

**ON ORACLE INEQUALITIES
RELATED TO A POLYNOMIAL FITTING**

Y. CAO AND Y. GOLUBEV

Université de Provence (Aix-Marseille 1) CMI
39 rue F. Joliot-Curie, 13453 Marseille, France
E-mail: cao@cmi.univ-mrs.fr, Youri.Golubev@cmi.univ-mrs.fr

We study a classical problem of fitting of an unknown regression function by polynomials. At the very core of our approach is the method of unbiased risk estimation which is used for data-driven choice of degree of the fitting polynomial. We derive non-asymptotic upper bounds for the mean square risk of fitting related to this technique.

Key words: polynomial fitting, mean square risk, oracle inequality, unbiased risk estimation.

2000 Mathematics Subject Classification: Primary 62G05, 62G20; secondary 62C20.

1. Introduction

The main goal in this paper is to give an exposition of elementary methods used in the theory of oracle inequalities related to projection estimators. To be more precise, we consider a classical problem of polynomial fitting going back to Gauss and Legendre. The mathematical model of this problem is widely used in practice and admits a simple and transparent statistical interpretation. Suppose we are given n design points X_1, \dots, X_n in \mathbb{R}^1 and the noisy data

$$(1) \quad Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where ε_i are i.i.d. random variables with $\mathbf{E}\varepsilon_i = 0$ and a finite variance $\mathbf{E}\varepsilon_i^2 = \sigma^2$. The regression function $f: \mathbb{R}^1 \rightarrow \mathbb{R}^1$ is unknown and our goal is to find, based on the data at hand $X, Y = (X_1, Y_1), \dots, (X_n, Y_n)$, a polynomial $p_{X,Y}(x)$, which provides a good approximation of $f(x)$. Since the fitting polynomial is a random function, we measure the quality of approximation by the mean square error

$$r^n(f, p_{X,Y}) = \mathbf{E}_f d^n(f, p_{X,Y}),$$

©2005 by Allerton Press, Inc. Authorization to photocopy individual items for internal or personal use, or the internal or personal use of specific clients, is granted by Allerton Press, Inc. for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of \$50.00 per copy is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923.

where \mathbf{E}_f stands for the expectation with respect to the measure generated by Y_i , $i = 1, \dots, n$, and

$$d^n(f, p_{X,Y}) = \frac{1}{n} \sum_{i=1}^n [f(X_i) - p_{X,Y}(X_i)]^2.$$

The traditional least squares technique, which is computationally efficient and relies on numerical linear algebra, provides a natural approach to the minimization of $r^n(f, p_{X,Y})$. From the mathematical viewpoint, this technique can be described as follows. Let \mathcal{P}^q be the set of all polynomials of degree q ,

$$\mathcal{P}^q = \left\{ p(x) : p(x) = \sum_{k=0}^q a_k x^k, \quad a_k \in \mathbb{R}^1 \right\}.$$

Then we find a polynomial providing the best fit to the data Y within \mathcal{P}^q :

$$p_{X,Y}^q(x) = \arg \min_{p \in \mathcal{P}^q} \frac{1}{n} \sum_{i=1}^n [Y_i - p(X_i)]^2.$$

The question, when and why $p_{X,Y}^q(x)$ can be viewed as a minimizer of $r^n(f, p_{X,Y})$, is at the very core of the mathematical foundations of learning theory and we refer the interested reader to Cucker and Smale (2001) for detail. Once the family of fitting polynomials $p_{X,Y}^q(x)$, $q = 1, \dots, n$, has been computed, the next natural step is to determine “the best” polynomial within this family. This should be done automatically by a data-driven method based on the available data. Before discussing traditional approaches to this problem, let us make a standard linear transformation of the data Y letting to simplify numerous technical details.

Denote by $\{\pi_k(x), k = 0, 1, \dots\}$ the system of orthogonal polynomials associated with the design points X_i , $i = 1, \dots, n$:

- $\pi_k(x)$ is a polynomial of degree k ,
-

$$\frac{1}{n} \sum_{i=1}^n \pi_k(X_i) \pi_l(X_i) = \delta_{lk},$$

where $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ otherwise.

For simplicity, it is assumed that the design points are different, so one can construct n orthogonal polynomials. Compute the following statistics:

$$Z_k = \frac{1}{n} \sum_{i=1}^n \pi_k(X_i) Y_i, \quad k = 0, \dots, n-1,$$

and notice that we can represent these new data as follows

$$(2) \quad Z_k = \theta_k + \varepsilon \xi_k, \quad k = 0, \dots, n-1,$$

where

$$\theta_k = \frac{1}{n} \sum_{i=1}^n \pi_k(X_i) f(X_i), \quad \xi_k = \frac{1}{\sigma \sqrt{n}} \sum_{i=1}^n \pi_k(X_i) \varepsilon_i,$$

and $\varepsilon = \sigma/\sqrt{n}$. Notice also that ξ_k are zero mean and uncorrelated with $\mathbf{E}\xi_k\xi_l = \delta_{kl}$. It is easy to see that the models (2) and (1) are equivalent, since

$$Y_i = \sum_{k=0}^{n-1} Z_k \pi_k(X_i), \quad f(X_i) = \sum_{k=0}^{n-1} \theta_k \pi_k(X_i), \quad \varepsilon_i = \sigma \sum_{k=0}^{n-1} \xi_k \pi_k(X_i).$$

Thus the polynomial fitting problem is equivalent to estimating θ_k , $k = 0, \dots, n-1$, based on the noisy data Z_k . Notice also that the risk of the fitting is computed as

$$d^n(f, p_{X,Y}) = \frac{1}{n} \mathbf{E}_\theta \sum_{k=0}^{n-1} [\theta_k - \hat{\theta}_k]^2 = \frac{1}{n} \mathbf{E}_\theta \|\theta - \hat{\theta}\|^2,$$

where

$$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^n p_{X,Y}(X_i) \pi_k(X_i).$$

Therefore from now on we deal with the statistical model (2). Moreover, in order to simplify substantially some technical details, we will assume that the ξ_k are i.i.d. $\mathcal{N}(0, 1)$. Our statistical analysis of this model relies essentially on a special class of estimators defined by

$$\hat{\theta}_k = h_k Z_k, \quad k = 0, \dots, n-1,$$

where the sequence $h_k \in [0, 1]$ may depend on the observations. For brevity, we will denote this estimator by $\hat{\theta} = hZ$. In what follows, the sequence h is called a *filter*. If h is a given filter which does not depend on the observations, then the mean square risk of the estimator $\hat{\theta} = hZ$ is computed very easily,

$$R(\theta, h) \stackrel{\text{def}}{=} \mathbf{E}_\theta \|hZ - \theta\|^2 = \sum_{k=0}^{n-1} (1 - h_k)^2 \theta_k^2 + \varepsilon^2 \sum_{k=0}^{n-1} h_k^2.$$

In the present paper, we concentrate on the so-called projection estimators with $h_k = \mathbf{1}(k \leq q)$. Motivation of this class is related to the fact that the least squares polynomial fitting $p_{X,Y}^q(x)$ is equivalent to the estimator $\hat{\theta}_k = \mathbf{1}(k \leq q) Z_k$. For the parameter q we use the term *bandwidth*. Denote for brevity by \mathcal{H}^m the set of all projection filters with the bandwidth less than m , i.e.,

$$\mathcal{H}^m = \left\{ h_k : h_k = \mathbf{1}(k \leq q), \quad 0 < q \leq m-1 \right\}.$$

Notice that the problem of the data-driven choice of the degree of a fitting polynomial can be reformulated as the data-driven choice of a filter within \mathcal{H}^m . A traditional approach to this problem is based on the principle of unbiased risk estimation (see Stein (1981)), which has a simple heuristic motivation. If θ_k^2 were known, the best h^* would be computed as $h^* = \arg \min_{h \in \mathcal{H}^m} R(\theta, h)$. Therefore the principal idea of the method is to minimize the unbiased estimate of $R(\theta, h)$. This approach, going back to Akaike (1973), Mallows (1973), and Shibata (1981),

can be easily implemented, since θ_k^2 may be estimated by $Z_k^2 - \varepsilon^2$. Thus, we arrive at the famous Akaike's criterion

$$(3) \quad \bar{h}^A = \arg \min_{h \in \mathcal{H}^n} \left\{ \sum_{k=0}^{n-1} [1 - h_k]^2 Z_k^2 + 2\varepsilon^2 \sum_{k=0}^{n-1} h_k \right\}.$$

The main goal in this paper is to provide a panorama of oracle inequalities related to Akaike's method and its natural extensions. We begin with non-asymptotic concentration inequalities for $\|\bar{h}^A Z - \theta\|^2$ expressed in terms of the oracle risk

$$(4) \quad R(\theta) = R(\theta, h^*) = \inf_{h \in \mathcal{H}^n} R(\theta, h).$$

Some of these inequalities (Theorem 1) are not surprisingly new (see, for instance, Kneip (1994)). However, we prove them by elementary technique based on the Doob inequality. This technique, improving the second order terms in the oracle inequalities, was originally proposed by Golubev (2004) and Golubev and Levit (2004).

