ISSN 1066-5307, Mathematical Methods of Statistics, 2010, Vol. 19, No. 3, pp. 219–245. © Allerton Press, Inc., 2010.

Oracle Convergence Rate of Posterior under Projection Prior and Bayesian Model Selection

A. Babenko[1] **and E. Belitser**[1*]

[1] *Math. Inst., Utrecht Univ., The Netherlands*

Received July 23, 2009; in final form, April 26, 2010

Abstract—We apply the Bayes approach to the problem of projection estimation of a signal observed in the Gaussian white noise model and we study the rate at which the posterior distribution concentrates about the true signal from the space ℓ_2 as the information in observations tends to infinity. A benchmark is the rate of a so-called oracle projection risk, i.e., the smallest risk of an unknown true signal over all projection estimators. Under an appropriate hierarchical prior, we study the performance of the resulting (appropriately adjusted by the empirical Bayes approach) posterior distribution and establish that the posterior concentrates about the true signal with the oracle projection convergence rate. We also construct a Bayes estimator based on the posterior and show that it satisfies an oracle inequality. The results are nonasymptotic and uniform over ℓ_2. Another important feature of our approach is that our results on the oracle projection posterior rate are always stronger than any result about posterior convergence with the minimax rate over all nonparametric classes for which the corresponding projection oracle estimator is minimax over this class. We also study implications for the model selection problem, namely, we propose a Bayes model selector and assess its quality in terms of the so-called false selection probability.

Key words: Bayes approach, Bayes model selector, false selection probability, oracle projection posterior rate, posterior-randomized estimator.

2000 Mathematics Subject Classification: primary 62G20, 62C10; secondary 62G05.

DOI: 10.3103/S1066530710030026

1. INTRODUCTION

Suppose we observe $X = (X_1, X_2, \ldots)$, where

$$X_i = \theta_i + \frac{\xi_i}{\sqrt{n}}, \quad i = 1, 2, \ldots, \tag{1}$$

$\theta = (\theta_i)_{i\in\mathbb{N}} = (\theta_1, \theta_2, \ldots) \in \ell_2$ is an unknown parameter of interest, the noise variables ξ_i are independent, identically distributed $\mathcal{N}(0,1)$ random variables, the parameter n is the noise intensity and reflects the increase of information in the data X as $n \to \infty$. The goal is to make an inference on θ on the basis of the observed data $X = X^{(n)}$. Many quantities actually depend on information parameter n, but for the sake of notational simplicity we will often skip this dependence.

Model (1) is known to be a Gaussian white noise model and it arises in various statistical settings, for example, nonparametric regression model: $Y_i = f(t_i) + \epsilon_i$, $t_i = i/n$, $i = 1, \ldots, n$, $f(\cdot) \in L_2[0,1]$, the ϵ_i's are independent standard normal. The mean vector θ in model (1) can be thought of as Fourier coefficients of signal f with respect to some orthonormal basis from the above regression model. In [5] the asymptotic equivalence between the nonparametric regression model and model (1) was established under some mild conditions. Model (1) is in fact the sequence version of the continuous white noise model, which is widely used in communication theory and signal transmission; see [16], [13], [3], [14]. More generally, model (1) can be considered as the Gaussian sequence version of the generalized linear

*E-mail: e.belitser@uu.nl

Gaussian model on a separable Hilbert space $\mathbb{H}$ as introduced in [4], which covers many Gaussian frameworks. Many typical functional classes can be related to the geometric sets in ℓ_2 via the isometry between $\mathbb{H}$ and ℓ_2, for example, Sobolev classes to ellipsoids, Besov classes to some special type of ℓ_p-bodies.

In a minimax setup, one usually assumes that signal θ belongs to some set $\Theta_\beta \subseteq \ell_2$ with parameter $\beta \in \mathcal{B}$, which typically has the meaning of signal smoothness. Suppose we want to estimate parameter θ measuring the quality of an estimator $\widehat{\theta}$ by a risk function $R(\widehat{\theta}, \theta) = R_n(\widehat{\theta}, \theta)$. Then a benchmark in this statistical problem is the *minimax risk* $r(\Theta_\beta) = r_n(\Theta_\beta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_\beta} R_n(\widehat{\theta}, \theta)$ and the goal would be to find a so-called minimax estimator, i.e., the one attaining the minimax risk, asymptotically or up to a constant factor. For example, in case of the Sobolev ball $\Theta_\beta = \Theta_\beta(Q) = \{\theta\colon \sum_{i=1}^{\infty} i^{2\beta}\theta_i^2 \le Q\}$ and ℓ_2-norm loss $R_n(\widehat{\theta}, \theta) = E_\theta\|\widehat{\theta} - \theta\|^2$ this problem was solved by Pinsker [16]. We use this risk function throughout this paper. Suppose the smoothness parameter β is not known, that is we have a family of models $\{\Theta_\beta, \beta \in \mathcal{B}\}$ and we only know that $\theta \in \Theta_\beta$ for some $\beta \in \mathcal{B}$. In fact, we assume $\theta \in \cup_{\beta\in\mathcal{B}}\Theta_\beta \subseteq \ell_2$. Then the problem of *adaptation* arises: construct a minimax estimator if it exists and which is called minimax adaptive, without the knowledge of smoothness parameter β. The adaptation problem was first studied in [8] and by many authors afterwards.

There is a way to look at the adaptation problem from another perspective. Namely, a framework of so-called oracle inequalities has recently been developed. Suppose we are given a family of estimators $\widehat{\Theta} = \widehat{\Theta}(\mathcal{N}) = \{\widehat{\theta}(N), N \in \mathcal{N}\}$. Slightly abusing notations (cf. the minimax risk $r(\Theta_\beta)$), introduce the *oracle risk* at the signal value θ as the best performance we can achieve by using the family of estimators $\widehat{\Theta}$:

$$r(\theta) = r_n(\theta) = r_n(\theta, \widehat{\Theta}) = \inf_{\widehat{\theta}\in\widehat{\Theta}} R_n(\widehat{\theta}, \theta) = \inf_{N\in\mathcal{N}} R_n(\widehat{\theta}(N), \theta) = R_n(\widehat{\theta}(N_o), \theta). \tag{2}$$

The oracle $N_o = N_o(\theta) = N_o(\theta, n)$ (or the corresponding sequence $N_o^{(k)}$ for which the above infimum is attained) and the oracle risk $r_n(\theta, \widehat{\Theta})$ are our new benchmarks in this approach. The goal is to construct an estimator $\widehat{\theta} = \widehat{\theta}(\widehat{N})$ for some $\widehat{N} = \widehat{N}(X) \in \mathcal{N}$ (although, in principle $\widehat{\theta}$ may not be in the class $\widehat{\Theta}$) mimicking the oracle N_o such that for some positive C_n, the following oracle inequality is satisfied:

$$R_n(\widehat{\theta}(\widehat{N}), \theta) \le C_n r_n(\theta, \widehat{\Theta}) \tag{3}$$

for every $\theta \in \Theta_0 \subseteq \ell_2$. Certainly, $C_n \ge 1$ and the above oracle inequality becomes stronger as C_n gets closer to 1 and the set Θ_0 gets "bigger". Sometimes adding a small penalty term $\bar{P}_n$ (e.g., c/n) to the right-hand side of the oracle inequality makes Θ_0 become the whole space ℓ_2. When properly motivated, one can study other more complicated forms of oracle inequalities, for example, for every $\theta \in \Theta_0$,

$$R_n(\widehat{\theta}(\widehat{N}), \theta) \le C_n \inf_{N\in\mathcal{N}} \Big\{R_n(\widehat{\theta}(N), \theta) + P_n(N)\Big\} + \bar{P}_n$$

with some positive penalty terms $P_n(N)$ and $\bar{P}_n$; see [7], [6], [11]. In this paper, we stick to the basic oracle inequality of the form (3) with $C_n = C > 1$, some absolute constant, and some additional penalty term of order $1/n$. An oracle approach to optimality of estimators was probably first studied in [15] (although without referring to it by the term "oracle" at the time) within the estimators class of ordered linear smoothers and then developed in the series of works by Donoho and Johnstone; see also [4], [7], [12].

The question arises how these two adaptation optimality frameworks, minimax over the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$ and oracle over the estimators family $\widehat{\Theta}(\mathcal{N})$, are related to each other. We discuss this issue in more detail in Section 4. For now we just mention that the oracle approach is stronger if the family $\widehat{\Theta}(\mathcal{N})$ is chosen appropriately. In particular, the oracle approach will imply the results on adaptive minimax. A reasonable choice of this family means it is neither too rich nor too poor in the sense that the family $\widehat{\Theta}(\mathcal{N})$ contains the minimax estimators over all Θ_β, $\beta \in \mathcal{B}$, and there is an estimator satisfying (3) for a sufficiently large Θ_0.

Now consider the Bayes approach. Given the statistical model $X \sim P_\theta = P_\theta^{(n)}$, we put a prior π on θ, which leads to the posterior distribution $P(\theta \mid X)$, the main quantity of interest for the Bayesian analysis.

A Bayesian procedure is regarded to have good asymptotic frequentist properties if the corresponding posterior distribution $P(\theta \mid X)$ concentrates about θ_0 as $n \to \infty$ from the point of view of measure $P_{\theta_0}^{(n)}$. In other words, we assume that $X \sim P_{\theta_0} = P_{\theta_0}^{(n)}$ for some "true" $\theta_0 = (\theta_{0i})_{i\in\mathbb{N}}$ and we want the corresponding posterior distribution $P(\theta \mid X)$ to concentrate about θ_0 as $n \to \infty$, at least for a good Bayesian procedure. To characterize the quality of Bayesian procedures, we look at the rate at which the neighborhood of θ_0 may decrease, while still capturing the most of the posterior mass. To be more precise, a positive sequence r_n is called the *posterior rate* if for any $M_n \to \infty$

$$P\{r_n^{-1}\|\theta - \theta_0\|^2 \geq M_n \mid X\} \to 0 \quad \text{as } n \to \infty \tag{4}$$

in P_{θ_0}-probability. We want this to hold for a sequence r_n converging to zero as fast as possible, for all $\theta_0 \in \Theta$ with set Θ as large as possible and preferably uniformly over $\theta_0 \in \Theta$ if possible.

For example, if we knew that $\theta_0 \in \Theta_\beta$ for nonparametric class Θ_β with smoothness parameter $\beta \in \mathcal{B}$, then the typical benchmark for the posterior rate r_n in (4) would be the minimax risk $r_n(\Theta_\beta) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta_\beta} R(\widehat{\theta}, \theta)$. For a prior $\theta \sim \pi_\beta$ we want the resulting posterior to satisfy the relation (4) with the posterior rate equal to the minimax rate $r_n(\Theta_\beta)$, [9] seems to be the first result of such kind. The relation (4) should preferably be established uniformly over $\theta_0 \in \Theta_\beta$ if possible; if not, then uniformly over a smaller set but as "close" to Θ_β as possible. Suppose the smoothness parameter β is not known, that is we are given a family of models $\{\Theta_\beta, \beta \in \mathcal{B}\}$ and we only know that $\theta \in \Theta_\beta$ for some $\beta \in \mathcal{B}$. Then we can apply the Bayes approach to this adaptation problem. Namely, given a family of priors $\{\pi_\beta,\ \beta \in \mathcal{B}\}$ and a prior λ on $\beta \in \mathcal{B}$, we can design a two-level hierarchical prior π on the pair (θ, β): $\theta \mid \beta \sim \pi_\beta$, $\beta \sim \lambda$. Prior π leads thus to the posterior $P(\theta \mid X)$ and if the true $\theta_0 \in \Theta_\beta$, then the posterior rate adaptiveness of the designed Bayesian procedure means that the relation (4) still holds with $r_n = r_n(\Theta_\beta)$, preferably uniformly over $\theta_0 \in \Theta_\beta$. In case of Sobolev classes Θ_β, $\beta \in \mathcal{B}$, this problem has been studied in [2], where also the issue of uniformity was discussed.

In this paper we develop an oracle optimality framework for the Bayes approach. Any prior π on signal θ leads to the posterior distribution $P(\theta \mid X)$. Recall that $r_n(\theta) = r_n(\theta, \widehat{\Theta}) = \inf_{\widehat{\theta}\in\widehat{\Theta}} R_n(\widehat{\theta}, \theta)$ is the oracle risk at θ as defined by (2) and $\theta_0 \in \Theta_0$. Then $r_n(\theta_0)$, $\theta_0 \in \Theta_0$, is said to be the *posterior oracle rate* (with respect to the class of estimators $\widehat{\Theta}$) if for any $M_n \to \infty$

$$P\big\{r_n^{-1}(\theta_0)\|\theta - \theta_0\|^2 \geq M_n \mid X\big\} \to 0 \quad \text{as } n \to \infty \tag{5}$$

in P_{θ_0}-probability, preferably uniformly over $\theta_0 \in \Theta_0$ if possible. The bigger the set Θ_0, the stronger this property is. The construction of the prior π is as follows. Suppose we are given a family of priors $\{\pi_N,\ N \in \mathcal{N}\}$ with the interpretation that we would use the prior π_k if we knew that the oracle $N_o = k$ for the true signal value θ_0. Since we do not know the oracle N_o, we put a prior on N as well. In doing so, we design a two-level hierarchical prior π on the pair (θ, N): $\theta \mid N \sim \pi_N$, $N \sim \lambda$. An advantage of this prior is that we can also make some statistical inference on N by looking at the resulting posterior $P(N \mid X)$, which can be regarded as Bayesian model selection.

Again it is important to understand how the posterior oracle rate is related to the posterior minimax rate. Suppose that for any $\beta \in \mathcal{B}$ there exists an $N_\beta \in \mathcal{N}$ such that $\widehat{\theta}(N_\beta) \in \widehat{\Theta}(\mathcal{N})$ is minimax over Θ_β. In this case we say that the family $\widehat{\Theta}(\mathcal{N})$ *covers* the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$. Then clearly the result on the posterior oracle convergence rate is stronger than the result on the adaptive posterior convergence with the minimax rate, at least for all $\theta_0 \in \Theta_0 \cap \big(\cup_{\beta\in\mathcal{B}} \Theta_\beta\big)$. It would be desirable to obtain the result (5) for a set Θ_0 which contains the whole scale $\cup_{\beta\in\mathcal{B}}\Theta_\beta$.

In this paper, we introduce the notion of oracle projection convergence rate, propose an appropriate hierarchical prior, construct several different estimators using the posterior, study their oracle properties, establish the oracle projection convergence rate for the resulting (appropriately adjusted by the empirical Bayes approach) posterior distribution and finally address the problem of Bayesian model selection. In fact, we fully implement the program of oracle estimation and Bayes oracle posterior optimality described above for the projection estimators family $\widehat{\Theta}(\mathcal{N})$ in the Gaussian white noise model. The class of projection estimators we consider is parameterized by the so-called cut-off parameter $N \in \mathbb{N}$, which

can be thought of as a model selector. According to the above general scheme, we designed a two-level hierarchical prior on (θ, N): conditionally $\theta \mid N \sim \pi_N$ and $N \sim \lambda$. Conditional priors π_N's are all normal, which makes it possible to compute many quantities, discrete prior λ on the cut-off parameter N satisfies certain conditions. Despite its simplicity, this projection estimators family turns out to be a very good choice. Indeed, this family is not too massive — we established the oracle relations of types (3) and (5) uniformly over $\Theta_0 = \ell_2$. This implies that our results are always stronger than any results on adaptive posterior convergence rate with the minimax rate over any scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the estimators family $\widehat{\Theta}(\mathcal{N})$. On the other hand, this estimators family is rich enough to cover some important scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ such as Sobolev ellipsoids, exponential ellipsoids, hyperrectangles and tail classes; see Section 4 below. This is another appealing feature of this approach: the family $\widehat{\Theta}(\mathcal{N})$ can cover many scales at once. Therefore all the results on adaptive posterior convergence rate with the minimax rate over all these scales follow immediately from our results. All these results will also be uniform over the corresponding scales in ℓ_2, since our results are uniform over the whole space ℓ_2. In particular, our results are stronger than those in [2], where the Sobolev scale was considered. Besides, the uniformity is not an issue anymore as it was in [2]. All the results are nonasymptotic and uniform over ℓ_2.

