Theory of Stochastic Processes
Vol. 17 (33), no. 2, 2011, pp. 16–24

A. BABENKO AND E. BELITSER

ORACLE WIENER FILTERING OF A GAUSSIAN SIGNAL

We study the problem of filtering a Gaussian process whose trajectories, in some sense, have an unknown smoothness β_0 from the white noise of small intensity ϵ. If we knew the parameter β_0, we would use the Wiener filter which has the meaning of oracle. Our goal is now to mimic the oracle, i.e., construct such a filter without the knowledge of the smoothness parameter β_0 that has the same quality (at least with respect to the convergence rate) as the oracle. It is known that in the pointwise minimax estimation, the adaptive minimax rate is worse by a log factor as compared to the nonadaptive one. By constructing a filter which mimics the oracle Wiener filter, we show that there is no loss of quality in terms of rate for the Bayesian counterpart of this problem - adaptive filtering problem.

1. Introduction

Suppose we observe a continuous-in-time signal $f(t) \in L_2[0,1]$ in the white noise model (see Ibragimov and Khasminski (1981), Evromovich (1999), Johnstone (2004), Tsybakov (2008)):

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

where $X_\epsilon(\cdot)$ is an observation process, $W(t)$ is a standard Wiener process, ϵ is the noise intensity. The goal is to recover the signal $f(t_0)$ at a point $t_0 \in [0,1]$, based on the observation $(X^{(\epsilon)}(t),\, t \in [0,1])$, in the asymptotic setup as $\epsilon \to 0$. From now on we will skip the index ϵ in the notations. The estimation quality of an estimator $\hat{f}(t_0)$ is measured by the (quadratic) risk function $R(\hat{f}(t_0), f(t_0)) = E|\hat{f}(t_0) - f(t_0)|^2$, where the estimator $\hat{f}(t_0)$ is a measurable function of observations and the expectation depends on whether the unknown signal $f(t)$ is deterministic or random.

Recall that, given an orthonormal basis $\{\phi_i,\, i \in \mathbb{N}\}$ in $L_2[0,1]$, the model (2) can be translated into an equivalent sequence model

$$X_i = \theta_i + \epsilon\xi_i, \quad i \in \mathbb{N}, \tag{2}$$

with the observations $X_i = \int_0^1 \phi_i(t)dX_\epsilon(t)$, the unknown Fourier coefficients of the signal $\theta_i = \int_0^1 \phi_i(t)f(t)dt$, independent Gaussian noises $\xi_i = \int_0^1 \phi_i(t)dW(t)$. While interesting in communication theory in its own right, in case of deterministic f model (1) also provides a good approximation to a variety of curve estimation problems, for example regression estimation and density estimation problems. For such problems $\epsilon = n^{-1/2}$, where n is the sample size.

The prior knowledge about the signal $f(t)$ can basically be modelled in two ways: either the signal f is assumed to be deterministic and to belong to a given functional class $\mathcal{F}_\beta$, where the parameter β typically has a meaning of the signal smoothness; or the signal $f(t)$ is assumed to be random according to a certain prior distribution π_β, i.e., it is a stochastic process with distribution π_β. Both approaches are actually connected to each other within the decision theoretic framework: according to the von Neumann minimax theorem, the minimax risk over $\mathcal{F}_\beta$ (which is $r(\mathcal{F}_\beta) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}_\beta} R(\hat{f}(t_0), f(t_0))$) in

2000 *Mathematics Subject Classification.* Primary 60G35, 62M20; Secondary 62C10, 93E11.
Key words and phrases. Bayesian oracle, Gaussian process, minimax pointwise risk, Wiener filter.

the first setting is equal to the Bayes risk corresponding to the so-called least favorable prior π_β in the second setting.

This interplay between minimax and Bayesian frameworks has been studied by Li and Zhao (2002) and later Babenko and Belitser (2009) for the case of a Sobolev ellipsoid $\mathcal{F}_\beta \subset L_2[0,1]$ and a certain choice of normal prior π_β (a family of priors in Babenko and Belitser (2009)). The signal $f(t)$ from the Bayesian perspective becomes a certain Gaussian process of smoothness (in some sense) β, the estimation problem becomes the problem of filtering of this Gaussian process of smoothness β from white noise and the corresponding Bayes estimator is nothing else but the Wiener filter. It was shown that this Bayes estimator attains this minimax rate over the Sobolev ellipsoid $\mathcal{F}_\beta$, i.e., the prior π_β is least favorable (at least in terms of the convergence rate) for the problem of minimax (over $\mathcal{F}_\beta$) pointwise estimation of the signal value $f(t_0)$. In a way, this means that prior π_β adequately models the deterministic condition $f \in \mathcal{F}_\beta$ and the problem of filtering of this Gaussian process from white noise mimics the problem of pointwise minimax estimation of a regression function f from the Sobolev ellipsoid $\mathcal{F}_\beta$.

The construction of the minimax Bayes estimators in Li and Zhao (2002) and in Babenko and Belitser (2009) is based on the knowledge of the smoothness parameter β. If the parameter β is unknown, then the problem of adaptive estimation of the signal $f(t_0)$ at point t_0 arises. In the minimax setup, Lepski (1990, 1991, 1992) showed that a penalty log factor in the minimax risk is unavoidable for the problem of adaptive pointwise estimation and proposed an adaptive estimator attaining this adaptive (i.e., degraded by the log factor) minimax rate.