The second class of problem addressed here is related to the fact that the noise variance is hardly known in practice. In order to make (3) feasible from the practical viewpoint, one can estimate the unknown noise variance by

$$\hat{\varepsilon}^2(h) = \frac{1}{n} \sum_{k=0}^{n-1} [1 - h_k]^2 Z_k^2$$

and plug-in this estimator in (3). Thus we get the following adaptive filter:

$$\tilde{h} = \arg \min_{h \in \mathcal{H}^n} \left\{ \sum_{k=0}^{n-1} (1 - h_k)^2 Z_k^2 \left(1 + \frac{2}{n} \sum_{s=0}^{n-1} h_s \right) \right\}.$$

In the present paper, we will study two counterparts of \tilde{h} :

$$(5) \quad \bar{h}^B = \arg \min_{h \in \mathcal{H}^{n/4}} \left\{ \sum_{k=0}^{n-1} (1 - h_k)^2 Z_k^2 \left(1 - \frac{2}{n} \sum_{s=0}^{n-1} h_s \right)^{-1} \right\}$$

and

$$(6) \quad \bar{h}^C = \arg \min_{h \in \mathcal{H}^{n/4}} \left\{ \sum_{k=0}^{n-1} (1 - h_k)^2 Z_k^2 \left(1 - \frac{1}{n} \sum_{s=0}^{n-1} h_s \right)^{-2} \right\}.$$

We would like to draw attention to the fact that here, by some technical reasons, we use the class $\mathcal{H}^{n/4}$, which is evidently embedded in \mathcal{H}^n , and we will control the risks of $\bar{h}^B Z$ and $\bar{h}^C Z$ in terms of

$$(7) \quad R^+(\theta) = \min_{h \in \mathcal{H}^{n/4}} R(\theta, h).$$

From a heuristic viewpoint the methods \bar{h}^B , \bar{h}^C can be viewed as approximations of \bar{h} provided that $\sum_{k=0}^{n-1} h_k \ll n$. Indeed, in this case by the Taylor expansion

$$1 + \frac{2}{n} \sum_{k=0}^{n-1} h_k \approx \left(1 - \frac{2}{n} \sum_{k=0}^{n-1} h_k\right)^{-1} \approx \left(1 - \frac{1}{n} \sum_{k=0}^{n-1} h_k\right)^{-2}$$

and we easily get (5) and (6).

Along with projection filters we will deal with their convex combinations. Formally, this class of filters is defined by

$$\mathcal{C}_M^m = \left\{ h : h_k = \sum_{s=1}^M \lambda_s \mathbf{1}(k \leq N_s), \lambda_s \geq 0, \sum_{s=1}^M \lambda_s = 1, N_s \in [1, m-1] \right\}.$$

The data-driven choice of a filter within this class relies on the same principle of the unbiased risk estimation, and we will provide an oracle inequality for this approach.

This paper is organized as follows. Oracle inequalities are summarized in Section 2 and their proofs are given in Section 3. In order to illustrate numerically oracle inequalities, we provide in Section 4 some simulation results.

2. Oracle Inequalities

2.1. THE AKAIKE METHOD. We start our series of oracle inequalities with the classical Akaike method defined by (3).

Theorem 1. *Uniformly in $\theta \in \mathbb{R}^n$ and $\gamma \in (0, 1)$*

$$(8) \quad \mathbf{E}_\theta \|\bar{h}^A Z - \theta\|^2 \leq \frac{R(\theta)}{1-\gamma} + \frac{2\varepsilon^2}{1-\gamma} \left[\frac{1}{\gamma} + \frac{1}{U^{-1}(\gamma/2)} \right],$$

where

$$(9) \quad U(x) = -1 - \frac{\log(1-2x)}{2x}, \quad x \in (0, \frac{1}{2}),$$

and the oracle risk $R(\theta)$ is defined by (4).

The statistical meaning of this theorem is very transparent. Let us consider two typical situations:

- parametric estimation $R(\theta) \asymp \varepsilon^2$,
- nonparametric estimation $R(\theta) \gg \varepsilon^2$.

In the first case, taking, for instance, $\gamma = \frac{1}{2}$, we get the following upper bound:

$$\mathbf{E}_\theta \|\bar{h}^A Z - \theta\|^2 \leq CR(\theta),$$

where C is a constant. It means that when f is a polynomial of a given small degree, we cannot mimic well the oracle risk, but the losses are not very crucial. On the other hand, in case of nonparametric estimation, (8) reveals that the Akaike criterion works nice. Indeed, assuming that γ is small, we get from (8) and the Taylor formula

$$\mathbf{E}_\theta \|\bar{h}^A Z - \theta\|^2 \lesssim R(\theta) + \gamma R(\theta) + 6\gamma^{-1}\varepsilon^2,$$

and minimizing the right-hand side of this inequality with respect to γ , we arrive at

$$\mathbf{E}_\theta \|\bar{h}^A Z - \theta\|^2 \lesssim R(\theta) \left[1 + \sqrt{\frac{24\varepsilon^2}{R(\theta)}} \right].$$

This upper bound certifies that the Akaike method mimics well the oracle risk.

While Theorem 1 provides only an upper bound for $\mathbf{E}_\theta \|\bar{h}^A Z - \theta\|^2$, the next theorem controls large deviations of $\|\bar{h}^A Z - \theta\|^2$ from $R(\theta)$.

Theorem 2. *Uniformly in $\theta \in \mathbb{R}^n$ and $\gamma \in (0, 1)$,*

$$(10) \quad \mathbf{P}_\theta \left\{ (1 - \gamma) \|\bar{h}^A Z - \theta\|^2 > (1 + \gamma) R(\theta) + \frac{8\varepsilon^2 x}{U^{-1}(\gamma/2)} \right\} \leq 4e^{-x}.$$

2.2. MODEL SELECTION APPROACH. Theorem 1 can be viewed as a special case of oracle inequalities related to model selection methods. We refer the interested reader to Barron, Birgé and Massart (1999) and Birgé and Massart (2001) for a motivation and mathematical background of this approach. For the polynomial fitting problem, the model selection approach provides us with the following data-driven filter:

$$(11) \quad \bar{h}^K = \arg \min_{h \in \mathcal{H}^n} \left\{ \sum_{k=0}^{n-1} [1 - h_k]^2 Z_k^2 + K\varepsilon^2 \sum_{k=0}^{n-1} h_k \right\},$$

where $K > 1$ is a constant. Let

$$(12) \quad R^K(\theta) = \min_{h \in \mathcal{H}^n} \left\{ \sum_{k=0}^{n-1} [1 - h_k] \theta_k^2 + (K - 1)\varepsilon^2 \sum_{k=0}^{n-1} h_k^2 \right\}.$$

The statistical properties of this approach are described by

Theorem 3. *Uniformly in $\theta \in \mathbb{R}^n$ we have*

(i) *for $K \geq 2$ and for any $\alpha > 0$*

$$\mathbf{E}_\theta \|\bar{h}^K Z - \theta\|^2 \leq (1 + 2\alpha) \left[R^K(\theta) + \frac{(1 + \alpha)\varepsilon^2}{\alpha} + \frac{2\varepsilon^2}{U^{-1}(\alpha)} \right];$$

(ii) *for $1 < K < 2$ and for any $\alpha \in (0, K - 1)$*

$$\mathbf{E}_\theta \|\bar{h}^K Z - \theta\|^2 \leq \frac{1 + \alpha}{K - 1 - \alpha} \left[R^K(\theta) + \frac{2(1 + \alpha)\varepsilon^2}{2 - K + 2\alpha} + \frac{K\varepsilon^2}{U^{-1}(\alpha)} \right].$$

It can be shown that these upper bounds cannot be substantially improved uniformly in $\theta \in \mathbb{R}^n$ (see Golubev (2004)).