Recall that our prior includes a prior on parameter N, which enables us to perform the Bayesian model selection to mimic the oracle N_o (in our case N_o is the dimension of the oracle model) by using the posterior $P(N \mid X)$; for example, by taking the maximum a posteriori probability selector $\widehat{N}_{\text{MAP}}$ or by simply generating from $P(N \mid X)$ the posterior randomized selector. To assess the quality of model selectors, we introduce the notion of *false selection probability* $\text{FSP}(\widehat{N}, \tau, \theta_0) = P_{\theta_0}\{\widehat{N} \notin \mathcal{N}(\tau, \theta_0)\}$, where $\mathcal{N}(\tau, \theta_0) \subseteq \mathbb{N}$ is the index set of the so-called admissible models, τ is a predetermined *tolerance* parameter, which describes how lenient the definition of admissible model is. We investigate the performance of the proposed Bayes model selectors with respect to this criterion and establish that the false selection probability at point θ_0 is small if the oracle risk at point θ_0 is not "too parametric"; roughly speaking, $r_n(\theta_0) \geq Cn^{-1}$ for some big constant C.

The proposed methodology can in principle be extended to other statistical models. In the present paper, we however consider the simplest Gaussian infinite-dimensional framework, which allows us to illustrate the main ideas of the proposed approach without too much technicalities and to derive closed expressions of many quantities and constants involved. As to the results on the oracle posterior rate, we are not aware of other studies of such kind. To the best of our knowledge, all the results about posterior rates obtained until now are actually global, typically such results are related to the minimax rates for the estimation problem over some nonparametric smoothness classes; cf. [9]. The point is that the existing methods for the derivation of posterior concentration rates are all based on some global quantities, like entropy characteristics and existence of uniform tests, and lead therefore to global posterior rates. Oracle rates are on the contrary intrinsically local: the oracle rate depends on the true value θ_0 and it is typically small for "good" signal values. The derivation of the results on oracle concentration rates requires therefore the development of new techniques and the present paper is intended to make a first step in this direction.

2. PRELIMINARIES

Denote the probability measure of X from the model (1) by $P_\theta = P_\theta^{(n)}$. In case θ is a stochastic element as in the Bayesian analysis, denote by P_θ the conditional probability measure of X given θ, by $P(\theta \mid X)$ the conditional distribution of θ given X and by P the joint probability measure of (X, θ). The same notation applies to the expectation operation. Denote by $I\{S\}$ the indicator function of the set S and by $|S|$ the cardinality of the set S, $\mathbb{N} = \{1, 2, \ldots\}$, i.e., zero is not included in $\mathbb{N}$. For a constant $c \in \mathbb{R}$, $c\theta = (c\theta_i)_{i \in \mathbb{N}}$.

Introduce the class of projection estimators $\{\widehat{\theta}(N),\ N \in \mathbb{N}\}$ which is parametrized by a so-called cut-off parameter N. It is the dimension of the approximating linear subspace $S_N = \{(s_k)_{k \in \mathbb{N}} \colon s_k = 0,\ k > N\}$ on which we project the data X to obtain the corresponding projection estimator: for an $N \in \mathbb{N}$,

$$\widehat{\theta}(N) = \Pi_{S_N} X = (\widehat{\theta}_i(N))_{i \in \mathbb{N}}, \qquad \widehat{\theta}_i(N) = X_i I\{i \leq N\}, \quad i \in \mathbb{N}. \tag{6}$$

The performance of the estimator $\widehat{\theta}(N)$ at point θ is measured by the risk function

$$\mathcal{R}_n(N) = \mathcal{R}_n(N,\theta) = R(\widehat{\theta}(N),\theta) = E_\theta\|\widehat{\theta}(N)-\theta\|^2.$$

It can be easily checked that

$$\mathcal{R}_n(N,\theta) = E_\theta\sum_{i=1}^{\infty}(\widehat{\theta}_i(N)-\theta_i)^2 = E_\theta\Big[\sum_{i=1}^{N}\frac{\xi_i^2}{n}+\sum_{i=N+1}^{\infty}\theta_i^2\Big] = \frac{N}{n}+\sum_{i=N+1}^{\infty}\theta_i^2. \tag{7}$$

For a fixed $\theta\in\ell_2$, we define the oracle $N_o = N_o(\theta)\in\mathbb{N}$ and the *oracle projection risk* (or just the *oracle risk*) $r_n = r_n(\theta) = \mathcal{R}_n(N_o,\theta)$ by the following relation:

$$r_n = r_n(\theta) = \mathcal{R}_n(N_o,\theta) = \frac{N_o}{n}+\sum_{i=N_o+1}^{\infty}\theta_i^2 = \min_{N\in\mathbb{N}}\mathcal{R}_n(N,\theta). \tag{8}$$

As is easy to see, the oracle $N_o(\theta)$ is well defined for any $\theta\in\ell_2$ and it is not an estimator, since it depends on the unknown θ. Recall that zero is not included in $\mathbb{N}$, which ensures that the oracle risk is always positive, in fact $\mathcal{R}_n(N_o,\theta)\ge n^{-1}$ and $N_o\ge 1$. It is not restrictive since n^{-1} is the best (parametric) rate in case at least one coordinate $\theta_i\ne 0$. We want our problem to be at least as difficult as the parametric one, so we will have to avoid the trivial case $\theta=(0,0,\ldots)$ anyway. For example, if we allow $N=0$, then our results will hold only for $\theta\in\Theta_0=\{\theta\in\ell_2\colon \mathcal{R}_n(N_o,\theta)\ge n^{-1}\}$. Alternatively, we can add a penalty term $1/n$ to the oracle risk and take the resulting sum as our new oracle benchmark.

Now our goal is to propose a two-level hierarchical prior on (θ,N) in the manner we described in the Introduction, so that we can study the resulting posterior $P(\theta\mid X)$ and the properties of an estimator $\widehat{\theta}$ constructed by using the posterior distribution $P(\theta\mid X)$ and $P(N\mid X)$ under the probability measure P_{θ_0} for some fixed ("true") $\theta_0\in\ell_2$.

First, consider the case of a fixed cut-off parameter N. Let $\{\pi_N(\theta),\ N\in\mathbb{N}\}$, be a family of priors on θ defined as follows:

$$\theta_i\mid N \overset{ind}{\sim}\mathcal{N}(0,\tau_i^2(N)),\qquad \tau_i^2(N) = n^{-1}I\{i\le N\},\quad i\in\mathbb{N}, \tag{9}$$

with the convention from now on that if $Z\sim\mathcal{N}(c,0)$, then $Z=c$ with probability 1. $\pi_N(\theta)$ is the product measure of normal distributions $\mathcal{N}(0,\tau_i^2(N))$, $i\in\mathbb{N}$. Recall that $X_i\mid\theta\overset{\text{ind}}{\sim}\mathcal{N}(\theta_i,n^{-1})$, so that the posterior distribution $P_N(\theta\mid X)$ corresponding to the prior $\pi_N(\theta)$ defined by (9) is readily obtained:

$$\theta_i\mid(X,N)\overset{ind}{\sim}\mathcal{N}\Big(\frac{X_i}{2}I\{i\le N\},\frac{n^{-1}}{2}I\{i\le N\}\Big),\quad i\in\mathbb{N}. \tag{10}$$

Here we used the elementary fact that if $Z\mid Y\sim\mathcal{N}(Y,\sigma^2)$ and $Y\sim\mathcal{N}(\mu,\tau^2)$, then

$$Y\mid Z\sim\mathcal{N}\Big(\frac{Z\tau^2+\mu\sigma^2}{\tau^2+\sigma^2},\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\Big).$$

The Bayes estimator with respect to the prior $\pi_N(\theta)$ is $\tilde{\theta}=E_N(\theta\mid X)=(\tilde{\theta}_i)_{i\in\mathbb{N}}$. Notice that $\tilde{\theta}=\frac{1}{2}\widehat{\theta}(N)$, that is $\tilde{\theta}_i=\frac{1}{2}\widehat{\theta}_i(N)$, $i\in\mathbb{N}$, where $\widehat{\theta}(N)$ is the projection estimator (6) with cut-off N. It follows therefore that the Bayes estimator $\tilde{\theta}$ converges to $\theta_0/2$ with the rate $\mathcal{R}_n(N,\theta_0)$ (up to a constant). Besides, it is not so difficult to see (for example, by applying the conditional Chebyshev inequality) that the posterior $P_N(\theta\mid X)$ concentrates about $\theta_0/2$ as $n\to\infty$ in P_{θ_0}-probability with the rate $\mathcal{R}_n(N,\theta_0)$. There is nothing special about factor $1/2$, in fact we can get any other factor in $(0,1)$ by taking the variances in the product prior π_N as $\tau_i^2(N)=Cn^{-1}I\{i\le N\}$, $i\in\mathbb{N}$, for some constant $C>0$. The reason for this is that our prior π_N is normal with zero mean so that the corresponding Bayes estimator is always a shrinkage estimator and thus it is always going to be a fraction of the projection estimator $\widehat{\theta}(N)$. To fix this problem, instead of (9), take the prior $\pi_{N,\mu}$ defined as

$$\theta_i\mid N\overset{ind}{\sim}\mathcal{N}(\mu_i(N),\tau_i^2(N)),\quad \mu_i(N)=\mu_iI\{i\le N\},\quad \tau_i^2(N)=n^{-1}I\{i\le N\},\quad i\in\mathbb{N},$$

with the idea that $\mu(N) = (\mu_i(N))_{i\in\mathbb{N}}$ would model a possible shift. The corresponding posterior is easy to derive in the same way as for (10):

$$\theta_i \mid (X, N) \stackrel{ind}{\sim} \mathcal{N}\Big(\frac{X_i + \mu_i}{2} I\{i \le N\}, \frac{n^{-1}}{2} I\{i \le N\}\Big), \quad i \in \mathbb{N}. \tag{11}$$

For now μ_i's are some parameters. To estimate these, apply the empirical Bayes approach, i.e., we use the marginal (of X) maximum likelihood estimates for μ_i's. Marginal distribution of X: $X_i \mid \mu(N) \stackrel{ind}{\sim} \mathcal{N}\big(\mu_i(N), n^{-1} + \tau_i^2(N)\big)$, $i \in \mathbb{N}$, so that trivially $\widehat{\mu}_i = X_i$, $i \in \mathbb{N}$. Thus, we take as posterior the conditional distribution $P'_N(\theta \mid X)$ defined as

$$\theta_i \mid (X, N) \stackrel{ind}{\sim} \mathcal{N}\Big(X_i I\{i \le N\}, \frac{n^{-1}}{2} I\{i \le N\}\Big), \quad i \in \mathbb{N}. \tag{12}$$

Now we can make all our inference on the basis of this empirical Bayes posterior $P'_N(\theta \mid X)$. Clearly, the corresponding (empirical) Bayes estimator coincides with the projection estimator $\widehat{\theta}(N)$ and this empirical Bayes posterior distribution concentrates about θ_0 with projection rate $\mathcal{R}_n(N, \theta_0)$ in P_{θ_0}-probability.

Clearly, if we knew the oracle N_o defined by (8), we would of course use the prior π_{N_o}. Since the oracle is not known we step to the next level in our Bayesian analysis by putting a prior on N as well. Formally, we now regard the distribution $\pi_N(\theta)$ to be the conditional distribution on θ given N, and next put a prior $\lambda = \lambda_\alpha$ on $N \in \mathbb{N}$, $N \sim \lambda$:

$$P(N = k) = \lambda_k = c(\alpha) e^{-\alpha k}, \quad k \in \mathbb{N},$$

where $c(\alpha) = (1 - e^{-\alpha})e^{\alpha}$ is the normalizing constant and the constant $\alpha > 0$ is to be specified later. Note that the prior λ is a geometric distribution $\lambda_k = p(1-p)^{k-1}$, $k \in \mathbb{N}$, with parameter $p = 1 - e^{-\alpha}$. In doing so, we introduced a two-level hierarchical prior $\pi = \pi_\alpha$ on (θ, N):

$$(\theta, N) \sim \pi \quad \Longleftrightarrow \quad \theta \mid N \sim \pi_N, \quad N \sim \lambda_\alpha, \tag{13}$$

with λ_α given above and π_N defined by (9). This leads to the posterior distributions $P(\theta \mid X) = P_\alpha(\theta \mid X)$ and $P(N \mid X) = P_\alpha(N \mid X)$. Although many quantities (like π_α, λ_α and $P_\alpha(\theta \mid X)$) depend on the parameter $\alpha > 0$, we typically skip this subscript to avoid complicated notations, unless we occasionally want to emphasize this dependence. There is a small notational abuse: π_α is the mixture of π_N's over $N \sim \lambda_\alpha$.

The posterior distribution $P(\theta \mid X)$ can be expressed as follows (for a measurable set $S \subset \ell_2$):

$$P(\theta \in S \mid X) = \sum_{k\in\mathbb{N}} P(\theta \in S \mid X, N = k) P(N = k \mid X), \tag{14}$$

where the conditional distribution $P(\theta \in S \mid X, N = k)$ is defined by (10) with $N = k$.

Recall that the family of priors π_N, $N \in \mathbb{N}$, makes the resulting posterior concentrate about $\theta_0/2$ in P_{θ_0}-probability. The same phenomenon occurs for the posterior $P(\theta \mid X)$ corresponding to the prior π_α, which is not surprising since π_α is merely a mixture of π_N's. We fix this problem by using the empirical Bayes approach as we did this above for the prior π_N. Namely, instead of posterior (14) we base the inference on the empirical Bayes posterior distribution

$$P'(\theta \in S \mid X) = \sum_{k\in\mathbb{N}} P'(\theta \in S \mid X, N = k) P(N = k \mid X), \tag{15}$$

where the conditional distribution $P'(\theta \mid X, N = k)$ is defined by (12) with $N = k$.

Using the above posterior distributions, we finally construct some estimators for the signal θ. From now on denote by $\widehat{\theta} = (\widehat{\theta}_i)_{i\in\mathbb{N}}$ an estimator of signal θ, which is in general a measurable function of X, α and n: $\widehat{\theta} = \widehat{\theta}(X, \alpha, n)$. First define a version of empirical Bayes estimator:

$$\widehat{\theta} = E'(\theta \mid X) \tag{16}$$

with the conditional expectation E' taken with respect to the empirical Bayes posterior distribution (15). The next estimator is in fact a version of the projection estimator with a posterior-randomized cut-off parameter $\widehat{N}$:

$$\widehat{N} \sim P(N \mid X), \qquad \widehat{\theta} = \widehat{\theta}(\widehat{N}) = \widehat{\theta}(\widehat{N}, X, \alpha), \tag{17}$$

where the random variable $\widehat{N} = \widehat{N}(X, \alpha)$ is drawn from the posterior distribution $P(N \mid X) = P_\alpha(N \mid X)$, $\widehat{\theta}(N)$ is the projection estimator defined by (6) with cut-off N. Notice that $\widehat{\theta}$ defined by (17) can be expressed as $\widehat{\theta} = 2E(\theta \mid X, N = k)\big|_{k=\widehat{N}}$.

We conclude this section with a couple of remarks.

Remark 1. Notice that as $\theta \in \ell_2$, then $\widehat{\theta}(N)$ with $N > n$ are all inadmissible for sufficiently large n. Indeed, assume that $N > n$, then $\mathcal{R}_n(N, \theta) > 1$ while $\mathcal{R}_n\big(\lfloor\sqrt{n}\rfloor, \theta\big) \to 0$ as $n \to \infty$. Therefore one could in principle restrict the set of possible values for the cut-off parameter N to the set $\mathcal{N}_n = \{1, \dots, n\}$ instead of $\mathbb{N}$. The set $\mathcal{N}_n$ of possible values for the cut-off parameter N would arise by itself in case we had a high-dimensional version of the model (1) instead of infinite-dimensional, i.e., $i = 1, \dots, n$ in (1), the observation $X = (X_1, \dots, X_n)$ and the unknown parameter $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$. Typically, there is not much difference between the above described high-dimensional and our infinite-dimensional models and actually many papers deal with the high-dimensional rather than the infinite-dimensional case. However, there is a significant difference between these two situations when the intensity of noise n^{-1} is unknown. The high-dimensional situation becomes problematic, while in the infinite-dimensional model one can base an estimate of n^{-1} on the observations $X_{l+1}, X_{l+2}, \dots, X_{l+p}$ for sufficiently large integers l, p, which are approximately $\mathcal{N}(0, n^{-1})$-distributed for large l. In fact, one can achieve an arbitrary precision in estimating n^{-1}, uniformly over some Sobolev ball $\Theta_\beta(Q)$ for some fixed $\beta, Q > 0$. In this paper we however assume parameter n to be known and thus consider the more general infinite-dimensional case $N \in \mathbb{N}$.