In this note we consider a Bayesian version of the adaptive pointwise estimation of the signal $f(t_0)$, which is in fact an adaptive filtering problem. We assume that the signal $f(t)$ is distributed according a prior π_{β_0} with an unknown "true smoothness" β_0. Clearly, the Bayes estimator $\hat{f}_{\beta_0}(t_0) = E(f(t_0)|X)$, which is the Wiener filter, cannot be used because β_0 is unknown. We regard $\hat{f}_{\beta_0}(t_0)$ as a Bayesian oracle (or oracle Wiener filter) and its risk becomes our benchmark, which we call the oracle Bayes risk. The main goal is then to mimic the Bayesian oracle, that is to find such a procedure $\hat{f}(t_0)$ without using the knowledge of β_0 whose Bayes risk with respect to the prior π_{β_0} is within the constant factor of the oracle Bayes risk. The same type of Bayesian adaptation problem but for the global L_2 estimation of the signal f was studied by Belitser and Enikeeva (2008). It was shown that the empirical Bayes approach mimics the Bayesian oracle whose rate coincides in this case with the nonadaptive global minimax rate. In the pointwise minimax estimation, there is a peculiarity as earlier mentioned: the adaptive minimax rate is worse by a log factor as compared to the nonadaptive one. In this paper we show that a better situation occurs for the Bayesian counterpart of this problem: there is no log penalty factor for the Bayesian pointwise adaptation. We construct a marginal likelihood (an empirical Bayes) estimator $\hat{\beta}$ for the smoothness parameter β_0 which leads to the plug-in procedure $\hat{f}_{\hat{\beta}}(t_0)$ for the signal value $f(t_0)$ and show that it mimics the Bayesian oracle without any loss in the convergence rate.

Of course, one could study this problem from the Bayesian perspective from the very beginning, without relating it to the pointwise minimax estimation. Then the considered problem has the following interpretation: we want to filter a Gaussian process whose trajectories have an unknown smoothness β_0 from the white noise. If we knew parameter β_0, we would use the Wiener filter, which has the meaning of oracle. Our goal is now to mimic the oracle, i.e., construct such a filter without the knowledge of smoothness parameter β_0 that has the same quality (at least with respect to the convergence rate) as the oracle.

2. Preliminaries

Recall the sequence observation model (2) and that we are given a basis $\{\phi_i,\, i \in \mathbb{N}\}$ so that our signal $f(t)$ has a Fourier representation:

$$f(t) = \sum_{i=1}^{\infty} \theta_i \phi_i(t), \quad t \in [0,1], \tag{3}$$

with the Fourier coefficients $\theta = (\theta_1, \theta_2, \dots) \in \ell_2$.

Assume from now on that $\sup_{t\in[0,1]} |\phi_i(t)| \le 1$. Next assume that θ is a Gaussian element from the space ℓ_2 with the distribution $\pi_{\beta,\delta}$ for a $\beta > 1/2$ and $\delta < 2\beta - 1$, where the distribution $\pi_{\beta,\delta}$ is determined as follows: $\theta \sim \pi_{\beta,\delta}$ means that

$$\theta_i \overset{ind}{\sim} \mathcal{N}(0, \tau_i^2(\beta,\delta)), \quad \tau_i^2(\beta,\delta) = \tau_i^2(\beta,\delta,\epsilon) = i^{-2\beta+\delta}\epsilon^{\delta/\beta}, \quad i \in \mathbb{N}. \tag{4}$$

Remark 2.1. All the results of this paper can be easily generalized to the case

$$\tau_i^2(\beta,\delta,\epsilon) = A(i,\epsilon) i^{-2\beta+\delta}\epsilon^{\delta/\beta}, \quad i \in \mathbb{N},$$

where $0 < A_1 \le A(i,\epsilon) \le A_2 < \infty$ uniformly in $i \in \mathbb{N}$ and $\epsilon > 0$.

Remark 2.2. Suppose that $\delta = 0$ (i.e., $\tau_i(\beta, 0) = i^{-2\beta}$) and $\{\phi_i,\, i \in \mathbb{N}\}$ is the trigonometric basis. The parameter β has a meaning of smoothness since the fractional derivative (in Weyl sense) of process (3) of order $\beta - 1/2 - \gamma$ for any $\gamma > 0$ is a well defined Gaussian process.

For a fixed $\beta > 1/2$, we thus defined a family of distributions $\{\pi_{\beta,\delta},\, \delta < 2\beta - 1\}$. Clearly, the signal $f(t)$ is a well defined centered Gaussian process with the covariance function $\mathrm{Cov}(t,s) = \sum_{i=1}^{\infty} \tau_i^2(\beta,\delta)\phi_i(t)\phi_i(s)$. If the basis functions $\phi_i(t)$'s are continuous on $[0,1]$ and $L_2[0,1]$-orthogonal, then (3) is a Karhunen-Loève decomposition of the Gaussian process $f(t)$. If $\{\phi_i,\, i \in \mathbb{N}\}$ is the usual trigonometric basis of $L_2[0,1]$, the resulting Gaussian process $f(t)$ is also stationary. Notice that the only prior which does not depend on the parameter ϵ is $\pi_{\beta,0}$, the one considered in Li and Zhao (2002).