2.3. CONVEX LINEAR COMBINATIONS OF PROJECTION METHODS. Let

$$(13) \quad \bar{h}^M = \arg \min_{h \in \mathcal{C}_M^n} \left\{ \sum_{k=0}^{n-1} (1 - h_k)^2 Z_k^2 + 2\varepsilon^2 \sum_{k=0}^{n-1} h_k \right\}$$

be the data-driven filter within the family of all linear combinations of $M \geq 2$ projection filters.

Theorem 4. *Uniformly in $\theta \in \mathbb{R}^n$ and $\gamma \in (0, 1)$*

$$\mathbf{E}_\theta \|\bar{h}^M Z - \theta\|^2 \leq \frac{1}{1-\gamma} \left\{ \min_{h \in \mathcal{C}_M^n} R(\theta, h) + 2\varepsilon^2 \left[\frac{M}{\gamma} + \frac{M}{U^{-1}(\gamma/2)} \right] \right\}.$$

Notice that this upper bound is good only when the number of projection estimators involved in the convex combination is small. For the case of large M we refer the reader to Nemirovski (2000).

2.4. ADAPTIVE ESTIMATION WITH UNKNOWN NOISE VARIANCE. In this section, we focus on the adaptive fitting with unknown noise variance. As was mentioned above, we use in this case the data-driven filters defined by (5), (6) and our goal is to control their risks in terms of the oracle risk given by (7).

Theorem 5. *Uniformly in $\theta \in \mathbb{R}^n$ and $\gamma \in (0, 1)$*

$$\mathbf{E}_\theta \|\bar{h}^B Z - \theta\|^2 \leq \frac{1}{1-\gamma} \left\{ R^+(\theta) \left[1 + \frac{C_1 R^+(\theta)}{n\varepsilon^2} \right] + \frac{2\varepsilon^2}{\gamma} + \frac{2\varepsilon^2}{U^{-1}(\gamma/2)} + \varepsilon^2 \right\},$$

where C_1 is a constant.

Theorem 6. *Uniformly in $\theta \in \mathbb{R}^n$ and $\gamma \in (0, 1)$*

$$\mathbf{E}_\theta \|\bar{h}^C Z - \theta\|^2 \leq \frac{1}{1-\gamma} \left\{ R^+(\theta) \left[1 + \frac{C_2 R^+(\theta)}{n\varepsilon^2} \right] + \frac{2\varepsilon^2}{\gamma} + \frac{2\varepsilon^2}{U^{-1}(\gamma/2)} + \varepsilon^2 \right\},$$

where C_2 is a constant.

Typically $R^+(\theta)$ is very close to $R(\theta)$ and the ratio $R^+(\theta)/(n\varepsilon^2)$ is small. Therefore both methods should work like the Akaike criterion in case of known noise variance. Our simulations in Section 4 confirm this fact.

3. Proofs

The cornerstone idea of the proofs presented in this paper is based on two simple probabilistic facts.

Lemma 1. *Let $W(t)$ be a Wiener process. Then for any $\alpha \geq 0$*

$$\mathbf{P} \left\{ \sup_{t \geq 0} \left[W(t) - \frac{\alpha t}{2} \right] > x \right\} \leq \exp(-\alpha x).$$

Proof. This well-known fact follows from the Doob inequality. It suffices to note that the random process $\zeta(t) = \exp[\alpha W(t) - \alpha^2 t/2]$ is a martingale. \square

Lemma 2. *Let ξ_s be i.i.d. $\mathcal{N}(0, 1)$. Then*

$$(14) \quad \mathbf{P} \left\{ \sup_{k \geq 0} \left[\sum_{s=0}^k (\xi_s^2 - 1) - U(\alpha)(k+1) \right] > x \right\} \leq e^{-\alpha x}, \quad \alpha \in (0, 1/2),$$

$$(15) \quad \mathbf{P} \left\{ \sup_{k \geq 0} \left[\sum_{s=0}^k (1 - \xi_s^2) - \alpha(k+1) \right] > x \right\} \leq e^{-\alpha x}, \quad \alpha > 0,$$

where $U(\cdot)$ is defined by (9).

Proof. Let $\mathcal{F}_k = \sigma(\xi_0, \dots, \xi_k)$. For a given $\alpha \in (0, \frac{1}{2})$, denote

$$\eta_k = \exp \left[\alpha \sum_{s=0}^k (\xi_s^2 - 1) - \alpha U(\alpha)(k+1) \right].$$

Since

$$\mathbf{E} \exp \left[\alpha \sum_{s=0}^k (\xi_s^2 - 1) - \alpha U(\alpha)(k+1) \right] = 1,$$

η_k is a martingale, then (14) follows from the Doob inequality.

In order to proof (15), we consider the martingale

$$\eta_k^* = \exp \left[\alpha \sum_{s=0}^k (1 - \xi_s^2) - \alpha U^*(\alpha)(k+1) \right],$$

where

$$U^*(\alpha) = 1 - \frac{1}{2\alpha} \log(1 + 2\alpha).$$

It is easy to see by the Taylor formula that $U^*(\alpha) \leq \alpha$. Therefore (15) is proved by the Doob inequality. \square

In the context of data-driven fitting, Lemmas 1 and 2 can be rephrased as follows.

Lemma 3. *Let ξ_k be i.i.d. $\mathcal{N}(0, 1)$. Then for any data-driven filter $\tilde{h} \in \mathcal{H}^n$*

$$(16) \quad \mathbf{E} \left[\sum_{k=0}^{n-1} (1 - \tilde{h}_k) \theta_k \xi_k - \alpha \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 \theta_k^2 \right]_+^p \leq \frac{\Gamma(p)}{(2\alpha)^p}$$

and

$$(17) \quad \mathbf{E} \left[\sum_{k=0}^{n-1} \tilde{h}_k (\xi_k^2 - 1) - \alpha \sum_{k=0}^{n-1} \tilde{h}_k^2 \right]_+^p \leq \frac{\Gamma(p)}{[U^{-1}(\alpha)]^p},$$

where $\Gamma(\cdot)$ is the Gamma function and $\alpha, p > 0$.

Proof. In order to proof (16), consider the Gaussian process $\xi(h) = \sum_{k=0}^{n-1} (1 - h_k) \theta_k \xi_k$ indexed by $h \in \mathcal{H}^n$. It is easy to see that this process coincides in distribution with $W \left(\sum_{k=0}^{n-1} (1 - h_k)^2 \theta_k^2 \right)$, where $W(\cdot)$ is a standard Wiener process. Therefore by Lemma 1 we immediately get

$$\begin{aligned} & \mathbf{E} \left[\sum_{k=0}^{n-1} (1 - \tilde{h}_k) \theta_k \xi_k - \alpha \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 \theta_k^2 \right]_+^p \\ & \leq \mathbf{E} \max_{h \in \mathcal{H}^n} \left[\sum_{k=0}^{n-1} (1 - h_k) \theta_k \xi_k - \alpha \sum_{k=0}^{n-1} (1 - h_k)^2 \theta_k^2 \right]_+^p \\ & = \mathbf{E} \max_{h \in \mathcal{H}^n} \left[W \left(\sum_{k=0}^{n-1} (1 - h_k)^2 \theta_k^2 \right) - \alpha \sum_{k=0}^{n-1} (1 - h_k)^2 \theta_k^2 \right]_+^p \\ & \leq \mathbf{E} \sup_{t \geq 0} [W(t) - \alpha t]_+^p \leq p \int_0^\infty x^{p-1} \exp(-2\alpha x) dx. \end{aligned}$$

