

# THE AZ ALGORITHM FOR LEAST SQUARES SYSTEMS WITH A KNOWN INCOMPLETE GENERALIZED INVERSE\*

VINCENT COPPÉ<sup>†</sup>, DAAN HUYBRECHS<sup>‡</sup>, ROEL MATTHYSEN<sup>†</sup>, AND MARCUS WEBB<sup>‡</sup>

**Abstract.** We introduce an algorithm for the least squares solution of a rectangular linear system  $Ax = b$ , in which  $A$  may be arbitrarily ill-conditioned. We assume that a complementary matrix  $Z$  is known such that  $A - AZ^*A$  is numerically low rank. Loosely speaking,  $Z^*$  acts like a generalized inverse of  $A$  up to a numerically low-rank error. We give several examples of  $(A, Z)$  combinations in function approximation, where we can achieve high-order approximations in a number of non-standard settings: the approximation of functions on domains with irregular shapes, weighted least squares problems with highly skewed weights, and the spectral approximation of functions with localized singularities. The algorithm is most efficient when  $A$  and  $Z^*$  have fast matrix-vector multiplication and when the numerical rank of  $A - AZ^*A$  is small.

**Key words.** least squares problems, ill-conditioned systems, randomized linear algebra, frames, low-rank matrix approximation

**AMS subject classifications.** Primary, 65F20; Secondary, 65F05, 68W20

**DOI.** 10.1137/19M1306385

**1. Introduction.** The topic of this paper arose from the study of function approximation using frames [2, 1]. Frames are sets of functions that, unlike a basis, may be redundant. This leads to improved flexibility in a variety of settings which we illustrate below, but it also comes at a cost: the approximation problem is equivalent to solving a linear system  $Ax = b$  that is, owing to the redundancy of the frame, ill-conditioned. The central result of [2, 1] is that the function at hand can nevertheless be approximated to high accuracy via a regularized least squares approximation. Thus, the linear system is rectangular,  $A \in \mathbb{C}^{M \times N}$  with  $M > N$ , and the cost of a direct solver scales as  $\mathcal{O}(MN^2)$ , where typically  $M$  is at least linear in  $N$ .

The particular type of frame that has received the most attention in this context corresponds to Fourier extension or Fourier continuation [3, 4, 12], where a smooth but nonperiodic function on a domain  $\Omega$  is approximated by a Fourier series on a larger bounding box. Here, one can think of the redundancy as corresponding to different extensions of  $f$  from  $\Omega$  to the bounding box. Fast algorithms for the construction of Fourier extension approximations were proposed by Lyon in [14] for univariate problems and by Matthysen and Huybrechs in [15, 16] for more general problems in one and two dimensions. These algorithms have the form of the proposed AZ algorithm in the special case where  $Z = A$ .

Thus, the AZ algorithm is a generalization of the algorithms proposed for Fourier extension in [14, 15, 16]. In this paper we analyze the algorithm from the point of view of linear algebra, rather than approximation theory. We supply error estimates when using techniques from randomized linear algebra for the first step of the AZ

\*Received by the editors December 12, 2019; accepted for publication (in revised form) June 15, 2020; published electronically August 27, 2020.

<https://doi.org/10.1137/19M1306385>

**Funding:** The work of the fourth author was supported by a postdoctoral fellowship from FWO Research Foundation Flanders.

<sup>†</sup>Department of Computer Science, KU Leuven, Leuven, Belgium (vincent.coppe@cs.kuleuven.be, daan.huybrechs@cs.kuleuven.be, roelmatthysen@gmail.com).

<sup>‡</sup>Department of Mathematics, University of Manchester, Manchester, UK (marcus.webb@manchester.ac.uk).

algorithm [11]. In contrast, the previous analysis in [15, 16] has focused exclusively on studying the numerical rank of this system, in the specific setting of Fourier extension problems. We characterize the scope of the algorithm and provide several examples in different settings. In our analysis we focus on the residual of the linear system, as its size is directly related to the approximation error in the examples given.

There is a rich body of literature in numerical linear algebra on solution methods for least squares problems. We refer the reader to the standard references [10, 13]. We note that direct solvers typically exhibit cubic complexity in the dimension of the problem: the aim of the AZ algorithm is to reduce that complexity by exploiting the specific structure of certain least squares problems. This leads to an algorithm with several steps, in which existing methods for least squares problems can be used for systems with lower rank. In particular we study SVD- and QR-based methods for step 1 of the algorithm. We briefly comment on the applicability of iterative solvers for least squares problems, such as LSQR [17] and LSMR [9], near the end of the paper.

The structure of the paper is as follows. We formulate the algorithm in section 2 and state some general algebraic properties. We analyze the residual of solvers for low-rank systems based on randomized SVD or randomized QR in section 3. After the analysis, we illustrate the algorithms with a sequence of examples: Fourier extension approximation in section 4, approximation using weighted bases in section 5, spline-based extension problems in section 6, and weighted least squares problems in section 7. A Jupyter notebook containing the Julia code of these examples is found in [7]. We end the paper with some concluding remarks in section 8.

**2. The AZ algorithm.** Given a linear system  $Ax = b$  with  $A \in \mathbb{C}^{M \times N}$  and an additional matrix  $Z \in \mathbb{C}^{M \times N}$ , the AZ algorithm consists of three steps.

---

**Algorithm 2.1** The AZ algorithm

---

**Input:**  $A, Z \in \mathbb{C}^{M \times N}$ ,  $b \in \mathbb{C}^M$

**Output:**  $x \in \mathbb{C}^N$  such that  $Ax \approx b$  in a least squares sense

- 1: Solve  $(I - AZ^*)Ax_1 = (I - AZ^*)b$
  - 2:  $x_2 \leftarrow Z^*(b - Ax_1)$
  - 3:  $x \leftarrow x_1 + x_2$
- 

It is stated that the output solves  $Ax \approx b$  in a least squares sense, though the algorithm does not specify which matrix  $Z$  to use, nor which solver to use in step 1, and of course the properties of the output depend on that. We will give precise statements in the analysis that follows for particular choices. Overall, the intention is to choose  $Z$  such that the matrix  $(I - AZ^*)A$  approximately has low rank. Since  $(I - AZ^*)A = A - AZ^*A$ , this property can be thought of as  $Z^*$  being an incomplete generalized inverse of  $A$ . We analyze a few choices of solvers in section 3, and we remark on the choice of  $Z$  in later sections. Remarkably, the following two statements are independent of these choices.

**LEMMA 2.1 (AZ lemma).** *Let  $\hat{x} = \hat{x}_1 + \hat{x}_2$  be output from the AZ algorithm. Then the final residual is equal to the residual of step 1.*

*Proof.* A simple expansion of the final residual yields

$$\begin{aligned} b - A\hat{x} &= b - A\hat{x}_1 - A\hat{x}_2 \\ &= b - A\hat{x}_1 - AZ^*(b - A\hat{x}_1) \\ &= (I - AZ^*)(b - A\hat{x}_1). \end{aligned}$$

This is precisely the residual from step 1 of the algorithm.  $\square$

Thus, the accuracy with which step 1 is solved determines the accuracy of the overall algorithm, at least when accuracy is measured in terms of the size of the residual. Furthermore, the computational cost of the algorithm is also shifted to step 1, because steps 2 and 3 only involve matrix-vector multiplication and addition.

In the following statement, we use the terminology of a *stable least squares fit*. A stable least squares fit corresponds to a solution vector that has itself a moderate norm and that yields a small residual. The relevance of a stable least squares fit in the setting of our examples later in this paper is that, if a stable least squares fit exists, then numerical methods exist that can reliably find such a solution, no matter how ill-conditioned the linear system is.

One way to compute a stable least squares fit is based on a regularized SVD, and this case was studied for function approximation using frames in [2]. The following lemma proves that the AZ algorithm can also, in principle, find a stable least squares fit (if it exists).

**LEMMA 2.2** (stable least squares fitting). *Let  $A \in \mathbb{C}^{M \times N}$ ,  $b \in \mathbb{C}^M$ , and suppose there exists  $\tilde{x} \in \mathbb{C}^N$  such that*

$$(2.1) \quad \|b - A\tilde{x}\|_2 \leq \tau, \quad \|\tilde{x}\|_2 \leq C$$

*for  $\tau, C > 0$ . Then there exists a solution  $\hat{x}_1$  to step 1 of the AZ algorithm such that the residual of the computed vector  $\hat{x} = \hat{x}_1 + \hat{x}_2$  satisfies*

$$(2.2) \quad \|b - A\hat{x}\|_2 \leq \|I - AZ^*\|_2 \tau, \quad \|\hat{x}\|_2 \leq C + \|Z^*\|_2 \tau.$$

*Proof.* Take  $\hat{x}_1 = \tilde{x}$ . Then Lemma 2.1 implies  $b - A\hat{x} = (I - AZ^*)(b - A\tilde{x})$ , which gives us the first inequality. The second inequality follows from  $\hat{x} = \tilde{x} + Z^*(b - A\tilde{x})$ .  $\square$

Expression (2.2) hints at the expected algebraic regime in which the AZ algorithm may work well, namely, when both  $A$  and  $Z^*$  (and therefore  $I - AZ^*$  as well) have moderately bounded norms. In that case, the lemma shows that a stable least squares fit can be recovered with norm and residual that are comparable to any other stable fit. This is also the regime in which the notion of a stable least squares fit itself is relevant. Indeed, if  $A$  is ill-conditioned but bounded, then it must have small singular values or, equivalently, a nontrivial (numerical) null space. The consequences are twofold: there may be multiple different solutions with comparable residual, i.e., a solution vector may not be unique, and there is a risk of returning a large solution vector  $x$ , which is often undesirable. The aim in the choice of the solver in step 1 and the choice of  $Z$  is to compute a stable least squares fit.

**2.1. Choosing the matrix  $Z$ .** In principle, the AZ algorithm can be used with any matrix  $Z$ . However, we must choose  $Z$  wisely if we want efficiency and accuracy. If the norm of  $Z^*$  is large, then Lemma 2.2 suggests that we may fail to obtain a stable least squares fit. If  $A - AZ^*A$  is not low rank, then the cost of solving the least squares system in step 1 may be expensive. Ideally, both  $A$  and  $Z^*$  also have fast matrix-vector multiplication.

Two extreme choices are  $Z = 0$  and  $Z = (A^\dagger)^*$ . In the case  $Z = 0$ , step 1 solves the original system, and step 2 returns zero. We do not gain any efficiency over simply solving the original system. In the latter case, it is exactly the opposite: if  $Z = (A^\dagger)^*$ , then the first step yields  $x_1 = 0$ , and the original system is solved in the second step using the pseudoinverse of  $A$ . This may be efficient, but we would need to know the pseudoinverse of  $A$ ! In addition, the solution vector may be very large.

Any other choice of  $Z$  leads to a situation in between: part of the solution is found using a direct solver in step 1, and part of it is found by multiplication with  $Z^*$  in step 2. We desire a matrix  $Z$  such that  $A - AZ^*A$  has low numerical rank, so that step 1 can be solved efficiently, and  $Z^*$  must be readily available with an efficient matrix-vector multiply if possible. Much like a preconditioner, we choose  $Z$  satisfying these properties using some a priori information about the underlying problem.

The following lemma gives a general relationship between  $A$  and  $Z$  which would guarantee  $A - AZ^*A$  to be numerically low rank.

