

## MIT Open Access Articles

*Matrix Probing and its Conditioning*

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

**Citation:** Chiu, Jiawei, and Laurent Demanet. "Matrix Probing and Its Conditioning." SIAM Journal on Numerical Analysis 50.1 (2012): 171–193. Web.

**As Published:** <http://dx.doi.org/10.1137/110825972>

**Publisher:** Society for Industrial and Applied Mathematics

**Persistent URL:** <http://hdl.handle.net/1721.1/71842>

**Version:** Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

**Terms of Use:** Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



## MATRIX PROBING AND ITS CONDITIONING\*

JIAWEI CHIU<sup>†</sup> AND LAURENT DEMANET<sup>‡</sup>

**Abstract.** When a matrix  $A$  with  $n$  columns is known to be well-approximated by a linear combination of basis matrices  $B_1, \dots, B_p$ , we can apply  $A$  to a random vector and solve a linear system to recover this linear combination. The same technique can be used to obtain an approximation to  $A^{-1}$ . A basic question is whether this linear system is well-conditioned. This is important for two reasons: a well-conditioned system means (1) we can invert it and (2) the error in the reconstruction can be controlled. In this paper, we show that if the Gram matrix of the  $B_j$ 's is sufficiently well-conditioned and each  $B_j$  has a high numerical rank, then  $n \propto p \log^2 n$  will ensure that the linear system is well-conditioned with high probability. Our main application is probing linear operators with smooth pseudodifferential symbols such as the wave equation Hessian in seismic imaging [L. Demanet et al., *Appl. Comput. Harmonic Anal.*, 32 (2012), pp. 155–168]. We also demonstrate numerically that matrix probing can produce good preconditioners for inverting elliptic operators in variable media.

**Key words.** randomized algorithms, pseudodifferential operators, preconditioners, conditioning, system identification, large deviation estimates

**AMS subject classifications.** 68W20, 15A12, 65F35, 65F08, 65N75, 65T40

**DOI.** 10.1137/110825972

**1. Introduction.** The earliest randomized algorithms include Monte Carlo integration and Monte Carlo Markov chains [1]. These are standard techniques in numerical computing with widespread applications from physics and econometrics to machine learning. However, they are often seen as the methods of last resort, because they are easy to implement but produce solutions of uncertain accuracy.

In the last few decades, a new breed of randomized algorithms has been developed by the computer science community. These algorithms remain easy to implement and have failure probabilities that are provably negligible. In other words, we have rigorous theory to ensure that these algorithms perform consistently well. Moreover, their time complexity can be as good as the most sophisticated deterministic algorithms, e.g., Karp–Rabin’s pattern matching algorithm [17] and Karger’s min-cut algorithm [16].

In recent years, equally attractive randomized algorithms are being developed in the numerical community. For example, in compressed sensing [4], we can recover sparse vectors with random measurement matrices and  $\ell^1$  minimization. Another interesting example is that we can build a good low rank approximation of a matrix by *applying it to random vectors* [14].

Our work carries a similar flavor: often, the matrix  $A$  can be approximated as a linear combination of a small number of matrices and the idea is to obtain these coefficients by applying  $A$  to a random vector or just a few of them. We call this “forward matrix probing.” What is even more interesting is that we can also probe for  $A^{-1}$  by applying  $A$  to a random vector. We call this “backward matrix probing” for a reason that will be clear in section 1.5.

---

\*Received by the editors February 28, 2011; accepted for publication (in revised form) November 1, 2011; published electronically February 9, 2012.

<http://www.siam.org/journals/sinum/50-1/82597.html>

<sup>†</sup>Corresponding author. Department of Mathematics, MIT, Cambridge, MA 02139 (jiawei@mit.edu). This author’s work was supported by an A\*STAR fellowship from Singapore.

<sup>‡</sup>Department of Mathematics, MIT, Cambridge, MA 02139 (laurent@math.mit.edu). This author’s work was supported in part by a grant from the National Science Foundation and the Alfred P. Sloan Foundation.

Due to approximation errors, the output of “backward probing,” denoted as  $C$ , is only an approximate inverse. Nevertheless, as we will see in section 4,  $C$  serves very well as a preconditioner for inverting  $A$ , and we believe that its performance could match that of multigrid methods for elliptic operators in smooth media.

The idea of “matrix probing” is not new. For example, Chan and Resasco [6] and Chan and Mathew [5] use the technique to approximate  $A$  with a sparse matrix. Another example is the work by Pfander, Rauhut, and Tanner [21], where the same idea is used in a way typical in compressed sensing. In the next section, we will see that their setup is fundamentally different from ours.

**1.1. Forward matrix probing.** Let  $\mathcal{B} = \{B_1, \dots, B_p\}$ , where each  $B_j \in \mathbb{C}^{m \times n}$  is called a basis matrix. Note that  $\mathcal{B}$  is specified in advance. Let  $u$  be a Gaussian or a Rademacher sequence, that is, each component of  $u$  is independent and is either a standard normal variable or  $\pm 1$  with equal probability.

Define the matrix  $L \in \mathbb{C}^{m \times p}$  such that its  $j$ th column is  $B_j u$ . Let  $A \in \mathbb{C}^{m \times n}$  be the matrix we want to probe and suppose  $A$  lies in the span of  $\mathcal{B}$ . Say

$$A = \sum_{i=1}^p c_i B_i \text{ for some } c_1, \dots, c_p \in \mathbb{C}.$$

Observe that  $Au = \sum_{i=1}^p c_i (B_i u) = Lc$ . Given the vector  $Au$ , we can obtain the coefficient vector  $c = (c_1, \dots, c_p)^T$  by solving the linear system

$$(1.1) \quad Lc = Au.$$

In practice,  $A$  is not exactly in the span of a small  $\mathcal{B}$  and (1.1) has to be solved in a least squares sense, that is,  $c = L^+(Au)$ , where  $L^+$  is the pseudoinverse of  $L$ .

We will assume that  $p \leq n$ . Otherwise there are more unknowns than equations and there is no unique solution if there is any. This differs from the setup in [21], where  $n \gg p$  but  $A$  is assumed to be a sparse linear combination of  $B_1, \dots, B_p$ .

