# Spectral cut-off regularizations for ill-posed linear models

E Chernousova, Yu Golubev

# Spectral Cut-off Regularizations for Ill-posed Linear Models

Chernousova, E.* and Golubev, Yu.†

**Abstract**

This paper deals with recovering an unknown vector $\beta$ from the noisy data $Y = X\beta + \sigma\xi$, where $X$ is a known $n \times p$ - matrix with $n \geq p$ and $\xi$ is a standard white Gaussian noise. In order to estimate $\beta$, a spectral cut-off estimate $\hat{\beta}^{\bar{m}}(Y)$ with a data-driven cut-off frequency $\bar{m}(Y)$ is used. The cut-off frequency is selected as a minimizer of the unbiased risk estimate of the mean square prediction error, i.e. $\bar{m} = \arg\min_m \big\{\|Y - X\hat{\beta}^m(Y)\|^2 + 2\sigma^2 m\big\}$. Assuming that $\beta$ belongs to an ellipsoid $\mathcal{W}$, we derive uppers bounds for the maximal risk $\sup_{\beta\in\mathbb{W}} \mathbf{E}\|\hat{\beta}^{\bar{m}}(Y) - \beta\|^2$ and show that $\hat{\beta}^{\bar{m}}(Y)$ is a rate optimal minimax estimator over $\mathcal{W}$.

**Keywords:** ill-posed linear model, spectral cut-off regularization, data-driven cut-off frequency, oracle inequality, minimax risk.

**2010 Mathematics Subject Classification:** Primary 62C99; secondary 62C10, 62C20, 62J05.

## 1 Introduction and main result

This paper deals with recovering an unknown vector $\beta \in \mathbb{R}^p$ from the noisy observations

$$Y = X\beta + \sigma\xi, \tag{1.1}$$

where $X$ is a known $n \times p$ - matrix with $n \geq p$ and $\xi$ is a standard white Gaussian noise. Let us emphasize that all results below can be extended

*Moscow Institute of Physics and Technology, Institutski per. 9, Dolgoprudny, 141700, Russsia, e-mail: lena-ezhova@rambler.ru

†Aix-Marseille Université, Centrale Marseille, I2M, UMR 7373, 13453 Marseille, France and Institute for Information Transmission Problems, e-mail: golubev.yuri@gmail.com.

to the case $p = \infty$ provided that $X\beta \in \ell_2$. For the sake of simplicity it is assumed also that the noise level $\sigma$ is known.

The standard way to estimate $\beta$ based on the observations (1.1) is to make use of the maximum likelihood estimate

$$\hat{\beta}(Y) = \arg\min_{\beta} \|Y - X\beta\|^2 = (X^\top X)^{-1} X^\top Y,$$

where here and in what follows $\|\cdot\|$ stands for the standard Euclidean norm.

It is well known that the mean square risk of this method is given by

$$\mathbf{E}\|\hat{\beta}(Y) - \beta\|^2 = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda(k)},$$

where $\lambda(k)$ are the eigenvalues of $X^\top X$, i.e.

$$X^\top X e_k = \lambda(k) e_k, \ k = 1, \ldots, p, \tag{1.2}$$

and so, it is clear that it may be very large when $X^\top X$ is ill-posed. In this case a regularization of $\hat{\beta}(Y)$ is required.

Nowadays, statisticians have at their disposal a very vast family of regularization methods (see e.g. [5]). In this paper, we focus on the so-called spectral cut-off regularizations which are computed with the help of the Singular Value Decomposition.

Let $e_k \in \mathbb{R}^p$ be the eigenvectors of $X^\top X$ (see (1.2)). Using this basis, we represent the vector of interest as follows :

$$\beta = \sum_{k=1}^{p} \langle \beta, e_k \rangle e_k = \sum_{k=1}^{p} \beta(k) e_k,$$

where $\langle \cdot, \cdot \rangle$ is the ordinary inner product in $\mathbb{R}^p$, and thus we obtain

$$Y = \sum_{k=1}^{p} \beta(k) X e_k + \sigma \xi. \tag{1.3}$$

Next, noticing that

$$e_k^* = \frac{X e_k}{\sqrt{\lambda(k)}}, \ k = 1, \ldots, p$$

is an orthonormal system in $\mathbb{R}^n$ and projecting (1.3) onto these basis vectors, we arrive at the following equivalent representation of $Y$

$$Z(k) \stackrel{\text{def}}{=} \langle Y, e_k^* \rangle = \sqrt{\lambda(k)} \beta(k) + \sigma \xi'(k), \ k = 1, \ldots, p, \tag{1.4}$$

where $\xi'$ is a standard white Gaussian noise.

Spectral cut-off estimators of $\beta$ are defined by

$$\hat{\beta}^m(Y) \stackrel{\text{def}}{=} \sum_{k=1}^{m} \frac{Z(k)}{\sqrt{\lambda(k)}} e_k = \sum_{k=1}^{m} \frac{\langle Y, e_k^* \rangle}{\sqrt{\lambda(k)}} e_k, \ m = 1, \ldots, p, \tag{1.5}$$

where integer $m$ is often called cut-off frequency.

In view of (1.4), statistical analysis of this method for a fixed cut-off frequency is rather simple. If the performance of $\hat{\beta}^m(Y)$ is measured by the prediction mean square error then one can check easily that

$$r(\beta, m) \stackrel{\text{def}}{=} \mathbf{E}\|X[\hat{\beta}^m(Y) - \beta]\|^2 = \sum_{k=m+1}^{p} \lambda(k)\beta^2(k) + \sigma^2 m. \tag{1.6}$$

On the other hand, for the standard mean square risk one obtains

$$R(\beta, m) \stackrel{\text{def}}{=} \mathbf{E}\|\hat{\beta}^m(Y) - \beta\|^2 = \sum_{k=m+1}^{p} \beta^2(k) + \sigma^2 \sum_{k=1}^{m} \frac{1}{\lambda(k)}. \tag{1.7}$$

Thus, we see that in the both cases the risks depend on $m$ and to get a good estimate of $\beta$ we have to select properly the cut-off frequency. Since $\beta$ is unknown, this selection must be data-driven. The standard approaches to the data-driven choice of $m$ are often based on the principle of the unbiased risk estimation and go back to [1, 13]. There are two basic methods.

The first one is related to the unbiased estimate of $\mathbf{E}\|X[\hat{\beta}^m(Y) - \beta]\|^2$ and yields the following cut-off frequency :

$$\bar{m} = \arg\min_m \Big\{ \|Y - X\hat{\beta}^m(Y)\|^2 + 2\sigma^2 m \Big\}. \tag{1.8}$$

Notice that there is a vast literature devoted to statistical analysis of this method. The most precise fact about its performance was firstly obtained in [11].

**Theorem 1** *Uniformly in* $\beta \in \mathbb{R}^p$

$$\mathbf{E}\|X(\hat{\beta}^{\bar{m}} - \beta)\|^2 \le r_\circ(\beta) + K\sigma^2 \sqrt{\frac{r_\circ(\beta)}{\sigma^2}}, \tag{1.9}$$

*where* $K$ *is a generic constant and* $r_\circ(\beta)$ *is the so-called oracle risk defined by*

$$r_\circ(\beta) \stackrel{\text{def}}{=} \min_m r(\beta, m).$$

Unfortunately, with the help of this theorem we cannot obtain good upper bounds for the risk of $\hat{\beta}^{\bar{m}}(Y)$ measured by $\mathbf{E}\|\hat{\beta}^{\bar{m}}(Y)-\beta\|^2$. Of course, one can bound this risk as follows

$$\mathbf{E}\|\hat{\beta}^{\bar{m}}(Y)-\beta\|^2 \le \lambda^{-1}(p)\mathbf{E}\|X(\hat{\beta}^{\bar{m}}(Y)-\beta)\|^2 \approx \lambda^{-1}(p)r_\circ(\beta), \qquad (1.10)$$

but it can be seen easily that the right-hand side in this equation may be very far from the oracle risk (see (1.7)) defined by

$$R_\circ(\beta) \stackrel{\text{def}}{=} \min_m R(\beta, m) \qquad (1.11)$$

when $X^\top X$ is ill-posed.

