Asymptotic minimax risk of predictive density estimation for non-parametric regression

XINYI XU[1] and FENG LIANG[2]

[1]*Department of Statistics, Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247, USA.
E-mail: xinyi@stat.osu.edu*
[2]*Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign,
IL 61820, USA. E-mail: liangf@uiuc.edu*

We consider the problem of estimating the predictive density of future observations from a non-parametric regression model. The density estimators are evaluated under Kullback–Leibler divergence and our focus is on establishing the exact asymptotics of minimax risk in the case of Gaussian errors. We derive the convergence rate and constant for minimax risk among Bayesian predictive densities under Gaussian priors and we show that this minimax risk is asymptotically equivalent to that among all density estimators.

Keywords: asymptotic minimax risk; convergence rate; non-parametric regression; Pinsker's theorem;
predictive density

1. Introduction

Consider the canonical non-parametric regression setup

$$Y(t_i) = f(t_i) + \sigma \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where f is an unknown function in $\mathcal{L}^2[0, 1]$, $t_i = i/n$ and the ε_i's are i.i.d. standard Gaussian random variables. We assume the noise level σ is known and, without loss of generality, set $\sigma = 1$ throughout.

Based on observing $Y = (Y(t_1), \ldots, Y(t_n))$, estimating f or various functionals of f has been the central problem in non-parametric function estimation. The asymptotic optimality of estimators is usually associated with the optimal rate of convergence in terms of minimax risk. A huge body of literature has been devoted to the evaluation of minimax risks under $\mathcal{L}^2$ loss over certain function spaces; see, for example, Pinsker [21], Ibragimov and Has'minskii [16], Golubev and Nussbaum [14], Efroimovich [8], Belitser and Levit [3,4] and Goldenshluger and Tsybakov [13]. An excellent survey of the literature in this area can be found in Efromovich [9].

Sometimes, instead of estimating f itself, one is interested in making statistical inference about future observations from the same process that generated $Y(t)$. A predictive distribution function assigns probabilities to all possible outcomes of a random variable. It thus provides a complete description of the uncertainty associated with a prediction. The minimaxity of predictive density estimators has been studied for finite-dimensional parametric models; see, for

example, Liang and Barron [18], George, Liang and Xu [11], Aslan [2] and George and Xu [12]. However, so far, few results have been obtained on predictive density estimation for non-parametric models. The major thrust of this paper is to establish the asymptotic minimax risk for predictive density estimation under Kullback–Leibler loss in the context of non-parametric regression. Our result closely parallels the well-known work by Pinsker [21] for non-parametric function estimation under $\mathcal{L}^2$ loss and provides a benchmark for studying the optimality of density estimates for non-parametric regression.

Let $\tilde{Y} = (\tilde{Y}(u_1), \ldots, \tilde{Y}(u_m))^t$ denote a vector of future observations from model (1.1) at locations $\{u_j\}_{i=1}^m$. To evaluate the performance of density prediction across the whole curve, we assume that the u_j's are equally spaced dense (that is, $m \geq n$) grids in [0, 1]. Given f, the conditional density $p(\tilde{y}|f)$ is a product of $N(\tilde{y}_j; f(u_j))$, where $N(\cdot; \mu)$ denotes a univariate Gaussian density function with mean μ and unit variance. Based on observing $Y = y$, we estimate $p(\tilde{y}|f)$ by a predictive density $\hat{p}(\tilde{y}|y)$, a non-negative function of $\tilde{y}$ that integrates to 1 with respect to $\tilde{y}$.

Common approaches to constructing $\hat{p}(\tilde{y}|y)$ includes the "plug-in" rule that simply substitutes an estimate $\hat{f}$ for f in $p(\tilde{y}|f)$,

$$p(\tilde{y}|\hat{f}) = \prod_{j=1}^n N(\tilde{y}_j; \hat{f}(u_j)), \tag{1.2}$$

and the Bayes rule that integrates f with respect to a prior π to obtain

$$\int p(\tilde{y}|f)\pi(f|y)\,\mathrm{d}f = \frac{\int p(y|f)p(\tilde{y}|f)\pi(f)\,\mathrm{d}f}{\int p(y|f)\pi(f)\,\mathrm{d}f}. \tag{1.3}$$

We measure the discrepancy between $p(\tilde{y}|f)$ and $\hat{p}(\tilde{y}|y)$ by the average Kullback–Leibler (KL) divergence

$$R(f, \hat{p}) = \frac{1}{m} E_{Y, \tilde{Y}|f} \log \frac{p(\tilde{Y}|f)}{\hat{p}(\tilde{Y}|Y)}. \tag{1.4}$$

Assuming that f belongs to a function space $\mathcal{F}$, such as a Sobolev space, we are interested in the minimax risk

$$R(\mathcal{F}) = \min_{\hat{p}} \max_{f \in \mathcal{F}} R(f, \hat{p}). \tag{1.5}$$

It is worth observing that in this framework, the densities of future observations $(\tilde{Y}_1, \ldots, \tilde{Y}_m)$ are estimated simultaneously by $\hat{p}(\tilde{y}|y)$. An alternative approach is to estimate the densities individually by $\{\hat{p}(\tilde{y}_j|y)\}_{j=1}^m$ with risk

$$\frac{1}{m} \sum_{j=1}^m E_{Y, \tilde{Y}|f} \log \frac{p(\tilde{Y}_j|f(u_j))}{\hat{p}(\tilde{Y}_j|Y)}. \tag{1.6}$$

When the u_j's are equally spaced and m goes to infinity, the risk above converges to

$$\int_0^1 E_{Y,\tilde{Y}|f} \log \frac{p(\tilde{Y}|f(u))}{\hat{p}(\tilde{Y}|Y)} \, du,$$

which can be interpreted as the integrated KL risk of prediction at a random location u in $[0, 1]$. This individual prediction problem can be studied in our simultaneous prediction framework with $\hat{p}(\tilde{y}|y)$ restricted to a product form, that is, $\hat{p}(\tilde{y}|y) = \prod_{j=1}^m \hat{p}(\tilde{y}_j|y)$. For example, the plug-in estimator (1.2) has such a product form and it is easy to check that its individual estimation risk (1.6) is the same as its simultaneous estimation risk (1.4). In general, simultaneous prediction considers a broader class of $\hat{p}$ than the one considered by individual prediction. Therefore, simultaneous prediction is more efficient since the corresponding minimax risk (1.5) is less than or equal to the one with individual prediction. This is distinct from estimating f itself under $\mathcal{L}^2$ loss where, due to the additivity of $\mathcal{L}^2$ loss, simultaneous estimation and individual estimation are equivalent.

This paper is organized as follows. In Section 2, we show that the problem of predictive density estimation for a non-parametric regression model can be converted to the one for a Gaussian sequence model with a constrained parameter space. Direct evaluation of the minimax risk is difficult because of the constraint on the parameter space. Therefore, in Section 3, we first derive the minimax risk over a special class of $\hat{p}$ that consists of predictive densities under Gaussian priors on the unconstrained parameter space $\mathbb{R}^n$. Then, in Section 4, we show that this minimax risk is asymptotically equivalent to the overall minimax risk. Finally, in Section 5, we provide two explicit examples of minimax risks over $\mathcal{L}^2$ balls and Sobolev spaces.

