IDEAL DENOISING FOR SIGNALS IN SUB-GAUSSIAN NOISE

SEBASTIAN E. FERRANDO[1] AND RANDALL PYKE[2]

1. Department of Mathematics, Ryerson University, Toronto. ferrando@ryerson.ca. Correspondence should be sent to this author at: Department of Mathematics, Ryerson University, 350 Victoria St., Toronto, ON, M5B 2K3, Canada.
2. Department of Mathematics, Simon Fraser University, Vancouver, BC, Canada. rpyke@sfu.ca

Date: September 27, 2006.

Abstract. Donoho and Johnstone introduced an algorithm and supporting inequality that allows the selection of an orthonormal basis for optimal denoising. The present paper concentrates in extending and improving this result, the main contribution is to incorporate a wider class of noise vectors. The class of strict sub-Gaussian random vectors allow us to obtain large deviation inequalities in a uniform way over all basis in a given library. The results are obtained maintaining the algorithmic properties of the original results.

Key words: Signal Denoising. L^2 spaces. Adaptive Basis Selection. Oracles for Adaptation. Thresholding of Wavelet Coefficients. Strict sub-Gaussian Noise.
Subject classification: 42C40, 93E11, 94A08

1. Introduction

The subject of this paper is oracle based denoising. Let s be a signal embedded in noise, we are interested in estimators $\hat{s}$, for the signal s, obtained by thresholding coefficients in an orthonormal basis $\mathcal{B}$ of $\mathbb{R}^n$. We consider the problem of optimal basis selection when there is available a library $\mathcal{L}$ of such bases from which to choose from. We will look for estimators which satisfy the following oracle-type inequality with high probability

$$||\hat{s} - s||_2^2 \leq \; c \min_{\mathcal{B} \in \mathcal{L}} \mathcal{R}(s, \mathcal{B}). \tag{1}$$

$\mathcal{R}(s, \mathcal{B})$ is the oracle risk for the basis $\mathcal{B}$, this last quantity is the average quadratic error incurred by an oracle estimator (see (2)). This last estimator makes use of knowledge of s and is of excellent quality but unavailable in practice. After proper re-scaling, it can be argued that an inequality of the above type is asymptotically optimal (in the number of samples) as the oracle risk decays in a best possible manner. Therefore, this type of inequality gives an apriori measure for the quality of the algorithm associated to estimators satisfying (1). We refer the reader to the bibliography (for example [1] and [3]) for background information.

In [1] Donoho and Johnstone introduced an algorithm that allows the selection of an orthonormal basis from a library of such bases. Their result concentrates on proving an inequality like the one described above. Their assumption on the noise vector is that its coordinates are independent identically distributed (i.i.d.) Gaussian random variables. The technique employed in [1] uses a general concentration inequality which they borrow from [6]. Other accounts of these results as well as improvements can be found in [3] and [5].

The main point of the present article is to extend the results of Donoho and Johnstone to a wider class of noise vectors. We borrow the noise set-up and related background results from [4]. We generalize the Gaussian hypothesis and only require the noise vector to satisfy a strict sub-Gaussian hypothesis (see Theorem 3). This set-up generalizes the main result from [1]; Theorem 2 characterizes sub-Gaussian random variables and gives an indication for the wider scope of our theorems (see Definition 6 for the precise set-up). For example, noise coordinates with the uniform distribution are included in our setting. Our results have been carefully crafted so that the algorithmic content of the original results have been preserved, in particular the thresholding parameters used are the ones used in the Gaussian case. We also take the opportunity to improve on the value of some of the key parameters appearing in the main inequality in [1]. Our proof follows the one in [1] but uses a classical argument to derive a more specific, relative to our noise vector, concentration inequality.

Let us comment on the essence of our approach. As a consequence of the way that risk is defined, it is easy to see that the oracle risk appearing in (1) uses very little knowledge about the noise distribution; it actually only uses the variances of the noise coordinates. In fact our results show that the key aspect of the noise distribution is the existence of an exponential second moment inequality. This result only depends on the tail decay of the Gaussian distribution. The effect of a change of coordinates has also to be considered. These observations lead us into the wide class of strict sub-Gaussian random vectors as a natural class were results of the type (1) can be proven.

The paper is organized as follows, Section 2 summarizes the main results from [1]. Section 3 describes the set-up of strict sub-Gaussian noise vectors. Our main result, Theorem 3, is then proved in Section 4. Section 5 briefly discusses some technical issues. Appendix A states some of the properties of sub-Gaussian and strict sub-Gaussian noise vectors that we require. Finally, Appendix B proves, for the reader's convenience, some intermediate results needed along the way.

2. Summary of Known Results

This section summarizes the main result from [1] and the associated algorithm. First we introduce some notation.
If $\mathcal{B} = \{e_1, \ldots, e_n\}$ is an orthonormal basis for $\mathbb{R}^n$, then for any vector $v \in \mathbb{R}^n$, $v_k[\mathcal{B}]$ denotes the k^{th} coordinate of v in the basis $\mathcal{B}$; $v_k[\mathcal{B}] = \langle v, e_k \rangle$. Here, $\langle v, u \rangle$ is the standard (Euclidean) inner product on $\mathbb{R}^n$. In particular, $\langle u, v \rangle = \sum_{k=1}^n u_k[\mathcal{B}] v_k[\mathcal{B}]$ for any orthonormal basis $\mathcal{B}$. $\mathcal{U} = \{u_k\}_{k=1}^n$ denotes the standard orthonormal basis of $\mathbb{R}^n$ and $\|v\|^2 = \langle v, v \rangle$.
The data is given in the form $y = s + z$ where $s, z \in \mathbb{R}^n$, s is the deterministic signal and z a noise vector whose coordinates $z_i[\mathcal{U}]$ are assumed to be i.i.d. Gaussian white noise. The common variance of the coordinates will be denoted by $\sigma^2 = \mathbb{E}(z_i^2[\mathcal{U}])$. Let $\mathcal{L}$ be a library of orthonormal bases of $\mathbb{R}^n$ and $\mathcal{M}_n$ the set of distinct vectors in $\mathcal{L}$. M_n will denote the cardinality of $\mathcal{M}_n$, i.e., M_n is the total number of distinct vectors occurring among all the bases in the library. Let $y[\mathcal{B}]$ be the original data transformed into the basis $\mathcal{B}$.

