On the empirical Bayes approach to adaptive filtering in the Gaussian model

Eduard Belitser[1] and Boris Levit[2]

[1]Mathematical Institute, Utrecht University
P.O. Box 80010, 3508 TA Utrecht, The Netherlands

[2]Department of Mathematics and Statistics, Queen's University
Kingston, Ontario, K7L 3N6, Canada

We discuss empirical Bayes approach to the problem of adaptive estimation of a linear functional of an infinite dimensional normal mean vector. In the continuous setting, this problem corresponds, via Fourier duality, to the pointwise recovery of a signal observed in the white Gaussian noise. The minimax approach to this problem is essentially equivalent to the Bayes filtering of a stationary Gaussian process corrupted by a white Gaussian noise.
The proposed method of adaptation essentially combines two classical approaches: the Wiener filter and the empirical Bayes approach. Our main purpose is to demonstrate how this approach works, in a prototypical nonparametric problem. We will also briefly discuss an interesting phenomenon of (Bayesian) under- and oversmoothing.

Key words: adaptation, Bayes risk, empirical Bayes approach, minimax risk, nonparametric model, Wiener filter.

2000 Mathematics Subject Classification: Primary 62G05, 62C12; Secondary 62M20, 93E11.

1 Introduction

Suppose we draw observations

$$X_i = \theta_i + \epsilon\,\xi_i, \quad i = 1, 2, \dots\,, \tag{1}$$

where the noise variables ξ_i's are independent standard Gaussian and $\epsilon > 0$ is a small parameter. The parameter $\theta = (\theta_1, \theta_2 \dots) \in \ell_2$ is unknown, and the goal is estimate a linear functional of θ, say $\Phi = \Phi(\theta)$. In the minimax setting, when θ is assumed to lie in a given symmetric convex set $\Theta \subset \ell_2$, this problem was thoroughly investigated by Ibragimov and Hasminskii (1984). Pinsker (1980) studied the problem of minimax estimation of θ in ℓ_2-norm, for ellipsoidal parameter sets Θ.

The interest to this problem is motivated by the following. Consider the prototypical white noise model:

$$dX_\epsilon(t) = f(t)dt + \epsilon\, dW(t), \quad 0 \le t \le 1, \tag{2}$$

where $X_t(\cdot)$ is the noise-corrupted observation process, $f(\cdot) \in L_2([0,1])$ is the unknown signal, $W(t)$ is a standard Brownian motion and $\epsilon > 0$ is a small parameter. The statistical estimation problem is to recover the signal $f(t)$ at a given point t_0, based on the observation $X_\epsilon(t)$. The above model arises as the limiting experiment in different curve estimation problems, where typically $\epsilon = n^{-1/2}$, with n being the sample size; see Nussbaum (1996) and Klemelä and Nussbaum (1998) for density estimation; Brown and Low (1996) for non-parametric regression.

Suppose that $\phi_i(\cdot)$, $i = 1, 2, \ldots$, form an orthonormal basis of $L_2[0,1]$. Then the problem can be transformed to the equivalent *sequence model* (1) in which $X_i = \int_0^1 \phi_i(t) dX_\epsilon(t)$ are the observations, $\theta_i = \int_0^1 \phi_i(t) f(t) dt$ – the unknown Fourier coefficients and $\xi_i = \int_0^1 \phi_i(t) dW(t)$. Then the signal can be recovered from the basis expansion $f(t) = \sum_{i=1}^\infty \theta_i \phi_i(t)$, with convergence in L_2-sense. If this series converges pointwisely, signal $f(t_0)$ at point t_0 is a given linear functional of the infinitely dimensional normal mean vector θ: $f(t_0) = \sum_{i=1}^\infty \theta_i \phi_i(t_0) = \Phi_{t_0}(\theta)$.

Thinking of estimates such as $\sum_{i=1}^{\hat{N}} X_i \phi_i(t_0)$, it is often the structure of the vector θ, such as the rate at which $\theta_i \to 0$, which defines the proper estimates, rather then the basis functions ϕ_i. In an adaptive setting, the choice of $\hat{N}$ can be driven by the data X_i. In both cases, the functions $\phi_i(t)$ play a subordinate role, although in the end they will affect the accuracy of estimation; e.g. $|\phi_k(t_0)|^2$'s will appear in the mean square error of the resulting estimate. Artiles (2001) provides a good introduction into this kind of problems.

To simplify our approach, we will concentrate on estimating

$$\Phi(\theta) = \sum_{i=1}^\infty \theta_i\,, \qquad \theta \in \ell_1\,. \tag{3}$$

When the classical trigonometric basis $\phi_k(x) = e^{i2\pi kx}$, $k = 0, \pm 1, \pm 2, \ldots$, is used, a minor technical adjustment is needed to reduce the problem again to (1) using the relations $\theta_k = \bar{\theta}_{-k}$, $k = \pm 1, \pm 2, \ldots$. In this special case we can assume, without loss of generality, that $t_0 = 0$ (indeed $|\phi_k(t)|^2 = |\phi_k(0)|^2 = 1$) so that again our functional takes on the form (3).

From now on we focus on estimation of the functional (3) based on the model (1) and assuming that $\theta \in \ell_1$. Moreover, in this paper we will be dealing exclusively with the quadratic risk function

$$R_\epsilon(\hat{\Phi}, \theta) = E_\theta^{(\epsilon)} (\hat{\Phi} - \Phi(\theta))^2.$$

Here by $E_\theta^{(\epsilon)}$ we mean the conditional expectation given θ. From now on we suppress the dependence of the expectation and some other quantities on ϵ.

There are basically two approaches most often used to study this problem: the classical Bayes approach often invoked within the framework of stationary random processes, and, more recently, the minimax approach which gained popularity after its use in Pinsker (1980). In fact, Pinsker (1980) combined both approaches, and it is this interplay that we will discuss next.

In the minimax approach, it is assumed that the vector θ belongs to Θ, a given compact symmetric subset of ℓ_2. To simplify our discussion, consider the special case of a given hyper-rectangle $\mathcal{H} = \mathcal{H}_a = \{\theta : |\theta_k| \le a_k\,,\ k = 1, 2, \ldots\}$, for a positive $a = (a_1, a_2, \ldots)$, $a_k \to 0$ as $k \to \infty$, and a general linear estimate $\bar{\Phi} = \sum_{k=1}^{\infty} h_k X_k$. Assuming that the corresponding series are converging, the minimax risk, under the restriction $\theta \in \mathcal{H}_a$, is bounded by

$$\sup_{\theta\in\mathcal{H}_a} R_\epsilon(\bar{\Phi}, \theta) = \sup_{\theta\in\mathcal{H}_a} \sum_{k=1}^{\infty} \theta_k^2 (h_k - 1)^2 + \epsilon^2 h_k^2 \le \sum_{k=1}^{\infty} a_k^2 (h_k - 1)^2 + \epsilon^2 h_k^2.$$

Therefore, a natural choice of a linear estimate is the one for which this upper bound is minimal:

$$\hat{\Phi} = \sum_{k=1}^{\infty} \frac{a_k^2}{a_k^2 + \epsilon^2} X_k, \tag{4}$$

with the corresponding maximal risk

$$\sup_{\theta\in\mathcal{H}_a} R_\epsilon(\hat{\Phi}, \theta) \le \epsilon^2 \sum_{k=1}^{\infty} \frac{a_k^2}{a_k^2 + \epsilon^2}. \tag{5}$$

The estimator (4) can be easily seen to be Bayes, with regards to the prior $\Lambda_a(\theta)$, according to which θ_k's are a priori independent $N(0, a_k^2)$-distributed.

Of course, in the Bayesian framework with Gaussian priors, the Bayes estimates are always linear. Note further that the expression (4) can be associated with the so called *Wiener filter* of the stationary Gaussian process $f(t) = \sum_k \theta_k e^{i2\pi kt}$ at point 0. The estimation in this setting is called signal filtration. The Bayes risk is also easy to calculate:

$$E_\Lambda R_\epsilon(\hat{\Phi}, \theta) = \epsilon^2 \sum_{i=1}^{\infty} \frac{a_i^2}{a_i^2 + \epsilon^2}. \tag{6}$$

Note, if the probability (with respect to the prior distribution Λ_a)

$$P_{\Lambda_a}\{\theta \in \mathcal{H}_a\} = 1 \tag{7}$$

then obviously

$$E_\Lambda R_\epsilon(\hat{\Phi}, \theta) \le \sup_{\theta\in\mathcal{H}_a} R_\epsilon(\hat{\Phi}, \theta),$$

and the estimate $\hat{\Phi}$ would thus be exactly minimax by (5)-(6). Of course, the relation (7) is, strictly speaking, impossible. However there are many interesting classes Θ and corresponding sequences a (often $a = a(\epsilon)$, i.e., as distinct from the classical

Bayes approach, prior depends on ϵ) for which this relation nearly holds, rendering that the estimate $\hat{\Phi}$ is asymptotically minimax; cf. Pinsker (1980), Ibragimov and Hasminskii (1984), Belitser and Levit (1995). Note that in such cases, the estimate $\hat{\Phi}$ is both Bayes and asymptotically minimax.

Such situation happens for instance for the hyper-rectangle $\mathcal{H}_a$, when a_i decrease exponentially: $a_i = Q^{1/2}\exp(-\gamma i/2)$, parameter γ has meaning of the amount of "smoothness". In fact, the minimax estimation over $\mathcal{H}_a$ is asymptotically equivalent to estimation over $\mathcal{E}_\gamma$ (see Belitser and Levit (1995)): $\mathcal{E}_\gamma = \mathcal{E}_\gamma(Q) = \big\{\theta : \sum_{i=1}^\infty e^{\gamma i}\theta_i^2 \le Q\big\}$. Such exponential ellipsoids in frequency domain correspond to analytic classes of functions in time domain; cf. Golubev and Levit (1996) in the context of density estimation. More generally, Lepski and Levit (1998) demonstrated that it is always brought about, in the white noise model, if the sequence a_i is *rapidly decreasing*. Lepski and Levit (1998) argue that such models are widely applicable, in particular in the situation when the "true model" is unknown, i.e. in the adaptive setting.

A momentous conclusion to draw from this discussion, is that, at least in the case of rapidly (exponentially) decreasing a_i there is essentially no significant difference between the minimax approach to the non-parametric regression and the Bayes filtering of Gaussian stationary process (with analytic realizations), based on the Wiener filter. Indeed, since the resulting estimates are the same, the difference between the two approaches, one could argue, is simply a matter of their justification. In this case, one could think of the two theories of Bayes and minimax non-parametric estimation as being essentially the same. Note that for other problems, such as estimation in the ℓ_2 norm, this holds even for models with polynomially decreasing a_i. However, for such models the situation is more complicated in the case of linear real-valued functionals or, equivalently, in the point-wise signal estimation since the best linear estimates are no longer asymptotically minimax in this case. For the rest of this paper we will concentrate on the simplest scenario $a_i = \exp(-\gamma i/2)$.