Remark 2. The above framework can be generalized as follows. Instead of $\{\widehat{\theta}(N),\ N \in \mathbb{N}\}$, we can consider the following more general family of estimators $\{\widehat{\theta}(I_k),\ k \in \mathbb{N}\}$:

$$\widehat{\theta}(I_k) = \Pi_{S_{I_k}} X = (\widehat{\theta}_i(I_k))_{i \in \mathbb{N}} \quad \text{with} \quad \widehat{\theta}_i(I_k) = X_i I\{i \in I_k\}, \quad i \in \mathbb{N},$$

the family of the projectors of the data on the subspaces $S_{I_k} = \{(s_i)_{i \in \mathbb{N}} \colon s_i \in \mathbb{R},\ i \in I_k;\ s_i = 0,\ i \notin I_k\}$. Here $\{I_k\}_{k \in \mathbb{N}}$ is a countable nested family of subsets of $\mathbb{N}$: $I_k \subset \mathbb{N}$, $I_k \subset I_{k+1}$, $|I_k| = M_k$. Finite families can be considered as well. So the corresponding nested family of finite-dimensional linear subspaces $\{S_{I_k}\}_{k \in \mathbb{N}}$ (i.e., $S_{I_k} \subset S_{I_{k+1}}$, $\dim(S_{I_k}) = M_k$, $k \in \mathbb{N}$) can be interpreted as a collection of finite-dimensional models, where model S_{I_k} corresponds to the choice $\{\theta_i,\ i \in I_k\}$ as "most significant" variables. We can assume without loss of generality that $I_k = \{1, \dots, M_k\}$. Indeed, we can always rearrange the coordinates of signal θ in such a way that this holds. Next, notice that the particular case $M_k = k$ is exactly the one we study in this paper. It is related to the so-called ordered variable selection problem, cf. [4]. However this case is most important since it corresponds to the most detailed slicing of the space ℓ_2 by a family $\{S_{I_k}\}_{k \in \mathbb{N}}$ of embedded pieces $S_{I_k} \subset S_{I_{k+1}}$, $k \in \mathbb{N}$, and thus the most difficult one to handle. The results and their proofs for any less detailed slicing case ($M_{k+1} > M_k + 1$ for some $k \in \mathbb{N}$) can be obtained along similar lines.

Another generalization concerns the high-dimensional case $\theta \in \mathbb{R}^n$ described in the previous remark. One can apply our approach to the situation corresponding to the complete variable selection, namely the family of linear subspaces $\{S_I,\ I \subseteq \mathcal{N}_n\}$, $S_I = \{(s_i)_{i \in \mathcal{N}_n} \in \mathbb{R}^n \colon s_i = 0, i \notin I\}$, $\mathcal{N}_n = \{1, \dots, n\}$. The conjecture is that there will be a price for the complete variable selection in the posterior convergence rate: as compared to the oracle risk, it is expected to be slower by a $\log n$ factor. This problem will be considered elsewhere.

Remark 3. As to the estimation of the parameter μ in (11), instead of the empirical Bayes approach one could also use pure Bayesian approach by putting some prior on μ. We believe this approach should work as well, but its theoretical treatment seems to be more difficult since it adds one more hierarchy level to the Bayesian analysis.

Remark 4. Apart from the empirical Bayes approach, one can propose two more ways to fix the posterior: shifting the posterior and rescaling the posterior by an appropriate factor. The first approach is to construct an estimator $\widehat{\theta}$ by using the posterior $P(N \mid X)$ which converges to θ_0 with oracle rate, and then shift the posterior $P(\theta \mid X)$ by the factor $\widehat{\theta}/2$ to make the resulting posterior distribution concentrate about the true value θ_0 with the oracle rate in P_{θ_0}-probability.

The second approach is based on rescaling the posterior distributions $P(\theta \mid X, N) = P_N(\theta \mid X)$ in (14) by factor 2. First consider the nonadaptive case when the parameter N is fixed. Introduce the conditional distribution $P''_N(\theta \mid X)$:

$$\theta_i \mid X \overset{ind}{\sim} \mathcal{N}\Big(X_i I\{i \le N\}, 2n^{-1} I\{i \le N\}\Big), \quad N, i \in \mathbb{N}. \tag{18}$$

Notice that the distribution $P''_N(\theta \mid X)$ is simply the distribution of $2\theta \mid X$ with $\theta \mid X \sim P_N(\theta \mid X)$, i.e., a rescaled version of $P_N(\theta \mid X)$ defined by (10). Then this newly defined posterior distribution will move towards θ_0 in the sense that its Bayes estimator will converge to θ_0 and this posterior itself will concentrate about θ_0 with the rate $\mathcal{R}_n(N, \theta_0)$, although with somewhat bigger variances in this resulting product posterior distribution.

Coming back to the adaptive case when we put a prior on N instead of posterior $P(\theta \mid X)$ defined by (14), we base the inference on the conditional distribution $P''(\theta \mid X)$, the rescaled version of $P(\theta \mid X)$ defined by (14) with the scale parameter 2:

$$P''(\theta \in S \mid X) = \sum_{k \in \mathbb{N}} P''(\theta \in S \mid X, N = k) P(N = k \mid X), \tag{19}$$

where the conditional distribution $P''(\theta \mid X, N) = P''_N(\theta \mid X)$ is defined by (18).

Remark 5. If one wants to avoid the randomization as in (17), then one could use an estimator of the form (17) but with a nonrandomized cut-off: $\widehat{\theta} = \widehat{\theta}(\widehat{N}_{\text{MAP}})$, with $\widehat{N}_{\text{MAP}} = \widehat{N}_{\text{MAP}}(X)$ being the so-called maximum a posteriori probability (MAP) estimator defined by

$$\widehat{N}_{\text{MAP}} = \widehat{N}_{\text{MAP}}(X) = \arg\max\{P(N = k \mid X),\, k \in \mathbb{N}\}, \tag{20}$$

i.e., $\widehat{N}_{\text{MAP}}$ is the mode of the posterior distribution $P(N = k \mid X)$. Since $\sum_{k \in \mathbb{N}} P(N = k \mid X) = 1$, the selector $\widehat{N}_{\text{MAP}}$ is well defined almost surely.

Remark 6. Another candidate for estimator $\widehat{\theta}$ of signal θ is possible: for example, $\widehat{\theta} = 2E(\theta \mid X)$, with the conditional expectation E taken with respect to the posterior (14). Factor 2 is due to the shrinkage phenomenon discussed above. Notice that $\widehat{\theta} = 2E(\theta \mid X) = E''(\theta \mid X)$ with the conditional expectation E'' taken with respect to the posterior (19). However this estimator is very much alike the estimator (16). This stems from the fact that the posteriors $P'(\theta \mid X)$ and $P''(\theta \mid X)$ are very similar, with the only difference between them being a constant factor in the conditional variances $\mathrm{Var}(\theta_i \mid X, N = k)$, which is $n^{-1} I\{i \le k\}$ for the posterior (15) and $2n^{-1} I\{i \le k\}$ for the posterior (19). Therefore we will not study the posterior (19) and the estimator $\widehat{\theta} = E''(\theta \mid X)$, as the derivation of their properties is exactly the same as for the posterior (15) and the estimator (16) respectively, with some modified constants in the proof.

3. MAIN RESULTS

This section contains the main results of the paper. The proofs are deferred to the last section.

3.1. Oracle properties of the Bayes estimator and oracle posterior rate

The first theorem claims that the estimator (17) mimics the oracle, i.e., satisfies an oracle inequality of the form (3). This estimator can therefore be used as a correcting shifting factor for the posterior distribution $P(\theta \mid X)$ later on.

Theorem 1. *Let $\theta_0 \in \ell_2$, let the prior π_α be defined by* (13) *with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$, and let the estimator $\widehat{\theta}$ be defined either by* (16) *or by* (17)*. Then there exist constants $K_1, K_2 > 0$ depending only on α such that*

$$E_{\theta_0}\|\widehat{\theta} - \theta_0\|^2 \le K_1 \mathcal{R}_n(N_o, \theta_0) + \frac{K_2}{n}.$$

Similar oracle inequalities are obtained in [10], in [7] for blockwise Stein estimators and in [4] for penalized estimators. The results in those papers are in fact stronger in some respects. The above theorem shows that Bayes estimators with appropriately chosen prior satisfy some projection oracle inequalities as well, which makes it possible to use these estimators as a shifting factor for the posterior distribution $P(\theta \mid X)$ later on. Besides, the theorem says that the posterior-randomized selector $\widehat{N}$ defined by (17) mimics the oracle in the sense that $\widehat{\theta}(\widehat{N})$ satisfies the oracle inequality of the theorem.

The next result, which is the main result in the paper, establishes that the posterior distributions $P(\theta \mid X)$ and $P'(\theta \mid X)$ concentrate about $\theta_0/2$ and θ_0 respectively in P_{θ_0}-probability with the posterior oracle projection rate $r_n(\theta_0)$ defined by (8), uniformly in $\theta_0 \in \ell_2$.

Theorem 2. *Let the oracle rate $r_n(\theta_0)$ be defined by* (8) *and the prior π_α be defined by* (13) *with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$. Then there exist constants $C_1, C_2 > 0$ depending only on α such that, for any $\theta_0 \in \ell_2$ and any $M > 0$,*

$$E_{\theta_0} P\Big\{\big\|\theta - \theta_0/2\big\|^2 \ge M r_n(\theta_0) \mid X\Big\} \le \frac{C_1}{M}, \quad E_{\theta_0} P'\Big\{\big\|\theta - \theta_0\big\|^2 \ge M r_n(\theta_0) \mid X\Big\} \le \frac{C_2}{M},$$

where posteriors $P(\theta \mid X)$ and $P'(\theta \mid X)$ are defined by (14) *and* (15) *respectively.*

Further, the parts of Theorems 1 and 2 dealing with the estimator (17) and the posterior (14) imply that an appropriately shifted posterior distribution $P(\theta \mid X)$ concentrates about θ_0 in P_{θ_0}-probability with the projection oracle posterior rate $r_n(\theta_0)$ defined by (8), uniformly in $\theta_0 \in \ell_2$.

Corollary 1. *Let the oracle rate $r_n(\theta_0)$ be defined by* (8)*, the prior π_α be defined by* (13) *with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$, and the estimator $\widehat{\theta}$ be defined either by* (16) *or by* (17)*. Then there exists a constant $C' > 0$ depending only on α such that for any $\theta_0 \in \ell_2$ and any $M > 0$*

$$E_{\theta_0} P\Big\{\big\|\theta - \theta_0 + \widehat{\theta}/2\big\|^2 \ge M r_n(\theta_0) \mid X\Big\} \le \frac{C'}{M},$$

where posteriors $P(\theta \mid X)$ are defined by (14)*.*

Remark 7. Note that the results of both above theorems are uniform in $\theta_0 \in \ell_2$ and nonasymptotic, i.e., also uniform in $n \in \mathbb{N}$.

Remark 8. The results for the estimator $\widehat{\theta} = E''(\theta \mid X)$ and for the rescaled posterior $P''(\theta \mid X)$ defined by (19) follow in the same way as for the estimator $\widehat{\theta} = E'(\theta \mid X)$ and for the empirical Bayes posterior $P'(\theta \mid X)$ defined by (15) with slightly different constants.

Remark 9. Take any $M_n > 0$ such that $M_n \to \infty$ as $n \to \infty$. Then Theorem 2 and Corollary 1 imply that, under the conditions of Corollary 1,

$$P'\Big\{\big\|\theta - \theta_0\big\|^2 \ge M_n r_n(\theta_0) \mid X\Big\} \to 0 \quad \text{and} \quad P\Big\{\big\|\theta - \theta_0 + \widehat{\theta}/2\big\|^2 \ge M_n r_n(\theta_0) \mid X\Big\} \to 0$$

as $n \to \infty$ in P_{θ_0}-probability, uniformly in $\theta_0 \in \ell_2$.

Remark 10. As was already discussed, with the prior we use, we need either to shift or rescale the resulting posterior distribution to make it concentrate with the oracle rate about θ_0 rather than about $\theta_0/2$. The shifting factor is data driven and good enough to ensure that the concentration rate still remains the same, which is the projection oracle rate. For example, the shifting factor based on the Bayes estimator $\widehat{\theta}$ defined by (16) is optimal in the sense of projection oracle properties. So the results with a data-driven optimal shifting or rescaling (in our case the scale parameter 2 is fixed) of the posterior are as good as any other results about the oracle posterior convergence rate when the posterior is not adjusted. Sometimes adjusting the posterior may be an intrinsic operation. For example, in the classical parametric statistics, the Bernstein–von Mises theorem is about an appropriately shifted posterior distribution. However, manipulating (for example, shifting or rescaling) the posterior distribution after a Bayesian analysis is not in the core of the traditional Bayesian paradigm. The empirical Bayes approach for adjusting the posterior is on the contrary well established and accepted in the Bayesian community. These remarks though are more of a philosophical, methodological rather than mathematical nature.

Remark 11. The obtained results establish in some sense the correct behavior of the posterior distributions of θ and N from the P_{θ_0}-perspective, which seems to make it possible to use them in solving another interesting challenging problem — the construction of an adaptive confidence set for θ_0 by using these posteriors. This is a subtle issue to address and will be considered elsewhere.

3.2. Bayesian Model Selection, Assessing the False Selection Probability

The next result concerns the model selection problem. For a $\tau > 0$, $\theta_0 \in \ell_2$ and the oracle $N_o = N_o(\theta_0)$ defined by (8), introduce the sets:

$$\mathcal{N}(\tau) \;=\; \mathcal{N}(\tau,\theta_0) = \big\{k \in \mathbb{N}\colon \mathcal{R}_n(k,\theta_0) \le \tau \mathcal{R}_n(N_o,\theta_0)\big\}, \tag{21}$$

$$\mathcal{N}^-(\tau) \;=\; \mathcal{N}^-(\tau,\theta_0) = \big\{k \in \mathbb{N}\colon k < N_o,\ \mathcal{R}_n(k,\theta_0) > \tau \mathcal{R}_n(N_o,\theta_0)\big\}, \tag{22}$$

$$\mathcal{N}^+(\tau) \;=\; \mathcal{N}^+(\tau,\theta_0) = \big\{k \in \mathbb{N}\colon k > N_o,\ \mathcal{R}_n(k,\theta_0) > \tau \mathcal{R}_n(N_o,\theta_0)\big\}. \tag{23}$$

For any $\tau \ge 1$, these sets form a partition of $\mathbb{N}$, i.e., they are disjoint and $\mathbb{N} = \mathcal{N}^-(\tau) \cup \mathcal{N}(\tau) \cup \mathcal{N}^+(\tau)$. From now on, we deal only with $\tau \ge 1$.

The interpretation of the set $\mathcal{N}(\tau,\theta_0)$ is clear — it specifies the set of acceptable models in the sense that the risks $\mathcal{R}_n(k,\theta_0)$ for all models $k \in \mathcal{N}(\tau,\theta_0)$ are all within the constant factor τ of the oracle risk $\mathcal{R}_n(N_o,\theta_0)$. We call this constant τ tolerance parameter. For a model selector $\tilde{N} = \tilde{N}(X) \in \mathbb{N}$ and a tolerance $\tau \ge 1$, we define our quality measure of model selectors — the false selection probability (FSP):

$$\mathrm{FSP}(\tilde{N},\tau,\theta_0) = \mathrm{FSP}(\tilde{N},\tau,\theta_0,n) = P_{\theta_0}\big\{\tilde{N} \notin \mathcal{N}(\tau,\theta_0)\big\} \tag{24}$$

with $\mathcal{N}(\tau,\theta_0)$ defined by (21).

Theorem 3. *Let the model selector $\widehat{N}(X,\alpha)$ be defined by* (17)*. Then, for any $\alpha \in [\frac{1}{6} - \log\big(\frac{2}{\sqrt{3}}\big), \frac{1}{2}]$ and τ such that*

$$\tau > \tau_-(\alpha) = 6\log\Big(\frac{2}{\sqrt{3}}\Big) + 6\alpha \quad \textit{and} \quad \tau \ge \tau_+(\alpha) = \frac{1}{2\alpha},$$

there exist constants $B_1, B_2, B_3 > 0$ depending only on α and τ such that for all $\theta_0 \in \ell_2$

$$\mathit{FSP}(\widehat{N},\tau,\theta_0) = P_{\theta_0}\big\{\widehat{N} \notin \mathcal{N}(\tau,\theta_0)\big\} \le \min\Big\{\frac{B_1}{n\mathcal{R}_n(N_o,\theta_0)}, \frac{B_2 n\mathcal{R}_n(N_o,\theta_0)}{e^{B_3 n\mathcal{R}_n(N_o,\theta_0)}}\Big\}.$$

Remark 12. In Theorem 3, the constants B_1 and B_2 decrease and B_3 increases as tolerance τ grows. The exact relations are given in the proof of the theorem.