LEMMA 2.3. *Suppose that  $A, Z \in \mathbb{C}^{M \times N}$  satisfy*

$$A = W + L_1 + E_1, \quad Z^* = W^\dagger + L_2 + E_2,$$

where  $\text{rank}(L_1), \text{rank}(L_2) \leq R$ , and  $\|E_1\|_F, \|E_2\|_F \leq \varepsilon$ . Here  $W^\dagger$  is the Moore–Penrose pseudoinverse. Then

$$A - AZ^*A = L + E,$$

where  $\text{rank}(L) \leq 3R$  and

$$(2.3) \quad \|E\|_F \leq \varepsilon (1 + \|I - AZ^*\|_2 + \|A\|_2^2) + \varepsilon^2 \|A\|_2.$$

The result is exactly the same if the norms on  $E_1$ ,  $E_2$ , and  $E$ , are changed to  $\|\cdot\|_2$ .

*Proof.* We simply expand  $A - AZ^*A$ . For clarity we expand terms gradually.

$$\begin{aligned} A - AZ^*A &= W + L_1 + E_1 - AZ^*W - AZ^*(L_1 + E_1) \\ &= W + L_1 + E_1 - AW^\dagger W - A(L_2 + E_2)W - AZ^*(L_1 + E_1) \\ &= W + L_1 + E_1 - WW^\dagger W - (L_1 + E_1)W^\dagger W \\ &\quad - A(L_2 + E_2)W - AZ^*(L_1 + E_1). \end{aligned}$$

Since  $W^\dagger$  is a generalized inverse of  $W$ , we have  $W - WW^\dagger W = 0$ , so

$$A - AZ^*A = L_1 + E_1 - (L_1 + E_1)W^\dagger W - A(L_2 + E_2)W - AZ^*(L_1 + E_1).$$

Now, writing  $A(L_2 + E_2)W = AL_2W + AE_2A - AE_2L_1 - AE_2E_1$  and splitting up the low-rank and small norm parts, we obtain

$$\begin{aligned} A - AZ^*A &= \underbrace{(I - AZ^* + AE_2)L_1 - L_1W^\dagger W - AL_2W}_L \\ &\quad + \underbrace{(I - AZ^*)E_1 - E_1WW^\dagger - AE_2A + AE_2E_1}_E. \end{aligned}$$

It is clear that  $\text{rank}(L) \leq 3R$  because of the three occurrences of  $L_1$  and  $L_2$ . The Frobenius norm of  $E$  is bounded above by

$$\begin{aligned} \|E\|_F &\leq \|(I - AZ^*)E_1\|_F + \|WW^\dagger E_1\|_F + \|AE_2A\|_F + \|AE_2E_1\|_F \\ &\leq \|E_1\|_F \|I - AZ^*\|_2 + \|WW^\dagger\|_2 \|E_1\|_F + \|E_2\|_F \|A\|_2^2 + \|E_1\|_F \|E_2\|_2 \|A\|_2. \end{aligned}$$

Here we have used the inequality  $\|BC\|_F \leq \|B\|_2 \|C\|_F$  for general rectangular matrices  $B$  and  $C$ .<sup>1</sup>

<sup>1</sup>The proof of this is a one-liner:  $\|BC\|_F^2 = \sum_j \|Bc_j\|_2^2 \leq \|B\|_2^2 \sum_j \|c_j\|_2^2 = \|B\|_2^2 \|C\|_F^2$ .

Because  $W^\dagger$  is the Moore–Penrose pseudoinverse of  $W$ , the matrix  $WW^\dagger$  is an orthogonal projection onto the image of  $W$ , so  $\|WW^\dagger\|_2 \leq 1$ . Furthermore,  $\|E_2\|_2 \leq \|E_2\|_F \leq \varepsilon$ , which enables us to arrive at the desired bound on  $\|E\|_F$  in (2.3).

In order to prove the same result with the Frobenius norms on  $E_1$ ,  $E_2$ , and  $E$  replaced by  $\|\cdot\|_2$ , it is readily checked that the exact same proof holds with all instances of  $F$  replaced by 2. In particular, we have the inequality  $\|BC\|_2 \leq \|B\|_2\|C\|_2$ .  $\square$

This lemma focuses on the rank of  $A - AZ^*A$ . Following Lemma 2.2, and justified further by (2.3), we also desire  $A$  and  $Z^*$  to be bounded, even if  $A^\dagger$  is not. In the decomposition of this lemma, this implies that both  $W$  and its pseudoinverse  $W^\dagger$  should be bounded, but  $L_1^\dagger$  need not be. In other words, if  $A$  is seen as consisting of a well-conditioned part and an ill-conditioned part, then  $Z^*$  acts like an approximate generalized inverse of the well-conditioned part.

In our examples,  $A$  and  $Z$  are determined by analytical means, which are application-specific. We do not explicitly compute the  $W$  and  $L_1, L_2$  matrices of the above lemma. The effectiveness of the algorithm merely relies on the fact that these matrices exist.

**3. Fast randomized algorithms for numerically low-rank systems.** In this section we discuss fast algorithms for the solution of the system  $Ax = b$ , where  $A \in \mathbb{C}^{M \times N}$ , with  $M \geq N$ , has *epsilon rank*  $r$ . We write  $\text{rank}_\varepsilon(A) = r$ , and in this paper it means that there exists  $L, E \in \mathbb{C}^{M \times N}$  such that

$$(3.1) \quad A = L + E, \text{ where } \text{rank}(L) = r \text{ and } \|E\|_F \leq \varepsilon.$$

It is important to note that we have used the Frobenius norm here, which implies  $\sum_{k>r} \sigma_k^2 \leq \varepsilon^2$ , where  $\sigma_{r+1}, \dots, \sigma_N$  are the  $N - r$  smallest singular values of  $A$ . This is necessary for our proofs of the error bounds. Throughout this section there are points where the Frobenius norm appears to have been used unnecessarily where the 2-norm could have been used, but in all cases we do not believe that the final bounds will be improved significantly by changing to the 2-norm. This is due to the fact that we are not aware of an effective version of Proposition 3.1 which bounds 2-norms of the relevant random matrices purely in terms of 2-norms of other matrices.

We make use of Gaussian random matrices,  $\Omega \sim \mathcal{N}(0, 1; \mathbb{R}^{n \times k})$ , which are  $n \times k$  matrices whose elements are independent Gaussian random variables with mean 0 and variance 1.

**PROPOSITION 3.1** (see [11]). *Let  $\Omega \sim \mathcal{N}(0, 1; \mathbb{R}^{n \times k})$ , and let  $S, T \in \mathbb{R}^{n \times k}$  be fixed matrices. Then for all  $u \geq 0$ ,*

$$\mathbb{E} \{ \|S^* \Omega T\|_F \} \leq \|S\|_F \|T\|_F, \quad \mathbb{P} \{ \|S^* \Omega T\|_F \geq (1 + u) \cdot \|S\|_F \|T\|_F \} \leq e^{-\frac{u^2}{2}}.$$

**PROPOSITION 3.2** (see [11]). *Let  $\Omega \sim \mathcal{N}(0, 1; \mathbb{R}^{r \times (r+p)})$  with  $p \geq 4$ . Then for all  $s \geq 1$ ,*

$$\mathbb{E} \{ \|\Omega^\dagger\|_F \} = \sqrt{\frac{r}{p-1}}, \quad \mathbb{P} \left\{ \|\Omega^\dagger\|_F \geq s \cdot \sqrt{\frac{3r}{p+1}} \right\} \leq s^{-p}.$$

Note that we intend to apply the results of this section to step 1 of the AZ algorithm. Thus, matrix  $A$  in this section actually corresponds to matrix  $A - AZ^*A$  in Algorithm 2.1.

**Algorithm 3.1** Truncated SVD solver [10]**Input:**  $A \in \mathbb{C}^{M \times N}$ ,  $b \in \mathbb{C}^M$ ,  $\varepsilon > 0$ **Output:**  $x \in \mathbb{C}^N$  such that  $Ax \approx b$ 1: Compute the SVD,  $A = U\Sigma V^*$  where

$$U\Sigma V^* = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1 & V_2 \end{pmatrix}^*$$

with  $0 \leq \Sigma_2 < \varepsilon I \leq \Sigma_1$ .2:  $x \leftarrow V_1 \Sigma_1^{-1} U_1^* b$ 

**3.1. Truncated SVD solvers.** Algorithm 3.1 is a standard method of solving a linear system using a truncated SVD [10]. It is based on computing the full SVD of the matrix  $A$  and subsequently discarding the singular values smaller than a threshold  $\varepsilon$ . Assuming that  $M = \mathcal{O}(N)$ , this algorithm has cubic complexity  $\mathcal{O}(N^3)$  even when the numerical rank of  $A$  is small. For systems of low rank, the randomized algorithms that follow have a more favorable complexity. Nevertheless, because this algorithm is suggested in [2, 1], we prove bounds on the residual which are similar in flavor to those in [2, 1].

LEMMA 3.3. *Let  $x$  be computed by Algorithm 3.1. Then*

$$\|b - Ax\|_2 \leq \inf_{v \in \mathbb{C}^N} \left\{ \|b - Av\|_2 + \varepsilon \cdot \|v\|_2 \right\}.$$

*Proof.* We substitute  $x = V_1 \Sigma_1^{-1} U_1^* b$  into the residual to obtain

$$b - Ax = (I - AV_1 \Sigma_1^{-1} U_1^*)b.$$

One can expand the block form of the SVD of  $A$  into  $A = U_1 \Sigma_1 V_1^* + U_2 \Sigma_2 V_2^*$ , and since the columns of  $V$  are orthonormal vectors, we know that  $V_2^* V_1 = 0$ . Therefore,

$$\begin{aligned} AV_1 \Sigma_1^{-1} U_1^* &= U_1 \Sigma_1 V_1^* V_1 \Sigma_1^{-1} U_1^* + U_2 \Sigma_2 V_2^* V_1 \Sigma_1^{-1} U_1^* \\ &= U_1 \Sigma_1 \Sigma_1^{-1} U_1^* = U_1 U_1^*. \end{aligned}$$

For any  $v \in \mathbb{C}^N$ , we can add and subtract  $(I - U_1 U_1^*)Av$  to get

$$b - Ax = (I - U_1 U_1^*)(b - Av) + (I - U_1 U_1^*)Av.$$

Since the columns of  $U$  are orthonormal, we have  $U_1^* U_2 = 0$  and  $U_1^* U_1 = I$ . Therefore,

$$\begin{aligned} b - Ax &= (I - U_1 U_1^*)(b - Av) + (I - U_1 U_1^*)(U_1 \Sigma_1 V_1^* + U_2 \Sigma_2 V_2^*)v \\ &= (I - U_1 U_1^*)(b - Av) + U_2 \Sigma_2 V_2^* v. \end{aligned}$$

Since  $\Sigma_2 < \varepsilon I$  by assumption,  $\|\Sigma_2\|_2 < \varepsilon$ . Also,  $I - U_1 U_1^*$  is an orthogonal projection onto the orthogonal complement of the range of  $U_1$ , so  $\|I - U_1 U_1^*\|_2 = 1$ . Furthermore,  $\|U_2\|_2 = \|V_2^*\|_2 = 1$ , so the bound on the norm of the residual readily follows.  $\square$

Randomized algorithms based on matrix-vector products with random vectors can compute a truncated SVD at a lower cost when the effective numerical rank is small. Algorithm 3.2 assumes that the user specifies a value  $R = r + p$ , where  $r$  is the

numerical rank of  $A$  and larger values of  $p$  yield higher probability of accuracy, as quantified in analysis that follows.

