Constructive subsampling of finite frames with applications in optimal function recovery

Felix Bartel, Martin Schäfer, Tino Ullrich*

Chemnitz University of Technology, Faculty of Mathematics, 09107 Chemnitz, Germany

Abstract

In this paper we present new constructive methods, random and deterministic, for the efficient subsampling of finite frames in $\mathbb{C}^m$. Based on a suitable random subsampling strategy, we are able to extract from any given frame with bounds $0 < A \leq B < \infty$ (and condition B/A) a similarly conditioned reweighted subframe consisting of merely $\mathcal{O}(m \log m)$ elements. Further, utilizing a deterministic subsampling method based on principles developed by Batson, Spielman, and Srivastava to control the spectrum of sums of Hermitian rank-1 matrices, we are able to reduce the number of elements to $\mathcal{O}(m)$ (with a constant close to one). By controlling the weights via a preconditioning step, we can, in addition, preserve the lower frame bound in the unweighted case. This permits the derivation of new quasi-optimal unweighted (left) Marcinkiewicz-Zygmund inequalities for $L_2(D, \nu)$ with constructible node sets of size $\mathcal{O}(m)$ for m-dimensional subspaces of bounded functions. Those can be applied e.g. for (plain) least-squares sampling reconstruction of functions, where we obtain new quasi-optimal results avoiding the Kadison-Singer theorem. Numerical experiments indicate the applicability of our results.

Keywords: finite frames, sampling, least squares recovery
2010 MSC: 41A10, 41A25, 41A60, 41A63, 42A10, 68Q25, 68W40, 94A20

1. Introduction

The paper mainly deals with the question of how to choose a well-conditioned subframe out of a given frame. The notion of a frame goes back to Duffin and Schaeffer [13]. Let H be a complex Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. A countable subset $(\boldsymbol{y}^i)_i$ in H is said to be a *frame* if there are constants $0 < A \leq B < \infty$ such that

$$A\|\boldsymbol{x}\|^2 \leq \sum_i |\langle \boldsymbol{x}, \boldsymbol{y}^i \rangle|^2 \leq B\|\boldsymbol{x}\|^2 \quad \text{for all} \quad \boldsymbol{x} \in H\,. \tag{1.1}$$

We are mostly interested in frames consisting of finitely many elements in some finite dimensional Hilbert space H, see e.g. Casazza and Kutyniok [7] for an introduction to finite frame theory. Systems of this kind may be represented by $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$ with $M \geq m$. The question of finding good subframes in such a system is rather fundamental and important for many applications ranging from graph sparsifiers [3, 32], the Kadison-Singer problem [24, 38], to optimal discretization and sampling recovery of multivariate functions [11, 20, 21, 27, 22, 34, 30]. In this context, let us also mention the possibility of generating 'approximations' of Hadamard matrices, a problem which has been considered in [12] for example. Subsampling of a tight Hadamard frame $(\boldsymbol{y}^i)_{i=1}^M$, where all entries of the $\boldsymbol{y}^i$ are ± 1 (or in a complex setting of modulus 1), may lead to an almost square Hadamard-type matrix with good condition.

*Corresponding author
Email addresses: felix.bartel@mathematik.tu-chemnitz.de (Felix Bartel), martin.schaefer@mathematik.tu-chemnitz.de (Martin Schäfer), tino.ullrich@mathematik.tu-chemnitz.de (Tino Ullrich)

As a first step, we put forward a simple random subsampling strategy in **Section 2**. For the index set $\{1,\dots,M\}$ of the intial system let us subsequently use the short-hand notation $[M]$. Theorem 2.1 shows that drawing elements $\boldsymbol{y}^i$, $i \in [M]$, according to the probabilities $\varrho_i := \|\boldsymbol{y}^i\|_2^2/(\sum_j \|\boldsymbol{y}^j\|_2^2)$ yields a reweighted subframe $(\varrho_i^{-1/2}\boldsymbol{y}^i)_{i\in J}$ with similar frame bounds, with high probability provided $|J| = \mathcal{O}(m \log m)$. In terms of computational complexity this strategy is very efficient. It is not optimal with respect to the number of frame elements, however.

This shortcoming is dealt with in **Section 3**. Here we formulate a deterministic algorithm (`BSS`) which yields reweighted subframes with an optimal order of $\mathcal{O}(m)$ frame elements. It is an extension of a subsampling method due to Batson, Spielmann, Srivastava [3] to the complex and non-tight case, which in the original version only applies to tight frames in $\mathbb{R}^m$. Let us remark that an extension to tight frames in $\mathbb{C}^m$ was already considered in [22, Cor. 2.1]. We further modify the subsampling method in [3] by allowing variable barrier shifts. As proved in Theorem 3.1, the `BSS` algorithm constructs in polynomial time for any frame $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$ with constants $0 < A \le B < \infty$ and $b > \kappa^2 \ge 1$, $\kappa = \kappa(A,B)$ being the value in (3.1), a subset $J \subset [M]$ and weights $s_i \ge 0$ such that

$$A\|\boldsymbol{x}\|_2^2 \le \sum_{i\in J} s_i |\langle \boldsymbol{x}, \boldsymbol{y}^i\rangle|^2 \le \gamma \cdot B\|\boldsymbol{x}\|_2^2 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{C}^m \tag{1.2}$$

with $|J| \le \lceil bm \rceil$ and $\gamma = \gamma(b,\kappa)$ as in (3.3). Combined with a suitable 'preconditioning' step, it can even achieve (1.2) for any $b > 1$ and with the better constant $\gamma = \gamma(b,1)$. In this variant (`BSS`$^\perp$), at the core, `BSS` is only applied to a tight frame. Since exact tightness can usually not be guaranteed in practice, due to numerical inaccuracies, it is nevertheless important for applications that `BSS` works stably also for non-tight frames, as ensured by Theorem 3.1.

Sections 2 and 3 lay the groundwork for the main **Section 4**, where we are concerned with the extraction of unweighted subframes. This is a much more difficult task. The existence of similarly conditioned subframes consisting of order $\mathcal{O}(m)$ elements is guaranteed by the solution of the famous Kadison-Singer problem by Markus, Spielmann, Srivastava [24]. To the knowledge of the authors, there are no general constructive polynomial-time methods available, however. Our approach to tackle this problem is to use the obtained results on weighted subframes (e.g. (1.2)) and try to control the weights s_i. By bounding those from above, we are able to preserve the lower frames bounds, which for many applications are the relevant ones.

One of the main results is Theorem 4.4. In Corollary 4.5, we obtain the non-weighted sparsification inequality (1.3), which holds true for arbitrary sets of vectors $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$. For any $1 + \frac{1}{m} < b \le \frac{M}{m}$ we can extract a subset $J \subset [M]$ of cardinality $|J| \le \lceil bm \rceil$ such that

$$\frac{1}{M}\sum_{i=1}^M |\langle \boldsymbol{x}, \boldsymbol{y}^i\rangle|^2 \le C_0 \frac{b^3}{(b-1)^3}\frac{1}{|J|}\sum_{i\in J} |\langle \boldsymbol{x}, \boldsymbol{y}^i\rangle|^2 \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{C}^m\,, \tag{1.3}$$

where $C_0 > 0$ is an absolute constant. The lower frame bound of an initial frame is thus preserved up to a constant $(b/(b-1))^3$. Earlier results in this direction, see Harvey, Olver [15], Nitzan, Olevskii, Ulanovski [28], Limonova, Temlyakov [22], and Nagel, Schäfer, T. Ullrich [27], are all non-constructive and have been initiated by the solution of the Kadison-Singer problem [24] in the form of Weaver's conjecture [38]. These results need additional restrictions on the size of the frame elements. They provide the upper frame bounds as well, however.

A **central message** in this paper is the fact that, at least for the lower bounds, we can argue much more elementary and do not need such deep results as Kadison-Singer. Our approach is semi-constructive (with an at times probabilistic component) and yields polynomial-time algorithms (e.g. `plainBSS`) that work with high probability and can be efficiently implemented. We developed a corresponding JULIA-package, availabe at `www.github.com/felixbartel/BSSsubsampling.jl`, and conducted numerical experiments with this code. A few results are presented in **Section 5**.

Some applications of our obtained subsampling results are discussed in **Section 6**. A first interesting consequence of (1.3) is the non-weighted left Marcinkiewicz-Zygmund inequality given by Theorem 6.2.

Assume we are given an m-dimensional space

$$V_m = \operatorname{span}\{\eta_1(\cdot), \cdots, \eta_m(\cdot)\} \tag{1.4}$$

of complex-valued functions on some non-empty set D. In case that $V_m \subset L_2(D,\nu)$ for a finite measure ν, (1.3) allows us to construct a set of nodes $\mathbf{X}_n = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n) \in D^n$ with $n \le \lceil bm \rceil$ for any $1 + \frac{1}{m} < b \le \frac{M}{m}$ in polynomial time (in m) such that

$$\|f\|^2_{L_2(D,\nu)} \le C_1 \frac{b^3}{(b-1)^3} \frac{1}{n} \sum_{i=1}^{n} |f(\boldsymbol{x}^i)|^2 \quad \text{for all} \quad f \in V_m\,. \tag{1.5}$$

Here $C_1 > 0$ is again an absolute constant. Inequalities like (1.5) have direct consequences for the sampling recovery of functions. With ν and V_m as before, let us consider a simple plain least squares recovery operator $S_{V_m}^{\boldsymbol{X}_n}$ for nodes $\boldsymbol{X}_n$ satisfying (1.5). It reconstructs any ν-measurable function $f : D \to \mathbb{C}$ in V_m via a plain least squares minimization at the nodes $\boldsymbol{X}_n$ and, according to Theorem 6.3, with

$$\|f - S_{V_m}^{\mathbf{X}_n} f\|^2_{L_2(D,\nu)} \le C_2 \frac{b^3}{(b-1)^3} e(f, V_m)^2_{\ell_\infty(D)}\,, \tag{1.6}$$

where $e(f, V_m)_{\ell_\infty(D)} = \inf_{g \in V_m} \|f - g\|_{\ell_\infty(D)}$ is the error of best approximation of f from V_m.

Inequalities of this type have been first established by Cohen and Migliorati [8], but with a larger number of samples, namely $n = \mathcal{O}(m \log(m))$. This has been improved by Temlyakov [34] to $n = \mathcal{O}(m)$ samples with unspecified constants. The mentioned results rely on weighted least squares algorithms, however, and Temlyakov posed the question in [34], if also classical plain least squares methods could be used. Theorem 6.3 gives an affirmative answer and even displays the dependence of the constant on the oversampling factor b. As a consequence we obtain from (1.6) for classes of bounded complex-valued functions $F \subset \ell_\infty(D)$, when optimizing over all m-dimensional reconstruction spaces, the new relation

$$g^{\mathrm{ls}}_{\lceil bm \rceil, m}(F, L_2(D,\nu)) \le C_3 \frac{b^{3/2}}{(b-1)^{3/2}} d_m(F, \ell_\infty(D)) \tag{1.7}$$

between the m-th Kolmogorov number d_m of the class F (see (6.11) for a definition) and the plain sampling numbers $g^{\mathrm{ls}}_{n,m}(F, L_2(D,\nu))$ defined in (6.10). Here recovery is restricted to canonical plain least squares operators using n samples acting on subspaces of dimension m. Since the quantities on the left-hand side of (1.7) are in general larger than the standard sampling numbers [9, (5.0.1)], where there are no such restrictions on the recovery, this slightly improves on recent results by Temlyakov, Theorems 1.1 and 1.2 in [34] as well as [22, Thm. 3.4], the latter joint work with Limonova. Interestingly, the b-dependent constant may be improved to $(b-1)^{-1}$ when allowing weighted least squares algorithms, cf. [22, Thm. 1.7] (or the original [10, Thm. 6.3]) for the case of real functions (an extension to the complex case has been given in [22, Rem. 3.2] but only for $b > 2$). In our case a distinction between real and complex $L_2(D,\nu)$ in (1.7) as in [22] is unnecessary due to the validity of Theorem 3.1 in the complex setting. Note that the right-hand side in (1.7) is of particular importance if the linear widths in L_2 are not square-summable [35, 36].

A related scenario is investigated in the recent papers [11, 20, 19, 25, 27]. Here one is interested in the recovery of functions from a reproducing kernel Hilbert space (RKHS) $H(K)$ in $L_2(D,\nu)$, where ν is allowed to be infinite. If we assume $H(K)$ to fulfill some natural assumptions (such as a finite trace of the kernel K and a compact embedding $\mathrm{Id}_{K,\nu} : H(K) \hookrightarrow L_2(D,\nu)$), our polynomial-time subsampling schemes allow to construct node sets $\boldsymbol{X}_n$ with $n \le \lceil bm \rceil$, $1 + \frac{1}{m} < b \le 2$, and a weight function w_m such that

$$\|f - S_{V_m, w_m}^{\boldsymbol{X}_n} f\|^2_{L_2(D,\nu)} \le C_4 \frac{1}{(b-1)^3} \log\Big(\frac{m}{p}\Big)\Big(\sigma^2_{m+1} + \frac{7}{m} \sum_{k=m+1}^{\infty} \sigma_k^2\Big) \|f\|^2_{H(K)} \tag{1.8}$$

for every $f \in H(K)$, with a probability exceeding $1 - \frac{3}{2}p$ for each $p \in (0, \frac{2}{3})$ (see Theorem 6.7). Here, $C_4 > 0$ is an absolute constant, $\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$ denote the singular numbers of the embedding $\mathrm{Id}_{K,\nu}$, and the

recovery operator $S^{\boldsymbol{X}_n}_{V_m,w_m}$ is a weighted least squares operator for the subspace V_m, as in (1.4), spanned by the left singular functions corresponding to the m largest singular numbers. The performance (1.8) is near-optimal as in [27]. The latter reference is the first which used the Weaver subsampling technique for the sampling recovery problem. However, it does not achieve the optimal rate. By a further refinement of the technique, established very recently in [11] by Dolbeault, Krieg, and M. Ullrich, the optimal rate (without additional log-term) has been found. In contrast to [11, 27] we have a semi-constructive method to generate the sampling nodes (offline step) that does not need the Kadison-Singer theorem in terms of the Weaver subsampling. In addition, the dependence on the oversampling factor b is displayed. An **open question** remains. Although in many relevant cases (like periodic Sobolev spaces with mixed smoothness) the recovery operator turns out to be a canonical plain least squares operator with equal weights (acting on the hyperbolic cross frequency subspace with nodes displayed in Figure 5.1) we do not know whether this is possible in general.

Throughout the paper, we will use the following **notation**. As usual, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{C}$ denote the natural (without 0), integer, real, and complex numbers. If not indicated otherwise $\log(\cdot)$ denotes the natural logarithm. For $m \in \mathbb{N}$ we further set $[m] := \{1, \dots, m\}$ and $\mathbb{N}_{\geq m} := \{m, m+1, \dots\}$. $\mathbb{C}^n$ shall denote the complex n-space and $\mathbb{C}^{m\times n}$ the set of complex $m \times n$-matrices. Vectors and matrices are usually typesetted boldface. For a vector $\boldsymbol{y} \in \mathbb{C}^n$ we introduce the tensor notation $\boldsymbol{y} \otimes \boldsymbol{y}$ for the matrix $\boldsymbol{y} \cdot \boldsymbol{y}^* \in \mathbb{C}^{n\times n}$, where $\boldsymbol{y}^* := \overline{\boldsymbol{y}}^\top$. More general, the adjoint of a matrix $\boldsymbol{L} \in \mathbb{C}^{m\times n}$ is denoted by $\boldsymbol{L}^*$. For the spectral norm we use $\|\boldsymbol{L}\|$ or $\|\boldsymbol{L}\|_{2\to 2}$ and we use $A \preceq B$ to denote that $B - A$ is positive semi-definite. Finally, we will write $\mathbb{E}(X)$ for the expectation of a random variable X and $\mathbb{P}(E)$ for the probability of an event E. In our case the probability measure $\mathbb{P}$ is a product measure $\mu^{\otimes n}$ (resp. $\varrho_m^{\otimes n}$) on D^n with a certain probability measure μ (resp. ϱ_m) on D, for both discrete and continuous domains D. The abbreviation i.i.d. refers to 'independent and identically distributed'. For a set D we denote with $\ell_\infty(D)$ the set of all bounded complex-valued functions on D. If D is ν-measurable we denote with $L_2(D, \nu)$ the space of all ν-measurable square-integrable functions (equivalence classes) on D.

2. Random weighted subsampling of finite frames

We begin with a random subsampling strategy that allows to extract 'good' subframes of $\mathcal{O}(m \log m)$ elements out of any given frame in $\mathbb{C}^m$. This goes back to Rudelson and Vershynin [31], see also Spielman and Srivastava [32], where the goal was to efficiently find a low rank approximation of a given matrix such that the error with respect to the spectral norm remains small. The method is rather simple since it relies on a random subselection where the discrete probability mass ϱ_i for selecting one particular frame element $\boldsymbol{y}^i$ (see Theorem 2.1) is directly linked to its contribution to the sum of the norms, i.e., the Frobenius norm $\|\boldsymbol{Y}\|_F^2$ of the matrix

$$\boldsymbol{Y} := \begin{bmatrix} \overline{(\boldsymbol{y}^1)^*} \\ \hline \vdots \\ \hline \overline{(\boldsymbol{y}^M)^*} \end{bmatrix} \in \mathbb{C}^{M\times m}. \tag{2.1}$$

Note that for a given frame $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$, with $M \geq m$ and $m \in \mathbb{N}$, this matrix represents the analysis operator of the frame and that

$$mA \leq \operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{Y}) = \|\boldsymbol{Y}\|_F^2 \leq m\|\boldsymbol{Y}^*\boldsymbol{Y}\|_{2\to 2} = m\lambda_{\max}(\boldsymbol{Y}^*\boldsymbol{Y}) \leq mB. \tag{2.2}$$

Our main result of this section relies on a matrix Chernoff bound proven by Tropp [37, Thm. 1.1] (see Theorem A.3 in the Appendix). It shows how one can randomly subsample a finite frame of arbitrary size in $\mathbb{C}^m$ to a weighted subframe with $\mathcal{O}(m \log m)$ elements while essentially keeping its stability properties.

Theorem 2.1. *Let $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$ be a frame with constants $0 < A \leq B < \infty$ (see (1.1)). Let further $p, t \in (0,1)$ and $n \in \mathbb{N}$ be such that*

$$n \geq \frac{3B}{At^2} m \log\left(\frac{2m}{p}\right).$$

Drawing n indices $J \subset [M]$ (with duplicates) i.i.d. according to the discrete probability density $\varrho_i = \|\boldsymbol{y}^i\|_2^2/\|\boldsymbol{Y}\|_F^2$, $i \in [M]$, then gives a rescaled random subframe $(\varrho_i^{-1/2}\boldsymbol{y}^i)_{i\in J}$ such that

$$(1-t)A\|\boldsymbol{a}\|_2^2 \leq \frac{1}{n}\sum_{i\in J}\left|\left\langle \boldsymbol{a}, \varrho_i^{-1/2}\boldsymbol{y}^i\right\rangle\right|^2 \leq (1+t)B\|\boldsymbol{a}\|_2^2 \quad \text{for all} \quad \boldsymbol{a} \in \mathbb{C}^m$$

with probability exceeding $1-p$.

Proof. The result is a direct consequence of Tropp's result in Lemma A.3. For a randomly chosen index $i \in [M]$ we define the rank-one random matrix $\boldsymbol{A}_i := \frac{1}{n}\varrho_i^{-1}(\boldsymbol{y}^i \otimes \boldsymbol{y}^i)$. Clearly, it holds

$$\lambda_{\max}(\boldsymbol{A}_i) = \lambda_{\max}\left(\frac{1}{n}\varrho_i^{-1}(\boldsymbol{y}^i \otimes \boldsymbol{y}^i)\right) = \frac{1}{n}\varrho_i^{-1}\|\boldsymbol{y}^i\|_2^2 = \frac{1}{n}\|\boldsymbol{Y}\|_F^2.$$

Furthermore, having n independent copies $(\boldsymbol{A}_i)_{i\in J}$, we obtain

$$\sum_{i\in J}\mathbb{E}\boldsymbol{A}_i = \sum_{i\in J}\mathbb{E}\left(\frac{1}{n}\varrho_i^{-1}(\boldsymbol{y}^i \otimes \boldsymbol{y}^i)\right) = \sum_{i\in J}\frac{1}{n}\boldsymbol{Y}^*\boldsymbol{Y} = \boldsymbol{Y}^*\boldsymbol{Y}.$$

This gives for $\mu_{\min} := \lambda_{\min}(\sum_{i\in J}\mathbb{E}\boldsymbol{A}_i)$ and $\mu_{\max} := \lambda_{\max}(\sum_{i\in J}\mathbb{E}\boldsymbol{A}_i)$ that

$$\mu_{\min} = \lambda_{\min}(\boldsymbol{Y}^*\boldsymbol{Y}) \geq A \quad \text{and} \quad \mu_{\max} = \lambda_{\max}(\boldsymbol{Y}^*\boldsymbol{Y}) = \|\boldsymbol{Y}^*\boldsymbol{Y}\|_{2\to 2}^2.$$

Since $\|\boldsymbol{Y}\|_F^2 = \operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{Y})$, Lemma A.3 and (2.2) gives

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i\in J}\varrho_i^{-1}\boldsymbol{y}^i \otimes \boldsymbol{y}^i\right) \geq (1+t)B\right) &\leq m\exp\left(-\frac{n\|\boldsymbol{Y}^*\boldsymbol{Y}\|_{2\to 2}}{\operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{Y})}\frac{t^2}{3}\right)\\ &\leq m\exp\left(-\frac{n}{m}\frac{t^2}{3}\right).\end{aligned}$$

For the smallest eigenvalue things are a bit different. Here we obtain

$$\begin{aligned}\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{n}\sum_{i\in J}\varrho_i^{-1}\boldsymbol{y}^i \otimes \boldsymbol{y}^i\right) \leq (1-t)A\right) &\leq m\exp\left(-\frac{nA}{\operatorname{tr}(\boldsymbol{Y}^*\boldsymbol{Y})}\frac{t^2}{2}\right)\\ &\leq m\exp\left(-\frac{An}{Bm}\frac{t^2}{2}\right).\end{aligned}$$

For the probability of our assertion we need the complement of the two events above:

$$1 - m\exp\left(-\frac{n}{m}\frac{t^2}{3}\right) - m\exp\left(-\frac{An}{Bm}\frac{t^2}{2}\right) \geq 1 - 2m\exp\left(-\frac{An}{Bm}\frac{t^2}{3}\right) \geq 1-p,$$

which follows from the assumption on n. ∎

Remark 2.2. *The rescaled random subframe $(\varrho_i^{-1/2}\boldsymbol{y}^i)_{i\in J}$ in Theorem 2.1 is an equal-norm frame. Thus, starting with a tight frame, we are able to construct an 'almost tight' frame with unit-norm (UNTF). These are important in robust data transmission and have proven notoriously difficult to construct, cf. [5, 6].*

3. Deterministic weighted subsampling of finite frames

We next present a deterministic subsampling algorithm for finite frames in $\mathbb{C}^m$ which we subsequently call (generalized) BSS algorithm. A version for real-valued tight frames in $\mathbb{R}^m$ was originally introduced by Batson, Spielman, and Srivastava in the context of graph sparsification [3]. It allows to extract from any given finite frame in $\mathbb{C}^m$ a comparably well-conditioned re-weighted subframe of cardinality $\mathcal{O}(m)$. This is the statement of Theorem 3.1 below which generalizes [3, Thm. 3.1].

In contrast to related non-weighted subsampling results, such as e.g. [27, Thm. 2.3] which are all based on Weaver's theorem, a deep result equivalent to the famous Kadison-Singer theorem [24], the proof of Theorem 3.1 is elementary and constructive. The underlying BSS algorithm lends itself to practical polynomial time implementation.