The proof of (17) follows from similar arguments and Lemma 2. \square

3.1. PROOF OF THEOREMS 1 AND 2. It is based on simple algebra formulas. Let \tilde{h} be an arbitrary data-driven filter. Then we have

$$\begin{aligned}
(18) \quad \|\tilde{h}Z - \theta\|^2 &= \sum_{k=0}^{n-1} (\tilde{h}_k Z_k - Z_k + \varepsilon \xi_k)^2 \\
&= \sum_{k=0}^{n-1} (\tilde{h}_k - 1)^2 Z_k^2 + 2\varepsilon \sum_{k=0}^{n-1} (\tilde{h}_k - 1) Z_k \xi_k + \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2 \\
&= \left\{ \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 Z_k^2 + 2\varepsilon^2 \sum_{k=0}^{n-1} \tilde{h}_k \right\} - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2 \\
&\quad + 2\varepsilon \sum_{k=0}^{n-1} (\tilde{h}_k - 1) \theta_k \xi_k + 2\varepsilon^2 \sum_{k=0}^{n-1} \tilde{h}_k (\xi_k^2 - 1).
\end{aligned}$$

On the other hand, for any $\gamma \in (0, 1)$ we can decompose the estimation error as follows:

$$\begin{aligned}
(19) \quad \|\tilde{h}Z - \theta\|^2 &= (1 - \gamma) \|\tilde{h}Z - \theta\|^2 + \gamma \|\tilde{h}Z - \theta\|^2 = (1 - \gamma) \|\tilde{h}Z - \theta\|^2 \\
&\quad + \gamma \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 \theta_k^2 + \gamma \varepsilon^2 \sum_{k=0}^{n-1} \tilde{h}_k^2 \xi_k^2 - 2\gamma \varepsilon \sum_{k=0}^{n-1} \tilde{h}_k (1 - \tilde{h}_k) \theta_k \xi_k \\
&= (1 - \gamma) \|\tilde{h}Z - \theta\|^2 + \gamma \varepsilon^2 \sum_{k=0}^{n-1} \tilde{h}_k^2 + \gamma \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 \theta_k^2 \\
&\quad + \gamma \varepsilon^2 \sum_{k=0}^{n-1} \tilde{h}_k^2 (\xi_k^2 - 1) - 2\gamma \varepsilon \sum_{k=0}^{n-1} \tilde{h}_k (1 - \tilde{h}_k) \theta_k \xi_k.
\end{aligned}$$

Recalling the definition of \bar{h}^A (see (3)), we have that for any given $h \in \mathcal{H}^n$

$$\begin{aligned}
&\left\{ \sum_{k=0}^{n-1} (1 - \bar{h}_k^A)^2 Z_k^2 + 2\varepsilon^2 \sum_{k=0}^{n-1} \bar{h}_k^A \right\} \leq \left\{ \sum_{k=0}^{n-1} (1 - h_k)^2 Z_k^2 + 2\varepsilon^2 \sum_{k=0}^{n-1} h_k \right\} \\
&\leq R(\theta, h) + \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2 + 2\varepsilon \sum_{k=0}^{n-1} (1 - h_k)^2 \theta_k \xi_k + \varepsilon^2 \sum_{k=0}^{n-1} (h_k^2 - 2h_k) (\xi_k^2 - 1).
\end{aligned}$$

Therefore, combining this inequality with (18) and (19), we obtain

$$\begin{aligned}
(20) \quad (1 - \gamma) \|\bar{h}^A Z - \theta\|^2 &\leq R(\theta, h) + 2\varepsilon \sum_{k=0}^{n-1} (\bar{h}_k^A - 1) \theta_k \xi_k - \gamma \sum_{k=0}^{n-1} (1 - \bar{h}_k^A)^2 \theta_k^2 \\
&\quad + (2 - \gamma) \varepsilon^2 \sum_{k=0}^{n-1} \bar{h}_k^A (\xi_k^2 - 1) - \gamma \varepsilon^2 \sum_{k=0}^{n-1} \bar{h}_k^A \\
&\quad + 2\varepsilon \sum_{k=0}^{n-1} (1 - h_k) \theta_k \xi_k - \varepsilon^2 \sum_{k=0}^{n-1} h_k (\xi_k^2 - 1).
\end{aligned}$$

Finally using Lemma 3 to control two remainder terms in the above inequality, we get

$$(1 - \gamma)\mathbf{E}_\theta \|\bar{h}^A Z - \theta\|^2 \leq R(\theta) + \frac{2\varepsilon^2}{\gamma} + \frac{(2 - \gamma)\varepsilon^2}{U^{-1}[\gamma/(2 - \gamma)]},$$

thus finishing the proof of Theorem 1. \square

In order to prove (10), we rewrite (20) as follows:

$$(21) \quad (1 - \gamma)\|\bar{h}^A Z - \theta\|^2 \leq (1 + \gamma)R(\theta) + \sum_{s=1}^4 \chi_s,$$

where

$$\begin{aligned} \chi_1 &= 2\varepsilon \sum_{k=0}^{n-1} (\bar{h}_k^A - 1)\theta_k \xi_k - \gamma \sum_{k=0}^{n-1} (1 - \bar{h}_k^A)^2 \theta_k^2, \\ \chi_2 &= 2\varepsilon \sum_{k=0}^n (1 - h_k)\theta_k \xi_k - \gamma \sum_{k=0}^{n-1} (1 - h_k)^2 \theta_k^2, \\ \chi_3 &= (2 - \gamma)\varepsilon^2 \sum_{k=0}^{n-1} \bar{h}_k^A (\xi_k^2 - 1) - \gamma \varepsilon^2 \sum_{k=0}^{n-1} \bar{h}_k^A, \\ \chi_4 &= \varepsilon^2 \sum_{k=0}^{n-1} h_k (1 - \xi_k^2) - \gamma \varepsilon^2 \sum_{k=0}^{n-1} h_k^2. \end{aligned}$$

It is easy to see that $U^{-1}(\gamma/2) \leq \gamma/2$. Then, by Lemma 1 we get for any $y \geq 0$

$$\begin{aligned} \mathbf{P}\left\{\chi_1 \geq \frac{\varepsilon^2 y}{U^{-1}(\gamma/2)}\right\} &\leq \mathbf{P}\left\{\chi_1 \geq \frac{\varepsilon^2 y}{\gamma/2}\right\} \leq e^{-y}, \\ \mathbf{P}\left\{\chi_2 \geq \frac{\varepsilon^2 y}{U^{-1}(\gamma/2)}\right\} &\leq \mathbf{P}\left\{\chi_2 \geq \frac{\varepsilon^2 y}{\gamma/2}\right\} \leq e^{-y}. \end{aligned}$$

On the other hand, Lemma 2 yields

$$\begin{aligned} \mathbf{P}\left\{\chi_3 \geq \frac{2\varepsilon^2 y}{U^{-1}(\gamma/2)}\right\} &\leq e^{-y}, \\ \mathbf{P}\left\{\chi_4 \geq \frac{\varepsilon^2 y}{U^{-1}(\gamma/2)}\right\} &\leq \mathbf{P}\left\{\chi_4 \geq \frac{\varepsilon^2 y}{2U^{-1}(\gamma/2)}\right\} \leq \mathbf{P}\left\{\chi_4 \geq \frac{\varepsilon^2 y}{\gamma}\right\} \leq e^{-y}. \end{aligned}$$

Therefore (10) follows from (21) and the trivial inequality

$$\mathbf{P}\left(\sum_{i=1}^4 \chi_i \geq z\right) \leq \sum_{i=1}^4 \mathbf{P}\left(\chi_i > z/4\right),$$

which holds true for any random variables χ_i . \square

3.2. PROOF OF THEOREM 3. Rewrite (18) in the following form:

$$(22) \quad \mathbf{E}_\theta \|\bar{h}^K Z - \theta\|^2 = \mathbf{E}_\theta \left\{ \sum_{k=0}^{n-1} (1 - \bar{h}_k^K)^2 Z_k^2 + (K-1)\varepsilon^2 \sum_{k=0}^{n-1} \bar{h}_k^K \right\} - n\varepsilon^2 \\ + 2\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^K (\xi_k^2 - 1) - 2\varepsilon \mathbf{E}_\theta \sum_{k=0}^{n-1} (1 - \bar{h}_k^K) \theta_k \xi_k + (2-K)\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^K.$$

According to the definition of \bar{h}^K ,

$$\mathbf{E}_\theta \left\{ \sum_{k=0}^{n-1} (1 - \bar{h}_k^K)^2 Z_k^2 + (K-1)\varepsilon^2 \sum_{k=0}^{n-1} \bar{h}_k^K \right\} \leq R^K(\theta) + n\varepsilon^2.$$