When the a_k's (or rather smoothness linkage parameter γ) are actually unknown, a much more realistic and difficult problem ensues, collectively known as *adaptive estimation.* The above discussion suggests that one could use either minimax approach or the Bayes approach to this problem. Until recently, the first of these approaches appeared to be dominant. Starting with Efromovich and Pinsker (1984), it gained momentum after a series of contributions of Lepski and his collaborators; see e.g. Lepski (1992), Lepski and Levit (1998) in the super-smooth case and further references therein.

Another approach is the one with the Bayes flavor, which we investigate in this paper. Assume again that the θ_i's are *a priori* independent normally distributed, $\theta_i \sim N(0, \sigma_i^2(\gamma_0))$ where as above $\sigma_i^2(\gamma) = \exp(-\gamma i)$. The exact behavior of the Bayes estimates $\hat{\Phi} = \hat{\Phi}_{\gamma_0}$ of our functional of interest (3) is derived in the next section. Next we consider a more complicated situation when the "smoothness" parameter γ_0 will be assumed fixed but unknown. An approach to tackle this problem is well known in statistics – the empirical Bayes approach, it was first introduced by Robbins (1955) in the classical parameter estimation problems. Belitser and Ghosal

(2001) studied a pure Bayesian adaptation approach to a Sobolev-type smoothness classes by putting a prior on the unknown smoothness, which boils down effectively to mixing over candidate models.

The idea of the empirical Bayes approach is rather simple and easy to implement. One passes to the marginal distribution of observations X_i, which are then independent normally $N(0, \epsilon^2 + \sigma_i^2(\gamma_0))$-distributed, and tries to estimate the unknown parameter γ_0, using for example the *marginalized* maximum likelihood estimates. The estimated value $\hat{\gamma}$ is used then to construct the empirical Bayes estimate, say $\hat{\Phi}_{\hat{\gamma}}$, of $\Phi(\theta)$. Our main result shows that the resulting estimate $\hat{\Phi}_{\hat{\gamma}}$ is asymptotically Bayes, when $\epsilon \to 0$, for any "true" but unknown γ_0. A peculiar feature of our study is that the true smoothness parameter γ_0 can be consistently estimated by applying empirical Bayes approach, while in the minimax context this is in general not possible.

During our presentation of these results at the meeting at CIRM, Luminy in 2001, Y. Golubev posed the following problem: does this approach lead to adaptive estimates also in the minimax sense, for instance, in the same sense as in Lepski and Levit (1998). We have not approved or disapproved this hypotheses, which therefore remains an interesting open question. In this respect, one can think of the above method as an alternative adaptive estimation procedure based on empirical Bayes approach. One can try to extend this approach to other statistical models by constructing such estimators with appropriately chosen prior in those models and study their asymptotic behavior in the minimax sense.

2 Main results

Recall that the quality of estimation is measured by the following risk function

$$R_\epsilon(\hat{\Phi}) = E_\pi E_\theta \big(\hat{\Phi} - \Phi(\theta)\big)^2 ,$$

where $\pi = \pi_{\gamma_0}$ is the following prior distribution on $\theta = (\theta_1, \theta_2, \ldots)$:

$$\theta_i \overset{\text{ind}}{\sim} N(0, \sigma_i^2(\gamma_0)) , \quad i = 1, 2, \ldots , \tag{8}$$

with $\sigma_i^2(\gamma) = e^{-\gamma i}$, $\gamma > 0$. An appealing aspect of (8) is its analytical tractability; it is of conjugate form, allowing analytical integration over θ for Bayesian marginalization. Exponential structure of variances of the prior distribution facilitates further the mathematical treatment of the problem.

If $Z|Y \sim N(Y, \tau^2)$ and $Y \sim N(\mu, \sigma^2)$, then

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

the marginal distribution of Z is also normal with mean μ and variance $\tau^2 + \sigma^2$. So, in our case the marginal distribution of $X = (X_1, X_2, \ldots)$ is described as follows:

$$X_i \overset{\text{ind}}{\sim} N(0, \epsilon^2 + \sigma_i^2(\gamma_0)) , \quad i = 1, 2, \ldots .$$

If we knew γ_0, we would use the Bayes estimator

$$\hat{\Phi}_B = \sum_{i=1}^{\infty} \hat{\theta}_i \,, \qquad \text{with} \quad \hat{\theta}_i = \frac{\sigma_i^2(\gamma_0) X_i}{\sigma_i^2(\gamma_0) + \epsilon^2} \,, \quad i = 1, 2, \dots \,, \tag{9}$$

which is also minimax over ellipsoid $\mathcal{E}_{\gamma_0}$:

$$\mathcal{E}_{\gamma_0} = \mathcal{E}_{\gamma_0}(Q) = \left\{ \theta : \ \sum_{i=1}^{\infty} e^{\gamma_0 i} \theta_i^2 \le Q \right\} .$$

Such exponential ellipsoids in frequency domain correspond to analytic classes of functions in time domain; more precisely functions f admitting a bounded analytic continuation into the strip $\{z \in \mathbb{C}, \ |\mathrm{Im}(z)| \le \gamma_0\}$ and $\int |f(x + i\gamma_0)|^2 dx \le Q$. In a way, the above framework represents a Bayesian counterpart of the minimax estimation problem over analytic class with γ_0 as "smoothness" parameter.

The corresponding Bayes signal estimator, coming back to the equivalent white noise model (2) under the trigonometric basis, $\hat{f}(t) = \sum_k \hat{\theta}_k e^{i2\pi kt}$ is in fact the Wiener filter of a stationary Gaussian process from a white Gaussian noise.

Instead of the Bayes estimator $\hat{\Phi}_B$, one can use a somewhat simpler projection estimator

$$\tilde{\Phi} = \tilde{\Phi}_{\gamma_0} = \sum_{i=1}^{N_{\gamma_0}} X_i \,, \tag{10}$$

with $N_\gamma = \lfloor \log \epsilon^{-2}/\gamma \rfloor$. The Bayes risk of this estimator, as the next theorem shows, has the same asymptotic behavior (in the first order as $\epsilon \to 0$)) as the risk of the Bayes estimator $\hat{\Phi}_B$. To avoid unnecessary technicalities in each inequality by N_γ we will mean either $\lfloor \log \epsilon^{-2}/\gamma \rfloor$ or $\lceil \log \epsilon^{-2}/\gamma \rceil$ (this would not influence the results) depending on the sign of the inequality; the relation $\left| N_\gamma - \frac{1}{\gamma} \log \epsilon^{-2} \right| \le 1$ will always hold.

All asymptotic relations and all symbols O and o below refer to, unless otherwise specified, $\epsilon \to 0$. Further, let E denote the mathematical expectation with respect to the random element (X, θ).

The following theorem describes the asymptotic performance of both estimators.

Theorem 1. *As $\epsilon \to 0$,*

$$R_\epsilon(\hat{\Phi}_B) = R_\epsilon(\tilde{\Phi}_{\gamma_0})(1 + o(1)) = \frac{\epsilon^2 \log \epsilon^{-2}}{\gamma_0}(1 + o(1)) \,,$$

where the Bayes estimator $\hat{\Phi}_B$ and the projection estimator $\tilde{\Phi}_{\gamma_0}$ are defined by (9) and (10) respectively.

Proof. We have

$$\begin{aligned} R_\epsilon(\hat{\Phi}_B) &= E\big[\hat{\Phi}_B - \Phi\big]^2 = E\left[\sum_{k=1}^{\infty} \frac{e^{-\gamma_0 k} X_k}{\epsilon^2 + e^{-\gamma_0 k}} - \theta_k \right]^2 \\ &= E\left[\sum_{k=1}^{\infty} \frac{\epsilon e^{-\gamma_0 k} \xi_k - \epsilon^2 \theta_k}{\epsilon^2 + e^{-\gamma_0 k}} \right]^2 = \epsilon^2 \sum_{k=1}^{\infty} \frac{\epsilon^2 e^{-\gamma_0 k} + e^{-2\gamma_0 k}}{\left(\epsilon^2 + e^{-\gamma_0 k}\right)^2} \,. \end{aligned}$$

Now, evaluate first the term. Note that the function $\frac{\epsilon^2 e^{-\gamma_0 x}}{\left(\epsilon^2+e^{-\gamma_0 x}\right)^2}$ is increasing for $x \in \left(0, x_\epsilon\right]$, with $x_\epsilon = \frac{\log \epsilon^{-2}}{\gamma_0}$, and decreasing afterwards. Therefore,

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{\epsilon^2 e^{-\gamma_0 k}}{\left(\epsilon^2+e^{-\gamma_0 k}\right)^2} &\leq \frac{\epsilon^2 e^{-\gamma_0 x_\epsilon}}{\left(\epsilon^2+e^{-\gamma_0 x_\epsilon}\right)^2} + \int_0^{\infty} \frac{\epsilon^2 e^{-\gamma_0 x} dx}{\left(\epsilon^2+e^{-\gamma_0 x}\right)^2} \\ &= \frac{1}{4} + \frac{1}{\gamma_0(1+\epsilon^2)}\,.\end{aligned}$$

In similar manner we evaluate the second term, since function $\frac{e^{-2\gamma_0 x}}{\left(\epsilon^2+e^{-\gamma_0 x}\right)^2}$ is decreasing in x,

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{e^{-2\gamma_0 k}}{\left(\epsilon^2+e^{-\gamma_0 k}\right)^2} &= \int_0^{\infty} \frac{e^{-2\gamma_0 x} dx}{\left(\epsilon^2+e^{-\gamma_0 x}\right)^2} - \beta_\epsilon \\ &= \frac{\log \epsilon^{-2}}{\gamma_0} + \frac{\log(1+\epsilon^{-2})}{\gamma_0} - \frac{1}{\gamma_0(1+\epsilon^2)} - \beta_\epsilon \\ &= \frac{\log \epsilon^{-2}}{\gamma_0} + O(1)\,,\end{aligned}$$

because

$$0 \leq \beta_\epsilon \leq \int_0^1 \frac{e^{-2\gamma_0 x} dx}{\left(\epsilon^2+e^{-\gamma_0 x}\right)^2} \leq \frac{1}{(1+\epsilon^2)^2}\,.$$

Collecting all the above relations, we obtain that, as $\epsilon \to 0$,

$$R_\epsilon(\hat{\Phi}_B) = \frac{\epsilon^2 \log \epsilon^{-2}}{\gamma_0} + O(\epsilon^2) = \frac{\epsilon^2 \log \epsilon^{-2}}{\gamma_0}(1+o(1))\,.$$

The risk of the estimator $\tilde{\Phi}_{\gamma_0}$ is easy to derive:

$$\begin{aligned}R(\tilde{\Phi}_{\gamma_0}) &= E\left[\tilde{\Phi}_{\gamma_0} - \Phi\right]^2 = E\left[\sum_{k=1}^{N_{\gamma_0}} (X_k - \theta_k) - \sum_{k=N_{\gamma_0}+1}^{\infty} \theta_k\right]^2 \\ &= \epsilon^2 N_{\gamma_0} + \sum_{k=N_{\gamma_0}+1}^{\infty} e^{-\gamma_0 k} = \frac{\epsilon^2 \log \epsilon^{-2}}{\gamma_0} + O(\epsilon^2)\,.\end{aligned}$$

The theorem is proved. □

Suppose now that the true γ_0 is unknown and we still want to estimate Φ – the issue of adapting to γ_0 rises. We apply the basic empirical Bayes approach (due to Robbins (1955)), which uses the observed data X to estimate the unknown parameter γ_0, and than proceeds as in a standard Bayesian analysis. That is, one simply replaces γ_0 in the Bayes estimator for Φ by an estimate $\hat{\gamma}$ (for example, obtained as

the value which (nearly) maximizes the marginal likelihood of X). This straightforward approach is frequently used and many simulation studies are performed, but its theoretical properties are usually not so easy to analyze. In traditional minimax context, when θ is assumed to belong to an ellipsoid $\mathcal{E}_{\gamma_0}$ of unknown smoothness γ_0, an adaptation problem in Gaussian noise was first studied by Efromovich and Pinsker (1984).