**1.2. Conditioning of  $L$ .** Whether (1.1) can be solved accurately depends on  $\text{cond}(L)$ , the condition number of  $L$ . This is the ratio between the largest and the smallest singular values of  $L$  and can be understood as how different  $L$  can stretch or shrink a vector.

Intuitively, whether  $\text{cond}(L)$  is small depends on the following two properties of  $\mathcal{B}$ :

1. The  $B_i$ 's “act differently” in the sense that  $\langle B_j, B_k \rangle \simeq \delta_{jk}$  for any  $1 \leq j, k \leq p$ .<sup>1</sup>
2. Each  $B_i$  has a high rank so that  $B_1 u, \dots, B_p u \in \mathbb{C}^n$  exist in a high dimensional space.

When  $\mathcal{B}$  possesses these two properties and  $p$  is sufficiently small compared to  $n$ , it makes sense that  $L$ 's columns,  $B_1 u, \dots, B_p u$ , are likely to be independent, thus guaranteeing that  $L$  is invertible, at least.

We now make the above two properties more precise. Let

$$(1.2) \quad M = L^* L \in \mathbb{C}^{p \times p} \text{ and } N = \mathbb{E}M.$$

Clearly,  $\text{cond}(M) = \text{cond}(L)^2$ . If  $\mathbb{E}M$  is ill-conditioned, there is little chance that  $M$  or  $L$  is well-conditioned. This can be related to Property 1 by observing that

$$(1.3) \quad N_{jk} = \mathbb{E}M_{jk} = \text{Tr}(B_j^* B_k) = \langle B_j, B_k \rangle.$$

<sup>1</sup>Note that  $\langle \cdot, \cdot \rangle$  is the Frobenius inner product and  $\delta_{jk}$  is the Kronecker delta.

If  $\langle B_j, B_k \rangle \simeq \delta_{jk}$ , then the Gram matrix  $N$  is approximately the identity matrix which is well-conditioned. Hence, a more quantitative way of putting Property 1 is that we have control over  $\kappa(B)$  defined as follows.

DEFINITION 1.1. Let  $\mathcal{B} = \{B_1, \dots, B_p\}$  be a set of matrices. Define its condition number  $\kappa(\mathcal{B})$  as the condition number of the matrix  $N \in \mathbb{C}^{p \times p}$ , where  $N_{jk} = \langle B_j, B_k \rangle$ .

On the other hand, Property 2 can be made precise by saying that we have control over  $\lambda(\mathcal{B})$  as defined below.

DEFINITION 1.2. Let  $A \in \mathbb{C}^{m \times n}$ . Define its weak condition number<sup>2</sup> as

$$\lambda(A) = \frac{\|A\| n^{1/2}}{\|A\|_F}.$$

Let  $\mathcal{B}$  be a set of matrices. Define its (uniform) weak condition number as

$$\lambda(\mathcal{B}) = \max_{A \in \mathcal{B}} \lambda(A).$$

We justify the nomenclature as follows. Suppose  $A \in \mathbb{C}^{n \times n}$  has condition number  $k$ ; then  $\|A\|_F^2 = \sum_{i=1}^n \sigma_i^2 \geq n\sigma_{\min}^2 \geq n\|A\|^2/k^2$ . Taking the square root, we obtain  $\lambda(A) \leq k$ . In other words, any well-conditioned matrix is also weakly well-conditioned. And like the usual condition number,  $\lambda(A) \geq 1$  because we always have  $\|A\|_F \leq n^{1/2}\|A\|$ .

The numerical rank of a matrix  $A$  is  $\|A\|_F^2/\|A\|^2 = n\lambda(A)^{-2}$ ; thus having a small  $\lambda(A)$  is the same as having a high numerical rank. We also want to caution the reader that  $\lambda(\mathcal{B})$  is defined very differently from  $\kappa(\mathcal{B})$  and is not a weaker version of  $\kappa(\mathcal{B})$ .

Using classical concentration inequalities, it was shown [9] that when  $\lambda(\mathcal{B})$  and  $\kappa(\mathcal{B})$  are fixed,  $p = \tilde{O}(n^{1/2})^3$  will ensure that  $L$  is well-conditioned with high probability.

In this paper, we establish a stronger result, namely, that  $p = \tilde{O}(n)$  suffices. The implication is that we can expect to recover  $\tilde{O}(n)$  instead of  $\tilde{O}(n^{1/2})$  coefficients. The exact statement is presented below.

THEOREM 1.3 (main result). Let  $C_1, C_2 > 0$  be numbers given by Remark B.1 in the appendix. Let  $\mathcal{B} = \{B_1, \dots, B_p\}$ , where each  $B_j \in \mathbb{C}^{m \times n}$ . Define  $L \in \mathbb{C}^{n \times p}$  such that its  $j$ th column is  $B_j u$ , where  $u$  is either a Gaussian or a Rademacher sequence. Let  $M = L^* L$ ,  $N = \mathbb{E}M$   $\kappa = \kappa(\mathcal{B})$ , and  $\lambda = \lambda(\mathcal{B})$ . Suppose

$$n \geq p(C\kappa\lambda \log n)^2 \text{ for some } C \geq 1.$$

Then

$$\mathbb{P}\left(\|M - N\| \geq \frac{t\|N\|}{\kappa}\right) \leq 2C_2 p n^{1-\alpha}, \text{ where } \alpha = \frac{tC}{eC_1}.$$

The number  $C_1$  is small.  $C_2$  may be large but it poses no problem because  $n^{-\alpha}$  decays very fast with larger  $n$  and  $C$ . With  $t = 1/2$ , we deduce that with high probability,

$$\text{cond}(M) \leq 2\kappa + 1.$$

In general, we let  $0 < t < 1$  and for the probability bound to be useful, we need  $\alpha > 2$ , which implies  $C > 2eC_1 > 1$ . Therefore the assumption that  $C \geq 1$  in the theorem can be considered redundant.

<sup>2</sup>Throughout the paper,  $\|\cdot\|$  and  $\|\cdot\|_F$  denote the spectral and Frobenius norms respectively.

<sup>3</sup>Note that  $\tilde{O}(n)$  denotes  $O(n \log^c n)$  for some  $c > 0$ . In other words, ignore log factors.