The existing analysis of randomized linear algebra (see [11] and references therein) focuses mostly on the accuracy of the matrix factorization. In this paper we are interested mainly in bounds of the residual of Algorithm 3.2. To our knowledge, it is not possible to derive residual bounds directly from the error of the factorization, so we derive the result directly using similar techniques to those in [11] for bounding expectation and tail probabilities of the random matrices involved.

---

**Algorithm 3.2** Randomized truncated SVD solver

---

**Input:**  $A \in \mathbb{C}^{M \times N}$ ,  $b \in \mathbb{C}^M$ ,  $R \in \{1, \dots, N\}$ ,  $\varepsilon > 0$

**Output:**  $x \in \mathbb{C}^N$  such that  $Ax \approx b$

- 1: Generate  $\Omega \sim \mathcal{N}(0, 1; \mathbb{R}^{N \times R})$
- 2:  $\tilde{A} \leftarrow A\Omega \in \mathbb{C}^{M \times R}$
- 3: Compute the SVD,  $\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^*$  where

$$\tilde{U}\tilde{\Sigma}\tilde{V}^* = \begin{pmatrix} \tilde{U}_1 & \tilde{U}_2 \end{pmatrix} \begin{pmatrix} \tilde{\Sigma}_1 & \\ & \tilde{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \tilde{V}_1 & \tilde{V}_2 \end{pmatrix}^*$$

with  $0 \leq \tilde{\Sigma}_2 < \varepsilon I \leq \tilde{\Sigma}_1$ . Here  $\tilde{U} \in \mathbb{C}^{M \times R}$ ,  $\tilde{\Sigma} \in \mathbb{R}^{R \times R}$ ,  $\tilde{V} \in \mathbb{C}^{R \times R}$ , but the dimensions of the blocks depend on the singular values.

- 4:  $y \leftarrow \tilde{V}_1 \tilde{\Sigma}_1^{-1} \tilde{U}_1^* b$
  - 5:  $x \leftarrow \Omega y$
- 

**THEOREM 3.4** (residual bounds for randomized truncated SVD solver). *Assume that  $A$  is such that  $\text{rank}_\varepsilon(A) = r$  (as defined in (3.1)), and let  $x \in \mathbb{C}^N$  come from step 5 of Algorithm 3.2 with  $R = r + p$  for  $p \geq 2$ . Then*

$$\|b - Ax\|_2 \leq \inf_{v \in \mathbb{C}^N} \left\{ \|b - Av\|_2 + \varepsilon \cdot (1 + \kappa) \cdot \|v\|_2 \right\},$$

where  $\kappa$  is a non-negative-valued random variable satisfying

$$\mathbb{E}\{\kappa\} \leq 2\sqrt{\frac{r}{p-1}}, \quad \mathbb{P}\left\{\kappa > (2+u) \cdot s \cdot \sqrt{\frac{3r}{p+1}}\right\} \leq s^{-p} + e^{-\frac{u^2}{2}}$$

for any  $s \geq 1$ ,  $u \geq 0$ . Loosely speaking,  $\kappa = \mathcal{O}(\sqrt{r})$  with a high probability which improves rapidly as  $p$  increases.

**Remark 3.5.** The probability distribution of  $\kappa$  is similar to that of random variables appearing in the factorization errors of [11, section 10]. Careful choices of  $s$  and  $u$  will give different bounds on the probability which depends on  $p$ . Following the example choices of parameters in [11], setting  $s = e$ ,  $u = 2 + \sqrt{2p}$ , and  $p = 20$  in this theorem shows that

$$\mathbb{E}\{\kappa\} \leq 0.459\sqrt{r}, \quad \mathbb{P}\{\kappa > 8.56\sqrt{r}\} \leq 4.13 \times 10^{-9}.$$

It might be tempting to let  $p$  grow linearly with respect to  $r$ , since then the expected value of  $\kappa$  is  $\mathcal{O}(1)$  instead of  $\mathcal{O}(\sqrt{r})$ , but this is overkill, and we find that in practice a fixed  $p$  such as 20 works well.

*Proof.* Note that we can write  $\tilde{A} = \tilde{U}_1 \tilde{\Sigma}_1 \tilde{V}_1^* + \tilde{U}_2 \tilde{\Sigma}_2 \tilde{V}_2^*$  with  $\|\tilde{\Sigma}_2\|_2 \leq \varepsilon$ , by step 3 of Algorithm 3.1. By assumption,  $A$  has epsilon rank  $r$ , so its SVD is of the form

$$A = \underbrace{U_1 \Sigma_1 V_1^*}_L + \underbrace{U_2 \Sigma_2 V_2^*}_E,$$

where  $\Sigma_1 \in \mathbb{R}^{r \times r}$ , the concatenated matrices  $(U_1 \ U_2)$  and  $(V_1 \ V_2)$  have orthonormal columns, and  $\|E\|_F = \|\Sigma_2\|_F \leq \varepsilon$ . Substituting  $x = \Omega \tilde{V}_1 \tilde{\Sigma}_1^{-1} \tilde{U}_1^* b$  into the residual, we get

$$\begin{aligned} b - Ax &= b - \tilde{A} \tilde{V}_1 \tilde{\Sigma}_1^{-1} \tilde{U}_1^* b \\ &= (I - \tilde{U}_1 \tilde{\Sigma}_1 \tilde{V}_1^* \tilde{V}_1 \tilde{\Sigma}_1^{-1} \tilde{U}_1^* - \tilde{U}_2 \tilde{\Sigma}_2 \tilde{V}_2^* \tilde{V}_1 \tilde{\Sigma}_1^{-1} \tilde{U}_1^*) b \\ &= (I - \tilde{U}_1 \tilde{U}_1^*) b \end{aligned}$$

by the identities for  $\tilde{U}_1, \tilde{U}_2, \tilde{V}_1, \tilde{V}_2$  which follow from the orthonormal columns of  $\tilde{U}$  and  $\tilde{V}$ . Now, for any  $v \in \mathbb{C}^N$ , if we write  $b = (b - Av) + Av = (b - Av) + Ev + Lv$ , then

$$(3.2) \quad b - Ax = (I - \tilde{U}_1 \tilde{U}_1^*)(b - Av + Ev + Lv).$$

Consider the matrices  $\Omega_1 = V_1^* \Omega$  and  $\Omega_2 = V_2^* \Omega$ . They are submatrices of the Gaussian matrix  $(V_1 \ V_2)^* \Omega$  (noting that the independence and Gaussian distribution of the elements are preserved by a unitary transformation). It follows that  $\Omega_1$  and  $\Omega_2$  are independent Gaussian matrices in  $\mathbb{R}^{r \times (r+p)}$  and  $\mathbb{R}^{(N-r) \times (r+p)}$ , are respectively. With probability 1, the rows of  $\Omega_1$  are linearly independent, so that the pseudoinverse is in fact a right inverse, i.e.,  $\Omega_1 \Omega_1^\dagger = I_{r \times r}$  with probability 1. Therefore, we can write  $L = U_1 \Sigma_1 V_1^* = U_1 \Sigma_1 \Omega_1 \Omega_1^\dagger V_1^* = U_1 \Sigma_1 V_1 \Omega \Omega_1^\dagger V_1^* = L \Omega \Omega_1^\dagger V_1^*$ . Since  $\tilde{A} = A \Omega$  (by step 2 of Algorithm 3.1), this implies  $L = (A - E) \Omega \Omega_1^\dagger V_1^* = \tilde{A} \Omega_1^\dagger V_1^* - U_2 \Sigma_2 \Omega_2 \Omega_1^\dagger V_1^*$ . Substituting this into (3.2) gives

$$\begin{aligned} (3.3) \quad b - Ax &= (I - \tilde{U}_1 \tilde{U}_1^*)(b - Av) \\ &\quad + (I - \tilde{U}_1 \tilde{U}_1^*)(E - U_2 \Sigma_2 \Omega_2 \Omega_1^\dagger V_1^*) v \\ &\quad + (I - \tilde{U}_1 \tilde{U}_1^*) \tilde{A} \Omega_1^\dagger V_1^* v. \end{aligned}$$

There are three terms here. Note that  $\|I - \tilde{U}_1 \tilde{U}_1^*\|_2 \leq 1$  since  $\tilde{U}_1$  has orthonormal columns, so the first term is bounded above by  $\|b - Av\|_2$ , and the second term is bounded above by  $(\|E\|_F + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F) \|v\|_2$ . The third term requires more manipulation, as follows. Using the decomposition  $\tilde{A} = \tilde{U}_1 \tilde{\Sigma}_1 \tilde{V}_1^* + \tilde{U}_2 \tilde{\Sigma}_2 \tilde{V}_2^*$  made at the start of the proof, we obtain the identity  $(I - \tilde{U}_1 \tilde{U}_1^*) \tilde{A} = \tilde{U}_2 \tilde{\Sigma}_2 \tilde{V}_2^*$ . Using this, the third term in (3.3) has norm that is readily confirmed to be bounded above by  $\|\tilde{\Sigma}_2\|_2 \|\Omega_1^\dagger\|_F \|v\|_2$ .

Combining our estimates for the three terms in (3.3) provides the deterministic bound

$$\begin{aligned} \|b - Ax\|_2 &\leq \|b - Av\|_2 + (\|E\|_F + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F) \|v\|_2 + \|\tilde{\Sigma}_2\|_2 \|\Omega_1^\dagger\|_F \|v\|_2 \\ &\leq \|b - Av\|_2 + \varepsilon (1 + \|\varepsilon^{-1} \Sigma_2 \Omega_2 \Omega_1^\dagger\|_F + \|\Omega_1^\dagger\|_F) \|v\|_2. \end{aligned}$$

Now we bound the expectation and tail probabilities of the random variable

$$\kappa = \|\varepsilon^{-1} \Sigma_2 \Omega_2 \Omega_1^\dagger\|_F + \|\Omega_1^\dagger\|_F.$$



Note that this has nonnegative value, as required. From Proposition 3.1 with  $S^* = \varepsilon^{-1}\Sigma_2$  and  $T = \Omega_1^\dagger$  which is independent of the Gaussian matrix  $\Omega_2$ , we obtain

$$\mathbb{E} \left\{ \|\varepsilon^{-1}\Sigma_2\Omega_2\Omega_1^\dagger\|_F \mid \Omega_1 \right\} \leq \|\Omega_1^\dagger\|_F,$$

since  $\|\varepsilon^{-1}\Sigma_2\|_F \leq 1$ . Therefore  $\mathbb{E} \{\kappa\} \leq 2\mathbb{E} \{\|\Omega_1^\dagger\|_F\}$ . Applying Proposition 3.2 to this yields the bound on the expectation of  $\kappa$ .

Now we turn to the bounds on tail probabilities of  $\kappa$ . Following [11, Theorem 10.8] with a crude simplification (in which we only consider Frobenius norms throughout the calculation), we condition on the event

$$\mathcal{E}_s = \left\{ \Omega_1 : \|\Omega_1^\dagger\|_F < s \cdot \sqrt{\frac{3r}{p+1}} \right\},$$

where  $s \geq 1$ . Proposition 3.2 implies that the probability that  $\mathcal{E}_s$  is empty is  $s^{-p}$ . Conditional on  $\mathcal{E}_s$ , Proposition 3.1 gives us the inequality

$$\mathbb{P} \left\{ \|\varepsilon^{-1}\Sigma_2\Omega_2\Omega_1^\dagger\|_F > (1+u) \cdot s \cdot \sqrt{\frac{3r}{p+1}} \mid \mathcal{E}_s \right\} \leq e^{-\frac{u^2}{2}}$$

for all  $u \geq 0$ . This implies

$$\mathbb{P} \left\{ \kappa > (2+u) \cdot s \cdot \sqrt{\frac{3r}{p+1}} \mid \mathcal{E}_s \right\} \leq e^{-\frac{u^2}{2}}.$$

Adding in the probability that  $\mathcal{E}_s$  is empty, in order to remove the conditioning, we arrive at the tail probability bound for  $\kappa$ .  $\square$

**3.2. Truncated pivoted QR solvers.** Similar to the case of the SVD, first we formulate a standard algorithm based on a full pivoted QR decomposition in Algorithm 3.3. A randomized algorithm with better complexity for systems of small numerical rank is Algorithm 3.4.