So, we replace γ_0 in the estimator $\tilde{\Phi}$ by an estimate $\hat{\gamma}$ obtained as the value which (nearly) maximizes the log of the trimmed marginal likelihood:

$$\log L_\epsilon(\gamma) = \log \left[\prod_{i=1}^{n_\epsilon} \frac{1}{\sqrt{2\pi(\epsilon^2 + e^{-\gamma i})}} \exp\Big\{ - \frac{X_i^2}{2(\epsilon^2 + e^{-\gamma i})} \Big\} \right],$$

over the interval $[\alpha_\epsilon, +\infty)$, where $n_\epsilon = \lceil \epsilon^{-2} \rceil$ and the sequence $\alpha_\epsilon > 0$, $\alpha_\epsilon \to 0$ (to be chosen later) as $\epsilon \to 0$. We need $\hat{\gamma}$ to be separated from zero, since, as we will see in the proofs, $\hat{\gamma}$ appears often in denominators.

Exactly, let $\hat{\gamma}$ be an estimator satisfying

$$\log L_\epsilon(\hat{\gamma}) \ge \sup_{\gamma \ge \alpha_\epsilon} \log L_\epsilon(\gamma) - \epsilon^2 .$$

Notice that it is of no importance whether $\log L_\epsilon(\gamma)$ attains maximum over $\gamma \ge \alpha_\epsilon$ or not. Equivalently, the estimator $\hat{\gamma}$ must satisfy

$$Z_\epsilon(\hat{\gamma}) \le \inf_{\gamma \ge \alpha_\epsilon} Z_\epsilon(\gamma) + \epsilon^2$$

with

$$Z_\epsilon(\gamma) = 2 \log \Big[\frac{L_\epsilon(\gamma_0)}{L_\epsilon(\gamma)} \Big] = \sum_{i=1}^{n_\epsilon} \left(\frac{(e^{-\gamma_0 i} - e^{-\gamma i}) X_i^2}{(\epsilon^2 + e^{-\gamma i})(\epsilon^2 + e^{-\gamma_0 i})} + \log \Big[\frac{\epsilon^2 + e^{-\gamma i}}{\epsilon^2 + e^{-\gamma_0 i}} \Big] \right). \quad (11)$$

The sequence $\alpha_\epsilon > 0$ can be any positive sequence converging to zero slowly enough. As is clear from the proof of Lemma 5, we need a technical requirement for $\hat{\gamma}$ to be bounded from below away from zero at least by a power of $1/\log \epsilon^{-2}$; to be exact, we assume that

$$\alpha_\epsilon \to 0 \qquad \text{and} \qquad \alpha_\epsilon \ge \big(\log \epsilon^{-2} \big)^{-p} .$$

for a fixed positive p. This does not restrict any generality in the asymptotic setup. Indeed, recall that the true $\gamma_0 > 0$. So for sufficiently small ϵ we have $0 < \alpha_\epsilon \le \gamma_0$ (namely for all $\epsilon < \epsilon_0(\gamma_0)$ with $\alpha_{\epsilon_0} = \gamma_0$), which implies in turn that $Z_\epsilon(\hat{\gamma}) \le Z_\epsilon(\gamma_0)$.

Now introduce estimator

$$\hat{\Phi} = \sum_{i=1}^{N_{\hat{\gamma}} + M_\epsilon} X_i = \sum_{i=1}^{N_{\hat{\gamma}}} X_i = \tilde{\Phi}_{\hat{\gamma}} , \quad (12)$$

where $M_\epsilon \in \mathbb{N}$ is a sequence such that, as $\epsilon \to 0$,

$$M_\epsilon \big(\log \epsilon^{-2} \big)^{-1} \to 0 \qquad \text{and} \qquad M_\epsilon \big(\log \log \epsilon^{-2} \big)^{-1} \to \infty .$$

For instance, one can take M_ϵ to be the whole part of $\left(\log\log\epsilon^{-2}\right)^2$. Here $\tilde{\gamma} = \hat{B}\hat{\gamma}$ can be thought of as a correction to the naive empirical Bayes smoothness $\hat{\gamma}$ by shrinking it with the shrinkage factor $\hat{B} = \frac{1}{1+M_\epsilon\hat{\gamma}/\log\epsilon^{-2}}$.

Below is the main theorem.

Theorem 2. *As $\epsilon \to 0$,*

$$R_\epsilon(\hat{\Phi}) = \frac{\epsilon^2\log\epsilon^{-2}}{\gamma_0}(1+o(1))\,,$$

where the estimator $\hat{\Phi}$ is given by (12).

Remark 1. One can think of the above result as a Bayesian "oracle". Indeed, suppose the class of estimators, $\mathcal{E} = \{\tilde{\Phi}_\gamma,\ \gamma > 0\}$, $\tilde{\Phi}_\gamma = \sum_{i=1}^{N_\gamma} X_i$, is given beforehand. The aim is, on the basis of data, to pick an estimator within the family $\mathcal{E}$ which is of the same quality as the oracle estimator $\tilde{\Phi}_{\gamma_0}$ that attains $\inf_{\tilde{\Phi}\in\mathcal{E}} R(\tilde{\Phi}) = R(\tilde{\Phi}_{\gamma_0}) = \frac{\epsilon^2\log\epsilon^{-2}}{\gamma_0}(1+o(1))$ and has the same asymptotic risk as the Bayes estimator when γ_0 is known. This is exactly what the estimator $\hat{\Phi} = \tilde{\Phi}_{\tilde{\gamma}}$ does.

Remark 2. When constructing estimator for Φ, we correct the estimator by adding the term M_ϵ to $N_{\hat{\gamma}}$ which corresponds to shrinking the empirical Bayes smoothness $\hat{\gamma}$. At first sight this seems to be a technical manipulation, coming from the proof of the theorem. However, the necessity for shrinking the empirical Bayes smoothness $\hat{\gamma}$ stems from the fact that $\hat{\gamma}$ is "more likely" to be bigger than γ_0 than to be smaller; see the discussion in Remark 9. Roughly speaking, there is more probability in "oversmoothness zone" $\Gamma_+ = \{\gamma:\ \gamma > \gamma_0\}$, which makes it somewhat more difficult to separate different γ's in that zone. By adding M_ϵ to $N_{\hat{\gamma}}$ we effectively shrink $\hat{\gamma}$, i.e. shift it towards "undersmoothness zone" $\Gamma_- = \{\gamma:\ 0 < \gamma < \gamma_0\}$ (where things are easier) away from Γ_+.

Remark 3. As is easy to see, taking a bigger constant p in the definition of the sequence α_ϵ means a less restrictive requirement on α_ϵ. But it has an adverse effect as well. When we say that a certain relation holds for sufficiently small ϵ, we mean that there exists ϵ_0 such that this relation holds for all $\epsilon \in (0, \epsilon_0]$. In several such assertion, a bigger p leads to a smaller ϵ_0. In these assertions the implicit ϵ_0 should actually depend on the constant p. We will, however, skip this dependence on p to ease the notations.

Remark 4. Notice that we take the trimmed marginal likelihood corresponding to n_ϵ first number of observations. The observations beyond n_ϵ are non-informative, since the signal there θ_i, $i > n_\epsilon$, is undetectable: the so called *signal-to-noise* ratio $\frac{\sigma_i(\gamma_0)}{\epsilon^2}$ is small. We can ignore these observations in our inference on γ_0. Another thing is that the log-likelihood $\log L_\epsilon(\gamma)$ with ∞ instead of n_ϵ blows up to infinity almost surely for all $\gamma > 0$, so that the maximization problem over γ does not make sense. Interestingly, the series, even with ∞ instead of n_ϵ, $-Z_\epsilon(\gamma) =$

$2\log\big[L_\epsilon(\gamma)/\log L_\epsilon(\gamma_0)\big]$ does converge almost surely. Often in the literature, the sequence model (1) is formulated from the very beginning for n observations with $\epsilon = n^{-1/2}$.

Remark 5. We can draw a further correspondence with the minimax estimation over the exponential ellipsoid of unknown smoothness. Namely, a next interesting problem would be to study the frequentist property of the constructed adaptive estimator in the minimax context, i.e., how does its maximal risk relates asymptotically to the minimax risk, under model (1) with θ from an ellipsoid $\mathcal{E}_{\gamma_0}$. In fact, for each $\gamma < \gamma_0$ there exists some positive $Q = Q(\gamma, \gamma_0)$ such that, with π_{γ_0}-probability 1, θ belongs to the exponential ℓ_2-ellipsoid $\mathcal{E}_\gamma(Q)$. The closer γ to γ_0, the bigger the corresponding Q.

3 Technical lemmas

In this section we provide several technical lemmas which we need below.

Recall that $\hat\gamma$ is a (near) minimizer of $Z_\epsilon(\gamma)$ given by (11). The process $Z_\epsilon(\gamma)$ is the sum of two monotone processes

$$Z_1(\gamma) = Z_1(\gamma) + Z_2(\gamma)\,.$$

The first term $Z_1(\gamma)$ is a monotone increasing stochastic process:

$$Z_1(\gamma) = \sum_{i=1}^{n_\epsilon} \frac{(e^{-\gamma_0 i} - e^{-\gamma i})X_i^2}{(\epsilon^2 + e^{-\gamma i})(\epsilon^2 + e^{-\gamma_0 i})} = \sum_{i=1}^{n_\epsilon} a_i(\gamma) Y_i^2$$

with $n_\epsilon = \lceil \epsilon^{-2} \rceil$,

$$a_i(\gamma) = \frac{e^{-\gamma_0 i} - e^{-\gamma i}}{\epsilon^2 + e^{-\gamma i}} \qquad \text{and} \qquad Y_i = \frac{X_i}{\sqrt{\epsilon^2 + e^{-\gamma_0 i}}}\,,$$

the Y_i's are independent standard normal random variables, $Z_1(\gamma_0) = 0$. The second term $Z_2(\gamma)$ is a deterministic monotone decreasing function:

$$Z_2(\gamma) = \sum_{i=1}^{n_\epsilon} \log\Big[\frac{\epsilon^2 + e^{-\gamma i}}{\epsilon^2 + e^{-\gamma_0 i}}\Big]\,,$$

with $Z_2(\gamma_0) = 0$. The next lemma gives a uniform minorant for this function for sufficiently small ϵ.