We remark that Rauhut and Tropp have a new result (a Bernstein-like tail bound) that may be used to refine the theorem. This will be briefly discussed in section 4.1, where we conduct a numerical experiment.

Note that when  $u$  is a Gaussian sequence,  $M$  resembles a Wishart matrix for which the distribution of the smallest eigenvalue is well-studied [11]. However, each row of  $L$  is not independent, so results from random matrix theory cannot be used in this way.

An intermediate result in the proof of Theorem 1.3 is the following. It conveys the essence of Theorem 1.3 and may be easier to remember.

**THEOREM 1.4.** *Assume the same setup as in Theorem 1.3. Suppose  $n = \tilde{O}(p)$ . Then*

$$\mathbb{E} \|M - N\| \leq C(\log n) \|N\| (p/n)^{1/2} \lambda \text{ for some } C > 0.$$

A numerical experiment in section 4.1 suggests that the relationship between  $p$  and  $n$  is not tight in the log factor. Our experiment shows that for  $\mathbb{E} \|M - N\| / \|N\|$  to vanish as  $p \rightarrow \infty$ ,  $n$  just needs to increase faster than  $p \log(np)$ , whereas Theorem 1.4 requires  $n$  to grow faster than  $p \log^2 n$ .

Next, we see that when  $L$  is well-conditioned, the error in the reconstruction is also small.

**PROPOSITION 1.5.** *Assume the same setup as in Theorem 1.3. Suppose  $A = \sum_{j=1}^p d_j B_j + E$ , where  $\|E\| \leq \varepsilon$  and assume with high probability*

$$\|M - N\| \leq \frac{t \|N\|}{\kappa} \text{ for some } 0 < t < 1.$$

Let  $c = L^+ A u$  be the recovered coefficients. Then with high probability

$$\left\| A - \sum_{j=1}^p c_j B_j \right\| \leq O \left( \varepsilon \lambda \left( \frac{\kappa p}{1-t} \right)^{1/2} \right).$$

If  $\varepsilon = o(p^{-1/2})$ , then the proposition guarantees that the overall error goes to zero as  $p \rightarrow \infty$ . Of course, a larger  $n$  and more computational effort are required.

**1.3. Multiple probes.** Fix  $n$  and suppose  $p > n$ .  $L$  is not going to be well-conditioned or even invertible. One way around this is to probe  $A$  with multiple random vectors  $u_1, \dots, u_q \in \mathbb{C}^n$  at one go, that is, to solve

$$L'c = A'u,$$

where the  $j$ th column of  $L'$  and  $A'u$  is, respectively,

$$\begin{pmatrix} B_j u_1 \\ \vdots \\ B_j u_q \end{pmatrix} \text{ and } \begin{pmatrix} A u_1 \\ \vdots \\ A u_q \end{pmatrix}.$$

For this to make sense,  $A' = I_q \otimes A$ , where  $I_q$  is the identity matrix of size  $q$ . Also define  $B'_j = I_q \otimes B_j$  and treat the above as probing  $A'$  assuming that it lies in the span of  $\mathcal{B}' = \{B'_1, \dots, B'_p\}$ .

Regarding the conditioning of  $L'$ , we can apply Theorem 1.3 to  $A'$  and  $\mathcal{B}'$ . It is an easy exercise (cf. Proposition A.1) to see that the condition numbers are unchanged,

that is,  $\kappa(\mathcal{B}) = \kappa(\mathcal{B}')$  and  $\lambda(\mathcal{B}) = \lambda(\mathcal{B}')$ . Applying Theorem 1.3 to  $A'$  and  $\mathcal{B}'$ , we deduce that  $\text{cond}(L) \leq 2\kappa + 1$  with high probability provided that

$$nq \propto p(\kappa\lambda \log n)^2.$$

Remember that  $A$  has only  $mn$  degrees of freedom; while we can increase  $q$  as much as we like to improve the conditioning of  $L$ , the problem setup does not allow  $p > mn$  coefficients. In general, when  $A$  has rank  $\tilde{n}$ , its degrees of freedom is  $\tilde{n}(m + n - \tilde{n})$  by considering its SVD.

**1.4. When to probe.** Matrix probing is an especially useful technique when the following holds:

1. We know that the probed matrix  $A$  can be approximated by a small number of basis matrices that are specified in advance. This holds for operators with smooth pseudodifferential symbols, which will be studied in section 3.
2. Each matrix  $B_i$  can be applied to a vector in  $\tilde{O}(\max(m, n))$  time using only  $\tilde{O}(\max(m, n))$  memory.

The second condition confers two benefits. First, the coefficients  $c$  can be recovered fast, assuming that  $u$  and  $Au$  are already provided. This is because  $L$  can be computed in  $\tilde{O}(\max(m, n)p)$  time and (1.1) can be solved in  $O(mp^2 + p^3)$  time by QR factorization or other methods. In the case where increasing  $m, n$  does not require a bigger  $\mathcal{B}$  to approximate  $A$ ,  $p$  can be treated as a constant and the recovery of  $c$  takes only  $\tilde{O}(\max(m, n))$  time.

Second, given the coefficient vector  $c$ ,  $A$  can be applied to any vector  $v$  by summing over  $B_i v$ 's in  $\tilde{O}(\max(m, n)p)$  time. This speeds up iterative methods such as GMRES and Arnoldi.

**1.5. Backward matrix probing.** A compelling application of matrix probing is computing the pseudoinverse  $A^+$  of a matrix  $A \in \mathbb{C}^{m \times n}$  when  $A^+$  is known to be well-approximated in the space of some  $\mathcal{B} = \{B_1, \dots, B_p\}$ . This time, we probe  $A^+$  by applying it to a random vector  $v = Au$ , where  $u$  is a Gaussian or Rademacher sequence that we generate.