To measure the quality of an estimator $\hat{f}(t_0)$, we use the risk function

$$R_\pi(\hat{f}(t_0)) = R(\hat{f}(t_0), f(t_0)) = E\big(\hat{f}(t_0) - f(t_0)\big)^2 = E_\pi E_\theta\big(\hat{f}(t_0) - f(t_0)\big)^2, \tag{5}$$

where by E_θ we denote the expectation with respect to the conditional distribution of X given θ, by E_π the expectation with respect to the distribution $\pi = \pi_{\beta,\delta}$ of θ defined by (4) and by E the expectation with respect to the joint distribution of (X, θ). If $f(t)$ is a Gaussian process described by (3) and (4), the optimal estimator $\tilde{f}(t_0)$ with respect to the risk R_π is the Bayes estimator

$$\tilde{f}(t_0) = \tilde{f}_\beta(t_0) = \tilde{f}_\beta(t_0, \delta, X) = E(f(t_0)|X) = \sum_{i=1}^{\infty} \phi_i(t_0) E(\theta_i|X) = \sum_{i=1}^{\infty} \phi_i(t_0)\tilde{\theta}_i, \tag{6}$$

where

$$\tilde{\theta}_i = \tilde{\theta}_i(\beta,\delta,X) = E(\theta_i|X) = E(\theta_i|X_i) = \frac{\tau_i^2(\beta,\delta)X_i}{\tau_i^2(\beta,\delta) + \epsilon^2}, \quad i \in \mathbb{N}.$$

Let us relate the above problem of filtering the signal $f(t)$ to the problem of minimax estimation of the signal over a Sobolev class of smoothness β. For a $\beta > 1/2$, introduce the Sobolev functional class

$$\mathcal{F}_\beta = \mathcal{F}_\beta(Q) = \Big\{ f(t),\, t \in [0,1] : f(t) = \sum_{i=1}^{\infty} \theta_i \phi_i(t),\, \theta \in \Theta_\beta(Q) \Big\},$$

where

$$\Theta_\beta(Q) = \Big\{ \theta : \sum_{i=1}^{\infty} i^{2\beta}\theta_i^2 \le Q \Big\}$$

is the Sobolev ellipsoid in space ℓ_2. Suppose that the signal f is deterministic and $f \in \mathcal{F}_\beta$. In this case, the risk of an estimator $\hat{f}(t_0) = \sum_{i=1}^\infty \hat{\theta}_i \phi_i(t_0)$ is

$$R_f^*(\hat{f}(t_0)) = R(\hat{f}(t_0), f(t_0)) = E_\theta\big(\hat{f}(t_0) - f(t_0)\big)^2, \quad \theta \in \Theta_\beta(Q). \tag{7}$$

Now we remind a result concerning the properties of the estimator $\tilde{f}_\beta(t_0)$, obtained in the paper of Babenko and Belitser (2009) with $\epsilon = n^{-1/2}$.

Theorem 2.1. *Let $\beta > 1/2$, the risks R_π and R_f^* be defined by (5) and (7) respectively, where a prior $\pi = \pi_{\beta,\delta}$ is from the family defined by (4) with $\delta < \min\{2\beta - 1, \beta + 1/2\}$. Let $\tilde{f}_\beta(t_0)$ be the corresponding Bayes estimator defined by (6) and $\sup_{t\in[0,1]} |\phi_i(t)| \le 1$, $i \in \mathbb{N}$. Then*

$$\sup_{f\in\mathcal{F}_\beta(Q)} R_f^*(\tilde{f}_\beta(t_0)) \le C_1 \epsilon^{(2\beta-1)/\beta} \quad \text{and} \quad R_\pi(\tilde{f}_\beta(t_0)) \le C_2 \epsilon^{(2\beta-1)/\beta},$$

where constant C_1 depends only on β, δ and Q and constant C_2 only on β and δ.

Remark 2.3. The precise description of the values of the constants C_1 and C_2, can be found in Babenko and Belitser (2009): the assertions of the above theorem hold for any

$$C_1 \ge \frac{B\left(\frac{4\beta-2\delta-1}{2\beta-\delta}, \frac{1}{2\beta-\delta}\right) + QB\left(\frac{2\beta-1}{2\beta-\delta}, \frac{2\beta-2\delta+1}{2\beta-\delta}\right)}{2\beta-\delta} \quad \text{and} \quad C_2 \ge \frac{\pi}{(2\beta-\delta)\sin(\pi/(2\beta-\delta))},$$

for all sufficiently small $\epsilon \le \epsilon_0 = \epsilon_0(C_1, C_2)$, where $B(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du$, $\alpha,\beta > 0$, is the beta function.

The above theorem implies that the chosen prior adequately reflects the requirement $f \in \mathcal{F}_\beta(Q)$: both frequentist and Bayes risks of the Bayes estimator $\tilde{f}_\beta(t_0)$ have the same convergence rate uniformly over the Sobolev class $\mathcal{F}_\beta(Q)$. It is known that, under appropriate assumptions on $\phi(t_0) = (\phi_1(t_0), \phi_2(t_0), \ldots)$, this rate is in fact minimax: the minimax risk $r_\epsilon(\mathcal{F}_\beta(Q)) = \inf_{\tilde{f}} \sup_{f\in\mathcal{F}_\beta(Q)} R_f^*(\tilde{f}_\beta(t_0))$ is of order $\epsilon^{(2\beta-1)/\beta}$ and it is sharp (see Donoho and Low (1992)), i.e., there exist $0 < C_1 \le C_2 < \infty$, such that

$$C_1 \le \liminf_{\epsilon\to 0} \epsilon^{-(2\beta-1)/\beta} r_\epsilon(\mathcal{F}_\beta(Q)) \le \limsup_{\epsilon\to 0} \epsilon^{-(2\beta-1)/\beta} r_\epsilon(\mathcal{F}_\beta(Q)) \le C_2.$$

The lower bound in the above relations holds only for the so called "nonparametric" $\phi(t_0)$; a simple example is $|\phi_i(t_0)| \ge \kappa > 0$, $i \in \mathbb{N}$. We should mention here the difference of the pointwise (quadratic) minimax rate with the global L_2-minimax rate (with respect to the risk $E_f\|\hat{f} - f\|^2$, where $\|\cdot\|$ is the usual norm in $L_2[0,1]$) over the Sobolev ellipsoid $\mathcal{F}_\beta$, the latter being $\epsilon^{4\beta/(2\beta+1)}$. However, if the smoothness parameter β is unknown, Lepski (1990, 1991, 1992) (see also Tsybakov (1998)) showed that a penalty log factor is unavoidable: the so called adaptive (quadratic) rate of convergence (see the exact definition in Tsybakov (1998)) is $\big(\epsilon^2 \log(1/\epsilon)\big)^{(2\beta-1)/2\beta}$.