Theorem 3.1. *Let $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$ be a frame (1.1) with frame constants $0 < A \le B < \infty$ and let $b > \kappa^2 \ge 1$ with*

$$\kappa := \left(\frac{B}{2A} + \frac{1}{2}\right) + \sqrt{\left(\frac{B}{2A} + \frac{1}{2}\right)^2 - 1}\,. \tag{3.1}$$

Then the BSS algorithm in Subsection 3.2 computes a subset $J \subset [M]$ with $|J| \le \lceil bm \rceil$ and nonnegative weights s_i, $i \in J$, such that

$$A\|\boldsymbol{a}\|_2^2 \le \sum_{i\in J} s_i |\langle \boldsymbol{a}, \boldsymbol{y}^i\rangle|^2 \le \gamma \cdot B\|\boldsymbol{a}\|_2^2 \quad \text{for all} \quad \boldsymbol{a} \in \mathbb{C}^m \tag{3.2}$$

with

$$\gamma := \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)(\sqrt{b}-\kappa)}\,. \tag{3.3}$$

Remark 3.2. *(i) The BSS algorithm computes the index subset J and the corresponding weights s_i in $\mathcal{O}(bMm^3)$. An implementation and runtime analysis is given in Subsection 3.2, see also [3, Sec.3]. Better guarantees on the bound can be obtained by a 'preconditioning' of the frame, given by Lemma 3.4. The resulting algorithm is called BSS$^\perp$. In BSS$^\perp$ also the restriction $b > \kappa^2$ can be evaded. Some empirical results are presented in Section 5.*

(ii) The theorem neither gives control over the weights s_i nor provides an unweighted version of itself. The latter would actually be useful for applications. We refer to [27] for an unweighted result which is called 'Weaver subsampling' and relies on the Kadison-Singer theorem [24]. In Section 4 below we will use a special construction from Lemma 4.3 to deduce an unweighted version that preserves the left frame bound, cf. Corollary 4.5.

3.1. The principal structure of the BSS algorithm

The frame property of the vectors $(\boldsymbol{y}^i)_{i=1}^M$ can be formulated as

$$A\boldsymbol{I} \preceq \sum_{i=1}^M \boldsymbol{y}^i(\boldsymbol{y}^i)^* \preceq B\boldsymbol{I}\,,$$

where $\boldsymbol{I}$ denotes the identity matrix in $\mathbb{C}^{m\times m}$ (see the notation paragraph for the meaning of $\preceq$). Furthermore, condition (3.2) of the subsampled frame can be rewritten as

$$A\boldsymbol{I} \preceq \sum_{i\in J} s_i \boldsymbol{y}^i(\boldsymbol{y}^i)^* \preceq \gamma \cdot B\boldsymbol{I}\,. \tag{3.4}$$

The idea of the BSS algorithm is to build the sum $\sum_{i\in J} s_i \boldsymbol{y}^i(\boldsymbol{y}^i)^*$ iteratively in $n := \lceil bm \rceil$ steps. Starting with the zero-matrix $\boldsymbol{A}^{(0)} := 0$, a sequence of Hermitian matrices

$$\boldsymbol{A}^{(0)},\ \boldsymbol{A}^{(1)},\ \boldsymbol{A}^{(2)},\ \ldots,\ \boldsymbol{A}^{(n)} \tag{3.5}$$

is computed via rank-1 updates of the form

$$\boldsymbol{A}^{(k)} = \boldsymbol{A}^{(k-1)} + t^{(k)} \boldsymbol{y}^{i^{(k)}} (\boldsymbol{y}^{i^{(k)}})^* \,, \quad k \in [n] \,,$$

with suitably selected indices $i^{(k)} \in [M]$ and weights $t^{(k)} > 0$. After n iterations we have thus constructed a matrix $\boldsymbol{A}^{(n)}$ of the form

$$\boldsymbol{A}^{(n)} = \sum_{k=1}^{n} t^{(k)} \boldsymbol{y}^{i^{(k)}} (\boldsymbol{y}^{i^{(k)}})^* = \sum_{i \in J} \tilde{s}_i \boldsymbol{y}^i (\boldsymbol{y}^i)^* \tag{3.6}$$

where

$$\tilde{s}_i := \sum_{k : i^{(k)} = i} t^{(k)} \quad \text{and} \quad J := \left\{ i^{(k)} : k = 1, \ldots, n \right\} .$$

Clearly, $|J| \leq n = \lceil bm \rceil$. During the whole process the spectra of the constructed matrices $\boldsymbol{A}^{(k)}$ are controlled by means of so-called spectral barriers, i.e., numbers $l^{(k)}, u^{(k)} \in \mathbb{R}$ such that

$$\sigma(\boldsymbol{A}^{(k)}) \subset (l^{(k)}, u^{(k)}) \,, \quad k \in \{0, \ldots, n\} \,. \tag{3.7}$$

Whereas the precise location of the eigenvalues of $\boldsymbol{A}^{(k)}$ may not be known, in this way we have enclosed their location in open intervals $(l^{(k)}, u^{(k)})$, in particular it holds

$$l^{(k)} \boldsymbol{I} \preceq \boldsymbol{A}^{(k)} \preceq u^{(k)} \boldsymbol{I} \,.$$

The algorithm starts with initial barriers $l^{(0)} < 0$ and $0 < u^{(0)}$ for $\boldsymbol{A}^{(0)} = 0$. From each step to the next, the barriers are then shifted to the right, most simply by certain fixed lengths $\delta_L > 0$ and $\delta_U > 0$. In the kth iteration we thus have $l^{(k)} = l^{(0)} + k\delta_L$ and $u^{(k)} = u^{(0)} + k\delta_U$ (see Figure 3.1). For each $\boldsymbol{A}^{(k)}$ the indices and weights $i^{(k+1)}$ and $t^{(k+1)}$ are further chosen such that (3.7) remains valid for the updated matrix $\boldsymbol{A}^{(k+1)}$. This is the main technical challenge and requires some preparation, carried out in Subsections 3.3 and 3.4. Under these conditions, the final matrix $\boldsymbol{A}^{(n)}$ then has property (3.7) for

$$l^{(n)} = l^{(0)} + n\delta_L \quad \text{and} \quad u^{(n)} = u^{(0)} + n\delta_U \,.$$

As shown in Subsection 3.5, for the 'right' choice of $l^{(0)}$, $u^{(0)}$, δ_L, and δ_U we end up with final barriers satisfying

$$l^{(n)} > 0 \quad \text{and} \quad \frac{u^{(n)}}{l^{(n)}} \leq \gamma \cdot \frac{B}{A} \,.$$

This finally allows to rescale the weights $\tilde{s}_i$ in (3.6) appropriately, giving the desired weights s_i such that (3.4) is fulfilled.

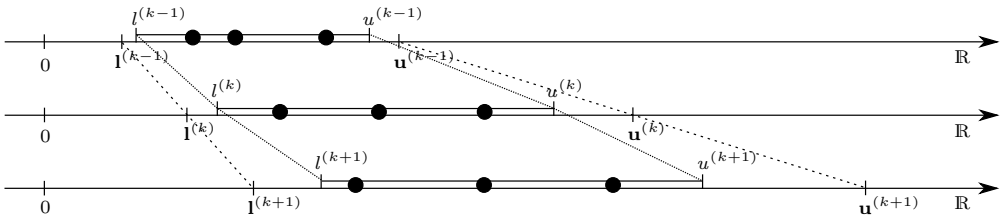

Figure 3.1: Spectral shifting via `constant` and *variable* barrier shifts.

3.2. A concrete implementation and runtime analysis

Before providing a profound theoretical basis for the BSS algorithm (and with that also a rigorous proof of Theorem 3.1), starting in Subsection 3.3, let us first present a concrete numerical implementation. The subsequent version, Algorithm 1, was implemented for the purpose of empirical analysis (see Section 5). Instead of fixed barrier shifts δ_L and δ_U it uses variable shifts $\delta_L^{(k)}$ and $\delta_U^{(k)}$ depending on the iteration step k.

A JULIA code is available at `www.github.com/felixbartel/BSSsubsampling.jl`.

Algorithm 1 BSS

Input: Frame $\boldsymbol{y}^1, \dots, \boldsymbol{y}^M \in \mathbb{C}^m$ with frame bounds $0 < A \le B < \infty$;
Oversampling factor $b > \kappa^2$ with κ as in (3.1); Stability factor $\Delta \ge 0$.

Output: Nonnegative weights s_i such that $\sqrt{s_1}\boldsymbol{y}^1, \dots \sqrt{s_M}\boldsymbol{y}^M$ is a frame with $|\{i : s_i > 0\}| \le \lceil bm \rceil$ and bounds $0 < A \le B\gamma(1+\Delta) < \infty$.
($\gamma := \gamma(b, \kappa)$ is the value from (3.3).)

Put $n := \lceil bm \rceil$ and $\kappa := \kappa(A, B)$ as in (3.1). Further, set $\boldsymbol{A}^{(0)} := 0$,

$$l^{(0)} := -m\frac{\sqrt{b}\kappa}{1+\Delta}\,, \quad u^{(0)} := m\frac{b+\sqrt{b}}{\sqrt{b}-1}\frac{B}{A}\,, \quad \delta_L^{(0)} := \frac{1}{1+\Delta}\,, \quad \delta_U^{(0)} := \frac{\sqrt{b}+1}{\sqrt{b}-1}\frac{B}{A}\,.$$

▷ $\boldsymbol{A}^{(0)} \in \mathbb{C}^{m\times m}$ is the zero matrix, $l^{(0)}$, $u^{(0)}$ associated lower and upper spectral barriers. The initial barrier shifts are given by $\delta_L^{(0)}$, $\delta_U^{(0)}$.

for $k = 1$ **to** n **do**

Compute the eigenvalues $\lambda_1^{(k-1)}, \dots, \lambda_m^{(k-1)}$ of $\boldsymbol{A}^{(k-1)}$.

Compute the so-called lower and upper potentials (see Definition 3.5)

$$\epsilon_L^{(k-1)} := \Phi_{l^{(k-1)}}(\boldsymbol{A}^{(k-1)}) = \textstyle\sum_{j=1}^m \left(\lambda_j^{(k-1)} - l^{(k-1)}\right)^{-1},$$
$$\epsilon_U^{(k-1)} := \Phi^{u^{(k-1)}}(\boldsymbol{A}^{(k-1)}) = \textstyle\sum_{j=1}^m \left(u^{(k-1)} - \lambda_j^{(k-1)}\right)^{-1}.$$

Put $\delta_L^{(k-1)} := \left(\frac{1}{\delta_L^{(0)}} - \kappa\epsilon_L^{(0)} + \kappa\epsilon_L^{(k-1)}\right)^{-1}$ and $\delta_U^{(k-1)} := \left(\frac{1}{\delta_U^{(0)}} + \epsilon_U^{(0)} - \epsilon_U^{(k-1)}\right)^{-1}$.

Increment $l^{(k-1)}$ and $u^{(k-1)}$: $l^{(k)} := l^{(k-1)} + \delta_L^{(k-1)}$, $u^{(k)} := u^{(k-1)} + \delta_U^{(k-1)}$.

Compute the factors

$$f_L^{(k-1)} := \Phi_{l^{(k)}}(\boldsymbol{A}^{(k-1)}) = \textstyle\sum_{j=1}^m \left(\lambda_j^{(k-1)} - l^{(k)}\right)^{-1},$$
$$f_U^{(k-1)} := \Phi^{u^{(k)}}(\boldsymbol{A}^{(k-1)}) = \textstyle\sum_{j=1}^m \left(u^{(k)} - \lambda_j^{(k-1)}\right)^{-1}.$$

for $j = 1$ **to** M **do**

Compute

$$L^{(k-1)}(\boldsymbol{y}^j) := \frac{(\boldsymbol{y}^j)^*(\boldsymbol{A}^{(k-1)} - l^{(k)}\boldsymbol{I})^{-2}\boldsymbol{y}^j}{f_L^{(k-1)} - \epsilon_L^{(k-1)}} - (\boldsymbol{y}^j)^*(\boldsymbol{A}^{(k-1)} - l^{(k)}\boldsymbol{I})^{-1}\boldsymbol{y}^j\,,$$
$$U^{(k-1)}(\boldsymbol{y}^j) := \frac{(\boldsymbol{y}^j)^*(\boldsymbol{A}^{(k-1)} - u^{(k)}\boldsymbol{I})^{-2}\boldsymbol{y}^j}{\epsilon_U^{(k-1)} - f_U^{(k-1)}} - (\boldsymbol{y}^j)^*(\boldsymbol{A}^{(k-1)} - u^{(k)}\boldsymbol{I})^{-1}\boldsymbol{y}^j\,.$$

if $L^{(k-1)}(\boldsymbol{y}^j) - U^{(k-1)}(\boldsymbol{y}^j) \ge \frac{\Delta}{2M}\left(1 - \frac{1}{\sqrt{b}}\right)$ **then**

denote this index by $i^{(k)}$.

break

end if

end for

15: Compute
$t^{(k)} := 2\big(L^{(k-1)}(\boldsymbol{y}^{i^{(k)}}) + U^{(k-1)}(\boldsymbol{y}^{i^{(k)}})\big)^{-1}$,
$\tilde{s}_{i^{(k)}} := \tilde{s}_{i^{(k)}} + t^{(k)}$,
$\boldsymbol{A}^{(k)} := \boldsymbol{A}^{(k-1)} + t^{(k)} \boldsymbol{y}^{i^{(k)}} (\boldsymbol{y}^{i^{(k)}})^*$.
16: **end for**
17: **return** rescaled weights $s_i := \frac{1}{2}\Big(\frac{A}{l^{(n)}} + \frac{B\gamma(1+\Delta)}{u^{(n)}}\Big)\tilde{s}_i$ for $i = 1, \dots, M$.

As explained in Subsection 3.1, we want to produce a sequence (3.5) of matrices $\boldsymbol{A}^{(k)}$ which fulfill the spectral condition (3.7) for the respective spectral barriers $l^{(k)}$ and $u^{(k)}$. Algorithm 1 accomplishes this. For the details we refer to Subsection 3.5. A crucial step is the index selection in line 10. The condition there guarantees that the chosen $i^{(k)}$ and the subsequently computed $t^{(k)}$ lead to a new updated matrix $\boldsymbol{A}^{(k)}$ (in line 15) which fulfills (3.7) as the matrices $\boldsymbol{A}^{(0)}, \dots, \boldsymbol{A}^{(k-1)}$ did before. According to Corollary 3.9, proved below, essential for this is $L^{(k-1)}(\boldsymbol{y}^{i^{(k)}}) \geq U^{(k-1)}(\boldsymbol{y}^{i^{(k)}})$ and $t^{(k)} \in [(L^{(k-1)}(\boldsymbol{y}^{i^{(k)}}))^{-1}, (U^{(k-1)}(\boldsymbol{y}^{i^{(k)}}))^{-1}]$. To avoid numerical issues in the selection, which might occur due to calculation inaccuracies, the stability parameter $\Delta \geq 0$ comes into play. It ensures that $L^{(k-1)}(\boldsymbol{y}^{i^{(k)}}) \geq U^{(k-1)}(\boldsymbol{y}^{i^{(k)}})$ can be verified, via the condition in line 10, in a numerically stable manner.

Remark 3.3. *Algorithm 1 also works for fixed barrier shifts. We can skip the update in line 5 and always use $\delta_L^{(0)}$ and $\delta_U^{(0)}$ in the subsequent incrementation step in line 6. The advantage of variable shifts is a sharper containment of the spectrum (see illustration in Fig. 3.1).*

By including a preceding orthogonalization procedure, it is possible to allow arbitrarily small oversampling factors $b > 1$. Further the guarantees on the bounds improve. The modified algorithm is called $\texttt{BSS}^{\perp}$. It is based on the following simple observation.

Lemma 3.4. *For every matrix $\boldsymbol{Y} \in \mathbb{C}^{M \times m}$ with $M \geq m$ there is a matrix $\tilde{\boldsymbol{Y}} \in \mathbb{C}^{M \times m}$ such that*

$$\mathrm{R}(\tilde{\boldsymbol{Y}}) \supset \mathrm{R}(\boldsymbol{Y}), \quad \tilde{\boldsymbol{Y}}^* \tilde{\boldsymbol{Y}} = \boldsymbol{I}, \quad \textit{and} \quad \|\tilde{\boldsymbol{Y}}\|_F^2 = m\,,$$

where $\mathrm{R}(\tilde{\boldsymbol{Y}})$ and $\mathrm{R}(\boldsymbol{Y})$ denote the range in $\mathbb{C}^M$ of the respective operators.

Proof. The matrix $\tilde{\boldsymbol{Y}}$ is constructed by applying the Gram-Schmidt algorithm to the columns of $\boldsymbol{Y}$. If we end up with less than m vectors, which happens if $\mathrm{rank}(\boldsymbol{Y}) < m$, we orthogonally extend them, which is possible since $M \geq m$. ∎

Algorithm 2 $\texttt{BSS}^{\perp}$

Input: Frame $\boldsymbol{y}^1, \dots, \boldsymbol{y}^M \in \mathbb{C}^m$ with frame bounds $0 < A \leq B < \infty$;
Oversampling factor $b > 1$; Stability factor $\Delta \geq 0$.

Output: Nonnegative weights s_i such that $\sqrt{s_1}\boldsymbol{y}^1, \dots \sqrt{s_M}\boldsymbol{y}^M$ is a frame with $|\{i : s_i > 0\}| \leq \lceil bm \rceil$ and bounds $0 < A \leq B\gamma(1+\Delta) < \infty$.
(γ is the value from (3.3) for $\kappa = 1$.)

1: Let $\boldsymbol{Y} \in \mathbb{C}^{M \times m}$ be the matrix with rows $\boldsymbol{y}^1, \dots, \boldsymbol{y}^M$ and construct $\tilde{\boldsymbol{Y}} \in \mathbb{C}^{M \times m}$ as in Lemma 3.4 via Gram-Schmidt orthogonalization of the columns of $\boldsymbol{Y}$.
2: **return** weights $s_1, \dots, s_M$, calculated by applying BSS (Algorithm 1) to the rows of $\tilde{\boldsymbol{Y}}$.

Note that the rows $\tilde{\boldsymbol{y}}^1, \dots, \tilde{\boldsymbol{y}}^M$ of $\tilde{\boldsymbol{Y}}$, constructed in line 1 of Algorithm 2, form a tight frame. Hence, in lines 2 Algorithm 1 can be applied for arbitrarily small $b > 1$. In fact, the frame property of the initial

system $(\boldsymbol{y}^i)_{i=1}^M$ is not needed for this. It is possible to run $\texttt{BSS}^\perp$ for any input vector sequence $(\boldsymbol{y}^i)_{i=1}^M$ in $\mathbb{C}^m$, satisfying $M \geq m$. The returned weights s_i always fulfill $|\{i : s_i \neq 0\}| \leq \lceil bm \rceil$ and it always holds

$$\sum_{i=1}^{M} \left|\langle \boldsymbol{a}, \boldsymbol{y}^i \rangle\right|^2 \leq \sum_{i=1}^{M} s_i \left|\langle \boldsymbol{a}, \boldsymbol{y}^i \rangle\right|^2 \leq \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}(1+\Delta) \sum_{i=1}^{M} \left|\langle \boldsymbol{a}, \boldsymbol{y}^i \rangle\right|^2 \tag{3.8}$$

for all $a \in \mathbb{C}^m$. Assuming the input sequence was a frame, we then further deduce

$$A\|\boldsymbol{a}\|_2^2 \leq \sum_{i=1}^{M} s_i \left|\langle \boldsymbol{a}, \boldsymbol{y}^i \rangle\right|^2 \leq \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}(1+\Delta)B\|\boldsymbol{a}\|_2^2 \quad \text{for all} \quad \boldsymbol{a} \in \mathbb{C}^m \,.$$

To verify (3.8), let us first reformulate this inequality as

$$\|\boldsymbol{Y}\boldsymbol{a}\|_2^2 \leq \|\boldsymbol{S}^{\frac{1}{2}}(\boldsymbol{Y}\boldsymbol{a})|_J\|_2^2 \leq \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}(1+\Delta)\|\boldsymbol{Y}\boldsymbol{a}\|_2^2 \,, \tag{3.9}$$

where $J \coloneqq \{i : s_i \neq 0\}$, $(\boldsymbol{Y}\boldsymbol{a})|_J$ stands for the restriced vector $([\boldsymbol{Y}\boldsymbol{a}]_i)_{i\in J} \in \mathbb{C}^{|J|}$, and $\boldsymbol{S} \coloneqq \operatorname{diag}(s_i)_{i\in J} \in \mathbb{C}^{|J|\times|J|}$. By Theorem 3.1, applying BSS (Algorithm 1) to $(\tilde{\boldsymbol{y}}^i)_{i=1}^M$ yields s_i such that $|J| = |\{i : s_i \neq 0\}| \leq \lceil bm \rceil$ and

$$\|\boldsymbol{a}\|_2^2 \leq \sum_{i=1}^{M} s_i |\langle \boldsymbol{a}, \tilde{\boldsymbol{y}}^i \rangle|^2 \leq \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}(1+\Delta)\|\boldsymbol{a}\|_2^2 \quad \text{for all} \quad \boldsymbol{a} \in \mathbb{C}^m \,.$$

Therefore, by the orthogonality of $\tilde{\boldsymbol{Y}}$, for all $\boldsymbol{a} \in \mathbb{C}^m$

$$\|\tilde{\boldsymbol{Y}}\boldsymbol{a}\|_2^2 \leq \|\boldsymbol{S}^{\frac{1}{2}}(\tilde{\boldsymbol{Y}}\boldsymbol{a})|_J\|_2^2 \leq \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}(1+\Delta)\|\tilde{\boldsymbol{Y}}\boldsymbol{a}\|_2^2 \,.$$

Using $\mathrm{R}(\tilde{\boldsymbol{Y}}) \supset \mathrm{R}(\boldsymbol{Y})$, as guaranteed by Lemma 3.4, we may finally replace $\tilde{\boldsymbol{Y}}$ in this last inequality with the original $\boldsymbol{Y}$, leading to (3.9).

Runtime analysis of Algorithm 1. The singular value decomposition (SVD) of $\boldsymbol{A}^{k-1}$ in the kth iteration step has a complexity of $\mathcal{O}(m^3)$. Having the SVD decomposition at hand, matrix-vector products with $(\boldsymbol{A}^{(k-1)} - l^{(k)}\boldsymbol{I})^{-1}$, $(\boldsymbol{A}^{(k-1)} - l^{(k)}\boldsymbol{I})^{-2}$, $(\boldsymbol{A}^{(k-1)} - u^{(k)}\boldsymbol{I})^{-1}$, and $(\boldsymbol{A}^{(k-1)} - u^{(k)}\boldsymbol{I})^{-2}$ are computable in $\mathcal{O}(m^2)$. To eventually decide, which index is selected in line 10, $L^{(k-1)}(\boldsymbol{y}^i)$ and $U^{(k-1)}(\boldsymbol{y}^i)$ in the worst case need to be computed for all $i \in [M]$, . This thus may require $\mathcal{O}(Mm^2)$ multiplication steps. All in all, taking into account $M \geq m$, each iteration can be performed in $\mathcal{O}(Mm^2)$ time. Since the number of iterations is $\lceil bm \rceil$, the total time of the algorithm is $\mathcal{O}(bMm^3)$. In our implementation we used a random procedure to traverse the indices $i \in [M]$ and noticed that this speeds up the algorithm, see Section 5, Experiment 3.

Instead of computing the singular value decomposition from scratch every iteration it is possible to update it continuously, cf. [4, 26], which we have not implemented.

3.3. Spectral analysis of rank-1 updates

In this subsection we analyze from a general perspective, how the spectrum $\sigma(\boldsymbol{A})$ of a Hermitian matrix $\boldsymbol{A} \in \mathbb{C}^{m\times m}$ changes under a rank-1 update of the form

$$\boldsymbol{A} \rightsquigarrow \boldsymbol{A}' \coloneqq \boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^* \tag{3.10}$$

with $\boldsymbol{v} \in \mathbb{C}^m$ and $t \in \mathbb{R}$. This question has already been discussed for the real setting in [3, Sec. 3.1 & 3.2]. Our analysis here is analogous, however, we go a bit more into detail. In the end, we can derive precise conditions in Lemmas 3.7, 3.8, and Corollary 3.9.