Like in section 1.1, define  $L \in \mathbb{C}^{n \times p}$  such that its  $j$ th column is  $B_j v = B_j Au$ . Suppose  $A^+ = \sum_{i=1}^p c_i B_i$  for some  $c_1, \dots, c_p \in \mathbb{C}$ . Then the coefficient vector  $c$  can be obtained by solving

$$(1.4) \quad Lc = A^+ v = A^+ Au.$$

The right-hand side is  $u$  projected onto  $\text{null}(A)^\perp$ , where  $\text{null}(A)$  is the nullspace of  $A$ . When  $A$  is invertible,  $A^+ Au$  is simply  $u$ . We call this “backward matrix probing” because the generated random vector  $u$  appears on the opposite side of the matrix being probed in (1.4). The equation suggests the following framework for probing  $A^+$ .

ALGORITHM 1 (backward matrix probing). *Suppose  $A^+ = \sum_{i=1}^p c_i B_i$ . The goal is to retrieve the coefficients  $c_1, \dots, c_p$ .*

1. *Generate  $u \sim N(0, 1)^n$  independently and identically distributed (iid).*
2. *Compute  $v = Au$ .*
3. *Filter away  $u$ 's components in  $\text{null}(A)$ . Call this  $\tilde{u}$ .*
4. *Compute  $L$  by setting its  $j$ -column to  $B_j v$ .*
5. *Solve for  $c$  the system  $Lc = \tilde{u}$  in a least squares sense.*

In order to perform the filtering in step 3 efficiently, prior knowledge of  $A$  may be needed. For example, if  $A$  is the Laplacian with periodic boundary conditions, its nullspace is the set of constant functions and step 3 amounts to subtracting the

mean from  $u$ . A more involved example can be found in [9]. In this paper, we invert the wave equation Hessian, and step 3 entails building an illumination mask. Further comments on [9] are located in section 4.5 of this paper.

For the conditioning of  $L$ , we may apply Theorem 1.3 with  $\mathcal{B}$  replaced with  $\mathcal{B}_A := \{B_1A, \dots, B_pA\}$  since the  $j$ th column of  $L$  is now  $B_jAu$ . Of course,  $\kappa(\mathcal{B}_A)$  and  $\lambda(\mathcal{B}_A)$  can be very different from  $\kappa(\mathcal{B})$  and  $\lambda(\mathcal{B})$ ; in fact,  $\kappa(\mathcal{B}_A)$  and  $\lambda(\mathcal{B}_A)$  seem much harder to control because it depends on  $A$ . Fortunately, as we shall see in section 3.5, knowing the “order” of  $A^+$  as a pseudodifferential operator helps in keeping these condition numbers small.

When  $A$  has a high dimensional nullspace but has comparable nonzero singular values,  $\lambda(\mathcal{B}_A)$  may be much larger than is necessary. By a change of basis, we can obtain the following tighter result.

**COROLLARY 1.6.** *Let  $C_1, C_2 > 0$  be numbers given by Remark B.1 in the appendix. Let  $A \in \mathbb{C}^{m \times n}$ ,  $\tilde{n} = \text{rank}(A)$ , and  $\mathcal{B}_A = \{B_1A, \dots, B_pA\}$ , where each  $B_j \in \mathbb{C}^{n \times m}$ . Define  $L \in \mathbb{C}^{n \times p}$  such that its  $j$ th column is  $B_jAu$ , where  $u \sim N(0, 1)^n$  iid. Let  $M = L^*L$ ,  $N = \mathbb{E}M$ ,  $\kappa = \kappa(\mathcal{B}_A)$ , and  $\lambda = (\tilde{n}/n)^{1/2}\lambda(\mathcal{B}_A)$ . Suppose*

$$\tilde{n} \geq p(C\kappa\lambda \log \tilde{n})^2 \text{ for some } C \geq 1.$$

Then

$$\mathbb{P}\left(\|M - N\| \geq \frac{t\|N\|}{\kappa}\right) \leq (2C_2p)\tilde{n}^{1-\alpha}, \text{ where } \alpha = \frac{tC}{eC_1}.$$

Notice that  $\tilde{n} = \text{rank}(A)$  has taken the role of  $n$ , and our  $\lambda$  is now  $\max_{1 \leq j \leq p} \frac{\tilde{n}^{1/2}\|B_jA\|}{\|B_jA\|_F}$ , which ignores the  $n - \tilde{n}$  zero singular values of each  $B_jA$  and can be much smaller than  $\lambda(\mathcal{B}_A)$ .

**2. Proofs.**

**2.1. Proof of Theorem 1.3.** Our proof is decoupled into two components: one linear algebraic and one probabilistic. The plan is to collect all the results that are linear algebraic, deterministic in nature, then appeal to a probabilistic result developed in the appendix.

To facilitate the exposition, we use a different notation for this section. We use lowercase letters as *superscripts* that run from 1 to  $p$  and Greek symbols as subscripts that run from 1 to  $n$  or  $m$ . For example, the set of basis matrices is now  $\mathcal{B} = \{B^1, \dots, B^p\}$ .

Our linear algebraic results concern the following variables:

1. Let  $T^{jk} = B^{j*}B^k \in \mathbb{C}^{n \times n}$  and  $T_{\xi\eta} \in \mathbb{C}^{p \times p}$  such that the  $(j, k)$ th entry of  $T_{\xi\eta}$  is the  $(\xi, \eta)$ th entry of  $T^{jk}$ .
2. Let  $Q = \sum_{1 \leq \xi, \eta \leq n} T_{\xi\eta}^* T_{\xi\eta}$ .
3. Let  $S = \sum_{j=1}^p B^j B^{j*} \in \mathbb{C}^{m \times m}$ .
4. Let  $F$  and  $G$  be block matrices  $(T_{\xi\eta})_{1 \leq \xi, \eta \leq n}$  and  $(T_{\xi\eta}^*)_{1 \leq \xi, \eta \leq n}$ , respectively.

The reason for introducing  $T$  is that  $M$  can be written as a quadratic form in  $T_{\xi\eta}$  with input  $u$ :

$$M = \sum_{1 \leq \xi, \eta \leq n} u_\xi u_\eta T_{\xi\eta}.$$

Since  $u_\xi$  has unit variance and zero mean,  $N = \mathbb{E}M = \sum_{\xi=1}^n T_{\xi\xi}$ .