Therefore, in order to get good upper bounds for $\mathbf{E}\|\hat{\beta}^{\tilde{m}}(Y)-\beta\|^2$, the cut-off frequency $\tilde{m}$ is selected as a minimizer of the unbiased risk estimate of $\mathbf{E}\|\hat{\beta}^{m}(Y)-\beta\|^2$, i.e.,

$$\tilde{m} = \arg\min_m \left\{ \|\hat{\beta}(Y)-\hat{\beta}^m(Y)\|^2 + 2\sigma^2 \sum_{k=1}^m \frac{1}{\lambda(k)} \right\}. \qquad (1.12)$$

The risk of this method is controlled by following oracle inequality (see, [4]).

**Theorem 2** *Suppose $\lambda(k) \ge \lambda(1)k^{-\alpha}$ for some $\alpha > 0$. Then uniformly in $\beta \in \mathbb{R}^p$*

$$\mathbf{E}\|\hat{\beta}^{\tilde{m}}(Y)-\beta\|^2 \le R_\circ(\beta) + C(\alpha)\sigma^2 \left[\frac{R_\circ(\beta)}{\sigma^2}\right]^{(2\alpha+1)/(2\alpha+2)}, \qquad (1.13)$$

*where $C(\alpha) \ge K\alpha$ is a constant depending on $\alpha$ and the oracle risk $R_\circ(\beta)$ is defined by* (1.11).

Comparing (1.9) and (1.13) one can see the principal difference between these oracle inequalities: the remainder term in (1.9) doesn't depend on $\alpha$, whereas the one in (1.13) goes to infinity as $\alpha \to \infty$. This observation confirms a suspicion that $\hat{\beta}^{\tilde{m}}(Y)$ may perform poorly when $X^\top X$ is ill-posed. This effect can be seen in simulation, see e.g. [4] or Section 2 below.

In order to improve the performance of the unbiased risk estimation method in the ill-posed case, penalties heavier than $2\sigma^2\sum_{k=1}^m \lambda^{-1}(k)$ should be used, see for details [4, 9]. It is shown in these papers that the estimate $\hat{\beta}^{\tilde{m}^+}(Y)$ with

$$\tilde{m}^+ = \arg\min_m \left\{ \|\hat{\beta}(Y)-\hat{\beta}^m(Y)\|^2 + 2\sigma^2 \left[\sum_{k=1}^m \frac{1}{\lambda(k)} + Pen_\lambda(m)\right] \right\} \qquad (1.14)$$

works better than $\hat{\beta}^{\bar{m}}(Y)$ provided that the additional penalty $Pen_\lambda(m)$ is properly chosen. Unfortunately, computing good penalties $Pen_\lambda(m)$ in (1.14) is a time-consuming numerical problem (see, e.g. [4, 9]).

This is why the main goal in this paper is to improve significantly Inequality (1.10) and to show that the standard spectral cut-off estimator

$$\hat{\beta}^{\bar{m}}(Y) = \sum_{k=1}^{\bar{m}} \frac{\langle Y, e_k^* \rangle}{\sqrt{\lambda(k)}} e_k, \ \bar{m} = \arg\min_m \Big\{ \|Y - X\hat{\beta}^m(Y)\|^2 + 2\sigma^2 m \Big\} \quad (1.15)$$

has a reasonable risk.

Unfortunately, in general case, point-wise oracle inequalities having the form

$$\mathbf{E}\|\hat{\beta}^{\bar{m}}(Y) - \beta\|^2 \le R_\circ(\beta) + remainder\ terms$$

cannot be proved.

However, as we will see below, it is possible to bound from above the maximal risk $\sup_{\beta\in\mathcal{W}} \mathbf{E}\|\hat{\beta}^{\bar{m}}(Y) - \beta\|^2$ with the help of $\sup_{\beta\in\mathcal{W}} R_\circ(\beta)$, where $\mathcal{W}$ is an ellipsoid in $\mathbb{R}^p$ defined by

$$\mathcal{W} = \Big\{ \beta : \sum_{k=1}^{p} w(k)\beta^2(k) \le 1 \Big\}.$$

Here $w(k)$, $k = 1, \dots, p$ is a positive increasing sequence.

In order to obtain such upper bounds, we will need some basic notions and facts related to the minimax estimation theory over ellipsoids. Nowadays, the minimax approach to ill-posed linear models is well-developed and literature on this topic is very vast, see e.g. [2, 3, 6, 12, 14], where additional references can be found.

With $\mathcal{W}$ we associate the following risks :

$$r(\mathcal{W}, m) \stackrel{\text{def}}{=} \max_{\beta\in\mathcal{W}} r(\beta, m), \quad R(\mathcal{W}, m) \stackrel{\text{def}}{=} \max_{\beta\in\mathcal{W}} R(\beta, m),$$

and

$$r_\circ(\mathcal{W}) \stackrel{\text{def}}{=} \min_m r(\mathcal{W}, m), \quad R_\circ(\mathcal{W}) \stackrel{\text{def}}{=} \min_m R(\mathcal{W}, m).$$

**Proposition 1** *For any integer* $m$

$$r(\mathcal{W}, m) = \frac{\lambda(m+1)}{w(m+1)} + \sigma^2 m, \quad R(\mathcal{W}, m) = \frac{1}{w(m+1)} + \sigma^2 \sum_{k=1}^{m} \frac{1}{\lambda(k)}.$$

*Proof.* It follows immediately from (1.6) and (1.7). □

Let us emphasize that in general case the spectral cut-off estimates are not sharp minimax over ellipsoids, they are only rate optimal. Minimax estimates optimal up to a constant have been obtained in the seminal paper [15].

To simplify some technical details, it is assumed in what follows that

$$\lambda(k) = Lk^{-\alpha} \quad \text{and} \quad w(k) = k^{\gamma}/P,$$

where $\alpha, \gamma, L, P$ are some positive constants. In order to make computations more transparent, let us assume also that the noise level is small, i.e. $\sigma \to 0$. Then we obtain easily

$$r(\mathcal{W}, m) = LP(m+1)^{-\gamma-\alpha} + \sigma^2 m$$

and denoting for brevity

$$q \stackrel{\text{def}}{=} \frac{1}{\gamma + \alpha + 1},$$

we get a simple algebra the minimax cut-off frequency

$$m_r(P, \sigma^2) = \arg\min_m \left\{ PL(m+1)^{-\gamma-\alpha} + \sigma^2 m \right\} = (1 + o(1)) \left[ \frac{(\gamma+\alpha)PL}{\sigma^2} \right]^q.$$

With this cut-off frequency we obtain the following formula for the minimax risk

$$r_\circ(\mathcal{W}) = (1 + o(1)) \left( \frac{q}{1-q} \right)^{1-q} \sigma^2 \left( \frac{PL}{\sigma^2} \right)^q. \tag{1.16}$$

