Lemma 3.

**Lemma 13** *Suppose that $\{S_\ell\}$ is a collection of linear function space on $[0,1]$. Suppose that for each $\ell$, there is a basis $B_{\ell,1}, ..., B_{\ell,m_\ell}$ for $S_\ell$ such that* *there exist constants $T_1$ and $T_2$ such that for $\theta = (\theta_1, ..., \theta_{m_\ell}) \in R^{m_\ell}$, (19) and (20) hold as in Lemma 7. Let*

$$f_\theta = \theta' B, \tag{53}$$

*and $\Theta_j \subset R^{m_\ell}$, where $j$ depends on $\ell$, then Assumption 6 holds with*

$$A_j = 9.64 \frac{T_1}{T_2} + 0.06 \tag{54}$$

*and $m_j = m_\ell$.*

Proof: straightforward modification of the proof for Lemma 1 by Yang and Barron (1998).

Now let's prove Lemma 3 by verifying the assumptions in Theorem 2. To verify Assumption 6, we will apply Lemma 13. Note that (19) and (20) hold with $T_1 = 1$ and $T_2 = \gamma_q$ by the following fact taken from Schumaker p.145 (4.86),

$$\begin{cases} \|\theta' B\|_\infty \leq \|\theta\|_\infty \\ \|\theta' B\| \geq \frac{\gamma_q}{\sqrt{m_\ell}} \|\theta\| \end{cases}$$

where $\gamma_q = 1/(\sqrt{q}(2q+1)9^{q-1})$, so by Lemma 13, Assumption 6 holds for $A_j$ and $m_j$ in (13).

Now before verifying Assumptions 7 and 3, we need to find $j$ and $\beta$ so that $R_{j,n}(\beta, p_{f_0})$ is small and then compute $\epsilon_n$ so that (4) holds with $p_0 = p_{f_0}$, $p_\theta = p_{f_\theta}$ for $j_n = j$ and $\beta_n = \beta$. Before we start to find such $j_n$ and $\beta_n$, note that for the $C_j$ specified in (13), $\sum_j e^{-C_j} = e^{-2}/(1-e^{-1})^3 < 1$ as required. Now we can start to find $j$ and $\beta$ so that $R_{j,n}(\beta, p_{f_0})$ is small. To control the approximation errors $D(p_{f_0} \| p_{f_\beta})$ and $V(p_{f_0} \| p_{f_\beta})$ in $R_{j,n}(\beta, p_{f_0})$, we use the following fact from Schumaker p.231 (6.50): for $q \geq s+1$, there exists $\beta \in R^{m_j}$ such that

$$\begin{cases} \|D^r(f_0 - f_\beta)\|_\infty \leq \alpha_q \left(\frac{1}{k+1}\right)^{s-r} M_0, & \text{for } 0 \leq r \leq s-1 \\ \|D^s f_\beta\|_\infty \leq \alpha_q M_0, & \end{cases} \tag{55}$$

where $M_0 = \max_{0 \leq r \leq s} \|D^r f_0\|_{L_\infty}$. Note that by (43) and (44), the inequality in (55) for the case $r = 0$ is sufficient to control the errors $D(p_{f_0} \| p_{f_\beta})$ and $V(p_{f_0} \| p_{f_\beta}))$ if $\beta \in \Theta_j$. So now we need to make sure $\max_{0 \leq r \leq q-1} \|D^r f_\beta\|_{L_\infty} \leq L$ and $\|f_\beta\|_\infty \leq M$. For the first condition to hold, we need to make $L$ large enough. Let

$$q = s + 1 \stackrel{\text{def}}{=} q^*$$

and $L = L^*$, the smallest integer such that

$$L \geq 2^s + \alpha_{q^*} M_0 + M_0, \tag{56}$$

then by (55), $\max_{0\leq r\leq q-1} \|D^r \log p_\beta\|_{L_\infty} \leq L^*$. For the second condition to hold, by (55) it sufficies to require that

$$\alpha_{q^*} \left(\frac{1}{k+1}\right)^s M_0 + \|f_0\|_\infty \leq M, \tag{57}$$

which holds for large $k$ since we assume $M > \|f_0\|_\infty$. Now suppose that (57) holds, then $\beta \in \Theta_j$ and by (55) and the definition of $\eta_j$, we can find two constants $c_1$ and $c_2$ which depend only on $s$, $M_0$ and $M$ so that

$$R_{j,n}(\beta, p_{f_0}) \leq c_1 \left(\frac{1}{k+1}\right)^{2s} + c_2 \frac{k}{n}.$$