With the matrix determinant lemma (Lemma A.1 in the Appendix) the characteristic polynomial $p_{\boldsymbol{A}'}(\lambda) = \det(\lambda \boldsymbol{I} - \boldsymbol{A}')$ of $\boldsymbol{A}'$ in (3.10) can be calculated explicitly. For $\lambda \notin \sigma(\boldsymbol{A})$

$$p_{\boldsymbol{A}'}(\lambda) = \det(\lambda \boldsymbol{I} - \boldsymbol{A}) \left(1 - t \boldsymbol{v}^* (\lambda \boldsymbol{I} - \boldsymbol{A})^{-1} \boldsymbol{v}\right) = p_{\boldsymbol{A}}(\lambda) \left(1 - t \sum_{j=1}^{m} \frac{|\langle \boldsymbol{v}, \boldsymbol{u}^j \rangle|^2}{\lambda - \lambda_j}\right),$$

where $\{\lambda_1, \ldots, \lambda_m\}$ are the, not necessarily distinct, eigenvalues of $\boldsymbol{A}$ and $\{\boldsymbol{u}^j\}_{j=1}^m$ is a corresponding orthonormal basis of eigenvectors. The eigenvalues of $\boldsymbol{A}'$ can be obtained from the associated characteristic equation. In case $t = 0$, we obtain as solutions the eigenvalues of $\boldsymbol{A}$, i.e., the roots of $p_{\boldsymbol{A}}(\lambda)$. In case $t \neq 0$, we obtain the roots of $p_{\boldsymbol{A}}(\lambda)$ of at least second order together with the solutions of the so-called secular equation

$$\sum_{j=1}^{m} \frac{|\langle \boldsymbol{v}, \boldsymbol{u}^j \rangle|^2}{\lambda - \lambda_j} = \frac{1}{t}.$$

A discussion of this equation yields the following insight: If $t \neq 0$ the eigenvalues of $\boldsymbol{A}'$ interlace the eigenvalues of $\boldsymbol{A}$, shifted to the left when $t < 0$ and shifted to the right when $t > 0$. Moreover, the shifts occur continuously in t. In the limit $t \to \infty$, the largest eigenvalue tends to ∞ and the corresponding eigenvector to $\boldsymbol{v}$. For $t \to -\infty$, the smallest eigenvalue tends to $-\infty$, while the corresponding eigenvector again tends to $\boldsymbol{v}$. In case of algebraic multiplicities, always merely one of the respective eigenvalues moves, the rest remain at their old position.

With this, we already have a good qualitative picture of what happens to $\sigma(\boldsymbol{A})$ when applying a rank-1 update (3.10). Next, we want to quantify how far the eigenvalues are shifted depending on the size of t. Again we follow [3] and utilize so-called potential functions.

Definition 3.5 (cf. [3, Def. 3.2])**.** *Let $\boldsymbol{I}$ denote the identity matrix in $\mathbb{C}^{m \times m}$. For $l, u \in \mathbb{R}$ and a Hermitian matrix $\boldsymbol{A} \in \mathbb{C}^{m \times m}$ with eigenvalues $\{\lambda_1, \ldots, \lambda_m\} \subset \mathbb{R}$ the* lower and upper potential functions *$\Phi_l(\boldsymbol{A})$ and $\Phi^u(\boldsymbol{A})$ are given by*

$$\Phi_l(\boldsymbol{A}) := \operatorname{tr}([\boldsymbol{A} - l\boldsymbol{I}]^{-1}) = \sum_{i=1}^{m} \frac{1}{\lambda_i - l},$$
$$\Phi^u(\boldsymbol{A}) := \operatorname{tr}([u\boldsymbol{I} - \boldsymbol{A}]^{-1}) = \sum_{i=1}^{m} \frac{1}{u - \lambda_i}.$$

When l and u are lower, respectively upper, barriers for $\sigma(\boldsymbol{A})$, i.e., when $l < \lambda_{\min}(\boldsymbol{A})$, respectively $\lambda_{\max}(\boldsymbol{A}) < u$, these potential functions serve well as measures for the distance of $\sigma(\boldsymbol{A})$ to the respective barriers. Note that the so-measured distance counts in the whole spectrum of $\boldsymbol{A}$, i.e., the location of all eigenvalues matters. Further, it holds: the larger the distance, the lower the potentials, and the smaller the distance, the larger the potentials.

From the qualitative discussion above it is clear that the upper potential $\Phi^u(\boldsymbol{A}')$ becomes smaller when t decreases and larger when t increases. The lower potential $\Phi_l(\boldsymbol{A}')$ behaves the other way round, it becomes smaller when t increases and larger when t decreases. Based on the Sherman-Morrison formula, we now precisely quantify the change of the potentials.

Lemma 3.6. *Suppose $\boldsymbol{A} \in \mathbb{C}^{m \times m}$ is Hermitian and let $\boldsymbol{v} \in \mathbb{C}^m$ be a vector.*

(i) For $l \notin \sigma(\boldsymbol{A})$ and $t \neq -(\boldsymbol{v}^*[\boldsymbol{A} - l\boldsymbol{I}]^{-1}\boldsymbol{v})^{-1}$

$$\Phi_l(\boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^*) = \Phi_l(\boldsymbol{A}) - \frac{t\boldsymbol{v}^*[\boldsymbol{A} - l\boldsymbol{I}]^{-2}\boldsymbol{v}}{1 + t\boldsymbol{v}^*[\boldsymbol{A} - l\boldsymbol{I}]^{-1}\boldsymbol{v}}.$$

(ii) For $u \notin \sigma(\boldsymbol{A})$ and $t \neq (\boldsymbol{v}^*(u\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v})^{-1}$

$$\Phi^u(\boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^*) = \Phi^u(\boldsymbol{A}) + \frac{t\boldsymbol{v}^*(u\boldsymbol{I} - \boldsymbol{A})^{-2}\boldsymbol{v}}{1 - t\boldsymbol{v}^*(u\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{v}}.$$

Proof. (i): The lower potential has the form

$$\Phi_l(\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*) = \operatorname{tr}\left(\left[\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*-l\boldsymbol{I}\right]^{-1}\right).$$

Using the Sherman-Morrison formula and properties of the trace, we obtain

$$\begin{aligned}\Phi_l(\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*) &= \operatorname{tr}\left([\boldsymbol{A}-l\boldsymbol{I}]^{-1} - \frac{t[\boldsymbol{A}-l\boldsymbol{I}]^{-1}\boldsymbol{v}\boldsymbol{v}^*[\boldsymbol{A}-l\boldsymbol{I}]^{-1}}{1+t\boldsymbol{v}^*[\boldsymbol{A}-l\boldsymbol{I}]^{-1}\boldsymbol{v}}\right)\\ &= \operatorname{tr}\left([\boldsymbol{A}-l\boldsymbol{I}]^{-1}\right) - \frac{t\operatorname{tr}\left([\boldsymbol{A}-l\boldsymbol{I}]^{-1}\boldsymbol{v}\boldsymbol{v}^*[\boldsymbol{A}-l\boldsymbol{I}]^{-1}\right)}{1+t\boldsymbol{v}^*[\boldsymbol{A}-l\boldsymbol{I}]^{-1}\boldsymbol{v}}\\ &= \Phi_l(\boldsymbol{A}) - \frac{t\boldsymbol{v}^*[\boldsymbol{A}-l\boldsymbol{I}]^{-2}\boldsymbol{v}}{1+t\boldsymbol{v}^*[\boldsymbol{A}-l\boldsymbol{I}]^{-1}\boldsymbol{v}}.\end{aligned}$$

(ii): Analogous to (i). ∎

Next, we turn our attention to barrier shifts, i.e., modifications of the barriers of the form $l \rightsquigarrow l' := l+\delta_L$ and $u \rightsquigarrow u' := u+\delta_U$. Our concrete goal is to specify shifts for which the spectrum of the updated matrix $\boldsymbol{A}'$ in (3.10) is enclosed in (l',u') and the distance of $\sigma(\boldsymbol{A}')$ to the barriers has not decreased. In other words, we want to find $l',u' \in \mathbb{R}$ such that

$$\sigma(\boldsymbol{A}') \subset (l',u'), \quad \Phi_{l'}(\boldsymbol{A}') \le \Phi_l(\boldsymbol{A}), \quad \text{and} \quad \Phi^{u'}(\boldsymbol{A}') \le \Phi^u(\boldsymbol{A}). \tag{3.11}$$

In Lemma 3.7 below, which corresponds to [3, Lem. 3.3], we first handle the lower potential.

Lemma 3.7 (cf. [3, Lem. 3.3])**.** *Let $\boldsymbol{A} \in \mathbb{C}^{m\times m}$ be Hermitian and assume $\Delta_L := \lambda_{\min}(\boldsymbol{A}) - l > 0$ for $l \in \mathbb{R}$, $\boldsymbol{v} \in \mathbb{C}^m$ be any vector. Further let $\delta_L \in (-\infty, \Delta_L)$ and additionally assume $L_{\boldsymbol{A}}(\boldsymbol{v}; l, \delta_L) > 0$ in case $\delta_L > 0$, where $L_{\boldsymbol{A}}(\boldsymbol{v}; l, \delta_L) := \infty$ for $\delta_L = 0$ and otherwise*

$$L_{\boldsymbol{A}}(\boldsymbol{v}; l, \delta_L) := \frac{\boldsymbol{v}^*(\boldsymbol{A}-(l+\delta_L)\boldsymbol{I})^{-2}\boldsymbol{v}}{\Phi_{l+\delta_L}(\boldsymbol{A}) - \Phi_l(\boldsymbol{A})} - \boldsymbol{v}^*(\boldsymbol{A}-(l+\delta_L)\boldsymbol{I})^{-1}\boldsymbol{v}. \tag{3.12}$$

Then precisely for $t \ge L_{\boldsymbol{A}}(\boldsymbol{v}; l, \delta_L)^{-1}$ (with $\infty^{-1} = 0$)

$$\Phi_{l+\delta_L}(\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*) \le \Phi_l(\boldsymbol{A}) \quad \textit{and} \quad \lambda_{\min}(\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*) > l+\delta_L. \tag{3.13}$$

For $0 < \delta_L < \Delta_L$ with $L_{\boldsymbol{A}}(\boldsymbol{v}; l, \delta_L) \le 0$ there are no $t \in \mathbb{R}$ satisfying (3.13)*.*

Proof. Let $l' = l+\delta_L$. Due to $\delta_L < \Delta_L$, we have $\lambda_{\min}(\boldsymbol{A}) > l'$ and in particular $l' \notin \sigma(\boldsymbol{A})$. Hence, from Lemma 3.6, we directly derive

$$\Phi_{l'}(\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*) - \Phi_l(\boldsymbol{A}) = (\Phi_{l'}(\boldsymbol{A}) - \Phi_l(\boldsymbol{A})) - \frac{t\boldsymbol{v}^*[\boldsymbol{A}-l'\boldsymbol{I}]^{-2}\boldsymbol{v}}{1+t\boldsymbol{v}^*[\boldsymbol{A}-l'\boldsymbol{I}]^{-1}\boldsymbol{v}}.$$

Note further that always $\boldsymbol{v}^*(\boldsymbol{A}-l'\boldsymbol{I})^{-2}\boldsymbol{v} > 0$ and $\boldsymbol{v}^*(\boldsymbol{A}-l'\boldsymbol{I})^{-1}\boldsymbol{v} > 0$.

The case $\delta_L = 0$ is clear. Here $\Phi_l(\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*) \le \Phi_l(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*) > l$ precisely for $t \ge 0 = L_{\boldsymbol{A}}(\boldsymbol{v}; l, 0)^{-1}$.

Next we turn to $\delta_L > 0$. Here we have the additional assumption $L_{\boldsymbol{A}} := L_{\boldsymbol{A}}(\boldsymbol{v}; l, \delta_L) > 0$. We hence have $t \ge L_{\boldsymbol{A}}^{-1} > 0$ which implies $0 < 1/t \le L_{\boldsymbol{A}}$. We conclude

$$\frac{t\boldsymbol{v}^*[\boldsymbol{A}-l'\boldsymbol{I}]^{-2}\boldsymbol{v}}{1+t\boldsymbol{v}^*[\boldsymbol{A}-l'\boldsymbol{I}]^{-1}\boldsymbol{v}} \ge \frac{\boldsymbol{v}^*[\boldsymbol{A}-l'\boldsymbol{I}]^{-2}\boldsymbol{v}}{L_{\boldsymbol{A}}+\boldsymbol{v}^*[\boldsymbol{A}-l'\boldsymbol{I}]^{-1}\boldsymbol{v}} = \Phi_{l'}(\boldsymbol{A}) - \Phi_l(\boldsymbol{A}). \tag{3.14}$$

In addition, $\lambda_{\min}(\boldsymbol{A}+t\boldsymbol{v}\boldsymbol{v}^*) > \lambda_{\min}(\boldsymbol{A}) > l'$ due to $t > 0$.

Finally, if $\delta_L < 0$ then $\Phi_{l'}(\boldsymbol{A}) < \Phi_l(\boldsymbol{A})$ and thus $L_{\boldsymbol{A}} < -\boldsymbol{v}^*(\boldsymbol{A}-l'\boldsymbol{I})^{-1}\boldsymbol{v} < 0$. Hence, in the range $0 > t \ge L_{\boldsymbol{A}}^{-1}$ we have $1/t \le L_{\boldsymbol{A}} < -\boldsymbol{v}^*(\boldsymbol{A}-l'\boldsymbol{I})^{-1}\boldsymbol{v}$ and (3.14) is valid. For $t \ge 0$ (3.14) is also valid

since then we have a negative right-hand side and a positive left-hand side there. Also, clearly $\lambda_{\min}(\boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^*) \geq \lambda_{\min}(\boldsymbol{A}) > l'$ in case $t \geq 0$. If $t < 0$ we argue by contradiction. Assuming $\lambda_{\min}(\boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^*) \leq l'$, by continuity since $\lambda_{\min}(\boldsymbol{A}) > l'$, there would be t' with $t \leq t' < 0$ and $\lambda_{\min}(\boldsymbol{A} + t'\boldsymbol{v}\boldsymbol{v}^*) = l'$. But $\Phi_{l'}(\boldsymbol{A} + t'\boldsymbol{v}\boldsymbol{v}^*) \leq \Phi_{l'}(\boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^*) \leq \Phi_l(\boldsymbol{A}) < \infty$, which is a contradiction.

For the last statement, assume that $0 < \delta_L < \Delta_L$ and $L_{\boldsymbol{A}} \leq 0$. Then necessarily $t > 0$ for (3.13). Further $\Phi_{l'}(\boldsymbol{A}) > \Phi_l(\boldsymbol{A})$ and as a consequence $L_{\boldsymbol{A}} > -\boldsymbol{v}^*[\boldsymbol{A} - l'\boldsymbol{I}]^{-1}\boldsymbol{v}$. However $1/t > 0 \geq L_{\boldsymbol{A}}$ now contradicts (3.14) since $\frac{1}{t} + \boldsymbol{v}^*[\boldsymbol{A} - l'\boldsymbol{I}]^{-1}\boldsymbol{v} > L_{\boldsymbol{A}} + \boldsymbol{v}^*[\boldsymbol{A} - l'\boldsymbol{I}]^{-1}\boldsymbol{v} > 0$. ∎

In the next lemma, which corresponds to [3, Lem. 3.4], we handle the upper potential, where the situation is dual to the one before.

Lemma 3.8 (cf. [3, Lem. 3.4])**.** *Let $\boldsymbol{A} \in \mathbb{C}^{m\times m}$ be Hermitian and assume $\Delta_U := u - \lambda_{\max}(\boldsymbol{A}) > 0$ for $u \in \mathbb{R}$, $\boldsymbol{v} \in \mathbb{C}^m$ be any vector. Further let $\delta_U \in (-\Delta_U, \infty)$ and additionally assume $U_{\boldsymbol{A}}(\boldsymbol{v}; u, \delta_U) < 0$ in case $\delta_U < 0$, where $U_{\boldsymbol{A}}(\boldsymbol{v}; u, \delta_U) := \infty$ for $\delta_U = 0$ and otherwise*

$$U_{\boldsymbol{A}}(\boldsymbol{v}; u, \delta_U) := \frac{\boldsymbol{v}^*[(u+\delta_U)\boldsymbol{I} - \boldsymbol{A}]^{-2}\boldsymbol{v}}{\Phi^u(\boldsymbol{A}) - \Phi^{u+\delta_U}(\boldsymbol{A})} + \boldsymbol{v}^*[(u+\delta_U)\boldsymbol{I} - \boldsymbol{A}]^{-1}\boldsymbol{v}\,. \tag{3.15}$$

Then precisely for $t \leq U_{\boldsymbol{A}}(\boldsymbol{v}; u, \delta_U)^{-1}$ (with $\infty^{-1} = 0$)

$$\Phi^{u+\delta_U}(\boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^*) \leq \Phi^u(\boldsymbol{A}) \quad \text{and} \quad \lambda_{\max}(\boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^*) < u + \delta_U\,. \tag{3.16}$$

For $-\Delta_U < \delta_U < 0$ with $U_{\boldsymbol{A}}(\boldsymbol{v}; u, \delta_U) \geq 0$ there are no $t \in \mathbb{R}$ satisfying (3.16)*.*

Proof. Using duality, the proof carries over from Lemma 3.7 in a strictly analogous manner. ∎

Let us now assume that $\boldsymbol{A} \in \mathbb{C}^{m\times m}$ is Hermitian with $\sigma(\boldsymbol{A}) \subset (l, u)$. Then, for any fixed vector $\boldsymbol{v} \in \mathbb{C}^m$ and any given δ_U fulfilling the assumptions of Lemma 3.8, condition (3.16) can be ensured by choosing t small enough. Similarly, condition (3.13) in Lemma 3.7 can always be ensured for large t if δ_L satisfies the assumptions of this lemma. It is unclear, however, if there exist $t \in \mathbb{R}$ which fulfill both conditions simultaneously, i.e., values for t which ensure (3.11).

Combining Lemma 3.7 and Lemma 3.8 the subsequent corollary arises. It provides a precise condition when such t exist and determines their precise range.

Corollary 3.9. *Let $\boldsymbol{v} \in \mathbb{C}^m$ be a vector and $\boldsymbol{A} \in \mathbb{C}^{m\times m}$ a Hermitian matrix with lower and upper barriers $l, u \in \mathbb{R}$. Let further $\delta_L, \delta_U \in \mathbb{R}$ with $\delta_L < \Delta_L$ and $\delta_U > -\Delta_U$, where $\Delta_L := \lambda_{\min}(\boldsymbol{A}) - l$ and $\Delta_U := u - \lambda_{\max}(\boldsymbol{A})$. Then the following conditions are equivalent.*

(i) With $L_{\boldsymbol{A}} := L_{\boldsymbol{A}}(\boldsymbol{v}; u, \delta_U)$ and $U_{\boldsymbol{A}} := U_{\boldsymbol{A}}(\boldsymbol{v}; u, \delta_U)$ as in (3.12) *and* (3.15)*, respectively,*

$$\operatorname{sgn}\delta_L = \operatorname{sgn} L_{\boldsymbol{A}}\,, \quad \operatorname{sgn}\delta_U = \operatorname{sgn} U_{\boldsymbol{A}}\,, \quad L_{\boldsymbol{A}}^{-1} \leq U_{\boldsymbol{A}}^{-1}\,. \tag{3.17}$$

(ii) There exist $t \in \mathbb{R}$ such that (3.11) *is fulfilled for $\boldsymbol{A}' := \boldsymbol{A} + t\boldsymbol{v}\boldsymbol{v}^*$, $l' := l + \delta_L$, and $u' := u + \delta_U$.*

In case that (i) and (ii) hold true, (3.11) *is fulfilled precisely for $t \in [L_{\boldsymbol{A}}^{-1}, U_{\boldsymbol{A}}^{-1}]$.*

Proof. The corollary is a consequence of Lemma 3.7 and Lemma 3.8. To see this, note that $\delta_U > 0$ implies $U_{\boldsymbol{A}} > 0$ and that, similarly, $\delta_L < 0$ implies $L_{\boldsymbol{A}} < 0$. ∎

3.4. The well-determined update step

Let us now focus back on the `BSS` algorithm. The main result of this subsection, Lemma 3.11, is the key tool to keep control of the spectra of the matrices $\boldsymbol{A}^{(k)}$ in (3.5). It provides a condition on the barrier shifts δ_L and δ_U to guarantee that, when the spectral window $(l^{(k)}, u^{(k)})$ of $\boldsymbol{A}^{(k)}$ is shifted to $(l^{(k+1)}, u^{(k+1)})$, there is at least one frame vector $\boldsymbol{y}^i$ such that condition (3.17) in Corollary 3.9 is fulfilled. Hence, there exists $t > 0$ such that (3.11) is fulfilled for the update $\boldsymbol{A}^{(k+1)} = \boldsymbol{A}^{(k)} + t\boldsymbol{y}^i(\boldsymbol{y}^i)^*$, in particular $\sigma(\boldsymbol{A}^{(k+1)}) \subset (l^{(k+1)}, u^{(k+1)})$.

For the proof of Lemma 3.11 we need two auxiliary results. The first result is Lemma A.2 from the Appendix. The second auxiliary result is [3, Claim 3.6], which we recall for convenience here.

Lemma 3.10. *Let δ_L, $\epsilon_L > 0$, $l \in \mathbb{R}$, $\{\lambda_1, \dots, \lambda_m\} \subset \mathbb{R}$. If $\lambda_i > l$ for $i \in [m]$, $0 \leq \sum_i (\lambda_i - l)^{-1} \leq \epsilon_L$, and $1/\delta_L - \epsilon_L \geq 0$, then*

$$\frac{\sum_i (\lambda_i - l - \delta_L)^{-2}}{\sum_i (\lambda_i - l - \delta_L)^{-1} - \sum_i (\lambda_i - l)^{-1}} - \sum_i \frac{1}{\lambda_i - l - \delta_L} \geq \frac{1}{\delta_L} - \sum_i \frac{1}{\lambda_i - l}\,. \tag{3.18}$$

With this, we are ready to prove Lemma 3.11, which plays the same role in the proof of Theorem 3.1 as [3, Lem. 3.5] in the proof of [3, Thm. 3.1].

Lemma 3.11. *Let $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$ be a frame with frame bounds $0 < A \leq B < \infty$. Let further $\boldsymbol{A} \in \mathbb{C}^{m \times m}$ be Hermitian with $\sigma(\boldsymbol{A}) \subset (l, u)$ for $l, u \in \mathbb{R}$ and with corresponding potentials $\Phi_l(\boldsymbol{A})$, $\Phi^u(\boldsymbol{A})$. If the quantities δ_L, δ_U, ϵ_L, $\epsilon_U > 0$ satisfy the condition*

$$0 < \frac{B}{A}\Big(\frac{1}{\delta_U} + \epsilon_U\Big) \leq \frac{1}{\delta_L} - \kappa\epsilon_L\,, \quad \epsilon_L \geq \Phi_l(\boldsymbol{A})\,, \quad \epsilon_U \geq \Phi^u(\boldsymbol{A})\,, \tag{3.19}$$

where $\kappa = \kappa(A, B)$ is as in (3.1), then there exists an index $i \in [M]$ such that condition (3.17) in Corollary 3.9 is fulfilled for $\boldsymbol{y}^i$, and the indices with this property are precisely those where $L_{\boldsymbol{A}}(\boldsymbol{y}^i) \geq U_{\boldsymbol{A}}(\boldsymbol{y}^i)$ for $L_{\boldsymbol{A}}(\boldsymbol{y}^i) = L_{\boldsymbol{A}}(\boldsymbol{y}^i; l, \delta_L)$ and $U_{\boldsymbol{A}}(\boldsymbol{y}^i) = U_{\boldsymbol{A}}(\boldsymbol{y}^i; u, \delta_U)$ as in (3.12) and (3.15). The corresponding rank-1 updates $\boldsymbol{A}' := \boldsymbol{A} + t\boldsymbol{y}^i(\boldsymbol{y}^i)^*$ fulfill (3.11) with $l' = l + \delta_L$ and $u' = u + \delta_U$ for each $t > 0$ with

$$L_{\boldsymbol{A}}(\boldsymbol{y}^i) \geq 1/t \geq U_{\boldsymbol{A}}(\boldsymbol{y}^i)\,.$$

Proof. First note that, according to our assumptions, we have $\epsilon_L \geq \Phi_l(\boldsymbol{A})$ and $0 < \delta_L^{-1} - \kappa\epsilon_L$, which implies

$$\delta_L < (\kappa\epsilon_L)^{-1} \leq \frac{1}{\kappa}\Phi_l(\boldsymbol{A})^{-1} \leq \frac{1}{\kappa}\Delta_L\,, \tag{3.20}$$

where $\Delta_L = \lambda_{\min}(\boldsymbol{A}) - l$ and the last estimate is due to

$$\Phi_l(\boldsymbol{A}) = \sum_{i=1}^m \frac{1}{\lambda_i - l} \geq \frac{1}{\lambda_{\min}(\boldsymbol{A}) - l} = \Delta_L^{-1}\,.$$

In particular, $\delta_L < \Delta_L$ since $\kappa \geq 1$. Further $\delta_U > 0 > -\Delta_U$ for $\Delta_U = u - \lambda_{\max}(\boldsymbol{A})$, wherefore the quantities $L_{\boldsymbol{A}}(\boldsymbol{y}^i)$ and $U_{\boldsymbol{A}}(\boldsymbol{y}^i)$ are well-defined (see (3.12) and (3.15)).