The same technique is used in computing $R_\circ(\mathcal{W})$. We have

$$R(\mathcal{W}, m) = P(m+1)^{-\gamma} + \frac{\sigma^2}{L} \sum_{k=1}^{m} k^{\alpha}$$

and therefore we get the following minimax cut-off frequency

$$m_R(P, \sigma^2) = \arg\min_m \left\{ P(m+1)^{-\gamma} + \frac{\sigma^2}{L} \sum_{k=1}^{m} k^{\alpha} \right\} = (1 + o(1)) \left( \frac{\gamma PL}{\sigma^2} \right)^q$$

and the minimax risk

$$\begin{aligned} R_\circ(\mathcal{W}) =& (1 + o(1)) \frac{\gamma^{-q\gamma}}{q(\alpha+1)} P \left( \frac{\sigma^2}{PL} \right)^{q\gamma} \\ =& (1 + o(1)) \frac{\gamma^{-q\gamma}}{q(\alpha+1)} \frac{\sigma^2}{L} \left( \frac{\sigma^2}{PL} \right)^{-q(1+\alpha)}. \end{aligned} \tag{1.17}$$

Comparing (1.17) and (1.16), we see that the minimax risks $r_\circ(\mathcal{W})$ and $R_\circ(\mathcal{W})$ are related as follows :

$$R_\circ(\mathcal{W}) = (1+o(1))\frac{\gamma^{-q\gamma}(\gamma+\alpha)^{(1+\alpha)(q-1)}}{q(\alpha+1)}\frac{\sigma^2}{L}\left[\frac{r_\circ(\mathcal{W})}{\sigma^2}\right]^{\alpha+1}. \tag{1.18}$$

The second important remark is that the optimal minimax cut-off frequencies have the same order, i.e.,

$$\lim_{\sigma\to 0}\frac{m_R(P,\sigma^2)}{m_r(P,\sigma^2)} = \left(\frac{\gamma}{\gamma+\alpha}\right)^{1/(1+\alpha+\gamma)}. \tag{1.19}$$

In practice, we cannot make use of the minimax cut-off frequencies because they strongly depend on the ellipsoid parameters that are hardly known in practice. However, Equations (1.18) and (1.19) may be viewed as a heuristic motivation of $\hat{\beta}^{\bar{m}}(Y)$. These equations show that there are strong links between the minimax risks $r(\mathcal{W})$ and $R(\mathcal{W})$ as well between the cut-off frequencies $m_R(P,\sigma^2)$ and $m_r(P,\sigma^2)$. So, there is a hope the cut-off frequency $\bar{m}(Y)$, which is nearly optimal (see Theorem 1) when the risk is measured by the prediction error $\mathbf{E}\|X[\hat{\beta}^{\bar{m}}(Y)-\beta]\|^2$, is also good for the risk $\mathbf{E}\|\hat{\beta}^{\bar{m}}(Y)-\beta\|^2$.

The next theorem provides a mathematical justification of this conjecture.

**Theorem 3** *For any $\beta\in\mathcal{W}$*

$$\begin{aligned}\mathbf{E}\|\hat{\beta}^{\bar{m}}(Y)-\beta\|^2 \le & \frac{2\sigma^2}{L}\left[\frac{PL}{\sigma^2}\right]^{\alpha/(\alpha+\gamma)}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}+K\right]^{2\gamma/(\alpha+\gamma)} \\ & + \frac{\sigma^2}{(\alpha+1)L}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}+K\alpha\right]^{2\alpha+2} \\ & + \frac{K\sigma^2}{L\sqrt{2\alpha+1}}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}+K\alpha\right]^{2\alpha+1} + K\sigma^2,\end{aligned} \tag{1.20}$$

*where $K$ is a generic constant.*

With the help of this theorem one can easily check that $\hat{\beta}^{\bar{m}}(Y)$ is a rate optimal asymptotically minimax estimator over $\mathcal{W}$.

**Theorem 4** *As $\sigma\to 0$*

$$\sup_{\beta\in\mathcal{W}}\mathbf{E}\|\hat{\beta}^{\bar{m}}(Y)-\beta\|^2 \le (1+o(1))C(\alpha,\gamma)R_\circ(\mathcal{W}), \tag{1.21}$$

*where*

$$C(\alpha,\gamma)=\frac{\gamma^{\gamma/(1+\gamma+\alpha)}}{1+\alpha+\gamma}\Big[2(\alpha+1)+(\gamma+\alpha)^{(1+\alpha)(\alpha+\gamma)/(1+\alpha+\gamma)}\Big].$$

*Proof.* It is clear that $r_\circ(\beta)\le r_\circ(\mathcal{W})$ for any $\beta\in\mathcal{W}$, and thus by (1.16) and (1.17) we obtain

$$\begin{aligned}&\frac{2\sigma^2}{L}\left(\frac{PL}{\sigma^2}\right)^{\alpha/(\alpha+\gamma)}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}+K\right]^{2\gamma/(\alpha+\gamma)}\\&\quad=(1+o(1))\frac{2\sigma^2}{L}\left(\frac{PL}{\sigma^2}\right)^{\alpha/(\alpha+\gamma)}\left[\frac{r_\circ(\beta)}{\sigma^2}\right]^{\gamma/(\alpha+\gamma)}\\&=2(1+o(1))P\left(\frac{\sigma^2}{LP}\right)=2q(\alpha+1)\gamma^{q\gamma}(1+o(1))R_\circ(\mathcal{W}).\end{aligned}$$

Similarly, by (1.18) we have

$$\begin{aligned}\frac{\sigma^2}{(\alpha+1)L}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}+K\alpha\right]^{2\alpha+2}&=\frac{1+o(1)}{1+\alpha}\frac{\sigma^2}{L}\left[\frac{r_\circ(\beta)}{\sigma^2}\right]^{1+\alpha}\\&\le q\gamma^{q\gamma}(\gamma+\alpha)^{(1+\alpha)(1-q)}(1+o(1))R_\circ(\mathcal{W})\end{aligned}$$

and

$$\frac{\sigma^2}{(\alpha+1)L}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}+K\alpha\right]^{2\alpha+1}\le o(1)R_\circ(\mathcal{W}).$$

Substituting the above equations in (1.20), we complete the proof. □

**Remark.** The spectral cut-off method requires the SVD which may be time-consuming from a numerical viewpoint when $p$ is large. This is why for very large linear models simpler regularization techniques are usually used. A classic example of such a regularization is the famous Phillips-Tikhonov method [16] known in statistics as ridge regression. In order to select the regularization parameter in this method, the Generalized Cross Validation is usually used (see, e.g. [7, 8]). Unfortunately, in spite of a very long history of this idea, all known facts about its performance are related to upper bounds for the prediction error (see e.g. Theorem 1). In view of Theorem 3, there is a hope that bounds similar to (1.20) and (1.21) may be obtained for this method as well, and so, the use of the GCV will be justified.