2. Connection to Gaussian sequence models

Let $\{\phi_i\}_{i=1}^\infty$ be the orthonormal trigonometric basis of $\mathcal{L}^2[0, 1]$, that is,

$$\phi_0(t) \equiv 1, \qquad \begin{cases} \phi_{2k-1} = \sqrt{2}\sin(2\pi kx), \\ \phi_{2k} = \sqrt{2}\cos(2\pi kx), \end{cases} \qquad k = 1, 2, \dots.$$

Then, $f = \sum_{i=1}^\infty \theta_i \phi_i$, where $\theta_i = \int_0^1 f(t)\phi_i(t)\,dt$ is the coefficient with respect to the ith basis element ϕ_i. A function space $\mathcal{F}$ corresponds to a constraint on the parameter space of θ. In this paper, we consider function spaces whose parameter spaces Θ have ellipsoid constraints, that is,

$$\Theta(C) = \left\{ \theta : \sum_{i=1}^\infty a_i^2 \theta_i^2 \le C \right\}, \tag{2.1}$$

where $a_1 \le a_2 \le \cdots$ and $a_n \to \infty$.

We approximate f by a finite summation $f_n = \sum_{i=1}^n \theta_i \phi_i$. The bias incurred by estimating $p(\tilde{y}|f_n)$ instead of $p(\tilde{y}|f)$ can be expressed as

$$\text{Bias}(f, f_n) = \frac{1}{m} E_{\tilde{Y}|f} \log \frac{p(\tilde{Y}|f)}{p(\tilde{Y}|f_n)} = \frac{1}{2m} \sum_{j=1}^m [f(u_j) - f_n(u_j)]^2 = \frac{1}{2m} \sum_{i=n+1}^\infty \theta_i^2.$$

This bias is often negligible compared to the prediction risk (1.4); for example, it is of order $O(n^{-2\alpha})$ for Sobolov ellipsoids $\Theta(C, \alpha)$, as defined in (5.3). Therefore, from now on, we set $f = f_n$.

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^t$, Φ_A be a $n \times n$ matrix whose (i, j)th entry equals $\phi_j(t_i)$ and Φ_B be a $m \times n$ matrix whose (i, j)th entry equals $\phi_j(u_i)$. Then, $Y|\theta$ and $\tilde{Y}|\theta$ are two independent Gaussian vectors with $Y|\theta \sim N(\Phi_A\theta, \mathbf{I}_n)$ and $\tilde{Y}|\theta \sim N(\Phi_B\theta, \mathbf{I}_m)$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. Note that since the t_i's and u_j's are equally spaced, we have $\Phi_A^t \Phi_A = n\mathbf{I}_n$ and $\Phi_B^t \Phi_B = m\mathbf{I}_n$. Defining

$$X = \frac{1}{n}\Phi_A^t Y \quad \text{and} \quad \tilde{X} = \frac{1}{m}\Phi_B^t \tilde{Y}, \tag{2.2}$$

it is then easy to check that X and $\tilde{X}$ are independent and that

$$X|\theta \sim N(\theta, v_n\mathbf{I}_n) \quad \text{and} \quad \tilde{X}|\theta \sim N(\theta, v_m\mathbf{I}_n), \tag{2.3}$$

where $v_n = 1/n$ and $v_m = 1/m$. We refer to the model above as a *Gaussian sequence model* since its number of parameters is increasing at the same rate as the number of data points.

Consider the problem of predictive density estimation for the Gaussian sequence model (2.3). Let $\hat{p}(\tilde{x}|x)$ denote a predictive density function of $\tilde{x}$ given $X = x$. The incurred KL risk is defined to be

$$R(\theta, \hat{p}) = \frac{1}{m} E_{X, \tilde{X}|\theta} \log \frac{p(\tilde{X}|\theta)}{\hat{p}(\tilde{X}|X)}$$

and the corresponding minimax risk is given by

$$R(\Theta) = \inf_{\hat{p}} \sup_{\theta \in \Theta(C)} R(\theta, \hat{p}). \tag{2.4}$$

The following theorem states that the two minimax risks, the one associated with $(Y, \tilde{Y})$ from a non-parametric regression model and the one associated with $(X, \tilde{X})$ from a normal sequence model, are equivalent.

Theorem 2.1. $R(\mathcal{F}) = R(\Theta)$, *where $R(\mathcal{F})$ is defined in* (1.5) *and $R(\Theta)$ in* (2.4).

Proof. See the Appendix. $\square$

Remark. The idea of reducing a non-parametric regression model to a Gaussian sequence model via an orthonormal function basis has been widely used for non-parametric function estimation. Early references include Ibraginov and Has'minskii [15], Efromovich and Pinsker [10] and references therein. For recent developments, see Brown and Low [6], Nussbaum [19,20] and Johnstone [17]. Our proof of Theorem 2.1, given in the Appendix, implies that *simultaneous* estimation of predictive densities in these two models are equivalent. However, this equivalence does not hold for the *individual* estimation approach described in Section 1 because the product form of the density estimators, that is, $\hat{p}(\tilde{y}|y) = \prod_j \hat{p}(\tilde{y}_j|y)$, is not retained under the transformation.

3. *Linear* minimax risk

Direct evaluation of the minimax risk (2.4) is difficult because the parameter space $\Theta(C)$ is constrained. In this section, we first consider a subclass of density estimators that have simple forms and investigate the minimax risk over this subclass. In next section, we then show that the minimax risk over this subclass is asymptotically equivalent to the overall minimax risk R. Such an approach was first used in Pinsker [21] to establish a minimax risk bound for the function estimation problem. It inspired a series of developments, including Belitser and Levit [3,4], Tsybakov [22] and Goldenshluger and Tsybakov [13].

Recall that in the problem of estimating the mean of a Gaussian sequence model under $\mathcal{L}^2$ loss, diagonal linear estimators of the form $\hat{\theta}_i = c_i x_i$ play an important role. Indeed, Pinsker [21] showed that when the parameter space (2.1) is an ellipsoid, the minimax risk among diagonal linear estimators is asymptotically minimax among all estimators. Moreover, the results in Diaconis and Ylvisaker [7] imply that if such a diagonal linear estimator is Bayes, then the prior π must be a Gaussian prior with a diagonal covariance matrix. Similarly, in investigating the minimax risk of predictive density estimation, we first restrict our attention to a special class of $\hat{p}$ that are Bayes rules under Gaussian priors over the unconstrained parameter space $\mathbb{R}^n$. Due to the above connection, we call these predictive densities *linear* predictive densities and call the minimax risk over this class the *linear* minimax risk, even though 'linear' does not have any literal meaning in our setting.