Probabilistic inequalities applied to  $M$  will involve  $T_{\xi\eta}$ , which must be related to  $\mathcal{B}$ . The connection between these  $n$  by  $n$  matrices and  $p$  by  $p$  matrices lies in the identity

$$(2.1) \quad T_{\xi\eta}^{jk} = \sum_{\zeta=1}^m \overline{B_{\zeta\xi}^j} B_{\zeta\eta}^k.$$

The linear algebraic results are contained in the following propositions.

PROPOSITION 2.1. *For any  $1 \leq \xi, \eta \leq n$ ,*

$$T_{\xi\eta} = T_{\eta\xi}^*.$$

Hence,  $T_{\xi\xi}, N$  are all Hermitian. Moreover, they are positive semidefinite.

*Proof.* Showing that  $T_{\xi\eta} = T_{\eta\xi}^*$  is straightforward from (2.1). We now check that  $T_{\xi\xi}$  is positive semidefinite. Let  $v \in \mathbb{C}^p$ . By (2.1),  $v^* T_{\xi\xi} v = \sum_{\zeta} \sum_{j,k} \overline{v^j} v^k \overline{B_{\zeta\xi}^j} B_{\zeta\xi}^k = \sum_{\zeta} |\sum_k v^k B_{\zeta\xi}^k|^2 \geq 0$ . It follows that  $N = \sum_{\xi} T_{\xi\xi}$  is also positive semidefinite.  $\square$

PROPOSITION 2.2.

$$Q^{jk} = \text{Tr}(B^{j*} S B^k) \text{ and } Q = \sum_{1 \leq \xi, \eta \leq n} T_{\xi\eta} T_{\xi\eta}^*.$$

*Proof.* By (2.1),  $Q^{jk} = \sum_l \langle T^{lj}, T^{lk} \rangle = \sum_l \text{Tr}(B^{j*} B^l B^{l*} B^k)$ . The summation and trace commute to give us the first identity. Similarly, the  $(j, k)$ th entry of  $\sum_{\xi\eta} T_{\xi\eta} T_{\xi\eta}^*$  is  $\sum_l \langle T^{kl}, T^{jl} \rangle = \sum_l \text{Tr}(B^{l*} B^k B^{j*} B^l)$ . Cycle the terms in the trace to obtain  $Q^{jk}$ .  $\square$

PROPOSITION 2.3. *Let  $u \in \mathbb{C}^p$  be a unit vector. Define  $U = \sum_{k=1}^p u^k B^k \in \mathbb{C}^{m \times n}$ . Then*

$$\|U\|_F^2 \leq \|N\|.$$

*Proof.*  $\|U\|_F^2 = \text{Tr}(U^* U) = \text{Tr}(\sum_{j,k} \overline{u^j} u^k B^{j*} B^k)$ . The sum and trace commute and due to (1.3),  $\|U\|_F^2 = \sum_{j,k} \overline{u^j} u^k N^{jk} \leq \|N\|$ .  $\square$

PROPOSITION 2.4.

$$\|Q\| \leq \|S\| \|N\|.$$

*Proof.*  $Q$  is Hermitian, so  $\|Q\| = \max_u u^* Q u$ , where  $u \in \mathbb{C}^p$  has unit norm. Now let  $u$  be an arbitrary unit vector and define  $U = \sum_{k=1}^p u^k B^k$ . By Proposition 2.2,  $u^* Q u = \sum_{j,k} \overline{u^j} u^k Q^{jk} = \text{Tr}(\sum_{j,k} \overline{u^j} u^k B^{j*} S B^k) = \text{Tr}(U^* S U)$ . Since  $S$  is positive definite, it follows from  $\|AB\|_F \leq \|A\| \|B\|_F$  that  $u^* Q u = \|S^{1/2} U\|_F^2 \leq \|S\| \|U\|_F^2$ . By Proposition 2.3,  $u^* Q u \leq \|S\| \|N\|$ .  $\square$

PROPOSITION 2.5. *For any  $1 \leq j \leq p$ ,*

$$\|B^j\| \leq \lambda n^{-1/2} \|N\|^{1/2}.$$

It follows that

$$\|Q\| = \left\| \sum_{\xi\eta} T_{\xi\eta} T_{\xi\eta}^* \right\| \leq p \lambda^2 \|N\|^2 / n.$$

*Proof.* We begin by noting that  $\|N\| \geq \max_j |N^{jj}| = \max_j \langle B^j, B^j \rangle = \max_j \|B^j\|_F^2$ . From Definition 1.2,  $\|B^j\| \leq \lambda n^{-1/2} \|B^j\|_F \leq \lambda n^{-1/2} \|N\|^{1/2}$  for any  $1 \leq j \leq p$ , which is our first inequality. It follows that  $\|S\| \leq \sum_{j=1}^p \|B^j\|^2 \leq p \lambda^2 \|N\| / n$ . Apply Propositions 2.4 and 2.2 to obtain the second inequality.  $\square$

PROPOSITION 2.6. *F, G are Hermitian, and*

$$\max(\|F\|, \|G\|) \leq \lambda^2 \|N\| (p/n).$$

*Proof.* That  $F, G$  are Hermitian follows from Proposition 2.1. Define  $F' = (T^{jk})$  another block matrix. Since reindexing the rows and columns of  $F$  does not change its norm,  $\|F\| = \|F'\|$ . By Proposition 2.5,  $\|F'\|^2 \leq \sum_{j,k=1}^p \|T^{jk}\|^2 \leq \sum_{j,k=1}^p \|B^j\|^2 \|B^k\|^2 \leq \lambda^4 \|N\|^2 (p/n)^2$ . The same argument works for  $G$ .  $\square$

We now combine the above linear algebraic results with a probabilistic result in Appendix B. Prepare to apply Proposition B.6 with  $A_{ij}$  replaced with  $T_{\xi\eta}$ . Note that  $R = \sum_{\xi\eta} T_{\xi\eta} T_{\xi\eta}^* = Q$  by Proposition 2.2. Bound  $\sigma$  using Propositions 2.5 and 2.6:

$$\begin{aligned} \sigma &= C_1 \max(\|Q\|^{1/2}, \|R\|^{1/2}, \|F\|, \|G\|) \\ &\leq C_1 \|N\| \max((p/n)^{1/2} \lambda, (p/n) \lambda^2) \\ &\leq C_1 \|N\| (p/n)^{1/2} \lambda. \end{aligned}$$

The last step goes through because our assumption on  $n$  guarantees that  $(p/n)^{1/2} \lambda \leq 1$ . Finally, apply Proposition B.6 with  $t\|N\|/\kappa = e\sigma u$ . The proof is complete.