## 2 Simulation study

To illustrate numerically how does the spectral cut-off method with different data-driven cut-off frequencies work, the following simulations have been carried out. For given $A \in [0, 20]$, 20000 replications of the observations

$$Z(k) = \sqrt{\lambda(k)}\mu(k) + \xi'(k),\ k = 1, \dots, 300$$

were generated. Here $\xi'$ is a standard white Gaussian noise, $\mu \in \mathbb{R}^{300}$ is a Gaussian vector (independent from $\xi'$) with independent components and

$$\mathbf{E}\mu(k) = 0, \quad \mathbf{E}\mu^2(k) = A \exp\left(-\frac{k^2}{2\Sigma^2}\right).$$

To simulate the spectral cut-off estimator, the unknown vector $\mu$ is estimated as follows:

$$\hat{\mu}^m(k) = \frac{Z(k)}{\sqrt{\lambda(k)}}\mathbf{1}\{k \le m\}.$$

With the help of the Monte-Carlo method the following risks were computed

- the oracle risk
$$r_{or}(A) = \min_m \mathbf{E}\|\mu - \hat{\mu}^m\|^2,$$
- the risk related to the cut-off frequency minimizing the unbiased estimate of $\mathbf{E}\|\mu - \hat{\mu}^m\|^2$ , i.e.,
$$r_{inv}(A) = \mathbf{E}\|\mu - \hat{\mu}^{\tilde{m}}\|^2,$$
where
$$\tilde{m} = \arg\min_m \left\{ \|\lambda^{-1/2} \cdot Z - \hat{\mu}^m\|^2 + 2\sum_{k=1}^{m} \frac{1}{\lambda(k)} \right\}, \tag{2.1}$$
- the risk related to the cut-off frequency minimizing the unbiased estimate of $\mathbf{E}\|\sqrt{\lambda}(\mu - \hat{\mu}^m)\|^2$, i.e.,
$$r_{dir}(A) = \mathbf{E}\|\mu - \hat{\mu}^{\bar{m}}\|^2,$$
where
$$\bar{m} = \arg\min_m \left\{ \|Z - \sqrt{\lambda} \cdot \hat{\mu}^m\|^2 + 2m \right\}, \tag{2.2}$$

- the risk related to the cut-off frequency minimizing the over-penalized unbiased risk estimator, i.e. ,

$$r_{ovp}(A) = \mathbf{E}\|\mu - \hat{\mu}^{\bar{m}_+}\|^2,$$

where

$$\bar{m}_+ = \arg\min_m \Big\{ \|Z - \sqrt{\lambda}\cdot\hat{\mu}^m\|^2 + 2m + \sqrt{2m\log(m)} \Big\}. \tag{2.3}$$

A motivation of this method and, in particular, the use of the additional penalty $\sqrt{2m\log(m)}$ can be found in [4].

Finally, the following oracle efficiencies were computed

$$\theta_{inv}(A) = \frac{r_{or}(A)}{r_{inv}(A)},\ \theta_{dir}(A) = \frac{r_{or}(A)}{r_{dir}(A)},\ \theta_{ovp}(A) = \frac{r_{or}(A)}{r_{ovp}(A)}.$$

In order to compare graphically the above methods of selection of the cut-off frequency, the data $\big\{A, \theta_{inv}(A), \theta_{dir}(A), \theta_{ovp}(A)\big\}$ were plotted on Figures 1, 2, 3, and 4.

Looking at these pictures we see that when $X^\top X$ becomes ill-posed, the spectral cut-off method with the cut-off frequency from (2.2) works significantly better than the cut-off frequency minimizing the unbiased estimate of $\mathbf{E}\|\mu - \hat{\mu}^m\|^2$. Let us emphasize also that in the case of ill-posed $X^\top X$ the data-driven cut-off frequency from (2.3) improves the risk of the spectral cut-off estimate.

# 3 Proofs

## 3.1 Auxiliary facts

The proof of Theorem 3 is based on two cornerstone facts. The first one is used to control the bias of $\hat{\beta}^{\bar{m}}(Y)$ and has the following form:

**Lemma 1** *For any $\beta \in \mathcal{W}$ and any integer $m = 1, \dots, p$*

$$\sum_{k=m}^{p} \beta^2(k) \le 2L^{-\gamma/(\alpha+\gamma)} P^{\alpha/(\alpha+\gamma)} \Big[ \sum_{k=m}^{p} \beta^2(k)\lambda(k) \Big]^{\gamma/(\alpha+\gamma)}. \tag{3.1}$$

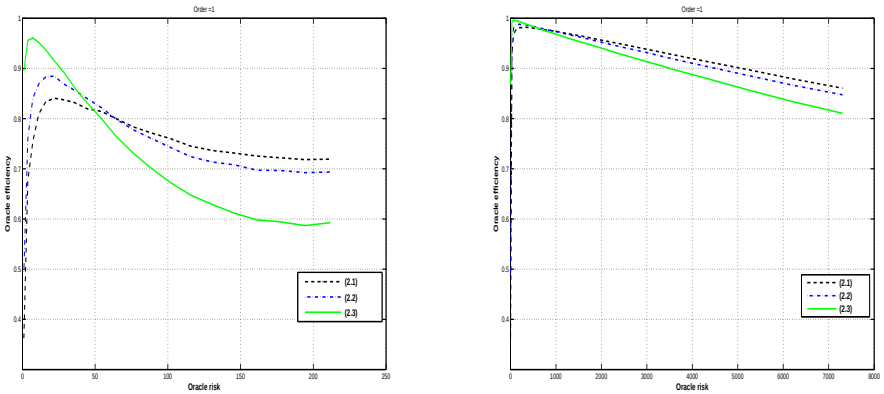


Figure 1: Oracle efficiencies of data-driven cut-off frequencies from (2.1), (2.2), and (2.3) for $\lambda(k) = k^{-1}$ (left panel $\Sigma = 10$; right panel $\Sigma = 100$).

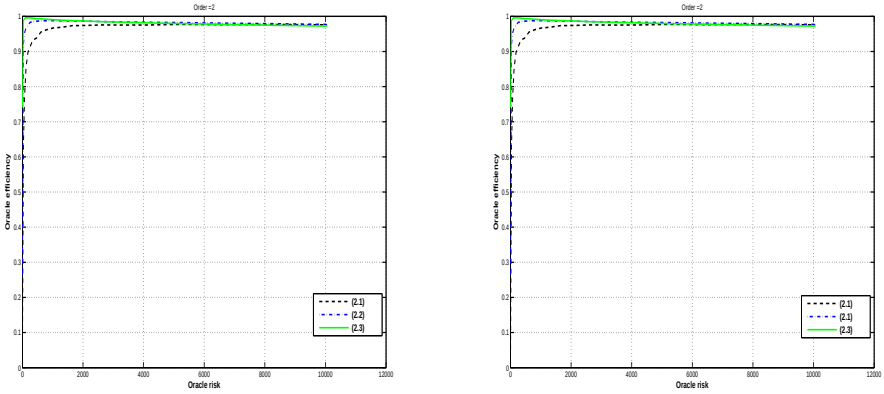


Figure 2: Oracle efficiencies of data-driven cut-off frequencies from (2.1), (2.2), and (2.3) for $\lambda(k) = k^{-2}$ (left panel $\Sigma = 10$; right panel $\Sigma = 100$).

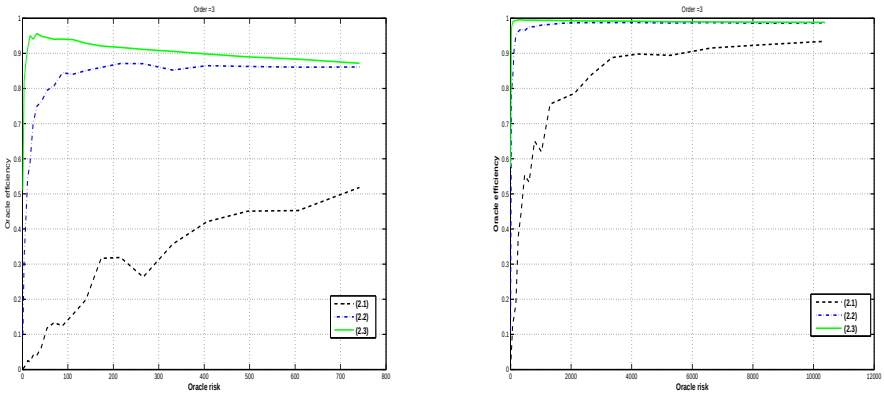


Figure 3: Oracle efficiencies of data-driven cut-off frequencies from (2.1), (2.2), and (2.3) for $\lambda(k) = k^{-3}$ (left panel $\Sigma = 10$; right panel $\Sigma = 100$).