Thus, combining this inequality with (19) and (22), we get

$$(23) \quad (1-\gamma)\mathbf{E}_\theta \|\bar{h}^K Z - \theta\|^2 \leq R^K(\theta) + 2\varepsilon \mathbf{E}_\theta \sum_{k=0}^{n-1} (\bar{h}_k^K - 1) \theta_k \xi_k \\ - \gamma \mathbf{E}_\theta \sum_{k=0}^{n-1} (1 - \bar{h}_k^K)^2 \theta_k^2 + (2-\gamma)\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^K (\xi_k^2 - 1) \\ - (2-\gamma)\alpha\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^K + [(2-\gamma)(1+\alpha) - K]\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^K.$$

If $K \geq 2$, we choose

$$\gamma = \frac{2\alpha}{1+\alpha}.$$

It is clear that in this case $(2-\gamma)(1+\alpha) - K \leq 0$ and we obtain from (23) and Lemma 3 the following upper bound:

$$\frac{1-\alpha}{1+\alpha} \mathbf{E}_\theta \|\hat{\theta}(\bar{h}^K) - \theta\|^2 \leq R^K(\theta) + \frac{(1+\alpha)\varepsilon^2}{\alpha} + \frac{2\varepsilon^2}{(1+\alpha)U^{-1}(\alpha)},$$

thus proving the first part of the theorem. On the other hand, if $K < 2$, we take

$$\gamma = \frac{2+2\alpha-K}{1+\alpha}$$

and by the same arguments we arrive at

$$\frac{K-1-\alpha}{1+\alpha} \mathbf{E}_\theta \|\bar{h}^K Z - \theta\|^2 \leq R^K(\theta) + \frac{2(1+\alpha)\varepsilon^2}{2+2\alpha-K} + \frac{K\varepsilon^2}{(1+\alpha)U^{-1}(\alpha)}. \quad \square$$

3.3. PROOF OF THEOREM 4. We begin with a simple generalization of Lemma 3 for the class of all convex combinations of M projection filters.

Lemma 4. Let ξ_k be i.i.d. $\mathcal{N}(0, 1)$ and $p \geq 1$. Then for any $\alpha, p > 0$ and any data-driven filter $\hat{h} \in \mathcal{C}_M^n$ we have

$$(24) \quad \mathbf{E} \left[\sum_{k=0}^{n-1} (1 - \hat{h}_k) \theta_k \xi_k - \alpha \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \theta_k^2 \right]_+^p \leq \frac{\Gamma(p) M^p}{(2\alpha)^p}$$

and

$$(25) \quad \mathbf{E} \left[\sum_{k=0}^{n-1} \hat{h}_k (\xi_k^2 - 1) - \alpha \sum_{k=0}^{n-1} \hat{h}_k^2 \right]_+^p \leq \frac{\Gamma(p) M^p}{[U^{-1}(\alpha)]^p}.$$

Proof. Since $\hat{h} = \sum_{s=1}^M \lambda_s \tilde{h}_s$, where $\tilde{h}_s \in \mathcal{H}^n$, we have by (16)

$$\begin{aligned} & \mathbf{E} \left[\sum_{k=0}^{n-1} (1 - \hat{h}_k) \theta_k \xi_k - \alpha \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \theta_k^2 \right]_+^p \\ &= \mathbf{E} \left[\sum_{k=0}^{n-1} \sum_{s=1}^M \lambda_s (1 - \tilde{h}_{s_k}) \theta_k \xi_k - \alpha \sum_{k=0}^{n-1} \left(\sum_{s=1}^M \lambda_s (1 - \tilde{h}_{s_k}) \right)^2 \theta_k^2 \right]_+^p \\ &\leq \mathbf{E} \left[\sum_{s=1}^M \lambda_s \left(\sum_{k=0}^{n-1} (1 - \tilde{h}_{s_k}) \theta_k \xi_k - \alpha \lambda_s \sum_{k=0}^{n-1} (1 - \tilde{h}_{s_k})^2 \theta_k^2 \right) \right]_+^p \\ &\leq M^{p-1} \mathbf{E} \sum_{s=1}^M \lambda_s^p \left[\sum_{k=0}^{n-1} (1 - \tilde{h}_{s_k}) \theta_k \xi_k - \alpha \lambda_s \sum_{k=0}^{n-1} (1 - \tilde{h}_{s_k})^2 \theta_k^2 \right]_+^p \leq \frac{\Gamma(p) M^p}{(2\alpha)^p}. \end{aligned}$$

Inequality (25) is proved by the same argument and the inequality

$$U^{-1}(\lambda x) \geq \lambda U^{-1}(x),$$

which holds true for any $\lambda \in (0, 1)$. Its proof follows directly from the Taylor formula

$$U(\lambda x) = \sum_{k=2}^{\infty} \frac{(2\lambda x)^{k-1}}{k} \leq \lambda \sum_{k=2}^{\infty} \frac{(2x)^{k-1}}{k} = \lambda U(x)$$

and from the fact that $U^{-1}(\cdot)$ is a monotone function. \square

Lemma 5. Let ξ_k be i.i.d. $\mathcal{N}(0, 1)$. Then for any $\alpha > 0$ and any data-driven filter $\hat{h} \in \mathcal{C}_M^n$

$$(26) \quad \mathbf{E} \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \theta_k \xi_k - \alpha \mathbf{E} \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \theta_k^2 \leq \frac{1}{2\alpha},$$

$$(27) \quad \mathbf{E} \sum_{k=0}^{n-1} (\hat{h}_k)^2 (1 - \xi_k^2) - \alpha \mathbf{E} \sum_{k=0}^{n-1} (\hat{h}_k)^2 \leq \frac{1}{\alpha}.$$

Proof. For a sequence x_k denote by Δx_k the differences between subsequent elements of x , i.e., $\Delta x_k = x_{k+1} - x_k$. Notice that for $\hat{h} \in \mathcal{C}_M^n$ we have $\hat{h}_0 = 1$ and $\hat{h}_k \geq \hat{h}_{k+1}$. Therefore using the Abel formula (discrete integration by parts), we obtain

$$\begin{aligned}
& \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \theta_k \xi_k - \alpha \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \theta_k^2 \\
&= - \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \Delta \sum_{s=k}^{n-1} \theta_s \xi_s - \alpha \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \theta_k^2 \\
&= \sum_{k=1}^{n-1} \sum_{s=k}^{n-1} \theta_s \xi_s \Delta (1 - \hat{h}_{k-1})^2 - \alpha \sum_{k=0}^{n-1} (1 - \hat{h}_k)^2 \theta_k^2 \\
&= \sum_{k=1}^{n-1} \left[\sum_{s=k}^{n-1} \theta_s \xi_s - \alpha \sum_{s=k}^{n-1} \theta_s^2 \right] \Delta (1 - \hat{h}_{k-1})^2 \\
&\leq \sum_{k=1}^{n-1} \sup_p \left[\sum_{s=p}^{n-1} \theta_s \xi_s - \alpha \sum_{s=p}^{n-1} \theta_s^2 \right] \Delta (1 - \hat{h}_{k-1})^2 = \sup_k \left[\sum_{s=k}^{n-1} \theta_s \xi_s - \alpha \sum_{s=k}^{n-1} \theta_s^2 \right].
\end{aligned}$$

The rest of the proof of (26) follows from Lemma 1 (see also the proof of Lemma 3). Inequality (27) can be proved by the same argument based on the Abel formula and (15). \square

Proof of Theorem 4. It follows the main lines of that of Theorem 1. By definition of \bar{h}^M we have

$$\begin{aligned}
& \mathbf{E}_\theta \left\{ \sum_{k=0}^{n-1} (1 - \bar{h}_k^M)^2 Z_k^2 + 2\varepsilon^2 \sum_{k=0}^{n-1} \bar{h}_k^M \right\} \\
&\leq \min_{h \in \mathcal{C}_M^n} \mathbf{E}_\theta \left\{ \sum_{k=0}^{n-1} (1 - h_k)^2 Z_k^2 + 2\varepsilon^2 \sum_{k=0}^{n-1} h_k \right\} = \inf_{h \in \mathcal{C}_M^n} R(\theta, h) + \varepsilon^2 n.
\end{aligned}$$