If the oracle quantity $n\mathcal{R}_n(N_o,\theta_0)$ is large enough, it is easily seen from the above theorem that the false selection probability for the selector $\widehat{N}$ is small, i.e., $\widehat{N}$ selects a "good" model (from $\mathcal{N}(\tau,\theta_0)$) with high probability. The oracle quantity $n\mathcal{R}_n(N_o,\theta_0) \geq K$ if there are at least K coordinates in vector θ_0 which are not less than n^{-1}. If $n\mathcal{R}_n(N_o,\theta_0)$ is small, θ_0 is too close to the zero signal for the selector $\widehat{N}$ to perform well. However, according to the above remark, even for such θ_0's, the method does a good job if we lower our tolerance requirement: making tolerance parameter τ bigger leads to a smaller false selection probability.

Another possibility is the so-called maximum a posteriori probability (MAP) model selector $\widehat{N}_{\text{MAP}}(X)$, the mode of the posterior distribution $P(N=k \mid X)$ defined by (20). Recall that it also depends on the parameter α. It is expected that this selector lives on the set $\mathcal{N}(\tau,\theta_0)$, at least for "good" θ_0's. The following assertion elaborates on this.

Theorem 4. *Let the model selector $\widehat{N}_{MAP}(X)$ be defined by* (20)*. Then, under the conditions of Theorem* 3 *and for all $\theta_0 \in \ell_2$,*

$$\begin{aligned} FSP(\widehat{N}_{MAP},\tau,\theta_0) &= P_{\theta_0}\{\widehat{N}_{MAP} \notin \mathcal{N}(\tau,\theta_0)\} \leq \bigl(1+|\mathcal{N}(\tau,\theta_0)|^{1/2}\bigr)^2 P_{\theta_0}\{\widehat{N} \notin \mathcal{N}(\tau,\theta_0)\} \\ &\leq \bigl(1+\sqrt{\tau n\mathcal{R}_n(N_o,\theta_0)}\bigr)^2 \min\Bigl\{\frac{B_1}{n\mathcal{R}_n(N_o,\theta_0)}, \frac{B_2 n\mathcal{R}_n(N_o,\theta_0)}{e^{B_3 n\mathcal{R}_n(N_o,\theta_0)}}\Bigr\}. \end{aligned}$$

Although the above result claims somewhat worse properties of the MAP selector $\widehat{N}_{\text{MAP}}$ as compared to the posterior randomized selector $\hat{N}$, it seems actually that the MAP selector should have at least the same quality as $\hat{N}$ since it does perform better in simulations. So, it is more the method of the proof of the properties for the MAP selector that is not precise enough rather than the selector itself.

Remark 13. In principle, we are interested in the smallest value from the set $\mathcal{N}(\tau)$ in order to select the model with the smallest dimension. So, we can correct our MAP selector slightly to incorporate this requirement. For some positive sequence $\delta = \delta_n$, let

$$\hat{N}_\delta = \min\bigl\{k \in \mathbb{N}\colon P(N=k \mid X) \geq \delta_n\bigr\}$$

with the convention that $\min\{\emptyset\} = \infty$. Then we can take the selector

$$\tilde{N} = \min\{\widehat{N}_{\text{MAP}}, \hat{N}_\delta\}.$$

However, it is not clear what would be an appropriate choice for the sequence δ_n. From the proof of the above theorem, it seems that the sequence δ_n should be of the order $\bigl(1+\sqrt{\tau n\mathcal{R}_n(N_o,\theta_0)}\bigr)^2$. The oracle risk is, however, unknown, so in practice one should take its reasonable data driven empirical analog, say, $\hat{\delta}_n = c\widehat{N}_{\text{MAP}}^{-1}$.

Remark 14. Suppose a family of estimators $\widehat{\Theta}(\mathcal{N})$ covers some scale $\{\Theta_\beta,\ \beta \in \mathcal{B}\}$ in the sense we discussed in the Introduction, so that the estimator $\widehat{\theta}(N_\beta)$ is minimax over Θ_β. Then we can relate the above model selection problem to the problem of selecting the smoothness parameter $\beta \in \mathcal{B}$ in case we assume that the signal θ is from the scale $\{\Theta_\beta,\ \beta \in \mathcal{B}\}$. Suppose the above correspondence $\beta \to N_\beta$ can be inverted: $N \to \beta_N$, $N \in \mathcal{N}$. Namely, for a given scale of models $\{\Theta_\beta, \beta \in \mathcal{B}\}$ parametrized by $\mathcal{B}$, we select a model $\widehat{\beta} = \beta_{\widehat{N}}$ with $\widehat{N}$ constructed by using $P(N \mid X)$. If our selector $\widehat{N}$ satisfies (3), the model $\widehat{\beta}$ is then close to the best model in terms of the risk function $R(\widehat{\theta}(N_\beta),\theta)$, i.e., the oracle model $\beta_o = \beta_{N_o}$ for which $\inf_{\beta\in\mathcal{B}} R(\widehat{\theta}(N_\beta),\theta) = \inf_{N\in\mathcal{N}} R(\widehat{\theta}(N),\theta) = R(\widehat{\theta}(N_o),\theta)$.

For a example, in case of the Sobolev ellipsoids scale described in the next section $N_\beta = \lfloor cn^{1/(2\beta+1)} \rfloor$, so that $\widehat{\beta} = \frac{\log n}{2\log(\widehat{N}/c)} - \frac{1}{2}$ can be taken to be a smoothness selector.

4. POSTERIOR RATE: MINIMAX VERSUS ORACLE

In this section we discuss the relation between the oracle over the family $\widehat{\Theta}(\mathcal{N})$ and minimax over the scale $\{\Theta_\beta,\ \beta\in\mathcal{B}\}$ approaches to the optimality. The two main factors are the family of estimators $\widehat{\Theta}(\mathcal{N})$ and the scale $\{\Theta_\beta,\ \beta\in\mathcal{B}\}$. Other important factors are whether there exists an estimator $\widehat{\theta}$ satisfying the oracle inequality (3) and how large the set Θ_0 is for which this oracle inequality holds.

As was already mentioned in the Introduction, to make the notion of oracle estimator sensible, the family $\widehat{\Theta}(\mathcal{N})$ should neither be too poor, nor too rich. This is also discussed in [10]. Indeed, on the one hand, we want this family to contain some "good" estimators about which we know that they perform well over nonparametric classes Θ_β, $\beta\in\mathcal{B}$. Suppose the family $\widehat{\Theta}(\mathcal{N})$ is sufficiently rich to contain the minimax estimators over all Θ_β, $\beta\in\mathcal{B}$, i.e., for any $\beta\in\mathcal{B}$, there exists an $N_\beta\in\mathcal{N}$ such that $\widehat{\theta}(N_\beta)$ is minimax over the class Θ_β (the family $\widehat{\Theta}(\mathcal{N})$ covers the scale $\{\Theta_\beta,\ \beta\in\mathcal{B}\}$):

$$\sup_{\theta\in\Theta_\beta} R_n(\widehat{\theta}(N_\beta),\theta)\le C'_n r_n(\Theta_\beta),$$

ideally with $C'_n=1+o(1)$ as $n\to\infty$, otherwise with some uniform constant $1\le C'_n=C'<\infty$. This implies that $\inf_{N\in\mathcal{N}}\sup_{\theta\in\Theta_\beta} R(\widehat{\theta}(N),\theta)\le C'_n r(\Theta_\beta)$ for all $\beta\in\mathcal{B}$. Then certainly the oracle approach is stronger as the oracle risk at each point $\theta\in\Theta_\beta$ can only be smaller than a multiple of the minimax risk $r_n(\Theta_\beta)$.

On the other hand, if the family $\widehat{\Theta}(\mathcal{N})$ is too rich, then it may not be possible to find an estimator $\widehat{\theta}$ satisfying the relation (3) for a reasonable Θ_0. Instead, the relation (3) may hold only for a "thin" set Θ_0, while we would certainly like Θ_0 to be as big as possible, ideally containing all Θ_β's. However, it is enough to assume the set Θ_0 to be sufficiently "massive" to contain at least all the worst representatives from Θ_β in the following sense:

$$\sup_{\theta\in\Theta_\beta} R(\widehat{\theta},\theta)\le C_n\sup_{\theta\in\Theta_\beta}\inf_{N\in\mathcal{N}} R(\widehat{\theta}(N),\theta).$$

Then, combining these relations, we derive the minimax adaptivity of estimator $\widehat{\theta}$:

$$\sup_{\theta\in\Theta_\beta} R(\widehat{\theta},\theta)\le C_n\sup_{\theta\in\Theta_\beta}\inf_{N\in\mathcal{N}} R(\widehat{\theta}(N),\theta)\le C_n\inf_{N\in\mathcal{N}}\sup_{\theta\in\Theta_\beta} R(\widehat{\theta}(N),\theta)\le C_nC'_n r(\Theta_\beta).$$

The same reasoning applies to the results about posterior convergence rate. If the family $\widehat{\Theta}(\mathcal{N})$ covers the scale $\{\Theta_\beta,\beta\in\mathcal{B}\}$, then the results on oracle posterior convergence rate of the form (5) are stronger than the results on adaptive minimax convergence rate of the posterior simply because the oracle rate can only be smaller or the same up to a multiple than the corresponding minimax risk over Θ_β, at least for all $\theta_0\in\Theta_0\cap\big(\cup_{\beta\in\mathcal{B}}\Theta_\beta\big)$. Recall that (5) holds for $\theta_0\in\Theta_0$, so it is desirable that $\cup_{\beta\in\mathcal{B}}\Theta_\beta\subseteq\Theta_0$. In our case, $\widehat{\Theta}(\mathcal{N})$ is the class of projection estimators and $\Theta_0=\ell_2$, so that our results on oracle projection posterior convergence are always stronger than all the adaptation results about the posterior convergence with the minimax rate, simultaneously over all the scales where the corresponding minimax rates are attained by projection estimators. Besides, the uniformity of all the results on adaptive minimax convergence rate of the posterior over all these scales follows immediately from the uniformity of our results over the whole space ℓ_2.

Summarizing, once we establish that the projection estimators family $\widehat{\Theta}(\mathcal{N})$ defined by (6) covers the scale $\{\Theta_\beta,\beta\in\mathcal{B}\}$, i.e., the minimax rate over Θ_β for each $\beta\in\mathcal{B}$ is attained by a projection estimator from $\widehat{\Theta}(\mathcal{N})$, then, by applying Theorem 2, we immediately derive the following result.

Theorem 5. *Let the class of projection estimators $\widehat{\Theta}(\mathcal{N})$ defined by (6) cover the scale $\{\Theta_\beta,\beta\in\mathcal{B}\}$, let $r_n(\Theta_\beta)=\inf_{\widehat{\theta}}\sup_{\theta\in\Theta_\beta} R_n(\widehat{\theta},\theta)$ be the minimax risk over the class Θ_β and let the prior π_α be*

defined by (13) *with* $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$. *Then there exist constants* $C_1, C_2 > 0$ *depending only on* α *and* Θ_β *such that for any* $M > 0$,

$$E_{\theta_0} P\Big\{ \big\|\theta - \theta_0/2\big\|^2 \ge M r_n(\Theta_\beta) \mid X \Big\} \le \frac{C_1}{M}, \quad E_{\theta_0} P'\Big\{ \big\|\theta - \theta_0\big\|^2 \ge M r_n(\Theta_\beta) \mid X \Big\} \le \frac{C_2}{M},$$

uniformly in $\theta_0 \in \Theta_\beta$, *where the posteriors* $P(\theta \mid X)$ *and* $P'(\theta \mid X)$ *are defined by* (14) *and* (15) *respectively.*

As a consequence we derive that the posterior $P'(\theta \mid X)$ and the appropriately shifted posterior $P(\theta \mid X)$ will concentrate about θ_0 uniformly over $\theta_0 \in \Theta_\beta$.

Corollary 2. *Under the conditions of Corollary* 1, *for any* $M_n > 0$ *such that* $M_n \to \infty$

$$P'\Big\{ \big\|\theta - \theta_0\big\|^2 \ge M_n r_n(\Theta_\beta) \mid X \Big\} \to 0 \quad \textit{and} \quad P\Big\{ \big\|\theta - \theta_0 + \widehat{\theta}/2\big\|^2 \ge M_n r_n(\Theta_\beta) \mid X \Big\} \to 0$$

as $n \to \infty$ *in* P_{θ_0}*-probability, uniformly in* $\theta_0 \in \Theta_\beta$.

Consider a couple of examples of nonparametric scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ for which the minimax rate is attained by a projection estimator. Denote $\lfloor a \rfloor = \max\{k \in \mathbb{Z},\, k \le a\}$ for $a \in \mathbb{R}$.

Sobolev ellipsoids. Consider the Sobolev ellipsoids $\Theta_\beta(Q) = \{\theta\colon \sum_{i=1}^\infty i^{2\beta}\theta_i^2 \le Q\}$, $\beta > 0$. In [2] a result on the posterior convergence with the minimax rate for the Sobolev ellipsoid $\Theta_\beta(Q)$ is given. It is well known that the corresponding minimax rate is $r_n(\beta) = n^{-2\beta/(2\beta+1)}$; see, for example, [16] or [3]. The uniformity of the main claim in [2] was obtained only for sufficiently "small" ellipsoids, although it could have been established for the original ellipsoid $\Theta_\beta(Q)$ as well with somewhat more careful analysis. The prior on θ in [2] was different and based on putting a joint prior on the pair (θ, β) rather than on (θ, N), to model the unknown smoothness parameter β. However, it is easy to see that the projection estimator $\widehat{\theta}(N_\beta)$ defined by (6), with $N_\beta = \lfloor cn^{1/(2\beta+1)} \rfloor$, is minimax with respect to the convergence rate:

$$\begin{aligned} \sup_{\theta \in \Theta_\beta(Q)} R(\widehat{\theta}(N_\beta), \theta) &= \sup_{\theta \in \Theta_\beta(Q)} \Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^\infty \theta_i^2 \Big) \le \frac{N_\beta}{n} + \sup_{\theta \in \Theta_\beta(Q)} \Big(\sum_{i=N_\beta+1}^\infty \frac{\theta_i^2 i^{2\beta}}{N_\beta^{2\beta}} \Big) \\ &\le \frac{N_\beta}{n} + \frac{Q}{N^{2\beta}} = C n^{-2\beta/(2\beta+1)}. \end{aligned}$$

Then Theorem 5 implies that the appropriately shifted posterior distribution $P(\theta \mid X)$ and the posterior $P'(\theta \mid X)$ both concentrate in P_{θ_0}-probability about the true signal θ_0 with at least the minimax rate $r_n(\Theta_\beta(Q)) = n^{-2\beta/(2\beta+1)}$ uniformly in $\theta_0 \in \Theta_\beta(Q)$. Thus our (oracle posterior convergence rate) results are stronger than those (adaptive minimax posterior convergence rate) in [2].

Exponential ellipsoids. Suppose our nonparametric class $\Theta_\beta(Q)$ is from the scale of the exponential ellipsoids: $\Theta_\beta(Q) = \{\theta\colon \sum_{k=1}^\infty e^{2\beta k}\theta_k^2 \le Q\}$, $\beta > 0$. By the same arguments we derive that the projection estimator $\widehat{\theta}(N)$ defined by (6), with $N_\beta = \lfloor \log n/(2\beta) \rfloor$, is minimax with respect to the convergence rate for the exponential ellipsoid $\Theta_\beta(Q)$. Indeed, in this case

$$\begin{aligned} \sup_{\theta \in \Theta_\beta(Q)} R(\widehat{\theta}(N_\beta), \theta) &= \sup_{\theta \in \Theta_\beta(Q)} \Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^\infty \theta_i^2 \Big) \le \frac{N_\beta}{n} + \sup_{\theta \in \Theta_\beta(Q)} \Big(\sum_{k=N_\beta+1}^\infty \frac{\theta_k^2 e^{2\beta k}}{e^{2\beta N_\beta}} \Big) \\ &\le \frac{\log n}{2\beta n} + \frac{Q}{n} \le \frac{C \log n}{n}, \end{aligned}$$

which is of the same order as the minimax rate $r_n(\Theta_\beta(Q)) = \log n/n$ over the exponential ellipsoid $\Theta_\beta(Q)$, cf. [3]. Thus, according to Theorem 5, the appropriately shifted posterior distribution $P(\theta \mid X)$ and the posterior $P'(\theta \mid X)$ both concentrate in P_{θ_0}-probability about the true signal θ_0 with at least the minimax rate $r_n(\Theta_\beta(Q)) = \log n/n$ uniformly over $\theta_0 \in \Theta_\beta(Q)$.