This lower bound holds also for the minimax risk over the periodic Sobolev functional class of smoothness β (which is a subset of the usual Sobolev functional class, so the same lower bound holds for the Sobolev class); see Evromovich (1999) and Tsybakov (2009). In this case, ellipsoid $\widetilde{\Theta}_\beta(Q) = \big\{\theta : \sum_{k=1}^\infty k^{2\beta}(\theta_{2k}^2 + \theta_{2k+1}^2) \le Q\big\}$ is used instead of $\Theta_\beta(Q)$ and the sequence $\phi(t_0) = (\phi_1(t_0), \phi_2(t_0), \ldots)$ is such that $\phi_{2k}^2(t_0) + \phi_{2k+1}^2(t_0) = 2$, $k \in \mathbb{N}$. If $\{\phi_k,\, k \in \mathbb{N}\}$ is the standard trigonometric basis: $\phi_1(t) = 1$, $\phi_{2k}(t) = \sqrt{2}\cos(2\pi k t)$, $\phi_{2k+1}(t) = \sqrt{2}\sin(2\pi k t)$, $k \in \mathbb{N}$, the above relations on $\phi(t_0) = (\phi_1(t_0), \phi_2(t_0), \ldots)$, $k \in \mathbb{N}$, are indeed fulfilled.

3. Adaptive filtering by empirical Bayes

In what follows, we consider a Bayesian version of the adaptive estimation of the functional $\Phi(\theta)$, which is in fact an adaptive filtering problem. Namely, we assume that $\theta \sim \pi_\beta$, i.e., θ is a random element distributed according a prior π_β with unknown "true smoothness" β.

We restrict ourselves to the simplest (and the most natural from the Bayesian point of view) prior $\pi_\beta = \pi_{\beta,0}$ from the family of priors $\{\pi_{\beta,\delta},\, \delta < 2\beta - 1\}$, i.e., we take $\delta = 0$ in the expression of the variances $\tau_i^2(\beta, \delta)$. Slightly abusing the notations, we denote

$$\tau_i^2(\beta) = \tau_i^2(\beta, 0) = i^{-2\beta}, \quad i \in \mathbb{N},$$

in what follows. It is possible in principle to consider the general case $\delta < 2\beta - 1$ as well, but the mathematical treatment becomes more involved.

From now on we denote by $\beta_0 > 1/2$ the true value of the unknown parameter β. Since β_0 is unknown, the Bayes estimator $\tilde{f}_\beta(t_0)$ given by (6) cannot be used and it plays now a role of the *Bayesian oracle*. We call its risk $R_\pi(\tilde{f}_\beta(t_0))$ the *oracle Bayes risk*. Instead, we are going to use a plug-in estimator $\tilde{f}_{\hat{\beta}}(t_0)$ for the signal value $f(t_0)$, where $\hat{\beta}$ is an estimator of the smoothness parameter β_0.

Recall that from the Bayesian perspective,

$$X_i \overset{ind}{\sim} \mathcal{N}(0, \epsilon^2 + \tau_i^2(\beta_0)), \quad i \in \mathbb{N}, \tag{8}$$

where β_0 denotes the true value of the unknown smoothness parameter. Let $L_\epsilon(\beta) = L_\epsilon(\beta, X)$ be the marginal likelihood of the data $X = (X_i)_{i\in\mathbb{N}}$:

$$L_\epsilon(\beta) = \prod_{i=1}^{\infty} \frac{1}{\sqrt{2\pi\big(\tau_i^2(\beta) + \epsilon^2\big)}} \exp\Big\{ - \frac{X_i^2}{2\big(\tau_i^2(\beta) + n^{-1}\big)}\Big\}.$$

Maximizing the function $L_\epsilon(\beta)$ is equivalent to minimizing $Z_\epsilon(\beta) = -2\log L_\epsilon(\beta)$. To avoid complications in defining the minimum of $Z_\epsilon(\beta)$ under the events $\{Z_\epsilon(\beta) = \pm\infty\}$, for some fixed reference value $\bar{\beta} > 0$ it is convenient to introduce $\bar{Z}_\epsilon(\beta) = Z_\epsilon(\beta, \bar{\beta}) = -2\log\frac{L_\epsilon(\beta)}{L_\epsilon(\bar{\beta})}$, which is finite almost surely. For any set $S_\epsilon \subseteq (0, +\infty)$, define the marginal likelihood estimator of β restricted to the set S_ϵ:

$$\hat{\beta} = \hat{\beta}(S_\epsilon) = \hat{\beta}(S_\epsilon, X, \epsilon) = \arg\min_{\beta\in S_\epsilon} \bar{Z}_\epsilon(\beta). \tag{9}$$

This means that $Z_\epsilon(\hat{\beta}(S_\epsilon)) \leq Z_\epsilon(\beta')$ for all $\beta' \in S_\epsilon$, or equivalently $Z_\epsilon(\hat{\beta}(S_\epsilon), \beta') \leq 0$ for all $\beta' \in S_\epsilon$. We will use a finite set S_ϵ so that the above estimator is well defined.