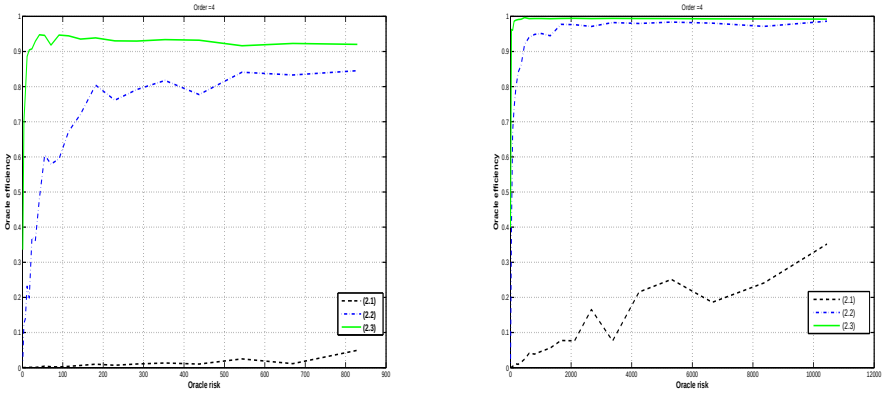


Figure 4: Oracle efficiencies of data-driven cut-off frequencies from (2.1), (2.2), and (2.3) for $\lambda(k) = k^{-4}$ (left panel $\Sigma = 10$; right panel $\Sigma = 100$).

*Proof.* Let us prove (3.1) for $p = \infty$. Since $\beta \in \mathcal{W}$ we have that for any $x \geq 1$

$$\begin{aligned}\sum_{k=m}^{\infty} \beta^2(k) \leq & \sum_{k=m}^{mx} \beta^2(k) + \sum_{k>mx} \beta^2(k) \\ \leq & \frac{1}{\lambda(mx)} \sum_{k=m}^{mx} \lambda(k)\beta^2(k) + \frac{1}{w(xm)} \\ \leq & \frac{x^\alpha}{\lambda(m)} \sum_{k=m}^{\infty} \lambda(k)\beta^2(k) + \frac{1}{w(m)x^\gamma} \\ = & \frac{1}{w(m)} \left[ x^\alpha \frac{w(m)}{\lambda(m)} \sum_{k=m}^{\infty} \lambda(k)\beta^2(k) + x^{-\gamma} \right].\end{aligned} \tag{3.2}$$

Our next step is to minimize the right-hand side in this equation in $x \geq 1$. Denote for brevity

$$S(m) = \frac{w(m)}{\lambda(m)} \sum_{k=m}^{\infty} \lambda(k)\beta^2(k)$$

and note that $S(m) \leq 1$ since $\beta \in \mathcal{W}$. So, we can choose $x$ as a root of the equation

$$x^\alpha S(m) = x^{-\gamma},$$

or, equivalently, $x = \big[S(m)\big]^{-1/(\alpha+\gamma)}$. Hence, substituting this in (3.2), we obtain

$$\begin{aligned}\sum_{k=m}^{\infty} \beta^2(k) \leq & \frac{2}{w(m)} \left[ \frac{w(m)}{\lambda(m)} \sum_{k=m}^{\infty} \lambda(k)\beta^2(k) \right]^{\gamma/(\alpha+\gamma)} \\ = & 2[w(m)]^{-\alpha/(\alpha+\gamma)} [\lambda(m)]^{-\gamma/(\alpha+\gamma)} \left[ \sum_{k=m}^{\infty} \lambda(k)\beta^2(k) \right]^{\gamma/(\alpha+\gamma)} \\ = & 2P^{\alpha/(\alpha+\gamma)} L^{-\gamma/(\alpha+\gamma)} \left[ \sum_{k=m}^{\infty} \lambda(k)\beta^2(k) \right]^{\gamma/(\alpha+\gamma)}.\end{aligned}$$

Equation (3.1) follows now from this inequality if we assume that $\beta(k) = 0,\ k > p$. $\square$

The following lemma plays an essential role in controlling the variance of $\hat{\beta}^{\bar{m}}(Y)$.

**Lemma 2** *For any $\beta$ and any $\mu > 0$*

$$\mathbf{E}\exp\Big\{\mu\sqrt{r(\beta,\bar{m})}\Big\} \le \exp\Big\{\mu\sqrt{r_\circ(\beta)} + K\sigma\mu + K\sigma^2\mu^2\Big\}, \qquad (3.3)$$

*where $K$ is a generic constant.*

**Proof.** It follows immediately from the definition of $\bar{m}$ (see (1.4), (1.5), and (1.8)) that for any given positive integer $m$ the following inequality holds

$$-\sum_{i=1}^{\bar{m}} Z^2(i) + 2\sigma^2\bar{m} \le -\sum_{i=1}^{m} Z^2(i) + 2\sigma^2 m$$

or, equivalently,

$$\begin{aligned} r(\beta,\bar{m}) \le r(\beta,m) &+ \sigma^2\sum_{i=1}^{\bar{m}}[\xi'^2(i)-1] + 2\sigma\sum_{i=\bar{m}+1}^{p}\sqrt{\lambda(i)}\beta(i)\xi'(i) \\ &-\sigma^2\sum_{i=1}^{m}[\xi'^2(i)-1] - 2\sigma\sum_{i=m+1}^{p}\sqrt{\lambda(i)}\beta(i)\xi'(i). \end{aligned}$$

We can easily derive from this inequality the following one

$$\begin{aligned} (1-\epsilon)r(\beta,\bar{m}) \le r(\beta,m) &+ \sigma^2\sup_{k\ge1}\Big\{\sum_{i=1}^{k}[\xi'^2(i)-1] - \epsilon k\Big\} \\ &+ 2\sigma\sup_{k\ge1}\Big\{\sum_{i=k+1}^{p}\sqrt{\lambda(i)}\beta(i)\xi'(i) - \frac{\epsilon}{2\sigma}\sum_{i=k+1}^{p}\lambda(i)\beta^2(i)\Big\} \qquad (3.4) \\ &+ \sigma^2\sum_{i=1}^{m}[1-\xi'^2(i)] - 2\sigma\sum_{i=m+1}^{p}\sqrt{\lambda(i)}\beta(i)\xi'(i), \end{aligned}$$

where $\epsilon \in (0,1)$.