**2.2. Sketch of the proof for Theorem 1.4.** Follow the proof of Proposition B.6. Letting  $s = \log n$ , we obtain

$$\begin{aligned} \mathbb{E} \|M - N\| &\leq (\mathbb{E} \|M - N\|^s)^{1/s} \\ &\leq C_1 (2C_2 np)^{1/s} s \max(\|Q\|^{1/2}, \|R\|^{1/2}, \|F\|, \|G\|) \\ &\leq C(\log n) \|N\| (p/n)^{1/2} \lambda. \end{aligned}$$

**2.3. Proof of Proposition 1.5.** Recall that  $A$  is approximately the linear combination  $\sum_{j=1}^p d^j B^j$ , while  $\sum_{j=1}^p c^j B^j$  is the recovered linear combination. We shall first show that the recovered coefficients  $c$  is close to  $d$ :

$$\begin{aligned} \|d - c\| &= \|L^+ Au - c\| \\ &= \|L^+(Lc + Eu) - c\| \\ &= \|L^+ Eu\| \\ &\leq \varepsilon \|u\| \left( \frac{\kappa}{(1-t)\|N\|} \right)^{1/2}. \end{aligned}$$

Let  $v$  be a unit  $n$ -vector. Let  $L'$  be a  $n \times p$  matrix such that its  $j$ th column is  $B^j v$ . Now,

$$Av - \sum_{j=1}^p c^j B^j v = (L'd + Ev) - L'c = Ev + L'(d - c).$$

Combining the two equations, we have

$$(2.2) \quad \left\| A - \sum_{j=1}^p c^j B^j \right\| \leq \varepsilon + \varepsilon \|L'\| \|u\| \left( \frac{\kappa}{(1-t)\|N\|} \right)^{1/2}.$$

With overwhelming probability,  $\|u\| = O(\sqrt{n})$ . The only term left that needs to be bounded is  $\|L'\|$ . This turns out to be easy because  $\|B^j\| \leq \lambda n^{-1/2} \|N\|^{1/2}$  by Proposition 2.5 and

$$\|L'\|^2 \leq \sum_{j=1}^p \|B^j v\|^2 \leq \lambda^2 \|N\| p/n.$$

Substitute this into (2.2) to finish the proof.

**2.4. Proof of Corollary 1.6.** Let  $u \sim N(0, 1)^n$  iid. Say  $A$  has a singular value decomposition  $E\Lambda F^*$ , where  $\Lambda$  is diagonal. Do a change of basis by letting  $u' = F^*u \sim N(0, 1)^n$  iid,  $B'_j = F^*B_jE$ , and  $\mathcal{B}'_\Lambda = \{B'_1\Lambda, \dots, B'_p\Lambda\}$ . Equation (1.1) is reduced to  $L'c = \Lambda u'$ , where the  $j$ th column of  $L'$  is  $B'_j\Lambda u'$ .

Since Frobenius inner products  $\|\cdot\|$  and  $\|\cdot\|_F$  are all preserved under unitary transformations, it is clear that  $\kappa(\mathcal{B}'_\Lambda) = \kappa(\mathcal{B}_A)$  and  $\lambda(\mathcal{B}'_\Lambda) = \lambda(\mathcal{B}_A)$ . Essentially, for our purpose here, we may pretend that  $A = \Lambda$ .

Let  $\tilde{n} = \text{rank}(A)$ . If  $A$  has a large nullspace, i.e.,  $\tilde{n} \ll \min(m, n)$ , then  $B'_j\Lambda$  has  $n - \tilde{n}$  columns of zeros and many components of  $u'$  are never transmitted to the  $B'_j$ 's anyway. We may therefore truncate the length of  $u'$  to  $\tilde{n}$ , let  $\tilde{B}_j \in \mathbb{C}^{n \times \tilde{n}}$  be  $B'_j\Lambda$  with its columns of zeros chopped away, and apply Theorem 1.3 with  $\mathcal{B}$  replaced with  $\tilde{\mathcal{B}} := \{\tilde{B}_1, \dots, \tilde{B}_p\}$ . Observe that  $\kappa(\tilde{\mathcal{B}}) = \kappa(\mathcal{B}'_\Lambda)$ , whereas  $\lambda(\tilde{\mathcal{B}}) = (\tilde{n}/n)^{1/2}\lambda(\mathcal{B}'_\Lambda)$  because  $\|\tilde{B}_j\|_F = \|B'_j\Lambda\|_F$  and  $\|\tilde{B}_j\| = \|B'_j\Lambda\|$  but  $\tilde{B}_j$  has  $\tilde{n}$  instead of  $n$  columns. The proof is complete.

**3. Probing operators with smooth symbols.**

**3.1. Basics and assumptions.** We begin by defining what a pseudodifferential symbol is.

DEFINITION 3.1. *Every linear operator  $A$  is associated with a pseudodifferential symbol  $a(x, \xi)$  such that for any  $u : \mathbb{R}^d \rightarrow \mathbb{R}$ ,*

$$(3.1) \quad Au(x) = \int_{\xi \in \mathbb{R}^d} e^{2\pi i \xi \cdot x} a(x, \xi) \hat{u}(\xi) d\xi,$$

where  $\hat{u}$  is the Fourier transform of  $u$ , that is,  $\hat{u}(\xi) = \int_{x \in \mathbb{R}^d} u(x) e^{-2\pi i \xi \cdot x} dx$ .

We refrain from calling  $A$  a “pseudodifferential operator” at this point because its symbol has to satisfy some additional constraints that will be covered in section 3.5. What is worth noting here is the Schwartz kernel theorem, which shows that every linear operator  $A : \mathcal{S}(\mathbb{R}^d) \rightarrow \mathcal{S}'(\mathbb{R}^d)$  has a symbol representation as in (3.1) and in that integral,  $a(x, \xi) \in \mathcal{S}'(\mathbb{R}^d \times \mathbb{R}^d)$  acts as a distribution. Recall that  $\mathcal{S}$  is the Schwartz space and  $\mathcal{S}'$  is its dual or the space of tempered distributions. The interested reader may refer to [12] or [25] for a deeper discourse.