Lemma 1. *There exist positive constants C, c and ϵ_0 (dependent only on γ_0) such that for all $0 < \epsilon \le \epsilon_0$*

$$Z_2(\gamma) \ge C\Big(\frac{\gamma_0}{\gamma} - 1\Big)\big(\log \epsilon^{-2}\big)^2 - c\log\epsilon^{-2}\,, \qquad \alpha_\epsilon \le \gamma < \gamma_0\,.$$

$$Z_2(\gamma) \ge C\Big(\frac{\gamma_0}{\gamma} - 1\Big)\big(\log \epsilon^{-2}\big)^2\,, \qquad \gamma \ge \gamma_0.$$

Proof. Recall that $\left|N_\gamma - \frac{1}{\gamma}\log\epsilon^{-2}\right| \le 1$ and if $\alpha_\epsilon \le \gamma < \gamma_0$, then $N_{\gamma_0} \le N_\gamma$ and $N_\gamma \le n_\epsilon$ for sufficiently small ϵ due to the restriction on the sequence α_ϵ. Using this, we obtain the following bound

$$\begin{aligned}
Z_2(\gamma) &= \sum_{i=1}^{n_\epsilon} \log\Big[\frac{\epsilon^2 + e^{-\gamma i}}{\epsilon^2 + e^{-\gamma_0 i}}\Big] \ge \sum_{i=1}^{N_\gamma} \log\Big[\frac{\epsilon^2 + e^{-\gamma i}}{\epsilon^2 + e^{-\gamma_0 i}}\Big] \\
&\ge \sum_{i=1}^{N_{\gamma_0}} \log\Big[\frac{e^{-\gamma i}}{2e^{-\gamma_0 i}}\Big] + \sum_{i=N_{\gamma_0}+1}^{N_\gamma} \log\Big[\frac{e^{-\gamma i}}{2\epsilon^2}\Big] \\
&= \sum_{i=1}^{N_{\gamma_0}} \big((\gamma_0 - \gamma)i - \log 2\big) + \sum_{i=N_{\gamma_0}+1}^{N_\gamma} \big(\log\epsilon^{-2} - \gamma i - \log 2\big) \\
&= \sum_{i=1}^{N_{\gamma_0}} \gamma_0 i - \sum_{i=1}^{N_\gamma} \gamma i + \sum_{i=N_{\gamma_0}+1}^{N_\gamma} \log\epsilon^{-2} - N_\gamma \log 2
\end{aligned}$$

for sufficiently small ϵ. The right hand side of the last inequality is equal to

$$\begin{aligned}
&\frac{\gamma_0 N_{\gamma_0}^2 - \gamma N_\gamma^2}{2} + (N_\gamma - N_{\gamma_0})\log\epsilon^{-2} - N_\gamma \log 2 \\
&= \frac{\gamma_0 - \gamma}{2\gamma\gamma_0}\big(\log\epsilon^{-2}\big)^2 - N_{\gamma_0}\log 2 - (N_\gamma - N_{\gamma_0})\log 2 + O(1) \\
&= \frac{\gamma_0 - \gamma}{2\gamma\gamma_0}\big(\log\epsilon^{-2}\big)^2 - \frac{\log 2}{\gamma_0}\log\epsilon^{-2} - \frac{\gamma_0 - \gamma}{\gamma\gamma_0}\log\epsilon^{-2} + O(1)
\end{aligned}$$

for sufficiently small ϵ.

For $\gamma \ge \gamma_0$ we have

$$\begin{aligned}
-Z_2(\gamma) &= \sum_{i=1}^{n_\epsilon} \log\Big[\frac{\epsilon^2 + e^{-\gamma_0 i}}{\epsilon^2 + e^{-\gamma i}}\Big] \\
&\le \sum_{i=1}^{N_\gamma} \Big(\log\Big[\frac{e^{-\gamma_0 i}}{e^{-\gamma i}}\Big] + \log\Big[\frac{1 + \epsilon^2 e^{\gamma_0 i}}{1 + \epsilon^2 e^{\gamma i}}\Big]\Big) \\
&\quad + \sum_{i=N_\gamma+1}^{N_{\gamma_0}} \log\Big[\frac{2e^{-\gamma_0 i}}{\epsilon^2}\Big] + \sum_{i=N_{\gamma_0}+1}^{\infty} \log\Big[\frac{\epsilon^2 + e^{-\gamma_0 i}}{\epsilon^2}\Big] \\
&\le \sum_{i=1}^{N_\gamma} (\gamma - \gamma_0)i + (N_{\gamma_0} - N_\gamma)\log(2\epsilon^{-2}) + \sum_{i=N_{\gamma_0}+1}^{\infty} \log\Big[1 + \frac{e^{-\gamma_0 i}}{\epsilon^2}\Big]
\end{aligned}$$

The first term is bounded by

$$(\gamma - \gamma_0)\frac{N_\gamma(N_\gamma + 1)}{2} = \frac{\gamma - \gamma_0}{2\gamma^2}\big(\log\epsilon^{-2}\big)^2 + O(1)\,,$$

the second and third by

$$\frac{\gamma-\gamma_0}{\gamma_0\gamma}\log(2\epsilon^{-2})+\epsilon^{-2}\sum_{i=N_{\gamma_0}+1}^{\infty}e^{-\gamma_0 i}+O(1)=\frac{\gamma-\gamma_0}{\gamma_0\gamma}\log\epsilon^{-2}+O(1)\,.$$

The lemma follows. □

The following lemma provides a bound on $EZ_1(\gamma)$.

Lemma 2. *For all* $\gamma\in(0,\gamma_0]$

$$EZ_1(\gamma)=\sum_{i=1}^{\infty}a_i(\gamma)\geq-\frac{\log\epsilon^{-2}+2}{\gamma}\,.$$

Proof. We have

$$\begin{aligned}-\sum_{i=1}^{n_\epsilon}a_i(\gamma) &= \sum_{i=1}^{n_\epsilon}\frac{e^{-\gamma i}-e^{-\gamma_0 i}}{\epsilon^2+e^{-\gamma i}}\leq\sum_{i=1}^{\infty}\frac{e^{-\gamma i}}{\epsilon^2+e^{-\gamma i}}\\ &\leq \sum_{i=1}^{N_\gamma}\frac{e^{-\gamma i}}{\epsilon^2+e^{-\gamma i}}+\sum_{i=N_\gamma+1}^{\infty}\frac{e^{-\gamma i}}{\epsilon^2+e^{-\gamma i}}\\ &\leq N_\gamma+\epsilon^{-2}\sum_{i=N_\gamma+1}^{\infty}e^{-\gamma i}\leq N_\gamma+\frac{1}{\gamma}\,.\end{aligned}$$

□

Introduce

$$a^2(\gamma)\stackrel{def}{=}\sum_{i=1}^{n_\epsilon}a_i^2(\gamma)=\frac{\mathrm{Var}(Z_\epsilon(\gamma))}{2}\,.$$

The following lemma provides bounds on $a^2(\gamma)$.

Lemma 3. *There exist positive constants* h, H *and* $\epsilon_0=\epsilon_0(\gamma_0)$, *such that for all* $0<\epsilon\leq\epsilon_0$ *and* $\gamma\in[\alpha_\epsilon,\gamma_0)$

$$\frac{h\log\epsilon^{-2}}{\gamma}-\frac{1}{\gamma_0-\gamma}\leq a^2(\gamma)\leq\frac{H\log\epsilon^{-2}}{\gamma}\,.$$

Proof. We bound $a^2(\gamma)$ first from above:

$$\begin{aligned}a^2(\gamma) &= \sum_{i=1}^{n_\epsilon}a_i^2(\gamma)=\sum_{i=1}^{n_\epsilon}\frac{(e^{-\gamma_0 i}-e^{-\gamma i})^2}{(\epsilon^2+e^{-\gamma i})^2}\\ &\leq \sum_{i=1}^{N_\gamma}1+C\epsilon^{-4}\sum_{N_\gamma+1}^{\infty}e^{-2\gamma i}\leq N_\gamma+\frac{c}{\gamma}\leq(c+1)N_\gamma\,,\end{aligned}$$

Now recall that, by the restriction on the sequence α_ϵ, $N_\gamma \le n_\epsilon$ for sufficiently small ϵ. Therefore,

$$\begin{aligned} a^2(\gamma) &= \sum_{i=1}^{n_\epsilon} \frac{(e^{-\gamma_0 i} - e^{-\gamma i})^2}{(\epsilon^2 + e^{-\gamma i})^2} \qquad (13) \\ &\ge \sum_{i=1}^{N_\gamma} \frac{e^{-2\gamma_0 i} - 2e^{-(\gamma_0+\gamma)i} + e^{-2\gamma i}}{4e^{-2\gamma i}} \\ &\ge \frac{1}{4}(e^{2(\gamma-\gamma_0)N_\gamma} + N_\gamma) - \frac{1}{2}\sum_{i=1}^{N_\gamma} e^{(\gamma-\gamma_0)i} \qquad (14) \end{aligned}$$

for sufficiently small ϵ. The lemma follows. □

Remark 6. The above two lemmas provide bounds for $EZ_1(\gamma)$ and $\mathrm{Var}\big(Z_1(\gamma)\big) = \mathrm{Var}\big(Z_\epsilon(\gamma)\big)$ only for the case $0 < \gamma < \gamma_0$. In further considerations we will not need bounds for the case $\gamma > \gamma_0$. It is however of some interest to look at the asymptotic behavior of the process $Z_\epsilon(\gamma)$ for $\gamma > \gamma_0$. As a matter of fact, both the expectation and the oscillation (the variance) of the process $Z_\epsilon(\gamma)$ become of a much bigger order compared to the case $\gamma < \gamma_0$. Another peculiarity is that this asymptotic behavior is essentially determined by just one term $a_{N_\gamma}(\gamma)Y^2_{N_\gamma}$ for each $\gamma > \gamma_0$, so one can not use results based on central limit theorem approximations.

Indeed, we evaluate $EZ_1(\gamma)$ as follows: for $\gamma > \gamma_0$,

$$\begin{aligned} EZ_1(\gamma) &= \sum_{i=1}^{n_\epsilon} \frac{e^{-\gamma_0 i} - e^{-\gamma i}}{\epsilon^2 + e^{-\gamma i}} \ge \sum_{i=1}^{N_\gamma} \frac{e^{-\gamma_0 i} - e^{-\gamma i}}{\epsilon^2 + e^{-\gamma i}} \\ &\ge \sum_{i=1}^{N_\gamma} \Big(\frac{e^{-\gamma_0 i}}{2e^{-\gamma i}} - 1\Big) \ge c e^{(\gamma-\gamma_0)N_\gamma} - N_\gamma \\ &\ge c\big(\epsilon^{-2}\big)^{1-\frac{\gamma_0}{\gamma}} - \frac{\log \epsilon^{-2}}{\gamma} \end{aligned}$$

for sufficiently small ϵ. Similarly to (14), we obtain for $\gamma > \gamma_0$ that

$$\begin{aligned} a^2(\gamma) &\ge \frac{1}{4}(e^{2(\gamma-\gamma_0)N_\gamma} + N_\gamma) - \frac{1}{2}\sum_{i=1}^{N_\gamma} e^{(\gamma-\gamma_0)i} \\ &\ge C\big(\epsilon^{-4}\big)^{1-\frac{\gamma_0}{\gamma}} - \frac{c\big(\epsilon^{-2}\big)^{1-\frac{\gamma_0}{\gamma}}}{\gamma - \gamma_0} \end{aligned}$$

for sufficiently small ϵ.