Remark 3.1. Certainly, this is not the only possible way to estimate the smoothness parameter. The estimator is easy to implement in practice and it has an appealing feature that it is based on the fundamental principle in statistics: maximization of the likelihood. However, the analytic treatment of this approach is somewhat involved, even in our case for conjugate pair of normal model and normal prior. One can in principle try to find another estimator for β (for example, by using the method of moments), which should be good enough to plug in $\tilde{f}_\beta(t_0)$ and is easier to treat.

Remark 3.2. The problem of estimating the smoothness parameter β_0 is an auxiliary problem, but it is of interest on its own right. This a peculiar problem of parametric estimation with infinitely many non-identically distributed normal observations and a peculiar asymptotics: in our case $\epsilon \to 0$ some information parameter involved in the variances of the observations, and not the traditional size of the observation sample. In fact, we use a version of the maximum likelihood method. Similar approach was previously considered by Belitser and Enikeeva (2008) for another estimation problem, the signal estimation in the ℓ_2-norm.

Remark 3.3. The accompanying problem of estimating the smoothness parameter β_0 by the empirical Bayes procedure $\hat{\beta}$ can be seen as a Bayesian counterpart of the inference problem on the smoothness parameter β_0. In the minimax frequentist setting, it is impossible to estimate the smoothness in any meaningful sense, while it is a well defined problem from the Bayesian point of view as the problem of estimating a parameter β of the prior distribution π_{β_0}.

From now on we define the set S_ϵ to be as follows:

$$S_\epsilon = \{1/2 + \kappa_\epsilon + k\delta_\epsilon,\ k = 0, 1, \dots, M_\epsilon - 1\}, \tag{10}$$

where the positive sequences $\kappa_\epsilon \to 0$, $\delta_\epsilon \to 0$ and $M_\epsilon \in \mathbb{N}$, $M_\epsilon \to \infty$ so that $M_\epsilon\delta_\epsilon \to \infty$, as $\epsilon \to 0$.

The next lemma describes the quality of the smoothness estimator $\hat{\beta}$.

Lemma 3.1. *Let $\hat{\beta}$ be defined by (9), S_ϵ by (10) with $\delta_\epsilon = o\big(1/\log(1/\epsilon)\big)$ as $\epsilon \to 0$. Then there exist $c_1 = c_1(\beta_0)$ and $\epsilon_0 = \epsilon_0(\beta_0)$ such that for any $\beta \in S_\epsilon$ such that $|\beta - \beta_0| \geq 2\delta_\epsilon$ the inequality*

$$P\{\hat{\beta} = \beta\} \leq \exp\big\{ - c_1\delta_\epsilon^2\big(\log 1/\epsilon)\big)^2\epsilon^{-1/\beta_0}\big\}$$

holds for all $\epsilon \leq \epsilon_0$.

The proof is rather tedious and lengthy, but we omit it since it essentially follows the same line as the proof of Lemma 2 in Belitser and Enikeeva (2008), with the substitutions ϵ^2 and $\beta_0 - 1/2$ instead of quantities n^{-1} and β_0 respectively in Lemma 2 of Belitser and Enikeeva (2008).

Next, denote $\hat{f}(t_0, \beta) = \tilde{f}_\beta(t_0)$ and introduce the empirical Bayesian plug-in estimator for the signal value $f(t_0)$:

$$\hat{f} = \hat{f}(t_0, \hat{\beta}) = \tilde{f}_{\hat{\beta}}(t_0) = \sum_{i=1}^{\infty} \frac{\phi_i(t_0)\tau_i^2(\hat{\beta})X_i}{\tau_i^2(\hat{\beta}) + \epsilon^2}, \tag{11}$$

with $\tilde{f}_\beta(t_0)$ defined by (6), $\hat{\beta} = \hat{\beta}(S_\epsilon)$ defined by (9) and S_ϵ defined by (10).

To avoid uninteresting cases (when the signal value $f(t_0)$ is close to zero), we assume that

$$R_\pi(\tilde{f}_{\beta_0}(t_0)) \geq c\epsilon^\alpha$$

for some $c, \alpha > 0$. This requirement is not restrictive since it will be fulfilled if there exists an $i \in \{1, 2, \dots, \lfloor\epsilon^{-2}\rfloor\}$ such that $\phi_i^2(t_0) \geq 2c > 0$. Indeed, as $\beta_0 > 1/2$, then

$$R_\pi(\tilde{f}_{\beta_0}(t_0)) = \sum_{i=1}^{\infty} \frac{\phi_i^2(t_0)}{i^{2\beta_0} + \epsilon^{-2}} \geq c\epsilon^{4\beta_0}.$$

The next theorem claims that, under the above condition and very mild conditions on the choice of the set S_ϵ, the adaptive empirical Bayes estimator $\hat{f}$ mimics the Bayesian oracle, i.e., its Bayes risk is asymptotically not worse than the Bayesian oracle risk.