The last line in this equation can be easily controlled since for any $\mu > 0$

$$\begin{aligned} &\mathbf{E}\exp\Big\{\mu\Big[\sigma^2\sum_{i=1}^{m}\big[1-\xi'^2(i)\big] - 2\sigma\sum_{i=m+1}^{p}\sqrt{\lambda(i)}\beta(i)\xi'(i)\Big]\Big\} \\ &\le \exp\Big[\sigma^2 m\mu - \frac{m}{2}\log\big(1+2\mu\sigma^2\big) + 2\sigma^2\mu^2\sum_{i=m+1}^{p}\lambda(i)\beta^2(i)\Big] \qquad (3.5) \\ &\le \exp\Big[\sigma^4 m\mu^2 + 2\sigma^2\mu^2\sum_{i=m+1}^{p}\lambda(i)\beta^2(i)\Big] \le \exp\big[2\sigma^2\mu^2 r(\beta,m)\big]. \end{aligned}$$

To bound from above $\sup_{k\geq 1}$ at the right-hand side in (3.4), we make use of the following inequalities (see Lemma 2 in [10])

$$\mathbf{P}\Big\{\max_{k\geq 1}\Big[\sum_{i=1}^{k}\big[\xi'^2(i)-1\big]-U(\epsilon)k\Big]\geq x\Big\}\leq \exp(-\epsilon x),$$
$$\mathbf{P}\Big\{\max_{k\geq 1}\Big[\sum_{i=k}^{p}\xi'(i)\beta(i)\sqrt{\lambda(i)}-\frac{\epsilon}{2}\sum_{i=k}^{p}\beta^2(i)\lambda(i)\Big]\geq x\Big\}\leq \exp(-\epsilon x),$$

where $\epsilon\in\mathbb{R}^+$ and

$$U(\epsilon)=-\frac{\epsilon+\log[1-2\epsilon]_+/2}{\epsilon}.$$

With the help of these inequalities and a simple algebra one can check easily that for any $\epsilon\in(0,1/4)$

$$\mathbf{E}\exp\Big\{\lambda\sigma^2\sup_{k\geq 1}\Big[\sum_{i=1}^{k}[\xi'^2(i)-1]-\epsilon k\Big]\Big\}\leq\frac{1}{[1-K\lambda\sigma^2/\epsilon]_+} \tag{3.6}$$

and

$$\begin{aligned}\mathbf{E}\exp\Big\{2\lambda\sigma\sup_{k\geq 1}\Big[\sum_{i=k+1}^{p}\sqrt{\lambda(i)}\beta(i)\xi'(i)-\frac{\epsilon}{2\sigma}\sum_{i=k+1}^{p}\lambda(i)\beta^2(i)\Big]\Big\}\\ \leq\frac{1}{[1-K\lambda\sigma^2/\epsilon]_+}.\end{aligned} \tag{3.7}$$

It follows from the Hölder inequality that for any random variables $\zeta_1,\zeta_2,\zeta_3$ with bounded exponential moments the following inequality holds

$$\mathbf{E}\exp\big[\lambda\zeta_1+\lambda\zeta_2+\lambda\zeta_3\big]\leq\prod_{i=1}^{3}\Big\{\mathbf{E}\exp\big[3\lambda\zeta_i\big]\Big\}^{1/3}.$$

With this equation and (3.4)-(3.7) we obtain

$$\begin{aligned}\mathbf{E}\exp\big\{(1-\epsilon)\lambda r(\beta,\bar m)\big\}\leq\exp\Big\{&\lambda r(\beta,m)+K\sigma^2\lambda^2 r(\beta,m)\\ &-\frac{2}{3}\log\Big[1-\frac{K\lambda\sigma^2}{\epsilon}\Big]_+\Big\}\end{aligned}$$

or equivalently,

$$\begin{aligned}\mathbf{E}\exp\big\{\lambda r(\beta,\bar m)\big\}\leq\exp\Big\{&\lambda r(\beta,m)+\frac{\epsilon\lambda r(\beta,m)}{1-\epsilon}\\ &+\frac{K\sigma^2\lambda^2 r(\beta,m)}{(1-\epsilon)^2}-\frac{2}{3}\log\Big[1-\frac{K\lambda\sigma^2}{(1-\epsilon)\epsilon}\Big]_+\Big\}.\end{aligned}$$

Minimizing the right-hand side at this equation in $m$ and recalling that

$$r_\circ(\beta) \stackrel{\text{def}}{=} \min_m r(\beta, m),$$

we arrive at the following inequality

$$\begin{aligned}
&\mathbf{E}\exp\bigl\{\lambda[r(\beta,\bar m) - r_\circ(\beta)]\bigr\} \\
&\quad\le \exp\biggl\{K\sigma^2\lambda^2 r_\circ(\beta) + K\epsilon\lambda r_\circ(\beta) - \frac{2}{3}\log\biggl[1 - \frac{K\lambda\sigma^2}{\epsilon}\biggr]_+\biggr\},
\end{aligned} \tag{3.8}$$

that holds for any $\epsilon \in (0, 1/2)$.

The next step is to minimize the right-hand side at (3.8) in $\epsilon \in (0, 1/2)$. Suppose

$$\lambda \le \frac{1}{4K\sigma^2}. \tag{3.9}$$

Since $r_\circ(\beta)/\sigma^2 \ge 1$, therefore

$$\epsilon = K\lambda\sigma^2 + \frac{\sigma}{4\sqrt{r_\circ(\beta)}} \le \frac{1}{2},$$

and substituting this $\epsilon$ in (3.8), we obtain

$$\begin{aligned}
\mathbf{E}\exp\bigl\{\lambda[r(\beta,\bar m) - r_\circ(\beta)]\bigr\} &\le \exp\biggl\{K\sigma^2\lambda^2 r_\circ(\beta) + K\sigma\lambda\sqrt{r_\circ(\beta)} \\
&\qquad + \frac{2}{3}\log\bigl[1 + 4K\lambda\sigma\sqrt{r_\circ(\beta)}\bigr]\biggr\} \\
&\le \exp\biggl\{K\sigma^2\lambda^2 r_\circ(\beta) + K\sigma\lambda\sqrt{r_\circ(\beta)}\biggr\}.
\end{aligned} \tag{3.10}$$

Let $\zeta$ be a nonnegative random variable. Then for any $\mu, \lambda > 0$ we have

$$\begin{aligned}
\mathbf{E}\exp\bigl\{\mu\sqrt{\zeta}\bigr\} &= \int_0^\infty \exp(\mu\sqrt{x} - \lambda x + \lambda x)\, d\mathbf{P}\{\zeta \le x\} \\
&\le \exp\bigl\{\max_{x\ge 0}\bigl[\mu\sqrt{x} - \lambda x\bigr]\bigr\} \int_0^\infty \exp(\lambda x)\, d\mathbf{P}\{\zeta \le x\} \\
&\le \exp\biggl\{\frac{\mu^2}{4\lambda} + \log\bigl[\mathbf{E}\exp(\lambda\zeta)\bigr]\biggr\}.
\end{aligned}$$

and combining this equation with (3.9) and (3.10), we have

$$\begin{aligned}
&\mathbf{E}\exp\biggl\{\mu\sqrt{r(\beta,\bar m)}\biggr\} \\
&\quad\le \exp\biggl\{\min_{\lambda\le 1/(4K\sigma^2)}\biggl[\frac{\mu^2}{4\lambda} + \lambda r + K\sigma\sqrt{r_\circ(\beta)}\lambda + K\sigma^2 r_\circ(\beta)\lambda^2\biggr]\biggr\}.
\end{aligned} \tag{3.11}$$

Note that choosing

$$\lambda = \frac{1}{4K\sigma^2},$$

we obtain from (3.11) that for any $\mu \in \mathbb{R}^+$

$$\mathbf{E}\exp\Big\{\mu\sqrt{r(\beta,\bar{m})}\Big\} \le \exp\Big\{K\mu^2\sigma^2 + K\frac{r_\circ(\beta)}{\sigma^2} + K\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}\Big\}. \qquad (3.12)$$

On the other hand, if $\mu$ is small, namely,

$$\mu \le \frac{2\sqrt{r_\circ(\beta)}}{4K\sigma^2},$$

then with

$$\lambda = \frac{\mu}{2\sqrt{r_\circ(\beta)}}$$

we get from (3.11)

$$\mathbf{E}\exp\Big\{\mu\sqrt{r(\beta,\bar{m})}\Big\} \le \exp\big\{\mu\sqrt{r_\circ(\beta)} + K\mu^2\sigma^2 + K\mu\sigma\big\}. \qquad (3.13)$$

Finally notice that if

$$\mu \ge \frac{2\sqrt{r_\circ(\beta)}}{4K\sigma^2},$$

then

$$\mu\sqrt{r_\circ(\beta)} + K\mu^2\sigma^2 + K\mu\sigma \ge K\mu^2\sigma^2 + K\frac{r_\circ(\beta)}{\sigma^2} + K\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}$$

and thus in view of (3.12), Equation (3.13) holds for any $\mu \in \mathbb{R}^+$. $\square$

The following simple technical lemma is used together with Lemma 2 to control the variance of $\hat{\beta}^{\bar{m}}(Y)$.