---

**Algorithm 3.3** Truncated pivoted QR solver [10]

---

**Input:**  $A \in \mathbb{C}^{M \times N}$ ,  $b \in \mathbb{C}^M$ ,  $r \in \{1, \dots, N\}$

**Output:**  $x \in \mathbb{C}^N$  such that  $Ax \approx b$

- 1: Compute a column pivoted QR decomposition  $A\Pi = QR$ , with block forms,

$$\Pi = \begin{pmatrix} \Pi_1 & \Pi_2 \end{pmatrix}, \quad Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}, \quad R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix},$$

where  $\Pi_1 \in \mathbb{C}^{N \times r}$ ,  $\Pi_2 \in \mathbb{C}^{N \times N-r}$ ,  $Q_1 \in \mathbb{C}^{M \times r}$ ,  $Q_2 \in \mathbb{C}^{M \times N-r}$ ,  $R_{11} \in \mathbb{C}^{r \times r}$ ,  $R_{22} \in \mathbb{C}^{(N-r) \times (N-r)}$ .

- 2:  $x \leftarrow \Pi_1 R_{11}^{-1} Q_1^* b$
- 

LEMMA 3.6. *Let  $x$  be computed by Algorithm 3.3. Then*

$$\|b - Ax\|_2 \leq \inf_{v \in \mathbb{C}^N} \left\{ \|b - Av\|_2 + \|R_{22}\|_2 \cdot \|v\|_2 \right\}.$$

*Proof.* We substitute  $x = \Pi_1 R_{11}^{-1} Q_1^* b$  into the residual to obtain

$$b - Ax = (I - A\Pi_1 R_{11}^{-1} Q_1^*)b.$$

Note that  $A\Pi_1 = Q_1 R_{11}$ . Therefore,  $A\Pi_1 R_{11}^{-1} Q_1^* = Q_1 Q_1^*$ , and so  $b - Ax = (I - Q_1 Q_1^*)b$ . For any  $v \in \mathbb{C}^N$  we add and subtract  $(I - Q_1 Q_1^*)Av$  to obtain

$$b - Ax = (I - Q_1 Q_1^*)(b - Av) + (I - Q_1 Q_1^*)Av.$$

Let us deal with the  $(I - Q_1 Q_1^*)Av$  term. Note that  $\Pi$  is merely a permutation matrix, so  $\Pi\Pi^T = I$ . Therefore,

$$\begin{aligned} (I - Q_1 Q_1^*)Av &= (I - Q_1 Q_1^*) A\Pi\Pi^T v \\ &= (I - Q_1 Q_1^*) QR\Pi^T v \\ &= \begin{pmatrix} 0 & Q_2 \end{pmatrix} R\Pi^T v \\ &= \begin{pmatrix} 0 & Q_2 R_{22} \end{pmatrix} \Pi^T v \\ &= Q_2 R_{22} \Pi_2^T v. \end{aligned}$$

Therefore,  $b - Ax = (I - Q_1 Q_1^*)(b - Av) + Q_2 R_{22} \Pi_2^T v$ , and the bound follows readily from  $\|I - Q_1 Q_1^*\|_2 \leq 1$ ,  $\|Q_2\|_2 \leq 1$  and  $\|\Pi_2^T\|_2 \leq 1$ .  $\square$

---

**Algorithm 3.4** Randomized truncated pivoted QR solver

---

**Input:**  $A \in \mathbb{C}^{M \times N}$ ,  $b \in \mathbb{C}^M$ ,  $R \in \{1, \dots, N\}$ ,  $\varepsilon > 0$

**Output:**  $x \in \mathbb{C}^N$  such that  $Ax \approx b$

- 1: Generate  $\Omega \sim \mathcal{N}(0, 1; \mathbb{R}^{N \times R})$
- 2:  $\tilde{A} \leftarrow A\Omega \in \mathbb{C}^{M \times R}$
- 3: Compute a column pivoted QR decomposition  $\tilde{A}\tilde{\Pi} = \tilde{Q}\tilde{R}$  with block forms,

$$\tilde{\Pi} = \begin{pmatrix} \tilde{\Pi}_1 & \tilde{\Pi}_2 \end{pmatrix}, \quad \tilde{Q} = \begin{pmatrix} \tilde{Q}_1 & \tilde{Q}_2 \end{pmatrix}, \quad \tilde{R} = \begin{pmatrix} \tilde{R}_{11} & \tilde{R}_{12} \\ 0 & \tilde{R}_{22} \end{pmatrix}$$

with  $0 \leq \text{diag}(\tilde{R}_{22}) < \varepsilon I \leq \text{diag}(\tilde{R}_{11})$ . Here  $\tilde{\Pi} \in \mathbb{R}^{M \times R}$ ,  $\tilde{Q} \in \mathbb{C}^{M \times R}$ ,  $\tilde{R} \in \mathbb{C}^{R \times R}$ , but the dimensions of the blocks depend on the diagonal entries of  $\tilde{R}$ .

- 4:  $y \leftarrow \tilde{\Pi}_1 \tilde{R}_{11}^{-1} \tilde{Q}_1^* b$
  - 5:  $x \leftarrow \Omega y$
- 

**THEOREM 3.7** (residual bounds for randomized pivoted QR solver). *Assume that  $A$  is such that  $\text{rank}_\varepsilon(A) = r$  (as defined in (3.1)), and let  $x \in \mathbb{C}^N$  come from step 5 of Algorithm 3.4 with  $R = r + p$  for  $p \geq 2$ . Then*

$$\|b - Ax\|_2 \leq \inf_{v \in \mathbb{C}^N} \left\{ \|b - Av\|_2 + \varepsilon \cdot (1 + \kappa) \cdot \|v\|_2 \right\},$$

where  $\kappa$  is a non-negative-valued random variable satisfying

$$\mathbb{E}\{\kappa\} \leq (1 + \sqrt{r+p}) \sqrt{\frac{r}{p-1}}, \quad \mathbb{P}\left\{\kappa > (1 + \sqrt{r+p} + u) \cdot s \cdot \sqrt{\frac{3r}{p+1}}\right\} \leq s^{-p} + e^{-\frac{u^2}{2}}$$

for any  $s \geq 1$ ,  $u \geq 0$ . Loosely speaking,  $\kappa = \mathcal{O}(r)$  with a high probability which improves rapidly as  $p$  increases.

*Remark 3.8.* The bounds here on  $\kappa$  are larger than those found for the randomized truncated SVD algorithm in Theorem 3.4. This is related to the possibility that the diagonals of  $\tilde{R}_{22}$  in the pivoted QR algorithm can bear no relation to the underlying numerical rank of  $\tilde{A}$ . This is discussed in [10, section 5.4.3], and therein it is noted that “Nevertheless, in practice, small trailing  $R$ -submatrices almost always emerge that correlate well with the underlying [numerical] rank.” Hence, in practice, we would expect  $\kappa = \mathcal{O}(\sqrt{r})$  as in the randomized truncated SVD algorithm.

*Proof.* By assumption,  $A$  has epsilon rank  $r$ , so its SVD is of the form

$$A = \underbrace{U_1 \Sigma_1 V_1^*}_L + \underbrace{U_2 \Sigma_2 V_2^*}_E,$$

where  $\Sigma_1 \in \mathbb{R}^{r \times r}$ , the concatenated matrices  $(U_1 \ U_2)$  and  $(V_1 \ V_2)$  have orthonormal columns, and  $\|E\|_F = \|\Sigma_2\|_F \leq \varepsilon$ . Substituting  $x = \Omega \tilde{\Pi}_1 \tilde{R}_{11}^{-1} \tilde{Q}_1^* b$  into the residual gives

$$b - Ax = b - \tilde{A} \tilde{\Pi}_1 \tilde{R}_{11}^{-1} \tilde{Q}_1^* b.$$

Note that  $\tilde{A} \tilde{\Pi}_1 = \tilde{Q}_1 \tilde{R}_{11}$ ; therefore,  $\tilde{A} \tilde{\Pi}_1 \tilde{R}_{11}^{-1} \tilde{Q}_1^* = \tilde{Q}_1 \tilde{Q}_1^*$ , so  $b - Ax = (I - \tilde{Q}_1 \tilde{Q}_1^*)b$ . Now, for any  $v \in \mathbb{C}^N$ , if we write  $b = (b - Av) + Av = (b - Av) + Ev + Lv$ , then

$$(3.4) \quad b - Ax = (I - \tilde{Q}_1 \tilde{Q}_1^*)(b - Av + Ev + Lv).$$

Using the same notation and result derived in the proof of Theorem 3.4, we have  $L = \tilde{A} \Omega_1^\dagger V_1^* - U_2 \Sigma_2 \Omega_2 \Omega_1^\dagger V_1^*$ . Substituting this into (3.4) gives

$$(3.5) \quad \begin{aligned} b - Ax &= (I - \tilde{Q}_1 \tilde{Q}_1^*)(b - Av) \\ &+ (I - \tilde{Q}_1 \tilde{Q}_1^*)(E - U_2 \Sigma_2 \Omega_2 \Omega_1^\dagger V_1^*)v \\ &+ (I - \tilde{Q}_1 \tilde{Q}_1^*) \tilde{A} \Omega_1^\dagger V_1^* v. \end{aligned}$$

There are three terms here of which the 2-norm needs to be bounded. Note that  $\|I - \tilde{Q}_1 \tilde{Q}_1^*\|_2 \leq 1$  since  $\tilde{Q}_1$  has orthonormal columns, so the first term is bounded above by  $\|b - Av\|_2$ , and the second term is bounded above by  $(\|E\|_F + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F) \|v\|_2$ . The third term requires more manipulation, as follows. Following the same reasoning as in the proof of Lemma 3.6, we have  $(I - \tilde{Q}_1 \tilde{Q}_1^*) \tilde{A} = \tilde{Q}_2 \tilde{R}_{22} \tilde{\Pi}_2^T$ . Using this, the third term in (3.5) has norm that is readily confirmed to be bounded above by  $\|\tilde{R}_{22}\|_F \|\Omega_1^\dagger\|_F \|v\|_2$ .