2.1. Oracle in basis $\mathcal{B}$.

Let w, θ, ζ be the coordinate vectors of y, s, z, respectively, in some basis $\mathcal{B}$. The only probabilistic hypothesis needed in the computation that follows is the assumption that $\mathbb{E}(\zeta_i) = 0$. Let $\hat{\theta}$ be the oracle estimate for θ;

$$\hat{\theta}_i = \delta_i \; w_i, \quad \delta_i = \delta_i(\theta_i) \in \{0, 1\}.$$

The Oracle risk in basis $\mathcal{B}$, $\mathcal{R}(s, \mathcal{B})$, is given by

$$\mathcal{R}(s, \mathcal{B}) \equiv \min_{\delta_i} \mathbb{E}(\|\hat{\theta} - \theta\|^2) = \sum_{i=1}^n \min(s_i^2[\mathcal{B}], \sigma_i^2), \tag{2}$$

this equality is easy to prove and well known. It will be convenient to introduce the best risk in the library $\mathcal{L}$;

$$\mathcal{R}^\star(s, \mathcal{L}) = \min_{\mathcal{B} \in \mathcal{L}} \mathcal{R}(s, \mathcal{B}). \tag{3}$$

These quantities depend on knowledge of θ (hence the name *oracle*) and are materially unavailable for denoising purposes.
Choose $\lambda' > 8$ and set

$$\Lambda'_n = \Lambda'_n(\lambda') = (\lambda' \; \sigma \; (1 + t_n))^2, \text{ where } t_n = \sqrt{2 \; \log M_n}. \tag{4}$$

Remark 1. *The " $'$" in λ' and Λ'_n was not used originally in* [1]*; we use it to differentiate their values from our values of λ and Λ_n used here. Similar remarks also apply to other symbols which use " $'$" below.*

We now describe a procedure to obtain an optimal best basis estimate. Define the entropy functional

$$\mathcal{E}'_\lambda(y, \mathcal{B}) = \sum_i \min(y_i^2[\mathcal{B}], \Lambda'_n).$$

Let $\hat{\mathcal{B}}'$ be the best orthogonal basis according to this entropy;

$$\hat{\mathcal{B}}' = \arg\ \min_{\mathcal{B}\in\mathcal{L}} \mathcal{E}_\lambda(y, \mathcal{B}).$$

For given threshold t define the thresholding function, acting on a real number c, by $\eta_t(c) = c\ \mathbf{1}_{\{|c|>t\}}$ where $\mathbf{1}_{\{|c|>t\}}$ is the characteristic function of the set $(-\infty, -t) \cup (t, \infty)$. Given the basis $\hat{\mathcal{B}}'$, we then apply hard thresholding to obtain the empirical best estimate $\hat{s}'$;

$$\hat{s}'_i[\hat{\mathcal{B}}'] = \eta_{\sqrt{\Lambda'_n}}(y_i[\hat{\mathcal{B}}']).$$

A measure of the quality of the above procedure, which approximates s by means of $\hat{s}'$, is given by the following theorem from [1].

Theorem 1. *Given data $y = s + z$ as described above, then, with probability exceeding $\pi_n = 1 - e/M_n$:*

$$||\hat{s}' - s||_2^2 \leq \frac{\lambda'\ \Lambda'_n}{\sigma^2\ (\lambda' - 8)} \min_{\mathcal{B}\in\mathcal{L}} \mathbb{E}(||\hat{s}'_{\mathcal{B}} - s||^2) = \frac{\lambda'^3\ (1 + t_n)^2}{(\lambda' - 8)} \mathcal{R}^\star(s, \mathcal{L}). \tag{5}$$

The indicated *minimum* is over all ideal hard thresholding estimates working in all bases of the library, i.e. in basis $\mathcal{B}$ the coordinates of $\hat{s}'_{\mathcal{B}}$ are just given by $y_i[\mathcal{B}]\ \mathbf{1}_{\{|s_i[\mathcal{B}|>\sigma\}}$. We provide more details in Appendix B.
The relevance of having the magnitude $\min_{\mathcal{B}\in\mathcal{L}} \mathbb{E}(||\hat{s}'_{\mathcal{B}} - s||^2)$ as an upper bound to the L^2 error of the approximation, as in (5), is discussed in [1]. For the case when $\mathcal{L}$ consists of a single orthonormal basis, Theorem 1 specializes, up to a constant factor, to a main result in [2].

3. Strict Sub-Gaussian Noise

Here we introduce the definitions needed to characterize our hypothesis on the noise vector z.

Definition 1. A random variable ξ is called *sub-Gaussian* if there exists a number $a \in [0, \infty)$ such that the inequality

$$\mathbb{E}\Big(\exp\Big(\lambda\,\xi\Big)\Big) \leq \exp\left(\frac{a^2\ \lambda^2}{2}\right)$$

holds for all $\lambda \in \mathbb{R}^1$. The class of all sub-Gaussian random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$ is denoted by $Sub(\Omega)$.

Introduce the notation

$$\tau(\xi) = \inf\{a \geq 0 : \mathbb{E}\left(\exp\left(\lambda\,\xi\right)\right) \leq \exp\left(\frac{a^2\ \lambda^2}{2}\right)\ \lambda \in \mathbb{R}^1\},$$

$\tau(\xi)$ is called the *sub-Gaussian standard of the random variable* ξ. We say that the sub-Gaussian random variable ξ is *standardized to one* if $\tau(\xi) = 1$. It can be seen that $Sub(\Omega)$ is a Banach space under the norm $\tau(\cdot)$ and that (see [4])

$$\mathbb{E}(\xi^2) \leq \tau^2(\xi). \tag{6}$$

Definition 2. A random variable η majorizes a random variable ξ in distribution if there exists $x_0 \geq 0$ such that

$$P(|\xi| \geq x) \leq P(|\eta| \geq x) \text{ for all } x > x_0.$$

The following theorem characterizes sub-Gaussian random variables.

Theorem 2. *A random variable ξ is sub-Gaussian if and only if it has mean zero and ξ is majorized in distribution by a zero-mean Gaussian random variable.*

Definition 3. A random vector $\xi \in \mathbb{R}^n$ is called *sub-Gaussian* if there exists a symmetric nonnegative definite operator $B : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\mathbb{E}\Big(\exp\langle u, \xi\rangle\Big) \leq \exp\left(\frac{1}{2}\langle Bu, u\rangle\right) \tag{7}$$

holds for all $u \in \mathbb{R}^n$. We call the operator B the *companion operator* for ξ. The class of all sub-Gaussian random vectors is denoted by $Sub(\Omega, \mathbb{R}^n)$. Notice that $Sub(\Omega, \mathbb{R}) = Sub(\Omega)$.

Lemma 1. *Let $\xi \in \mathbb{R}^n$ be a random vector and $\mathcal{B}$ be an orthonormal basis. If $\xi_k[\mathcal{B}] \in Sub(\Omega)$ $k = 1, \ldots, n$, then $\xi \in Sub(\Omega, \mathbb{R}^n)$.*

Notice that Lemma 1 reduces the problem of constructing elements of $Sub(\Omega, \mathbb{R}^n)$ to the problem of constructing elements of $Sub(\Omega)$. See Appendix A for explicit examples and [4] for detailed information on $Sub(\Omega, \mathbb{R}^n)$.
In some sense the assumption $z \in Sub(\Omega, \mathbb{R}^n)$ is the most natural for the noise vector, nonetheless we use stronger hypothesis as introduced below. The reasons for requiring stronger hypothesis are explained in Section 5.

Definition 4. A sub-Gaussian random variable ξ is called *strictly sub-Gaussian* if $\sigma^2 \equiv \mathbb{E}(\xi^2) = \tau^2(\xi)$. The class of all strictly sub-Gaussian variables is denoted by $SSub(\Omega)$.