Under a Gaussian prior $\pi_S(\theta) = N(0, S)$, where $S = \text{diag}(s_1, \ldots, s_n)$ and $s_i \geq 0$ for $i = 1, \ldots, n$, the linear predictive density $\hat{p}_S$ is given by

$$\hat{p}_S(\tilde{x}|x) = \int_{\mathbb{R}^n} p(\tilde{x}|\theta)\pi_S(\theta|x) \, d\theta = \frac{\int_{\mathbb{R}^n} p(x|\theta)p(\tilde{x}|\theta)\pi_S(\theta) \, d\theta}{\int_{\mathbb{R}^n} p(x|\theta)\pi_S(\theta) \, d\theta}. \tag{3.1}$$

Note that $\hat{p}_S$ is not a Bayes estimator for the problem described in Section 2 because the prior distribution $N(0, S)$ is supported on $\mathbb{R}^n$ instead of on the ellipsoidal space Θ. Nonetheless, $\hat{p}_S$ is a valid predictive density function.

The following lemma provides an explicit form of the average KL risk of $\hat{p}_S$.

Lemma 3.1. *The average Kullback–Leibler risk* (1.4) *of $\hat{p}_S$ is given by*

$$R(\theta, \hat{p}_S) = \frac{n}{2m} \log \frac{v_n}{v_{n+m}} + \frac{1}{2m} \sum_{i=1}^{n} \left[\log \frac{v_{n+m} + s_i}{v_n + s_i} + \frac{v_{n+m} + \theta_i^2}{v_{n+m} + s_i} - \frac{v_n + \theta_i^2}{v_n + s_i} \right], \tag{3.2}$$

where $v_{n+m} = 1/(n+m)$.

Proof. Let $\hat{p}_U$ denote the posterior predictive density under the uniform prior $\pi_U \equiv 1$, namely,

$$\hat{p}_U(\tilde{x}|x) = \left(\frac{1}{2\pi v_{n+m}} \right)^{n/2} \exp\left(-\frac{\|\tilde{x} - x\|^2}{2v_{n+m}} \right).$$

Then, by [11], Lemma 2, the average KL risk of $\hat{p}_S$ is given by

$$R(\theta, \hat{p}_S) = R(\theta, \hat{p}_U) - \frac{1}{m} E \log m_S(W; v_{n+m}) + \frac{1}{m} E \log m_S(X; v_n), \tag{3.3}$$

where

$$W = \frac{v_m X + v_n \tilde{X}}{v_{n+m}} \sim N(\theta, v_{n+m} I)$$

and $m_S(x; \sigma^2)$ denotes the marginal distribution of $X|\theta \sim N_n(\theta, \sigma^2 I)$ under the normal prior π_S. It is easy to check that

$$R(\theta, \hat{p}_U) = \frac{1}{m} E \log \frac{p(\tilde{x}|\theta)}{\hat{p}_U(\tilde{x}|x)} = \frac{n}{2m} \log \frac{v_n}{v_{n+m}} \tag{3.4}$$

and

$$E \log m_S(W; v_{n+m}) = -\frac{n}{2m} \sum_{i=1}^n \log[2\pi(v_{n+m} + s_i)] - \frac{1}{2m} \sum_{i=1}^n \frac{v_{n+m} + \theta_i^2}{v_{n+m} + s_i}, \tag{3.5}$$

$$E \log m_S(X; v_n) = -\frac{n}{2m} \sum_{i=1}^n \log[2\pi(v_n + s_i)] - \frac{1}{2m} \sum_{i=1}^n \frac{v_n + \theta_i^2}{v_n + s_i}. \tag{3.6}$$

The lemma then follows immediately by combining equations (3.3)–(3.6). $\qquad\square$

We denote the linear minimax risk over all $\hat{p}_S$ by $R_L(\Theta)$, that is,

$$R_L(\Theta) = \inf_S \sup_{\theta \in \Theta(C)} R(\theta, \hat{p}_S). \tag{3.7}$$

This linear minimax risk is not directly tractable because the inside maximization is over a constrained space $\Theta(C)$. In the following theorem, we first show that we can switch the order of inf and sup in equation (3.7) and then evaluate R_L using the Lagrange multiplier method.

The following notation will be useful throughout. Let $\tilde{\lambda}(C, v_n, v_{n+m})$ denote a solution of the equation

$$\sum_{i=1}^n a_i^2 \left[(v_n - v_{n+m}) \sqrt{1 + \frac{4\tilde{\lambda}/a_i^2}{v_n - v_{n+m}}} - (v_n + v_{n+m}) \right]_+ = 2C, \tag{3.8}$$

where $[x]_+ = \sup(x, 0)$, and let $\tilde{\theta}_i^2$ be

$$\tilde{\theta}_i^2 = \frac{1}{2} \left[(v_n - v_{n+m}) \sqrt{1 + \frac{4\tilde{\lambda}/a_i^2}{v_n - v_{n+m}}} - (v_n + v_{n+m}) \right]_+ \tag{3.9}$$

for $i = 1, 2, \ldots, n$.

Theorem 3.2. *Suppose that the parameter space $\Theta(C)$ is an ellipsoid, as defined in* (2.1). *The linear minimax risk is then given by*

$$R_L(\Theta) = \inf_S \sup_{\theta \in \Theta(C)} R(\theta, \hat{p}_S) = \sup_{\theta \in \Theta(C)} \inf_S R(\theta, \hat{p}_S) \tag{3.10}$$

$$= \frac{n}{2m} \log \frac{v_n}{v_{n+m}} + \frac{1}{2m} \sum_{i=1}^{n} \log \frac{v_{n+m} + \tilde{\theta}_i^2}{v_n + \tilde{\theta}_i^2}, \tag{3.11}$$

where $\tilde{\theta}_i^2$ is defined as in (3.9). *The linear minimax estimator $\hat{p}_{\tilde{V}}$ is the Bayes predictive density under a Gaussian prior*

$$\pi_{\tilde{V}}(\theta) = N(0, \tilde{V}), \qquad where \ \tilde{V} = \mathrm{diag}(\tilde{\theta}_1^2, \tilde{\theta}_2^2, \dots, \tilde{\theta}_n^2), \tag{3.12}$$

namely,

$$\hat{p}_{\tilde{V}}(\tilde{x}|x) = N(\theta_{\tilde{V}}, \Sigma_{\tilde{V}}),$$

with

$$\theta_{\tilde{V}} = \left(\frac{\tilde{\theta}_1^2}{\tilde{\theta}_1^2 + v_n} x_1, \dots, \frac{\tilde{\theta}_n^2}{\tilde{\theta}_n^2 + v_n} x_n \right)',$$

$$\Sigma_{\tilde{V}} = \mathrm{diag}\left(\frac{\tilde{\theta}_1^2 v_n}{\tilde{\theta}_1^2 + v_n} + v_m, \dots, \frac{\tilde{\theta}_n^2 v_n}{\tilde{\theta}_n^2 + v_n} + v_m \right).$$

Proof. We first prove equality (3.11). It is easy to check that for any fixed θ, $R(\theta, \hat{p}_S)$ achieves its minimum at $S = \mathrm{diag}(\theta_1^2, \dots, \theta_n^2)$, and

$$\inf_S R(\theta, \hat{p}_S) = \frac{n}{2m} \log \frac{v_{n+m}}{v_n} + \frac{1}{2m} \sum_{i=1}^{n} \log \frac{v_{n+m} + \theta_i^2}{v_n + \theta_i^2}.$$

To calculate the maximum of the above quantity over $\theta \in \Theta(C)$, one needs to solve

$$\sup \left\{ \sum_{i=1}^{n} \log \frac{v_{n+m} + \theta_i^2}{v_n + \theta_i^2} : \sum_{i=1}^{n} a_i^2 \theta_i^2 \le C \right\}.$$

With the Lagrangian

$$\mathcal{L} = \sum_{i=1}^{n} \log \frac{v_{n+m} + \theta_i^2}{v_n + \theta_i^2} - \frac{1}{\lambda} \left(\sum_{i=1}^{n} a_i^2 \theta_i^2 - C \right),$$

simple calculation reveals that the maximum is attained at $\tilde{\theta}_i$ given by (3.9).