Now, analogous to the proof of [3, Lem. 3.5], we aim to show

$$\sum_i L_{\boldsymbol{A}}(\boldsymbol{y}^i) \geq \sum_i U_{\boldsymbol{A}}(\boldsymbol{y}^i)\,. \tag{3.21}$$

On the one hand, using Lemma A.2, we have

$$\begin{aligned}
\sum_i U_{\boldsymbol{A}}(\boldsymbol{y}^i) &= \frac{\sum_i (\boldsymbol{y}^i)^*((u+\delta_U)\boldsymbol{I} - \boldsymbol{A})^{-2}\boldsymbol{y}^i}{\Phi^u(\boldsymbol{A}) - \Phi^{u+\delta_U}(\boldsymbol{A})} + \sum_i (\boldsymbol{y}^i)^*((u+\delta_U)\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{y}^i \\
&\leq B\left[\frac{\operatorname{tr}((u+\delta_U)\boldsymbol{I} - \boldsymbol{A})^{-2}}{\Phi^u(\boldsymbol{A}) - \Phi^{u+\delta_U}(\boldsymbol{A})} + \operatorname{tr}((u+\delta_U)\boldsymbol{I} - \boldsymbol{A})^{-1}\right] \\
&= B\left[\frac{\sum_i (u + \delta_U - \lambda_i)^{-2}}{\sum_i (u - \lambda_i)^{-1} - \sum_i (u + \delta_U - \lambda_i)^{-1}} + \Phi^{u+\delta_U}(\boldsymbol{A})\right].
\end{aligned}$$

The denominator of the first term in the brackets can be estimated as follows,

$$\sum_i \frac{1}{u - \lambda_i} - \sum_i \frac{1}{u + \delta_U - \lambda_i} = \sum_i \frac{\delta_U}{(u - \lambda_i)(u + \delta_U - \lambda_i)} > \sum_i \frac{\delta_U}{(u + \delta_U - \lambda_i)^2}\,.$$

Taking into account $\epsilon_U \geq \Phi^u(\boldsymbol{A}) > \Phi^{u+\delta_U}(\boldsymbol{A})$, we obtain altogether

$$\sum_i U_{\boldsymbol{A}}(\boldsymbol{y}^i) < B\Big(\frac{1}{\delta_U} + \epsilon_U\Big)\,. \tag{3.22}$$

On the other hand, again using Lemma A.2, we have

$$\begin{aligned}\sum_i L_{\boldsymbol{A}}(\boldsymbol{y}^i) &= \frac{\sum_i (\boldsymbol{y}^i)^*(\boldsymbol{A}-(l+\delta_L)\boldsymbol{I})^{-2}\boldsymbol{y}^i}{\Phi_{l+\delta_L}(\boldsymbol{A})-\Phi_l(\boldsymbol{A})} - \sum_i (\boldsymbol{y}^i)^*(\boldsymbol{A}-(l+\delta_L)\boldsymbol{I})^{-1}\boldsymbol{y}^i \\ &\geq A\bigg[\frac{\operatorname{tr}(\boldsymbol{A}-(l+\delta_L)\boldsymbol{I})^{-2}}{\Phi_{l+\delta_L}(\boldsymbol{A})-\Phi_l(\boldsymbol{A})}\bigg] - B\Big[\operatorname{tr}(\boldsymbol{A}-(l+\delta_L)\boldsymbol{I})^{-1}\Big] \\ &= A\bigg[\frac{\sum_i(\lambda_i-l-\delta_L)^{-2}}{\sum_i(\lambda_i-l-\delta_L)^{-1}-\sum_i(\lambda_i-l)^{-1}} - \sum_i \frac{1}{\lambda_i-l-\delta_L}\bigg] - (B-A)\Phi_{l+\delta_L}(\boldsymbol{A})\,,\end{aligned}$$

and the term in brackets can further be estimated by Lemma 3.10. The assumptions of this lemma are satisfied, in particular $\delta_L^{-1} - \epsilon_L \geq \delta_L^{-1} - \kappa\epsilon_L > 0$ due to $\kappa \geq 1$. Using (3.18) leads to

$$\sum_i L_{\boldsymbol{A}}(\boldsymbol{y}^i) \geq A\Big(\frac{1}{\delta_L} - \Phi_l(\boldsymbol{A})\Big) - (B-A)\Phi_{l+\delta_L}(\boldsymbol{A})\,. \tag{3.23}$$

We distinguish two cases, $A = B$ and $A < B$, to derive

$$\sum_i L_{\boldsymbol{A}}(\boldsymbol{y}^i) \geq A\Big(\frac{1}{\delta_L} - \kappa\epsilon_L\Big)\,. \tag{3.24}$$

If $A = B$ then $\kappa = 1$ and (3.24) follows directly from (3.23), since $\epsilon_L \geq \Phi_l(\boldsymbol{A})$. If $A < B$ the argument is a bit more involved. We then have $\kappa > 1$. From (3.20) we further deduce $\delta_L < \kappa^{-1}(\lambda_i - l)$ for $i = 1, \ldots, m$. This allows for the estimate

$$\Phi_{l+\delta_L}(\boldsymbol{A}) = \sum_i \frac{1}{\lambda_i - l - \delta_L} \leq \sum_i \frac{1}{\lambda_i - l}\Big(1 - \frac{1}{\kappa}\Big)^{-1} = \Phi_l(\boldsymbol{A})\frac{\kappa}{\kappa-1}\,.$$

Plugging this relation into (3.23) then also implies (3.24), namely

$$\sum_i L_{\boldsymbol{A}}(\boldsymbol{y}^i) \geq \frac{A}{\delta_L} - \Phi_l(\boldsymbol{A})\Big[A + (B-A)\frac{\kappa}{\kappa-1}\Big] \geq A\Big(\frac{1}{\delta_L} - \kappa\epsilon_L\Big)\,,$$

since κ as in (3.1) fulfills

$$A + (B-A)\frac{\kappa}{\kappa-1} = A\kappa\,.$$

Hence, we have finally established (3.22) and (3.24) and, in view of assumption (3.19), these two results yield (3.21).

As a consequence, there exists at least one $i \in [M]$ so that $L_{\boldsymbol{A}}(\boldsymbol{y}^i) \geq U_{\boldsymbol{A}}(\boldsymbol{y}^i)$. Furthermore, $U_{\boldsymbol{A}}(\boldsymbol{y}^i) > 0$ due to $\delta_U > 0$, which in turn implies $L_{\boldsymbol{A}}(\boldsymbol{y}^i) > 0$. Since also $\delta_L > 0$ by assumption, condition (3.17) in Corollary 3.9 is satisfied for $\boldsymbol{y}^i$. The rest follows from Corollary 3.9, which can indeed be applied due to $\delta_L < \Delta_L$ and $\delta_U > -\Delta_U$ as established at the beginning of this proof. ∎

With Lemma 3.11 we are now prepared to give a proof for Theorem 3.1.

3.5. Discussion of Algorithm 1 and proof of Theorem 3.1

We use the BSS algorithm (Algorithm 1) to construct the matrices $\boldsymbol{A}^{(k)}$ in (3.5). According to the algorithm, $\boldsymbol{A}^{(0)} = 0$ as it should be. Further, due to $l^{(0)} < 0 < u^{(0)}$, these initial values are valid spectral barriers for $\boldsymbol{A}^{(0)}$. The initial choice of barrier shifts fulfill $\delta_L^{(0)} > 0$ and $\delta_U^{(0)} > 0$. For the initial potentials $\epsilon_L^{(0)}$, $\epsilon_U^{(0)}$ we have

$$\epsilon_L^{(0)} = \Phi_{l^{(0)}}(\boldsymbol{A}^{(0)}) = -\frac{m}{l^{(0)}} \quad \text{and} \quad \epsilon_U^{(0)} := \Phi^{u^{(0)}}(\boldsymbol{A}^{(0)}) = \frac{m}{u^{(0)}}\,.$$

Hence, also $\epsilon_L^{(0)} > 0$, $\epsilon_U^{(0)} > 0$, and the quantities $\delta_L^{(0)}$, $\delta_U^{(0)}$, $\epsilon_L^{(0)}$, $\epsilon_U^{(0)}$ satisfy condition (3.19) of Lemma 3.11 with respect to $\boldsymbol{A}^{(0)}$. But even more holds true. For the further analysis of the algorithm, it is useful to note that for all $k \in \{0, \dots, n\}$

$$\frac{1}{\delta_L^{(k)}} - \kappa\epsilon_L^{(k)} - \frac{B}{A}\Big(\frac{1}{\delta_U^{(k)}} + \epsilon_U^{(k)}\Big) = \Delta\Big(1 - \frac{1}{\sqrt{b}}\Big) \geq 0\,. \tag{3.25}$$

This can be easily checked for $k = 0$ by a direct calculation. The definition of the variable barrier shifts $\delta_L^{(k)}$ and $\delta_U^{(k)}$ in line 5 of the algorithm further ensures that this expression does not change with k.

Let us now assume that for some arbitrary $k \in \{0, \dots, n-1\}$

$$\sigma(\boldsymbol{A}^{(k)}) \subset (l^{(k)}, u^{(k)}) \quad , \quad \delta_L^{(k)}, \delta_U^{(k)} > 0\,, \tag{3.26}$$

which was already checked for $k = 0$. Then, due to (3.25), condition (3.19) of Lemma 3.11 is fulfilled with respect to $\boldsymbol{A}^{(k)}$. Hence there is an index i such that condition (3.17) in Corollary 3.9 is satisfied for $\boldsymbol{y}^i$ and $L^{(k)}(\boldsymbol{y}^i) \geq U^{(k)}(\boldsymbol{y}^i)$, where (see line 9 of the BSS algorithm)

$$L^{(k)}(\boldsymbol{y}^i) = L_{\boldsymbol{A}^{(k)}}\big(\boldsymbol{y}^i; l^{(k)}, \delta_L^{(k)}\big) \quad \text{and} \quad U^{(k)}(\boldsymbol{y}^i) = U_{\boldsymbol{A}^{(k)}}\big(\boldsymbol{y}^i; u^{(k)}, \delta_U^{(k)}\big)\,.$$

Further, based on (3.25) and looking into the proof of Lemma 3.11, it holds

$$\frac{1}{A}\sum_{j=1}^{M}\big[L^{(k)}(\boldsymbol{y}^j) - U^{(k)}(\boldsymbol{y}^j)\big] = \frac{1}{\delta_L^{(k)}} - \kappa\epsilon_L^{(k)} - \frac{B}{A}\Big(\frac{1}{\delta_U^{(k)}} + \epsilon_U^{(k)}\Big) \geq \Delta\Big(1 - \frac{1}{\sqrt{b}}\Big)\,,$$

which implies that there is always even an index $i \in [M]$ such that $\boldsymbol{y}^i$ satisfies

$$L^{(k)}(\boldsymbol{y}^i) - U^{(k)}(\boldsymbol{y}^i) \geq \frac{\Delta}{M}\Big(1 - \frac{1}{\sqrt{b}}\Big)\,.$$

Hence, an index i which satisfies the selection condition in line 10 is found stably in each iteration of the BSS algorithm. According to Lemma 3.11, such an index in particular satisfies condition (3.17) of Corollary 3.9.

An update of $\boldsymbol{A}^{(k)}$ to $\boldsymbol{A}^{(k+1)}$ as in line 15 of the BSS algorithm thus yields a matrix with

$$\sigma(\boldsymbol{A}^{(k+1)}) \subset (l^{(k+1)}, u^{(k+1)}) \quad \text{and} \quad 0 < \epsilon_L^{(k+1)} \leq \epsilon_L^{(k)} \quad \text{and} \quad 0 < \epsilon_U^{(k+1)} \leq \epsilon_U^{(k)}\,.$$

From this and $\delta_L^{(k)}$, $\delta_U^{(k)} > 0$, we deduce

$$\delta_L^{(k+1)} \geq \delta_L^{(k)} > 0 \quad \text{and} \quad 0 < \delta_U^{(k+1)} \leq \delta_U^{(k)}\,. \tag{3.27}$$

Hence, (3.26) is fulfilled for $k+1$. Inductively, this proves that (3.26) is fulfilled for $k = 0, \dots, n$. In particular, $\sigma(\boldsymbol{A}^{(n)}) \subset (l^{(n)}, u^{(n)})$ and due to (3.27)

$$l^{(n)} = l^{(0)} + \sum_{k=0}^{\lceil bm \rceil - 1} \delta_L^{(k)} \geq l^{(0)} + bm\delta_L^{(0)} = -m\kappa\sqrt{b} + bm > 0\,,$$

where $b > \kappa^2$ was used in the last estimation step. Hence,

$$0 < l^{(n)} \quad \text{and} \quad l^{(n)} \boldsymbol{I} \preceq \sum_{i=1}^{M} \tilde{s}^{(i)} \boldsymbol{y}^i (\boldsymbol{y}^i)^* \preceq u^{(n)} \boldsymbol{I} \,.$$

The system $(\sqrt{\tilde{s}^{(i)}}\boldsymbol{y}^i)_{i=1}^M$, where the $\tilde{s}^{(i)}$ are the weights from the output line 17 before the rescaling, is thus a frame with bounds $0 < l^{(n)} \leq u^{(n)} < \infty$. Finally, we can estimate with (3.27)

$$\begin{aligned}\frac{u^{(n)}}{l^{(n)}} &= \frac{u^{(0)} + \sum_{i=0}^{n-1} \delta_U^{(i)}}{l^{(0)} + \sum_{i=0}^{n-1} \delta_L^{(i)}} \leq \frac{u^{(0)} + \lceil bm \rceil \delta_U^{(0)}}{l^{(0)} + \lceil bm \rceil \delta_L^{(0)}} = \frac{\delta_U^{(0)}}{\delta_L^{(0)}} + \frac{u^{(0)} - l^{(0)} \delta_U^{(0)} / \delta_L^{(0)}}{l^{(0)} + \lceil bm \rceil \delta_L^{(0)}} \\ &\leq \frac{\delta_U^{(0)}}{\delta_L^{(0)}} + \frac{u^{(0)} - l^{(0)} \delta_U^{(0)} / \delta_L^{(0)}}{l^{(0)} + bm \delta_L^{(0)}} = \frac{u^{(0)} + bm \delta_U^{(0)}}{l^{(0)} + bm \delta_L^{(0)}} = \frac{B}{A} \gamma (1 + \Delta) \,.\end{aligned}$$

The rescaled system $(\sqrt{s^{(i)}}\boldsymbol{y}^i)_{i=1}^M$ with the actual output weights $s^{(i)}$ from Algorithm 1 has thus frame bounds in the range $[A, B\gamma(1+\Delta)]$ due to

$$\begin{aligned} l^{(n)} \frac{1}{2} \Big(\frac{A}{l^{(n)}} + \frac{B\gamma(1+\Delta)}{u^{(n)}} \Big) &= \frac{1}{2} \Big(A + \frac{l^{(n)}}{u^{(n)}} B\gamma(1+\Delta) \Big) \geq A \,, \\ u^{(n)} \frac{1}{2} \Big(\frac{A}{l^{(n)}} + \frac{B\gamma(1+\Delta)}{u^{(n)}} \Big) &= \frac{1}{2} \Big(\frac{u^{(n)}}{l^{(n)}} A + B\gamma(1+\Delta) \Big) \leq B\gamma(1+\Delta) \,. \end{aligned}$$

This finishes the proof of Theorem 3.1, choosing $\Delta = 0$. ∎

4. Non-weighted subsampling of finite frames

We now turn to non-weighted versions of the subsampling strategies in Sections 2 and 3. Our approach is to give estimates on the occurring weights. In this way, we are able to save the lower frame bounds. For many applications those are the important ones as they ensure the stable reconstruction of any vector $\boldsymbol{a} \in \mathbb{C}^m$ from its frame coefficients $\langle \boldsymbol{a}, \boldsymbol{y}^i \rangle$. Results in this section will be of the following form:

Given vectors $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^M$, we seek inequalities of the type

$$\frac{1}{M} \sum_{i=1}^{M} |\langle \boldsymbol{a}, \boldsymbol{y}^i \rangle|^2 \leq \frac{C}{|J|} \sum_{i \in J} |\langle \boldsymbol{a}, \boldsymbol{y}^i \rangle|^2 \quad \text{for all} \quad \boldsymbol{a} \in \mathbb{C}^m \tag{4.1}$$

for $J \subset [M]$ and some fixed constant $C > 0$. If the initial $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^M$ satisfy a lower frame bound, (4.1) gives that the vectors $\boldsymbol{y}^i$, $i \in J$, satisfy a lower frame bound as well.

For the non-weighted version of the random subsampling in Theorem 2.1 the construction of $\tilde{\boldsymbol{Y}}$ is covered by Lemma 3.4. We obtain the following result with $|J| = \mathcal{O}(m \log m)$.

Theorem 4.1. *Let $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$ be a frame and $c, p, t \in (0,1)$ and $n \in \mathbb{N}$ be such that*

$$n \geq \frac{3}{ct^2} m \log \left(\frac{m}{p} \right) .$$

Drawing n indices $J \subset [M]$ (with duplicates) i.i.d. according to the discrete probability density $\varrho_i = (1 - c)/M + c \cdot \|\tilde{\boldsymbol{y}}^{\boldsymbol{i}}\|_2^2 / m$ gives

$$\frac{1}{M} \sum_{i=1}^{M} \big| \langle \boldsymbol{a}, \boldsymbol{y}^i \rangle \big|^2 \leq \frac{1}{(1-c)(1-t)} \frac{1}{|J|} \sum_{i \in J} \big| \langle \boldsymbol{a}, \boldsymbol{y}^i \rangle \big|^2 \quad \textit{for all} \quad \boldsymbol{a} \in \mathbb{C}^m$$

with probability exceeding $1 - p$.

Proof. Similar to Theorem 2.1, the result follows from Tropp's result in Lemma A.3. Here it is applied to the random rank-1 matrices $\boldsymbol{A}_i := \frac{1}{n}\varrho_i^{-1}\tilde{\boldsymbol{y}}^i \otimes \tilde{\boldsymbol{y}}^i$, where $\tilde{\boldsymbol{y}}^1, \dots, \tilde{\boldsymbol{y}}^M \in \mathbb{C}^m$ are the rows of the matrix $\tilde{\boldsymbol{Y}}$ obtained according to Lemma 3.4 from $\boldsymbol{Y}$, the analysis operator (2.1) of $(\boldsymbol{y}^i)_{i=1}^M$. The matrices $\boldsymbol{A}_i$ satisfy

$$\lambda_{\max}(\boldsymbol{A}_i) = \frac{1}{n}\varrho_i^{-1}\|\tilde{\boldsymbol{y}}^i\|_2^2 \le \frac{m}{cn}\,.$$

For $n = |J|$ independent copies $(\boldsymbol{A}_i)_{i\in J}$, due to the orthogonality of $\tilde{\boldsymbol{Y}}$, we further have

$$\sum_{i\in J}\mathbb{E}\boldsymbol{A}_i = \sum_{i\in J}\mathbb{E}\left(\frac{1}{n}\varrho_i^{-1}\tilde{\boldsymbol{y}}^i \otimes \tilde{\boldsymbol{y}}^i\right) = \sum_{i\in J}\frac{1}{n}\tilde{\boldsymbol{Y}}^*\tilde{\boldsymbol{Y}} = \boldsymbol{I}\,,$$

where $\boldsymbol{I}$ is the $m \times m$ dimensional identity matrix. Thus, $\mu_{\min} = \lambda_{\min}(\sum_{i\in J}\mathbb{E}\boldsymbol{A}_i) = 1$ and Lemma A.3 gives

$$\mathbb{P}\left(\lambda_{\min}\left(\frac{1}{n}\sum_{i\in J}\varrho_i^{-1}\tilde{\boldsymbol{y}}^i \otimes \tilde{\boldsymbol{y}}^i\right) \le 1-t\right) \le m\exp\left(-\frac{cn}{m}\frac{t^2}{3}\right)\,,$$

which is smaller than p by the assumption on n. Using $\varrho_i \ge (1-c)/M$, we obtain

$$\|\tilde{\boldsymbol{Y}}\boldsymbol{a}\|_2^2 = \|\boldsymbol{a}\|_2^2 \le \frac{1}{1-t}\frac{1}{n}\sum_{i\in J}\varrho_i^{-1}|\langle \boldsymbol{a}, \tilde{\boldsymbol{y}}^i\rangle|^2 \le \frac{M}{(1-c)(1-t)}\frac{1}{n}\|(\tilde{\boldsymbol{Y}}\boldsymbol{a})|_J\|_2^2$$

for all $\boldsymbol{a} \in \mathbb{C}^m$ with probability exceeding $1-p$. By the arguments in (3.9) and after we may replace $\tilde{\boldsymbol{Y}}$ with the original $\boldsymbol{Y}$ to obtain the assertion. ∎

Next, we assume that we have a Bessel sequence in $\mathbb{C}^m$ with elements that are norm-bounded from below. Applying Algorithm 2 ($\texttt{BSS}^\perp$) then yields a non-weighted inequality of type (4.1) with $|J| = \mathcal{O}(m)$. In Section 5 this algorithm is used in the experiments 1-3.

Lemma 4.2. *Let $(\boldsymbol{y}^i)_{i=1}^M$ be a Bessel sequence in $\mathbb{C}^m$, i.e., a set of vectors satisfying the upper bound in (1.1) for some $B > 0$. Further assume $M \ge m$ and $\|\boldsymbol{y}^i\|_2^2 \ge \beta m/M$ for some $\beta > 0$ and all $i \in [M]$. Then, for any $b > 1$, there exists a subset $J \subset [M]$ with $|J| \le \lceil bm\rceil$ (without duplicates) such that*

$$\frac{1}{M}\sum_{i=1}^M \left|\langle \boldsymbol{a}, \boldsymbol{y}^i\rangle\right|^2 \le \frac{B}{\beta}\frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}\frac{1}{m}\sum_{i\in J}\left|\langle \boldsymbol{a}, \boldsymbol{y}^i\rangle\right|^2 \quad \textit{for all} \quad \boldsymbol{a} \in \mathbb{C}^m\,.$$

Proof. Applying Algorithm 2 ($\texttt{BSS}^\perp$) to the sequence $(\boldsymbol{y}^i)_{i=1}^M$ yields weights $s_i \ge 0$, where $|\{i : s_i \ne 0\}| \le \lceil bm\rceil$. Recall that, by the discussion of Algorithm 2, its application to any input sequence $(\boldsymbol{y}^i)_{i=1}^M$ is possible provided $M \ge m$. We obtain (3.8). Taking into account the Bessel property of $(\boldsymbol{y}^i)_{i=1}^M$ and choosing $\Delta = 0$ in Algorithm 2 yields

$$\sum_{i=1}^M \left|\langle \boldsymbol{a}, \boldsymbol{y}^i\rangle\right|^2 \le \sum_{i\in J} s_i \left|\langle \boldsymbol{a}, \boldsymbol{y}^i\rangle\right|^2 \le \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}B\|\boldsymbol{a}\|_2^2 \tag{4.2}$$

for $J := \{i : s_i \ne 0\}$ and all $\boldsymbol{a} \in \mathbb{C}^m$. Setting $\boldsymbol{a} = \boldsymbol{y}^j$ for $j \in J$, we obtain by the assumption $\|\boldsymbol{y}^j\|_2^2 \ge \beta m/M$ and the upper estimate in (4.2)

$$s_j \le \frac{(\sqrt{b}+1)^2 B}{(\sqrt{b}-1)^2\|\boldsymbol{y}^j\|_2^2} \le \frac{B}{\beta}\frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}\frac{M}{m}\,.$$

Thus, by the lower estimate in (4.2), we obtain the assertion. ∎

The condition on the norms $\|\boldsymbol{y}^i\|_2$ in Lemma 4.2 can be dropped with a more elaborate subsampling strategy, `PlainBSS` (see below) instead of $\texttt{BSS}^\perp$. The 'preconditioning' in `PlainBSS` is based on Lemma 4.3 rather than Lemma 3.4. The final result is stated in Corollary 4.5. The price we pay for this is the dependence of the constant in terms of the oversampling factor b. It deteriorates to $(b-1)^{-3}$ while in the previous result it is $(b-1)^{-2}$.