**Lemma 3** *Let $\eta \ge 1$ be a random variable such that*

$$\mathbf{E}\exp(\lambda\sqrt{\eta}) \le \exp(A\lambda + B\lambda^2)$$

*for any $\lambda > 0$. Then for any $p \ge 1$*

$$\mathbf{E}\eta^p \le [A + (2p-1)B]^{2p}. \qquad (3.14)$$

*Proof.* Let us consider the following function

$$f(x) = \log^{2p}(x),\ x \geq 0$$

and compute its second derivative

$$\begin{aligned} f''(x) &= -\frac{2p\log^{2p-1}(x)}{x^2} + \frac{2p(2p-1)\log^{2p-2}(x)}{x^2} \\ &= \frac{2p\log^{2p-2}(x)[(2p-1)-\log(x)]}{x^2}. \end{aligned}$$

So, $f(x)$ is concave on $[\exp(2p-1), \infty)$. Therefore noticing that

$$\eta^p = \frac{f[\exp(\lambda\sqrt{\eta})]}{\lambda^{2p}}$$

and using Jensen's inequality, we obtain that for any $\lambda \geq 2p-1$

$$\mathbf{E}\eta^p \leq \frac{f[\mathbf{E}\exp(\lambda\sqrt{\eta})]}{\lambda^{2p}} = \frac{[A\lambda + B\lambda^2]^{2p}}{\lambda^{2p}} = [A + B\lambda]^{2p}.$$

Minimizing the right-hand side at this equation in $\lambda \geq 2p-1$ we finish the proof. $\square$

In order to control the variance of $\hat{\beta}^{\bar{m}}(Y)$, we will need to bound from above $\mathbf{E}\zeta(\bar{m})$, where

$$\zeta(m) \stackrel{\text{def}}{=} \sum_{i=1}^{m} \frac{\xi'^2(i) - 1}{\lambda(i)}.$$

We will do this with the help of following fact. Let $\eta(t)$ be a separable zero mean random process on $\mathbb{R}^+$. Denote for brevity

$$\Delta_\eta(u,v) = \eta(u) - \eta(v).$$

**Lemma 4** *Let $\sigma^2(u),\ u \in \mathbb{R}^+$, be a continuous strictly increasing function with $\sigma^2(0) = 0$. Then for any $\mu > 0$,*

$$\begin{aligned} \log \mathbf{E}\exp\left\{\mu \max_{0<u\leq t} \frac{\Delta_\eta(u,t)}{\sigma(t)}\right\} \leq \frac{\log(2)\sqrt{2}}{\sqrt{2}-1} + \\ + \max_{0<u<v\leq t} \max_{|z|\leq\sqrt{2}/(\sqrt{2}-1)} \log \mathbf{E}\exp\left\{z\mu\frac{\Delta_\eta(u,v)}{\bar{\Delta}_\sigma(v,u)}\right\}, \end{aligned} \tag{3.15}$$

*where $\bar{\Delta}_\sigma(v,u) = \sqrt{|\sigma^2(v) - \sigma^2(u)|}$.*

**Proof.** It is similar to the one of Dudley's entropy bound (see, e.g., [17]) and can be found in [9]. $\square$

**Lemma 5** *Assume that* $\lambda(k)/\lambda(1) \geq 2^{-k}$. *Then for any* $\epsilon \in (0, 1/4)$

$$\mathbf{E} \max_{m \geq 1} \big[\zeta(m) - \epsilon \mathbf{E}\zeta^2(m)\big]_+ \leq \frac{K}{\epsilon}.$$

**Proof.** Let us apply Lemma 4 for the random process $\eta(k) = \zeta(k + k_0) - \zeta(k_0),\ k = 0, \ldots, t$ and $\sigma^2(k) = \mathbf{E}[\zeta(k + k_0) - \zeta(k_0)]^2$. Using a Taylor expansion, we obtain from (3.15)

$$\mathbf{E} \exp\Big\{\mu \sigma^{-1}(t) \max_{k=1,\ldots,t} \eta(k)\Big\} \leq K \exp(K\mu^2),\ \mu \in (0, \mu_\circ), \tag{3.16}$$

where

$$\mu_\circ = \frac{\sqrt{2} - 1}{4\sqrt{2}}.$$

Let

$$m_k^\epsilon = \max\Big\{m :\ \mathbf{E}\zeta^2(m) \leq \frac{2^k}{\epsilon^2}\Big\}.$$

Then we have

$$\begin{aligned} \mathbf{E} \max_{m \geq 1} \big[\zeta(m) - \epsilon \mathbf{E}\zeta^2(m)\big]_+ &\leq \mathbf{E} \sum_{k=0}^{p} \max_{m \in [m_k^\epsilon, m_{k+1}^\epsilon)} \big[\zeta(m) - \epsilon \mathbf{E}\zeta^2(m)\big]_+ \\ &\leq \sum_{k=0}^{p} \mathbf{E}\Big\{\zeta(m_k^\epsilon) + \max_{m \in [m_k^\epsilon, m_{k+1}^\epsilon)} [\zeta(m) - \zeta(m_k^\epsilon)] - \epsilon \mathbf{E}\zeta^2(m_k^\epsilon)\Big\}_+. \end{aligned} \tag{3.17}$$

Denote for brevity

$$\Sigma_k^\epsilon = \Big[2 \sum_{k=m_k^\epsilon + 1}^{m_{k+1}^\epsilon} \frac{1}{\lambda^2(k)}\Big]^{1/2}.$$

and notice that

$$\frac{2^{k/2}}{4\epsilon} \leq \Sigma_k^\epsilon \leq \frac{2^{k/2}}{\epsilon}. \tag{3.18}$$

To control the right-hand side at (3.17), we make use of the following inequality :

$$\mathbf{E}\big\{\zeta - x\big\}_+ \leq \mu^{-1} \exp(-\mu x) \mathbf{E} \exp(\mu\zeta),\ x > 0,$$

that holds true for any random variable $\zeta$ and any $\mu > 0$. With this inequality, (3.16), and (3.18) we arrive at

$$\begin{aligned}
&\mathbf{E}\Big\{\zeta(m_k^\epsilon) + \max_{m\in[m_k^\epsilon, m_{k+1}^\epsilon)}[\zeta(m) - \zeta(m_k^\epsilon)] - \epsilon\mathbf{E}\zeta^2(m_k^\epsilon)\Big\}_+ \\
&= \Sigma_k^\epsilon\,\mathbf{E}\left\{\frac{\zeta(m_k^\epsilon)}{\Sigma_k^\epsilon} + \frac{1}{\Sigma_k^\epsilon}\max_{m\in[m_k^\epsilon, m_{k+1}^\epsilon)}[\zeta(m) - \zeta(m_k^\epsilon)] - \epsilon\frac{\mathbf{E}\zeta^2(m_k^\epsilon)}{\Sigma_k^\epsilon}\right\}_+ \\
&\le K\Sigma_k^\epsilon \exp\left[-K\epsilon\frac{\mathbf{E}\zeta^2(m_k^\epsilon)}{\Sigma_k^\epsilon}\right] \le K\epsilon^{-1}2^{k/2}\exp\Big\{-K2^{k/2}\Big\}.
\end{aligned}$$