Hyperrectangles. Consider the so-called hyperrectangles in ℓ_2:

$$\Theta_\beta(Q) = \Big\{\theta\colon |\theta_k| \le \sqrt{Q}k^{-\beta},\, k\in\mathbb{N}\Big\}, \quad \beta > 1/2.$$

It is not difficult to show that the minimax convergence rate over this class is $r_n(\Theta_\beta(Q)) = n^{-(2\beta-1)/(2\beta)}$. We derive that the projection estimator $\widehat{\theta}(N_\beta)$ defined by (6), with $N_\beta = \lfloor cn^{1/(2\beta)}\rfloor$, is minimax with respect to the convergence rate for the hyperrectangle $\Theta_\beta(Q)$. Indeed, in this case

$$\begin{aligned}\sup_{\theta\in\Theta_\beta(Q)} R(\widehat{\theta}(N_\beta),\theta) &= \sup_{\theta\in\Theta_\beta(Q)}\Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^{\infty}\theta_i^2\Big) \le \frac{N_\beta}{n} + \sum_{k=N_\beta+1}^{\infty}\frac{Q}{k^{2\beta}}\\ &\le \frac{N_\beta}{n} + \frac{Q}{(2\beta-1)N_\beta^{2\beta-1}} \le Cn^{-(2\beta-1)/(2\beta)},\end{aligned}$$

which is of the same order as the minimax rate over the hyperrectangle $\Theta_\beta(Q)$ $r_n(\beta) = n^{-(2\beta-1)/(2\beta)}$. Theorem 5 implies that the appropriately shifted posterior distribution $P(\theta\mid X)$ and the posterior $P'(\theta\mid X)$ both concentrate in P_{θ_0}-probability about the true signal θ_0 with at least the minimax rate $r_n(\Theta_\beta(Q)) = n^{-(2\beta-1)/(2\beta)}$ uniformly in $\theta_0\in\Theta_\beta(Q)$.

Tail classes. Finally consider the so-called tail classes:

$$\Theta_\beta(Q) = \Big\{\theta\colon \sum_{k=m}^{\infty}\theta_k^2 \le Qm^{-\beta},\, m\in\mathbb{N}\Big\}, \quad \beta > 0.$$

Since hyperrectangle with parameters $\beta' > 1/2$ and Q' can be embedded into a tail class with parameters $\beta' = 2\beta - 1$ and some $Q = Q(\beta, Q')$, we obtain that the minimax risk over the tail class $\Theta_\beta(Q)$ is at least $r_n(\beta) = n^{-\beta/(1+\beta)}$. Now we derive that the projection estimator $\widehat{\theta}(N)$ defined by (6), with $N_\beta = \lfloor cn^{1/(1+\beta)}\rfloor$, is minimax with respect to the convergence rate for the tail class $\Theta_\beta(Q)$. Indeed, in this case

$$\sup_{\theta\in\Theta_\beta(Q)} R(\widehat{\theta}(N_\beta),\theta) = \sup_{\theta\in\Theta_\beta(Q)}\Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^{\infty}\theta_i^2\Big) \le \frac{N_\beta}{n} + \frac{Q}{(N_\beta+1)^\beta} \le Cn^{-\beta/(1+\beta)},$$

which is of the same order as the minimax rate $r_n(\beta) = n^{-\beta/(1+\beta)}$ over the tail class $\Theta_\beta(Q)$. Therefore Theorem 5 implies that the appropriately shifted posterior distribution $P(\theta\mid X)$ and the posterior $P'(\theta\mid X)$ both concentrate in P_{θ_0}-probability about the true signal θ_0 with at least the minimax rate $r_n(\Theta_\beta(Q)) = n^{-\beta/(1+\beta)}$ uniformly in $\theta_0\in\Theta_\beta(Q)$.

5. TECHNICAL RESULTS

In this section we provide a couple of technical lemmas used in the proofs of the main results.

Lemma 3. *Let $\theta_0\in\ell_2$, $\tau > 0$ and let $\widehat{\theta}(k)$, $k\in\mathbb{N}$, be defined by* (6)*, $\widehat{N}$ by* (17) *and the set $\mathcal{N}(\tau)$ by* (21)*. Then*

$$\begin{aligned}E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 I\big\{\widehat{N}\in\mathcal{N}(\tau)\big\}\Big] &\le 2\tau\mathcal{R}_n(N_o,\theta_0),\\ E_{\theta_0}\Big[\sum_{k\in\mathcal{N}(\tau)}\|\widehat{\theta}(k) - \theta_0\|^2 P(N=k\mid X)\Big] &\le 2\tau\mathcal{R}_n(N_o,\theta_0).\end{aligned}$$

Proof. Recall that $\xi_i = \sqrt{n}(X_i - \theta_0) \overset{ind}{\sim} \mathcal{N}(0,1)$, $i\in\mathbb{N}$, under $X\sim P_{\theta_0}$. Write

$$E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 I\big\{\widehat{N}\in\mathcal{N}(\tau)\big\}\Big] = E_{\theta_0}\Big[\Big(\sum_{i=1}^{\widehat{N}}\frac{\xi_i^2}{n} + \sum_{i=\widehat{N}+1}^{\infty}\theta_{0i}^2\Big) I\big\{\widehat{N}\in\mathcal{N}(\tau)\big\}\Big]$$

$$= E_{\theta_0}\Big[\Big(\sum_{i=1}^{\widehat{N}}\frac{\xi_i^2}{n}\Big)I\big\{\widehat{N}\in\mathcal{N}(\tau)\big\}\Big] + E_{\theta_0}\Big[\Big(\sum_{i=\widehat{N}+1}^{\infty}\theta_{0i}^2\Big)I\big\{\widehat{N}\in\mathcal{N}(\tau)\big\}\Big]. \tag{25}$$

Now we bound the both terms in the right-hand side of (25) by $\tau\mathcal{R}_n(N_o,\theta_0)$. Since $\mathcal{R}_n(k,\theta_0) = \frac{k}{n} + \sum_{i=k+1}^{\infty}\theta_{0i}^2 \le \tau\mathcal{R}_n(N_o,\theta_0)$ for any $k\in\mathcal{N}(\tau)$, we obtain

$$N_{\max} = \max\{\mathcal{N}(\tau)\} \le \tau n\mathcal{R}_n(N_o,\theta_0).$$

This implies the bound for the first term in the right-hand side of (25):

$$E_{\theta_0}\Big[\Big(\sum_{i=1}^{\widehat{N}}\frac{\xi_i^2}{n}\Big)I\big\{\widehat{N}\in\mathcal{N}(\tau)\big\}\Big] \le E_{\theta_0}\sum_{i=1}^{N_{\max}}\frac{\xi_i^2}{n} = \frac{N_{\max}}{n} \le \tau\mathcal{R}_n(N_o,\theta_0).$$

Finally, we evaluate the second term in the right-hand side of (25) as follows:

$$E_{\theta_0}\Big[\Big(\sum_{i=\widehat{N}+1}^{\infty}\theta_{0i}^2\Big)I\big\{\widehat{N}\in\mathcal{N}(\tau)\big\}\Big] = E_{\theta_0}\Big[\sum_{k\in\mathcal{N}(\tau)}\Big(\sum_{i=k+1}^{\infty}\theta_{0i}^2\Big)I\{\widehat{N}=k\}\Big]$$

$$= \sum_{k\in\mathcal{N}(\tau)}\Big(\sum_{i=k+1}^{\infty}\theta_{0i}^2\Big)E_{\theta_0}I\{\widehat{N}=k\} \le \sum_{k\in\mathcal{N}(\tau)}\mathcal{R}_n(k,\theta_0)P_{\theta_0}\{\widehat{N}=k\} \le \tau\mathcal{R}_n(N_o,\theta_0),$$

since $\mathcal{R}_n(k,\theta_0)\le\tau\mathcal{R}_n(N_o,\theta_0)$ for all $k\in\mathcal{N}(\tau)$. Combining the last two relations with (25) completes the proof of the first inequality. The second inequality follows similarly. □

Lemma 4. *Let $\widehat{\theta}(k)$, $k\in\mathbb{N}$, be defined by* (6) *and $\widehat{N}$ by* (17). *Then for any $\theta_0\in\ell_2$ and any $k\in\mathbb{N}$,*

$$E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N})-\theta_0\|^2 I\big\{\widehat{N}=k\big\}\Big] \le \Big(\mathcal{R}_n(k,\theta_0)+\frac{1}{n}\Big)\big(E_{\theta_0}P(N=k\mid X)\big)^{1/2},$$

$$E_{\theta_0}\Big[\|\widehat{\theta}(k)-\theta_0\|^2 P(N=k\mid X)\Big] \le \Big(\mathcal{R}_n(k,\theta_0)+\frac{1}{n}\Big)\big(E_{\theta_0}P(N=k\mid X)\big)^{1/2}.$$

Proof. The following fact is well known. Let $Z_1,\dots,Z_k$ be iid $\mathcal{N}(0,\sigma^2)$ random variables, then

$$\Big[E\Big(\sum_{i=1}^{k}Z_i^2\Big)^2\Big]^{\frac12} = \Big[E\Big(\sum_{i=1}^{k}Z_i^4+\sum_{i\ne j}Z_i^2Z_j^2\Big)\Big]^{\frac12} = \big(3k\sigma^4+(k^2-k)\sigma^4\big)^{1/2} \le (k+1)\sigma^2.$$

Applying this fact and the Cauchy–Schwarz inequality, we evaluate

$$E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N})-\theta_0\|^2 I\big\{\widehat{N}=k\big\}\Big] = E_{\theta_0}\Big[\sum_{i=1}^{\infty}(\widehat{\theta}_i(k)-\theta_{0i})^2 I\big\{\widehat{N}=k\big\}\Big]$$

$$= E_{\theta_0}\Big[\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}+\sum_{i=k+1}^{\infty}\theta_{0i}^2\Big)I\big\{\widehat{N}=k\big\}\Big]$$

$$\le \Big[E_{\theta_0}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)^2\Big]^{1/2}\Big[E_{\theta_0}I\big\{\widehat{N}=k\big\}\Big]^{1/2} + \Big[\sum_{i=k+1}^{\infty}\theta_{0i}^2\Big]E_{\theta_0}I\big\{\widehat{N}=k\big\}$$

$$\le \frac{k+1}{n}\Big[E_{\theta_0}P(N=k\mid X)\Big]^{1/2} + \Big[\sum_{i=k+1}^{\infty}\theta_{0i}^2\Big]E_{\theta_0}P(N=k\mid X) \tag{26}$$

$$\le \Big(\mathcal{R}_n(k,\theta_0)+\frac{1}{n}\Big)\big(E_{\theta_0}P(N=k\mid X)\big)^{1/2},$$

since $\widehat{N}$ is generated from the conditional distribution $P(N\mid X)$. The second inequality follows similarly. □

Remark 15. In fact, we proved a slightly stronger assertion (26) as compared to the final statement of the above lemma. Besides, summing up over all values of $\widehat{N} = k$, $k \in \mathbb{N}$, we derive that, for any $\theta_0 \in \ell_2$,

$$E_{\theta_0}\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 \le \sum_{k\in\mathbb{N}} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\big(E_{\theta_0} P(\widehat{N} = k \mid X)\big)^{1/2}.$$

Recall our Bayesian scheme:

$$X \mid (\theta, N) \sim P_\theta, \quad \theta \mid N \sim \pi_N, \quad N \sim \lambda.$$

Two-level hierarchical prior on (θ, N) leads to a joint distribution on (X, θ, N), which in turn gives rise to the conditional marginal $P(X \mid N)$ and the posterior distributions $P(\theta \mid X)$ and $P(N \mid X)$, at which we can look from the perspective of P_{θ_0}-measure of X. Since $P(X \mid \theta, N)$, $P(\theta \mid N)$ are both the products of normals, it is easy to derive due to the conjugacy that the conditional marginal $P(X \mid N)$ is also the product of normals $\mathcal{N}(0, \tau_i^2(N) + n^{-1})$.

Define

$$a_i(k) = (\tau_i^2(k) + n^{-1})^{-1}, \quad a_i(k, k_0) = a_i(k) - a_i(k_0), \quad k, k_0, i \in \mathbb{N}, \tag{27}$$

where $\tau_i^2(k)$ is defined by (9). The following lemma concerns the posterior distribution $P(N \mid X)$.

Lemma 5. *For any $k, k_0 \in \mathbb{N}$ and any $\theta_0 \in \ell^2$,*

$$E_{\theta_0} P(N = k \mid X) \le \frac{\lambda_k}{\lambda_{k_0}} \prod_{i=1}^{\infty} \Big[\frac{a_i(k)}{a_i(k_0)(1 + n^{-1}a_i(k, k_0))}\Big]^{1/2} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{\infty} \frac{a_i(k, k_0)\theta_{0i}^2}{1 + n^{-1}a_i(k, k_0)}\Big\},$$

where $a_i(k)$ and $a_i(k, k_0)$ are defined by (27).

Proof. Applying the martingale convergence theorem and the dominated convergence theorem in the same manner as in the proof of Lemma 2 of [2], we have that

$$E_{\theta_0} P(N = k \mid X) = \lim_{m\to\infty} E_{\theta_0} P(N = k \mid X_1, \ldots, X_m). \tag{28}$$

Since the conditional marginal $P(X_1, \ldots, X_m \mid N = k)$ is the product of normal distributions $\mathcal{N}(0, \tau_i^2(k) + n^{-1})$, $i = 1, \ldots, m$, with densities $f_i(x_i \mid N = k)$ respectively, it is not difficult to compute the posterior probability

$$\begin{aligned} P(N = k \mid X_1, \ldots, X_m) &= \frac{\prod_{i=1}^m f_i(X_i \mid N = k) P(N = k)}{\sum_{l\in\mathbb{N}} \prod_{i=1}^m f_i(X_i \mid N = l) P(N = l)} \\ &= \frac{\lambda_k \prod_{i=1}^m \frac{1}{\sqrt{(\tau_i^2(k) + n^{-1})}} \exp\Big\{ -\frac{X_i^2}{2(\tau_i^2(k) + n^{-1})}\Big\}}{\sum_{l\in\mathbb{N}} \lambda_l \prod_{i=1}^m \frac{1}{\sqrt{(\tau_i^2(l) + n^{-1})}} \exp\Big\{ -\frac{X_i^2}{2(\tau_i^2(l) + n^{-1})}\Big\}}. \end{aligned}$$

Obviously

$$E_{\theta_0} P(N = k \mid X_1, X_2, \ldots, X_m) \le \frac{\lambda_k}{\lambda_{k_0}} \prod_{i=1}^m \sqrt{\frac{a_i(k)}{a_i(k_0)}}\; E_{\theta_0} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^m X_i^2 a_i(k, k_0)\Big\}.$$

Using this and the elementary identity

$$E\Big(\exp\Big[-\frac{b}{2}Y^2\Big]\Big) = \frac{1}{\sqrt{1 + b\sigma^2}} \exp\Big[-\frac{\mu^2 b}{2(1 + b\sigma^2)}\Big]$$

for $Y \sim \mathcal{N}(\mu, \sigma^2)$ and $b > -\sigma^{-2}$ (under P_{θ_0}, $X_i \overset{ind}{\sim} \mathcal{N}(\theta_{0i}, n^{-1})$, $i \in \mathbb{N}$), we derive

$$E_{\theta_0} P(N = k \mid X_1, X_2, \ldots, X_m)$$
$$\leq \frac{\lambda_k}{\lambda_{k_0}} \prod_{i=1}^{m} \left[\frac{a_i(k)}{a_i(k_0)(1 + n^{-1} a_i(k, k_0))} \right]^{1/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{m} \frac{a_i(k, k_0)\theta_{0i}^2}{1 + n^{-1} a_i(k, k_0)} \right\}.$$

Combining the last relation with (28) completes the proof of the lemma. □

Note that the above lemma holds for any prior variances $\tau_i^2(k)$'s and any $k_0 \in \mathbb{N}$ including $k_0 = N_o$, the oracle cut-off defined by (8). Taking $k_0 = N_o$ and $\tau_i^2(k)$ defined by (9), we obtain the following corollary.

Corollary 3. *Let $\tau_i^2(k)$, $k \in \mathbb{N}$, be defined by* (9).