Therefore, combining this inequality with (18) and (19), we get

$$\begin{aligned}
& (1 - \gamma) \mathbf{E}_\theta \| \bar{h}^M Z - \theta \|^2 \leq \inf_{h \in \mathcal{C}_M^n} R(\theta, h) \\
&+ 2\varepsilon\gamma \mathbf{E}_\theta \sum_{k=0}^{n-1} (1 - \bar{h}_k^M)^2 \theta_k \xi_k - \gamma^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} (1 - \bar{h}_k^M)^2 \theta_k^2 \\
&+ 2(1 - \gamma)\varepsilon \mathbf{E}_\theta \sum_{k=0}^{n-1} (\bar{h}_k^M - 1) \theta_k \xi_k - \gamma(1 - \gamma) \mathbf{E}_\theta \sum_{k=0}^{n-1} (1 - \bar{h}_k^M)^2 \theta_k^2 \\
&+ 2\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^M (\xi_k^2 - 1) - \gamma(1 - \gamma)\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} [\bar{h}_k^M]^2 \\
&+ \gamma\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} [\bar{h}_k^M]^2 (1 - \xi_k^2) - \gamma^2 \varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} [\bar{h}_k^M]^2.
\end{aligned}$$

Finally, we use Lemmas 4 and 5 to control the remainder terms in the right-hand side of the above inequality. So, using that $M \geq 2$ and $\gamma \in (0, 1)$, we get

$$(1 - \gamma)\mathbf{E}_\theta \|\bar{h}^M Z - \theta\|^2 \leq \inf_{h \in \mathcal{C}_M^n} R(\theta, h) + 2\varepsilon^2 + \frac{2M(1 - \gamma)\varepsilon^2}{\gamma} \\ + \frac{2M\varepsilon^2}{U^{-1}[\gamma(1 - \gamma)/2]} + \gamma^2\varepsilon^2,$$

thus finishing the proof. \square

3.4. PROOF OF THEOREM 5. In the rest of the paper, C will denote a generic constant, which may change even within the same equation. Denote for brevity

$$\|h\|_1 = \sum_{k=0}^{n-1} h_k.$$

The proofs of Theorems 5 and 6 rely on two simple ideas. The first one is related to the fact that a good adaptive filter \hat{h} cannot have a very large bandwidth. It means, for instance, that the norm $\|\hat{h}\|_1$ can be controlled by $\|h^*\|_1$, where $h^* = \arg \min_{h \in \mathcal{H}^{n/4}} R(\theta, h)$. The second idea is that the variance estimate $\|\hat{h}Z - Z\|^2/n$ always provides a good *upper bound* for the unknown noise variance.

Lemma 6. *For some constant C*

$$(28) \quad \mathbf{E}_\theta \|\bar{h}^B\|_1^2 \leq C[R(\theta, h^*)/\varepsilon^2]^2.$$

Proof. It is clear from the definition of \bar{h}^B that

$$\frac{\|\bar{h}^B Z - Z\|^2}{1 - 2\|\bar{h}^B\|_1/n} \leq \frac{\|h^* Z - Z\|^2}{1 - 2\|h^*\|_1/n}.$$

This yields immediately

$$(29) \quad \|\bar{h}^B Z - Z\|^2 + \frac{2}{n}\|\bar{h}^B Z - Z\|^2\|\bar{h}^B\|_1 - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2 \leq \frac{\|h^* Z - Z\|^2}{1 - 2\|h^*\|_1/n} - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2.$$

It is easy to check by a simple algebra that

$$(30) \quad \mathbf{E}_\theta \left[\frac{\|h^* Z - Z\|^2}{1 - 2\|h^*\|_1/n} - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2 \right]^2 \leq \frac{5[R(\theta, h^*)]^2}{(1 - 2\|h^*\|_1/n)^2}.$$

Therefore it remains to bound from below the left-side of (29). We have

$$(31) \quad \|\bar{h}^B Z - Z\|^2 + \frac{2}{n}\|\bar{h}^B Z - Z\|^2\|\bar{h}^B\|_1 - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2$$

$$\begin{aligned}
&\geq -\varepsilon^2 \|\bar{h}^B\|_1 \left[\frac{7}{6} \left(1 + \frac{2\|\bar{h}^B\|_1}{n} \right) - \frac{2}{n} \sum_{k=0}^{n-1} \xi_k^2 \right]_+ \\
&\quad - \left(1 + \frac{2\|\bar{h}^B\|_1}{n} \right) \left[-2\varepsilon \sum_{k=0}^{n-1} (1 - \bar{h}_k^B) \xi_k \theta_k - \sum_{k=0}^{n-1} (1 - \bar{h}_k^B)^2 \theta_k^2 \right]_+ \\
&\quad - \varepsilon^2 \left(1 + \frac{2\|\bar{h}^B\|_1}{n} \right) \left[\sum_{k=0}^{n-1} \bar{h}_k^B (\xi_k^2 - 1) - \frac{1}{6} \sum_{k=0}^{n-1} \bar{h}_k^B \right]_+
\end{aligned}$$

By Lemma 3 we obtain

$$\begin{aligned}
(32) \quad &\mathbf{E}_\theta \left(1 + \frac{2\|\bar{h}^B\|_1}{n} \right)^2 \left[-2\varepsilon \sum_{k=0}^{n-1} (1 - \bar{h}_k^B) \xi_k \theta_k - \sum_{k=0}^{n-1} (1 - \bar{h}_k^B)^2 \theta_k^2 \right]_+^2 \leq C\varepsilon^4, \\
&\mathbf{E}_\theta \left(1 + \frac{2\|\bar{h}^B\|_1}{n} \right)^2 \left[\sum_{k=0}^{n-1} \bar{h}_k^B (\xi_k^2 - 1) - \frac{1}{6} \sum_{k=0}^{n-1} \bar{h}_k^B \right]_+^2 \leq C.
\end{aligned}$$

Denote

$$A = \left\{ \xi_0, \dots, \xi_{n-1} : \frac{2}{n} \sum_{k=0}^{n-1} \xi_k^2 - \frac{7}{4} \geq \frac{1}{8} \right\}.$$

Since $\|\bar{h}^B\|_1 \leq n/4$, we obviously get

$$\mathbf{1}(A) \left[\frac{7}{6} \left(1 + \frac{2\|\bar{h}^B\|_1}{n} \right) - \frac{2}{n} \sum_{k=0}^{n-1} \xi_k^2 \right]_+^2 = 0$$

and

$$\mathbf{1}(A^c) \left[\frac{7}{6} \left(1 + \frac{2\|\bar{h}^B\|_1}{n} \right) - \frac{2}{n} \sum_{k=0}^{n-1} \xi_k^2 \right]_+^2 \leq \mathbf{1}(A^c) \left[\frac{7}{6} \left(1 + \frac{2\|\bar{h}^B\|_1}{n} \right) \right]^2 \leq 4\mathbf{1}(A^c).$$

Therefore with (29)–(32) we obtain

$$\mathbf{E}_\theta \|\bar{h}^B\|_1^2 \leq \frac{5[R(\theta, h^*)/\varepsilon^2]^2}{(1 - 2\|h^*\|_1/n)^2} + C + n^2 \mathbf{P}(A^c).$$

To finish the proof, it suffices to note that in view of the Markov inequality

$$\mathbf{P}(A^c) = \mathbf{P} \left\{ \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} (1 - \xi_k^2) \geq \frac{\sqrt{n}}{16} \right\} \leq \exp(-Cn). \quad \square$$

Lemma 7. *Let $\tilde{h} \in \mathcal{H}^n$ be an arbitrary data-driven projection filter. Then*

$$(33) \quad \mathbf{E}_\theta \left[\varepsilon^2 - \frac{\|\tilde{h}Z - Z\|^2}{n} \right]_+^2 \leq \frac{C\varepsilon^4}{n} + \frac{C\varepsilon^4}{n^2} \mathbf{E}_\theta \|\tilde{h}\|_1^2.$$

Proof. We have

$$(34) \quad \varepsilon^2 - \frac{\|\tilde{h}Z - Z\|^2}{n} = \frac{\varepsilon^2}{n} \sum_{k=0}^{n-1} (1 - \xi_k^2) + \frac{2\varepsilon^2}{n} \sum_{k=0}^{n-1} \tilde{h}_k^2 - \frac{1}{n} \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 \theta_k^2 \\ - \frac{2\varepsilon}{n} \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 \theta_k \xi_k + \frac{\varepsilon^2}{n} \sum_{k=0}^{n-1} \tilde{h}_k (\xi_k^2 - 1) - \frac{\varepsilon^2}{n} \sum_{k=0}^{n-1} \tilde{h}_k^2.$$