Lemma 4.3. *Let $\boldsymbol{Y} \in \mathbb{C}^{M\times m}$ be a matrix and $K \in \{0,\dots,M\}$. Then there is a matrix $\tilde{\boldsymbol{Y}} \in \mathbb{C}^{M\times m'}$ with $m' \in \{K,\dots,K+m\}$ and rows $\tilde{\boldsymbol{y}}^1,\dots,\tilde{\boldsymbol{y}}^M \in \mathbb{C}^{m'}$ such that*

$$\mathrm{R}(\tilde{\boldsymbol{Y}}) \supset \mathrm{R}(\boldsymbol{Y}), \quad \tilde{\boldsymbol{Y}}^*\tilde{\boldsymbol{Y}} = \boldsymbol{I}, \quad \text{and} \quad \|\tilde{\boldsymbol{y}}^i\|_2^2 \geq \frac{K}{M},$$

where $\boldsymbol{I}$ is the $m'\times m'$ dimensional identity matrix.

Proof. Let us denote the columns of $\boldsymbol{Y}$ with $\boldsymbol{c}^1,\dots,\boldsymbol{c}^m$. Further define columns in $\mathbb{C}^M$ by

$$d^k = \frac{1}{\sqrt{M}}\Big[\exp\Big(2\pi \mathrm{i} k\frac{j}{M}\Big)\Big]_{j=1}^M$$

for $k=1,\dots,K$, which are the first K columns of a Fourier matrix. By construction the system $(\boldsymbol{d}^k)_{k=1}^K$ is orthonormal. It can hence be extended by appropriate vectors $\tilde{\boldsymbol{c}}^1,\dots,\tilde{\boldsymbol{c}}^l$ to an orthonormal basis of

$$\mathrm{span}\{\boldsymbol{d}^1,\dots,\boldsymbol{d}^K,\boldsymbol{c}^1,\dots,\boldsymbol{c}^m\}.$$

Those can be constructed e.g. via the Gram-Schmidt algorithm. Finally, we set up

$$\tilde{\boldsymbol{Y}} := \big[\,\boldsymbol{d}^1\,\big|\cdots\big|\,\boldsymbol{d}^K\,\big|\,\tilde{\boldsymbol{c}}^1\,\big|\cdots\big|\,\tilde{\boldsymbol{c}}^l\,\big] = \begin{bmatrix} \overline{(\tilde{\boldsymbol{y}}^1)^*} \\ \hline \vdots \\ \hline \overline{(\tilde{\boldsymbol{y}}^M)^*} \end{bmatrix} \in \mathbb{C}^{M\times(K+l)},$$

which fulfills the stated conditions. ∎

Theorem 4.4. *Let $(\boldsymbol{y}^i)_{i=1}^M$ be a sequence of vectors in $\mathbb{C}^m$ and $K \in \{0,\dots,M\}$. Then, for any $b>1$, a set of indices $J \subset [M]$ (without duplicates) can be constructed (in polynomial time) such that $|J| \leq \lceil b(K+m)\rceil$ and*

$$\frac{1}{M}\sum_{i=1}^M \big|\langle \boldsymbol{a},\boldsymbol{y}^i\rangle\big|^2 \leq \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}\frac{m}{K}\frac{1}{m}\sum_{i\in J}\big|\langle \boldsymbol{a},\boldsymbol{y}^i\rangle\big|^2 \quad \text{for all} \quad \boldsymbol{a}\in\mathbb{C}^m. \tag{4.3}$$

Proof. We construct the vectors $\tilde{\boldsymbol{y}}^1,\dots,\tilde{\boldsymbol{y}}^M$ according to Lemma 4.3. They form a tight frame in $\mathbb{C}^{m'}$ with $m' \in \{K,\dots,K+m\}$ and $\|\tilde{\boldsymbol{y}}^i\|_2^2 \geq \frac{K}{M}$ for all $i\in[M]$. We can thus apply Lemma 4.2 ($\texttt{BSS}^\perp$, which in effect is here `BSS`) with $B=1$. We obtain a subset $J\subset[M]$ with $|J| \leq \lceil bm'\rceil \leq \lceil b(K+m)\rceil$ (without duplicates) such that

$$\frac{1}{M}\sum_{i=1}^M \big|\langle \boldsymbol{a},\boldsymbol{y}^i\rangle\big|^2 \leq \frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}\frac{1}{K}\sum_{i\in J}\big|\langle \boldsymbol{a},\boldsymbol{y}^i\rangle\big|^2 \quad \text{for all} \quad \boldsymbol{a}\in\mathbb{C}^m$$

which finishes the proof. ∎

A result in terms of the 'real' oversampling factor b' in Theorem 4.4, determined by $\lceil b'm\rceil = \lceil b(K+m)\rceil$, is given in Corollary 4.5.

Corollary 4.5. *Let $\boldsymbol{y}^1,\dots,\boldsymbol{y}^M \in \mathbb{C}^m$ be vectors with $m\in\mathbb{N}$. Further, take $b' > 1+\frac{1}{m}$ and assume $M \geq \lceil b'm\rceil$. We then obtain indices $J'\subset[M]$ with $|J'| \leq \lceil b'm\rceil$ such that*

$$\frac{1}{M}\sum_{i=1}^M \big|\langle \boldsymbol{a},\boldsymbol{y}^i\rangle\big|^2 \leq 89\frac{(b'+1)^2}{(b'-1)^3}\frac{1}{m}\sum_{i\in J'}\big|\langle \boldsymbol{a},\boldsymbol{y}^i\rangle\big|^2 \quad \text{for all} \quad \boldsymbol{a}\in\mathbb{C}^m.$$

Proof. The idea is to apply Theorem 4.4 for specifically chosen $K \in \mathbb{N}$ and $b > 1$, such that $\lceil b(m+K) \rceil \leq \lceil b'm \rceil$ for the given b' and the prefactor in (4.3) becomes small. Theorem 4.4 yields the prefactor

$$\frac{(\sqrt{b}+1)^2}{(\sqrt{b}-1)^2}\frac{m}{K} = \frac{(\sqrt{b}+1)^4}{(b-1)^2}\frac{m}{K} \leq 4\frac{(b+1)^2}{(b-1)^2}\frac{m}{K} =: C(K)\,.$$

Choosing b and K such that $b' = b\frac{m+K}{m}$ gives

$$b = b'/(1+K/m)\,, \quad b+1 = (b'+1+K/m)/(1+K/m)\,, \quad b-1 = (b'-1-K/m)/(1+K/m)\,,$$

and hence

$$C(K) = 4\Big(\frac{b'+1+\frac{K}{m}}{b'-1-\frac{K}{m}}\Big)^2\frac{m}{K}\,. \tag{4.4}$$

We now choose $K^\star = \lceil \frac{(b'-1)m}{8} \rceil \in [M]$. Assuming $b' \geq 1 + 4/m$, we can then bound

$$\frac{b'-1}{8} \leq \frac{K^\star}{m} \leq \frac{b'-1}{4}\,.$$

Further, since $b' - 1 - K^\star/m > 0$, we arrive at the estimate

$$C(K^\star) \leq 4\Big(\frac{b'+1+(b'-1)/4}{b'-1-(b'-1)/4}\Big)^2\frac{8}{b'-1} = \frac{32}{b'-1}\Big(\frac{5b'/4+3/4)}{3(b'-1)/4}\Big)^2 \leq 32\Big(\frac{5}{3}\Big)^2\frac{(b'+1)^2}{(b'-1)^3} \leq 89\frac{(b'+1)^2}{(b'-1)^3}\,. \tag{4.5}$$

Next, we consider the cases $b' = 1+2/m$ and $b' = 1+3/m$ separately, where in both $K^\star = 1$. The associated b are given by $b = 1 + 1/(m+1)$ and $b = 1 + 2/(m+1)$. Further $1/m = (b'-1)/2$ and $1/m = (b'-1)/3$. Inserting these values into (4.4), we obtain estimates for $C(K^\star)$ as in (4.5). The prefactors, being 72 and 48, are even smaller than 89. Finally, to extend the estimate (4.5) to the whole range $b' > 1 + \frac{1}{m}$, note that the right-hand side of (4.5) is increasing for $b' \searrow 1$. Taking into account $\lceil b'm \rceil = m + k + 1$ for each $k \in \mathbb{N}$ and $b' \in (1 + \frac{k}{m}, 1 + \frac{k+1}{m}]$, we are finished. ∎

Building on the proof of Corollary 4.5, we now formulate Algorithm 3 (`PlainBSS`). Like Algorithm 1 (`BSS`) and Algorithm 2 ($\texttt{BSS}^\perp$), it is polynomial in time.

Algorithm 3 `PlainBSS`

Input:	Vectors $\boldsymbol{y}^1, \dots, \boldsymbol{y}^M \in \mathbb{C}^m$ with $m \in \mathbb{N}$ and $M \geq m+2$; Oversampling factor b' s.t. $m+2 \leq \lceil b'm \rceil \leq M$; Stability factor $\Delta \geq 0$.
Output:	Indices $J \subset [M]$ such that $\lvert J \rvert \leq \lceil b'm \rceil$ and $\frac{1}{M}\sum_{i=1}^M \lvert\langle \boldsymbol{a}, \boldsymbol{y}^i\rangle\rvert^2 \leq 89\frac{(b'+1)^2}{(b'-1)^3}\frac{1+\Delta}{m}\sum_{i\in J}\lvert\langle \boldsymbol{a}, \boldsymbol{y}^i\rangle\rvert^2$.

1: Compute $K^\star$ and b from b' as in the proof of Corollary 4.5.
2: Construct vectors $\tilde{\boldsymbol{y}}^1, \dots, \tilde{\boldsymbol{y}}^M \in \mathbb{C}^{m'}$ with $K = K^\star$ according to Lemma 4.3, where the initial $\boldsymbol{Y} \in \mathbb{C}^{M\times m}$ is the matrix (2.1) with rows $(\boldsymbol{y}^1)^*, \dots, (\boldsymbol{y}^M)^*$.
3: Apply Algorithm 1 (`BSS`) to $\tilde{\boldsymbol{y}}^1, \dots, \tilde{\boldsymbol{y}}^M$ with oversampling factor b and stability factor Δ to obtain weights $s_1, \dots, s_M$.
4: **return** indices $J := \{i : s_i \neq 0\}$.

For a better runtime, it might sometimes be advantageous to combine `BSS` subsampling with a preceding random subsampling step. Theorem 2.1 could be used, for instance, to quickly reduce the number of vectors to $\mathcal{O}(m\log(m))$ in case of very large M. In the following corollary such a two-step procedure is used to construct a unit-norm frame with very few (close to m) elements and well-behaved frame bounds. Here it is crucial that the `BSS` algorithm returns no duplicates, which is used in the proof.

Corollary 4.6. *Assume that the vectors $\boldsymbol{y}^1, \dots, \boldsymbol{y}^M \in \mathbb{C}^m, m \in \mathbb{N}$, form a tight frame and let $b' > 1 + \frac{1}{m}$. Further choose $p, t \in (0,1)$ and draw*

$$n := \left\lceil \frac{3}{t^2} m \log\left(\frac{2m}{p}\right) \right\rceil$$

indices $J \subset [M]$ (with duplicates) i.i.d. according to the discrete probability density $\varrho_i = \|\boldsymbol{y}^i\|_2^2 / \|\boldsymbol{Y}\|_F^2$. In case $n > \lceil b'm \rceil$, those can further be subsampled using `BSS` *(with oversampling factor b') giving $J' \subset J$ with $|J'| \le \lceil b'm \rceil$ and a unit-norm frame $(\boldsymbol{y}^i / \|\boldsymbol{y}^i\|_2)_{i \in J'}$ satisfying*

$$\frac{(1-t)(b'-1)^3}{89(b'+1)^2} \|\boldsymbol{a}\|_2^2 \le \sum_{i \in J'} \left| \left\langle \boldsymbol{a}, \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|_2} \right\rangle \right|^2 \le (1+t) \left\lceil \frac{3 \log(2m/p)}{t^2} \right\rceil \|\boldsymbol{a}\|_2^2$$

for all $\boldsymbol{a} \in \mathbb{C}^m$ with probability exceeding $1-p$. Otherwise, when $n \le \lceil b'm \rceil$, the frame $(\boldsymbol{y}^i / \|\boldsymbol{y}^i\|_2)_{i \in J'}$ with $J' = J$ already satisfies $|J'| \le \lceil b'm \rceil$ and (4.6) for all $\boldsymbol{a} \in \mathbb{C}^m$ with probability exceeding $1-p$.

Proof. By (2.2) and $(\boldsymbol{y}^i)_{i=1}^M$ forming a tight frame, we have $\|\boldsymbol{Y}\|_F^2 = mA$. By Theorem 2.1 we first obtain a subframe with $n = |J|$ elements such that

$$\frac{1-t}{m} \|\boldsymbol{a}\|_2^2 \le \frac{1}{n} \sum_{i \in J} \left| \left\langle \boldsymbol{a}, \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|_2} \right\rangle \right|^2 \le \frac{1+t}{m} \|\boldsymbol{a}\|_2^2 . \tag{4.6}$$

Next, if we apply Algorithm 3 (`PlainBSS`) to this subframe, we obtain $J' \subset J$ with $|J'| \le \lceil b'm \rceil$ such that

$$\frac{1}{n} \sum_{i \in J} \left| \left\langle \boldsymbol{a}, \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|_2} \right\rangle \right|^2 \le 89 \frac{(b'+1)^2}{(b'-1)^3} \frac{1}{m} \sum_{i \in J'} \left| \left\langle \boldsymbol{a}, \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|_2} \right\rangle \right|^2 ,$$

which is used in the lower frame bound. For the upper frame bound we use that J' has no duplicates, wherefore

$$\frac{1}{m} \sum_{i \in J'} \left| \left\langle \boldsymbol{a}, \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|_2} \right\rangle \right|^2 \le \left\lceil \frac{3 \log(2m/p)}{t^2} \right\rceil \frac{1}{n} \sum_{i \in J} \left| \left\langle \boldsymbol{a}, \frac{\boldsymbol{y}^i}{\|\boldsymbol{y}^i\|_2} \right\rangle \right|^2 .$$

Here, the relation of n and m was used. Last, we use the upper frame bound (4.6) and obtain the assertion. ∎

5. Numerical results

In this section we test the unweighted `BSS`, `BSS`$^\perp$, and `PlainBSS` (Algorithms 1, 2, and 3) in practice. Note that there are further recent attempts to reduce the sampling budget in least squares approximations in practice, see [14]. A survey on different probabilistic sampling strategies for sparse recovery of multivariate functions can be found in [2] (here especially Sec. 1.4 provides many further references). In addition, let us mention [1], where Adcock and Brugiapaglia give theoretical and empirical evidence of the near-optimal performance of simple Monte Carlo sampling for the recovery of smooth functions in high dimensions.

For the first three experiments, we use the rows of a d-dimensional Fourier matrix as initial frame, i.e.,

$$\boldsymbol{y}^i = \left[\frac{1}{\sqrt{M}} \exp(2\pi \mathrm{i} \langle \boldsymbol{k}, \boldsymbol{x}^i \rangle) \right]_{\boldsymbol{k} \in I} \quad \text{for} \quad i \in [M] , \tag{5.1}$$

where $I \subset \mathbb{Z}^d$ are $|I| = m$ frequencies determining the dimension of the frame elements and the nodes $\boldsymbol{X} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^M) \subset \mathbb{C}^d$ determine their number. In the experiments, we will have a look at different choices for these frequencies I and nodes $\boldsymbol{X}$. Note that construction (5.1) gives an equal-norm frame.

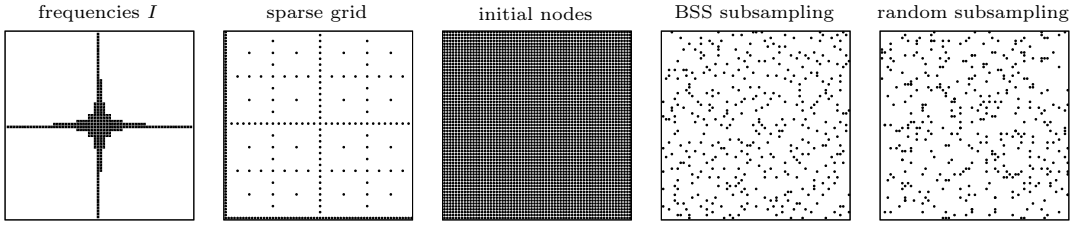

Figure 5.1: Two-dimensional experiment with sparse grid.

Experiment 1. We choose dimension $d = 2$ and, in the frequency domain, we use a so-called dyadic hyperbolic cross

$$I = H_R^d = \bigcup_{\substack{\boldsymbol{l} \in \mathbb{N}_0^d \\ \|\boldsymbol{l}\|_1 = R}} \hat{G}_{\boldsymbol{l}} \quad \text{with} \quad \hat{G}_{\boldsymbol{l}} = \bigtimes_{j=1}^{d} \hat{G}_{l_j} \quad \text{and} \quad \hat{G}_l = \mathbb{Z} \cap (-2^{l-1}, 2^{l-1}] \,,$$

which occurs naturally when approximating in Sobolev spaces with mixed smoothness, cf. [9]. Here, we use $R = 6$, which results in 256 frequencies. In spatial domain, the canonical candidate are sparse grids:

$$S_R^d = \bigcup_{\substack{j \in \mathbb{N}_0^d \\ \|\boldsymbol{l}\|_1 = R}} G_{\boldsymbol{l}} \quad \text{with} \quad G_{\boldsymbol{l}} = \bigtimes_{j=1}^{d} G_{l_j} \quad \text{and} \quad G_l = 2^{-l}(\mathbb{Z} \cap [0, 2^l)) \,.$$

Sparse grids have the minimal amount of nodes $n = m$ and reconstruct every frequency $\boldsymbol{k} \in H_R^d$, i.e., $A > 0$. Precise estimates on the frame bounds of these matrices are found in [18, Thm. 3.1].

To test the `BSS` algorithm we use an initial 65×65 equispaced grid

$$\boldsymbol{X} = \left\{ \frac{i}{\sqrt[d]{M}} : \boldsymbol{i} \in \{0, \dots, \sqrt[d]{M} - 1\}^d \right\} ,$$

which has $M = 4225$ nodes and is exact ($A = B = 1$) for the M frequencies $\boldsymbol{k} \in \{-(\sqrt[d]{M}-1)/2, \dots, (\sqrt[d]{M}-1)/2\}^d$, cf. [29, Sec. 4.4.3], in particular for the given dyadic hyperbolic cross. These initial frequencies and nodes can be seen in the first three graphs of Figure 5.1.

On the resulting frame constructed according to (5.1) we apply the unweighted `BSS` algorithm (discarding the weights s_i) with a target oversampling of $b = 1.5$ to obtain the subset J and compute the new frame bounds. For comparison, we draw a random subset (with replacement) of the same size and compute the frame bounds as well. Note, that we do not have theoretical bounds for these few random nodes. The results are depicted in the two rightmost graphs of Figure 5.1.

Since $\|\boldsymbol{y}^i\|_2^2 = m$, we obtain by Lemma 4.2 the theoretical lower frame bound $A = (\sqrt{b}-1)^2/(\sqrt{b}+1)^2 = 0.01021$ (cf. Lemma 4.2) where we observe $A = 0.06672$ in the experiment. This is better by a factor of 4 when compared to random subsampling, where we obtain a lower frame bound of $A = 0.02379$. Furthermore, the `BSS` algorithm gives a smaller upper frame constant than random subsampling, but this is not covered by our theory. The lower frame bound of the `BSS` subsampled nodes is bigger than the lower frame bound of the sparse grid. Even using the next biggest sparse grid with $n = 576$ nodes this still holds, as the frame bounds are $A = 0.06126$ and $B = 14.44698$.

Following [18] the frame bounds worsen for the sparse grids in higher dimensions. We conducted the same experiment in five dimensions with dyadic hyperbolic cross with $m = 1002$ frequencies with the following outcome:

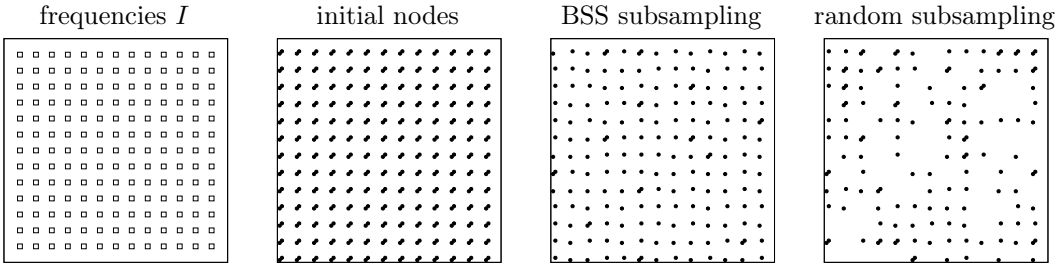

Figure 5.2: Two-dimensional experiment with frequencies on the grid

		b	n	A	B
sparse grids	S_5^5	1.00	1002	0.00009	89.5249
	S_5^6	2.96	2972	0.00063	74.5446
	S_5^7	8.46	8472	0.00158	63.5213
Frolov nodes		1.02	1021	0.00008	3.13560
		2.05	2051	0.08128	2.14287
		4.08	4093	0.37502	1.79493
BSS		1.01	1013	0.00012	3.69835
		1.50	1503	0.04333	2.99637
		2.00	2004	0.10659	2.61729
		2.50	2505	0.16325	2.39153
		2.96	2966	0.20790	2.24841
		3.50	3507	0.25682	2.10744
		4.08	4089	0.30101	2.00187

We cannot set $b = 1$ with the BSS algorithm, but already for $b = 1.01$ we achieve a slightly better lower frame bound A than for the sparse grid. When b increases is where the BSS algorithm shows its advantage as the frame bounds become progressively better.

Experiment 2. As the components of the frame elements $\boldsymbol{y}^i$ are continuous, we have similar frame elements for close nodes $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$. For the next experiment, we again are in dimension $d = 2$ and choose the full grid of frequencies $I = [-6, 6] \cap \mathbb{Z}^2$ with $m = 169$ frequencies for which the full grid of $13 \times 13 = 169$ nodes is barely exact. For the nodes we use two 13×13 point grids where one is slightly moved by $[0.01, 0.01]^\mathsf{T}$, which is depicted in the two leftmost plots of Figure 5.2. This setting is a union of two tight frames, itself a tight frame, where each element has a close duplicate which occur as pairs. A reasonable subsampling technique would pick at least one out of each pair. We set a target oversampling factor of $b = 1.1$ and apply the unweighted BSS algorithm and random subsampling for comparison. The results are depicted in the two rightmost graphs of Figure 5.2.

As in the first experiment, we have the theoretical lower frame bound $A = (\sqrt{b}-1)^2/(\sqrt{b}+1)^2 = 0.00057$ (cf. Lemma 4.2) where we observe $A = 0.49471$ in the experiment. For random subsampling we do not pick one frame element of each pair creating holes which spoil the lower frame bound. In fact, the subsampling is not even a frame anymore as $A = 0$.