Combining our estimates for the three terms in (3.5) provides the deterministic bound

$$\begin{aligned} \|b - Ax\|_2 &\leq \|b - Av\|_2 + (\|E\|_F + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_F) \|v\|_2 + \|\tilde{R}_{22}\|_F \|\Omega_1^\dagger\|_F \|v\|_2 \\ &\leq \|b - Av\|_2 + \varepsilon (1 + \|\varepsilon^{-1} \Sigma_2 \Omega_2 \Omega_1^\dagger\|_F + \|\varepsilon^{-1} \tilde{R}_{22}\|_F \|\Omega_1^\dagger\|_F) \|v\|_2. \end{aligned}$$

We can bound  $\|\varepsilon^{-1} \tilde{R}_{22}\|_F$  by  $\sqrt{r+p}$  by the following argument. Column pivoting ensures that the first column of  $\tilde{R}_{22}$  has the greatest norm of all its columns [10, section 5.4.2]. Therefore, since  $\tilde{R}_{22}$  has at most  $r+p$  columns, we have

$$\|\tilde{R}_{22}\|_F^2 \leq (r+p) \left| [\tilde{R}_{22}]_{11} \right|^2 \leq (r+p) \varepsilon^2.$$

The proof is completed if we can bound the expectation and tail probabilities of the random variable

$$\kappa = \|\varepsilon^{-1} \Sigma_2 \Omega_2 \Omega_1^\dagger\|_F + \sqrt{r+p} \|\Omega_1^\dagger\|_F$$

with the appropriate bounds. This can be done by almost exactly the same analysis as in the proof of Theorem 3.4.  $\square$

**3.3. Computational cost of the randomized solvers.** Both Algorithm 3.2 and Algorithm 3.4 consist of 5 steps. We break down the computational cost for each step. Recall that the input parameters are  $A \in \mathbb{C}^{M \times N}$ ,  $b \in \mathbb{C}^M$ ,  $R \in \{1, \dots, N\}$ ,  $\varepsilon > 0$ . We will describe the costs using big  $\mathcal{O}$  notation while regarding  $A$ ,  $b$ , and  $\varepsilon$  as fixed quantities whose properties are hidden in the big  $\mathcal{O}$ . The number of operations required to apply  $A$  to a vector is denoted  $T_{\text{mult},A}$ .

1. Generate  $N \cdot R$  Gaussian random numbers:  $\mathcal{O}(N \cdot R)$ .
2. Apply the matrix  $A$  to  $R$  vectors:  $\mathcal{O}(R \cdot T_{\text{mult},A})$ .
3. Compute the SVD or pivoted QR factorization of an  $M \times R$  matrix:  $\mathcal{O}(M \cdot R^2)$ .
4. Apply a sequence of matrices with one dimension smaller than  $R$  and the other smaller than  $M$ :  $\mathcal{O}(R \cdot M)$ .
5. Apply a matrix of size  $N \times R$  to a vector:  $\mathcal{O}(R \cdot N)$ .

The total computational cost in floating point operations is, therefore,

$$\mathcal{O}(RT_{\text{mult},A} + R^2M).$$

Hence, these algorithms have improved computational cost when compared to their nonrandomized counterparts (Algorithm 3.1 and Algorithm 3.3) when  $R = o(N)$ , i.e., when the numerical rank grows slower than  $N$  with increasing  $N$ . The non-randomized algorithms require  $\mathcal{O}(MN^2)$  operations. Further gains are possible in the randomized methods if  $A$  supports fast matrix-vector multiplication.

**4. Fourier extension.** We proceed by illustrating the AZ algorithm for a number of cases. In particular, we show how to identify a suitable  $Z$  matrix in different settings having to do with function approximation. The matrix  $Z^*$  should be close to a pseudoinverse of  $A$  up to a numerically low-rank error—recall Lemma 2.3. For function approximation, in practice this means that  $Z^*b$  are coefficients yielding an accurate approximation for a space of functions whose complement is of low dimension (relative to the degree of the approximation).

The first case is that of function approximation using Fourier extension frames, where a severely ill-conditioned matrix  $A$  is obtained by discretizing the problem. Here it is known that, up to a real-valued normalization factor,  $Z = A$ : see [15, Algorithm 2] and [16, Algorithm 1]. Compared to these references, we emphasize here how  $Z$  arises from the so-called canonical dual frame, which yields extremely accurate results for functions compactly supported within the inner domain  $\Omega$ . The complement of this space is of low dimension related to the Hausdorff measure of the boundary [16] which leads to the low rank of  $A - AZ^*A$ .

**4.1. Approximation with the Fourier extension frame.** We refer the reader to [5] for an introduction to frames and to [2, 1] for an exposition on their use in function approximation. We merely recall that a system of functions  $\Phi = \{\phi_k\}_{k=1}^\infty$  is a frame for a separable Hilbert space  $H$  if it satisfies the frame condition: there exist constants  $0 < A, B < \infty$  such that

$$(4.1) \quad A\|f\|_H^2 \leq \sum_{k=1}^\infty \langle f, \phi_k \rangle_H^2 \leq B\|f\|_H^2.$$

For comparison, note that for a complete orthonormal basis (4.1) holds with  $A = B = 1$ . However, unlike a basis, a frame may be redundant, i.e., linearly dependent, causing ill-conditioning in the discretization process.

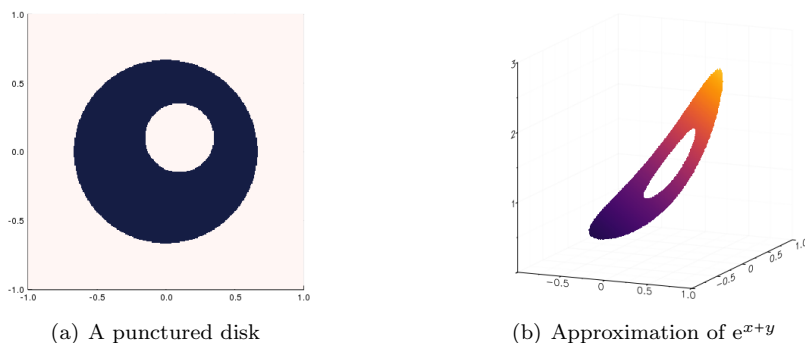


FIG. 4.1. *Fourier extension approximation of the function  $f(x, y) = e^{x+y}$  on a punctured disk. The function is represented by a periodic Fourier series in the bounding box  $[-1, 1]^2$ .*

To fix one frame, we consider a  $d$ -dimensional complete orthonormal Fourier basis  $\Phi$  on the box  $[-1, 1]^d$ ,

$$\Phi = \{\phi_{\mathbf{n}}\}_{\mathbf{n} \in \mathbb{Z}^d} \quad \text{with} \quad \phi_{\mathbf{n}}(\mathbf{x}) = 2^{-d/2} e^{i\pi \mathbf{n} \cdot \mathbf{x}}, \quad \mathbf{x} \in [-1, 1]^d.$$

The restriction of this basis to a smaller domain  $\Omega \subset [-1, 1]^d$  yields a Fourier extension frame for  $L^2(\Omega)$  (see [12] and [2, Example 1]). Any smooth function  $f$  defined on  $\Omega$  is well approximated in this frame, even if it is not periodic (which may not be a well-defined concept for an irregularly shaped  $\Omega$ ). An example is shown in Figure 4.1.

The approximation problem requires the truncation of the infinite frame. Thus, assuming that  $N^{1/d}$  is an integer, define the subset of  $N$  functions  $\Phi_N = \{\phi_{\mathbf{n}}(\mathbf{x})\}_{\mathbf{n} \in I_N}$ , with

$$I_N = \left\{ \mathbf{n} = (n_1, n_2, \dots, n_d) \in \mathbb{Z}^d : -\frac{1}{2}N^{1/d} \leq n_1, \dots, n_d \leq \frac{1}{2}N^{1/d} \right\},$$

where we consider only  $\mathbf{x} \in \Omega$ . The best approximation to  $f$  in the norm of  $L^2(\Omega)$  leads to a linear system  $Ax = B$  where  $A$  is the Gram matrix  $G$  of  $\Phi_N$  with elements

$$(4.2) \quad G_{\mathbf{n}, \mathbf{m}} = \langle \phi_{\mathbf{n}}, \phi_{\mathbf{m}} \rangle_{L^2(\Omega)}, \quad \mathbf{n}, \mathbf{m} \in I_N.$$

Since inner products on  $\Omega$  may be difficult to compute, especially for multidimensional domains, we resort to a *discrete least squares approximation* instead. Given a set of points  $\{\mathbf{x}_m\}_{m=1}^M \subset \Omega$ , this leads to the linear system  $Ax = B$  with<sup>2</sup>

$$(4.3) \quad A_{m,n} = \phi_n(\mathbf{x}_m), \quad \text{and} \quad B_m = f(\mathbf{x}_m), \quad n = 1, \dots, N, \quad m = 1, \dots, M.$$

It follows from the theory in [2, 1] that this linear system has solutions that are stable least squares fits, and one such solution can be found using a truncated (and thereby regularized) SVD of  $A$ .

**4.2. The canonical dual frame.** The concept of dual frames does not play a large role in the analysis in [2, 1], but it is important in the theory of frames itself, and, as we shall see, it is also relevant for the application of the AZ algorithm. A frame  $\Phi = \{\phi_k\}_{k=1}^\infty$  for a Hilbert space  $H$  has a *dual frame*  $\Psi = \{\psi_k\}_{k=1}^\infty$  if

$$(4.4) \quad f = \sum_{k=1}^{\infty} \langle f, \psi_k \rangle \phi_k \quad \forall f \in H.$$

<sup>2</sup>The index  $\mathbf{n}$  of  $\phi_{\mathbf{n}}$  in the definition of the truncated frame  $\Phi_N$  is a multi-index. For simplicity, in this expression we have identified each multi-index  $\mathbf{n} \in I_N$  with an integer  $n$  between 1 and  $N$ .

Convergence is understood in the norm of  $H$ . The above expansion simplifies if  $\Phi$  is an orthonormal basis: in that case  $\Psi = \Phi$ . Yet, since frames may be redundant, it is important to remark that (4.4) may not be the only representation of  $f$  in the frame. Correspondingly, dual frames are not necessarily unique.

The so-called *canonical dual frame* plays a special role among all dual frames. For the case of Fourier extension, this dual frame corresponds to the Fourier series on  $[-1, 1]^d$  of  $\bar{f}$ , defined as the extension of  $f$  by zero. The canonical dual frame expansion is readily obtained and easily truncated. For example, one can sample  $\bar{f}$  in a regular grid on  $[-1, 1]^d$  and use the FFT. Unfortunately, though the expansion converges in norm, it does not actually yield an accurate approximation for most  $f$ . Since extension by zero introduces a discontinuity along the boundary of  $\Omega$ , in general the Fourier series of  $\bar{f}$  exhibits the Gibbs phenomenon.

Nevertheless, the canonical dual frame expansion does lead to accurate approximations for a large class of functions, namely, those that are compactly support on  $\Omega$ . In that case, their extension by zero results in a smooth and periodic function, and rapid convergence of the Fourier series of  $\bar{f}$  follows. Thus, we have a simple and efficient FFT-based solver for a large subclass of functions, from which we can derive a suitable matrix  $Z$ .

**4.3. Characterizing the Z-matrix: Continuous duality.** We will first illustrate the properties of an approximation computed using  $A$  equal to the Gram matrix (4.2). In Figure 4.2 the computational domain is  $\Omega = [-\frac{3}{4}, -\frac{1}{4}] \cup [0, \frac{1}{2}]$ . The frame is the restriction of the normalized Fourier basis on  $[-1, 1]$  to that domain. The left panel shows the approximation to  $e^x$  on this nonconnected domain. The right panel shows the singular values of the Gram matrix for  $N = 50$ . The singular values cluster near 1 and near 0.<sup>3</sup>

Any matrix  $A$  with clustered singular values as shown in Figure 4.2(b) is amenable to the AZ algorithm with the simple choice  $Z = A$ , regardless of the underlying application. Indeed, if  $A$  has left and right singular vectors  $u_k$  and  $v_k$ , respectively, with singular value  $\sigma_k$ , one has that

$$(A - AZ^*A)v_k = (A - AA^*A)v_k = \sigma_k(1 - \sigma_k^2)v_k.$$

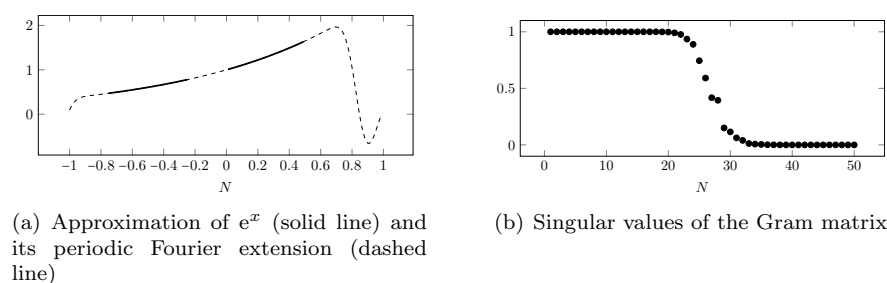


FIG. 4.2. Fourier extension approximation of the function  $f(x) = e^x$  on the domain  $[-\frac{3}{4}, -\frac{1}{4}] \cup [0, \frac{1}{2}]$ , using orthonormal Fourier series on  $[-1, 1]$  ( $N = 50$ ). The best approximation was computed using the Gram matrix (4.2). Its singular values cluster exponentially near 1 and near 0.