Theorem 3.1. *Suppose $R_\pi(\tilde{f}_{\beta_0}(t_0)) \geq c\epsilon^\alpha$ for some $c, \alpha > 0$. Let $\hat{f}(t_0, \hat{\beta})$ be defined by (11), constant $c_1 = c_1(\beta_0)$ be from Lemma 3.1, and the sequences κ_ϵ, δ_ϵ and M_ϵ from the definition (10) of the set S_ϵ be such that $\delta_n = o\big(1/\log(1/\epsilon)\big)$ and*

$$\kappa_\epsilon^{-1}M_\epsilon \exp\big\{ - c_1\delta_\epsilon^2\big(\log(1/\epsilon)\big)^2\epsilon^{-1/\beta_0}/2\big\} = o(\epsilon^\alpha)$$

as $\epsilon \to 0$. Then for some $C = C(\beta_0, \alpha)$

$$R_\pi(\hat{f}(t_0, \hat{\beta})) \leq R_\pi(\tilde{f}_{\beta_0}(t_0))(4 + o(1)) \leq C\epsilon^{(2\beta_0 - 1)/\beta_0}(1 + o(1)) \quad \textit{as} \quad \epsilon \to 0.$$

Remark 3.4. Many choices of sequences δ_ϵ, κ_ϵ and M_ϵ satisfying the conditions of the theorem are possible. For example, $\delta_\epsilon = 1/(\log(1/\epsilon))^2$, $\kappa_\epsilon = 1/\log(1/\epsilon)$ and $M_\epsilon = (\log(1/\epsilon))^3$ will do. In fact, there is no need to take a sequence κ_ϵ converging to zero faster than $1/\log(1/\epsilon)$ since already for $\beta_0 = \kappa_\epsilon = 1/\log(1/\epsilon)$ the risk will not converge to zero. Neither does it make sense to take the sequence M_ϵ converging to infinity faster than a sequence for which $\log(1/\epsilon) = o(M_\epsilon \delta_\epsilon)$, since already for $\beta_0 = \log(1/\epsilon)$ the risk will have the unimprovable parametric convergence rate ϵ^2.

Proof. Write

$$R_\pi(\hat f(t_0,\hat\beta)) = E(\hat f(t_0,\hat\beta) - f(t_0))^2 = E\Big[(\hat f(t_0,\hat\beta) - f(t_0))^2 I\{|\hat\beta-\beta_0| \geq 2\delta_\epsilon\}\Big] + E\Big[(\hat f(t_0,\hat\beta) - f(t_0))^2 I\{|\hat\beta-\beta_0| < 2\delta_\epsilon\})\Big] = T_1 + T_2.$$

First notice that

$$\begin{aligned} T_2 &= E\Big[(\hat f(t_0,\hat\beta) - f(t_0))^2 I\{|\hat\beta-\beta_0| < 2\delta_\epsilon\})\Big] \\ &\leq \sum_{\beta\in S_\epsilon:\,|\beta-\beta_0|<2\epsilon_n} E(\hat f(t_0,\beta) - f(t_0))^2 \\ &\leq 4 \max_{\beta:\,|\beta-\beta_0|<2\delta_\epsilon} E(\hat f(t_0,\beta) - f(t_0))^2, \end{aligned} \tag{12}$$

since there are at most 4 β's from the set S_ϵ such that $|\beta-\beta_0| < 2\delta_\epsilon$. Now, recall that $E(X_i-\theta_i)^2 = \epsilon^2$ and $E\theta_i^2 = i^{-2\beta_0}$. Therefore,

$$\begin{aligned} E(\hat f(t_0,\beta) - f(t_0))^2 &= E_\pi E_\theta(\hat f(t_0,\beta) - f(t_0))^2 \\ &= \sum_{i=1}^\infty \frac{\epsilon^2\phi_i^2(t_0)\tau_i^4(\beta)}{(\tau_i^2(\beta)+\epsilon^2)^2} + E_\pi\Big(\sum_{i=1}^\infty \frac{\epsilon^2\phi_i(t_0)\theta_i}{\tau_i^2(\beta)+\epsilon^2}\Big)^2 \\ &= \sum_{i=1}^\infty \frac{\phi_i^2(t_0)\epsilon^{-2}}{(i^{2\beta}+\epsilon^{-2})^2} + \sum_{i=1}^\infty \frac{\phi_i^2(t_0)i^{4\beta-2\beta_0}}{(i^{2\beta}+\epsilon^{-2})^2} \\ &= \sum_{i=1}^\infty \frac{\phi_i^2(t_0)}{i^{2\beta}+\epsilon^{-2}} + \sum_{i=1}^\infty \frac{\phi_i^2(t_0)(i^{4\beta-2\beta_0}-i^{2\beta})}{(i^{2\beta}+\epsilon^{-2})^2}. \end{aligned} \tag{13}$$

Next, let $K_\epsilon = \lfloor \epsilon^{-\alpha/(2\beta_0-1)-\gamma} \rfloor$ for some fixed $\gamma > 0$ (for example, take $\gamma = 1$). Then, as $\epsilon \to 0$,

$$\sum_{i=K_\epsilon+1}^\infty i^{-2\beta_0} \leq \sum_{i=K_\epsilon+1}^\infty i^{-2(\beta_0-2\delta_\epsilon)} = o(\epsilon^\alpha) = o\big(R_\pi(\tilde f_{\beta_0}(t_0))\big)$$

due to the condition $R_\pi(\tilde f_{\beta_0}(t_0)) \geq c\epsilon^\alpha$. Using this relation and the elementary inequality $a - 1 \geq 1 - a^{-1}$ for any $a > 0$, we obtain that, uniformly over $|\beta-\beta_0| < 2\delta_\epsilon$ with $\delta_\epsilon = o(1/\log(1/\epsilon))$,