- *If $k < N_o$, then*

$$E_{\theta_0}\big[P(N = k \mid X)\big] \leq \frac{\lambda_k}{\lambda_{N_o}} \exp\left\{ -\frac{n}{6}\Big(\sum_{i=k+1}^{N_o} \theta_{0i}^2 - \frac{6 \log\big(\frac{2}{\sqrt{3}}\big)(N_o - k)}{n} \Big) \right\}.$$

- *If $k > N_o$, then*

$$E_{\theta_0}\big[P(N = k \mid X)\big] \leq \frac{\lambda_k}{\lambda_{N_o}} \exp\left\{ \frac{n}{2} \sum_{i=N_o+1}^{k} \theta_{0i}^2 \right\}.$$

Proof. Indeed, using relations (27) and (9), we compute

$$a_i(k) = \frac{n}{2} I\{i \leq k\} + n I\{i > k\}, \quad a_i(k, k_0) = -\frac{n}{2} I\{k_0 < i \leq k\} + \frac{n}{2} I\{k < i \leq k_0\}$$

and substitute these values in the right-hand side of the inequality of Lemma 5. □

Notice that in case $k = N_o$ we obtain a trivial useless bound $E_{\theta_0}\big[P(N = N_o \mid X)\big] \leq 1$.

Lemma 6. *Let the prior π_α be defined by* (13) *with $\alpha \in [\frac{1}{6} - \log\big(\frac{2}{\sqrt{3}}\big), \frac{1}{2}]$, the projection risk $\mathcal{R}_n(k, \theta)$ be defined by* (7) *and the oracle cut-off N_o be defined by* (8).

- *For any $\tau_1 \geq 6 \log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha \geq 1$ and any $k < N_o$,*

$$E_{\theta_0}\big[P(N = k \mid X)\big] \leq \exp\Big\{ -\frac{n}{6}\big(\mathcal{R}_n(k, \theta_0) - \tau_1 \mathcal{R}_n(N_o, \theta_0) \big) \Big\}.$$

- *For any $\tau_2 \geq \frac{1}{2\alpha} \geq 1$ and any $k > N_o$,*

$$E_{\theta_0}\big[P(N = k \mid X)\big] \leq \exp\big\{ -n\alpha\big(\mathcal{R}_n(k, \theta_0) - \tau_2 \mathcal{R}_n(N_o, \theta_0) \big) \big\}.$$

Proof. To prove the lemma, we use $\frac{\lambda_k}{\lambda_{N_o}} = e^{-\alpha(k - N_o)}$ and Corollary 3. For $k < N_o$,

$$\begin{aligned} E_{\theta_0} P(N = k \mid X) &\leq \exp\left\{ -\frac{n}{6}\Big(\sum_{i=k+1}^{N_o} \theta_{0i}^2 - \frac{(6 \log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha)(N_o - k)}{n} \Big) \right\} \\ &\leq \exp\left\{ -\frac{n}{6}\Big(\sum_{i=k+1}^{\infty} \theta_{0i}^2 - \tau_1 \sum_{i=N_o+1}^{\infty} \theta_{0i}^2 - \frac{\tau_1 N_o - k}{n} \Big) \right\} \\ &= \exp\left\{ -\frac{n}{6}\big(\mathcal{R}_n(k, \theta_0) - \tau_1 \mathcal{R}_n(N_o, \theta_0) \big) \right\} \end{aligned}$$

since $\tau_1 \geq 6\log\left(\frac{2}{\sqrt{3}}\right) + 6\alpha \geq 1$. Similarly, for $k > N_o$, we have that

$$\begin{aligned} E_{\theta_0} P(N = k \mid X) &\leq \exp\left\{ -n\alpha\Big(-\frac{1}{2\alpha} \sum_{i=N_o+1}^{k} \theta_{0i}^2 + \frac{(k-N_o)}{n} \Big) \right\} \\ &\leq \exp\left\{ -n\alpha\Big(\sum_{i=k+1}^{\infty} \theta_{0i}^2 - \tau_2 \sum_{i=N_o+1}^{\infty} \theta_{0i}^2 + \frac{k-\tau_2 N_o}{n} \Big) \right\} \\ &= \exp\left\{ -n\alpha\big(\mathcal{R}_n(k,\theta_0) - \tau_2 \mathcal{R}_n(N_o,\theta_0) \big) \right\} \end{aligned}$$

because $\tau_2 \geq \frac{1}{2\alpha} \geq 1$. □

Remark 16. The above exponential inequalities describe essentially a "correct" frequentist behavior of the posterior distribution $P(N \mid X)$. The bigger the difference is between the risk $\mathcal{R}_n(k,\theta_0)$ at point k and the oracle risk $\mathcal{R}_n(N_o,\theta_0)$, the smaller the exponential bound for $E_{\theta_0} P(N = k \mid X)$ becomes.

For brevity sake, denote

$$R_-(\theta_0,\beta,\tau,\tau_1,n) = \sum_{k\in\mathcal{N}^-(\tau)} n\mathcal{R}_n(k,\theta_0) \exp\left\{-\beta n\big(\mathcal{R}_n(k,\theta_0) - \tau_1\mathcal{R}_n(N_o,\theta_0)\big)\right\}, \tag{29}$$

$$R_+(\theta_0,\beta,\tau,\tau_2,n) = \sum_{k\in\mathcal{N}^+(\tau)} n\mathcal{R}_n(k,\theta_0) \exp\left\{-\beta n\big(\mathcal{R}_n(k,\theta_0) - \tau_2\mathcal{R}_n(N_o,\theta_0)\big)\right\}, \tag{30}$$

where the projection risk $\mathcal{R}_n(k,\theta)$ is defined by (7), the oracle cut-off $N_o = N_o(\theta_0)$ is defined by (8), the sets $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ are defined by (22) and (23) respectively.

Lemma 7. *Let the quantities $R_-(\theta_0,\beta,\tau,\tau_1,n)$ and $R_+(\theta_0,\beta,\tau,\tau_2,n)$ be defined by* (29) *and* (30) *respectively.*

(i) *For any $\tau > \tau_1 > 0$, $\beta > 0$, the following inequality holds:*

$$R_-(\theta_0,\beta,\tau,\tau_1,n) \leq \min\left\{\frac{c_1}{B^2}, \frac{\big(n\mathcal{R}_n(N_o,\theta_0)\big)^2}{e^{Bn\mathcal{R}_n(N_o,\theta_0)}}\right\},$$

where $B = B(\tau,\tau_1,\beta) = \beta(\tau-\tau_1)/\tau$, $c_1 = 4e^{-2}$.

(ii) *For any $\tau > \tau_2 > 0$, $\beta > 0$, the following inequality holds:*

$$R_+(\theta_0,\beta,\tau,\tau_2,n) \leq \min\left\{g(D), \frac{c_3\big(n\mathcal{R}_n(N_o,\theta_0)\big)^2}{e^{Dn\mathcal{R}_n(N_o,\theta_0)}}\right\} \leq \min\left\{\frac{c_2}{D^2}, \frac{c_3\big(n\mathcal{R}_n(N_o,\theta_0)\big)^2}{e^{Dn\mathcal{R}_n(N_o,\theta_0)}}\right\},$$

where $g(u) = e^{-1}u^{-2} + e^{-u}(e^u-1)^{-2}$, $D = D(\tau,\tau_2,\beta) = \beta(\tau-\tau_2)/\tau$, $c_2 = e^{-1}+1$ and $c_3 = (e+e^2)/4$.

Proof. Since $\tau > \tau_1$, we obtain for all $k \in \mathcal{N}^-(\tau)$ that

$$\begin{aligned} \mathcal{R}_n(k,\theta_0) - \tau_1\mathcal{R}_n(N_o,\theta_0) &= \frac{\tau_1}{\tau}(\mathcal{R}_n(k,\theta_0) - \tau\mathcal{R}_n(N_o,\theta_0)) + \Big(1-\frac{\tau_1}{\tau}\Big)\mathcal{R}_n(k,\theta_0) \\ &\geq \Big(1-\frac{\tau_1}{\tau}\Big)\mathcal{R}_n(k,\theta_0). \end{aligned} \tag{31}$$

Recall the notation $B = B(\tau,\tau_1,\beta) = \frac{\beta(\tau-\tau_1)}{\tau} > 0$ and let $a_k = a_k(\theta_0,n) = n\mathcal{R}_n(k,\theta_0)$. Let b be any constant such that $0 < b \leq B$. Certainly $a_k \geq a_{N_o} = n\mathcal{R}_n(N_o,\theta_0)$ by the definition of oracle. Then, using (31),

$$R_-(\theta_0,\beta,\tau,\tau_1,n) \leq \sum_{k\in\mathcal{N}^-(\tau)} n\mathcal{R}_n(k,\theta_0)\exp\{-Bn\mathcal{R}_n(k,\theta_0)\} = \sum_{k\in\mathcal{N}^-(\tau)} a_k \exp\{-Ba_k\}$$

$$\leq\ e^{-(B-b)a_{N_o}} \sum_{k\in\mathcal{N}^-(\tau)} a_k \exp\{-ba_k\}. \tag{32}$$

Consider separately two cases: $1 \leq N_o \leq b^{-1}$ and $N_o > b^{-1}$. First assume that $1 \leq N_o \leq b^{-1}$, then obviously

$$\sum_{k\in\mathcal{N}^-(\tau)} a_k \exp\{-ba_k\} \leq |\mathcal{N}^-(\tau)| \max_{x\geq 0}\{xe^{-bx}\} \leq N_o(be)^{-1} = b^{-2}e^{-1}. \tag{33}$$

Now suppose that $N_o > b^{-1}$. Let us look at the term $\sum_{k\in\mathcal{N}^-(\tau)} a_k \exp\{-ba_k\}$. First note that $a_k = n\mathcal{R}_n(k,\theta_0) \geq n\mathcal{R}_n(N_o,\theta_0) \geq N_o > b^{-1}$ for all $k \in \mathcal{N}^-(\tau)$. The function xe^{-bx}, $x \geq 0$, is increasing on the interval $[0, b^{-1}]$ and decreasing afterwards. This implies that $a_k \exp\{-ba_k\} \leq N_o \exp\{-bN_o\}$. Recall also that $|\mathcal{N}^-(\tau)| \leq N_o$. Using these relations, we derive the following bound:

$$\begin{aligned}\sum_{k\in\mathcal{N}^-(\tau)} a_k \exp\{-ba_k\}\ &\leq\ |\mathcal{N}^-(\tau)| N_o \exp\{-bN_o\} \leq N_o^2 \exp\{-bN_o\} \\ &\leq\ \max_{x\geq 0}\{x^2 e^{-bx}\} = 4(be)^{-2}.\end{aligned} \tag{34}$$

Combining (33) and (34) leads to

$$\sum_{k\in\mathcal{N}^-(\tau)} a_k \exp\{-ba_k\} \leq \max\left\{b^{-2}e^{-1}, 4(be)^{-2}\right\} = 4(be)^{-2},$$

which, together with (32), implies that for any $b \in (0, B]$

$$R_-(\theta_0,\beta,\tau,\tau_1,n) \leq 4e^{-2}e^{-(B-b)a_{N_o}}b^{-2}. \tag{35}$$

We have a family of upper bounds for different values of $b \in (0, B]$. Now we minimize the upper bound over $b \in (0, B]$. The upper bound in (35) is of the form $Ce^{a_{N_o}b}b^{-2}$, with constants $C = 4e^{-2}e^{-Ba_{N_o}}$, $a_{N_o} = n\mathcal{R}_n(N_o,\theta_0)$. The minimum of function $Ce^{a_{N_o}b}b^{-2}$ over $b \in (0, B]$ is attained at $\min\{B, 2/a_{N_o}\}$:

$$\min_{b\in(0,B]} Ce^{a_{N_o}b}b^{-2} = \min\left\{\frac{Ce^{a_{N_o}B}}{B^2}, \frac{Ce^2a_{N_o}^2}{4}\right\}. \tag{36}$$

Thus the corresponding sharpest upper bound becomes

$$\min_{b\in(0,B]} \left\{4e^{-2}e^{-(B-b)a_{N_o}}b^{-2}\right\} = \min\left\{4e^{-2}B^{-2}, \big(n\mathcal{R}_n(N_o,\theta_0)\big)^2 e^{-Bn\mathcal{R}_n(N_o,\theta_0)}\right\},$$

which, combined with (35), establishes part (i) of the lemma.

Let us prove part (ii) of the lemma. We derive in the same manner as in (31) that, for all $k \in \mathcal{N}^+(\tau)$,

$$\beta n\big(\mathcal{R}_n(k,\theta_0) - \tau_2\mathcal{R}_n(N_o,\theta_0)\big) \geq \beta n\Big(1 - \frac{\tau_2}{\tau}\Big)\mathcal{R}_n(k,\theta_0) = Da_k, \tag{37}$$

where $D = D(\tau,\tau_2,\beta) = \frac{(\tau-\tau_2)\beta}{\tau} > 0$ and $a_k = n\mathcal{R}_n(k,\theta_0)$ as before. Using (37), we obtain that for any $0 < d \leq D$

$$R_+(\theta_0,\beta,\tau,\tau_2,n) \leq \sum_{k\in\mathcal{N}^+(\tau)} a_k \exp\{-Da_k\} \leq e^{-(D-d)a_{N_0}} \sum_{k\in\mathcal{N}^+(\tau)} a_k \exp\{-da_k\}. \tag{38}$$

Next, introduce the sets

$$\begin{aligned}S_1(d) = S_1(d,\tau,\theta_0)\ &=\ \{k \in \mathcal{N}^+(\tau,\theta_0)\colon k \leq d^{-1}\}, \\ S_2(d) = S_2(d,\tau,\theta_0)\ &=\ \{k \in \mathcal{N}^+(\tau,\theta_0)\colon k > d^{-1}\},\end{aligned}$$

so that $\mathcal{N}^+(\tau) = S_1(d) \cup S_2(d)$. Obviously, $|S_1(d)| \le d^{-1}$. Besides, recall that $k \le a_k$ for all $k \in \mathbb{N}$ and the function xe^{-dx}, $d \ge 0$, is decreasing on $[d^{-1}, +\infty)$. Thus, as $a_k \ge k > d^{-1}$, $a_k \exp\{-da_k\} \le k \exp\{-dk\}$ for all $k \in S_2(d)$. Using this, we obtain that

$$\begin{aligned}
\sum_{k\in\mathcal{N}^+(\tau)} a_k \exp\{-da_k\} &\le \sum_{k\in S_1(d)} a_k \exp\{-da_k\} + \sum_{k\in S_2(d)} a_k \exp\{-da_k\} \\
&\le |S_1(d)| \max_{x\ge 0}\{xe^{-dx}\} + \sum_{k\in S_2(d)} k \exp\{-dk\} \\
&\le d^{-2}e^{-1} + \sum_{k=N_o+1}^{\infty} k \exp\{-dk\} = d^{-2}e^{-1} + \frac{e^{-d(N_o+1)}e^{-d}}{(1-e^{-d})^2} \\
&\le d^{-2}e^{-1} + \frac{e^{-3d}}{(1-e^{-d})^2} = \frac{e^{-1}}{d^2} + \frac{e^{-d}}{(e^d-1)^2} = g(d) \\
&\le \frac{e^{-1}}{d^2} + \frac{e^{-d}}{d^2} \le \frac{e^{-1}+1}{d^2}, \qquad (39)
\end{aligned}$$

which, combined with (38), implies that for any $0 < d \le D$

$$R_+(\theta_0, \beta, \tau, \tau_2, n) \le (e^{-1}+1)e^{-(D-d)a_{N_0}} d^{-2}. \qquad (40)$$

Again we have a family of upper bounds for different values of $d \in (0, D]$. Using (36) (this time with $C = (e^{-1}+1)e^{-Da_{N_o}}$), we minimize the upper bound over $d \in (0, D]$ in the same way as before:

$$\min_{d\in(0,D]} \left\{(e^{-1}+1)e^{-(D-d)a_{N_0}} d^{-2}\right\} = \min\left\{\frac{e^{-1}+1}{D^2}, \frac{e+e^2}{4}\big(n\mathcal{R}_n(N_o,\theta_0)\big)^2 e^{-Dn\mathcal{R}_n(N_o,\theta_0)}\right\}.$$

Besides, by taking $d = D = \frac{\beta(\tau-\tau_2)}{\tau}$ in (38) and (39), we also have a bound

$$R_+(\theta_0, \beta, \tau, \tau_2, n) \le \frac{e^{-1}}{D^2} + \frac{e^{-D}}{(e^D-1)^2} = g(D) \le \frac{e^{-1}+1}{D^2}.$$

Finally combine (40) with the last two relations to establish part (ii) of the lemma. □

Remark 17. The minimum in the right-hand sides of the inequalities of the above lemma is attained by the second term if $\min\{B, 2/a_{N_o}\} = 2/a_{N_o}$ and $\min\{D, 2/a_{N_o}\} = 2/a_{N_o}$, i.e., the oracle risk is sufficiently large $a_{N_o} = n\mathcal{R}_n(N_o,\theta_0) \ge \max\{2/B, 2/D\}$ or $\mathcal{R}_n(N_o,\theta_0) \ge 2\max\{B^{-1}, D^{-1}\}n^{-1}$. Thus, the oracle risk should be larger than the parametric rate with sufficiently large constant. Clearly, the upper bounds in the right-hand sides of the inequalities of the above lemma will therefore be small if $n\mathcal{R}_n(N_o,\theta_0)$ is large, which is typically the case for the so-called "nonparametric" θ_0's:

$$n\mathcal{R}_n(N_o,\theta_0) \to \infty \qquad \text{as} \quad n \to \infty,$$

i.e., the oracle risk is of a bigger order than the parametric rate n^{-1} as $n \to \infty$.