Since

$$(35) \quad \mathbf{E} \left[\frac{1}{n} \sum_{k=0}^{n-1} (\xi_k^2 - 1) \right]^2 = \frac{2}{n},$$

it remains to control the last two lines in the right-hand side of (34). By Lemma 3 we get

$$\mathbf{E}_\theta \left[-2\varepsilon \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 \theta_k \xi_k - \sum_{k=0}^{n-1} (1 - \tilde{h}_k)^2 \theta_k^2 \right]^2 \leq C\varepsilon^4, \\ \mathbf{E}_\theta \left[\sum_{k=0}^{n-1} \tilde{h}_k (\xi_k^2 - 1) - \sum_{k=0}^{n-1} \tilde{h}_k^2 \right]^2 \leq C.$$

Finally combining these inequalities with (34), (35), we arrive at (33). \square

Proof of Theorem 5. We can easily complete it with Lemmas 6 and 7. The idea is to bound from below the right-hand side of (29). First of all note that from (29) we get

$$(36) \quad \mathbf{E}_\theta \|\bar{h}^B Z - Z\|^2 + \frac{2}{n} \mathbf{E}_\theta \|\bar{h}^B Z - Z\|^2 \|\bar{h}^B\|_1 - n\varepsilon^2 \\ = \mathbf{E}_\theta \|\bar{h}^B Z - Z\|^2 - \varepsilon^2 n + 2\varepsilon^2 \mathbf{E}_\theta \|\bar{h}^B\|_1 \\ + 2\mathbf{E}_\theta \|\bar{h}^B\|_1 \left[\frac{\|\bar{h}^B Z - Z\|^2}{n} - \varepsilon^2 \right] \leq \frac{R(\theta, h^*)}{1 - 2\|\bar{h}^*\|_1/n}.$$

The first three terms in the left-hand side can be controlled by a simple algebra and Lemma 3. Using (19), we get

$$\mathbf{E}_\theta \|\bar{h}^B Z - Z\|^2 - \varepsilon^2 n + 2\varepsilon^2 \mathbf{E}_\theta \|\bar{h}^B\|_1 \\ = \mathbf{E}_\theta \|\bar{h}^B Z - \theta\|^2 + 2\varepsilon \mathbf{E}_\theta \sum_{k=0}^{n-1} (1 - \bar{h}_k^B) \theta_k \xi_k + 2\varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^B (1 - \xi_k^2) \\ = (1 - \gamma) \mathbf{E}_\theta \|\bar{h}^B Z - \theta\|^2 + 2\varepsilon \mathbf{E}_\theta \sum_{k=0}^{n-1} (1 - \bar{h}_k^B) \theta_k \xi_k + \gamma \mathbf{E}_\theta \sum_{k=0}^{n-1} (1 - \bar{h}_k^B)^2 \theta_k^2 \\ + (2 - \gamma) \varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^B (1 - \xi_k^2) + \gamma \varepsilon^2 \mathbf{E}_\theta \sum_{k=0}^{n-1} \bar{h}_k^B.$$

Next in view of Lemma 3, we obtain

$$(37) \quad \mathbf{E}_\theta \|\bar{h}^B Z - Z\|^2 - \varepsilon^2 n + 2\varepsilon^2 \mathbf{E}_\theta \|\bar{h}^B\|_1 \geq (1-\gamma) \mathbf{E}_\theta \|\bar{h}^B Z - \theta\|^2 - \frac{2\varepsilon^2}{\gamma} - \frac{2\varepsilon}{U^{-1}(\gamma/2)}.$$

The last term in the left-hand side of (36) can be bounded by the Cauchy-Schwarz inequality and Lemmas 6 and 7,

$$\begin{aligned} \mathbf{E}_\theta \|\bar{h}^B\|_1 \left[\frac{\|\bar{h}^B Z - Z\|^2}{n} - \varepsilon^2 \right] &\geq -\mathbf{E}_\theta \|\bar{h}^B\|_1 \left[\varepsilon^2 - \frac{\|\bar{h}^B Z - Z\|^2}{n} \right]_+ \\ &\geq -\frac{CR^+(\theta)}{\sqrt{n}} - \frac{C[R^+(\theta)]^2}{\varepsilon^2 n} \geq -\frac{C[R^+(\theta)]^2}{\varepsilon^2 n} - \varepsilon^2. \end{aligned}$$

This inequality and (36), (37) complete the proof of the theorem. \square

3.5. PROOF OF THEOREM 6. It is quite similar to that of Theorem 5. The only difference is related to inequality (29), which is now replaced by

$$(38) \quad \|\bar{h}^C Z - Z\|^2 + \frac{2}{n} \|\bar{h}^C Z - Z\|^2 \|\bar{h}^C\|_1 - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2 \leq \frac{\|h^* Z - Z\|^2}{(1 - \|h^*\|_1/n)^2} - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2.$$

This inequality follows immediately from the definition of \bar{h}^C and the elementary inequality

$$\frac{1}{(1-x/2)^2} \geq 1+x, \quad x \in [0, 2).$$

It is not hard to check that

$$(39) \quad \mathbf{E}_\theta \left[\frac{\|h^* Z - Z\|^2}{(1 - \|h^*\|_1/n)^2} - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2 \right]^2 \leq \frac{5[R(\theta, h^*)]^2}{(1 - \|h^*\|_1/n)^4},$$

$$(40) \quad \mathbf{E}_\theta \left[\frac{\|h^* Z - Z\|^2}{(1 - \|h^*\|_1/n)^2} - \varepsilon^2 \sum_{k=0}^{n-1} \xi_k^2 \right] \leq \frac{R(\theta, h^*)}{(1 - \|h^*\|_1/n)^2}.$$

Therefore we get from (38) and (39) (see the proof of Lemma 6)

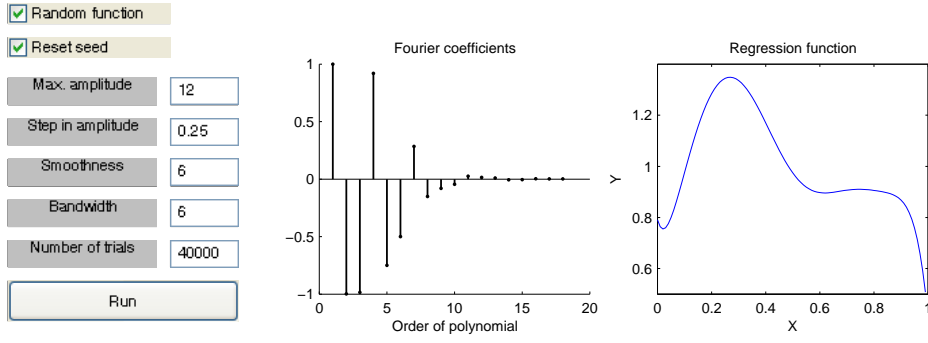
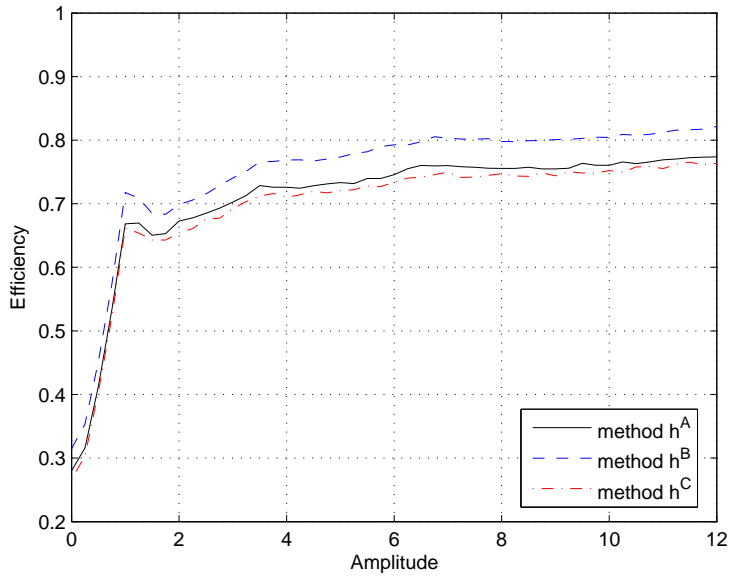
$$\mathbf{E}_\theta \|\bar{h}^C\|_1^2 \leq C[R(\theta, h^*)/\varepsilon^2]^2.$$