Definition 5. A random vector ξ is called *strictly sub-Gaussian* if

$$\mathbb{E}\Big(\exp\langle\, u, \xi\rangle\Big) \leq \exp\left(\frac{1}{2}\langle Bu, u\rangle\right)$$

holds for all $u \in \mathbb{R}^n$ where B is the covariance operator of ξ, namely

$$\langle Bu, v\rangle = \mathbb{E}\Big(\langle \xi, u\rangle\langle \xi, v\rangle\Big) \quad \forall u, v \in \mathbb{R}^n.$$

The class of all strictly sub-Gaussian random vectors is denoted by $SSub(\Omega, \mathbb{R}^n)$. Note that $SSub(\Omega, \mathbb{R}) = SSub(\Omega)$ and, of course, $SSub(\Omega, \mathbb{R}^n) \subseteq Sub(\Omega, \mathbb{R}^n)$.
Lemma 3 in Appendix A shows under what conditions $SSub(\Omega)$ is closed under linear combinations. Lemma 5 in Appendix A reduces the problem of constructing elements of $SSub(\Omega, \mathbb{R}^n)$ to the problem of constructing independent elements of $SSub(\Omega)$. For these reasons, it is enough to provide examples of elements in $SSub(\Omega)$, this is done in Appendix A.

As indicated before, $\mathcal{U} = \{u_k\}_{k=1}^n$ denotes the standard basis of $\mathbb{R}^n$. We assume the data is of the form $y = s + z$ where $s \in \mathbb{R}^n$ is the signal and z is the noise which is assumed to satisfy the following assumption.

Definition 6. We say that the vector valued random variable z (i.e. the noise vector) is *strict sub-Gaussian white noise* if it satisfies:

$z \in SSub(\Omega, \mathbb{R}^n)$, the random variables $z_k[\mathcal{U}]$ are uncorrelated, and $\mathbb{E}(z_k^2[\mathcal{U}]) = \sigma^2\ \forall k$.

Remark 2. *All our results remain valid if $\mathcal{U}$, in Definition 6, is replaced by any other orthonormal basis. It follows from Lemma 5 that a sufficient condition for z to be strict sub-Gaussian white noise is given by:*

$z_k[\mathcal{U}] \in SSub(\Omega, \mathbb{R})$, *the random variables* $z_k[\mathcal{U}]$ *are jointly independent, and*

$$\mathbb{E}(z_k[\mathcal{U}]^2) = \sigma^2.$$

4. Library of Bases and Strict Sub-Gaussian Noise

Set

$$\delta_n(\lambda) \equiv 4\ \lambda\ \left(1 + (2+3\beta)\ \log M_n\right).$$

We have left β unspecified as this is convenient for some developments and comparisons. For the sake of simplicity the reader can set $\beta = 1$. We will use the following threshold

$$\Lambda_n = \Lambda_n(\lambda) = \sigma^2 \delta_n(\lambda),\ \ \lambda > 2. \tag{8}$$

With this new parameter we repeat the notions from Section 2, first introduce the empirical entropy as follows

$$\mathcal{E}_\lambda(y, \mathcal{B}) = \sum_i \min(y_i^2[\mathcal{B}], \Lambda_n).$$

Let $\hat{\mathcal{B}}$ be the best orthogonal basis according to this entropy;

$$\hat{\mathcal{B}} = \arg\ \min_{\mathcal{B}\in\mathcal{L}} \mathcal{E}_\lambda(y, \mathcal{B}).$$

Apply hard thresholding with $\eta_t(y) = y\mathbf{1}_{\{|y|>t\}}$ to obtain the empirical best estimate $\hat{s}^\star$;

$$\hat{s}_i^\star[\hat{\mathcal{B}}] = \eta_{\sqrt{\Lambda_n}}(y_i[\hat{\mathcal{B}}]).$$

For the present set-up, in which Λ_n is direction independent, the *complexity functional* from (33) can be expressed as follows;

$$K(\tilde{s}, s) = \|\tilde{s} - s\|_2^2 + \Lambda_n \min_{\mathcal{B}\in\mathcal{L}} \sum_{\{i,\ \tilde{s}_i[\mathcal{B}]\neq 0\}} 1\ = \|\tilde{s} - s\|_2^2 +\ \Lambda_n\ N_{\mathcal{L}}(\tilde{s})$$

where

$$N_{\mathcal{L}}(\tilde{s}) = \min_{\mathcal{B}} \#\{e_i \in \mathcal{B}\ :\ \tilde{s}_i[\mathcal{B}] \neq 0\}.$$

Let s^o denote a signal of minimum theoretical complexity;

$$K(s^o, s) =\ \min_{\tilde{s}} K(\tilde{s}, s). \tag{9}$$

Several key properties of the above notions, and their proofs, are given in Appendix B.

In the next theorem we will make use the following notation

$$j_0 \equiv \max(N_{\mathcal{L}}(s^o), 1).$$

Remark 3. *In Theorem 3 we will assume* $(1+(2+3\beta)\ \log M_n) \leq M_n^\beta$ *which holds for* $M_n \geq 15$ *when* $\beta = 1$*, a condition that will hold for all practical purposes. If this condition is not assumed, the only effect on our result will be to replace the above* $\delta_n(\lambda)$ *by* $\delta_n(\lambda) = 6.44\ \lambda\ (1 + 2(1+\beta)\ \log M_n)$*.*

Here is our main result.

Theorem 3. *Given data $y = s + z$, we assume z is strict sub-Gaussian white noise (as in Definition 6) and $(1 + (2 + 3\beta)\ \log M_n) \leq M_n^{\beta}$. Then, after setting $\beta = 1$, with probability exceeding $\pi_n = 1 - e/M_n^{j_0}$:*

$$||\hat{s}^{\star} - s||_2^2 \leq \frac{2\ \lambda}{(\lambda - 2)} \min_{\mathcal{B}} \sum_i \min(s_i^2[\mathcal{B}], \Lambda_n) \tag{10}$$

$$\leq \frac{2\ \lambda\ \delta_n(\lambda)}{(\lambda - 2)}\ \mathcal{R}^{\star}(s, \mathcal{L}). \tag{11}$$

Proof. The first moves in the proof are intended to improve the values of some key parameters. Define

$$k = \frac{\lambda}{4}\Lambda_n \Big(N_{\mathcal{L}}(\hat{s}^{\star}) + N_{\mathcal{L}}(s^o) \Big), \tag{12}$$

$$\hat{k}^{\star} = ||\hat{s}^{\star} - s||_2^2 + k, \tag{13}$$

$$k^o = ||s^o - s||_2^2 + k. \tag{14}$$

Let B_1^o and $\widehat{B}_1^{\star}$ be the bases that are realized in $N_{\mathcal{L}}(s^o)$ and $N_{\mathcal{L}}(\hat{s}^{\star})$ respectively, and $B^o \subset B_1^o$ and $\widehat{B}^{\star} \subset \widehat{B}_1^{\star}$ be those basis vectors in B_1^o and $\widehat{B}_1^{\star}$ with which s^o and $\hat{s}^{\star}$ have nonzero inner products (i.e., the vectors corresponding to the nonzero coordinates of $s^o[B^o]$ and $\hat{s}^{\star}[\widehat{B}^{\star}]$). Let S be the subspace spanned by the vectors in $B^o \cup \widehat{B}^{\star}$. Let P_S be orthogonal projection onto S. Thus,

$$\dim S \leq N_{\mathcal{L}}(s^o) + N_{\mathcal{L}}(\hat{s}^{\star}).$$

From (39), in Appendix B, we have

$$K(\hat{s}^{\star}, s) \leq K(s^o, s) + 2\langle z, \hat{s}^{\star} - s^o \rangle. \tag{15}$$