4 Estimation of smoothness

Although the problem of estimating γ_0 is not a primary goal in this paper, it is interesting on its own right. In fact, from the Bayesian perspective, observations (1)

are simply

$$X_i = e^{-\frac{\gamma_0 i}{2}}\eta_i + \epsilon\,\xi_i \quad i = 1, 2, \dots ,$$

where the η_i's and ξ_i's are independent standard normal random variables, $\gamma_0 > 0$ is unknown parameter. Thus, the X_i's are independent but linked to each other through parameter γ_0, each observation carries different amount information about γ_0. No wonder that, from the Bayesian perspective, the true smoothness parameter γ_0 can be consistently estimated by applying empirical Bayes approach, while in the minimax context this is in general not possible. In this section we also discuss shortly the phenomenon of under- and oversmoothing, which correspond to events $\{\hat\gamma < \gamma_0\}$ and $\{\hat\gamma > \gamma_0\}$ with $\hat\gamma$ being a (near) maximum marginal likelihood estimator of γ_0.

First introduce some notations. Denote for nonnegative numbers D and k

$$b_\epsilon(k) = b_\epsilon(k, D) = \frac{\gamma_0(D\log\log\epsilon^{-2} + k)}{\log\epsilon^{-2}}\,,$$

$$\mathcal{K} = \mathcal{K}(D) = \{k \in \mathbb{N} :\ b_\epsilon(k) < 1\} = \{k \in \mathbb{N} :\ k < N_{\gamma_0} - D\log\log\epsilon^{-2}\}\,,$$

and for $k \in \mathcal{K} = \big\{0, 1, \dots, \lfloor N_{\gamma_0} - D\log\log\epsilon^{-2}\rfloor\big\}$ define finally

$$\gamma_u(k) = \gamma_u(k, D, \epsilon) = \frac{\gamma_0}{1 - b_\epsilon(k)}\,.$$

Notice that $\gamma_u(k) \ge \gamma_0$ and $\gamma_u(k) \to \gamma_0$ as $\epsilon \to 0$ for all $D > 0$ and $k \in \mathcal{K}(D)$.

Lemma 4. *For any fixed $m > 0$ and any $D > 4/\gamma_0$ there exists $\epsilon_0 = \epsilon_0(\gamma_0, m, D)$ such that for all $0 < \epsilon < \epsilon_0$ and $k \in \mathcal{K}$*

$$\begin{aligned} P\big\{N_{\hat\gamma} \;&\le\; N_{\gamma_0} - \big(k + D\log\log\epsilon^{-2}\big)\big\} \le P\big\{\hat\gamma \ge \gamma_u(k)\big\} \\ &\le\; (2/\pi)^{m/2}\big(\log\epsilon^{-2}\big)^{-m} e^{-\gamma_0 m k/2}\,. \end{aligned}$$

Proof. Using the formula for N_γ, we obtain the following implication:

$$\Big\{N_{\hat\gamma} \le N_{\gamma_0} - \big(k + D\log\log\epsilon^{-2}\big)\Big\} \subseteq \big\{\hat\gamma \ge \gamma_u(k)\big\}$$

for $k \in \mathcal{K}$. Since $0 = Z_\epsilon(\gamma_0) = Z_1(\gamma_0) + Z_2(\gamma_0)$ and $\hat\gamma$ is a near minimizer of $Z_\epsilon(\gamma)$, we have further that $Z_\epsilon(\hat\gamma) \le Z_\epsilon(\gamma_0) + \epsilon^2 = \epsilon^2$ for sufficiently small ϵ. Indeed, $\inf_{\gamma\ge\alpha_\epsilon} Z_\epsilon(\gamma) \le Z_\epsilon(\gamma_0)$ if $\gamma_0 \ge \alpha_\epsilon$, which holds for all $\epsilon \in (0, \epsilon_\gamma)$ with ϵ_γ defined by $\alpha_{\epsilon_\gamma} = \gamma_0$). Therefore

$$\big\{\hat\gamma \ge \gamma_u(k)\big\} \subseteq \Big\{\inf_{\gamma\ge\gamma_u(k)} Z_\epsilon(\gamma) \le Z_\epsilon(\hat\gamma)\Big\} \subseteq \Big\{\inf_{\gamma\ge\gamma_u(k)} Z_\epsilon(\gamma) \le \epsilon^2\Big\}\,.$$

Thus it is sufficient to prove that for any fixed $m > 0$ and D, $D > 4/\gamma_0$, there exists $\epsilon_0 = \epsilon(\gamma_0, m, D)$ such that for all ϵ, $0 < \epsilon \le \epsilon_0$ and $k \in \mathcal{K}$

$$P\Big\{\inf_{\gamma\ge\gamma_u(k)} Z_\epsilon(\gamma) \le \epsilon^2\Big\} \le (2/\pi)^{m/2}\big(\log\epsilon^{-2}\big)^{-m} e^{-\gamma_0 m k/2}\,. \tag{15}$$

Without loss of generality suppose that $m \in \mathbb{N}$. Recall that

$$|N_\gamma - \gamma^{-1} \log \epsilon^{-2}| \le 1 \,.$$

As $k \in \mathcal{K}$ and $\gamma \ge \gamma_u(k)$, $\gamma \ge \gamma_u(k) > \gamma_0$. Therefore, for any fixed $0 \le l \le m$ there exists a constant $c = c(\gamma_0, m)$ such that for sufficiently small ϵ

$$\begin{aligned} Z_1(\gamma) &= \sum_{i=1}^{n_\epsilon} a_i(\gamma) Y_i^2 \ge a_{N_\gamma + l}(\gamma) Y_{N_\gamma+l}^2 \\ &= \frac{e^{-\gamma_0(N_\gamma+l)} - e^{-\gamma(N_\gamma+l)}}{\epsilon^2 + e^{-\gamma(N_\gamma+l)}} Y_{N_\gamma+l}^2 \\ &\ge c\big((\epsilon^{-2})^{1-\frac{\gamma_0}{\gamma}} - 1\big) Y_{N_\gamma+l}^2 \end{aligned} \tag{16}$$

since $N_\gamma + l \le n_\epsilon$ for sufficiently small ϵ. In fact N_γ yields, as one can easily check, the $\max_{i \ge 1} a_i(\gamma)$ up to a constant term.

Notice now that the Y_i's are independent standard normal random variables. Therefore the following trivial inequality holds for any positive δ and $M, m \in \mathbb{N}$:

$$P\Big\{ \max_{M \le i \le M+m} Y_i^2 \le \delta^2 \Big\} = \Big(P\big\{ Y_M^2 \le \delta^2 \big\} \Big)^m \le (2/\pi)^{m/2} \delta^m \,. \tag{17}$$

Since process $Z_1(\gamma)$ is monotonically increasing, we have further that

$$\begin{aligned} \Big\{ \inf_{\gamma \ge \gamma_u(k)} Z_\epsilon(\gamma) \le \epsilon^2 \Big\} &\subseteq \Big\{ \inf_{\gamma \ge \gamma_u(k)} Z_1(\gamma) \le \sup_{\gamma \ge \gamma_u(k)} \big(- Z_2(\gamma) \big) + \epsilon^2 \Big\} \\ &\subseteq \Big\{ Z_1(\gamma_u(k)) \le \sup_{\gamma \ge \gamma_u(k)} \big(- Z_2(\gamma) \big) + \epsilon^2 \Big\} \\ &\subseteq \Big\{ \max_{N_{\gamma_u(k)} \le i \le N_{\gamma_u(k)}+m} Y_i^2 \le \delta^2 \Big\} \cup A_\epsilon \,, \end{aligned} \tag{18}$$

with

$$A_\epsilon = A_{\epsilon,\delta} = \Big\{ \max_{N_{\gamma_u(k)} \le i \le N_{\gamma_u(k)}+m} Y_i^2 > \delta^2,\ Z_1(\gamma_u(k)) \le \sup_{\gamma \ge \gamma_u(k)} \big(- Z_2(\gamma) \big) + \epsilon^2 \Big\} .$$

By using Lemma 1 and (16), we obtain for the constants $C = C(\gamma_0)$ from Lemma 1 and $c = c(\gamma_0, m)$ from (16)

$$\begin{aligned} A_\epsilon &\subseteq \big\{ c\big((\epsilon^{-2})^{1-\frac{\gamma_0}{\gamma_u(k)}} - 1 \big) \delta^2 \le C \big(\log \epsilon^{-2} \big)^2 + \epsilon^2 \big\} \\ &\subseteq \big\{ e^{\gamma_0 (D \log\log \epsilon^{-2} + k)} \le C c^{-1} \big(\log \epsilon^{-2} \big)^2 \delta^{-2} + c^{-1} \epsilon^2 \delta^{-2} + 1 \big\} \end{aligned}$$

Take now

$$\delta = \big(\log \epsilon^{-2} \big)^{-1} e^{-\gamma_0 k/2} \,. \tag{19}$$

Then, if $D > \frac{4}{\gamma_0}$ and ϵ is sufficiently small,

$$\begin{aligned} A_\epsilon &\subseteq \big\{ e^{\gamma_0 (D \log\log \epsilon^{-2} + k)} \le C c^{-1} \big(\log \epsilon^{-2} \big)^2 \delta^{-2} + c^{-1} \epsilon^2 \delta^{-2} + 1 \big\} \\ &\subseteq \big\{ \big(\log \epsilon^{-2} \big)^{D\gamma_0} \le C c^{-1} \big(\log \epsilon^{-2} \big)^4 + c^{-1} \epsilon^2 \big(\log \epsilon^{-2} \big)^2 e^{-\gamma_0 k/2} + e^{-\gamma_0 k} \big\} = \varnothing \end{aligned}$$

and consequently $P(A_\epsilon) = 0$ for sufficiently small ϵ, i.e. for all ϵ such that $0 < \epsilon \le \epsilon_0(\gamma_0, m, D)$. Combining this with (17), (18) and (19), we obtain (15), which completes the proof. □

Remark 7. The last lemma characterizes in fact the convergence rate of $\hat{\gamma}$ to γ_0 from above, the so called "oversmoothing" effect. The result is given in terms of N_γ, which is suitable in the sequel when proving the main result. We can however formulate the result for $\hat{\gamma}$ as well. For any $c > 0$ there exists $0 < x_0 < 1$ such that $(1-x)^{-1} \le 1 + (1+c)x$ for all $x \in (0, x_0)$. Recall also that in the conditions of the above lemma $D\gamma_0 > 4$. Now fix any $c > 0$. Then for any $k \in \mathcal{K}$ we obtain that

$$\begin{aligned} \{\hat{\gamma} \ge \gamma_u(k)\} &= \Big\{\hat{\gamma} \ge \frac{\gamma_0}{1 - b_\epsilon(k)}\Big\} \supseteq \{\hat{\gamma} \ge \gamma_0 + \gamma_0(1+c)b_\epsilon(k)\} \\ &\supseteq \big\{ \log \epsilon^{-2}(\hat{\gamma} - \gamma_0) > 4(1+c)\gamma_0 \log\log \epsilon^{-2} + \gamma_0^2 k\big\} \end{aligned}$$

for sufficiently small ϵ. The precise final assertion is then as follows. For any fixed $m > 0$ any $D > 4/\gamma_0$ and any $\nu > 0$ there exists positive $\epsilon_0 = \epsilon_0(\gamma_0, m, D, \nu)$ such that for all $0 < \epsilon < \epsilon_0$ and $k \in \mathcal{K}(D)$

$$P\big\{ \log \epsilon^{-2}(\hat{\gamma} - \gamma_0) > \gamma_0^2 D(1+\nu)\log\log \epsilon^{-2} + \gamma_0^2 k\big\} \le \frac{(2/\pi)^{m/2}}{\big(\log \epsilon^{-2}\big)^m e^{\gamma_0 m k/2}}\,. \quad (20)$$

As one can see the rate of convergence of $\hat{\gamma}$ to γ_0 from above is $\frac{\log \epsilon^{-2}}{\log\log \epsilon^{-2}}$ (take $k = 0$).