Next, we prove equality (3.10), that is, that the order of inf and sup can be exchanged. Note that for any diagonal matrix $\tilde{S}$, we have

$$\sup_{\theta \in \Theta(C)} R(\hat{p}_{\tilde{S}}, \theta) \geq \inf_S \sup_{\theta \in \Theta(C)} R(\hat{p}_S, \theta) \geq \sup_{\theta \in \Theta(C)} \inf_S R(\hat{p}_S, \theta). \tag{3.13}$$

Therefore, if there exists an $\tilde{S}$ such that

$$\sup_{\theta \in \Theta(C)} R(\hat{p}_{\tilde{S}}, \theta) - \sup_{\theta \in \Theta(C)} \inf_S R(\hat{p}_S, \theta) \leq 0,$$

then all of the inequalities in (3.13) become equalities.

If we let $\tilde{S} = \operatorname{diag}(\tilde{\theta}_1^2, \ldots, \tilde{\theta}_n^2)$, then

$$R(\hat{p}_{\tilde{S}}, \theta) - \sup_{\theta \in \Theta(C)} \inf_S R(\hat{p}_S, \theta) = \frac{1}{2m} \sum_{i=1}^n \frac{(v_n - v_{n+m})(\theta_i^2 - \tilde{\theta}_i^2)}{(v_n + \tilde{\theta}_i^2)(v_{n+m} + \tilde{\theta}_i^2)}$$

$$= \frac{1}{2m} \frac{\sum_{i=1}^n a_i^2 \theta_i^2 - C}{\tilde{\lambda}},$$

where the second equality holds because $\sum_{i=1}^n a_i^2 \tilde{\theta}_i^2 = C$ and $\tilde{\theta}_i^2$ is a solution to

$$\frac{\partial \mathcal{L}}{\partial \theta_i^2} = \frac{v_n - v_{n+m}}{(v_n + \theta_i^2)(v_{n+m} + \theta_i^2)} - \frac{a_i^2}{\tilde{\lambda}} = 0.$$

Since $\theta \in \Theta(C)$ implies that $\sum_{i=1}^n a_i^2 \theta_i^2 \leq C$, we have

$$\sup_{\theta \in \Theta(C)} R(\hat{p}_{\tilde{S}}, \theta) - \sup_{\theta \in \Theta(C)} \inf_S R(\hat{p}_S, \theta) \leq \frac{1}{2m} \frac{C - C}{\tilde{\lambda}} = 0,$$

which completes the proof. $\square$

Remark. Note that $a_1 \leq a_2 \leq \cdots$, so we have $\tilde{\theta}_i^2 = 0$ for $i > N$, where

$$N = \sup\left\{ i : a_i^2 \leq \tilde{\lambda}\left(\frac{1}{v_{m+n}} - \frac{1}{v_n} \right) = m\tilde{\lambda} \right\}. \tag{3.14}$$

This implies that the prior distribution corresponding to the linear minimax estimator, that is, $\pi_{\tilde{V}}(\theta) = \prod_{i=1}^n N(0, \tilde{\theta}_i^2)$, puts a point mass at zero for θ_i for all $i > N$.

4. Asymptotic minimax risk

In this section, we turn to establishing the asymptotic behavior of the minimax risk $R(\Theta)$ over all predictive density estimators. By definition, $R(\Theta) \leq R_L(\Theta)$. We extend the approach in [3] to show that the difference between $R(\Theta)$ and $R_L(\Theta)$ vanishes as the number of observations n

goes to infinity. Therefore, the overall minimax risk is asymptotically equivalent to the linear minimax risk. This also implies that the Gaussian prior $\pi_{\tilde{V}}$ defined in (3.12) is asymptotically least favorable.

The following lemma provides a lower bound for the overall minimax risk $R(\Theta)$ under some conditions.

Lemma 4.1. *Let $\{s_i^2\}_{i=1}^n$ be a sequence such that for some $\alpha > 0$,*

$$\sum_{i=1}^n a_i^2 s_i^2 + \left[-8\alpha \left(\sum_{i=1}^n a_i^4 s_i^4 \right) \log v_n \right]^{1/2} \le C. \tag{4.1}$$

Then, as $n \to \infty$, the minimax risk $R(\Theta)$ has the following lower bound:

$$R(\Theta) \ge \frac{n}{2m} \log \frac{v_n}{v_{n+m}} + \frac{1}{2m} \sum_{i=1}^n \log \frac{v_{n+m} + s_i^2}{v_n + s_i^2} + O(v_n^\alpha).$$

Proof. See the Appendix. □

Note that, as shown in the proof, for a posterior density with a Gaussian prior $\pi_S = N(0, S)$, where $S = \text{diag}(s_1, \ldots, s_n)$, condition (4.1) guarantees π_S to have most of its mass inside Θ, in the sense that $\pi_S(\Theta^c) \le v_n^{2\alpha}$ for some $\alpha > 0$.

With the lower bound in the above lemma, we are ready to prove the main result in this paper, which shows that the overall minimax risk $R(\Theta)$ is asymptotically equivalent to the linear minimax risk $R_L(\Theta)$.

Theorem 4.2. *Suppose that Θ is the ellipsoid defined in (2.1) and $\tilde{\theta}^2$ is defined in (3.9). If $m = O(n)$ and*

$$\log(1/v_n) \sum_{i=1}^n a_i^4 \tilde{\theta}_i^4 = o(1), \qquad as\ v_n \to 0, \tag{4.2}$$

then

$$\lim_{v_n \to 0} \frac{R(\Theta)}{R_L(\Theta)} = 1. \tag{4.3}$$

Proof. By definition, $R(\Theta) \le R_L(\Theta)$. So, to prove this theorem, it suffices to show that as $v_n \to 0$,

$$R(\Theta) \ge R_L(\Theta)\big(1 - o(1)\big).$$

For a fixed constant $\alpha > 1$, let $\gamma = \frac{1}{C}[8\alpha \log(1/v_n) \sum_{i=1}^n a_i^4 \tilde{\theta}_i^4]^{1/2}$ and let $b_i^2 = \tilde{\theta}_i^2(1 + \gamma)^{-1}$ for $i = 1, \ldots, n$. It is easy to check that the sequence $\{b_i\}_{i=1}^n$ satisfies the condition (4.1). There-

fore, by Theorem 4.1,

$$
R(\Theta) \geq \frac{n}{2m} \log \frac{v_n}{v_{n+m}} + \frac{1}{2m} \sum_{i=1}^{n} \log \frac{v_{n+m} + b_i^2}{v_n + b_i^2} + \mathrm{O}(v_n^\alpha)
$$

$$
= R_L(\Theta) - \frac{1}{2m} \sum_{i=1}^{n} \log \frac{(v_n + b_i^2)(v_{n+m} + \tilde{\theta}_i^2)}{(v_{n+m} + b_i^2)(v_n + \tilde{\theta}_i^2)} + \mathrm{O}(v_n^\alpha) \qquad \text{as } v_n \to 0.
$$

(4.4)

Next, we will derive the convergence rate of $R_L(\Theta)$ and show that the other terms are of smaller order.