We also have the upper bound

$$\begin{aligned} 2\langle z\, ,\, \hat{s}^{\star} - s^o \rangle &= 2\langle P_S z\, ,\, \hat{s}^{\star} - s^o \rangle \\ &= 2\langle P_S z\, ,\, \hat{s}^{\star} - s \rangle + 2\langle P_S z\, ,\, s - s^o \rangle \\ &\leq 2\|P_S z\| \sqrt{\hat{k}^{\star}} + 2\|P_S z\| \sqrt{k^o}. \end{aligned} \tag{16}$$

Consider the events

$$A_1 = \{\omega\ :\ \|P_S z(\omega)\|_2 \leq \frac{\sqrt{\hat{k}^{\star}}}{\lambda} \text{ and } \|P_S z(\omega)\|_2 \leq \frac{\sqrt{k^o}}{\lambda}\}$$

and

$$A_2 = \{\omega\ :\ N_{\mathcal{L}}(\hat{s}^{\star}) + N_{\mathcal{L}}(s^o) = 0\}.$$

We will prove (10) for any $\omega \in A \equiv A_1 \cup A_2$. To this end it is enough to prove that

$$||\hat{s}^{\star} - s||_2^2 \leq \frac{2K(s^o, s)}{(1 - \frac{2}{\lambda})}, \qquad \lambda > 2 \text{ for all } \omega \in A. \tag{17}$$

The fact that (17) implies (10) follows from

$$\begin{aligned} K(s^o, s) = \min_{\tilde{s}} K(\tilde{s}, s) &= \min_{\mathcal{B}} \sum_i \min(s_i^2[\mathcal{B}], \Lambda_n) \\ &\leq \delta_n(\lambda) \cdot \min_{\mathcal{B}} \sum_i \min(s_i^2[\mathcal{B}], \sigma^2) \\ &= \delta_n(\lambda) \cdot \mathcal{R}^{\star}(s, \mathcal{L}), \end{aligned} \tag{18}$$

for a proof of (18) see Appendix B.

Next we proceed to prove (17). Consider first $w \in A_1$,

$$\begin{aligned} 2\|P_S z\|_2\sqrt{\hat{k}^\star} + 2\|P_S z\|_2\sqrt{k^o} \;&\leq\; 2\frac{\hat{k}^\star}{\lambda} + 2\frac{k^o}{\lambda} \\ &= \frac{2}{\lambda}\|\hat{s}^\star - s\|_2^2 + \frac{2}{\lambda}\|s^o - s\|_2^2 + \Lambda_n\Big(N_{\mathcal{L}}(\hat{s}^\star) + N_{\mathcal{L}}(s^o)\Big). \end{aligned} \tag{19}$$

Using this and (15) we obtain

$$\begin{aligned} &\|\hat{s}^\star - s\|_2^2 + \Lambda_n N_{\mathcal{L}}(\hat{s}^\star) \;\leq \\ &\qquad \|s^o - s\|_2^2 + \Lambda_n N_{\mathcal{L}}(s^o) + \frac{2}{\lambda}\|\hat{s}^\star - s\|_2^2 + \frac{2}{\lambda}\|s^o - s\|_2^2 + \Lambda_n N_{\mathcal{L}}(\hat{s}^\star) + \Lambda_n N_{\mathcal{L}}(s^o), \end{aligned} \tag{20}$$

which reduces to

$$\|\hat{s}^\star - s\|_2^2 \;\leq\; \frac{2K(s^o, s)}{(1 - \frac{2}{\lambda})}, \qquad \lambda > 2.$$

Consider now $w \in A_2$, clearly then $s^o = \hat{s}^\star = 0$, therefore

$$\|\hat{s}^\star - s\|_2^2 \;= \|s\|_2^2 \;= K_{\mathcal{B}}(s^o, s) \leq \frac{2\ K_{\mathcal{B}}(s^o, s)}{(1 - \frac{2}{\lambda})},$$

where we used $\lambda > 2$. Therefore, (17) has been proven for all $w \in A$.

In order to complete the proof, it remains to obtain an upper bound for $P(A^c)$. Let $C(j, M_n)$ denote the collection of all subsets consisting of j vectors chosen from the M_n vectors of $\mathcal{M}_n$. By an abuse of notation we will write $\widehat{S} \in C(j, M_n)$ to mean the subspace spanned by an element of $C(j, M_n)$,

$$A^c \;\subseteq\; \big\{\omega \;:\; \|P_S z(\omega)\|_2 \geq \frac{\sqrt{\Lambda_n}\sqrt{N_{\mathcal{L}}(s^o) + N_{\mathcal{L}}(\hat{s}^\star)}}{2\sqrt{\lambda}}\big\} \subseteq \bigcup_{j=j_0}^{M_n} B_j$$

where

$$B_j \;\equiv\; \big\{\omega \;:\; \sup_{\widehat{S} \in C(j, M_n)} \|P_{\widehat{S}} z(\omega)\|_2 \geq \frac{\sqrt{\Lambda_n}\sqrt{j}}{2\sqrt{\lambda}}\;\big\}.$$

Let

$$a = \frac{\sqrt{\Lambda_n}\sqrt{j}}{2\sqrt{\lambda}} \tag{21}$$

and for a fixed subspace $\widehat{S}_1 \in C(j, M_n)$ of dimension d, with $d \leq j$, define

$$D_j \;=\; \big\{\omega \;:\; \|P_{\widehat{S}_1} z(\omega)\|_2 \geq a\,\big\}.$$

We will obtain the following bound

$$P(D_j) \leq M_n^{-j} M_n^{-j_0\beta}. \tag{22}$$

With this common bound, i.e. independent of the particular $\widehat{S}_1$, we have

$$P(B_j) \leq \#C(j, M_n) M_n^{-j} M_n^{-j\beta} = \left(\begin{array}{c} M_n \\ j \end{array}\right) M_n^{-j} M_n^{-j_0\beta}.$$

Therefore,

$$P(A^c) \leq \sum_{j=j_0}^{M_n} P(B_j) \leq \sum_{j=j_0}^{M_n} \left(\begin{array}{c} M_n \\ j \end{array}\right) M_n^{-j} M_n^{-j_0\beta} \leq M_n^{-j_0\beta} \sum_{j=1}^{M_n} \frac{1}{j!} \leq \frac{e}{M_n^{-j_0\beta}},$$

which will conclude the proof upon setting $\beta = 1$. Thus, remains to prove the bound on (22). It is here where the hypothesis on the noise vector are put to use. Let $\{e_1, \dots, e_d\}$ be an orthonormal basis of $\widehat{S}_1$. Extend $\{e_1, \dots, e_d\}$ to an orthonormal basis $\mathcal{E} = \{e_1, \dots, e_d, e_{d+1}, \dots, e_n\}$ of $\mathbb{R}^n$. Then

$$z = \sum_{k=1}^{n} \langle z, e_k \rangle \, e_k, \qquad \text{and} \qquad \xi = P_{\widehat{S}_1} z = \sum_{k=1}^{d} \langle z, e_k \rangle \, e_k. \tag{23}$$