The term “pseudodifferential” arises from the fact that differential operators have very simple symbols. For example, the Laplacian has the symbol  $a(x, \xi) = -4\pi^2 \|\xi\|^2$ . Another example is

$$Au(x) = u(x) - \nabla \cdot \alpha(x) \text{grad } u(x) \text{ for some } \alpha(x) \in C^1(\mathbb{R}^d).$$

Its symbol is

$$(3.2) \quad a(x, \xi) = 1 + \alpha(x)(4\pi^2 \|\xi\|^2) - \sum_{k=1}^d (2\pi i \xi_k) \partial_{x_k} \alpha(x).$$

Clearly, if the media  $\alpha(x)$  is smooth, so is the symbol  $a(x, \xi)$  smooth in both  $x$  and  $\xi$ , an important property which will be used in section 3.3.

For practical reasons, we make the following assumptions about  $u : \mathbb{R}^d \rightarrow \mathbb{R}$  on which symbols are applied:

1.  $u$  is periodic with period 1, so only  $\xi \in \mathbb{Z}^d$  will be considered in the Fourier domain.
2.  $u$  is bandlimited, say,  $\hat{u}$  is supported on  $\Xi := [-\xi_0, \xi_0]^d \subseteq \mathbb{Z}^d$ . Any summation over the Fourier domain is by default over  $\Xi$ .<sup>4</sup>

---

<sup>4</sup>To have an even number of points per dimension, one can use  $\Xi = [-\xi_0, \xi_0 - 1]^d$ , for example. We leave this generalization to the reader and continue to assume  $\xi \in [-\xi_0, \xi_0]^d$ .

3.  $a(x, \xi)$  and  $u(x)$  are only evaluated at  $x \in X \subset [0, 1]^d$  which are points uniformly spaced apart. Any summation over  $x$  is by default over  $X$ . Subsequently, (3.1) reduces to a discrete and finite form:

$$(3.3) \quad Au(x) = \sum_{\xi \in \Xi} e^{2\pi i \xi \cdot x} a(x, \xi) \hat{u}(\xi).$$

We like to call  $a(x, \xi)$  a “discrete symbol.” Some tools are already available for manipulating such symbols [10].

**3.2. User friendly representations of symbols.** Given a linear operator  $A$ , it is useful to relate its symbol  $a(x, \xi)$  to its matrix representation in the Fourier basis. This helps us understand the symbol as a matrix and also exposes easy ways of computing the symbols of  $A^{-1}$ ,  $A^*$ , and  $AB$  using standard linear algebra software.

By a matrix representation  $(A_{\eta\xi})$  in Fourier basis, we mean of course that  $\widehat{Au}(\eta) = \sum_{\xi} A_{\eta\xi} \hat{u}(\xi)$  for any  $\eta$ . We also introduce a more compact form of the symbol:  $\hat{a}(j, \xi) = \int_x a(x, \xi) e^{-2\pi i j \cdot x} dx$ . The next few results are pedagogical and listed for future reference.

**PROPOSITION 3.2.** *Let  $A$  be a linear operator with symbol  $a(x, \xi)$ . Let  $(A_{\eta\xi})$  and  $\hat{a}(j, \xi)$  be as defined above. Then*

$$A_{\eta\xi} = \int_x a(x, \xi) e^{-2\pi i(\eta-\xi)x} dx; \quad a(x, \xi) = e^{-2\pi i \xi x} \sum_{\eta} e^{2\pi i \eta x} A_{\eta\xi};$$

$$A_{\eta\xi} = \hat{a}(\eta - \xi, \xi); \quad \hat{a}(j, \xi) = A_{j+\xi, \xi}.$$

*Proof.* Let  $\eta = \xi + j$  and apply the definitions.  $\square$

**PROPOSITION 3.3** (trace). *Let  $A$  be a linear operator with symbol  $a(x, \xi)$ . Then*

$$\text{Tr}(A) = \sum_{\xi} \hat{a}(0, \xi) = \sum_{\xi} \int_x a(x, \xi) dx.$$

**PROPOSITION 3.4** (adjoint). *Let  $A$  and  $C = A^*$  be linear operators with symbols  $a(x, \xi)$ ,  $c(x, \xi)$ . Then*

$$\hat{c}(j, \xi) = \overline{\hat{a}(-j, j + \xi)}; \quad c(x, \xi) = \sum_{\eta} \int_y \overline{a(y, \eta)} e^{2\pi i(\eta-\xi)(x-y)} dy.$$

**PROPOSITION 3.5** (composition). *Let  $A, B$ , and  $C = AB$  be linear operators with symbols  $a(x, \xi)$ ,  $b(x, \xi)$ ,  $c(x, \xi)$ . Then*

$$\hat{c}(j, \xi) = \sum_{\zeta} \hat{a}(j + \xi - \zeta, \zeta) \hat{b}(\zeta - \xi, \xi);$$

$$c(x, \xi) = \sum_{\zeta} \int_y e^{2\pi i(\zeta-\xi)(x-y)} a(x, \zeta) b(y, \xi) dy.$$

We leave it to the reader to verify the above results.

**3.3. Symbol expansions.** The idea is that when a linear operator  $A$  has a smooth symbol  $a(x, \xi)$ , only a few basis functions are needed to approximate  $a$ , and

correspondingly only a small  $\mathcal{B}$  is needed to represent  $A$ . This is not new; see, for example, [10]. In this paper, we consider the separable expansion

$$a(x, \xi) = \sum_{jk} c_{jk} e_j(x) g_k(\xi).$$

This is the same as expanding  $A$  as  $\sum_{jk} c_{jk} B_{jk}$ , where the symbol for  $B_{jk}$  is  $e_j(x)g_k(\xi)$ . With an abuse of notation, let  $B_{jk}$  also denote its matrix representation in Fourier basis. Given our assumption that  $\xi \in [-\xi_0, \xi_0]^d$ , we have  $B_{jk} \in \mathbb{C}^{n \times n}$ , where  $n = (2\xi_0 + 1)^d$ . As its symbol is separable,  $B_{jk}$  can be factorized as

$$(3.4) \quad B_{jk} = \mathcal{F} \text{diag}(e_j(x)) \mathcal{F}^{-1} \text{diag}(g_k(\xi)),$$

where  $\mathcal{F}$  is the unitary Fourier matrix. An alternative way of viewing  $B_{jk}$  is that it takes its input  $\hat{u}(\xi)$ , multiplies it by  $g_k(\xi)$ , and convolves it with  $\hat{e}_j(\eta)$ , the Fourier transform of  $e_j(x)$ . There is also an obvious algorithm to apply  $B_{jk}$  to  $u(x)$  in  $\tilde{O}(n)$  time as outlined below. As mentioned in section 1.4, this speeds up the recovery of the coefficients  $c$  and makes matrix probing a cheap operation.