Using the fact that $\tilde{\theta}_i^2 = 0$ for $i > N$ (see (3.14)), we can rewrite $R_L(\Theta)$ as

$$
R_L(\Theta) = \frac{n}{2m} \log \frac{v_n}{v_{n+m}} + \frac{1}{2m} \sum_{i=1}^{N} \log \frac{v_{n+m} + \tilde{\theta}_i^2}{v_n + \tilde{\theta}_i^2} + \frac{1}{2m} \sum_{i=N}^{n} \log \frac{v_{n+m}}{v_n}
$$

$$
= \frac{1}{2m} \sum_{i=1}^{N} \log \frac{(v_{n+m} + \tilde{\theta}_i^2) v_n}{(v_n + \tilde{\theta}_i^2) v_{n+m}}
$$

$$
= \frac{1}{2m} \sum_{i=1}^{N} \log \left(1 + \frac{(v_n - v_{n+m}) \tilde{\theta}_i^2}{(v_n + \tilde{\theta}_i^2) v_{n+m}} \right).
$$

When $m = \mathrm{O}(n)$, we have $v_n - v_{n+m} = \mathrm{O}(v_n)$ and $v_n + v_{n+m} = \mathrm{O}(v_n)$. Therefore, by means of a Taylor expansion,

$$
R_L = \mathrm{O} \left(\frac{1}{2m} \sum_{i=1}^{N} \frac{(v_n - v_{n+m}) \tilde{\theta}_i^2}{(v_n + \tilde{\theta}_i^2) v_{n+m}} \right) \geq \mathrm{O} \left(\frac{1}{m} \right).
$$

(4.5)

Similarly, since $b_i = \tilde{\theta}_i^2 = 0$ for $i > N$, the second term in (4.4) can be written as

$$
\frac{1}{2m} \sum_{i=1}^{n} \log \frac{(v_n + b_i^2)(v_{n+m} + \tilde{\theta}_i^2)}{(v_{n+m} + b_i^2)(v_n + \tilde{\theta}_i^2)} = \frac{1}{2m} \sum_{i=1}^{N} \log \frac{(v_n + b_i^2)(v_{n+m} + \tilde{\theta}_i^2)}{(v_{n+m} + b_i^2)(v_n + \tilde{\theta}_i^2)}.
$$

For every $1 \leq i \leq N$, we have

$$
\log \frac{(v_n + b_i^2)(v_{n+m} + \tilde{\theta}_i^2)}{(v_{n+m} + b_i^2)(v_n + \tilde{\theta}_i^2)} = \log \left(\frac{[(1 + \gamma) v_n + \tilde{\theta}_i^2](v_{n+m} + \tilde{\theta}_i^2)}{[(1 + \gamma) v_{n+m} + \tilde{\theta}_i^2](v_n + \tilde{\theta}_i^2)} \right)
$$

$$
= \log \left(1 + \gamma \frac{(v_n - v_{n+m}) \tilde{\theta}_i^2}{(v_n + \tilde{\theta}_i^2)(v_{n+m} + \tilde{\theta}_i^2) + \gamma v_n (v_{n+m} + \tilde{\theta}_i^2)} \right)
$$

$$
\leq \log \left(1 + \gamma \frac{(v_n - v_{n+m}) \tilde{\theta}_i^2}{(v_n + \tilde{\theta}_i^2) v_{n+m}} \right).
$$

Again using a Taylor expansion, as well as the condition that $\gamma = o(1)$, we obtain

$$\frac{1}{2m} \sum_{i=1}^{n} \log\left(1 + \gamma \frac{(v_n - v_{n+m})\tilde{\theta}_i^2}{(v_n + \tilde{\theta}_i^2)v_{n+m}}\right) = O\left(\frac{\gamma}{2m} \sum_{i=1}^{n} \frac{(v_n - v_{n+m})\tilde{\theta}_i^2}{(v_n + \tilde{\theta}_i^2)v_{n+m}}\right) = o(R_L). \qquad (4.6)$$

Finally, since $m = O(n)$, by choosing $\alpha > 1$, the last term in (4.4) satisfies

$$v_n^{\alpha} = o(1). \qquad (4.7)$$

Combining (4.4)–(4.7), the theorem then follows. $\qquad\qquad\square$

5. Examples

In this section, we apply Theorems 3.2 and 4.2 to establish asymptotic behaviors of minimax risks over some constrained parameter spaces. In particular, we consider the asymptotics over $\mathcal{L}^2$ balls and Sobolev ellipsoids.

Example 1. Suppose that $m = n$ and θ is restricted in an $\mathcal{L}^2$ ball,

$$\Theta(C) = \left\{\theta : \sum_{i=1}^{n} \theta_i^2 \le C\right\}. \qquad (5.1)$$

The $\mathcal{L}^2$ ball can be considered as a variant of the ellipsoid (2.1) with $a_1 = a_2 = \cdots = a_n = 1$ and $a_{n+1} = a_{n+2} = \cdots = \infty$. Although the values of the a_i's here depend on n, the proofs of the above theorems are still valid. It is easy to see that N defined in (3.14) is equal to n and that $\tilde{\theta}_1^2 = \tilde{\theta}_2^2 = \cdots = \tilde{\theta}_n^2 = C/n$. Therefore,

$$(\log n) \sum_{i=1}^{n} a_i^4 \tilde{\theta}_i^4 = (\log n) \cdot \frac{C^2}{n} = o(1).$$

By Theorem 4.2, the minimax risk among all predictive density estimators is asymptotically equivalent to the minimax risk among *linear* density estimators. Furthermore, by Theorem 3.2,

$$\lim_{n\to\infty} R(\Theta(C)) = \lim_{n\to\infty} R_L(\Theta(C)) = \frac{1}{2}\log 2 + \frac{1}{2}\log \frac{1/(2n) + C/n}{1/n + C/n} = \frac{1}{2}\log\frac{1+2C}{1+C}.$$

Note that this minimax risk is strictly smaller than the minimax risk over the class of plug-in estimators since, for any plug-in density $\hat{p}(\tilde{x}|\hat{\theta})$,

$$R(\theta, \hat{p}) = \frac{1}{n}E\log\frac{p(\tilde{x}|\theta)}{p(\tilde{x}|\hat{\theta})} = \frac{1}{n}E\left[-\frac{\|x - \theta\|^2 - \|x - \hat{\theta}\|^2}{2/n}\right] = \frac{1}{2}E\|\hat{\theta} - \theta\|^2 \qquad (5.2)$$

and by Pinsker's theorem, the minimax risk of estimating θ under squared error loss is $C/(1+C)$, which is larger than $\log\frac{1+2C}{1+C}$, by the fact that $x > \log(1 + x)$ for any $x > 0$.