Evidently,

$$\begin{array}{rcll} \xi_k[\mathcal{E}] & = & z_k[\mathcal{E}], & k = 1, \dots, d, \\ \xi_k[\mathcal{E}] & = & 0, & k = d+1, \dots, n. \end{array}$$

Let $\mathcal{U} = \{u_1, \dots, u_n\}$ be the standard basis of $\mathbb{R}^n$. Then $z = \sum_{k=1}^{n} \langle z, u_k \rangle \, u_k$ and

$$\xi_k[\mathcal{E}] = \langle z, e_k \rangle = \sum_{j=1}^{n} z_j[\mathcal{U}] \langle u_j, e_k \rangle, \;\; k = 1, \dots, d.$$

Since $P_{\widehat{S}_1}$ is a linear operator, by Lemma 4 from Appendix A, $\xi \in SSub(\Omega, \mathbb{R}^d)$. Let $R = [P_{\widehat{S}_1}]_{\mathcal{U}}^{\mathcal{E}}$ be the matrix representation of $P_{\widehat{S}_1}$ acting from the basis $\mathcal{U}$ to the basis $\mathcal{E}$. Then $[R]_{ij} = \langle u_j, e_i \rangle \equiv r_{ij}$. Note that $\sum_{j=1}^{n} r_{ij}^2 = \|e_i\|_2^2 = 1$, $\forall i = 1, \dots, n$. Consider $k = 1, \dots, d$,

$$\begin{array}{rcll} \tau^2(\xi_k[\mathcal{E}]) & = & \mathbb{E}(\xi_k^2[\mathcal{E}]), & \xi_k[\mathcal{E}] \in SSub(\Omega) \\ & = & \mathbb{E}\Big((\sum_{j=1}^{n} r_{kj} \; z_j[\mathcal{U}])^2 \Big) & \\ & = & \sum_{j=1}^{n} r_{kj}^2 \; \mathbb{E}(z_j^2[\mathcal{U}]), & z_j[\mathcal{U}] \text{ are uncorrelated} \\ & = & \sum_{j=1}^{n} r_{kj}^2 \; \tau^2(z_j[\mathcal{U}]), & z_j[\mathcal{U}] \in SSub(\Omega) \\ & = & \sum_{j=1}^{n} \sigma_j^2 \; r_{kj}^2, & \tau^2(z_j[\mathcal{U}]) = \mathbb{E}(z_j^2[\mathcal{U}]) = \sigma_j^2 \end{array} \tag{24}$$

Under the hypothesis of the present theorem, $\tau^2(z_j[\mathcal{U}]) = \sigma_j^2 = \sigma^2, \; j = 1, \dots, n$.

Let Q be the covariance operator of ξ and q_{ij} be the matrix entries of $[Q]_{\mathcal{E}}^{\mathcal{E}}$. Since $\xi_k[\mathcal{E}] = 0$ for $k = d+1, \dots, n$, $b_{ij} = 0$ if either i or j is greater than d. For $i, j \in \{1, \dots, d\}$,

$$q_{ij} = \mathbb{E}\Big(\xi_i[\mathcal{E}]\xi_j[\mathcal{E}]\Big) = \mathbb{E}\Big(z_i[\mathcal{E}]z_j[\mathcal{E}]\Big) = \mathbb{E}\Big(\langle z, e_i\rangle\langle z, e_j\rangle\Big) =$$

$$\mathbb{E}\Big(\langle(\sum_{k=1}^{n} z_k[\mathcal{U}]u_k, e_i\rangle)\langle\sum_{k'=1}^{n} z_{k'}[\mathcal{U}]u_{k'}, e_j\rangle\Big) = \mathbb{E}\Big(\sum_{k=1}^{n} z_k[\mathcal{U}]\langle u_k, e_i\rangle \sum_{k'=1}^{n} z_{k'}[\mathcal{U}]\langle u_{k'}, e_j\rangle\Big) =$$

$$\sum_{k,k'} \mathbb{E}\Big(z_k[\mathcal{U}]z_{k'}[\mathcal{U}]\langle u_k, e_i\rangle\langle u_{k'}, e_j\rangle\Big) = \sum_{k=1}^{n} \sigma_k^2 \; r_{ki} \; r_{kj},$$

where we have used the fact that the $z_k[\mathcal{U}]$ are uncorrelated; $\mathbb{E}(z_k[\mathcal{U}]z_{k'}[\mathcal{U}]) = 0,\ k \neq k'$. In the present white noise case $\tau^2(z_j[\mathcal{U}]) = \sigma_j^2 = \sigma^2,\ j = 1, \ldots, n$. We use the fact that $\sum_1^n\ r_{ki}\ r_{kj} = \langle e_i,\, e_j\rangle = \delta_{ij}$ to obtain

$$q_{kk} \ = \ \sigma^2 \ = \ \tau^2(\xi_k[\mathcal{E}]),\ k = 1, \ldots, d \tag{25}$$

$$q_{ij} \ = \ 0, \quad i \neq j; \tag{26}$$

and so Q is diagonal. The above considerations will allow us to apply Theorem 4 in Appendix A. For convenience introduce the notation $Y = ||P_{\widehat{S}_1} z||_2$. With a as above, using Markov's inequality we obtain;

$$P(Y \geq a) = P\Big(\exp(\theta Y^2) \geq \exp(\theta a^2)\Big) \leq \exp(-\theta a^2)\ \mathbb{E}\Big(\exp(\theta Y^2)\Big), \tag{27}$$

where $\theta > 0$ is an arbitrary parameter.
Consider (29) in Appendix A with $r > 2\ \sigma^2$ and take $\theta = 1/r$. Let $\rho \equiv 2\sigma^2/r$ so $0 < \rho < 1$. Using $q_{k,k} = 0$ for $k > d$ and (27) we obtain
(28)

$$P(Y \geq a) \leq \exp(-a^2/r)\ \mathbb{E}\Big(\exp(1/r\sum_{k=1}^{n}\xi_k^2)\Big) = \exp(-a^2/r)\ \mathbb{E}\Big(\exp(1/r\sum_{k=1}^{d}\xi_k^2)\Big) \leq$$

$$\exp(-a^2/r)\prod_{k=1}^{d}\Big(1 - \frac{2\sigma^2}{r}\Big)^{-1/2} \leq \exp(-a^2\rho/2\sigma^2)\ (1-\rho)^{-d/2} \leq$$

$$\exp(-a^2\rho/2\sigma^2)\ (1-\rho)^{-j/2}$$

Notice that in the last inequality we have used the fact that $d \leq j$ and $\rho < 1$. The function $G(\rho) \equiv \exp(-a^2\rho/2\sigma^2)\ (1-\rho)^{-j/2}$ is minimized at

$$\rho' = 1 - \frac{j\ \sigma^2}{a^2},$$

Here, $a^2 = \Lambda_n j/4\lambda > j\sigma^2$ and so $\rho' \in (0, 1)$. Evaluating $G(\rho)$ at $\rho = \rho'$ and using equations (8), (21) and (28) we obtain (notice that $j \geq j_0$):

$$P(D_j) = P(Y \geq a) \leq G(\rho') = \exp(j/2)\exp(-a^2/2\sigma^2)\Big(\frac{a^2}{j\sigma^2}\Big)^{j/2} =$$

$$\exp(j/2)\ \exp(-j/2)\ \exp(-j/2\ (2+3\beta)\log M_n)\ (1 + (2+3\beta)\log M_n)^{j/2}$$

$$\leq M_n^{-j} M_n^{-j_0\beta}$$

where we have used $(1 + (2+3\beta)\ \log M_n) \leq M_n^\beta$. □

Presumably, the need to consider the event A_2 in the above proof has been overlooked in [1].