To make $R_{j,n}$ small, take $k = k_n$ such that $c_3 n^{1/(1+2s)} \leq k_n \leq c_4 n^{1/(1+2s)}$, where $c_3$ and $c_4$ are positive constants, then

$$R_{j,n}(\beta, p_{f_0}) \leq \left(\frac{c_1}{c_3^{2s}} + c_2 c_4\right) n^{-2s/(1+2s)} \leq \epsilon_n^2$$

if we take

$$\epsilon_n^2 = \left(\frac{c_1}{c_3^{2s}} + c_2 c_4\right) n^{-2s/(1+2s)} \stackrel{\text{def}}{=} c_5 n^{-2s/(1+2s)}. \tag{58}$$

Note that since $k = k_n \to \infty$ as $n \to \infty$, (57) holds for large $n$. Therefore, the above argument is valid for large $n$. Now take $j_n = (k_n, q^*, L^*)$, $\beta_n = \beta$ and $\epsilon_n$ as in (58), then (4) holds for large $n$.

Now let's verify Assumptions 7 and 3 with $d_{j_n}(\eta, \theta) = \|f_\eta - f_\theta\|_\infty$ for all $\eta$, $\theta \in \Theta_{j_n}$ for the above $j_n$, $\beta_n$ and $\epsilon_n$. Note that (12) holds with $K_0' = 1$ as required, so Assumption 7 holds. Also, by Theorem IV and Theorem XIV of Kolmogrov and Tihomirov (1961), there exists an $\epsilon_n$-net $F_{\epsilon_n}$ for $\Theta_{j_n}$ with respect to $d_{j_n}$ so that

$$\begin{aligned} \log \operatorname{card}(F_{\epsilon_n}) &\leq c_{q^*,L^*} \left(\frac{1}{\epsilon_n}\right)^{1/(q^*-1)} \\ &= c_{q^*,L^*} \left(\frac{1}{\epsilon_n}\right)^{1/s} \leq c_{q^*,L^*}(k_n+1) \leq c_{q^*,L^*} m_{j_n}. \end{aligned}$$

Therefore, Assumption 3 holds with $K_3 = e^{2L^*}$, $K_4 = e^{c_{q^*,L^*}}$ and $b_1 = 0$.

Now let's verify Assumption 4 for the above $d_{j_n}$, $j_n$, $\beta_n$ and $\epsilon_n$. To bound $\pi_{j_n}(B_{d_{j_n},j_n}(\beta_n,\epsilon_n))$, we will show that

$$\left\{\theta \in R^{m_{j_n}} : \|\theta-\beta_n\|_\infty \le c_6\left(\frac{1}{k_n+1}\right)^s\right\} \subset B_{d_{j_n},j_n}(\beta_n,\epsilon_n), \tag{59}$$

where $c_6 = \min(1, c_3^s\sqrt{c_5}/2)$. For $\|\theta-\beta_n\|_\infty \le c_6(1/(k_n+1))^s$,

$$\|f_\theta - f_{\beta_n}\|_\infty \le \|\theta-\beta_n\|_\infty \le 2\frac{c_3^s\sqrt{c_5}/2}{c_3^s}n^{-s/(1+2s)} = \epsilon_n$$

since $k_n \ge c_3 n^{1/(1+2s)}$. Now it remains to show that $\theta \in \Theta_{k_n,q^*,L^*}$. Note that for $0 \le r < s$,

$$\begin{aligned}
\|D^r f_\theta\|_\infty &= \|D^r\theta' B\|_\infty \\
&\le \|D^r(\theta' B - \beta_n' B)\|_\infty + \|D^r(\beta_n' B - f_0)\|_\infty + \|D^r f_0\|_\infty \\
&\overset{(31),(55)}{\le} 2^r(k_n+1)^r\|\theta-\beta_n\|_\infty + \alpha_{q^*}\left(\frac{1}{k_n+1}\right)^{s-r} M_0 + M_0 \\
&\le \left(\frac{1}{k_n+1}\right)^{s-r}(2^r + \alpha_{q^*}M_0) + M_0 \overset{(56)}{\le} L^*
\end{aligned}$$

and for $r = s$,

$$\begin{aligned}
\|D^s f_\theta\|_{L_\infty} &= \|D^s\theta' B\|_{L_\infty} \\
&\le \|D^s(\theta' B - \beta_n' B)\|_{L_\infty} + \|D^s\beta_n' B\|_{L_\infty} \\
&\overset{(31),(55)}{\le} 2^s + \alpha_{q^*}M_0 \overset{(56)}{\le} L^*.
\end{aligned}$$