<sup>3</sup>The properties of the Gram matrix are well understood for the case where  $\Omega$  is a regular subinterval of  $[-1, 1]$ . In that case, the Gram matrix is also known as the *prolate matrix* [18]. For example, it is known that the size of the cluster near 1 is determined by the size of  $\Omega$  relative to that of  $[-1, 1]$ . Moreover, the clustering is exponential [8]. The experiment shows that these properties are preserved for a nonconnected subset  $\Omega$ .

If  $A$  has all singular values between 0 and 1 and has  $l$  of them between  $\epsilon$  and  $1 - \epsilon$ , with  $0 < \epsilon < 1/2$ , then  $A - AZ^*A$  has at most  $l$  singular values larger than  $\epsilon$ .

The clustering of singular values precisely near the unit value 1 is a consequence of the normalization of the Fourier series on  $[-1, 1]$ . A different normalization of the basis functions is easily accommodated, since it only results in a diagonal scaling of the Gram matrix. In that case matrix  $Z$  may be modified using the inverse of that scaling. That is, starting from a valid combination of  $A$  and  $Z$ , we can use the combination

$$A_1 = DA, \quad Z_1 = (D^{-1})^* Z,$$

so that their product  $Z_1^* A_1 = Z^* D^{-1} D A = Z^* A$  remains unchanged.

**4.4. Characterizing the Z-matrix: Discrete duality.** The Discrete linear system (4.3) represents a more significant change to the linear system than a mere renormalization of the Gram matrix. The matrix elements are given not in terms of inner products but in terms of discrete function evaluations. Thus, we have to consider duality with respect to evaluation in the discrete grid. Fortunately, Fourier series are orthogonal with respect to evaluations on a periodic equispaced grid. Consider a set of  $L$  normalized and univariate Fourier basis functions and their associated (periodic) equispaced grid  $\{x_l\}_{l=1}^L$ ; then we have

$$(4.5) \quad \sum_{l=1}^L \phi_i(x_l) \overline{\phi_j(x_l)} = L \delta_{i-j}, \quad 1 \leq i, j \leq L.$$

Equivalently, as is well known, the full DFT matrix  $F$  of size  $L \times L$  with entries  $\phi_i(x_l)$  has inverse  $F^*/L$ . The discrete least squares problem (4.3) follows by choosing a subset of  $M < L$  points that belong to a subdomain  $\Omega$  and a subset of  $N < M$  basis functions. Thus,  $A \in \mathbb{C}^{M \times N}$  is a submatrix of  $F$ , with rows corresponding to the selected points and columns corresponding to the selected basis functions. In this discrete setting, owing to (4.5), we can choose

$$(4.6) \quad Z = A/L.$$

An interpretation can be given as follows. The multiplication with  $A$  corresponds to an extension from  $N$  to  $L$  Fourier coefficients, followed by the DFT of length  $L$ , and followed by the restriction to  $M$  points in  $\Omega$ . Multiplication with  $Z^*$  corresponds to extension by zero-padding in the time domain, followed by the inverse DFT, and restriction in the frequency domain. The vector  $x = Z^*b$  is an accurate solution of the system  $Ax = b$  in the special case where the sampled function is compactly supported on  $\Omega$ . Indeed, in that case zero-padding results in a smooth and periodic function, for which the discrete inverse Fourier transform gives an accurate approximation.

In Figure 4.3 we show the singular value patterns of  $A$ ,  $Z^*$ , and  $A - AZ^*A$  for the discrete Fourier extension from  $[-\frac{1}{2}, \frac{1}{2}]$  to  $[-1, 1]$ , with  $N = 201$ ,  $M \approx 2N$ , and  $L \approx 2M$ . The singular values of  $A$  cluster near  $\sqrt{L}$ , and those of  $Z = A/L$  near  $1/\sqrt{L}$ . The rank of the plunge region is known to be  $\mathcal{O}(\log(N))$ . The matrix  $A - AZ^*A$  isolates this plunge region and thus has rank  $\mathcal{O}(\log(N))$ . Since applying  $A$  and  $Z^*$  is  $\mathcal{O}(N \log(N))$  while the rank of the problem in the first step is  $\mathcal{O}(\log(N))$ , the overall computational complexity of the AZ algorithm is in this case  $\mathcal{O}(N \log^2(N))$ . Even though the condition number of  $A$  grows exponentially with increasing  $N$ , the residual norm  $\|Ax - b\|$  is on the order of the truncation threshold  $\epsilon = 10^{-10}$ , while the coefficient norm remains bounded as predicted by Lemma 2.2. Note that the existence

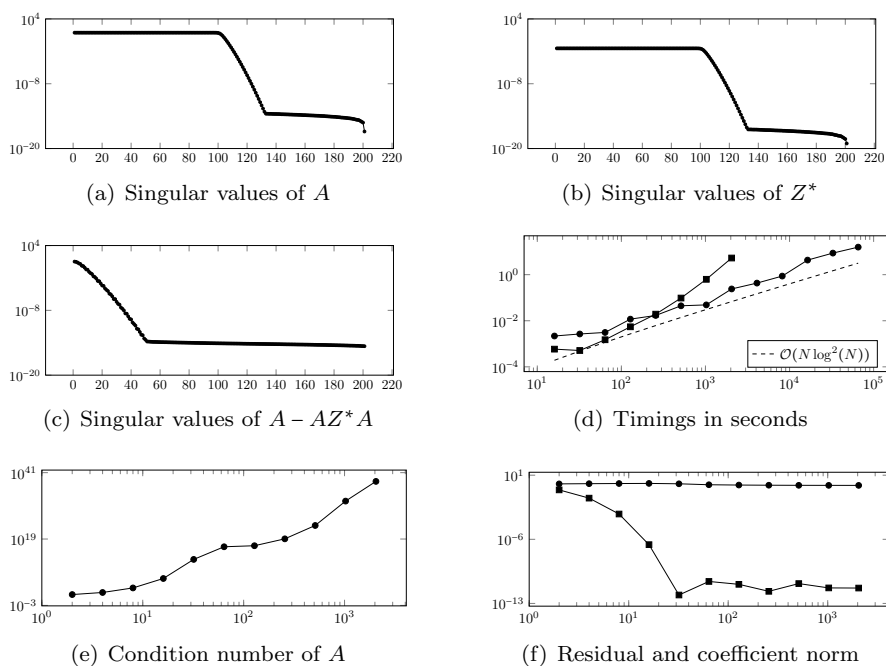


FIG. 4.3. Discrete Fourier extension approximation on the domain  $[-\frac{1}{2}, \frac{1}{2}]$ , using orthonormal Fourier series on  $[-1, 1]$ . In the three first panels  $N = 201$ . The fourth panel shows the timings in seconds for calculating the approximation to  $f(x) = x$  for  $N = 16, 32, \dots, 2^{16}$  using the AZ algorithm with a randomized SVD (dots) and a direct solver (squares). The dashed line shows  $\mathcal{O}(N \log^2(N))$ . The final two panels show the condition number of  $A$  (left), the residual  $\|Ax - b\|$  (right, squares), and coefficient norm  $\|x\|$  (right, dots) for increasing  $N$ .

of a stable least squares fit here is guaranteed analytically: the figure shows it is also recovered numerically by AZ.

The setting is entirely the same in two dimensions, but the plunge region is larger relative to the size of the overall approximation problem [16]. The AZ algorithm was used to produce the two-dimensional example in Figure 4.1.

**4.5. Generalization to other extension frames.** The setting can be generalized to other bases that satisfy a discrete orthogonality condition similar to (4.5), for example, orthogonal polynomials with their roots as sampling points. We consider the approximation of a smooth function on  $[-\frac{1}{2}, \frac{1}{2}]$ , using Legendre polynomials up to degree  $N - 1$  which are orthogonal on  $[-1, 1]$ . We sample the function in  $M$  Legendre nodes in the subinterval  $[-\frac{1}{2}, \frac{1}{2}]$ . In order to ensure oversampling, these nodes are the roots of a Legendre polynomial of higher degree  $L$ , where  $L$  is chosen such that the restricted point set has size  $M = L/2 > N$ . The resulting singular value patterns of  $A$ ,  $Z$ , and  $A - AZ^*A$  are shown in the first three panels of Figure 4.4. The singular values of  $A$  and  $Z$  do not cluster around 1 as in the case of Fourier extension. Yet, we can still isolate the plunge region that is present in  $A$  by the matrix  $A - AZ^*A$ .

The discrete orthogonality condition of Legendre polynomials on the full grid of length  $L$  is

$$(4.7) \quad \sum_{l=1}^L w_l P_i(x_l) P_j(x_l) = h_i^2 \delta_{i-j}, \quad 0 \leq i, j \leq L-1,$$



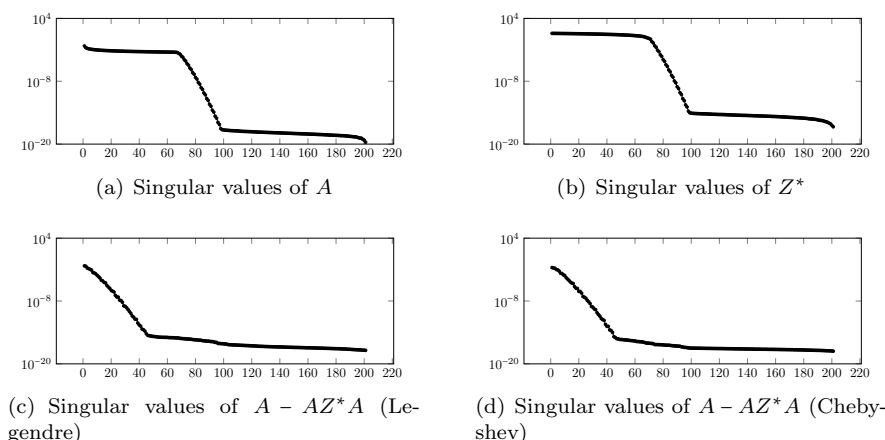


FIG. 4.4. Legendre extension approximation on the domain  $[-\frac{1}{2}, \frac{1}{2}]$ , using classical Legendre series on  $[-1, 1]$  for  $N = 201$ . The collocation points are the Legendre points, and an oversampling factor of 2 was chosen. The last panel (d) is analogous to (c) but is based on Chebyshev extension with Chebyshev roots.

where  $w_l$  are the Gauss–Legendre weights associated with the roots  $x_l$ , and  $h_i = \|P_i\|_{[-1,1]}$  is the norm of  $P_i$ . As in the Fourier case, there is a large  $L \times L$  matrix  $F$  with entries  $F_{i,j} = P_{i-1}(x_{j-1})$ . This matrix has inverse  $F^{-1} = DF^*W$ , where  $W$  and  $D$  are diagonal matrices with entries  $w_i$  and  $h_i^{-2}$ , respectively. The discrete least squares matrix  $A$  is a submatrix of  $F$ , with columns selected corresponding to the degrees of freedom  $(0, \dots, N-1)$  and rows selected corresponding to the points in the subinterval  $[-\frac{1}{2}, \frac{1}{2}]$ . We choose  $Z$  as the corresponding subblock of  $(F^{-1})^* = WFD$ . The size of the plunge region in this setting is not known in literature and is the topic of a separate study.