Experiment 3. As our algorithms do not depend on the dimension, for the next experiment, we choose $d = 25$. In frequency domain we choose $m = 500$ random frequencies in $[-1000, 1000]^{25} \cap \mathbb{Z}^{25}$. In time domain we use two different choice:

- We use a full grid with $M = 2001^{25} > 10^{80}$ nodes, which is exact for all possible frequencies.

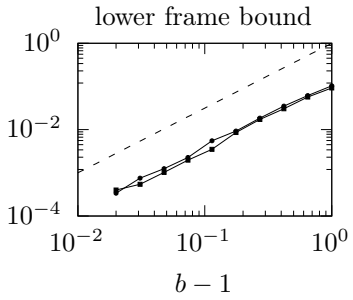

Figure 5.3: 25-dimensional experiment. Solid line with circles: lower frame bound A for the initial nodes being the full Grid. Solid line with squares: lower frame bound A for the initial nodes being drawn randomly. Dashed: $(b-1)^{3/2}$.

- We use $M = \lceil 6m\log(m)\rceil = 18\,644$ random nodes. In Lemma 6.1, we show that this gives frame bounds $A = 1/2$ and $B = 3/2$ with high probability.

For ten different choices of $b \in (1, 2]$ we use the unweighted `BSS` algorithm for the grid and unweighted `BSS`$^\perp$ for the random nodes. We compute the new frame bounds and count the inner iterations (i.i.) of the `BSS` algorithm in line 6. Further, we compute the theoretical frame bounds $1/B \cdot (\sqrt{b}-1)^2/(\sqrt{b}+1)^2$ from Lemma 4.2. The results are shown in the table below and Figure 5.3.

		Grid nodes $M = 2001^{25}$ $(b \approx 7 \cdot 10^{79})$ $A = B = 1$				**Random nodes** $M = 18\,644$ $(b \approx 37)$ $A = 0.70, B = 1.34$			
b	n	A	bound	B	i.i.	A	bound	B	i.i.
1.02	510	$3.72 \cdot 10^{-4}$	$2.45 \cdot 10^{-5}$	3.81	1.5	$2.70 \cdot 10^{-4}$	$1.83 \cdot 10^{-5}$	3.84	1.4
1.12	564	$5.59 \cdot 10^{-3}$	$8.02 \cdot 10^{-4}$	3.63	1.6	$4.68 \cdot 10^{-3}$	$5.99 \cdot 10^{-4}$	3.67	1.4
1.23	618	$1.46 \cdot 10^{-2}$	$2.67 \cdot 10^{-3}$	3.46	1.5	$1.26 \cdot 10^{-2}$	$2.00 \cdot 10^{-3}$	3.53	1.4
1.34	673	$2.63 \cdot 10^{-2}$	$5.33 \cdot 10^{-3}$	3.35	1.6	$2.28 \cdot 10^{-2}$	$3.98 \cdot 10^{-3}$	3.39	1.4
1.45	727	$3.92 \cdot 10^{-2}$	$8.58 \cdot 10^{-3}$	3.22	1.5	$3.56 \cdot 10^{-2}$	$6.40 \cdot 10^{-3}$	3.24	1.4
1.56	782	$5.14 \cdot 10^{-2}$	$1.23 \cdot 10^{-2}$	3.11	1.5	$4.89 \cdot 10^{-2}$	$9.15 \cdot 10^{-3}$	3.14	1.4
1.67	836	$6.63 \cdot 10^{-2}$	$1.63 \cdot 10^{-2}$	3.01	1.6	$5.77 \cdot 10^{-2}$	$1.21 \cdot 10^{-2}$	3.04	1.4
1.78	891	$7.90 \cdot 10^{-2}$	$2.05 \cdot 10^{-2}$	2.94	1.5	$7.02 \cdot 10^{-2}$	$1.53 \cdot 10^{-2}$	3.01	1.4
1.89	940	$9.02 \cdot 10^{-2}$	$2.49 \cdot 10^{-2}$	2.90	1.6	$7.96 \cdot 10^{-2}$	$1.86 \cdot 10^{-2}$	2.91	1.4
2.00	1000	$1.02 \cdot 10^{-1}$	$2.94 \cdot 10^{-2}$	2.82	1.6	$9.55 \cdot 10^{-2}$	$2.20 \cdot 10^{-2}$	2.83	1.4

The message of this experiment is twofold:

- The rate of A for $b \to 1$ is cubic in our theoretical results, cf. Lemma 4.2 and Theorem 4.4. In this experiment we observe the rate of $3/2$ which is even smaller that the bound for the weighted `BSS` algorithm, cf. Theorem 3.1.
- As the number of nodes in the grid is larger than the estimated number of atoms in the observable universe, we would expect a longer runtime for this example. The only difference in the computational effort could originate from the iterations in the inner loop of the `BSS` algorithm. From our theory we obtain M iterations in the worst case whereas we observe 1.5 iterations on average in both experiments.

Experiment 4. Here we deal with two-dimensional hyperbolic Chui-Wang wavelets $\psi_{\boldsymbol{j},\boldsymbol{k}}$, which are compactly supported and piecewise linear and $L_2([0,1]^d)$-normalized, see for instance [23] for the precise construction. We define the index sets

$$\mathcal{J}_n = \{(\boldsymbol{j},\boldsymbol{k}) \in \mathbb{N}_{-1}^d \times \mathbb{Z}^d : \boldsymbol{j} \geq -\mathbf{1}, |\boldsymbol{j}|_1 \leq N, \boldsymbol{k} \in I_{\boldsymbol{j}}\} \tag{5.2}$$

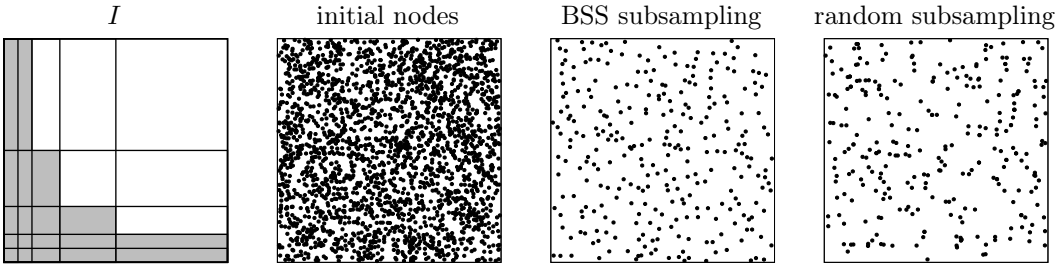

Figure 5.4: Two-dimensional hyperbolic wavelet transform

and

$$I_{\boldsymbol{j}} = \prod_{i=1}^{d} \begin{cases} \{0, 1, \dots 2^{j_i} - 1\} & \text{for } j_i \geq 0, \\ \{0\} & \text{for } j_i = -1. \end{cases}$$

The projection on the $\boldsymbol{j}$-component of this index set is displayed in the first picture in Figure 5.4 with $N = 3$. Drawing sufficiently many (M) nodes i.i.d. and uniformly at random ($M = \mathcal{O}(|\mathcal{J}_n| \log(|\mathcal{J}_n|))$) it has been shown in [23] that the corresponding frame $([\psi_{\boldsymbol{j},\boldsymbol{k}}(\boldsymbol{x}^i)]_{\boldsymbol{j},\boldsymbol{k}})_{i=1}^{M}$ has reasonably good frame bounds (see the second picture in Figure 5.4. In the previous experiments we only dealt with equal-norm frames. This is not given anymore in this particular frame such that we are forced to apply `PlainBSS` to extract a reasonable subframe with $b \approx 1.5$. The resulting nodes can be seen in the third picture of Figure 5.4.

The lower frame bound of the subsampled nodes can be estimated by Corollary 4.5: $A \leq \frac{0.01708(b-1)^3}{89(b+1)^2} \approx 3.84 \cdot 10^{-6}$ for $b = 1.5$. In practice we obtain a subsampled frame bound of $A = 3.21 \cdot 10^{-3}$, which indicates that the theoretical constants may be improved. Further, in the sanity check, `PlainBSS` is better by a factor of 10 when compared to random subsampling (last picture of Figure 5.4) and the upper frame bound does not differ much. Overall, this experiment demands for the tricky construction of Lemma 4.3 and shows its stable applicability.

6. Applications and discussion

Finally, we apply the subsampling results from the previous sections to the problem of L_2(-stable) recovery of multivariate complex-valued functions $f \colon D \to \mathbb{C}$. Those are assumed to be given on some measure space (D, ν) and the considered task shall be to recover f from sampling values

$$\boldsymbol{f}_n := (f(x^1), \dots, f(x^n)) \in \mathbb{C}^n \tag{6.1}$$

taken at certain sampling nodes

$$\boldsymbol{X}_n := (x^1, \dots, x^n) \in D^n \,. \tag{6.2}$$

In order to give sense to the point evaluation $f(x^i)$ (i.e. to ensure that it represents a continuous functional), we model f to belong either to a reproducing kernel Hilbert space (RKHS) $H(K)$ on D or to $\ell_\infty(D)$, the space of bounded functions on D. Since (6.1) and (6.2) is usually insufficient information for an exact reconstruction of f, we merely seek to find good approximants $\tilde{f}$ of f. The approximation shall take place in $L_2(D, \nu)$, the space of square-integrable complex-valued functions with

$$\langle f, g \rangle_{L_2} = \int_D f(x) \overline{g(x)} \, \mathrm{d}\nu(x) \quad \text{and} \quad \|f\|_{L_2} = \left(\int_D |f(x)|^2 \, \mathrm{d}\nu(x) \right)^{1/2} .$$

A possible way to recover f from the given data (6.1) and (6.2) is to apply a weighted least squares reconstruction operator

$$S^{\boldsymbol{X}_n}_{V_m,w_m} f := \operatorname*{arg\,min}_{g\in V_m} \sum_{i=1}^{n} w_m(x^i)|g(x^i) - f(x^i)|^2 \,, \tag{6.3}$$

for certain weights $w_m(x^i)$ and some m-dimensional reconstruction space $V_m \subset H(K) \subset L_2(D,\nu)$ (or, alternatively $V_m \subset \ell_\infty(D)$). Such operators perform well in many scenarios, depending on the utilized node set $\boldsymbol{X}_n$, the space V_m, and the weights $w_m(x^i)$. Of particular interest are plain least squares operators, where $w_m(x^i) = 1$ for $i \in [n]$. For those we will use the simpler notation $S^{\boldsymbol{X}_n}_{V_m}$.

To practically determine $S^{\boldsymbol{X}_n}_{V_m,w_m} f$, it is useful to employ a basis $(\eta_1,\dots,\eta_m)$ of V_m. The coefficients $\boldsymbol{c} \in \mathbb{C}^m$ of $S^{\boldsymbol{X}_n}_{V_m,w_m} f$ in this basis can then be obtained by solving

$$\boldsymbol{c} = \operatorname*{arg\,min}_{\boldsymbol{c}\in\mathbb{C}^m} \|\boldsymbol{L}_{n,m}\cdot\boldsymbol{c} - \boldsymbol{f}_n\|^2_{\ell_{2,w}} = \operatorname*{arg\,min}_{\boldsymbol{c}\in\mathbb{C}^m} \|\boldsymbol{W}_n(\boldsymbol{L}_{n,m}\cdot\boldsymbol{c} - \boldsymbol{f}_n)\|^2_{\ell_2} \,, \tag{6.4}$$

where $\boldsymbol{L}_{n,m} := (\eta_k(x^i))^{i=1\dots n}_{k=1\dots m} \in \mathbb{C}^{n\times m}$ and $\boldsymbol{W}_n := \operatorname{diag}(\sqrt{w_m(x^1)},\dots,\sqrt{w_m(x^n)}) \in \mathbb{C}^{n\times n}$. For this to make sense, the η_k are assumed to be proper functions, not equivalence classes, in $L_2(D,\nu)$ (see e.g. the explanation after (B.2)).

If the matrix $\widetilde{\boldsymbol{L}}_{n,m} := \boldsymbol{W}_n\boldsymbol{L}_{n,m}$ has full column rank, the solution of (6.4) is unique and can be expressed by

$$\boldsymbol{c} = (\widetilde{\boldsymbol{L}}_{n,m})^\dagger \boldsymbol{W}_n \mathbf{f}_n \,,$$

where $(\widetilde{\boldsymbol{L}}_{n,m})^\dagger$ is the Moore-Penrose pseudo-inverse of $\widetilde{\boldsymbol{L}}_{n,m}$ and has the explicit form

$$(\widetilde{\boldsymbol{L}}_{n,m})^\dagger = ((\widetilde{\boldsymbol{L}}_{n,m})^*\widetilde{\boldsymbol{L}}_{n,m})^{-1}(\widetilde{\boldsymbol{L}}_{n,m})^* \,. \tag{6.5}$$

6.1. Sampling recovery in spaces of finite measure

We first consider the case of reconstructing functions f from $\ell_\infty(D)$. We hereby assume that $L_2(D,\nu)$ is equipped with a finite measure ν. In this case, $\ell_\infty(D) \hookrightarrow L_2(D,\nu)$ and there is a constant $D_\nu > 0$ such that

$$\|f\|_{L_2(D,\nu)} \le D_\nu \|f\|_{\ell_\infty(D)} \quad \text{for all } f \in \ell_\infty(D) \,. \tag{6.6}$$

Our primary goal is now to derive unweighted (left) Marcinkiewicz-Zygmund inequalities for m-dimensional subspaces of $\ell_\infty(D)$. We let $m \in \mathbb{N}$ and $V_m := \operatorname{span}(\eta_k)_{k=1}^m$. The spanning set $(\eta_k)_{k=1}^m$ shall be a fixed orthonormal basis of V_m. Then we define a new sampling measure $d\mu$ as $d\mu := \varphi^{V_m}(\cdot)d\nu$ with

$$\varphi^{V_m}(x) := \frac{1}{2} + \frac{1}{2}\frac{\sum_{k=1}^m |\eta_k(x)|^2}{m} \tag{6.7}$$

and draw random sampling nodes according to this measure, which is independent of the chosen orthonormal basis since φ^{V_m} in (6.7) is unique up to ν-null sets. This density first appeared in [30].

For a sufficiently large number of sampling nodes we have the following result.

Lemma 6.1. *Let $p,t \in (0,1)$ and let $\widetilde{\boldsymbol{X}}_M = (\tilde{x}^i)_{i=1}^M \in D^M$ be M nodes drawn independently (with duplicates) according to the probability measure μ on D given by* (6.7)*. In case*

$$M \ge \frac{4}{t^2} m \log\left(\frac{m}{p}\right) \tag{6.8}$$

it holds

$$(1-t)\|\boldsymbol{a}\|_2^2 \le \frac{1}{M}\|\widetilde{\boldsymbol{L}}_{M,m}\boldsymbol{a}\|_2^2 \quad \textit{for all } \boldsymbol{a} \in \mathbb{C}^m$$

with probability exceeding $1-p$*, where*

$$\widetilde{\boldsymbol{L}}_{M,m} := \begin{pmatrix} [\eta_1/\sqrt{\varphi^{V_m}}](\tilde{x}^1) & \cdots & [\eta_m/\sqrt{\varphi^{V_m}}](\tilde{x}^1) \\ \vdots & & \vdots \\ [\eta_1/\sqrt{\varphi^{V_m}}](\tilde{x}^M) & \cdots & [\eta_m/\sqrt{\varphi^{V_m}}](\tilde{x}^M) \end{pmatrix}.$$

Proof. Let $\boldsymbol{u}^i$, $i \in [M]$, denote the rows of $\widetilde{\boldsymbol{L}}_{M,m}$ and define $\boldsymbol{A}_i := \frac{1}{M}\boldsymbol{u}^i \otimes \boldsymbol{u}^i$. Then we have $\lambda_{\max}(\boldsymbol{A}_i) = \frac{\|\boldsymbol{u}^i\|_2^2}{M} \leq \frac{2m}{M}$ due to $\varphi^{V_m} \geq \frac{1}{2m}\sum_{k=1}^m |\eta_k|^2$ (see (6.7)). The matrix

$$\boldsymbol{H}_m := \frac{1}{M}\widetilde{\boldsymbol{L}}_{M,m}^*\widetilde{\boldsymbol{L}}_{M,m} = \frac{1}{M}\sum_{i=1}^M \boldsymbol{u}^i \otimes \boldsymbol{u}^i = \sum_{i=1}^M \boldsymbol{A}_i$$

is further Hermitian positive semi-definite and fulfills $\mathbb{E}(\boldsymbol{H}_m) = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix in $\mathbb{C}^{m\times m}$. The latter follows from the orthogonality of the function system $(\eta_i/\sqrt{\varphi^{V_m}})_{i=1}^m$ in $L_2(D,\mu)$.

Lemma A.3, applied with $\mu_{\min} = \mu_{\max} = 1$ and $R = 2m/M$, now states that

$$\lambda_{\min}(\boldsymbol{H}_m) \leq 1 - t$$

with probability not more than $m\exp(-Mt^2/(4m))$. If we choose M according to (6.8), this then yields

$$\frac{1}{M}\big\|\widetilde{\boldsymbol{L}}_{M,m}\boldsymbol{w}\big\|_2^2 = \boldsymbol{w}^*\boldsymbol{H}_m\boldsymbol{w} \geq (1-t)\|\boldsymbol{w}\|_2^2$$

with probability exceeding $1-p$. ∎

In the following, let $b > 1 + \frac{1}{m}$ be a fixed parameter. Further, let $\widetilde{\boldsymbol{X}}_M = (\tilde{x}^i)_{i=1}^M$ be a node sequence sampled according to Lemma 6.1 fulfilling $M \geq \lceil bm \rceil$. Applying the `plainBSS` algorithm to $\widetilde{\boldsymbol{X}}_M$ (i.e. the rows of $\widetilde{\boldsymbol{L}}_{M,m}$) with respect to b yields an index set $J \subset [M]$ with $|J| \leq \lceil bm \rceil$. Selecting the corresponding nodes in $\widetilde{\boldsymbol{X}}_M$, we obtain a subsequence $\boldsymbol{X}_n = (x^i)_{i=1}^n \subset \widetilde{\boldsymbol{X}}_M$ with $n \leq \lceil bm \rceil$.

Theorem 6.2. *Let (D,ν) be a finite measure space and $p,t \in (0,1)$. Let further V_m be an m-dimensional subspace of $\ell_\infty(D)$ for fixed $m \in \mathbb{N}$. Let further $\widetilde{\boldsymbol{X}}_M = (\tilde{x}^i)_{i=1}^M \in D^M$ and $\boldsymbol{X}_n = (x^i)_{i=1}^n \in D^n$ denote the node sets constructed as above for $b > 1 + \frac{1}{m}$, with $M = M_{p,t}$ satisfying* (6.8) *and $n \leq \lceil bm \rceil \leq M$. Then with probability exceeding $1-p$ for all $f \in V_m$*

$$\|f\|_{L_2(D,\nu)}^2 \leq \frac{2}{M(1-t)}\sum_{i=1}^M |f(\tilde{x}^i)|^2 \leq \frac{178(b+1)^2}{(b-1)^3(1-t)}\frac{1}{m}\sum_{i=1}^n |f(x^i)|^2 . \tag{6.9}$$

Proof. Let $\boldsymbol{a} \in \mathbb{C}^m$ be the coefficient vector of f with respect to $(\eta_i)_{i=1}^m$. The first inequality follows from Lemma 6.1 and the fact that $\varphi^{V_m} \geq 1/2$. We have

$$\|f\|_{L_2(D,\nu)}^2 = \|\boldsymbol{a}\|_2^2 \leq \frac{\|\widetilde{\boldsymbol{L}}_{M,m}\boldsymbol{a}\|_2^2}{M(1-t)} = \frac{1}{M(1-t)}\sum_{i=1}^M \frac{|f(\tilde{x}^i)|^2}{\varphi^{V_m}(\tilde{x}^i)} \leq \frac{2}{M(1-t)}\sum_{i=1}^M |f(\tilde{x}^i)|^2 .$$

An application of Corollary 4.5 proves the second inequality. ∎

From this, we can directly derive a recovery result for functions $f \in \ell_\infty(D)$. Earlier versions of this result can be found in [8] and [34]. The relation between the L_2 recovery error and the ℓ_∞ best approximation has been first established in [8]. The main contribution here is that we prove the existence of a plain least squares recovery operator. In fact, the following theorem is a consequence of [34, Thm. 2.1] together with our Theorem 6.2 above. For the convenience of the reader we give a proof.

Theorem 6.3. *Let $V_m \subset \ell_\infty(D)$ with dimension $m \in \mathbb{N}$ and $\boldsymbol{X}_n = (x^i)_{i=1}^n$ be as above, with $n \leq \lceil bm \rceil$ and $b > 1 + \frac{1}{m}$, fulfilling (6.9) (with high probability). For any $f \in \ell_\infty(D)$ the plain least squares operator $S_{V_m}^{\boldsymbol{X}_n}$ recovers f in $L_2(D,\nu)$ with the following error*

$$\|f - S_{V_m}^{\boldsymbol{X}_n} f\|_{L_2(D,\nu)}^2 \leq C_\nu \frac{b^3}{(b-1)^3} e(f, V_m)_{\ell_\infty(D)}^2 \,,$$

with a constant $C_\nu > 0$ that only depends on ν, where

$$e(f, V_m)_{\ell_\infty(D)} := \inf_{g \in V_m} \|f - g\|_{\ell_\infty(D)} \,.$$

Proof. By the triangle inequality, we obtain for any $g \in V_m$

$$\|f - S_{V_m}^{\boldsymbol{X}_n} f\|_{L_2(D,\nu)}^2 = \|f - g\|_{L_2(D,\nu)}^2 + \|g - S_{V_m}^{\boldsymbol{X}_n} f\|_{L_2(D,\nu)}^2 \,.$$

The fact that ν is a finite measure gives $\|f - g\|_{L_2(D,\nu)} \leq D_\nu \|f - g\|_{\ell_\infty(D)}$ (see (6.6)). The second summand equals $\|S_{V_m}^{\boldsymbol{X}_n}(g - f)\|_{L_2(D,\nu)}$, which can be estimated by Theorem 6.2, namely

$$\begin{aligned}
\|S_{V_m}^{\boldsymbol{X}_n}(g-f)\|_{L_2(D,\nu)}^2 &\leq \frac{Cb^2}{(b-1)^3(1-t)} \frac{1}{m} \sum_{i=1}^n |(S_{V_m}^{\boldsymbol{X}_n}(f-g))(x^i)|^2 \\
&\leq \frac{2Cb^2}{(b-1)^3(1-t)} \frac{1}{m} \sum_{i=1}^n |(S_{V_m}^{\boldsymbol{X}_n}(f-g))(x^i) - (f-g)(x^i)|^2 + |(f-g)(x^i)|^2 \\
&\leq \frac{4Cb^2}{(b-1)^3(1-t)} \frac{1}{m} \sum_{i=1}^n |(f-g)(x^i)|^2 \leq \frac{4\tilde{C}b^3}{(b-1)^3(1-t)} \|f-g\|_{\ell_\infty}^2 \,.
\end{aligned}$$

From the second to the third line, we hereby used

$$\sum_{i=1}^n |(S_{V_m}^{\boldsymbol{X}_n}(f-g))(x^i) - (f-g)(x^i)|^2 \leq \sum_{i=1}^n |(f-g)(x^i)|^2 \,.$$

Choosing $g \in V_m$ such that $\|f - g\|_{\ell_\infty} \leq 2e(f, V_m)_{\ell_\infty}$ yields the result. ∎

At last, let us define the following quantity for a function class $F \subset \ell_\infty(D)$,

$$g_{n,m}^{\mathrm{ls}}(F, L_2(D,\nu)) := \inf_{\substack{V_m \subset \ell_\infty(D) \\ \dim V_m = m}} \inf_{\boldsymbol{X}_n = (x^1, \cdots, x^n) \in D^n} \sup_{f \in F} \|f - S_{V_m}^{\boldsymbol{X}_n} f\|_{L_2(D,\nu)} \,. \tag{6.10}$$

It measures the error of an optimal plain least squares algorithm of the above type, using n nodes and an m-dimensional reconstruction space. Our last result of this subsection, Corollary 6.4, compares this quantity with the Kolmogorov number

$$d_m(F, \ell_\infty(D)) := \inf_{\substack{W \subset \ell_\infty(D) \\ \dim(W) = m}} \sup_{f \in F} \min_{w \in W} \|f - w\|_{\ell_\infty(D)} \tag{6.11}$$

of the class F. It is a direct consequence of Theorem 6.3. The constant $C_\nu > 0$ only depends on ν.