Using the above inequality and (38) and (40), we complete the proof of the theorem (see the proof of Theorem 5 for more detail). \square

4. Simulations

In this section, we illustrate numerically Theorems 1, 5, and 6. Our basic idea is to measure the statistical performance of a data-driven filter \tilde{h} by its *oracle efficiency* defined by

$$e_{\text{or}}(\theta, \tilde{h}) = \frac{\inf_{h \in \mathcal{H}^n} \mathbf{E}_\theta \|hZ - \theta\|^2}{\mathbf{E}_\theta \|\tilde{h}Z - \theta\|^2}.$$

FIGURE 1. Model parameters for \bar{h}^A , \bar{h}^B , and \bar{h}^C FIGURE 2. Oracle efficiencies of \bar{h}^A , \bar{h}^B , and \bar{h}^C

Obviously, we cannot compute this efficiency for all θ 's. Therefore we choose a sufficiently simple, but representative family of vectors θ . In what follows we will use the following linear family:

$$\theta_k(a) = \frac{a\varepsilon\mu_k}{1 + (k/W)^m},$$

where a is called amplitude, W bandwidth, and m smoothness, and the μ_k are i.i.d. random variables taking values $-1, +1$ with equal probabilities. We vary a in a large range and plot $e_{\text{or}}(\theta(a), \tilde{h})$ as a function of a . The parameters $m = 6$ and $W = 6$ are fixed. In other examples of (W, m) the authors looked at, simulations showed that the oracle efficiency exhibited a similar behavior. Notice that for

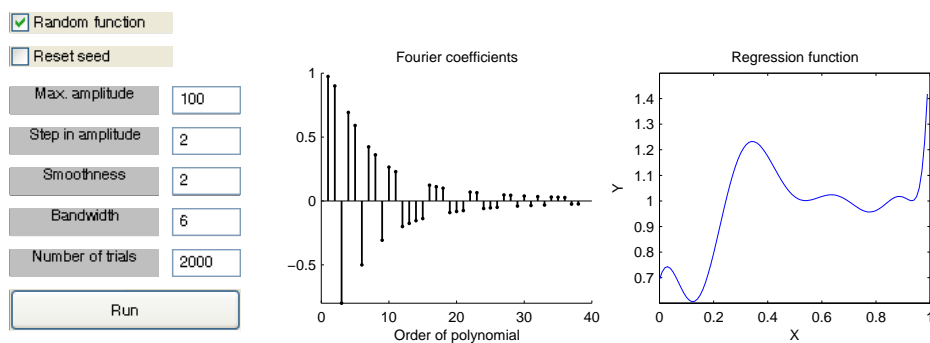


FIGURE 3. Model parameters for the convex combinations approach

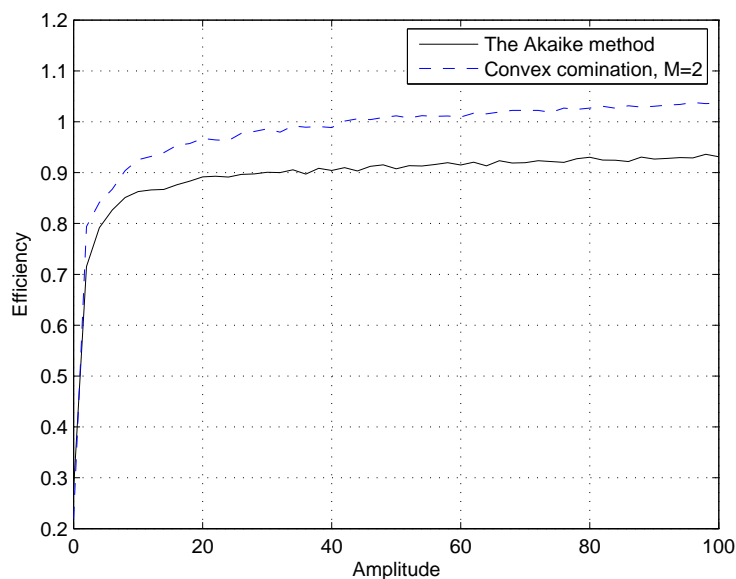


FIGURE 4. Oracle efficiency of the convex combination of two projection estimates

all data-driven filters considered in the paper, $e_{\text{or}}(\theta(a), \tilde{h})$ does not depend on ε . In our simulations, this function was computed by the Monte-Carlo method with 40000 replications. The parameters of the numerical experiment (the coefficients $\theta_k(a)$ and the corresponding regression function) are shown in Figure 1. Figure 2 represents graphically the oracle efficiencies for three methods: the classical Akaike method with known variance (solid line), the method (5) (dashed line), and (6) (dash-dot line). Even a quick look at this plot shows that all these methods work well: the minimal oracle efficiency is about 0.3 and for moderately large $a > 3$ it is greater than 0.7. We see also that the unknown noise variance does affect the oracle efficiency. Moreover, at the first glance, what we see here looks surprising:

the method (5) outperforms the Akaike criterion. In fact, this phenomenon is not very surprising, since it illustrates the fact that the principle of unbiased estimation is not optimal. There exist more efficient algorithms based on the idea of risk hull minimization (see Cavalier and Golubev (2006) for detail).

We finish this section with very short comments on numerical properties of the convex combination of projection methods. From the asymptotic viewpoint (see Theorems 4 and 1), this method should work better than the simple projection technique since

$$\min_{h \in \mathcal{C}_M^n} R(\theta, h) \leq \min_{h \in \mathcal{H}^n} R(\theta, h).$$

On the other hand, the remainder term in Theorem 4 is greater than the one in Theorem 1. So, for small data sets, the real advantages of the method based on convex combinations are not clear. Moreover, in majority of numerical simulations we looked at, the performances of both methods were similar. However, the convex combination exhibits better performance for slowly decreasing θ_k^2 . We illustrate this effect in the next numerical experiment with parameters shown in Figure 3. Here $m = 2$, $W = 6$; the number of Monte-Carlo replications is 2000. The oracle efficiencies of the simple projection Akaike filter (solid line) and the convex combination of two projection estimates (dashed line) are shown in Figure 4. We see that the convex combination works only slightly better than the standard method. On the other hand, we would like to point out that the numerical complexity of the convex combination approach is rather high.

References

- [1] H. Akaike (1973), *Information theory and an extension of the maximum likelihood principle*. In: *Proc. 2nd Intern. Symp. Inform. Theory* (P. N. Petrov and F. Csaki, eds.), Budapest, pp. 267–281.
- [2] A. Barron, L. Birgé, and P. Massart (1999), *Risk bounds for model selection via penalization*, *Probab. Theory Rel. Fields*, 113, 301–413.
- [3] L. Birgé and P. Massart (2001), *Gaussian model selection*, *J. European Math. Soc.*, 3, 203–268.
- [4] L. Cavalier and Yu. Golubev (2006), *Risk hull method and regularization by projections of ill-posed inverse problems*, *Ann. Statist.* (to appear).
- [5] F. Cucker and S. Smale (2001), *On the mathematical foundations of learning*, *Bull. Amer. Math. Soc.*, 39, 1–49.
- [6] Yu. Golubev (2004), *On the method of empirical risk minimization*, *Problems Inform. Transmission*, no. 2, 21–32.
- [7] Yu. Golubev and B. Levit (2004), *An oracle approach to adaptive estimation of linear functionals in a Gaussian model*, *Math. Methods Statist.*, 13, 392–408.
- [8] A. Kneip (1994), *Ordered linear smoothers*, *Ann. Statist.*, 22, 835–866.
- [9] C. L. Mallows (1973), *Some comments on C_p* , *Technometrics*, 15, 661–675.
- [10] A. Nemirovski (2000), *Topics in non-parametric statistics*. In: *Lectures on Probability Theory and Statistics*, Ecole d’Eté de Probabilité de Saint-Flour XXVIII-1998, Springer, Berlin–Heidelberg, pp. 85–277.
- [11] R. Shibata (1981), *An optimal selection of regression variables*, *Biometrika*, 68, 45–54.
- [12] C. M. Stein (1981), *Estimation of the mean of a multivariate normal distribution*, *Ann. Statist.*, 9, 1135–1151.

[Received September 2005]