The same methodology can be applied using Chebyshev polynomials and Chebyshev roots. Here, due to the relation between trigonometric polynomials and Chebyshev polynomials, there is a fast matrix-vector product for  $A$  and for  $Z$  based on the discrete cosine transform. The singular value patterns of  $A$  and  $Z^*$  are not shown for this case, but they are similar to the Fourier extension case in Figure 4.3. The singular value pattern of  $A - AZ^*A$  is in the bottom right panel of Figure 4.4. As in the Fourier extension case the rank of this matrix is  $\mathcal{O}(\log(N))$ .

Finally, the Chebyshev roots can also be replaced by the Chebyshev extremae, in which case a discrete orthogonality property based on Clenshaw–Curtis quadrature (as opposed to Gauss–Chebyshev) yields a suitable matrix  $Z$ . The different choice of discretization points corresponds to a different norm for the continuous approximation problem: we find the best approximation with respect to either the Legendre-weighted norm ( $w(x) = 1$ , Clenshaw–Curtis) or the Chebyshev-weighted norm ( $w(x) = \frac{1}{\sqrt{1-x^2}}$ , Gauss–Chebyshev).

**5. Weighted linear combinations of bases.** Let  $\Phi = \{\phi_k\}_{k=0}^\infty$  be a basis of function on  $[-1, 1]$ . Here, we consider *weighted sum frames* of the form

$$(5.1) \quad \Psi = w_1\Phi \cup w_2\Phi = \{w_1(x)\phi_k(x)\}_{k=0}^\infty \cup \{w_2(x)\phi_k(x)\}_{k=0}^\infty,$$

where  $w_1$  and  $w_2$  are two functions of bounded variation on  $[-1, 1]$  such that

$$(5.2) \quad w_1(x)^2 + w_2(x)^2 > 0 \quad \forall x \in [-1, 1].$$

**5.1. The canonical dual frame and  $(A, Z)$ .** Assume that the basis  $\Phi$ , truncated after  $N$  terms to  $\Phi_N$ , has an associated least squares grid  $\{x_m\}_{m=1}^M$  of length  $M > N$ . The corresponding discretization matrix  $A_\Phi \in \mathbb{C}^{M \times N}$  has entries  $A_{m,n} = \phi_n(x_m)$ . Assume also that the least squares system  $A_\Phi x = b$  is solved by  $Z_\Phi^* b$ . In other words, we assume that the basis  $\Phi$  has an  $(A_\Phi, Z_\Phi)$  combination such that  $Z_\Phi^*$  yields exact reconstruction for all functions in the span of  $\Phi_N$ .

If the basis  $\Phi_N$  is a Fourier basis or Chebyshev polynomial basis, then a least squares problem for this basis is solved by a truncated FFT or discrete cosine transform—in this case  $A_\Phi$  and  $Z_\Phi$  have efficient implementations.

The  $A$  matrix for the weighted sum frame is given by

$$(5.3) \quad A = \begin{pmatrix} W_1 A_\Phi & W_2 A_\Phi \end{pmatrix},$$

where  $W_{1,2} \in \mathbb{C}^{M \times M}$  are diagonal matrices with entries  $w_{\{1,2\}}(x_m)$  on the diagonal.

The  $Z$  matrix follows from the construction of a dual frame. The canonical dual frame for (5.1) is given by (see also [2, Example 3])

$$\tilde{\Phi} = \left\{ \frac{1}{|w_1(x)|^2 + |w_2(x)|^2} \phi_k(x) \right\}_{k=0}^{\infty}.$$

Note that condition (5.2) was imposed in order to guarantee the existence of this dual frame. We illustrate by example in what follows that a suitable  $Z$  matrix for the weighted sum frame is

$$(5.4) \quad Z = \begin{bmatrix} W^\dagger W_1 Z_\Phi & W^\dagger W_2 Z_\Phi \end{bmatrix},$$

where  $W \in \mathbb{R}^{M \times M}$  is a diagonal matrix with entries  $W_{m,m} = |w_1(x_m)|^2 + |w_2(x_m)|^2$ . Equation (5.4) is written in terms of the pseudoinverse  $W^\dagger$  since these diagonal entries might be zero.

**5.2. Weighted combination of bases.** Consider the function

$$f(x, y) = \cos(2\pi(x + y)) + \sqrt{x^2 + y^2} \sin(1 + 2\pi(x + y))$$

on the square domain  $[-1, 1]^2$ . We approximate  $f$  with a weighed sum frame

$$\Psi = \Phi \cup \sqrt{x^2 + y^2} \Phi,$$

where  $\Phi$  is a tensor product of Chebyshev polynomials.

The function  $f$  has a square root singularity at the origin. The approximation space is chosen such that it effectively captures the singularity. It is obvious that  $f$  can be well approximated in the span of the weighted sum frame. However, standard algorithms cannot be used to compute such an approximation if we are only allowed to sample  $f$  (rather than, for example, its smooth and singular part separately). We show that the AZ algorithm with the above choices of  $A$  and  $Z$  is an effective means to this. The results are shown in Figure 5.1.

The study of the size of the plunge region in this example, which demonstrates the computational complexity of the AZ algorithm, is the topic of a separate study. For example, see the forthcoming paper [20].

**5.3. Weighted combination of frames.** It is not necessary to assume that  $\Phi$  is a basis. We observe that the formulas remain valid if instead  $\Phi$  is a frame with a

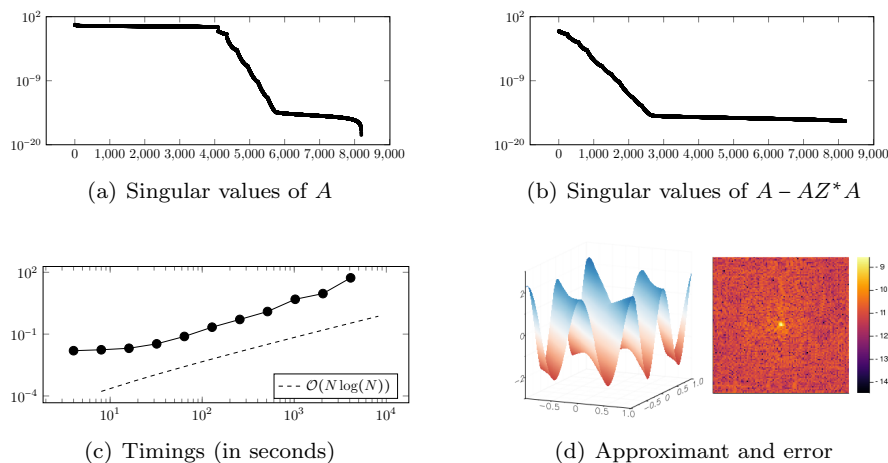


FIG. 5.1. Illustration of the AZ algorithm for approximation with  $\Phi \cup \sqrt{x^2 + y^2}\Phi$ , where  $\Phi$  is a  $64 \times 64$  tensor product of Chebyshev polynomials on  $[-1, 1]^2$ . The collocation points are the two-dimensional Cartesian product of a 128-point Chebyshev grid. In the last panel: the approximation of  $f(x, y) = \cos(2\pi(x + y)) + \sqrt{x^2 + y^2} \sin(1 + 2\pi(x + y))$  and the  $\log_{10}$  error of this approximation.

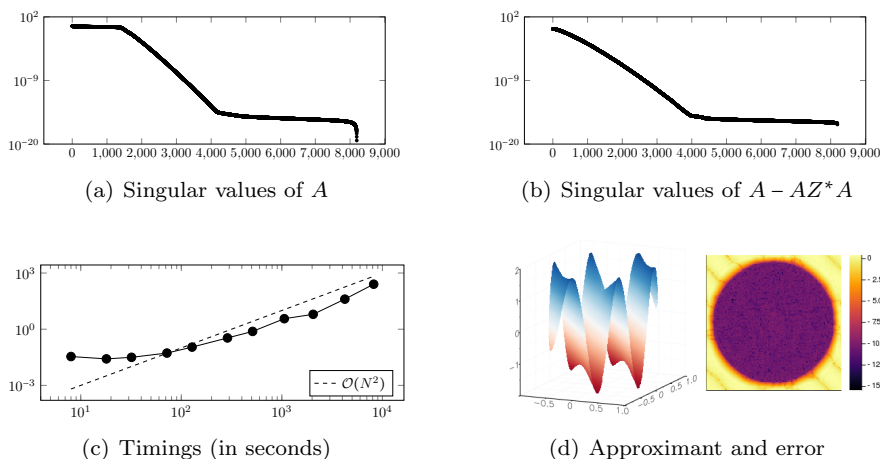


FIG. 5.2. Approximation with  $\Phi \cup \sqrt{x^2 + y^2}\Phi$  on a circular domain with center  $(0, 0)$  and radius 0.9, where  $\Phi$  is a  $64 \times 64$  tensor product of Chebyshev polynomials on  $[-1, 1]^2$ . The collocation points are the two-dimensional Cartesian product of a 128-point Chebyshev grid. In the last panel: the approximation of  $f(x, y) = \cos(2\pi(x + y)) + \sqrt{x^2 + y^2} \sin(1 + 2\pi(x + y))$  and the logarithmic error of this approximation.

suitable known  $(A_\Phi, Z_\Phi)$  combination. In that case, the matrix  $A$  given by (5.3) and  $Z$  given by (5.4) form a suitable  $(A, Z)$  combination for the weighted sum frame.

We approximate the same singular function  $f$  as in the previous section, but instead of using a rectangle we restrict samples of  $f$  to a disk with center  $(0, 0)$  and radius 0.9. Thus, we make a weighted linear combination of Chebyshev extension frames. A suitable  $(A, Z)$  combination for Chebyshev extension was described in section 4, and the experiment shows it can be used as part of a more complicated  $(A, Z)$  combination for the weighted combination. This is shown in Figure 5.2.

**6. B-spline extension frames.** In section 4 we considered extension frames using an orthogonal basis on a bounding box. In that section we emphasized the connection to continuous and discrete dual bases for the selection of  $Z$ . This connection is explored in more detail for extensions based on B-spline bases in [6]. Since B-splines are compactly supported, the collocation matrix  $A$  is highly sparse. B-splines are not orthogonal, but several different dual bases can be identified, and it is shown in [6] that each of these leads to a suitable  $Z$  matrix. Some choices of dual bases lead to a sparse  $Z$  and even a sparse matrix  $A - AZ^*A$ , with corresponding advantages for speed and efficiency. We refer to [6] for the analysis and examples.

**7. Weighted least squares approximation.** In this final example we show how an existing efficient solver for a least squares approximation can be used to solve weighted variants of the problem efficiently as well. Thus, assume that a least squares problem is given in the form

$$(7.1) \quad Ax = b,$$

where  $A \in \mathbb{C}^{M \times N}$  with  $M > N$ , and that a  $Z$  matrix is known that approximates  $A(A^*A)^{-1}$ , i.e., such that  $Z^*b$  solves the unweighted least squares problem. In this section we consider a Fourier series of length  $N$  and a discrete least squares approximation to a given function based on  $M > N$  equispaced samples. Owing to the continuous and discrete orthogonality properties of Fourier series, this problem is solved efficiently simply by computing the inverse FFT of length  $M$  and truncating the result to a vector of length  $N$ .