$$\begin{aligned} \sum_{i=1}^\infty \frac{\phi_i^2(t_0)|i^{4\beta-2\beta_0}-i^{2\beta}|}{(i^{2\beta}+\epsilon^{-2})^2} &= \sum_{i=1}^\infty \frac{\phi_i^2(t_0)i^{4\beta}|i^{-2\beta_0}-i^{-2\beta}|}{(i^{2\beta}+\epsilon^{-2})^2} \\ &\leq \sum_{i=1}^\infty \frac{\phi_i^2(t_0)i^{4\beta_0+8\delta_\epsilon}|i^{-2\beta_0}-i^{-2\beta}|}{(i^{2\beta_0+4\delta_\epsilon}+\epsilon^{-2})^2} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{\infty} \frac{\phi_i^2(t_0) i^{2\beta_0+8\delta_\epsilon} |1 - i^{-2(\beta-\beta_0)}|}{(i^{2\beta_0+4\delta_\epsilon} + \epsilon^{-2})^2} \\
&\le \sum_{i=1}^{K_\epsilon} \frac{\phi_i^2(t_0) i^{4\delta_\epsilon}(i^{4\delta_\epsilon} - 1)}{i^{2\beta_0} + \epsilon^{-2}} + \sum_{i=K_\epsilon+1}^{\infty} \frac{1}{i^{2\beta_0}} + \sum_{i=K_\epsilon+1}^{\infty} \frac{1}{i^{2\beta}} \\
&\le o(1) \sum_{i=1}^{K_\epsilon} \frac{\phi_i^2(t_0)}{i^{2\beta_0} + \epsilon^{-2}} + \sum_{i=K_\epsilon+1}^{\infty} \frac{2}{i^{2(\beta_0 - 2\delta_\epsilon)}} \\
&= o\big(R_\pi(\tilde f_{\beta_0}(t_0))\big) \qquad (14)
\end{aligned}
$$

as $\epsilon \to 0$.

Since $R_\pi(\tilde f_{\beta_0}(t_0)) \ge c\epsilon^\alpha$ and $\delta_\epsilon = o(1/\log(1/\epsilon))$ as $\epsilon \to 0$, it is not difficult to establish, similarly to (14), that

$$
\begin{aligned}
\sum_{i=1}^{\infty} \frac{\phi_i^2(t_0)}{i^{2\beta} + \epsilon^{-2}} - R_\pi(\tilde f_{\beta_0}(t_0)) &= \sum_{i=1}^{\infty} \frac{\phi_i^2(t_0)}{i^{2\beta} + \epsilon^{-2}} - \sum_{i=1}^{\infty} \frac{\phi_i^2(t_0)}{i^{2\beta_0} + \epsilon^{-2}} \\
&= o\big(R_\pi(\tilde f_{\beta_0}(t_0))\big),
\end{aligned}
$$

uniformly over $|\beta - \beta_0| < 2\delta_\epsilon$. By combining (12), (13), (14) and the last relation, we derive that, as $\epsilon \to 0$,

$$
T_2 \le 4 \max_{\beta:\, |\beta-\beta_0| < 2\delta_\epsilon} E\big(\hat f(t_0, \beta) - f(t_0)\big)^2 = R_\pi(\tilde f_{\beta_0}(t_0))(4 + o(1)).
$$

To finish the proof of the theorem, it remains to show that, as $n \to \infty$,

$$
T_1 = o(\epsilon^\alpha) = o\big(R_\pi(\tilde f_{\beta_0}(t_0))\big).
$$

Recall the elementary c_r-inequality $|a+b|^r \le c_r(|a|^r + |b|^r)$ for $r > 0$ and $c_r = 1$ if $r \le 1$ and $c_r = 2^{r-1}$ if $c_r > 1$. Using this and the Cauchy-Schwartz inequality, we obtain that

$$
\begin{aligned}
&\big(\hat f(t_0, \hat\beta) - f(t_0)\big)^4 \\
&\quad = \Big[\sum_{i=1}^{\infty} \phi_i(t_0) \Big(\frac{\tau_i^2(\hat\beta)(X_i - \theta_i)}{\tau_i^2(\hat\beta) + \epsilon^2} - \frac{\epsilon^2 \theta_i}{\tau_i^2(\hat\beta) + \epsilon^2}\Big)\Big]^4 \\
&\quad \le 8\Big[\sum_{i=1}^{\infty} \frac{\phi_i(t_0)\tau_i^2(\hat\beta)\xi_i\epsilon}{\tau_i^2(\hat\beta) + \epsilon^2}\Big]^4 + 8\Big[\sum_{i=1}^{\infty} \frac{\phi_i(t_0)\epsilon^2\theta_i}{\tau_i^2(\hat\beta) + \epsilon^2}\Big]^4 \\
&\quad \le 8 \sum_{\beta_k \in S_\epsilon} \Big[\sum_{i=1}^{\infty} \frac{\phi_i(t_0)\tau_i^2(\beta_k)\xi_i\epsilon}{\tau_i^2(\beta_k) + \epsilon^2}\Big]^4 + 8 \sum_{\beta_k \in S_\epsilon} \Big[\sum_{i=1}^{\infty} \frac{\phi_i(t_0)\epsilon^2\theta_i}{\tau_i^2(\beta_k) + \epsilon^2}\Big]^4.
\end{aligned}
$$

Recall the following fact. Let $Z_1, Z_2, \ldots$ be independent, $Z_i \sim \mathcal{N}(0, \sigma_i^2)$, with $\sum_{i=1}^\infty \sigma_i^2 < \infty$, then

$$
E\Big(\sum_{i=1}^{\infty} Z_i\Big)^4 \le 3\Big(\sum_{i=1}^{\infty} \sigma_i^2\Big)^2.
$$