Substituting this equation in (3.17), we finish the proof. $\square$

### 3.2 Proof of Theorem 3

From the definition of the spectral cut-off estimator (see (1.5)) and (1.4) it follows that

$$\begin{aligned}
\mathbf{E}\|\hat{\beta}^{\bar{m}}(Y) - \beta\|^2 = \mathbf{E}\sum_{k=\bar{m}+1}^{p}\beta^2(k) + \sigma^2\mathbf{E}\sum_{k=1}^{\bar{m}}\frac{1}{\lambda(k)} \\
+\sigma^2\mathbf{E}\sum_{i=1}^{\bar{m}}\frac{\xi'^2(i)-1}{\lambda(i)}.
\end{aligned} \tag{3.19}$$

To bound from above the first term at the right-hand side of this equation, we combine Lemma 1, Jensen's inequality, and Lemma 2. So, we have

$$\begin{aligned}
\mathbf{E}\sum_{k=\bar{m}+1}^{p}\beta^2(k) \le& 2L^{-\gamma/(\alpha+\gamma)}P^{\alpha/(\alpha+\gamma)}\left[\mathbf{E}\sum_{k=\bar{m}+1}^{p}\beta^2(k)\lambda(k)\right]^{\gamma/(\alpha+\gamma)} \\
\le& 2L^{-\gamma/(\alpha+\gamma)}P^{\alpha/(\alpha+\gamma)}\left[\mathbf{E}r(\beta,\bar{m})\right]^{\gamma/(\alpha+\gamma)} \\
\le& 2L^{-\gamma/(\alpha+\gamma)}P^{\alpha/(\alpha+\gamma)}\left[\sqrt{r_\circ(\beta)} + K\sigma\right]^{2\gamma/(\alpha+\gamma)} \\
=& \frac{2\sigma^2}{L}\left[\frac{PL}{\sigma^2}\right]^{\alpha/(\alpha+\gamma)}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}} + K\right]^{2\gamma/(\alpha+\gamma)}.
\end{aligned} \tag{3.20}$$

The second term at the right-hand side in (3.19) may be bounded with the help of Lemmas 2 and 3. Since $\bar{m} \le r(\mathcal{W},\bar{m})/\sigma^2$, we obtain from (3.3) that for any $\mu > 0$

$$\mathbf{E}\exp\{\mu\sqrt{\bar{m}}\} \le \exp\Big\{\mu\sqrt{r_\circ(\beta)/\sigma^2} + K\mu + K\mu^2\Big\}. \tag{3.21}$$

With the help of this equation and (3.14), we get

$$\mathbf{E}\sum_{k=1}^{\bar{m}}\frac{1}{\lambda(k)} \le \frac{1}{(\alpha+1)L}\mathbf{E}\bar{m}^{\alpha+1} \le \frac{1}{(\alpha+1)L}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}+K\alpha\right]^{2(1+\alpha)}. \quad (3.22)$$

Similar arguments may be are used in controlling the last term at the right-hand side of (3.19). By Lemma 5 we get that for any $\epsilon \in (0, 1/4)$

$$\mathbf{E}\sum_{i=1}^{\bar{m}}\frac{\xi^2(i)-1}{\lambda(i)} \le \epsilon\mathbf{E}\sum_{i=1}^{\bar{m}}\frac{1}{\lambda^2(i)}+\frac{K}{\epsilon}.$$

Next, minimizing the right-hand side at this equation in $\epsilon \in (0, 1/4)$, we obtain with the help of (3.21), Lemma 3, and a simple algebra

$$\begin{aligned}\mathbf{E}\sum_{i=1}^{\bar{m}}\frac{\xi^2(i)-1}{\lambda(i)} &\le K\left[\mathbf{E}\sum_{k=1}^{\bar{m}}\frac{1}{\lambda^2(k)}\right]^{1/2}+K \le \frac{K}{L\sqrt{2\alpha+1}}\left[\mathbf{E}\bar{m}^{2\alpha+1}\right]^{1/2}+K \\ &\le \frac{K}{L\sqrt{2\alpha+1}}\left[\sqrt{\frac{r_\circ(\beta)}{\sigma^2}}+K\alpha\right]^{2\alpha+1}+K.\end{aligned}$$

Finally, substituting this equation, (3.20), and (3.22) in (3.19), we finish the proof. □

# 4 Acknowledgments

This work is partially supported by Laboratory for Structural Methods of Data Analysis in Predictive Modeling, MIPT, RF Government grant, ag. 11.G34.31.0073; and RFBR research projects 13-01-12447 and 13-07-12111.

# References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Intern. Symp. Inf. Theory* 267–281.

[2] Donoho, D. L. (1995). Nonlinear solutions of linear inverse problems by wavelet-vaguelette decomposition. *J. of Appl. and Comp. Harmonic Anal.* **2** 101–126.

[3] Donoho, D. L. and Johnstone, I. M. (1996). Neoclassical minimax problems, thresholding and adaptive function estimation. *Bernoulli* **2**, 39–62.

[4] Cavalier, L. and Golubev Yu. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. of Statist.* Vol. 34 **4** 1653–1677.

[5] Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*. Mathematics and its Applications **375**. Kluwer Academic Publishers Group, Dordrecht.

[6] Ermakov M. S. (1989). Minimax estimation of the solution of an ill-posed convolution type problem. *Problems of Inform. Transmission* **25** 191–200.

[7] Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized crossvalidation as a method to choosing a good ridge parameter. *Technometrics* **21** 215-223.

[8] Golub, G. H. and von Matt, U. (1997). Generalized crossvalidation for large scale problems. *In Recent advances in total least squares techniques and errors in variable modeling, S. van Huffel ed.* SIAM, 139–148.

[9] Golubev, Yu. (2010). On universal oracle inequalities related to high dimensional linear models. *Ann. of Statist.* **38** 2751–2780.

[10] Golubev, Yu. (2012). Exponential weighting and oracle inequalities for projection estimates. *Problems of Inform. Trans.*, **48** 269-280.

[11] Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22** 835–866.

[12] Mair, B. and Ruymgaart, F. H. (1996). Statistical estimation in Hilbert scale. *SIAM J. Appl. Math.* **56** 1424–1444.

[13] Mallows, C. L. (1973). Some comments on $C_p$. *Technometrics* **15** 661–675.

[14] Koo, J. Y. (1993) Optimal rates of convergence for nonparametric statistical inverse problems. *Ann. of Statist.* **21** 590-599.

[15] Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Inform. Transmission*, **16**, 120-133.

[16] Tikhonov, A. N. and Arsenin, V. A. (1977). *Solution of Ill-posed Problems*. Translated from the Russian. Preface by translation editor Fritz John. Scripta Series in Mathematics. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York.

[17] Van der Vaart, A. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer-Verlag, New York.