Next, we consider a diagonal weight matrix  $W \in \mathbb{C}^{M \times M}$  with positive entries  $W_{i,i} = d_i > 0$  that associates a weight to each condition in the rows of  $A$ . The weighted least squares problem is

$$(7.2) \quad WAx = Wb.$$

When the linear system is overdetermined, the weighted least squares solution differs from the unweighted one. The solution is given by

$$x = (A^*W^*WA)^{-1}A^*W^*Wb.$$

The solution of (7.1) minimizes  $\|Ax - b\|$ ; the solution of (7.2) minimizes  $\|W(Ax - b)\|$ .

With an efficient solver for (7.1) at hand, the simplest solution method to solve the weighted problem is to ignore the weight matrix and to return the solution to (7.1). This may deviate from the weighted solution as follows:

$$\min_{1 \leq i \leq M} d_i \|Ax - b\| \leq \|W(Ax - b)\| \leq \max_{1 \leq i \leq M} d_i \|Ax - b\|.$$

The weighted and unweighted least squares problems may have radically different solutions if the ratio of the weights is large, and this is the setting we focus on. Assume without loss of generality that the large ratio of weights is due to having a number of very small weights. We choose a threshold  $\epsilon$ , and we define  $W_\epsilon$  as the diagonal matrix with entry  $i$  equal to zero if  $|d_i| < \epsilon$  and equal to  $d_i$  otherwise. We define the matrix  $\tilde{Z}$  in terms of the pseudoinverse of  $W_\epsilon$ ,

$$\tilde{Z} = W_\epsilon^\dagger Z.$$

We set out to solve  $\tilde{A}x = \tilde{b}$  with  $\tilde{A} = WA$  and  $\tilde{b} = Wb$ . With these choices, the AZ algorithm is given explicitly in Algorithm 7.1.

**Algorithm 7.1** The AZ algorithm for weighted approximation**Input:**  $A, Z \in \mathbb{C}^{M \times N}$ ,  $W = \text{diag}([d_1, \dots, d_M])$ , with  $d_i > 0$ ,  $b \in \mathbb{C}^M$ ,  $\epsilon > 0$ **Output:**  $x \in \mathbb{C}^N$  such that  $WAx \approx Wb$ 

- 1: Solve  $(I - WA(W_\epsilon^\dagger Z)^*)WAx_1 = (I - WA(W_\epsilon^\dagger Z)^*)Wb$
- 2:  $x_2 \leftarrow Z^*W_\epsilon^\dagger W(b - Ax_1)$
- 3:  $x \leftarrow x_1 + x_2$

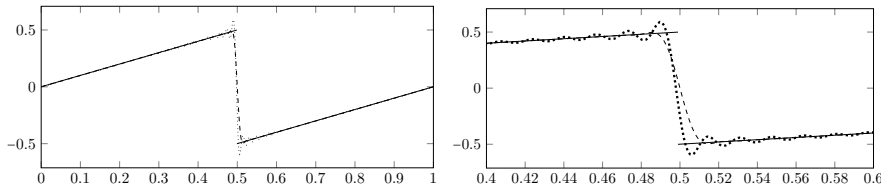


FIG. 7.1. Left panel: A function with a jump at  $x = 0.5$  (full), the unweighted approximation (dotted) containing the Gibbs phenomenon, and the weighted approximation (dashed). Right panel: Zoom of the left panel.

One interpretation of the algorithm is as follows. The known  $Z$  matrix is efficient, but it solves the wrong problem, namely, the unweighted one. It is used in step 2. Step 1 is slow, but it solves the right problem, namely, the weighted one. If  $\epsilon = 0$ , then  $W_\epsilon^\dagger = W^{-1}$ . In this case, the system in the first step has rank zero, and the problem is solved in step 2, yielding the solution to the unweighted problem: efficient, but possibly inaccurate. If on the other hand  $\epsilon > \|W\|$ , then  $W_\epsilon^\dagger = 0$ , and the problem is solved in step 1: accurate, but possibly slow. By varying  $\epsilon$ , one obtains a solution somewhere in between these two extreme cases. The tradeoff is between accuracy of the weighted problem and efficiency of the algorithm.

For an example we return to function approximation: we use a Fourier series to approximate a piecewise smooth, yet discontinuous function. This is known to suffer the Gibbs phenomenon, resulting in an oscillatory overshoot at the discontinuities. In order to obtain a smooth approximation to the smooth parts of  $f$ , we may want to assign a very small weight to the function accuracy near the discontinuity. This is a simple alternative to other smoothing methods, such as spectral filtering techniques.

The function shown in Figure 7.1 is periodic on  $[0, 1]$ , with a jump at 0.5. We use a Fourier series with  $N = 121$  terms on  $[0, 1]$ . The discrete weights are based on sampling the function  $w(x) = (x - 0.5)^2$ , which assigns greater weight to points further away from the discontinuity. We solve the weighted approximation problem with the AZ algorithm varying  $\epsilon$  from  $1e^{-6}$  to  $1e^{+1}$ . The results are shown in Figure 7.2. For small  $\epsilon$ , the rank of the system in the first step is small, but the solution deviates substantially from the true weighted solution. For larger  $\epsilon$ , the rank of the system in step 1 increases, but the AZ algorithm approximates the true solution better. Figure 7.1 shows that the AZ algorithm produces a nonoscillatory, smooth approximation to  $f$ , while Figure 7.2 shows that this approximation is highly accurate away from the jump.

**8. Concluding remarks.** We introduced a three step algorithm and showed that it can be used to efficiently solve various problems of interest in function approximation. The examples were chosen such that the matrix  $Z$  could be devised analytically. However, as the examples show, it is often possible to generate  $(A, Z)$

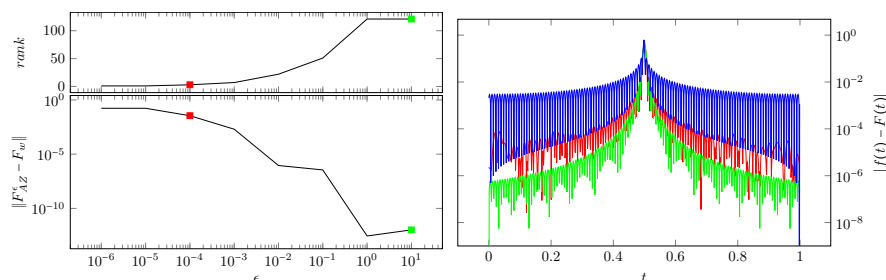


FIG. 7.2. Left panel. Top: the rank of step 1 of the AZ algorithm. Bottom: the norm of the difference between the solution given by the AZ algorithm and the solution of  $W Ax = W b$ . Right panel, from top to bottom: The difference between the weighed solution and the unweighted solution (blue, top), the error of the AZ solution with  $\epsilon = 10^{-4}$  (red, middle), and of the AZ solution with  $\epsilon = 10$  (green, bottom).

combinations for complicated approximation problems from  $(A, Z)$  combinations of simpler subproblems.

The efficiency of the AZ algorithm hinges on the numerical rank of the system that is solved in the first step. Plots of the rapidly decaying singular values are provided for each example. Theoretical bounds on these numerical ranks are a topic of ongoing research and beyond the scope of this paper. See [15, 16] for discussion of rank for the univariate and multivariate Fourier extension problem, and see [20] for discussion of rank for weighted combinations of certain univariate bases.

In our examples above, we have used the AZ algorithm in combination with a randomized low-rank SVD solver for step 1. Iterative solvers such as LSQR and LSMR [17, 9] can also be used for the first step of the AZ algorithm. Experiments indicate that they do not typically yield high accuracy, due to the ill-conditioning of the system. However, if only a few digits of accuracy are required, they can be several times more efficient than using a direct solver in step 1. In fact, low accuracy approximations using iterative methods have been described in extrapolation methods for band-limited signals [19], which is highly comparable to the Fourier extension example of section 4. Our results would suggest to use the iterative solver in step 1 of the AZ algorithm in this application, rather than using it for the original system  $Ax = b$ .

## REFERENCES

- [1] B. ADCOCK AND D. HUYBRECHS, *Frames and Numerical approximation II: Generalized Sampling*, e-print, 2018, <https://arxiv.org/abs/1802.01950>.
- [2] B. ADCOCK AND D. HUYBRECHS, *Frames and numerical approximation*, SIAM Rev., 61 (2019), pp. 443–473.
- [3] J. P. BOYD, *A comparison of numerical algorithms for Fourier extension of the first, second, and third kinds*, J. Comput. Phys., 178 (2002), pp. 118–160.
- [4] O. P. BRUNO, Y. HAN, AND M. M. POHLMAN, *Accurate, high-order representation of complex three-dimensional surfaces via Fourier continuation analysis*, J. Comput. Phys., 227 (2007), pp. 1094–1125.
- [5] O. CHRISTENSEN, *Frames and Bases: An Introductory Course*, Springer, Basel, 2008.
- [6] V. COPPÉ AND D. HUYBRECHS, *Efficient Function Approximation on General Bounded Domains Using Splines on a Cartesian Grid*, e-print, 2019, <https://arxiv.org/abs/1911.07894>.
- [7] V. COPPÉ, D. HUYBRECHS, R. MATTHYSEN, AND M. WEBB, *Examples to Accompany the Article Introducing the AZ Algorithm*, Github repository, 2020, <https://github.com/FrameFunVC/AZNotebook>.

- [8] A. EDELMAN, P. MCCORQUODALE, AND S. TOLEDO, *The future fast Fourier transform?*, SIAM J. Sci. Comput., 20 (1999), pp. 1094–1114.
- [9] D.C.-L. FONG AND M. SAUNDERS, *LSMR: An iterative algorithm for sparse least-squares problems*, SIAM J. Sci. Comput., 33 (2010), pp. 2950–2971.
- [10] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, MD, 2013.
- [11] N. HALKO, P. G. MARTINSSON, AND J. A. TROPP, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.
- [12] D. HUYBRECHS, *On the Fourier extension of nonperiodic functions*, SIAM J. Numer. Anal., 47 (2010), pp. 4326–4355.
- [13] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Classics Appl. Math., SIAM, Philadelphia, PA, 1996.
- [14] M. LYON, *A fast algorithm for Fourier continuation*, SIAM J. Sci. Comput., 33 (2011), pp. 3241–3260.
- [15] R. MATTHYSEN AND D. HUYBRECHS, *Fast algorithms for the computation of Fourier extensions of arbitrary length*, SIAM J. Sci. Comput., 38 (2016), pp. A899–A922.
- [16] R. MATTHYSEN AND D. HUYBRECHS, *Function approximation on arbitrary domains using Fourier frames*, SIAM J. Numer. Anal., 56 (2018), pp. 1360–1385.
- [17] C. C. PAIGE AND M. A. SAUNDERS, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.
- [18] D. SLEPIAN, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty V: The discrete case*, Bell Labs Tech. J., 57 (1978), pp. 1371–1430.
- [19] T. STROHMER, *On discrete band-limited signal extrapolation*, Contemp. Math., 190 (1995), pp. 323–337.
- [20] M. WEBB, *The Plunge Region in Approximation by Singularly Enriched Trigonometric Frames*, Tech. Report to appear, 2020.