Corollary 6.4. *Let F be a class of functions in $\ell_\infty(D)$. Then for $m \in \mathbb{N}$ and $b > 1 + \frac{1}{m}$*

$$g_{\lceil bm \rceil, m}^{\mathrm{ls}}(F, L_2(D,\nu)) \leq C_\nu \frac{b^{3/2}}{(b-1)^{3/2}} d_m(F, \ell_\infty(D)) \,.$$

This estimate improves on a recent result by Temlyakov [34], where the quantity $g_{\lceil bm \rceil, m}^{\mathrm{ls}}(F, L_2(D,\nu))$ is related to a modified Kolmogorov width which includes an additional restriction on the admissible subspaces in (6.11).

6.2. Sampling recovery in reproducing kernel Hilbert spaces

We next consider functions from a RKHS $H(K)$ with a finite trace kernel K. The RKHS is assumed to be compactly embedded into $L_2(D,\nu)$ via a Hilbert-Schmidt embedding $\mathrm{Id}_{K,\nu}$. The measure ν does not have to be a finite measure as in Subsection 6.1. In this setting, the number of non-zero singular values of $\mathrm{Id}_{K,\nu}$ is countable and, under the additional assumption that the subspace $\mathrm{Id}_{K,\nu}(H(K))$ of $L_2(D,\nu)$ is infinite-dimensional, also infinite. We thus have a sequence $(\sigma_k)_{k=1}^{\infty}$ of strictly positive singular numbers, which we order in descending order. The associated left and right singular functions shall be denoted by $(\eta_k)_{k=1}^{\infty}$ and $(e_k)_{k=1}^{\infty}$ (see Appendix B for more details). We follow the course of [11, 20, 19, 25, 27], where this setting was considered as well.

A natural reconstruction space in this scenario is $V_m := \mathrm{span}(\eta_k)_{k=1}^m$ spanned by the first m left singular functions associated to the m largest singular numbers of $\mathrm{Id}_{K,\nu}$. Appropriate nodes and weights for $S_{V_m,w_m}^{\boldsymbol{X}_n}$ are constructed in a two-step procedure, similar to the node generation in Subsection 6.1. The initial node set $\widetilde{\boldsymbol{X}}_M$ is drawn according to a probability measure $d\varrho_m := \varrho_m(\cdot)d\nu$ with

$$\varrho_m(x) := \frac{1}{2}\left(\frac{1}{m}\sum_{k=1}^{m}|\eta_k(x)|^2 + \frac{K(x,x)-\sum_{k=1}^m|e_k(x)|^2}{\int_D K(x,x)d\nu(x)-\sum_{k=1}^m\sigma_k^2}\right) \tag{6.12}$$

as density function. The corresponding weight function is

$$w_m : D\to[0,\infty) \quad , \quad w_m(x) := \begin{cases} \varrho_m(x)^{-1/2} & , \varrho_m(x)\neq 0\,, \\ 0 & , \varrho_m(x)=0\,. \end{cases}$$

The set $\boldsymbol{X}_n$ is then again obtained from $\widetilde{\boldsymbol{X}}_M$ by `PlainBSS`. Note that (6.12) is well-defined for all $m\in\mathbb{N}$ due to the positivity of the singular numbers.

Generation of sampling nodes.

Step 1 (Initial nodes). Let $m\in\mathbb{N}$ and $b>1+\frac{1}{m}$ and fix parameters $p,t\in(0,1)$. Then, with

$$M = \max\left\{\left\lceil\frac{4}{t^2}m\log\left(\frac{m}{p}\right)\right\rceil, \lceil bm\rceil\right\}, \tag{6.13}$$

an initial random sampling set

$$\widetilde{\boldsymbol{X}}_M := (\tilde{x}^1,\dots,\tilde{x}^M)\in D^M \tag{6.14}$$

is drawn, each node independently according to the measure $d\varrho_m$ with density (6.12).

For the associated weights almost surely $w_m(\tilde{x}^i)>0$. Furthermore, with probability exceeding $1-p$ the rows of the matrix $\frac{1}{\sqrt{M}}\widetilde{\boldsymbol{L}}_{M,m}$ where (cf. [27] replacing n with M)

$$\widetilde{\boldsymbol{L}}_{M,m} = \begin{pmatrix} [w_m\eta_1](\tilde{x}^1) & \cdots & [w_m\eta_m](\tilde{x}^1) \\ \vdots & & \vdots \\ [w_m\eta_1](\tilde{x}^M) & \cdots & [w_m\eta_m](\tilde{x}^M) \end{pmatrix},$$

represent a finite frame with lower frame bound $(1-t)$. This can be formulated as

$$(1-t)\|\boldsymbol{a}\|_2^2 \le \frac{1}{M}\left\|\widetilde{\boldsymbol{L}}_{M,m}\boldsymbol{a}\right\|_2^2 \quad \text{for all} \quad \boldsymbol{a}\in\mathbb{C}^m \tag{6.15}$$

and follows from Lemma 6.5 below.

Lemma 6.5. *Let $p,t \in (0,1)$ and let $\widetilde{\boldsymbol{X}}_M = (\tilde{x}^1,\dots,\tilde{x}^M) \in D^M$ be M nodes drawn independently (with duplicates) according to the measure $d\varrho_m$ given by* (6.12)*. In case*

$$M \geq \frac{4}{t^2} m \log\left(\frac{m}{p}\right)$$

(6.15) *holds with probability exceeding* $1-p$*.*

Proof. Based on Lemma A.3, analogous to the proof of Lemma 6.1. ∎

For any $\tilde{p} \in (0,1)$ we also have

$$\Big\| \frac{1}{M} (\widetilde{\boldsymbol{\Phi}}_{M,m})^* \widetilde{\boldsymbol{\Phi}}_{M,m} \Big\|_{2\to 2} \leq 2\sigma_{m+1}^2 + \frac{42}{M} \log\Big(2^{\frac{3}{4}} \frac{M}{\tilde{p}}\Big) \sum_{j=m+1}^{\infty} \sigma_j^2 \tag{6.16}$$

with probability exceeding $1-\tilde{p}$ for the infinite matrix $\widetilde{\boldsymbol{\Phi}}_{M,m}$ given by

$$\widetilde{\boldsymbol{\Phi}}_{M,m} := \begin{pmatrix} [w_m e_{m+1}](\tilde{x}^1) & [w_m e_{m+2}](\tilde{x}^1) & \dots \\ \vdots & \vdots & \\ [w_m e_{m+1}](\tilde{x}^M) & [w_m e_{m+2}](\tilde{x}^M) & \dots \end{pmatrix}.$$

This is a consequence of the following lemma, itself a corollary of [25, Prop. 3.8] (Proposition A.5 in the Appendix).

Lemma 6.6. *Let $\boldsymbol{u}^i$, $i \in [M]$, be i.i.d. random sequences from $\ell_2(\mathbb{N})$ with $M \in \mathbb{N}_{\geq 3}$. Let further $R > 0$ such that $\|\boldsymbol{u}^i\|_2 \leq R$ almost surely and $\mathbb{E}(\boldsymbol{u}^i \otimes \boldsymbol{u}^i) = \boldsymbol{\Lambda}$ for each $i \in [M]$. Then for each $\tilde{p} \in (0,1)$, with probability exceeding $1-\tilde{p}$,*

$$\Big\| \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{u}^i \otimes \boldsymbol{u}^i \Big\|_{2\to 2} \leq 2\|\boldsymbol{\Lambda}\|_{2\to 2} + \frac{21R^2}{M} \log\Big(\frac{2^{\frac{3}{4}} M}{\tilde{p}}\Big).$$

Proof. In [25, Prop. 3.8] (Proposition A.5) we can choose $r > 1$ such that $\tilde{p} = 2^{\frac{3}{4}} M^{1-r}$. Then $\log(2^{\frac{3}{4}} M/\tilde{p}) = r\log(M)$ and, noting $8\kappa^2 \leq 21$, we obtain

$$\Big\| \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{u}^i \otimes \boldsymbol{u}^i - \boldsymbol{\Lambda} \Big\|_{2\to 2} \leq \|\boldsymbol{\Lambda}\|_{2\to 2} + \frac{21R^2}{M} \log\Big(\frac{2^{\frac{3}{4}} M}{\tilde{p}}\Big)$$

with probability exceeding $1-\tilde{p}$. The triangle inequality finally yields the result. ∎

To prove (6.16) with Lemma 6.6 the rows of $\widetilde{\boldsymbol{\Phi}}_{M,m}$ are interpreted as sequences $\boldsymbol{u}^i \in \ell_2(\mathbb{N})$, where $\|\boldsymbol{u}^i\|_2^2 \leq 2\sum_{k\geq m+1} \sigma_k^2$ for each $i \in [M]$. Those then satisfy

$$\mathbb{E}(\boldsymbol{u}^i \otimes \boldsymbol{u}^i) = \operatorname{diag}(\sigma_{m+1}^2, \sigma_{m+2}^2, \dots) =: \boldsymbol{\Lambda}_m,$$

with $\|\boldsymbol{\Lambda}_m\|_{2\to 2} = \sigma_{m+1}^2$, and applying Lemma 6.6 with $R^2 = 2\sum_{k\geq m+1} \sigma_k^2$ yields (6.16).

Step 2 (Subsampling). The `PlainBSS` algorithm is used to determine a set of indices $J \subset [M]$ and a subset of nodes $\boldsymbol{X}_n \subset \widetilde{\boldsymbol{X}}_M$ of cardinality $|\boldsymbol{X}_n| = |J| \leq \lceil bm \rceil$, where $b > 1 + \frac{1}{m}$ as in Step 1. We then build a submatrix $\widetilde{\boldsymbol{L}}_{J,m}$ of $\widetilde{\boldsymbol{L}}_{M,m}$ by selecting the corresponding rows of $\widetilde{\boldsymbol{L}}_{M,m}$. From (6.15) and Corollary 4.5 we get, with probability exceeding $1-p$, that

$$(1-t)\|\boldsymbol{a}\|_2^2 \leq \frac{1}{M}\|\widetilde{\boldsymbol{L}}_{M,m}\boldsymbol{a}\|_2^2 \leq \frac{89(b+1)^2}{(b-1)^3}\frac{1}{m}\|\widetilde{\boldsymbol{L}}_{J,m}\boldsymbol{a}\|_2^2 \quad \text{for all } \boldsymbol{a} \in \mathbb{C}^m.$$

In particular, $\widetilde{\boldsymbol{L}}_{J,m}$ then has full rank and the norm of the Moore-Penrose pseudo-inverse $(\widetilde{\boldsymbol{L}}_{J,m})^\dagger = ((\widetilde{\boldsymbol{L}}_{J,m})^*\widetilde{\boldsymbol{L}}_{J,m})^{-1}(\widetilde{\boldsymbol{L}}_{J,m})^*$ (see (6.5)) fulfills the estimate

$$\|(\widetilde{\boldsymbol{L}}_{J,m})^\dagger\|_{2\to 2}^2 \leq \frac{89(b+1)^2}{(b-1)^3}\frac{1}{1-t}\frac{1}{m}. \tag{6.17}$$

Performance analysis. The sampling reconstruction operator $S^{\boldsymbol{X}_n}_{V_m,w_m}$ defined in (6.3), with nodes $\boldsymbol{X}_n$ constructed according to the previous paragraph, yields a near-optimal reconstruction performance (cf. [20, 19, 25, 27]), with the advantage of a precise control of the oversampling factor b as well as a polynomial-time semi-constructive node generation procedure. A modified subsampling procedure was presented recently by Dolbeault, Krieg, and M. Ullrich [11], leading to an optimal reconstruction rate (without the log-term). It is a refinement of the non-constructive Weaver subsampling from [27] and cannot be made constructive by our approach, since for that the upper frame bounds need to be preserved as well.

Theorem 6.7. *Let $H(K)$ be a reproducing kernel Hilbert space on a measurable domain (D,ν) with a positive semi-definite ν-measurable kernel $K : D \times D \to \mathbb{C}$ of finite trace*

$$\int_D K(x,x)d\nu(x) < \infty\,.$$

Further, assume that $H(K)$ is separable and that the canonical Hilbert-Schmidt embedding

$$\mathrm{Id}_{K,\nu} : H(K) \to L_2(D,\nu)$$

has infinite rank. In this setting, fix $m \in \mathbb{N}_{\geq 10}$, $p \in (0, \frac{2}{3})$, $t = \frac{2}{3}$, and construct a node set

$$\boldsymbol{X}_n \subset D \quad \textit{with} \quad |\boldsymbol{X}_n| \leq \lceil bm \rceil$$

for $b > 1 + \frac{1}{m}$ according to the procedure described in the previous paragraph (applying `PlainBSS` *on $M \geq \max\{\frac{4}{t^2} m \log(\frac{m}{p}), \lceil bm \rceil\}$ randomly drawn nodes). The reconstruction $S^{\boldsymbol{X}_n}_{V_m,w_m}$ given by* (6.3) *then performs as*

$$\sup_{\|f\|_{H(K)}\leq 1} \|f - S^{\boldsymbol{X}_n}_{V_m,w_m} f\|^2_{L_2(D,\nu)} \leq \frac{4827}{\min\{b-1,1\}} \frac{(b+1)^2}{(b-1)^2} \log\left(\frac{m}{p}\right)\left(\sigma^2_{m+1} + \frac{7}{m}\sum_{k=m+1}^{\infty} \sigma^2_k\right) \tag{6.18}$$

with probability exceeding $1 - \frac{3}{2}p$, where $(\sigma_k)_{k=1}^{\infty}$ is the sequence of decreasingly ordered, square-summable, non-trivial singular numbers of $\mathrm{Id}_{K,\nu}$.

Proof. First recall that, with certain probabilities, the initial nodes associated to $\boldsymbol{X}_n$,

$$\widetilde{\boldsymbol{X}}_M = (\tilde{x}^1, \dots, \tilde{x}^M)\,, \tag{6.19}$$

fulfill property (6.15) for the matrix $\widetilde{\boldsymbol{L}}_{M,m}$ and (6.16) for $\widetilde{\boldsymbol{\Phi}}_{M,m}$. To prove (6.18), we now take $f \in H(K)$ with $\|f\|_{H(K)} \leq 1$ and denote with $P_m : L_2(D,\nu) \to V_m$ the orthogonal projection onto the reconstruction space V_m of $S^{\boldsymbol{X}_n}_{V_m,w_m}$. Since the operator $S^{\boldsymbol{X}_n}_{V_m,w_m}$ acts as identity on V_m, whence $S^{\boldsymbol{X}_n}_{V_m,w_m} P_m f = P_m f$, we have

$$\begin{aligned}\|f - S^{\boldsymbol{X}_n}_{V_m,w_m} f\|^2_{L_2} &= \|f - P_m f\|^2_{L_2} + \|S^{\boldsymbol{X}_n}_{V_m,w_m}(f - P_m f)\|^2_{L_2} \\ &\leq \sigma^2_{m+1} + \|(\widetilde{\boldsymbol{L}}_{J,m})^{\dagger}\|^2_{2\to 2} \sum_{i=1}^{n} w_m(x^i)^2 |f(x^i) - P_m f(x^i)|^2\,. \end{aligned} \tag{6.20}$$

Clearly, we also have

$$\sum_{i=1}^{n} w_m(x^i)^2 |f(x^i) - P_m f(x^i)|^2 \leq \sum_{i=1}^{M} w_m(\tilde{x}^i)^2 |f(\tilde{x}^i) - P_m f(\tilde{x}^i)|^2\,, \tag{6.21}$$

where $(\tilde{x}^i)_{i=1}^M$ are the initial nodes (6.19). For $f \in \mathcal{N}(\mathrm{Id}_{K,\nu})$, the null-space of $\mathrm{Id}_{K,\nu}$, the right-hand side of (6.21) vanishes almost surely, due to the separability of $H(K)$. For general f, following the proof in [25, Thm. 5.1], almost surely

$$\sum_{i=1}^{M} w_m(\tilde{x}^i)^2 |f(\tilde{x}^i) - P_m f(\tilde{x}^i)|^2 \leq \|(\widetilde{\boldsymbol{\Phi}}_{M,m})^* \widetilde{\boldsymbol{\Phi}}_{M,m}\|_{2\to 2} \tag{6.22}$$

and according to (6.16)

$$\Big\|(\widetilde{\mathbf{\Phi}}_{M,m})^*\widetilde{\mathbf{\Phi}}_{M,m}\Big\|_{2\to 2} \le 2M\sigma_{m+1}^2 + 42\log\Big(2^{\frac{3}{4}}\frac{M}{\tilde{p}}\Big)\sum_{k=m+1}^{\infty}\sigma_k^2\,. \tag{6.23}$$

Altogether, the estimates (6.20)-(6.23) together with the norm estimate (6.17) for $(\widetilde{\boldsymbol{L}}_{J,m})^\dagger$ yield

$$\|f - S^{\boldsymbol{X}_n}_{V_m,w_m} f\|_{L_2}^2 \le \sigma_{m+1}^2 + \frac{89(b+1)^2}{(b-1)^3}\frac{1}{1-t}\frac{1}{m}\Big(2M\sigma_{m+1}^2 + 42\log\Big(2^{\frac{3}{4}}\frac{M}{\tilde{p}}\Big)\sum_{k=m+1}^{\infty}\sigma_k^2\Big)\,. \tag{6.24}$$

Let us proceed to bound the second summand in (6.24), which can be rewritten as

$$\frac{(b+1)^2}{(b-1)^3}\frac{178}{t^2(1-t)}\Big(\frac{Mt^2}{m}\sigma_{m+1}^2 + \frac{21t^2}{m}\log\Big(2^{\frac{3}{4}}\frac{M}{\tilde{p}}\Big)\sum_{k=m+1}^{\infty}\sigma_k^2\Big)\,. \tag{6.25}$$

Plugging in (6.13) for M with, at the moment, still arbitrary $p,t\in(0,1)$, we get

$$\frac{M-1}{m} \le \max\Big\{\frac{4}{t^2}\log\Big(\frac{m}{p}\Big), b\Big\} \quad \text{and thus} \quad \frac{Mt^2}{m} \le \frac{4M}{M-1}\max\Big\{\log\Big(\frac{m}{p}\Big), \frac{bt^2}{4}\Big\}\,. \tag{6.26}$$

Using $\log(\frac{m}{p}) \le \frac{m}{\mathrm{e}p}$ as well as $\log(xy) \le x\log(y)$ for $x\ge 1$ and $y\ge \mathrm{e}$, we also have

$$\log\Big(2^{\frac{3}{4}}\frac{M}{\tilde{p}}\Big) \le \log\Big(2^{\frac{3}{4}}\frac{Mm}{\tilde{p}(M-1)}\max\Big\{\log\Big(\frac{m}{p}\Big), \frac{bt^2}{4}\Big\}\frac{4}{t^2}\Big) \le \frac{M}{M-1}\log\Big(2^{\frac{3}{4}}\frac{4m}{\tilde{p}t^2}\max\Big\{\frac{m}{\mathrm{e}p}, \frac{bt^2}{4}\Big\}\Big)\,. \tag{6.27}$$

For $t=\frac{2}{3}$ the denominator $t^2(1-t)$ in (6.25) becomes maximal. We get $\frac{178}{t^2(1-t)} = 1201.5$. Together with (6.26) we obtain

$$\frac{(b+1)^2}{(b-1)^3}\frac{178}{t^2(1-t)}\frac{Mt^2}{m}\sigma_{m+1}^2 \le \frac{4806(b+1)^2}{(b-1)^3}\frac{M}{M-1}\max\Big\{\log\Big(\frac{m}{p}\Big), \frac{b}{9}\Big\}\sigma_{m+1}^2$$

for the first part in (6.25). For the second part, we have $21t^2 = \frac{28}{3}$ and due to (6.27)

$$21t^2\log\Big(2^{\frac{3}{4}}\frac{M}{\tilde{p}}\Big) \le \frac{56}{3}\frac{M}{M-1}\log\Big(2^{\frac{3}{8}}\sqrt{\frac{9}{\mathrm{e}}}\sqrt{\frac{m}{\tilde{p}}}\max\Big\{\sqrt{\frac{m}{p}}, \sqrt{\frac{\mathrm{e}b}{9}}\Big\}\Big)\,.$$

Putting $C := 2^{\frac{3}{8}}\sqrt{\frac{9}{\mathrm{e}}} \approx 2.359...$, this yields

$$\frac{(b+1)^2}{(b-1)^3}\frac{178}{t^2(1-t)}\log\Big(2^{\frac{3}{4}}\frac{M}{\tilde{p}}\Big)21t^2 \le \frac{4806(b+1)^2}{(b-1)^3}\frac{M}{M-1}\log\Big(C\sqrt{\frac{m}{\tilde{p}}}\max\Big\{\sqrt{\frac{m}{p}}, \sqrt{\frac{\mathrm{e}b}{9}}\Big\}\Big)\frac{14}{3}\,.$$

Altogether, we can thus bound (6.25) by

$$\frac{4806(b+1)^2}{(b-1)^3}\frac{M}{M-1}\Big(\max\Big\{\log\Big(\frac{m}{p}\Big), \frac{b}{9}\Big\}\sigma_{m+1}^2 + \log\Big(C\sqrt{\frac{m}{\tilde{p}}}\max\Big\{\sqrt{\frac{m}{p}}, \sqrt{\frac{\mathrm{e}b}{9}}\Big\}\Big)\frac{14}{3m}\sum_{k=m+1}^{\infty}\sigma_k^2\Big)\,. \tag{6.28}$$

We now choose $p = 2\tilde{p} \le \frac{2}{3}$. For this choice $M \ge \lceil 9\cdot 10\cdot\log(\frac{3\cdot 10}{2})\rceil = 244$ is always fulfilled, taking into account $t=\frac{2}{3}$, and thus $\frac{M}{M-1} \le \frac{244}{243}$. We hence arrive at

$$\frac{4826(b+1)^2}{(b-1)^3}\Big(\max\Big\{\log\Big(\frac{m}{p}\Big), \frac{b}{9}\Big\}\sigma_{m+1}^2 + \log\Big(C\sqrt{2}\sqrt{\frac{m}{p}}\max\Big\{\sqrt{\frac{m}{p}}, \sqrt{\frac{\mathrm{e}b}{9}}\Big\}\Big)\frac{14}{3m}\sum_{k=m+1}^{\infty}\sigma_k^2\Big)\,.$$

Further,

$$\sqrt{\frac{m}{p}}\max\Big\{\sqrt{\frac{m}{p}},\sqrt{\frac{\mathrm{e}b}{9}}\Big\}\leq\max\Big\{\frac{m}{p},\frac{\mathrm{e}b}{9}\Big\}\leq\max\Big\{\frac{m}{p},\frac{b}{3}\Big\}.$$

Let us also estimate

$$\log\Big(C\sqrt{2}\max\Big\{\frac{m}{p},\frac{b}{3}\Big\}\Big)=\Big(\frac{\log(C\sqrt{2})}{\log(\max\{m/p,b/3\})}+1\Big)\log\Big(\max\Big\{\frac{m}{p},\frac{b}{3}\Big\}\Big),$$

where, in view of $m\in\mathbb{N}_{\geq 10}$ and $p\leq\frac{2}{3}$,

$$\frac{\log(C\sqrt{2})}{\log(\max\{m/p,b/3\})}\leq\frac{\log(C\sqrt{2})}{\log(m/p)}\leq\frac{\log(\sqrt{2}2^{\frac{3}{8}}\sqrt{\frac{9}{\mathrm{e}}})}{\log(15)}=:F\approx 0.445\ldots.$$

When we plug this into (6.28), we finally arrive at the bound

$$\frac{4826(b+1)^2}{(b-1)^3}\Big(\max\Big\{\log\Big(\frac{m}{p}\Big),\frac{b}{9}\Big\}\sigma_{m+1}^2+\log\Big(\max\Big\{\frac{m}{p},\frac{b}{3}\Big\}\Big)\frac{14(F+1)}{3m}\sum_{k=m+1}^{\infty}\sigma_k^2\Big)$$

for (6.25), with $\frac{14}{3}(F+1)\approx 6.744...<7$.