Remark 8. It is of course desirable to take a big m, but one should relate this to the corresponding ϵ_0 which depends on m. Tracing the proof of the above lemma, one can see that the bigger m is, the smaller ϵ_0 becomes: bigger m leads to smaller constant c in (16), which in turn makes the event A_ϵ impossible only for $0 < \epsilon \le \epsilon_0$ with a smaller ϵ_0.

On the other hand, for a fixed m we can make ϵ_0 bigger by taking a bigger $D = D(m)$, effectively at a sacrifice in size of the set $\mathcal{K}(D)$. As to the constant ν, the smaller ν we take, the smaller ϵ_0 we get.

In the proof of the following lemma we will make use of the following version of the result from Freedman (1999).

Proposition 1 (Freedman). *Let U_i's be independent $N(0,1)$ variables. Let c_i's be real numbers with $c^2 = \sum_{i=1}^n c_i^2 < \infty$. Let $\rho > 0$ with $\rho|c_i|/c^2 < 1$ for all i, and let $V = \sum_{i=1}^n c_i(U_i^2 - 1)$. Then*

$$P\{V > \rho\} \le \exp\big\{ -\rho^2/(12c^2)\big\} \quad \textit{and} \quad P\{V < -\rho\} \le \exp\big\{ -\rho^2/(12c^2)\big\}\,.$$

The corresponding result in Freedman (1999) (appears as Lemma 4 in this paper) deals with the infinite sums $c^2 = \sum_{i=1}^\infty c_i^2$ and $V = \sum_{i=1}^\infty c_i(U_i^2 - 1)$. We skip the proof of the above proposition because it is exactly the same as the proof of Lemma 4 in Freedman (1999).

Lemma 5. *There exists positive ϵ_0, C_0, C and ρ (all depending on γ_0 only) such that for all $0 < \epsilon < \epsilon_0$*

$$P\big\{ \log \epsilon^{-2}(\gamma_0 - \hat{\gamma}) \geq C_0 \big\} \leq C\big(\log \epsilon^{-2} \big)^p \epsilon^{\rho} .$$

Proof. Let $u_0 = \gamma_0 - \frac{C_0}{\log \epsilon^{-2}}$, $u_m = \alpha_\epsilon$, $u_m < u_{m-1} < \ldots < u_1 < u_0$ and $I_i = [u_i, u_{i-1}]$, $i = 1, \ldots, m$. Split the interval $I = \big[\alpha_\epsilon, \gamma_0 - \frac{C_0}{\log \epsilon^{-2}}\big] = \cup_{i=1}^m I_i$.

Now, recall that $Z_1(\gamma)$ strictly increasing, $Z_2(\gamma)$ strictly decreasing functions respectively, and $Z_\epsilon(\gamma_0) = 0$.

$$\begin{aligned} \big\{ \hat{\gamma} \leq \gamma_0 - C_0\big(\log \epsilon^{-2} \big)^{-1} \big\} &\subseteq \cup_{i=1}^m \big\{ \inf_{\gamma \in I_i} Z_\epsilon(\gamma) \leq Z_\epsilon(\gamma_0) + \epsilon^2 \big\} \\ &\subseteq \cup_{i=1}^m \big\{ \inf_{\gamma \in I_i} (Z_1(\gamma)) \leq \sup_{\gamma \in I_i} (-Z_2(\gamma)) + \epsilon^2 \big\} \\ &= \cup_{i=1}^m \big\{ Z_1(u_i) \leq (-Z_2(u_{i-1})) + \epsilon^2 \big\} . \end{aligned}$$

Introduce the centered version of the process $Z_1(\gamma)$:

$$\bar{Z}_1(\gamma) = Z_1(\gamma) - EZ_1(\gamma) .$$

Making use of technical Lemmas 1 and 2, we proceed as follows: for sufficiently small ϵ,

$$\begin{aligned} &\cup_{i=1}^m \big\{ Z_1(u_i) \leq (-Z_2(u_{i-1})) + \epsilon^2 \big\} \\ &\quad = \cup_{i=1}^m \big\{ \bar{Z}_1(u_i) \leq (-Z_2(u_{i-1})) - EZ_1(u_i) + \epsilon^2 \big\} \\ &\quad \subseteq \cup_{i=1}^m \Big\{ \bar{Z}_1(u_i) \leq -\frac{C(\gamma_0 - u_{i-1})\big(\log \epsilon^{-2} \big)^2}{u_{i-1}} + c \log \epsilon^{-2} + \frac{\log \epsilon^{-2} + 2}{u_i} + \epsilon^2 \Big\} \\ &\quad = \cup_{i=1}^m A_i , \end{aligned}$$

say, where constants $C = C(\gamma_0)$ and $c = c(\gamma_0)$ are from Lemma 1.

From the last two relations we have

$$P\big\{ \log \epsilon^{-2}(\gamma_0 - \hat{\gamma}) \geq C_0 \big\} \leq \sum_{i=1}^m P(A_i) . \tag{21}$$

Introduce $\Delta_\epsilon = \big(\log \epsilon^{-2} \big)^{-p}$, where constant p appears in the definition of α_ϵ. Now take the following values of $u_1, \ldots, u_{m-1}$ (u_0 and u_m are already defined): $u_1 = \gamma_0/2$, $u_i = u_{i-1} + \Delta_\epsilon$, $i = 2, 3, \ldots, m-1$, so that $u_{m-1} - u_m \leq \Delta_\epsilon$. We can easily evaluate the number $m = m_\epsilon$:

$$m \leq \frac{\gamma_0/2}{\Delta_\epsilon} + 2 = \frac{\gamma_0\big(\log \epsilon^{-2} \big)^p}{2} + 2 . \tag{22}$$

For the first event from the right hand side of (21), we obtain

$$\begin{aligned} P(A_1) &= P\Big\{ \bar{Z}_1(u_1) \leq -\frac{C(\gamma_0 - u_0)\big(\log \epsilon^{-2} \big)^2}{u_0} + c \log \epsilon^{-2} + \frac{\log \epsilon^{-2} + 2}{u_1} + \epsilon^2 \Big\} \\ &\leq P\Big\{ \bar{Z}_1(u_1) \leq -\frac{CC_0 \log \epsilon^{-2}}{\gamma_0} + \frac{(2 + c\gamma_0) \log \epsilon^{-2}}{\gamma_0} + \frac{4}{\gamma_0} + \epsilon^2 \Big\} . \end{aligned} \tag{23}$$

For sufficiently large $C_0 > 0$ (in fact depending on constants appearing in the bounds in Lemmas 1 and 3) and sufficiently small ϵ, by Lemma 3 we have that

$$\frac{(CC_0 - 2 - c\gamma_0)\log\epsilon^{-2}}{\gamma_0} - \frac{4}{\gamma_0} - \epsilon^2 > \frac{2H\log\epsilon^{-2}}{\gamma_0} \geq a^2(u_1)\,.$$

Combining (23) with the last relation implies that

$$P(A_1) \leq P\big\{\bar{Z}_1(u_1) \leq -a^2(u_1)\big\}\,. \tag{24}$$

Recall that

$$\bar{Z}_1(u_i) = Z_1(u_i) - EZ_1(u_i) = \sum_{k=1}^{n_\epsilon} a_k(u_i)(Y_k^2 - 1)\,,$$

$a^2(\gamma) = \sum_{k=1}^{n_\epsilon} a_k^2(\gamma)$, and $|a_k(\gamma)| \leq 1$ for all k and all $0 < \gamma < \gamma_0$. Therefore for all u_i's and all k, $a^2(u_i)|a_k(u_i)| \leq a^2(u_i)$ so that we are in the position to apply Proposition 1 to the right hand side of (24):

$$P(A_1) \leq P\big\{Z_1(u_1) - EZ_1(u_1) \leq -a^2(u_1)\big\} \leq \exp\Big\{-\frac{a^2(u_1)}{12}\Big\}\,. \tag{25}$$

By Lemma 2 we bound $a^2(u_1)$:

$$a^2(u_1) \geq \frac{h\log\epsilon^{-2}}{u_1} - \frac{1}{\gamma_0 - u_1} = \frac{2h\log\epsilon^{-2}}{\gamma_0} - \frac{2}{\gamma_0}$$

Using the last two relation, we obtain that

$$P(A_1) \leq \exp\Big\{-\frac{a^2(u_1)}{12}\Big\} \leq C_1\epsilon^\rho \tag{26}$$

with $C_1 = e^{1/(6\gamma_0)}$, $\rho = h/(3\gamma_0)$, for sufficiently large C_0 and sufficiently small ϵ.

Now, for $i = 2, \ldots, m$ we have obviously

$$\begin{aligned} P(A_i) &= P\Big\{\bar{Z}_1(u_i) \leq -\frac{C(\gamma_0 - u_{i-1})\big(\log\epsilon^{-2}\big)^2}{u_{i-1}} + c\log\epsilon^{-2} + \frac{\log\epsilon^{-2} + 2}{u_i} + \epsilon^2\Big\} \\ &\leq P\Big\{\bar{Z}_1(u_i) \leq -\frac{C(\gamma_0/2)\big(\log\epsilon^{-2}\big)^2}{u_{i-1}} + c\log\epsilon^{-2} + \frac{\log\epsilon^{-2} + 2}{u_i} + \epsilon^2\Big\}\,. \end{aligned}$$

Since $\Delta_\epsilon \leq \alpha_\epsilon = u_m < u_{m-1} < \cdots < u_2$, we have for $i = 2, \ldots, m$

$$\begin{aligned} &-\frac{C(\gamma_0/2)\big(\log\epsilon^{-2}\big)^2}{u_{i-1}} + \frac{\log\epsilon^{-2} + 2}{u_i} + c\log\epsilon^{-2} + \epsilon^2 \\ &= \frac{-C(\gamma_0/2)u_i\big(\log\epsilon^{-2}\big)^2 + (u_i + \Delta_\epsilon)(\log\epsilon^{-2} + 2)}{u_i(u_i + \Delta_\epsilon)} + c\log\epsilon^{-2} + \epsilon^2 \\ &\leq -\frac{C_2\big(\log\epsilon^{-2}\big)^2}{u_i + \Delta_\epsilon} + c\log\epsilon^{-2} + \epsilon^2 \leq -\frac{H\log\epsilon^{-2}}{u_i} \leq -a^2(u_i) \end{aligned}$$

for sufficiently small ϵ. Therefore, the last two inequalities imply that, similarly to (25), we can again apply Proposition 1:

$$P(A_i) \le P\big\{\bar{Z}_1(u_i) \le -a^2(u_i)\big\} \le \exp\Big\{-\frac{a^2(u_i)}{12}\Big\},$$

for $i = 2, \dots, m_\epsilon$ and sufficiently small ϵ. Now, by Lemma 3, we have that for $i = 2, \dots, m$

$$a^2(u_i) \ge \frac{h \log \epsilon^{-2}}{u_i} - \frac{1}{\gamma_0 - u_i} \ge \frac{h \log \epsilon^{-2}}{u_1} - \frac{1}{\gamma_0 - u_1} = \frac{2h \log \epsilon^{-2}}{\gamma_0} - \frac{2}{\gamma_0}.$$

Thus, similarly to (26), we obtain that

$$P(A_i) \le C_1 \epsilon^\rho, \qquad i = 2, \dots, m,$$

for sufficiently small ϵ. Combining the last relation with (21), (22), (26), we conclude that the lemma is proved. □

The next corollary in terms of $N_{\hat{\gamma}}$ and N_{γ_0} follows from the above lemma.