Also,

$$\begin{aligned}
\|f_\theta\|_\infty &= \|\theta' B\|_\infty \\
&\le \|\theta' B - \beta_n' B\|_\infty + \|\beta_n' B - f_0\|_\infty + \|f_0\|_\infty \\
&\overset{(31),(55)}{\le} \|\theta-\beta_n\|_\infty + \alpha_{q^*}\left(\frac{1}{k_n+1}\right)^s M_0 + \|f_0\|_\infty \\
&\le \left(\frac{1}{k_n+1}\right)^s(1+\alpha_{q^*}M_0) + \|f_0\|_\infty < M
\end{aligned}$$

for large $n$ since $k_n \to \infty$ and $\|f_0\|_\infty < M$. Therefore, $\theta \in \Theta_{k_n,q^*,L^*}$ and (59) holds. To bound $\pi_{j_n}(B_{d_{j_n},j_n}(\theta_1,\epsilon_n))$ in Assumption 4, note that for all $\epsilon > 0$ and for all $j$,

$$\{\theta \in \Theta_j : \|f_\theta - f_{\theta_1}\|_\infty \le \epsilon\} \subset \{\theta \in \Theta_j : \|\theta-\theta_1\|_\infty \le 2\beta_q^*\epsilon\}, \tag{60}$$

which follows from Lemma 4.2 by Ghosal, Ghosh and Van Der Vaart (1998):

$$\|\theta - \theta_1\|_\infty \leq \beta_q^* \|f_{\theta-\theta_1}\|_\infty,$$

and the fact that $f_{\theta-\theta_1} = f_\theta - f_{\theta_1}$. Then by (60) and by (59), we have

$$\begin{aligned}
\frac{\pi_{j_n}(B_{d_{j_n},j_n}(\beta_n, \epsilon_n))}{\pi_{j_n}(B_{d_{j_n},j_n}(\theta_1, \epsilon_n))} &\geq \frac{\left(c_6 \left(\frac{1}{k_n+1}\right)^s\right)^{k_n+q^*-1}}{(2\beta_{q^*}^* \epsilon_n)^{k_n+q^*-1}} \\
&\geq \left(\frac{c_6}{2\beta_{q^*}^* \epsilon_n \left(1 + \left(\frac{c_4\sqrt{c_5}/2}{\epsilon_n}\right)^{1/s}\right)^s}\right)^{k_n+q^*-1}.
\end{aligned}$$

For $n$ such that $0 < \epsilon_n \leq 1$,

$$\begin{aligned}
\frac{\pi_{j_n}(B_{d_{j_n},j_n}(\beta_n, \epsilon_n))}{\pi_{j_n}(B_{d_{j_n},j_n}(\theta_1, \epsilon_n))} &\geq \left(\frac{c_6}{2\beta_{q^*}^* \epsilon_n \left(\left(\frac{1}{\epsilon_n}\right)^{1/s} + \left(\frac{c_4\sqrt{c_5}/2}{\epsilon_n}\right)^{1/s}\right)^s}\right)^{k_n+q^*-1} \\
&= \left(\frac{c_6}{2\beta_{q^*}^*(1 + (c_4\sqrt{c_5}/2)^{1/s})^s}\right)^{k_n+q^*-1}.
\end{aligned}$$

Without loss of generality, we can assume that $2\beta_{q^*}^* > 1$, so it is clear that Assumption 4 holds with $K_5 = 2\beta_{q^*}^*(1 + (c_4\sqrt{c_5}/2)^{1/s})^s/c_6$ and $b_2 = 0$.

For Assumption 5, it should be clear that it holds with the $\epsilon_n$ specified in (58). Apply Theorem 2 and we have the result in Lemma 3.

## References


Barron, A., Birge, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301-413.

Barron, A., Schervish, M. and Wasserman, L. (1999). The consistency of posterior distributions in non-parametric problems. *Ann. Statist.* **27** 536-561.

Barron, A. and Sheu, C. (1991). Approximating densities by exponentional families. *Ann. Statist.* **19** 1347-1369.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23** 493-507.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1-67.

Doob, J.L. (1948). Application of the theory of martigales. *Coll. Int. du C. N. R. S. Paris.* 22-28

Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (1998). Convergence rates of posterior distributions. *Preprint.*

Kolmogorov, A.N. and Tikhomirov, V.M., (1961). $\epsilon$-entropy and $\epsilon$-capacity of sets in function spaces. *Amer. Math. Soc. Transl. Ser. 2* **17,** 277-364.

Schumaker, L.L. (1981). *Spline Functions.* Wiley, New York.

Shen, X. and Wasserman, L. (1998) Rates of convergence of posterior distributions. Technical Report 678, Dept. Statistics, Carnegie Mellon University.

Stone, C. (1982). Optimal global rates of convergence for non-parametric estimators. *Ann. Statist.* **10** 1040-1053.

Yang, Y. and Barron, A. (1998). An asymptotic property of model selection criteria. *IEEE Trans. Inform. Theory* **44** 95-116.

Department of Statistics
Carnegie Mellon Univerisity
Pittsburgh, PA 15213