Example 2. Suppose that $m = n$ and θ is restricted in a Sobolev ellipsoid

$$\Theta(C, \alpha) = \left\{ \theta : \sum_{i=1}^{\infty} a_i^2 \theta_i^2 \le C \right\}, \tag{5.3}$$

where $a_{2i} = a_{2i-1} = (2i)^{\alpha} (\alpha > 0)$ for $i = 1, 2, \ldots$. Then, by (3.14), we have $a_N^2/\tilde{\lambda}n \sim N^{2\alpha}/\tilde{\lambda}n \to 1$ as $n \to \infty$. Substituting this relation into equation (3.8) yields

$$2C \sim \sum_{i=1}^{N} i^{2\alpha} \left(\frac{1}{2n} \sqrt{1 + 8\tilde{\lambda}ni^{-2\alpha}} - \frac{3}{2n} \right)$$

$$= \frac{1}{2n} \sum_{i=1}^{N} i^{2\alpha} \left(\sqrt{1 + 8N^{2\alpha}i^{-2\alpha}} - 3 \right)(1 + o(1)).$$

Using the Taylor expression

$$\sqrt{1 + 8N^{2\alpha}i^{-2\alpha}} = \sum_{k=0}^{\infty} \frac{2\sqrt{2}(-1)^k (2k)!}{(1 - 2k)k!^2 32^k} \left(\frac{i}{N} \right)^{(2k-1)\alpha}$$

and the asymptotic relation

$$\sum_{i=1}^{N} i^r = \frac{N^{r+1}}{r+1} (1 + o(1)) \qquad \text{as } N \to \infty, r > -1,$$

we obtain

$$N = Mn^{1/(2\alpha+1)}(1 + o(1)) \quad \text{and} \quad \tilde{\lambda} = Mn^{-2\alpha/(2\alpha+1)}(1 + o(1)),$$

where

$$M = \left[4C \Big/ \left(\sum_{k=0}^{\infty} \frac{2\sqrt{2}(-1)^k (2k)!}{(1 - 2k)k!^2 32^k} \cdot \frac{1}{(2k+1)\alpha + 1} - \frac{3}{2\alpha + 1} \right) \right]^{1/(2\alpha+1)}.$$

Note that, by (3.9),

$$\tilde{\theta}_i^2 = \frac{1}{2} \left[\frac{1}{2n} \sqrt{1 + 8 \left(\frac{N}{i} \right)^{2\alpha}} - \frac{3}{2n} \right]_+ (1 + o(1)).$$

Therefore,

$$(\log n) \sum_{i=1}^{N} a_i^4 \tilde{\theta}_i^4 = O\left((\log n) \cdot \frac{N^{4\alpha+1}}{n^2} \right) = O\left((\log n) \cdot n^{-1/(2\alpha+1)} \right) = o(1).$$

By Theorem 4.2, the minimax risk among all predictive density estimators is asymptotically equivalent to the minimax risk among the *linear* density estimators. Furthermore, by Theorem 3.2,

$$R_L(\Theta(C,\alpha)) = \frac{1}{2}\log 2 + \frac{1}{2n}\sum_{i=1}^{N}\log\frac{1/(2n)+\tilde{\theta}_i^2}{1/n+\tilde{\theta}_i^2} + \frac{n-N}{2n}\log\frac{1}{2}$$

$$= \frac{1}{2n}\sum_{i=1}^{N}\log\frac{1/n+2\tilde{\theta}_i^2}{1/n+\tilde{\theta}_i^2}$$

$$= \frac{1}{2n}\sum_{i=1}^{N}\log\left(1+\frac{\tilde{\theta}_i^2}{1/n+\tilde{\theta}_i^2}\right).$$

It is difficult to calculate an explicit form of the optimal constant for the minimax risk due to the log function, but we can get an accurate bound for it. By Taylor expansion, there exists $x_i^* \in (0,1), i = 1, 2, \ldots, N$, such that

$$R_L(\Theta(C,\alpha)) = \frac{1}{2n}\sum_{i=1}^{N}\left(\frac{1}{1+x_i^*}\frac{\tilde{\theta}_i^2}{1/n+\tilde{\theta}_i^2}\right) \in \left(\frac{1}{4n}\sum_{i=1}^{N}\frac{\tilde{\theta}_i^2}{1/n+\tilde{\theta}_i^2}, \frac{1}{2n}\sum_{i=1}^{N}\frac{\tilde{\theta}_i^2}{1/n+\tilde{\theta}_i^2}\right).$$

Moreover,

$$\sum_{i=1}^{N}\frac{\tilde{\theta}_i^2}{1/n+\tilde{\theta}_i^2} = \sum_{i=1}^{N}\frac{1/(4n)\sqrt{1+8(N/i)^{2\alpha}}-3/(4n)}{1/n+1/(4n)\sqrt{1+8(N/i)^{2\alpha}}-3/(4n)} = K \cdot N,$$

where

$$K = 1 + \frac{1}{2(2\alpha+1)} - \frac{1}{2}\sum_{k=0}^{\infty}\frac{2\sqrt{2}(-1)^k(2k)!}{(1-2k)k!^2 32^k}\cdot\frac{1}{(2k+1)\alpha+1}.$$

Therefore,

$$\lim_{n\to\infty} n^{2\alpha/(2\alpha+1)}R(\Theta(C,\alpha)) = \lim_{n\to\infty} n^{2\alpha/(2\alpha+1)}R_L(\Theta(C,\alpha)) \in \left(\tfrac{1}{4}KM, \tfrac{1}{2}KM\right), \quad (5.4)$$

that is, the convergence rate is $n^{-2\alpha/(2\alpha+1)}$ and the convergence constant is between $\frac{1}{4}KM$ and $\frac{1}{2}KM$.

As in Example 1, we compare the asymptotics of this minimax risk with the one over the class of plug-in estimators, where the latter can be easily computed by (5.2) and the results in [21]. Direct comparison reveals that the convergence rates of both minimax risks are $n^{2\alpha/(2\alpha+1)}$ and the convergence constants can both be written in the form $C^{1/(2\alpha+1)}f(\alpha)$, where $f(\alpha)$ is a function depending only on α. Although it is hard to obtain an explicit representation for the convergence constant for the overall minimax risk, our simulation result in Figure 1 shows that it is strictly smaller than that over the class of plug-in estimators.
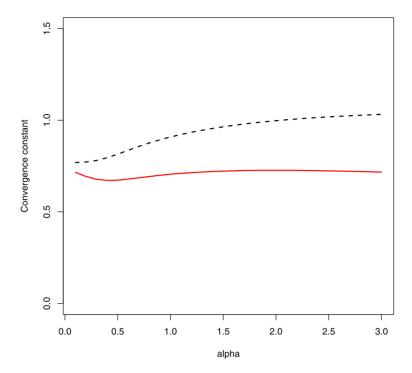
Figure 1. Convergence constants of the overall minimax risk (the lower red line) and the minimax risk over the class of plug-in estimators (the upper red line). Here, the sample size $n = 10\,000\,000$ and $C = 1$.