Corollary 1. *There exist positive constants $D = D(\gamma_0)$, $c = c(\gamma_0)$, $\rho = \rho(\gamma_0)$ and $\epsilon_0 = \epsilon_0(\gamma_0)$ such that for all $0 < \epsilon < \epsilon_0$*

$$P\big\{N_{\hat{\gamma}} \ge N_{\gamma_0} + D\big\} \le c\big(\log \epsilon^{-2}\big)^p \epsilon^\rho.$$

Proof. Take $D = \frac{2C_0}{\gamma_0^2}$, where constant C_0 comes from the above lemma. Note

$$\big\{N_{\hat{\gamma}} \ge N_{\gamma_0} + D\big\} = \big\{\log \epsilon^{-2}(\gamma_0 - \hat{\gamma}) \ge \hat{\gamma}\gamma_0 D\big\}$$

We bound the probability of the above event by the sum

$$P\big\{\log \epsilon^{-2}(\gamma_0 - \hat{\gamma}) \ge C_0\big\} + P\Big\{\log \epsilon^{-2}(\gamma_0 - \hat{\gamma}) \ge \hat{\gamma}\gamma_0 D, \hat{\gamma} \ge \gamma_0 - \frac{C_0}{\log \epsilon^{-2}}\Big\}.$$

The second term from the right hand side of the last inequality is bounded from above by

$$P\Big\{\log \epsilon^{-2}(\gamma_0 - \hat{\gamma}) \ge \frac{\gamma_0^2 D}{2}\Big\} = P\big\{\log \epsilon^{-2}(\gamma_0 - \hat{\gamma}) \ge C_0\big\},$$

for sufficiently small ϵ. Combining all the relations, we obtain that

$$P\big\{N_{\hat{\gamma}} \ge N_{\gamma_0} + D\big\} \le 2P\big\{\log \epsilon^{-2}(\gamma_0 - \hat{\gamma}) \ge C_0\big\}$$

for sufficiently small ϵ. Now, applying the lemma yields the result. □

Remark 9. The lemma claims the convergence rate $\log \epsilon^{-2}$ for $\hat{\gamma}$ to γ_0 from below. Compared with the rate of convergence of $\hat{\gamma}$ to γ_0 from above (see relation (20) in Remark 7), we see that the rate from below is faster by factor $\log\log \epsilon^{-2}$ than that from above.

Examining the proof of the last lemma, one can see that a further improvement upon the rate of convergence of $\hat{\gamma}$ to γ_0 from below seems feasible. We believe it should be up to the rate $(\log \epsilon^{-2})^{3/2}$ by using maximal inequality for the the process $Z_\epsilon(\gamma)$. This will be studied in detail elsewhere. Here we used rather rough estimates to derive the above rate of convergence, since it suffices in the proof of the main result.

We believe this nonsymmetric behavior of the estimator $\hat{\gamma}$ is intrinsic and connected to the embedded model structure: it is "easier" for the method to separate the true smoothness model from the less smooth models rather than from smoother models. In other words, the method tends to oversmooth rather than undersmooth, however both are under control. Similar phenomenon occurs in the minimax context in Lepski's adaptation method (see Lepski (1992) and further references therein) in somewhat more dramatic form. Namely, Lepski's methods detects undersmoothing easily, while oversmoothing can not be in general controlled. The latter problem in this method is handled by the special construction of estimator based on the comparisons of certain statistics evaluated at different values of the smoothness parameter from a fine grid.

5 Proof of the main theorem

We are now ready to prove the main theorem. Write

$$\begin{aligned} R_\epsilon(\hat{\Phi}) &= E(\hat{\Phi}-\Phi)^2 \\ &= E\Big[(\hat{\Phi}-\Phi)^2 I\{N_{\gamma_0}-M_\epsilon \le N_{\hat{\gamma}} \le N_{\gamma_0}+M_\epsilon\}\Big] \\ &\quad + E\Big[(\hat{\Phi}-\Phi)^2 I\{0 \le N_{\hat{\gamma}} < N_{\gamma_0}-M_\epsilon\}\Big] \\ &\quad + E\Big[(\hat{\Phi}-\Phi)^2 I\{N_{\gamma_0}+M_\epsilon < N_{\hat{\gamma}} \le N_{\alpha_\epsilon}\}\Big] = R_1+R_2+R_3\,, \quad (27) \end{aligned}$$

say. Here by $I\{E\}$ we denote the indicator of the event E. Let us evaluate each of these terms. Denote

$$I_1 = I\{N_{\gamma_0}-M_\epsilon \le N_{\hat{\gamma}} \le N_{\gamma_0}+M_\epsilon\}\,.$$

Using the elementary inequality

$$(a+b)^2 \le (1+\beta)\,a^2 + (1+\beta^{-1})\,b^2, \qquad 0<\beta\le 1\,,$$

we obtain

$$\begin{aligned} R_1 &= E\Big[\Big(\sum_{i=1}^{N_{\hat\gamma}+M_\epsilon} X_i - \sum_{i=1}^{\infty}\theta_i\Big)^2 I_1\Big] \\ &= E\Big[\Big(\sum_{i=1}^{N_{\gamma_0}}(X_i-\theta_i) + \sum_{i=N_{\gamma_0}+1}^{N_{\hat\gamma}+M_\epsilon}(X_i-\theta_i) - \sum_{i=N_{\hat\gamma}+M_\epsilon+1}^{\infty}\theta_i\Big)^2 I_1\Big] \\ &\le (1+\beta_\epsilon)\epsilon^2 E\Big[\Big(\sum_{i=1}^{N_{\gamma_0}}\xi_i\Big)^2 I_1\Big] \\ &\quad + 2(1+\beta_\epsilon^{-1})\epsilon^2 E\Big[\Big(\sum_{i=N_{\gamma_0}+1}^{N_{\hat\gamma}+M_\epsilon}\xi_i\Big)^2 I_1\Big] \\ &\quad + 2(1+\beta_\epsilon^{-1})E\Big[\Big(\sum_{i=N_{\hat\gamma}+M_\epsilon+1}^{\infty}\theta_i\Big)^2 I_1\Big] = R_{11}+R_{12}+R_{13}\,, \end{aligned} \tag{28}$$

where sequence $0<\beta_\epsilon\le 1$ is chosen in such a way that

$$\beta_\epsilon \to 0 \qquad \text{and} \qquad \beta_\epsilon^{-1} M_\epsilon \big(\log\epsilon^{-2}\big)^{-1} \to 0 \qquad \text{as} \quad \epsilon\to 0\,.$$

Such a choice of β_ϵ is possible because of the condition on M_ϵ: M_ϵ converges to infinity slower than $\log\epsilon^{-2}$. Then, as $\epsilon\to 0$,

$$R_{11} \le \epsilon^2 N_{\gamma_0}(1+o(1)) = \frac{\epsilon^2\log\epsilon^{-2}}{\gamma_0}(1+o(1))\,. \tag{29}$$

To bound R_{12}, note first that $N_{\hat\gamma}\le N_{\gamma_0}+M_\epsilon$ on I_1 and that M_l, $l=N_{\gamma_0}+1,\dots,N_{\gamma_0}+2M_\epsilon$, with $M_l=\sum_{i=N_{\gamma_0}+1}^{l}\xi_i$, is a martingale (with respect to the natural filtration). Applying the L_q-maximal inequality for submartingales (see for example Williams (1991), p. 143), we get

$$\begin{aligned} E\Big[\Big(\sum_{i=N_{\gamma_0}+1}^{N_{\hat\gamma}+M_\epsilon}\xi_i\Big)^2 I_1\Big] &\le E\Big[\max_{N_{\gamma_0}+1\le l\le N_{\gamma_0}+2M_\epsilon}\Big(\sum_{i=N_{\gamma_0}+1}^{l}\xi_i\Big)^2\Big] \\ &\le 4E\Big[\sum_{i=N_{\gamma_0}+1}^{N_{\gamma_0}+2M_\epsilon}\xi_i\Big]^2 = 8M_\epsilon\,, \end{aligned}$$

which implies that, as $\epsilon\to 0$,

$$R_{12} = O\big(\epsilon^2\beta_\epsilon^{-1}M_\epsilon\big) = o\big(\epsilon^2\log\epsilon^{-2}\big)\,. \tag{30}$$

In a similar way we bound R_{13}. $M_l'=\sum_{i=N_{\gamma_0}+1}^{l}\theta_i$, $l=N_{\gamma_0}+1,\dots,N_{\gamma_0}+2M_\epsilon$,

is also a martingale. Again, by the maximal inequality for submartingales, we obtain

$$\begin{aligned} E\Big[\Big(\sum_{i=N_{\hat\gamma}+M_\epsilon+1}^{\infty}\theta_i\Big)^2 I_1\Big] &= E\Big[\Big(\sum_{i=N_{\gamma_0}+1}^{\infty}\theta_i - \sum_{i=N_{\gamma_0}+1}^{N_{\hat\gamma}+M_\epsilon}\theta_i\Big)^2 I_1\Big] \\ &\le 2E\Big[\sum_{i=N_{\gamma_0}+1}^{\infty}\theta_i\Big]^2 + 2E\Big[\Big(\sum_{i=N_{\gamma_0}+1}^{N_{\hat\gamma}+M_\epsilon}\theta_i\Big)^2 I_1\Big] \\ &\le 2\gamma_0^{-1}\epsilon^2 + 16\epsilon^2 M_\epsilon\,. \end{aligned}$$

Thus, as $\epsilon \to 0$,

$$R_{13} = O\big(\beta_\epsilon^{-1}\epsilon^2 M_\epsilon\big) = o\big(\epsilon^2 \log \epsilon^{-2}\big)\,.$$