Parameter Improvements:

We now briefly compare the values of Λ_n' in (4) and Λ_n in (8). We will refer to the *energy* of a signal, by which we mean the square of its L^2-norm. define $e' \equiv \lambda'\Lambda_n'/\Big(\sigma^2(\lambda' - 8)\Big)$ and $e \equiv 2\ \lambda\Lambda_n/\Big(\sigma^2(\lambda - 2)\Big)$. The numbers e' and e are the coefficients in the right hand sides of (5) and (10) respectively. We will set $\beta = 1$ in these parameters. On the one hand, for a given λ' and setting $\lambda = \lambda'/4$, it is easy to see that

$$3.2\ \Lambda_n \leq \Lambda' \text{ and } e \leq 0.6\ e'$$

So in this case ($\lambda = \lambda'/4$) Theorem 3 constructs a more energetic estimate and reduces the upper bound by almost half. On the other hand, we can compare the improvements in the upper bounds for the errors by choosing λ such that $\Lambda_n \approx \Lambda'_n$. Therefore, if for purposes of comparison we force $\Lambda_n = \frac{2\lambda'^2\sigma^2}{5}\,(1 + 5\log M_n) \leq \Lambda'_n$, under this condition $\hat{s}^\star$ has a larger energy than $\hat{s}'$. It also follows that

$$\frac{e'}{e} = \frac{\lambda'(\lambda-2)}{(\lambda'-8)2\lambda} \text{ and } \lambda'^2 = 10\lambda.$$

As a numerical example we can take $\lambda' = 8.1$ which gives $e'/e > 28$. Therefore we obtain a more energetic estimate while reducing considerably the upper bound for the error.

5. Discussion

It should be possible to generalize our results to colored noise. It was mentioned previously that it is more natural to assume $z \in Sub(\Omega, \mathbb{R}^n)$ than the stronger hypothesis $z \in SSub(\Omega, \mathbb{R}^n)$. We now explain our use of this later assumption. Let ξ be the projection of z in some library subspace (as in (23)). It is then crucial to obtain useful estimates on the matrix elements of the companion matrix of ξ. The computation in (24) gives this estimate. One can always find analogous upper bounds (for matrix elements) under the sole assumption $z \in Sub(\Omega, \mathbb{R}^n)$, but these estimates are not useful for the purposes of this paper. On the other hand, this problem is not present for the case in which we only have to deal with a single basis as both B^o and $\widehat{B}^\star$ are both subsets of a common orthonormal set, namely $\mathcal{B}$. This implies that, for the single basis case, we can relax the hypothesis on the noise and only require it to be sub-Gaussian. A previous version of the present document presented a theorem with such a result, we did not include this theorem in our final version of the paper as it was shown to us (by a referee) that such a result did not provide optimal bounds. Moreover, when dealing with sub-Gaussian noise one has to address the issue of estimating $\tau^2(z)$.

Appendix A. Noise Results

In this appendix we state several basic results and provide examples related to our noise set-up, most of these results are used directly or indirectly in the main body of the paper. Most proofs can be found in [4].

Lemma 2. *Assume $\xi_1, \ldots, \xi_n$ are independent sub-Gaussian random variables. Then*

$$\tau^2\left(\sum_{k=1}^{n} \xi_k\right) \leq \sum_{k=1}^{n} \tau^2(\xi_k).$$

Lemma 3. *Let ξ and η be independent and in $SSub(\Omega)$. Then*

$$\tau^2(\xi+\eta) = \mathbb{E}\Big((\xi+\eta)^2\Big)$$

(so $\xi + \eta \in SSub(\Omega)$).

Lemma 4. *Assume $\xi \in SSub(\Omega, \mathbb{R}^n)$ and let A be a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $A\xi \in SSub(\Omega, \mathbb{R}^m)$.*

Lemma 5. *Suppose that $\xi_1, \dots, \xi_n$ are jointly independent elements of $SSub(\Omega)$ and $\{e_i\}_{i=1,\dots,n}$ an orthonormal basis in $\mathbb{R}^n$. Then*

$$\xi = \sum_{i=1}^{n} \xi_i \; e_i \in SSub(\Omega, \mathbb{R}^n).$$

Remark 4. *Notice that, given a collection of jointly independent elements from $SSub(\Omega)$, the previous lemma allows to construct new elements from $SSub(\Omega)$ by choosing different orthonormal basis of $\mathbb{R}^n$.*

Lemma 6. *One has $\xi \in SSub(\Omega, \mathbb{R}^n)$ if and only if $\langle u, \xi\rangle \in SSub(\Omega)$ for any $u \in \mathbb{R}^n$.*

Theorem 4. *Assume $\xi \in Sub(\Omega, \mathbb{R}^n)$ and let $B = (b_{k,j})_{k,j=1}^n$ be its companion matrix which we will assume to be a diagonal matrix. Then if $r > \; 2 \; \max_k b_{k,k}$*

$$\mathbb{E}\Big(e^{1/r \sum_{k=1}^n \xi_k^2}\Big) \; \leq \prod_{k=1}^{n} \Big(1 - \frac{2b_{k,k}}{r}\Big)^{-1/2}. \tag{29}$$

The next two results give sufficient conditions for a random variable ξ to be in $SSub(\Omega)$.

Proposition 1. *Let ξ be symmetric and $\mathbb{E}(\xi^2) < \infty$. If*

$$\mathbb{E}(\xi^{2k}) \leq (\mathbb{E}(\xi^2))^k \frac{(2k)!}{2^k x!}, \quad k \geq 1$$

then $\xi \in SSub(\Omega)$.

Proposition 2. *Assume the characteristic function $\psi(z) = \mathbb{E}(e^{iz\xi})$, $z \in \mathbb{C}$, of a zero mean random variable ξ is an entire function of finite order. If either $\{z : \psi(z) = 0\} = \emptyset$ or $\{z : \psi(z) = 0\} \subseteq \mathbb{R}$, then $\xi \in SSub(\Omega)$.*

Examples:

1) If ξ is a bounded random variable with zero mean then $\xi \in Sub(\Omega)$.

2) Gaussian random variables ξ with mean zero are in $SSub(\Omega)$.

3) Consider a random variable ξ and a positive real number m such that $P(\xi = m) = P(\xi = -m) = \frac{1-\mu}{2}$ and $P(\xi = 0) = \mu$. Then $\mu \in [0, 2/3]$ implies that $\xi \in SSub(\Omega)$. To prove this we argue as follows, first, we indicate that $\mathbb{E}(\xi^2) = m^2(1-\mu)$. Notice that for $\mu \in [0, 2/3]$

$$\frac{2^k k!}{(1-\mu)^{k-1}(2k)!} \leq \frac{6^k k!}{3(2k)!} \leq 1 \text{ for all } k \geq 1. \tag{30}$$

Now we compute

$$\mathbb{E}(e^{\lambda\xi}) = \mu + \frac{(1-\mu)}{2}(e^{m\lambda} - e^{-m\lambda}) = \tag{31}$$

$$1 + (1-\mu)\sum_{k=1}^{\infty} \frac{(m\lambda)^{2k}}{(2k)!} \leq 1 + \sum_{k=1}^{\infty} \frac{(1-\mu)^k (m\lambda)^2 k}{2^k k!} = e^{\frac{(1-\mu)m^2\lambda^2}{2}} = e^{\frac{\mathbb{E}(\xi^2)\lambda^2}{2}}.$$

Where we used (30) to obtain the inequality in (31). In general $\mathbb{E}(\xi^2) \leq \tau^2(\xi)$, hence the above computation shows $\mathbb{E}(\xi^2) = \tau^2(\xi)$.