Apply this relation and again the c_r-inequality (for $r = 1/2$) to get that

$$
\begin{aligned}
&\Big[E\big(\hat f(t_0, \hat\beta) - f(t_0)\big)^4\Big]^{1/2} \\
&\quad \le 2\sqrt{6M_\epsilon} \sum_{i=1}^{\infty} \frac{\phi_i^2(t_0)\epsilon^{-2}}{(i^{1+2\kappa_\epsilon} + \epsilon^{-2})^2} + 2\sqrt{6M_\epsilon} \sum_{i=1}^{\infty} \frac{\phi_i^2(t_0)\epsilon^4 i^{-2\beta_0}}{(\tau_i^2(\beta_{M_\epsilon}) + \epsilon^2)^2} \\
&\quad \le 2\sqrt{6M_\epsilon}\big(1 + (2\kappa_\epsilon)^{-1} + 1 + (2\beta_0 - 1)^{-1}\big) \\
&\quad \le c_2 \kappa_\epsilon^{-1} \sqrt{M_\epsilon}
\end{aligned}
$$

for sufficiently small ϵ. Using the above estimate and the Cauchy-Schwartz inequality,

$$\begin{aligned} T_1 &= E\Big[\big(\hat{f}(t_0,\hat{\beta}) - f(t_0)\big)^2 I\{|\hat{\beta}-\beta_0| \geq 2\delta_\epsilon\}\Big] \\ &\leq \Big[E\big(\hat{f}(t_0,\hat{\beta}) - f(t_0)\big)^4\Big]^{1/2}\Big[P\{|\hat{\beta}-\beta_0| \geq 2\delta_\epsilon\}\Big]^{1/2} \\ &\leq c_2\kappa_\epsilon^{-1}\sqrt{M_\epsilon}\Big[P\{|\hat{\beta}-\beta_0| \geq 2\delta_\epsilon\}\Big]^{1/2}. \end{aligned}$$

Since $\epsilon_n = o(1/\log n)$, by Lemma 3.1, we have that for all $\epsilon \leq \epsilon(\beta_0)$, with $c_1 = c_1(\beta_0)$ and $\epsilon_0(\beta_0)$ from Lemma 3.1,

$$P\{\hat{\beta} = \beta\} \leq \exp\big\{ -c_1\delta_\epsilon^2\big(\log 1/\epsilon)\big)^2\epsilon^{-1/\beta_0}\big\}$$

uniformly over all β such that $|\beta - \beta_0| \geq 2\delta_\epsilon$. Therefore,

$$\begin{aligned} P\big\{|\hat{\beta}-\beta_0| \geq 2\delta_\epsilon\big\} &= \sum_{\beta\in S_\epsilon:\, |\beta-\beta_0|\geq 2\delta_\epsilon} P\big\{\hat{\beta} = \beta\big\} \\ &\leq M_\epsilon \exp\big\{ -c_1\delta_\epsilon^2\big(\log 1/\epsilon)\big)^2\epsilon^{-1/\beta_0}\big\} \end{aligned}$$

for all $\epsilon \leq \epsilon(\beta_0)$. Combining the last relations with the condition of the theorem, we obtain that

$$T_1 \leq c_2\kappa_\epsilon^{-1}M_\epsilon \exp\big\{ -c_1\delta_\epsilon^2\big(\log(1/\epsilon)\big)^2\epsilon^{-1/\beta_0}/2\big\} = o(\epsilon^\alpha) = o\big(R_\pi(\tilde{f}_{\beta_0}(t_0))\big).$$

as $\epsilon \to 0$, which completes the proof. □

References

1. A. Babenko and E. Belitser, *On posterior pointwise convergence rate of a Gaussian signal under a conjugate prior*, Statist. Prob. Lett. **79** (2009), 670–675.
2. E. Belitser and F. Enikeeva, *Adaptive filtering of a random signal in Gaussian white noise*, Problemy Peredachi Informatsii **44** (2008), 39–51; translation in Probl. Inf. Transm. **44** (2008), 321–332.
3. E. Belitser and B. Levit, *On the empirical Bayes approach to adaptive filtering*, Math. Meth. Statist. **12** (2003), 131–154.
4. I. A. Ibragimov and R. Z. Khasminski, *Asymptotic Theory*, Springer, New York, 1981.
5. I. Johnstone, *Function Estimation and Gaussian Sequence Models*, Monograph draft, 2009, http://www-stat.stanford.edu/∼imj/
6. O. V. Lepski, *One problem of adaptive estimation in Gaussian white noise*, Theory Probab. Appl. **35** (1990), 459–470.
7. O. V. Lepski, *Asymptotic minimax adaptive estimation. 1. Upper bounds*, Theory Probab. Appl. **36** (1991), 645–659.
8. O. V. Lepski, *Asymptotic minimax adaptive estimation. 2. Statistical model without optimal adaptation. Adaptive estimators*, Theory Probab. Appl. **37** (1992), 468–481.
9. X. Li and L. H. Zhao, *Bayesian nonparametric point estimation under a conjugate prior*, Statist. Prob. Lett. **58** (2002), 23–30.
10. H. Robbins, *An empirical Bayes approach to statistics*, Proc. 3rd Berkeley Symp. on Math. Statist. and Prob. **1** (1956), Berkeley, Univ. of California Press, 157–164.
11. A. Tsybakov, *Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes*, Ann. Statist. **26** (1998), 2420–2469.
12. A. Tsybakov, *Introduction to nonparametric estimation*, Springer, 2009.

Department of Mathematics, Utrecht University, The Netherlands

Department of Mathematics, Eindhoven University of Technology, The Netherlands
E-mail address: e.n.belitser@tue.nl