The last relation, together with (28), (29) and (30), gives

$$R_1 \le \frac{\epsilon^2 \log \epsilon^{-2}}{\gamma_0}(1+o(1))\,. \tag{31}$$

Consider the term R_2 from (27). Denote $I_2 = I\{0 \le N_{\hat\gamma} < N_{\gamma_0} - M_\epsilon\}$. Calculate

$$\begin{aligned} R_2 &= E\Big[\Big(\sum_{i=1}^{N_{\hat\gamma}+M_\epsilon} X_i - \sum_{i=1}^{\infty}\theta_i\Big)^2 I_2\Big] \\ &= E\Big[\Big(\sum_{i=1}^{N_{\hat\gamma}+M_\epsilon}(X_i-\theta_i) - \sum_{i=N_{\hat\gamma}+M_\epsilon+1}^{\infty}\theta_i\Big)^2 I_2\Big] \\ &\le 2\epsilon^2 E\Big[\Big(\sum_{i=1}^{N_{\hat\gamma}+M_\epsilon}\xi_i\Big)^2 I_2\Big] + 2E\Big[\Big(\sum_{i=N_{\hat\gamma}+M_\epsilon+1}^{\infty}\theta_i\Big)^2 I_2\Big] = R_{21} + R_{22}\,. \end{aligned} \tag{32}$$

First we evaluate term R_{21}. Using the Hölder inequality, we have that

$$E\Big\{\Big(\sum_{i=1}^{N_{\hat\gamma}+M_\epsilon}\xi_i\Big)^2 I_2\Big\} \le \Big[E\Big\{\Big(\sum_{i=1}^{N_{\hat\gamma}+M_\epsilon}\xi_i\Big)^4 I_2\Big\}\Big]^{1/2}\Big[P\{I_2\}\Big]^{1/2}.$$

Recall that $N_{\hat\gamma} + M_\epsilon \le N_{\gamma_0}$ on I_2. Apply the maximal inequality for submartingales to get

$$E\Big[\Big(\sum_{i=1}^{N_{\hat\gamma}+M_\epsilon}\xi_i\Big)^4 I_2\Big] \le E\Big[\max_{1\le l\le N_{\gamma_0}}\Big(\sum_{i=1}^{l}\xi_i\Big)^4\Big] \le E\Big[\Big(\sum_{i=1}^{N_{\gamma_0}}\xi_i\Big)^4\Big] \le C N_{\gamma_0}^2$$

for some absolute constant C. Further, since the sequence M_ϵ converges to infinity faster than $\log\log\epsilon^{-2}$, we can apply Lemma 4 to the probability $P\{I_2\}$ with $k=0$ and $m=2$:

$$P\{I_2\} \le P\{N_{\hat\gamma} < N_{\gamma_0} - M_\epsilon\} \le c\big(\log\epsilon^{-2}\big)^{-2}\,.$$

for sufficiently small ϵ. Combining the last three relations, we conclude that

$$R_{21} = O\big(\epsilon^2\big) = o\big(\epsilon^2 \log \epsilon^{-2}\big)\,. \tag{33}$$

Somewhat more subtle arguments are needed to handle term R_{22}. By using the Hölder inequality and Lemma 4, we have that for any fixed m

$$\begin{aligned}
R_{22}/2 &= E\Big[\sum_{k=0}^{N_{\gamma_0}-M_\epsilon}\Big(\sum_{i=N_{\hat\gamma}+M_\epsilon}^{\infty}\theta_i\Big)^2 I\{N_{\hat\gamma} = N_{\gamma_0}-M_\epsilon-k\Big] \\
&= \sum_{k=0}^{N_{\gamma_0}-M_\epsilon} E\Big[\Big(\sum_{i=N_{\gamma_0}-k}^{\infty}\theta_i\Big)^2 I\{N_{\hat\gamma} = N_{\gamma_0}-M_\epsilon-k\}\Big] \\
&\le \sum_{k=0}^{N_{\gamma_0}-M_\epsilon}\Big(E\Big[\sum_{i=N_{\gamma_0}-k}^{\infty}\theta_i\Big]^4\Big)^{1/2}\Big(P\big\{N_{\hat\gamma}\le N_{\gamma_0}-M_\epsilon-k\big\}\Big)^{1/2} \\
&\le C\sum_{k=0}^{N_{\gamma_0}-M_\epsilon} e^{-\gamma_0(N_{\gamma_0}-k)}\big(\log\epsilon^{-2}\big)^{-m/2}e^{-\gamma_0 mk/4}
\end{aligned}$$

for sufficiently small ϵ. Taking $m = 5$ in the last inequality, we derive

$$R_{22} = o\big(\epsilon^2\log\epsilon^{-2}\big)\,.$$

So, (32), the bound for R_{21} (33) and the last bound for R_{22} ensure that R_2 is of a smaller order compared to R_1:

$$R_2 = R_{21} + R_{22} = o\big(\epsilon^2\log\epsilon^{-2}\big)\,. \tag{34}$$

It remains to show that $R_3 = o\big(\epsilon^2\log\epsilon^{-2}\big)$. Denote

$$I_3 = \{N_{\gamma_0} + M_\epsilon < N_{\hat\gamma} \le N_{\alpha_\epsilon}\}\,.$$

Next, similarly to the term R_2, we derive the following estimate:

$$R_3 \le 2\epsilon^2 E\Big[\Big(\sum_{i=1}^{N_{\hat\gamma}+M_\epsilon}\xi_i\Big)^2 I_3\Big] + 2E\Big[\Big(\sum_{i=N_{\hat\gamma}+M_\epsilon+1}^{\infty}\theta_i\Big)^2 I_3\Big] = R_{31} + R_{32}\,. \tag{35}$$

We handle the term R_{31} exactly in the same way as R_{21} with the difference that

$$N_{\hat\gamma} + M_\epsilon \le N_{\alpha_\epsilon} + M_\epsilon$$

on I_3 and we apply Corollary 1 to the probability $P(I_3)$. So, for any fixed $m > 0$

$$R_{31} \le c\epsilon^2(N_{\alpha_\epsilon}+M_\epsilon)\big(\log\epsilon^{-2}\big)^{p/2}\epsilon^{\rho/2} = c\epsilon^{2+\rho/2}\big(\big(\log\epsilon^{-2}\big)^{1+p/2}\alpha_\epsilon^{-1} + M_\epsilon\big)$$

for sufficiently small ϵ. Since α_ϵ converges to 0 and M_ϵ to infinity not faster than $\big(\log\epsilon^{-2}\big)^{-p}$ and $\log\epsilon^{-2}$ respectively, we obtain that

$$R_{31} = o\big(\epsilon^2 \log\epsilon^{-2}\big)\,. \tag{36}$$

The last term R_{32} is treated again in much the same way as R_{22}. By Corollary 1,

$$\begin{aligned} R_{32}/2 &= E\Big[\sum_{k=0}^{N_{\alpha_\epsilon}} \Big(\sum_{i=N_{\hat\gamma}+M_\epsilon}^{\infty} \theta_i\Big)^2 I\{N_{\hat\gamma} = N_{\gamma_0} + M_\epsilon + k\}\Big] \\ &= \sum_{k=0}^{N_{\alpha_\epsilon}} E\Big[\Big(\sum_{i=N_{\gamma_0}+2M_\epsilon+k}^{\infty} \theta_i\Big)^2 I\{N_{\hat\gamma} = N_{\gamma_0} + M_\epsilon + k\}\Big] \\ &\le \sum_{k=0}^{N_{\alpha_\epsilon}} \Big(E\Big[\sum_{i=N_{\gamma_0}+2M_\epsilon+k}^{\infty} \theta_i\Big]^4\Big)^{1/2} \Big(P\big\{N_{\hat\gamma} \ge N_{\gamma_0} + M_\epsilon + k\big\}\Big)^{1/2} \\ &\le C\sum_{k=0}^{N_{\alpha_\epsilon}} e^{-\gamma_0(N_{\gamma_0}+k)} \big(\log\epsilon^{-2}\big)^{p/2} \epsilon^{\rho/2} = o\big(\epsilon^2\log\epsilon^{-2}\big)\,. \end{aligned} \tag{37}$$

for sufficiently small ϵ. Thus, by (35),(36) and (37), we have that

$$R_3 = o\big(\epsilon^2 \log\epsilon^{-2}\big)\,.$$

Finally, combining (27), (31), (34) and the last relation proves the theorem.

Remark 10. It becomes clear from the proof of the theorem where the conditions on the sequence M_ϵ come from. On the one hand, M_ϵ should converge to infinity not slower than the sequence $\big(\log\log\epsilon^{-2}\big)^{-1}$ to make the application of Lemma 4 possible so that our estimator $\hat\gamma$ could do the job. On the other hand, M_ϵ should not converge to infinity too fast in order not to make the asymptotic risk exceed the asymptotic Bayes risk. In fact, a properly chosen M_ϵ does not disturb the first order asymptotic behavior, but it certainly effects the second order. In this context one may want to take the smallest M_ϵ among those that do not effect the first order risk asymptotics. From the proof of the theorem and the assertion of Lemma 4 we see that $M_\epsilon = D\log\log\epsilon^{-2}$, with sufficiently large D, can also be used in the estimator $\hat\Phi$.

References

[1] L. M. Artiles, *Adaptive Minimax Estimation in Classes of Smooth Functions*, PhD Thesis, Utrecht University, 2001.

[2] E. Belitser and S. Ghosal, *Adaptive Bayesian inference on the mean of an infinite dimensional normal distribution*, Ann. Statist., 31 (2003), to appear.

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

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

[5] S. Y. Efromovich and M. S. Pinsker, *Learning algorithm for nonparametric filtering*, Automat. Remote Control, 11 (1984), 1434–1440.

[6] D. Freedman, *On the Bernstein-von Mises theorem with infinite dimensional parameters*, Ann. Statist., 27 (1999), 1119–1140.

[7] G. K. Golubev and B. Y. Levit, *Asymptotically efficient estimation for analytic distributions*, Math. Meth. Statist., 5 (1996), 357–368.

[8] I. A. Ibragimov and R. Z. Hasminskii, *On nonparametric estimation of the value of a linear functional in Gaussian white noise*, Theory Probab. Appl., 29 (1984), 18–32.

[9] J. Klemelä and M. Nussbaum, *Constructive asymptotic equivalence of density estimation and Gaussian white noise*, Technical Report 53, Humbolt University, Berlin, 1998.

[10] O. V. Lepski and B. Y. Levit, *Adaptive minimax estimation of infinitely differentiable functions*, Math. Meth. Statist., 7 (1998), 123–156.

[11] O. V. Lepski, *On problems of adaptive estimation in white gaussian noise*, Advances in Soviet Math., 12 (1992), 87–106.

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

[13] M. S. Pinsker, *Optimal filtration of square-integrable signals in Gaussian noise*, Problems of Information Transmission, 16 (1980), 120–133.

[14] H. Robbins, *An empirical Bayes approach to statistics*, In: Proc. 3rd Berkeley Symp. on Math. Statist. and Prob., 1 (1955), Berkeley, Univ. of California Press, 157–164.

[15] D. Williams, *Probability with Martingales*, Cambridge University Press, Cambridge. 1991.