Appendix: Proofs

In this appendix, we provide the proofs of Theorem 2.1 and Lemma 4.1.

Proof of Theorem 2.1. Let Ψ be an $m \times m$ matrix whose (i, j)th entry equals $\phi_j(u_i)$. Since the ϕ_j's form an orthogonal basis for $\mathcal{L}^2$ and the u_i's are equally spaced, we have $\Psi^t \Psi = \mathbf{I}_m$. Consider the transformation $\frac{1}{m} \Psi^t \tilde{Y}$. Since the first n columns of Ψ are Φ_B, the first n elements of the transformed vector are just $\tilde{X}$, defined in (2.2), and we denote the remaining $(m - n)$ elements by $\tilde{Z}$. It is easy to check that $\tilde{X}|\theta \sim N_n(\theta, \frac{1}{m}\mathbf{I}_n)$ and $\tilde{Z} \sim N_{m-n}(\mathbf{0}, \frac{1}{m}\mathbf{I}_{m-n})$ are independent multivariate Gaussian variables, and the target density function $p(\tilde{y}|f)$ satisfies

$$p(\tilde{y}|f) = p(\tilde{x}, \tilde{z}|\theta) J_{\tilde{x}, \tilde{z}}(\tilde{y}), \tag{A.1}$$

where $J_{\tilde{x}, \tilde{z}}(\tilde{y})$ is the Jacobian for this transformation. Similarly, any predictor density estimator $\hat{p}(\tilde{y}|y)$ can be rewritten as

$$\hat{p}(\tilde{y}|y) = \hat{p}(\tilde{x}, \tilde{z}|x) J_{\tilde{x}, \tilde{z}}(\tilde{y}), \tag{A.2}$$

where X is a transformation of Y defined in (2.2). Note that the two predictive density functions on the left and right sides of the above equation may have different functional forms; however, to simplify the notation, we use the same symbol $\hat{p}$ to represent them when the context is clear.

Now, the average KL risk can be represented as

$$
\begin{aligned}
R(f, \hat{p}) &= E_{Y, \tilde{Y}|f} \log \frac{p(\tilde{Y}|f)}{\hat{p}(\tilde{Y}|Y)} \\
&= E_{X, \tilde{X}, \tilde{Z}|\theta} \log \frac{p(\tilde{X}, \tilde{Z}|\theta)}{\hat{p}(\tilde{X}, \tilde{Z}|X)},
\end{aligned}
\tag{A.3}
$$

where the second equality follows from (A.1) and (A.2). Since $\tilde{X}$ and $\tilde{Z}$ are independent, we can split $p(\tilde{x}, \tilde{z}|\theta)$ as

$$
p(\tilde{x}, \tilde{z}|\theta) = p(\tilde{x}|\theta) p(\tilde{z}),
\tag{A.4}
$$

where $p(\tilde{z})$ has a known distribution $N_{m-n}(\mathbf{0}, I_{m-n})$ Moreover, to evaluate the minimax risk, it suffices to consider predictive density estimators in the form

$$
\hat{p}(\tilde{x}, \tilde{z}|x) = \hat{p}(\tilde{x}|x) p(\tilde{z})
\tag{A.5}
$$

because any predictive density $\hat{p}(\tilde{x}, \tilde{z}|x)$ can be written as $\hat{p}(\tilde{x}, \tilde{z}|x) = \hat{p}(\tilde{x}|x)\hat{p}(\tilde{z}|x, \tilde{x})$, and if $\hat{p}(\tilde{z}|x, \tilde{x})$ is equal to $p(\tilde{z})$, then this density estimator is dominated by $\hat{p}(\tilde{x}|x)p(\tilde{z})$, due to the non-negativity of KL divergence.

Combining (A.3)–(A.5), we have

$$
R(f, \hat{p}) = E_{X, \tilde{X}|\theta} \log \frac{p(\tilde{X}|\theta)}{\hat{p}(\tilde{X}; X)} = R(\theta, \hat{p}).
$$

Consequently, the minimax risk in the non-parametric regression model is equal to the minimax risk in the Gaussian sequence model. □

Proof of Lemma 4.1. Let $\mathcal{Q}$ be the collection of all (generalized) Bayes predictive densities. Then, by [5], Theorem 5, $\mathcal{Q}$ is a complete class for the problem of predictive density estimation under KL loss. Therefore, the minimax risk among all possible density estimators is equivalent to the minimax risk among (generalized) Bayes estimators, namely,

$$
R(\Theta) = \inf_{\hat{p}} \sup_{\theta \in \Theta} R(\theta, \hat{p}) = \inf_{\hat{p} \in \mathcal{Q}} \sup_{\theta \in \Theta} R(\theta, \hat{p}).
$$

Consider a Gaussian distribution $\pi_S = N(0, S)$, where $S = \mathrm{diag}(s_1^2, \ldots, s_n^2)$ and the s_i's satisfy condition (4.1). Then,

$$
\begin{aligned}
R(\Theta) &= \inf_{\hat{p} \in \mathcal{Q}} \sup_{\theta \in \Theta} R(\theta, \hat{p}) \\
&\geq \inf_{\hat{p} \in \mathcal{Q}} \int_{\Theta} R(\theta, \hat{p}) \pi_S(\theta) \, d\theta
\end{aligned}
\tag{A.6}
$$

$$\geq \inf_{\hat{p}\in\mathcal{Q}} \int_{\mathbb{R}^n} R(\theta,\hat{p})\pi_S(\theta)\,\mathrm{d}\theta - \sup_{\hat{p}\in\mathcal{Q}} \int_{\Theta^c} R(\theta,\hat{p})\pi_S(\theta)\,\mathrm{d}\theta$$

$$\geq \inf_{\hat{p}\in\mathcal{Q}} \int_{\mathbb{R}^n} R(\theta,\hat{p})\pi_S(\theta)\,\mathrm{d}\theta - \sup_{\hat{p}\in\mathcal{Q}} \int_{\Theta^c} R(\theta,\hat{p})\pi_S(\theta)\,\mathrm{d}\theta. \tag{A.7}$$

The first term of (A.7) is the Bayes risk under π_S over the unconstrained parameter space $\mathbb{R}^n$. It is achieved by the linear predictive density $\hat{p}_S$; see [1]. Therefore,

$$\inf_{\hat{p}\in\mathcal{Q}} \int_{\mathbb{R}^n} R(\theta,\hat{p})\pi_S(\theta)\,\mathrm{d}\theta = \int_{\mathbb{R}^n} R(\theta,\hat{p}_S)\pi_S(\theta)\,\mathrm{d}\theta$$
$$= \frac{n}{2m}\log\frac{v_n}{v_{n+m}} + \frac{1}{2m}\sum_{i=1}^{n}\log\frac{v_{n+m}+s_i^2}{v_n+s_i^2}. \tag{A.8}$$