4) Consider a positive real number m and a random variable ξ uniformly distributed over $[-m, m]$. Then $\xi \in SSub(\Omega)$. To prove this we argue as follows, first, we indicate that $\mathbb{E}(\xi^2) = m^2/3$. Now we compute

$$\mathbb{E}(e^{\lambda\xi}) = \frac{1}{2m}\int_{-m}^{m} e^{\lambda x}dx = \sum_{k=0}^{\infty}\frac{(\lambda m)^{2k}}{(2k+1)!}, \tag{32}$$

using $6^k k! \leq (2k+1)!$ we can then estimate

$$\mathbb{E}(e^{\lambda\xi}) = 1 + \sum_{k=1}^{\infty}\frac{(\lambda m)^{2k}}{(2k+1)!} \leq 1 + \sum_{k=1}^{\infty}\frac{(\lambda m)^{2k}}{6^k k!} = e^{m^2\lambda^2/6} = e^{\mathbb{E}(\xi^2)\lambda^2/2}.$$

Therefore, it follows that $\tau^2(\xi) = \mathbb{E}(\xi^2)$.

Appendix B. Intermediate Results

Most of the results below appear explicitly or implicitly in [1], they are described here for the reader's convenience. We have used a direction dependent thresholding for the sake of generality.

Complexity functional: Introduce the definition,

$$K(\tilde{s}, s) \;\equiv\; \|\tilde{s} - s\|_2^2 + \min_{\mathcal{B}\in\mathcal{L}} \sum_{e_i\in\mathcal{B},\ \langle \tilde{s}, e_i\rangle \neq 0} \Lambda_n(e_i). \tag{33}$$

For the purposes of the computations in this appendix $\Lambda_n(e_i)$ could be taken to be any positive real number, in our applications we take $\Lambda_n(e_i)$ to be given by (8). Recall that $\hat{s}^\star$ was defined (in the basis $\hat{\mathcal{B}}$) by

$$\hat{s}_i^\star[\hat{\mathcal{B}}] = \eta_{\sqrt{\Lambda_n}}(y_i[\hat{\mathcal{B}}])$$

where $\hat{\mathcal{B}} = \text{arg}\ \text{min}_{\mathcal{B}\in\mathcal{L}}\mathcal{E}_\lambda(y, \mathcal{B}) = \text{arg}\ \text{min}_{\mathcal{B}\in\mathcal{L}} \sum_i \min(y_i^2[\mathcal{B}], \Lambda_n(e_i))$. We have the following properties:

$$\text{K1} \qquad \hat{s}^\star = \text{arg}\ \text{min}_{\tilde{s}} K(\tilde{s}, y) \tag{34}$$

$$\text{K2} \qquad K(\hat{s}^\star, s) \geq \|\hat{s}^\star - s\|_2^2 \tag{35}$$

$$\text{K3} \qquad \text{min}_{\tilde{s}}\, K(\tilde{s}, s) = \text{min}_{\mathcal{B}} \sum_{e_i\in\mathcal{B}} \min(s_i^2[\mathcal{B}], \Lambda_n(e_i)). \tag{36}$$

K2 is self-evident but we present proofs of the other two statements.

Proof of K1:

Fix y and basis $\mathcal{B}$. For any s define

$$K_{\mathcal{B}}(s, y) \equiv \|s - y\|_2^2 \;+ \sum_{e_i\in\mathcal{B}\ \langle s, e_i\rangle\neq 0} \Lambda_n(e_i).$$

Note that

$$\begin{aligned} K_{\mathcal{B}}(s,y) &= \sum_{i=1}^{n} K_{\mathcal{B},i}(s,y) \qquad \text{where} \\ K_{\mathcal{B},i}(s,y) &= |s_i[\mathcal{B}] - y_i[\mathcal{B}]|^2 + \Lambda_n(e_i)\ \chi(s_i[\mathcal{B}]), \\ &\quad \text{where}\ \ \chi(t) = 1 \text{ if } t \neq 0 \text{ and } \chi(t) = 0 \text{ if } t = 0. \end{aligned}$$

Define the vector q^i by

$$\begin{aligned} q^i_j[\mathcal{B}] &= a_i \delta_{ij} \\ a_i &= \begin{cases} y_i[\mathcal{B}], & |y_i[\mathcal{B}]| > \sqrt{\Lambda_n(e_i)} \\ 0 & |y_i[\mathcal{B}]| \leq \sqrt{\Lambda_n(e_i)} \end{cases} \end{aligned}$$

Let $u \neq q^i$. The following table exhausts all the possibilities for $K_{\mathcal{B},i}(u,y)$;

	$a_i = y_i$ (so $K_{\mathcal{B},i}(q^i,y) = \Lambda_n(e_i) < y_i^2$)	$a_i = 0$ (so $K_{\mathcal{B},i}(q^i,y) = y_i^2 \leq \Lambda_n(e_i)$)
$u_i = a_i$	$K_{\mathcal{B},i}(u,y) = \Lambda_n(e_i)$	$K_{\mathcal{B},i}(u,y) = y_i^2$
$u_i \neq a_i$	$K_{\mathcal{B},i}(u,y) = \|u_i - y_i\|^2 + \chi(u_i)\Lambda_n(e_i)$	$K_{\mathcal{B},i}(u,y) = \|u_i - y_i\|^2 + \Lambda_n(e_i)$

We see that $K_{\mathcal{B},i}(q^i,y) \leq K_{\mathcal{B},i}(u,y)$. Thus, q^i is a global minimizer of $K_{\mathcal{B},i}(s,y)$. By minimizing each $K_{\mathcal{B},i}(s,y)$,

$$s^{\star}_{\mathcal{B}} = q^1 + q^2 + \cdots + q^n$$

is a *global* minimizer of $K_{\mathcal{B}}(s,y)$.