For brevity sake, denote

$$P_-(\theta_0, \beta, \tau, \tau_1, n) = \sum_{k\in\mathcal{N}^-(\tau)} \exp\left\{-\beta n\big(\mathcal{R}_n(k,\theta_0) - \tau_1\mathcal{R}_n(N_o,\theta_0)\big)\right\}, \qquad (41)$$

$$P_+(\theta_0, \beta, \tau, \tau_2, n) = \sum_{k\in\mathcal{N}^+(\tau)} \exp\left\{-\beta n\big(\mathcal{R}_n(k,\theta_0) - \tau_2\mathcal{R}_n(N_o,\theta_0)\big)\right\}, \qquad (42)$$

where the projection risk $\mathcal{R}_n(k,\theta)$ is defined by (7), the oracle cut-off $N_o = N_o(\theta_0)$ is defined by (8), the sets $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ are defined by (22) and (23) respectively.

Corollary 4. *Under the conditions of Lemma 7, the following bounds hold:*

$$\begin{aligned} P_-(\theta_0,\beta,\tau,\tau_1,n) &\le \frac{1}{\tau}\min\Big\{\frac{c_1}{B^2 n\mathcal{R}_n(N_o,\theta_0)}, \frac{n\mathcal{R}_n(N_o,\theta_0)}{e^{Bn\mathcal{R}_n(N_o,\theta_0)}}\Big\} \le \frac{\min\big\{4(Be)^{-1},1\big\}}{\tau Be}, \\ P_+(\theta_0,\beta,\tau,\tau_2,n) &\le \min\Big\{\frac{g(D)}{\tau n\mathcal{R}_n(N_o,\theta_0)}, \frac{c_3 n\mathcal{R}_n(N_o,\theta_0)}{\tau e^{Dn\mathcal{R}_n(N_o,\theta_0)}}, \frac{e^{-DN_o}}{e^D-1}\Big\} \\ &\le \min\Big\{\frac{g(D)}{\tau}, \frac{1+e}{4D\tau}, \frac{e^{-D}}{e^D-1}\Big\}. \end{aligned}$$

Proof. Indeed, by using Lemma 7 and the facts that $n\mathcal{R}_n(N_o,\theta_0)\ge 1$, $\max_{x\ge 0}\{xe^{-Bx}\} = (Be)^{-1}$ and $\frac{1}{\mathcal{R}_n(k,\theta_0)} \le \frac{1}{\tau\mathcal{R}_n(N_o,\theta_0)}$ for all $k\notin\mathcal{N}(\tau,\theta_0)$, we obtain

$$\begin{aligned} P_-(\theta_0,\beta,\tau,\tau_1,n) &\le \frac{R_-(\theta_0,\beta,\tau,\tau_1,n)}{\tau n\mathcal{R}_n(N_o,\theta_0)} \le \tau^{-1}\min\Big\{\frac{c_1}{B^2 n\mathcal{R}_n(N_o,\theta_0)}, \frac{n\mathcal{R}_n(N_o,\theta_0)}{e^{Bn\mathcal{R}_n(N_o,\theta_0)}}\Big\} \\ &\le \tau^{-1}\min\Big\{\frac{4}{(Be)^2}, \frac{1}{Be}\Big\} = (\tau Be)^{-1}\min\big\{4(Be)^{-1},1\big\}. \end{aligned}$$

The inequality

$$P_+(\theta_0,\beta,\tau,\tau_2,n) \le \frac{1}{\tau}\min\Big\{\frac{g(D)}{n\mathcal{R}_n(N_o,\theta_0)}, \frac{c_3 n\mathcal{R}_n(N_o,\theta_0)}{e^{Dn\mathcal{R}_n(N_o,\theta_0)}}\Big\} \le \min\Big\{\frac{g(D)}{\tau}, \frac{1+e}{4D\tau}\Big\}$$

follows similarly from Lemma 7. Besides, due to (42) and (37) and the fact that $a_k\ge k$ for all $k\in\mathbb{N}$, we derive the bound

$$\begin{aligned} P_+(\theta_0,\beta,\tau,\tau_2,n) &\le \sum_{k\in\mathcal{N}^+(\tau)} \exp\{-Da_k\} \le \sum_{k=N_o+1}^{\infty}\exp\{-Dk\} \\ &\le \frac{e^{-D(N_o+1)}}{1-e^{-D}} \le \frac{e^{-DN_o}}{e^D-1}. \end{aligned}$$

Combining the last two relations completes the proof. □

6. PROOFS OF THE THEOREMS

Proof of Theorem 1. First recall that, according to (16),

$$\widehat{\theta} = E'(\theta \mid X) = \sum_{k\in\mathbb{N}} E'(\theta \mid X, N=k)P(N=k\mid X).$$

That is

$$\widehat{\theta}_i = E'(\theta_i \mid X) = \sum_{k\in\mathbb{N}} E'(\theta_i \mid X, N=k)P(N=k\mid X) = \sum_{k\in\mathbb{N}}\widehat{\theta}_i(k)P(N=k\mid X)$$

with $\widehat{\theta}_i(k) = X_i I\{i\le k\}$. Now, by Fubini's theorem and the Cauchy–Schwarz inequality,

$$\begin{aligned} E_{\theta_0}\|\widehat{\theta}-\theta_0\|^2 &= E_{\theta_0}\sum_{i\in\mathbb{N}}\Big(\sum_{k\in\mathbb{N}}\widehat{\theta}_i(k)P(N=k\mid X) - \theta_{0i}\Big)^2 \\ &\le E_{\theta_0}\sum_{i\in\mathbb{N}}\sum_{k\in\mathbb{N}}\big(\widehat{\theta}_i(k)-\theta_{0i}\big)^2 P(N=k\mid X) \\ &= E_{\theta_0}\sum_{k\in\mathbb{N}}\|\widehat{\theta}(k)-\theta_0\|^2 P(N=k\mid X). \end{aligned}$$

Fix a $\tau \geq 1$ to be chosen later and let the sets $\mathcal{N}(\tau)$, $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ be defined by (21), (22) and (23) respectively. Split the last sum in three sums over the sets $\mathcal{N}(\tau)$, $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ and apply Lemmas 3 and 4 to derive the bound for the estimator (16):

$$E_{\theta_0}\|\widehat{\theta} - \theta_0\|^2 \leq 2\tau\mathcal{R}_n(N_o, \theta_0) + T_1 + T_2, \tag{43}$$

for any $\theta_0 \in \ell_2$, where

$$T_1 = \sum_{k\in\mathcal{N}^-(\tau)} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\Big(E_{\theta_0}P(N=k \mid X)\Big)^{1/2},$$

$$T_2 = \sum_{k\in\mathcal{N}^+(\tau)} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\Big(E_{\theta_0}P(N=k \mid X)\Big)^{1/2}.$$

Similarly, applying Lemmas 3 and 4 to the estimator (17), we obtain the same bound for the estimator (17):

$$\begin{aligned} E_{\theta_0}\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 &= E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2\Big(I\big\{\widehat{N}\in\mathcal{N}(\tau)\big\} + I\big\{\widehat{N}\in\mathcal{N}^-(\tau)\big\} + I\big\{\widehat{N}\in\mathcal{N}^+(\tau)\big\}\Big)\Big] \\ &\leq 2\tau\mathcal{R}_n(N_o,\theta_0) + T_1 + T_2 \end{aligned} \tag{44}$$

for any $\theta_0 \in \ell_2$, with T_1 and T_2 defined above.

Now we apply Lemma 6 to evaluate both terms T_1 and T_2. Take some $\tau_1 \geq \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha \geq 1$ to be chosen later. The inequality $\tau_-(\alpha) \geq 1$ follows from the condition on α. Then, by using Lemma 6 and the definitions (29) and (41), we obtain

$$\begin{aligned} T_1 &\leq \sum_{k\in\mathcal{N}^-(\tau)} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\exp\Big\{-\frac{n}{12}\big(\mathcal{R}_n(k,\theta_0) - \tau_1\mathcal{R}_n(N_o,\theta_0)\big)\Big\} \\ &\leq n^{-1}\big(R_-(\theta_0, 1/12, \tau, \tau_1, n) + P_-(\theta_0, 1/12, \tau, \tau_1, n)\big). \end{aligned}$$

Assume now that τ and τ_1 are chosen in such a way that $\tau > \tau_1 \geq \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha$. Then we can apply part (i) of Lemma 7 and Corollary 4 to obtain

$$T_1 \leq \frac{C_1}{n}, \qquad C_1 = C_1(\tau,\tau_1) = \frac{c_1}{B^2} + \frac{1}{\tau B e}, \tag{45}$$

with $c_1 = 4e^{-2}$ and $B = B(\tau,\tau_1) = \frac{\tau-\tau_1}{12\tau}$.

To bound the term T_2, we apply Lemmas 6 and 7 again. Assume that τ and τ_2 are chosen in such a way that $\tau > \tau_2 \geq \tau_+(\alpha) = \frac{1}{2\alpha} \geq 1$. The inequality $\tau_+(\alpha) \geq 1$ follows from the condition on α. By applying consequently Lemma 6 and then Lemma 7 with Corollary 4, we obtain that

$$\begin{aligned} T_2 &\leq \sum_{k\in\mathcal{N}^+(\tau)} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\exp\Big\{-\frac{n\alpha}{2}\big(\mathcal{R}_n(k,\theta_0) - \tau_2\mathcal{R}_n(N^o,\theta_0)\big)\Big\} \\ &\leq n^{-1}\big(R_+(\theta_0, \alpha/2, \tau, \tau_2, n) + P_+(\theta_0, \alpha/2, \tau, \tau_2, n)\big) \leq \frac{C_2}{n}, \end{aligned} \tag{46}$$

$$C_2 = C_2(\tau,\tau_2,\alpha) = g(D) + \min\Big\{\frac{g(D)}{\tau}, \frac{e^{-D}}{e^D - 1}\Big\}$$

with $D = D(\tau,\tau_2,\alpha) = \frac{\alpha(\tau-\tau_2)}{2\tau}$ and $g(u) = e^{-1}u^{-2} + e^{-u}(e^u - 1)^{-2}$.

It remains to choose τ, τ_1, τ_2 so that $\tau > \tau_1 \geq \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha$ and $\tau \geq \tau_2 \geq \tau_+(\alpha) = \frac{1}{2\alpha}$. Take, for example, $\tau_1 = \tau_-(\alpha)$, $\tau_2 = \tau_+(\alpha)$ and $\tau = \max\{\tau_1,\tau_2\} + 1$. Finally, we combine the relations (43), (44), (45) and (46) to establish the statement of the theorem with $K_1 = 2\tau$ and $K_2 = C_1 + C_2$. The theorem is proved. □

Remark 18. For example, we take $\tau_1 = \frac{1}{2\alpha}$, $\tau_2 = (6\log\left(\frac{2}{\sqrt{3}}\right) + 6\alpha)$ with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$ such that $\tau_1 = \tau_2$, which leads to

$$\alpha = \frac{-6\log\left(\frac{2}{\sqrt{3}}\right) + \sqrt{36\log^2\left(\frac{2}{\sqrt{3}}\right) + 12}}{12} \approx 0.2255789$$

and $\tau_1 = \tau_2 = \frac{1}{2\alpha} \approx 2.21652$. Take further $\tau = \tau_1 + 7$. Then we compute the constants $K_1 \approx 18.433$ and $K_2 \approx 310.904$, or the following (non asymptotic) inequality holds true for all $\theta_0 \in \ell_2$:

$$E_{\theta_0}\|\widehat{\theta} - \theta_0\|^2 \le 19\mathcal{R}_n(N_o, \theta_0) + \frac{311}{n}.$$

Of course, this is not the optimal choice of the involved parameters $\alpha, \tau_1, \tau_2, \tau$.

Remark 19. We could also try to improve the constants K_1 and K_2 by refining the bounds in the proof. However, if we try to make K_1 as small as possible, the resulting K_2 becomes large. On the other hand, we can make K_2 smaller by sacrificing the constant K_1. Actually, the constant K_2 can be strongly improved if n is a reasonable number and we assume a nonparametric behavior of the oracle risk $\mathcal{R}_n(N_o, \theta_0)$, which means that $n\mathcal{R}_n(N_o, \theta_0)$ is large. This follows from the proof of the theorem, where we used the following uniform trivial bounds: $n\mathcal{R}_n(N_o, \theta_0) \ge 1$,

$$\begin{aligned} n\mathcal{R}_n(N_o, \theta_0)\exp\{-cn\mathcal{R}_n(N_o, \theta_0)\} &\le \max_{x\ge 0}\{xe^{-cx}\} = (ce)^{-1}, \\ \big(n\mathcal{R}_n(N_o, \theta_0)\big)^2\exp\{-cn\mathcal{R}_n(N_o, \theta_0)\} &\le \max_{x\ge 0}\{x^2e^{-cx}\} = 4(ce)^{-2}. \end{aligned}$$

Another possibility to improve the constants is by introducing a factor c in the expression for the prior variance $\tau_i^2(N) = cn^{-1}I\{i \le N\}$ in the definition of the prior (9). We also used a somewhat rough estimate $P^2(N = k \mid X) \le P(N = k \mid X)$ in some places of the proof.

Proof of Theorem 2. First consider the posterior (14). By the conditional Chebyshev inequality,

$$\begin{aligned} P\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\|^2 \ge Mr_n(\theta_0) \mid X\Big\} &= \sum_{k\in\mathbb{N}} P\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\|^2 \ge Mr_n(\theta_0) \mid X, N = k\Big\}P(N = k \mid X) \\ &\le \sum_{k\in\mathbb{N}} \frac{E\big(\|\theta - \frac{\theta_0}{2}\|^2 \mid X, N = k\big)}{Mr_n(\theta_0)}P(N = k \mid X), \end{aligned} \tag{47}$$

where the conditional distribution $P(\theta \mid X, N = k)$ is defined by (10). Analogously, for the posterior (15) we have that

$$P'\big\{\|\theta - \theta_0\|^2 \ge Mr_n(\theta_0) \mid X\big\} \le \sum_{k\in\mathbb{N}} \frac{E'\big(\|\theta - \theta_0\|^2 \mid X, N = k\big)}{Mr_n(\theta_0)}P(N = k \mid X), \tag{48}$$

where the conditional distribution $P'(\theta \mid X, N = k)$ is defined by (12).