To bound the second term of (A.7), note that for any Bayes predictive density $\hat{p}_\pi \in \mathcal{Q}$,

$$R(\theta,\hat{p}_\pi) = \frac{1}{m}E_{X,\tilde{X}|\theta}\log\frac{p(\tilde{X}|\theta)}{\int_\Theta p(\tilde{X}|\theta')\pi(\theta'|X)\,\mathrm{d}\theta'}$$
$$\leq \frac{1}{m}E_{X,\tilde{X}|\theta}\int_\Theta \log\frac{p(\tilde{X}|\theta)}{p(\tilde{X}|\theta')}\pi(\theta'|X)\,\mathrm{d}\theta' \tag{A.9}$$
$$= \frac{1}{m}E_{X|\theta}\int_\Theta \frac{\|\theta-\theta'\|^2}{2v_m}\pi(\theta'|X)\,\mathrm{d}\theta' $$
$$\leq \frac{1}{mv_m}E_{X|\theta}\int_\Theta (\|\theta\|^2+\|\theta'\|^2)\pi(\theta'|X)\,\mathrm{d}\theta' \tag{A.10}$$
$$\leq \frac{1}{mv_m}\left(\|\theta\|^2+\frac{C}{a_1^2}\right), \tag{A.11}$$

where (A.9) is due to Jensen's inequality, (A.10) is due to $\|\theta-\theta'\|^2 \leq 2\|\theta\|^2+2\|\theta'\|^2$ and (A.11) is due to

$$\int_\Theta \|\theta'\|^2\pi(\theta'|x)\,\mathrm{d}\theta' \leq \sup_{\theta'\in\Theta}\|\theta'\|^2$$
$$\leq \frac{1}{a_1^2}\sup_{\theta\in\Theta}\sum_{i=1}^{n}a_i^2\theta_i'^2 = \frac{C}{a_1^2}.$$

Therefore,

$$\sup_{\hat{p}\in\mathcal{Q}}\int_{\Theta^c} R(\theta,\hat{p})\pi_S(\theta)\,\mathrm{d}\theta \leq \frac{1}{mv_m}\left[\int_{\Theta^c}\|\theta\|^2\pi_S(\theta)\,\mathrm{d}\theta + \frac{C}{a_1^2}\pi_S(\Theta^c)\right], \tag{A.12}$$

where $\pi_S(\Theta^c) = \int_{\Theta^c} \pi_S(\theta)\,\mathrm{d}\theta$. Using the Cauchy–Schwarz inequality, we can further bound the right-hand side of (A.12) as follows:

$$\frac{1}{mv_m}\left[\int_{\Theta^c}\|\theta\|^2\pi_S(\theta)\,\mathrm{d}\theta + \frac{C}{a_1^2}\pi_S(\Theta^c)\right]$$

$$\leq \frac{1}{mv_m}\left[\sum_{i=1}^n\left(\int_{\Theta^c}\theta_i^4\pi_S(\theta)\,\mathrm{d}\theta\right)^{1/2}\sqrt{\pi_S(\Theta^c)} + \frac{C}{a_1^2}\pi_S(\Theta^c)\right]$$

$$= \frac{1}{mv_m}\left[\sqrt{3}\sqrt{\pi_S(\Theta^c)}\sum_{i=1}^n s_i^2 + \frac{C}{a_1^2}\pi_S(\Theta^c)\right]$$

$$\leq \frac{1}{mv_m}\left[\sqrt{3}\frac{C}{a_1}\sqrt{\pi_S(\Theta^c)} + \frac{C}{a_1}\pi_S(\Theta^c)\right].$$

Then, by [3], Proposition 2, which states that if $\epsilon_1, \ldots, \epsilon_m$ are independent Gaussian random variables with $E\epsilon_k = 0$ and $E\epsilon_k^2 = \sigma_k^2$, then

$$\mathbb{P}\left(\sum_{k=1}^m \epsilon_k^2 > Q\right) \leq \exp\left\{-\frac{(Q - \sum_{k=1}^m \sigma_k^2)^2}{4\sum_{k=1}^m \sigma_k^4}\right\},$$

we have

$$\sqrt{\pi_S(\Theta^c)} = \left[\mathbb{P}\left(\sum_{i=1}^n a_i^2\theta_i^2 > C\right)\right]^{1/2} \leq v_n^\alpha, \tag{A.13}$$

due to condition (4.1).

Combining (A.7), (A.8), (A.12) and (A.13), the theorem then follows immediately. □

Acknowledgements

References

[1] Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554.

[2] Aslan, M. (2006). Asymptotically minimax Bayes predictive densities. *Ann. Statist.* **34** 2921–2938.

[3] Belitser, E.N. and Levit, B.Y. (1995). On minimax filtering over ellipsoids. *Math. Methods Statist.* **3** 259–273.

[4] Belitser, E.N. and Levit, B.Y. (1996). Asymptotically minimax nonparametric regression in L_2. *Statistics* **28** 105–122. MR1405604

[5] Brown, L.D., George, E.I. and Xu, X. (2008). Admissible predictive density estimation. *Ann. Statist.* **36** 1156–1170. MR2418653

[6] Brown, L.D. and Low, M.G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958

[7] Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238

[8] Efromovich, S.Y. (1994). On adaptive estimation of nonlinear functionals. *Statist. Probab. Lett.* **19** 57–63. MR1253313

[9] Efromovich, S.Y. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer. MR1705298

[10] Efromovich, S.Y. and Pinsker, M.S. (1982). Estimation of square-integrable probability density of a random variable. *Probl. Inf. Transm.* **18** 175–189. MR0711898

[11] George, E.I., Liang, F. and Xu, X. (2006). Improved minimax prediction under Kullback–Leibler loss. *Ann. Statist.* **34** 78–91. MR2275235

[12] George, E.I. and Xu, X. (2008). Predictive density estimation for multiple regression. *Econometric Theory* **24** 1–17. MR2391619

[13] Goldenshluger, A. and Tsybakov, A.B. (2003). Optimal prediction for linear regression with infinitely many parameters. *J. Multivariate Anal.* **84** 40–60. MR1965822

[14] Golubev, G.K. and Nussbaum, M. (1990). A risk bound in Sobolev class regression. *Ann. Statist.* **18** 758–778. MR1056335

[15] Ibragimov, I.A. and Has'minskii, R.Z. (1977). On the estimation of an infinite-dimensional parameter in Gaussian white noise. *Soviet Math. Dokl.* **236** 1053–1055. MR0483232

[16] Ibragimov, I.A. and Has'minskii, R.Z. (1984). On nonparametric estimation of values of a linear functional in a Gaussian white noise. *Teor. Veroyatn. Primen.* **29** 19–32. MR0739497

[17] Johnstone, I.M. (2003). Function estimation and Gaussian sequence models. Draft of a Monograph.

[18] Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression and model selection. *IEEE Trans. Inform. Theory* **50** 2708–2726. MR2096988

[19] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430. MR1425959

[20] Nussbaum, M. (1999). Minimax risk: Pinsker bound. In *Encyclopedia of Statistical Sciences* (S. Kotz, ed.) 451–460. New York: Wiley.

[21] Pinsker, M.S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Probl. Inf. Transm.* **2** 120–133.

[22] Tsybakov, A.B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. MR1447735