This proves (6.18) with probability at least $1-p-\tilde{p}=1-\frac{3}{2}p$. ∎

Remark 6.8. *In the proof of Theorem 6.7 we chose $p=2\tilde{p}\leq\frac{2}{3}$. Of course, other choices are possible here. We would like to detail one particular scenario. First, recall Step 1 of the node generation process of $\boldsymbol{X}_n$. Here, condition* (6.15) *of $\widetilde{\boldsymbol{L}}_{M,m}$ could in fact be deterministically checked after the probabilistic generation of the initial nodes $\widetilde{\boldsymbol{X}}_M$ in* (6.14)*. Redrawing these nodes until this condition is fulfilled, which happens with high probability in polynomial time, we thus generate $\widetilde{\boldsymbol{X}}_M$ which satisfies* (6.15) *for $t=\frac{2}{3}$ and where M is as in* (6.13) *for $p\sim 1$. In practice, we could take for example (cf.* (6.13)*)*

$$M=\max\{\lceil 9m\log(m)\rceil+1,\lceil bm\rceil\}.$$

For this M the success probability of (6.15) *at each draw is strictly positive. Utilizing the reconstruction operator $S_{V_m,w_m}^{\boldsymbol{X}_n}$ with a* BSS*-downsampled node set $\boldsymbol{X}_n$ derived from such a (repeatedly redrawn) set of nodes $\widetilde{\boldsymbol{X}}_M$ then yields*

$$\sup_{\|f\|_{H(K)}\leq 1}\|f-S_{V_m,w_m}^{\boldsymbol{X}_n}f\|_{L_2(D,\nu)}^2\leq\frac{4831}{\min\{b-1,1\}}\frac{(b+1)^2}{(b-1)^2}\Big(\log(m)\sigma_{m+1}^2+\log\Big(\frac{m}{\sqrt{\tilde{p}}}\Big)\frac{7}{m}\sum_{k=m+1}^{\infty}\sigma_k^2\Big)$$

with probability exceeding $1-\tilde{p}$ for each $\tilde{p}\in(0,1)$. The proof is analogous to the proof of Theorem 6.7.

Acknowledgement

The authors would like to thank Moritz Moeller, who helped to implement the BSS algorithm and its modifications. They would also like to thank Daniel Potts, Vladimir N. Temlyakov, and André Uschmajew for fruitful discussions and David Krieg, Stefan Kunis, and Mario Ullrich for their comments and questions during the (online) school/conference 'Sampling Recovery and Related Problems' in May 2021. They are also very thankful for the valuable comments by Matthieu Dolbeault, which, in particular, resulted in an improved version of Lemma 4.3. Next to that, Felix Bartel would like to thank the Deutscher Akademischer Austauschdienst (DAAD) for funding his research scholarship.

References

[1] B. Adcock and S. Brugiapaglia. Is Monte Carlo a bad sampling strategy for learning smooth functions in high dimensions? *arXiv:math/2208.09045v2*, 2022.

[2] B. Adcock, J. M. Cardenas, N. Dexter, and S. Moraga. Towards optimal sampling for learning sparse approximation in high dimensions. *arXiv:math/2202.02360*, 2022.

[3] J. D. Batson, D. A. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. In *STOC'09—Proceedings of the 2009 ACM International Symposium on Theory of Computing*, pages 255–262. ACM, New York, 2009.

[4] J. R. Bunch and C. P. Nielsen. Updating the singular value decomposition. *Numer. Math.*, 31(2):111–129, 1978/79.

[5] P. G. Casazza, M. Fickus, and D. G. Mixon. Auto-tuning unit norm frames. *Appl. Comput. Harmon. Anal.*, 32(1):1–15, Jan. 2012.

[6] P. G. Casazza and J. Kovačević. Equal-norm tight frames with erasures. *Adv. Comput. Math.*, 18(2/4):387–430, 2003.

[7] P. G. Casazza and G. Kutyniok, editors. *Finite Frames*. Birkhäuser Boston, 2013.

[8] A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *SMAI J. Comput. Math.*, 3:181–203, 2017.

[9] D. Dũng, V. N. Temlyakov, and T. Ullrich. *Hyperbolic Cross Approximation*. Advanced Courses in Mathematics. CRM Barcelona. Birkhäuser/Springer, 2019.

[10] F. Dai, A. Prymak, A. Shadrin, V. Temlyakov, and S. Tikhonov. Entropy numbers and Marcinkiewicz-type discretization. *J. Funct. Anal.*, 281(6):109090, 2021.

[11] M. Dolbeault, D. Krieg, and M. Ullrich. A sharp upper bound for sampling numbers in L_2. *arXiv:math/2204.12621v2*, 2022.

[12] X. Dong and M. Rudelson. Approximately Hadamard matrices and Riesz bases in random frames. *arXiv:math/2207.07523*, 2022.

[13] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72:341–366, 1952.

[14] C. Haberstich, A. Nouy, and G. Perrin. Boosted optimal weighted least-squares. *Math. Comp.*, to appear.

[15] N. J. A. Harvey and N. Olver. Pipage rounding, pessimistic estimators and matrix concentration. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 926–945. ACM, New York, 2014.

[16] D. A. Harville. *Matrix algebra from a statistician's perspective*. Springer-Verlag, New York, 1997.

[17] M. Hein and O. Bousquet. Kernels, associated structures and generalizations. Technical Report 127, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2004.

[18] L. Kämmerer and S. Kunis. On the stability of the hyperbolic cross discrete Fourier transform. *Numer. Math.*, 117(3):581–600, 2011.

[19] L. Kämmerer, T. Ullrich, and T. Volkmer. Worst-case recovery guarantees for least squares approximation using random samples. *Constr. Approx.*, 54(2):295–352, 2021.

[20] D. Krieg and M. Ullrich. Function values are enough for L_2-approximation. *Found. Comput. Math.*, 21(4):1141–1151, 2021.

[21] D. Krieg and M. Ullrich. Function values are enough for L_2-approximation: Part II. *J. Complexity*, to appear.

[22] I. V. Limonova and V. N. Temlyakov. On sampling discretization in L_2. *arXiv:math/2009.10789v1*, 2020.

[23] L. Lippert, D. Potts, and T. Ullrich. Fast hyperbolic wavelet regression meets ANOVA. *arXiv e-prints*, pages 1–50, 2021.

[24] A. W. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families II: Mixed characteristic polynomials and the Kadison-Singer problem. *Ann. of Math. (2)*, 182(1):327–350, 2015.

[25] M. Moeller and T. Ullrich. L_2-norm sampling discretization and recovery of functions from RKHS with finite trace. *Sampling Theory, Signal Processing, and Data Analysis*, 19(2), July 2021.

[26] M. Moonen, P. Van Dooren, and J. Vandewalle. A singular value decomposition updating algorithm for subspace tracking. *SIAM J. Matrix Anal. Appl.*, 13(4):1015–1038, 1992.

[27] N. Nagel, M. Schäfer, and T. Ullrich. A new upper bound for sampling numbers. *Found. Comp. Math.*, Apr. 2021.

[28] S. Nitzan, A. Olevskii, and A. Ulanovskii. Exponential frames on unbounded sets. *Proc. Amer. Math. Soc.*, 144(1):109–118, 2016.

[29] G. Plonka, D. Potts, G. Steidl, and M. Tasche. *Numerical Fourier Analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2018.

[30] K. Pozharska and T. Ullrich. A note on sampling recovery of multivariate functions in the uniform norm. *SINUM*, to appear.

[31] M. Rudelson and R. Vershynin. Sampling from large matrices: an approach through geometric functional analysis. *J. ACM*, 54(4):Art. 21, 19, 2007.

[32] D. A. Spielman and N. Srivastava. Graph sparsification by effective resistances. *SIAM J. Comput.*, 40(6):1913–1926, 2011.

[33] I. Steinwart and C. Scovel. Mercers theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constr. Approx.*, 35, 2012.

[34] V. N. Temlyakov. On optimal recovery in L_2. *J. Complexity*, to appear.

[35] V. N. Temlyakov and T. Ullrich. Bounds on Kolmogorov widths and sampling recovery for classes with small mixed smoothness. *J. Complexity*, 67:101575, 2021.

[36] V. N. Temlyakov and T. Ullrich. Approximation of functions with small mixed smoothness in the uniform norm. *J. Approx. Theory*, 277:105718, 2022.

[37] J. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comp. Math.*, 12(4):389–434, 2011.

[38] N. Weaver. The Kadison-Singer problem in discrepancy theory. *Discrete Math.*, 278(1–3):227–239, 2004.

Appendix A. Matrix theory

A basic tool in Subsection 3.3 is the matrix determinant lemma (cf. [3, Lem. 2.2]). The complex version reads as follows.

Lemma A.1 (Matrix determinant lemma)**.** *If $\boldsymbol{A} \in \mathbb{C}^{m\times m}$ is nonsingular and $\boldsymbol{v} \in \mathbb{C}^m$ is a vector, then*

$$\det(\boldsymbol{A} + \boldsymbol{v}\boldsymbol{v}^*) = \det(\boldsymbol{A})(1 + \boldsymbol{v}^*\boldsymbol{A}^{-1}\boldsymbol{v})\,.$$

Lemma A.1 is a direct consequence of the Sherman-Morrison formula (see e.g. [16])

$$(\boldsymbol{A} + \boldsymbol{v}\boldsymbol{v}^*)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{v}\boldsymbol{v}^*\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^*\boldsymbol{A}^{-1}\boldsymbol{v}}\,,$$

which holds under the same assumptions as in Lemma A.1.

Another basic result, needed in the proof of Lemma 3.11, is the following statement.

Lemma A.2. *Let $(\boldsymbol{y}^i)_{i=1}^M \subset \mathbb{C}^m$ be a frame with frame bounds $0 < A \le B < \infty$. Let further $\boldsymbol{M} \in \mathbb{C}^{m\times m}$ be a positive semi-definite Hermitian matrix. Then*

$$A\,\mathrm{tr}(\boldsymbol{M}) \le \sum_{i=1}^M (\boldsymbol{y}^i)^*\boldsymbol{M}\boldsymbol{y}^i \le B\,\mathrm{tr}(\boldsymbol{M})\,.$$

Proof. First, observe that for an arbitrary matrix $\boldsymbol{M} \in \mathbb{C}^{m\times m}$

$$\sum_{i=1}^M (\boldsymbol{y}^i)^*\boldsymbol{M}\boldsymbol{y}^i = \sum_{i=1}^M \mathrm{tr}\left((\boldsymbol{y}^i)^*\boldsymbol{M}\boldsymbol{y}^i\right) = \sum_{i=1}^M \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{y}^i(\boldsymbol{y}^i)^*\right) = \mathrm{tr}\left(\boldsymbol{M}\Big(\sum_{i=1}^M \boldsymbol{y}^i(\boldsymbol{y}^i)^*\Big)\right).$$

Since $\boldsymbol{Y} := \sum_{i=1}^M \boldsymbol{y}^i(\boldsymbol{y}^i)^*$ is positive-definite Hermitian with $\sigma(\boldsymbol{Y}) \subset [A, B]$, there exists a unitary matrix $\boldsymbol{U}$ such that $\boldsymbol{D} := \boldsymbol{U}\boldsymbol{Y}\boldsymbol{U}^{-1}$ is diagonal with entries in the range $[A, B]$. We can hence conclude

$$\sum_{i=1}^M (\boldsymbol{y}^i)^*\boldsymbol{M}\boldsymbol{y}^i = \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{Y}\right) = \mathrm{tr}\left(\boldsymbol{U}\boldsymbol{M}\boldsymbol{Y}\boldsymbol{U}^{-1}\right) = \mathrm{tr}\left(\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{-1}\boldsymbol{D}\right).$$

Under the assumption that $\boldsymbol{M}$ is positive semi-definite Hermitian the transformation $\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{-1}$ is also positive semi-definite Hermitian. In particular, its diagonal entries are all nonnegative real numbers. As a consequence, we obtain the assertion

$$A\,\mathrm{tr}\left(\boldsymbol{M}\right) = A\,\mathrm{tr}\left(\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{-1}\right) \le \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{Y}\right) \le B\,\mathrm{tr}\left(\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{-1}\right) = B\,\mathrm{tr}\left(\boldsymbol{M}\right).$$

■

Appendix A.1. Concentration results for random matrices

Here we give some concentration inequalities which enable us to control the spectrum of sums of Hermitian positive semi-definite rank-1 matrices. In the finite case, these sums take the form

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{u}^i \otimes \boldsymbol{u}^i \quad \text{for a sequence of random vectors } \boldsymbol{u}^i \in \mathbb{C}^m,\ i \in [n]\,. \tag{A.1}$$

Recall the notation $[n] = \{1, \dots, n\}$. The basic assumption is always that the vectors $\boldsymbol{u}^i$ are drawn i.i.d. according to some probability distribution and that a uniform bound $\|\boldsymbol{u}^i\|_2 \le M$ is satisfied almost surely for every $i \in [n]$. Putting $\boldsymbol{A}_i := \frac{1}{n}\boldsymbol{u}^i \otimes \boldsymbol{u}^i$, we obtain a sequence $(\boldsymbol{A}_i)_{i=1}^n$ of i.i.d. Hermitian positive semi-definite random matrices which satisfy $\lambda_{\max}(\boldsymbol{A}_i) \le \frac{M^2}{n}$ almost surely. In this situation, a matrix Chernoff inequality proved by Tropp [37, Thm. 1.1] can be applied. We derive the following form.

Lemma A.3 (Matrix Chernoff, cf. [37, Thm. 1.1]). *For a sequence $(\boldsymbol{A}_i)_{i=1}^n \subset \mathbb{C}^{m\times m}$ of independent, Hermitian, positive semi-definite random matrices satisfying $\lambda_{\max}(\boldsymbol{A}_i) \leq R$ almost surely it holds*

$$\begin{aligned}\mathbb{P}\Big(\lambda_{\min}\Big(\sum_{i=1}^n \boldsymbol{A}_i\Big) \leq (1-t)\mu_{\min}\Big) &\leq m \exp\Big(-\frac{\mu_{\min}}{R}(t+(1-t)\log(1-t))\Big)\\ &\leq m\exp\Big(-\frac{\mu_{\min}t^2}{2R}\Big)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\Big(\lambda_{\max}\Big(\sum_{i=1}^n \boldsymbol{A}_i\Big) \geq (1+t)\mu_{\max}\Big) &\leq m \exp\Big(-\frac{\mu_{\max}}{R}(-t+(1+t)\log(1+t))\Big)\\ &\leq m\exp\Big(-\frac{\mu_{\max}t^2}{3R}\Big)\end{aligned}$$

for $t \in [0,1]$, where $\mu_{\min} \coloneqq \lambda_{\min}(\sum_{i=1}^n \mathbb{E}\boldsymbol{A}_i)$ and $\mu_{\max} \coloneqq \lambda_{\max}(\sum_{i=1}^n \mathbb{E}\boldsymbol{A}_i)$.

Proof. The first estimates are provided by [37, Thm. 1.1]. Based on the Taylor expansion

$$(1+t)\log(1+t) = t + \sum_{k=2}^{\infty} \frac{(-1)^k}{k(k-1)} t^k\,,$$

which holds true for $t \in [-1,1]$, we can further derive the inequalities

$$t + (1-t)\log(1-t) = \sum_{k=2}^{\infty} \frac{1}{k(k-1)} t^k \geq \frac{t^2}{2}$$

and

$$-t + (1+t)\log(1+t) = \sum_{k=2}^{\infty} \frac{(-1)^k}{k(k-1)} t^k \geq \frac{t^2}{2} - \frac{t^3}{6} \geq \frac{t^2}{3}$$

for the range $t \in [0,1]$. ∎

A concentration inequality for the case when the vectors $\boldsymbol{u}^i$ in (A.1) are infinite dimensional is given in [25, Thm. 1.1]. Here it is assumed that the $\boldsymbol{u}^i$ are i.i.d. random sequences from $\ell_2(\mathbb{N})$. Let us recite this result.

Theorem A.4 ([25, Thm. 1.1]). *Let $\boldsymbol{u}^i, i \in [n]$, be i.i.d. random sequences from $\ell_2(\mathbb{N})$. Let further $n \geq 3$, $M > 0$ such that $\|\boldsymbol{u}^i\|_2 \leq M$ almost surely and $\mathbb{E}(\boldsymbol{u}^i \otimes \boldsymbol{u}^i) = \boldsymbol{\Lambda}$ for $i \in [n]$ with $\|\boldsymbol{\Lambda}\|_{2\to 2} \leq 1$. Then*

$$\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{u}^i \otimes \boldsymbol{u}^i - \boldsymbol{\Lambda}\Big\|_{2\to 2} \geq t\Big) \leq 2^{\frac{3}{4}} n \exp\Big(-\frac{t^2 n}{21M^2}\Big)\,.$$

A useful rephrasing of Theorem A.4 is given by [25, Prop. 3.8]. It is as follows.

Proposition A.5 ([25, Prop. 3.8]). *Let $\boldsymbol{u}^i, i \in [n]$, be i.i.d. random sequences from $\ell_2(\mathbb{N})$. Let further $n \geq 3$, $r > 1$, $M > 0$ such that $\|\boldsymbol{u}^i\|_2 \leq M$ almost surely and $\mathbb{E}(\boldsymbol{u}^i \otimes \boldsymbol{u}^i) = \boldsymbol{\Lambda}$ for all $i \in [n]$. Then*

$$\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{u}^i \otimes \boldsymbol{u}^i - \boldsymbol{\Lambda}\Big\|_{2\to 2} \geq F\Big) \leq 2^{\frac{3}{4}}\, n^{1-r}\,,$$

where $F \coloneqq \max\Big\{\frac{8r\log n}{n} M^2\nu^2, \|\boldsymbol{\Lambda}\|_{2\to 2}\Big\}$ and $\nu = \frac{1+\sqrt{5}}{2}$.

Appendix B. The considered RKHS setting

In Subsection 6.2 we consider functions from a RKHS $H(K)$ on a non-empty measure space (D, ν). In the sequel, this setting is analyzed in more detail. First note that the space $H(K)$, as a RKHS, consists of proper point-wise defined functions. By definition, it is associated with a positive semi-definite Hermitian kernel $K : D \times D \to \mathbb{C}$ such that the reproducing property

$$f(x) = \langle f, K(\cdot, x) \rangle_{H(K)} \tag{B.1}$$

holds true for all $f \in H(K)$ and $x \in D$. Due to this property, sampling is not only a well-defined but even continuous operation in $H(K)$ (see e.g. [17]).

For our analysis of L_2-approximation, an embedding relation between $H(K)$ and $L_2(D, \nu)$ is crucial. It is guaranteed by the presumed finite trace of K, namely

$$\mathrm{tr}(K) := \int_D K(x, x)\, d\nu(x) < \infty\,.$$

Under such a condition, see [17] and [33, Lem. 2.3], for every $f \in H(K)$

$$\|f\|_2^2 = \int_D |\langle f, K(\cdot, x)\rangle|^2\, d\nu(x) \leq \int_D \|f\|_{H(K)}^2 \|K(\cdot, x)\|_{H(K)}^2\, d\nu(x) = \|f\|_{H(K)}^2 \cdot \mathrm{tr}(K)\,.$$

As a consequence, there is a compact embedding

$$\mathrm{Id}_{K,\nu} : H(K) \hookrightarrow L_2(D, \nu)\,,$$

which can be shown to be even Hilbert-Schmidt (see [33, Lem. 2.3]). Here, it is important to note that, in contrast to $H(K)$, the elements of $L_2(D, \nu)$ are not functions but ν-equivalence classes, with two functions $f, \tilde{f} : D \to \mathbb{C}$ considered ν-equivalent if $f(x) = \tilde{f}(x)$ for ν-almost every $x \in D$. $\mathrm{Id}_{K,\nu}$ may hence not be injective. In fact, $\mathrm{Id}_{K,\nu}$ is the restriction of Id_ν to $H(K)$, where Id_ν is the map that assigns to every $f : D \to \mathbb{C}$ the corresponding ν-equivalence class. Depending on the measure ν, the null-space $\mathcal{N}(\mathrm{Id}_{K,\nu})$ can thus be non-trivial. The choice $\nu = 0$ illustrates this, where $\mathcal{N}(\mathrm{Id}_{K,\nu}) = H(K)$.

Without loss of generality, to simplify the considerations in Subsection 6.2, it is further assumed that the subspace $\mathrm{Id}_{K,\nu}(H(K))$ of $L_2(D, \nu)$ is infinite dimensional (i.e. that $\mathrm{Id}_{K,\nu}$ has infinite rank). Under this condition, the sequence $(\sigma_k)_{k\in\mathbb{N}}$ of strictly positive singular numbers associated to $\mathrm{Id}_{K,\nu}$ is countably infinite. For this, note that $\mathrm{Id}_{K,\nu}(H(K))$ is a separable subspace of $L_2(D, \nu)$ due to the compactness of $\mathrm{Id}_{K,\nu}$.

We now fix orthonormal systems $(\eta_k)_{k\in\mathbb{N}} \subset L_2(D, \nu)$ and $(e_k)_{k\in\mathbb{N}} \subset H(K)$ of associated left and right singular functions such that

$$e_k = \sigma_k \eta_k \quad \text{for all} \quad k \in \mathbb{N}\,. \tag{B.2}$$

Whereas each e_k represents a point-wise function on D, the left singular functions η_k, as elements of $L_2(D, \nu)$, refer to ν-equivalence classes. However, we choose for each η_k the specific representative e_k/σ_k so that both systems $(\eta_k)_{k\in\mathbb{N}}$ and $(e_k)_{k\in\mathbb{N}}$ are comprised of proper functions satisfying (B.2) in a point-wise sense.

Let us next ask for basis properties of these systems. Clearly, $(\eta_k)_{k\in\mathbb{N}}$ is a basis for $\mathrm{Id}_{K,\nu}(H(K))$. The system $(e_k)_{k\in\mathbb{N}}$, on the other hand, is usually not a basis for $H(K)$ since it only corresponds to the non-trivial singular numbers of $\mathrm{Id}_{K,\nu}$. Under additional restrictions ensuring $\mathcal{N}(\mathrm{Id}_{K,\nu}) = \{0\}$ it would be, e.g. if the kernel K is continuous and bounded (i.e. a Mercer kernel). Generally, $H(K)$ decomposes in the form

$$H(K) = \overline{\mathrm{span}}\{e_1, e_2, \ldots\} \oplus_{\mathrm{orth}}^H \mathcal{N}(\mathrm{Id}_{K,\nu})\,.$$

A useful representation of K in terms of the functions $(e_k)_{k\in\mathbb{N}}$ can be obtained as follows. For each $y \in D$, first expand the function $K(\cdot, y) \in H(K)$ in the form

$$K(\cdot, y) = \sum_{k\in\mathbb{N}} c_k e_k + r_y$$

with an associated function $r_y \in \mathcal{N}(\mathrm{Id}_{K,\nu})$ and convergence of the sum in $H(K)$. For the coefficients calculate with (B.1)

$$c_k = \langle K(\cdot, y), e_k \rangle_{H(K)} = \overline{e_k(y)}\,.$$

Since convergence in $H(K)$ entails point-wise convergence, we obtain the representation

$$K(x, y) = \sum_{k \in \mathbb{N}} e_k(x)\overline{e_k(y)} + r_y(x)\,.$$

In case that $H(K)$ is separable, which is assumed in Theorem 6.7, it is shown in [17] and [33, Cor. 3.2] that $r_x(x)$ vanishes ν-almost everywhere. We hence have $K(x, x) = \sum_{k \in \mathbb{N}} |e_k(x)|^2$ for ν-almost every $x \in D$. This is a crucial ingredient in the proof of Theorem 6.7, see also [25, Sec. 4] and [19, Sec. 2].