Using these, we compute

$$\begin{aligned} E\Big[\big\|\theta - \frac{\theta_0}{2}\big\|^2 \mid X, N = k\Big] &= \sum_{i=1}^{\infty}\mathrm{Var}(\theta_i \mid X, N = k) + \sum_{i=1}^{\infty}\Big(E(\theta_i \mid X, N = k) - \frac{\theta_0}{2}\Big)^2 \\ &= \frac{k}{2n} + \frac{1}{4n}\sum_{i=1}^{k}\xi_i^2 + \sum_{i=k+1}^{\infty}\frac{\theta_{0_i}^2}{4}, \\ E'\Big[\big\|\theta - \theta_0\big\|^2 \mid X, N = k\Big] &= \frac{k}{2n} + \frac{1}{n}\sum_{i=1}^{k}\xi_i^2 + \sum_{i=k+1}^{\infty}\theta_{0_i}^2. \end{aligned} \tag{49}$$

The relations (47) and (49) imply that

$$
\begin{aligned}
&E_{\theta_0} P\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\|^2 \ge M r_n(\theta_0) \mid X\Big\} \\
&\le \frac{1}{M r_n(\theta_0)} E_{\theta_0}\Big[\sum_{k\in\mathbb{N}}\Big(\frac{k}{2n} + \frac{1}{4}\sum_{i=1}^{k}\frac{\xi_i^2}{n} + \sum_{i=k+1}^{\infty}\frac{\theta_{0i}^2}{4}\Big)P(N=k\mid X)\Big] \\
&\le \frac{\sum_{k\in\mathbb{N}}\mathcal{R}_n(k,\theta_0)E_{\theta_0}P(N=k\mid X)}{2Mr_n(\theta_0)} + \frac{E_{\theta_0}\Big[\sum_{k\in\mathbb{N}}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k\mid X)\Big]}{4Mr_n(\theta_0)}. \qquad (50)
\end{aligned}
$$

Similarly, from the relations (48) and (49) it follows that

$$
\begin{aligned}
&E_{\theta_0} P'\big\{\|\theta-\theta_0\|^2 \ge M r_n(\theta_0)\mid X\big\} \\
&\le \frac{\sum_{k\in\mathbb{N}}\mathcal{R}_n(k,\theta_0)E_{\theta_0}P(N=k\mid X)}{Mr_n(\theta_0)} + \frac{E_{\theta_0}\Big[\sum_{k\in\mathbb{N}}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k\mid X)\Big]}{Mr_n(\theta_0)}. \qquad (51)
\end{aligned}
$$

Fix a $\tau > 0$ to be chosen later and let the sets $\mathcal{N}(\tau)$, $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ be defined by (21), (22) and (23) respectively. Suppose τ, τ_1, τ_2 are chosen in such a way that $\tau > \tau_1 \ge \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha$ and $\tau \ge \tau_2 \ge \tau_+(\alpha) = \frac{1}{2\alpha}$. Clearly, $\mathcal{R}_n(k,\theta_0) \le \tau\mathcal{R}_n(N_o,\theta_0)$ for all $k \in \mathcal{N}(\tau)$, so that

$$
\begin{aligned}
\sum_{k\in\mathbb{N}}\mathcal{R}_n(k,\theta_0)E_{\theta_0}P(N=k\mid X) &\le \sum_{k\in\mathcal{N}(\tau)}\mathcal{R}_n(k,\theta_0)E_{\theta_0}P(N=k\mid X) \\
&\quad + \sum_{k\in\mathcal{N}^-(\tau)}\mathcal{R}_n(k,\theta_0)E_{\theta_0}P(N=k\mid X) + \sum_{k\in\mathcal{N}^+(\tau)}\mathcal{R}_n(k,\theta_0)E_{\theta_0}P(N=k\mid X) \\
&\le \tau\mathcal{R}_n(N_o,\theta_0)E_{\theta_0}P(N\in\mathcal{N}(\tau)\mid X) + T_1 + T_2 \\
&\le \tau\mathcal{R}_n(N_o,\theta_0) + \frac{C_1+C_2}{n}, \qquad (52)
\end{aligned}
$$

where the terms T_1 and T_2 are defined in the proof of Theorem 1, the constants $C_1 = C_1(\tau,\tau_1)$ and $C_2 = C_2(\tau,\tau_2,\alpha)$ are defined by (45) and (46) respectively. The last inequality follows from the bounds $T_1 \le C_1/n$ and $T_2 \le C_2/n$ established in the proof of Theorem 1 (relations (45) and (46)).

Since $\mathcal{R}_n(k,\theta_0) = \frac{k}{n} + \sum_{i=k+1}^{\infty}\theta_{0i}^2 \le \tau\mathcal{R}_n(N_o,\theta_0)$ for any $k \in \mathcal{N}(\tau)$, we obtain

$$N_{\max} = \max\{\mathcal{N}(\tau)\} \le \tau n \mathcal{R}_n(N_o,\theta_0).$$

This implies that

$$
\begin{aligned}
&E_{\theta_0}\Big[\sum_{k\in\mathbb{N}}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k\mid X)\Big] \\
&= E_{\theta_0}\Big[\sum_{k\in\mathcal{N}(\tau)}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k\mid X)\Big] + E_{\theta_0}\Big[\sum_{k\in\mathcal{N}^-(\tau)\cup\mathcal{N}^+(\tau)}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k\mid X)\Big] \\
&\le E_{\theta_0}\sum_{i=1}^{N_{\max}}\frac{\xi_i^2}{n}\sum_{k\in\mathcal{N}(\tau)}P(N=k\mid X) + \sum_{k\in\mathcal{N}^-(\tau)\cup\mathcal{N}^+(\tau)}\frac{k+1}{n}\Big(E_{\theta_0}P^2(N=k\mid X)\Big)^{1/2} \\
&\le E_{\theta_0}\sum_{i=1}^{N_{\max}}\frac{\xi_i^2}{n} + \sum_{k\in\mathcal{N}^-(\tau)\cup\mathcal{N}^+(\tau)}\Big(\mathcal{R}_n(k,\theta_0)+\frac{1}{n}\Big)\Big(E_{\theta_0}P(N=k\mid X)\Big)^{1/2} \\
&\le \frac{N_{\max}}{n} + T_1 + T_2 \le \tau\mathcal{R}_n(N_o,\theta_0) + \frac{C_1+C_2}{n}, \qquad (53)
\end{aligned}
$$

where we again used the relations (45) and (46) to bound the terms T_1 and T_2.

It remains to choose τ, τ_1, τ_2 so that $\tau > \tau_1 \geq \tau_-(\alpha) = 6\log\left(\frac{2}{\sqrt{3}}\right) + 6\alpha$ and $\tau > \tau_2 \geq \tau_+(\alpha) = \frac{1}{2\alpha}$. Take, for example, $\tau_1 = \tau_-(\alpha)$, $\tau_2 = \tau_+(\alpha)$ and $\tau = \max\{\tau_1, \tau_2\} + 1$. Thus the constants τ, C_1 and C_2 depend only on α.

Combining (50), (52) and (53) and taking into account that $r_n(\theta_0) = \mathcal{R}_n(N_o, \theta_0) \geq \frac{1}{n}$, we obtain that for any $\theta_0 \in \ell_2$

$$E_{\theta_0} P\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\|^2 \geq M r_n(\theta_0) \mid X\Big\} \leq \frac{3\tau r_n(\theta_0) + 3(C_1 + C_2)n^{-1}}{4M r_n(\theta_0)} \leq \frac{3(\tau + C_1 + C_2)}{4M}.$$

Finally, combining (51), (52) and (53) in a similar manner, we obtain that for any $\theta_0 \in \ell_2$

$$E_{\theta_0} P'\big\{\|\theta - \theta_0\|^2 \geq M r_n(\theta_0) \mid X\big\} \leq \frac{2\tau r_n(\theta_0) + 2(C_1 + C_2)n^{-1}}{M r_n(\theta_0)} \leq \frac{2(\tau + C_1 + C_2)}{M}.$$

The theorem follows. □

Proof of Corollary 1. Write

$$\begin{aligned} P\Big\{\big\|\theta - \theta_0 + \widehat{\theta}/2\big\|^2 \geq M r_n(\theta_0) \mid X\Big\} &= P\Big\{\big\|\theta - \theta_0 + \widehat{\theta}/2\big\| \geq \sqrt{M r_n(\theta_0)} \mid X\Big\} \\ \leq\ P\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\| \geq \frac{\sqrt{M r_n(\theta_0)}}{2} \mid X\Big\} &+ P\Big\{\Big\|\frac{\widehat{\theta}}{2} - \frac{\theta_0}{2}\Big\| \geq \frac{\sqrt{M r_n(\theta_0)}}{2} \mid X\Big\} \\ =\ P\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\|^2 \geq \frac{M r_n(\theta_0)}{4} \mid X\Big\} &+ P\Big\{\big\|\widehat{\theta} - \theta_0\big\|^2 \geq M r_n(\theta_0) \mid X\Big\}. \end{aligned}$$

Using the conditional Chebyshev inequality, Theorem 1 and the fact that always $r_n(\theta_0) = \mathcal{R}_n(N_o, \theta_0) \geq 1/n$, we bound the expectation of the second term in the right-hand side of the above inequality: for any $\theta_0 \in \ell_2$,

$$E_{\theta_0} P\Big\{\big\|\widehat{\theta} - \theta_0\big\|^2 \geq M r_n(\theta_0) \mid X\Big\} \leq \frac{E_{\theta_0}\|\widehat{\theta} - \theta_0\|^2}{M r_n(\theta_0)} \leq \frac{K_1 + K_2}{M}.$$

The corollary follows since, by Theorem 2, we have that, for any $\theta_0 \in \ell_2$,

$$E_{\theta_0} P\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\|^2 \geq \frac{M r_n(\theta_0)}{4} \mid X\Big\} \leq \frac{K_3}{M} \tag{54}$$

for some constant $K_3 > 0$ depending only on α. □

Proof of Theorem 3. Recall the definitions (22), (23) and the fact that the random variable $\widehat{N}$ is generated according to $P(N \mid X)$. We apply Lemma 6 and Corollary 4. For a given $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$, take some values τ, τ_1 and τ_2 such that $\tau > \tau_1 \geq \tau_- = \tau_-(\alpha) = 6\log\left(\frac{2}{\sqrt{3}}\right) + 6\alpha \geq 1, \tau > \tau_2 \geq \tau_+(\alpha) = \frac{1}{2\alpha} \geq 1$. Then, according to Lemma 6, we obtain

$$\begin{aligned} P_{\theta_0}\big\{\widehat{N} \notin \mathcal{N}(\tau)\big\} &= E_{\theta_0} P\big\{\widehat{N} \notin \mathcal{N}(\tau) \mid X\big\} \\ &= \sum_{k \in N_-(\tau) \cup N_+(\tau)} E_{\theta_0} P\big\{\widehat{N} = k \mid X\big\} \\ &\leq P_-(\theta_0, 1/6, \tau, \tau_1, n) + P_+(\theta_0, \alpha, \tau, \tau_1, n). \end{aligned}$$

Now fix values $\tau_1 = \tau_-(\alpha)$ and $\tau_2 = \tau_+(\alpha)$. Finally apply Corollary 4 to the right-hand side of the last inequality and use the trivial relation $\min\{a_1, b_1\} + \min\{a_2, b_2\} \leq \min\{a_1 + a_2, b_1 + b_2\}$ to get the statement of the theorem:

$$P_{\theta_0}\big\{\widehat{N} \notin \mathcal{N}(\tau)\big\} \leq \min\Big\{\frac{B_1}{n\mathcal{R}_n(N_o, \theta_0)}, \frac{B_2 n\mathcal{R}_n(N_o, \theta_0)}{e^{B_3 n\mathcal{R}_n(N_o, \theta_0)}}\Big\},$$

where

$$\begin{aligned} B_1 &= B_1(\alpha,\tau) = \frac{c_1 B^{-2} + g(D)}{\tau}, \\ B_2 &= B_2(\alpha,\tau) = \frac{1+c_3}{\tau}, \\ B_3 &= B_3(\alpha,\tau) = \min\{B, D\}, \end{aligned}$$

$c_1 = 4e^{-2}$, $c_3 = (e+e^2)/4$, $B = B(\tau, \tau_-(\alpha), 1/6) = \frac{\tau - \tau_-(\alpha)}{6\tau}$, $D = D(\tau, \tau_+(\alpha), \alpha) = \frac{\alpha(\tau - \tau_+(\alpha))}{\tau}$ and $g(D) = e^{-1}D^{-2} + e^{-D}(e^D - 1)^{-2}$. □

Remark 20. To get uniform constants, fix some $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$, take some value τ such that

$$\tau > \tau_-(\alpha) = 6\log\left(\frac{2}{\sqrt{3}}\right) + 6\alpha \qquad \text{and} \qquad \tau > \tau_+(\alpha) = \frac{1}{2\alpha}.$$

There are many choices possible, for example, take α such that $\tau_-(\alpha) = \tau_+(\alpha) = \frac{1}{2\alpha}$, which is $\alpha \approx 0.226$ and $\tau_-(\alpha) = \tau_+(\alpha) \approx 2.217$. Next, take $\tau = \tau_-(\alpha) + 2 \approx 4.217$, then $B_1 \approx 44.87$, $B_2 \approx 0.84$, $B_3 \approx 0.08$.

Proof of Theorem 4. By the definition of the MAP selector $\widehat{N}_{\text{MAP}}$,

$$\left\{P(N = \widehat{N}_{\text{MAP}} \mid X) \le \delta\right\} \subseteq \left\{P(N \notin \mathcal{N}(\tau,\theta_0) \mid X) \ge 1 - |\mathcal{N}(\tau,\theta_0)|\delta\right\},$$

$$\left\{\widehat{N}_{\text{MAP}} \notin \mathcal{N}(\tau,\theta_0), P(N = \widehat{N}_{\text{MAP}} \mid X) > \delta\right\} \subseteq \left\{P(N \notin \mathcal{N}(\tau,\theta_0) \mid X) \ge \delta\right\}.$$

Notice that $|\mathcal{N}(\tau,\theta_0)|$ is finite for any $\theta_0 \in \ell_2$. Next, apply the above relation and the Markov inequality:

$$\begin{aligned} P_{\theta_0}\left(\widehat{N}_{\text{MAP}} \notin \mathcal{N}(\tau,\theta_0)\right) &= P_{\theta_0}\left(\widehat{N}_{\text{MAP}} \notin \mathcal{N}(\tau,\theta_0), P_{\theta_0}(N = \widehat{N}_{\text{MAP}} \mid X) \le \delta\right) \\ &\quad + P_{\theta_0}\left(\widehat{N}_{\text{MAP}} \notin \mathcal{N}(\tau,\theta_0), P(N = \widehat{N}_{\text{MAP}} \mid X) > \delta\right) \\ &\le \frac{E_{\theta_0} P(N \notin \mathcal{N}(\tau,\theta_0) \mid X)}{1 - |\mathcal{N}(\tau,\theta_0)|\delta} + \frac{E_{\theta_0} P(N \notin \mathcal{N}(\tau,\theta_0) \mid X)}{\delta}. \end{aligned}$$

Take $\delta = \left(|\mathcal{N}(\tau,\theta_0)|^{1/2} + |\mathcal{N}(\tau,\theta_0)|\right)^{-1}$, apply Theorem 3 and recall that $|\mathcal{N}(\tau,\theta_0)| \le \tau n \mathcal{R}(N_o,\theta_0)$ to complete the proof. □

ACKNOWLEDGMENTS

This work was supported by the Netherlands Organization for Scientific Research (NWO).

REFERENCES

1. E. Belitser and F. Enikeeva, "Empirical Bayesian Test for the Smoothness", Math. Methods Statist. **17**, 1–18 (2008).
2. E. Belitser and S. Ghosal, "Adaptive Bayesian Inference on the Mean of an Infinite-Dimensional Normal Distribution", Ann. Statist. **31**, 536–559 (2003).
3. E. Belitser and B. Levit, "On Minimax Filtering over Ellipsoids", Math. Methods Statist. **3**, 259–273 (1995).
4. L. Birgé and P. Massart, "Gaussian Model Selection", J. Eur. Math. Soc. **3**, 203–268 (2001).
5. L. D. Brown and M. G. Low, "Asymptotic Equivalence of Nonparametric Regression and White Noise", Ann. Statist. **24**, 2384–2398 (1995).
6. L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov, "Oracle Inequalities for Inverse Problems", Ann. Statist. **30**, 843–874 (2002).
7. L. Cavalier and A. Tsybakov, "Penalized Blockwise Stein's Method, Monotone Oracles and Sharp Adaptive Estimation", Math. Methods Statist. **10**, 247–282 (2001).
8. S. Efromovich and M. Pinsker, "A Learning Algorithm for Nonparametric Filtering", Automat. Remote Control. **24**, 1434–1440 (1984).

9. S. Ghosal, J. K. Ghosh and A. W. van der Vaart, "Convergence Rates of Posterior Distributions", Ann. Statist. **28**, 500–531 (2000).
10. G. K. Golubev, "On a Method for Minimizing Empirical Risk", Problems Inform. Transmission. **40**, 202–211 (2004).
11. Y. Golubev and B. Levit, "An Oracle Approach to Adaptive Estimation of Linear Functionals in a Gaussian Model", Math. Methods Statist. **13**, 392–408 (2004).
12. M. Hoffmann and O. Lepski, "Random Rates in Anisotropic Regression", Ann. Statist. **28**, 325–396 (2002).
13. I. A. Ibragimov and R. Z. Khasminski, *Statistical Estimation: Asymptotic Theory* (Springer, New York, 1981).
14. I. Johnstone, *Function Estimation in Gaussian Noise: Sequence models* (Monograph draft, 1999), http://www-stat.stanford.edu/∼imj/.
15. A. Kneip, "Ordered Linear Smoothers", Ann. Statist. **22**, 835–866 (1994).
16. M. Pinsker, "Optimal Filtration of Square-Integrable Signal in Gaussian White Noise", Problems Inform. Transmission. **16**, 120–133 (1980).