Note that $s^{\star}_i[\mathcal{B}] = \eta_{\sqrt{\Lambda_n(e_i)}}(y_i[\mathcal{B}])$ in the notation above, and

$$K_{\mathcal{B}}(s^{\star}_{\mathcal{B}}, y) = \mathcal{E}_{\lambda}(y, \mathcal{B})$$

Therefore,

$$\begin{aligned} K_{\widehat{\mathcal{B}}}(s^{\star}_{\widehat{\mathcal{B}}}, y) &= \mathcal{E}_{\lambda}(y, \widehat{\mathcal{B}}) \\ &= \min_{\mathcal{B}} \mathcal{E}_{\lambda}(y, \mathcal{B}) \\ &= \min_{\mathcal{B}} K_{\mathcal{B}}(s^{\star}_{\mathcal{B}}, y) \\ &\geq K(\hat{s}^{\star}, y). \end{aligned} \tag{37}$$

To show the last inequality, pick any basis $\mathcal{B}_1$ and let $s^{\star}_{\mathcal{B}_1}$ be defined as above. Then,

$$\begin{aligned} K_{\mathcal{B}_1}(s^{\star}_{\mathcal{B}_1}, y) &= \|s^{\star}_{\mathcal{B}_1} - y\|_2^2 + \sum_{e_i \in \mathcal{B}_1\ \langle s^{\star}_{\mathcal{B}_1}, e_i\rangle \neq 0} \Lambda_n(e_i) \\ &\geq \|s^{\star}_{\mathcal{B}_1} - y\|_2^2 + \min_{\mathcal{B}} \sum_{e_i \in \mathcal{B}\ \langle s^{\star}_{\mathcal{B}_1}, e_i\rangle \neq 0} \Lambda_n(e_i) \\ &= K(s^{\star}_{\mathcal{B}_1}, y) \\ &\geq K(\hat{s}^{\star}, y). \end{aligned}$$

The following relations hold;

$$K_{\widehat{\mathcal{B}}}(s^{\star}_{\widehat{\mathcal{B}}}, y) \leq K_{\mathcal{B}}(s^{\star}_{\mathcal{B}}, y)\ \forall\, \mathcal{B}$$

$$K_{\mathcal{B}}(s^\star_{\mathcal{B}}, y) \ \leq \ K_{\mathcal{B}}(s, y) \ \ \forall\, s \ \ \forall\, \mathcal{B}$$

Pick any vector s and let $\mathcal{B}_2$ be the basis $\mathcal{B}$ that minimizes $\sum_{e_i \in \mathcal{B} \ \langle s, e_i\rangle \neq 0} \Lambda_n(e_i)$ (i.e., so $K(s, y) = K_{\mathcal{B}_2}(s, y)$). Then,

$$K(s^\star_{\widehat{\mathcal{B}}}, y) \leq K_{\widehat{\mathcal{B}}}(s^\star_{\widehat{\mathcal{B}}}, y) \leq K_{\mathcal{B}_2}(s^\star_{\mathcal{B}_2}, y) \leq K_{\mathcal{B}_2}(s, y) = K(s, y)$$

The first inequality was demonstrated above (in the equation following (56)), while the second and third inequalities follow from the two relations above. Thus,

$$K(s^\star_{\widehat{\mathcal{B}}}, y) \ \leq \ K(s, y) \ \ \forall\, s$$

Since the minimizer of $K(\cdot\,,\, y)$ is unique, $s^\star_{\widehat{\mathcal{B}}} = \hat{s}^\star$. This completes the proof of K1.

Proof of K3:

Let $\bar{s} = \arg\min_{\tilde{s}} K(\tilde{s}, s)$. Then, as shown above, $\bar{s}$ is defined in the basis $\hat{\mathcal{B}}(= \hat{\mathcal{B}}_s)$ by $\bar{s}_i = \eta_{\sqrt{\Lambda_n}}(s_i[\hat{\mathcal{B}}])$ where $\hat{\mathcal{B}} = \arg\min_{\mathcal{B}\in\mathcal{L}} \mathcal{E}_\lambda(s, \mathcal{B})$. Thus, $\min_{\tilde{s}} K(\tilde{s}, s) = K(\bar{s}, s) = K_{\hat{\mathcal{B}}}(\bar{s}, s) = \mathcal{E}_\lambda(s, \hat{\mathcal{B}})$, (by (56)), which proves K3.

Other relations

Let s^o denote a signal of minimum theoretical complexity (see (9)). Therefore,

$$K(\hat{s}^\star, y) \leq K(s^o, y).$$

Also,

$$K(\hat{s}^\star, y) = \|\hat{s}^\star - y\|_2^2 + \min_{\mathcal{B}\in\mathcal{L}} \sum_{i,\ \hat{s}^\star_i[\mathcal{B}]\neq 0} \Lambda_n(e_i) \tag{38}$$

$$= \langle \hat{s}^\star - y, \hat{s}^\star - y\rangle + \min_{\mathcal{B}\in\mathcal{L}} \sum_{i,\ \hat{s}^\star_i[\mathcal{B}]\neq 0} \Lambda_n(e_i) =$$

$$\langle \hat{s}^\star - s - z, \hat{s}^\star - s - z\rangle + \min_{\mathcal{B}\in\mathcal{L}} \sum_{i,\ \hat{s}^\star_i[\mathcal{B}]\neq 0} \Lambda_n(e_i)$$

$$= \langle \hat{s}^\star - s, \hat{s}^\star - s\rangle + \langle \hat{s}^\star - s, -z\rangle + \langle \hat{s}^\star - s, -z\rangle + \langle -z, -z\rangle + \min_{\mathcal{B}\in\mathcal{L}} \sum_{i,\ \hat{s}^\star_i[\mathcal{B}]\neq 0} \Lambda_n(e_i)$$

$$= K(\hat{s}^\star, s) + 2\langle z, s - \hat{s}^\star\rangle + \|z\|_2^2$$

Therefore,

$$\begin{aligned} K(\hat{s}^\star, s) &= K(\hat{s}^\star, y) - 2\langle z, s - \hat{s}^\star\rangle - \|z\|_2^2 \\ &\leq K(s^o, y) - 2\langle z, s - \hat{s}^\star\rangle - \|z\|_2^2 \\ &= \|s^o - (s+z)\|_2^2 + \min_{\mathcal{B}\in\mathcal{L}} \sum_{i,\ s^o_i[\mathcal{B}]\neq 0} \Lambda_n(e_i) - 2\langle z, s - \hat{s}^\star\rangle - \|z\|_2^2 \\ &= \|s^o - s\|_2^2 + \min_{\mathcal{B}\in\mathcal{L}} \sum_{i,\ s^o_i[\mathcal{B}]\neq 0} \Lambda_n(e_i) + 2\langle z, s - s^o\rangle + \|z\|_2^2 - 2\langle z, s - \hat{s}^\star\rangle - \|z\|_2^2 \\ &= K(s^o, s) + 2\langle z, \hat{s}^\star - s^o\rangle. \end{aligned} \tag{39}$$

Acknowledgements: We would like to thank Dr. John Baxter for his crucial input and help in the initial steps of the present paper.

References

[1] D. L. Donoho, I.M. Johnstone, Ideal denoising in an orthonormal basis chosen from a library of bases. *Comptes Rendus Aca. Sci. Paris A* **319** (1994) 1327-1322.

[2] D. L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage. Biometrika, **8**, 3 (1994), 425-255.

[3] S. Mallat, A Wavelet Tour of Signal Processing. Academic Press, New York, 1999.

[4] V.V. Buldygin and Yu. V. Kozachenko, *Metric Characterization of Random Variables and Random Processes.* Translations of Mathematical Monographs, American Mathematical Society, Vol. 188, 2000.

[5] H. Krim, D. Tucker, S. Mallat and D. Donoho, *On denoising and best signal representation.* IEEE Transactions on Information Theory, Vol. **45**, No. 7, Novembebr 1999, 2225-2238.

[6] M. Talagrand and M. Ledoux, *Probability in Banach Spaces*, Springer 1980.