ALGORITHM 2. Given vector  $u(x)$ , apply the symbol  $e_j(x)g_k(\xi)$ .

1. Perform FFT on  $u$  to obtain  $\hat{u}(\xi)$ .
2. Multiply  $\hat{u}(\xi)$  by  $g_k(\xi)$  elementwise.
3. Perform inverse FFT on the previous result, obtaining  $\sum_{\xi} e^{2\pi i \xi \cdot x} g_k(\xi) \hat{u}(\xi)$ .
4. Multiply the previous result by  $e_j(x)$  elementwise.

Recall that for  $L$  to be well-conditioned with high probability, we need to check whether  $N$ , as defined in (1.3), is well-conditioned or in a rough sense whether  $\langle B_j, B_k \rangle \simeq \delta_{jk}$ . For separable symbols, this inner product is easy to compute.

PROPOSITION 3.6. Let  $B_{jk}, B_{j'k'} \in \mathbb{C}^{n \times n}$  be matrix representations (in Fourier basis) of linear operators with symbols  $e_j(x)g_k(\xi)$  and  $e_{j'}(x)g_{k'}(\xi)$ . Then

$$\langle B_{jk}, B_{j'k'} \rangle = \langle e_j, e_{j'} \rangle \langle g_k, g_{k'} \rangle,$$

where  $\langle e_j, e_{j'} \rangle = \frac{1}{n} \sum_{i=1}^n \overline{e_j(x_i)} e_{j'}(x_i)$  and  $x_1, \dots, x_n$  are points in  $[0, 1]^d$  uniformly spaced, and  $\langle g_k, g_{k'} \rangle = \sum_{\xi} \overline{g_k(\xi)} g_{k'}(\xi)$ .

*Proof.* Apply Propositions 3.3, 3.4, and 3.5 with the symbols in the  $\hat{a}(\eta, \xi)$  form.  $\square$

To compute  $\lambda(\mathcal{B})$  as in Definition 1.2, we examine the spectrum of  $B_{jk}$  for every  $j, k$ . A simple and relevant result is as follows.

PROPOSITION 3.7. Assume the same setup as in Proposition 3.6. Then

$$\sigma_{\min}(B_{jk}) \geq \min_x |e_j(x)| \min_{\xi} |g_k(\xi)|; \quad \sigma_{\max}(B_{jk}) \leq \max_x |e_j(x)| \max_{\xi} |g_k(\xi)|.$$

*Proof.* In (3.4),  $\mathcal{F} \text{diag}(e^j(x)) \mathcal{F}^{-1}$  has singular values  $|e_j(x)|$  as  $x$  varies over  $X$ , defined at the end of section 3.1. The result follows from the min-max theorem.  $\square$

As an example, suppose  $a(x, \xi)$  is smooth and periodic in both  $x$  and  $\xi$ . It is well-known that a Fourier series is a good expansion scheme because the smoother  $a(x, \xi)$  is as a periodic function in  $x$ , the faster its Fourier coefficients decay, and less is lost when we truncate the Fourier series. Hence, we pick<sup>5</sup>

$$(3.5) \quad e_j(x) = e^{2\pi i j \cdot x}; \quad g_k(\xi) = e^{2\pi i k \cdot \varphi(\xi)},$$

where  $\varphi(\xi) = (\xi + \xi_0)/(2\xi_0 + 1)$  maps  $\xi$  into  $[0, 1]^d$ .

<sup>5</sup>Actually,  $\exp(2\pi i k \xi_0 / (2\xi_0 + 1))$  does not vary with  $\xi$ , and we can use  $\varphi(\xi) = \xi / (2\xi_0 + 1)$ .

Due to Proposition 3.6,  $N = \mathbb{E}M$  is a multiple of the identity matrix and  $\kappa(\mathcal{B}) = 1$ , where  $\mathcal{B} = \{B_{jk}\}$ . It is also immediate from Proposition 3.7 that  $\lambda(B_{jk}) = 1$  for every  $j, k$ , and  $\lambda(\mathcal{B}) = 1$ . The optimal condition numbers of this  $\mathcal{B}$  make it suitable for matrix probing.

**3.4. Chebyshev expansion of symbols.** The symbols of differential operators are polynomials in  $\xi$  and nonperiodic. When probing these operators, a Chebyshev expansion in  $\xi$  is in principle favored over a Fourier expansion, which may suffer from the Gibbs phenomenon. However, as we shall see,  $\kappa(\mathcal{B})$  grows with  $p$  and can lead to ill-conditioning.

For simplicity, assume that the symbol is periodic in  $x$  and that  $e_j(x) = e^{2\pi i j \cdot x}$ . Applying Proposition 3.2, we see that  $B_{jk}$  is a matrix with a displaced diagonal and its singular values are  $(g_k(\xi))_{\xi \in \Xi}$ . (Recall that we denote the matrix representation (in Fourier basis) of  $B_{jk}$  as  $B_{jk}$  as well.)

Let  $T_k$  be the  $k$ th Chebyshev polynomial. In one dimension, we can pick

$$(3.6) \quad g_k(\xi) = T_k(\xi/\xi_0) \text{ for } k = 1, \dots, K.$$

Define  $\|T_k\|_2 = (\int_{z=-1}^1 T_k(z)^2 dz)^{1/2}$ . By approximating sums with integrals,  $\lambda(B_{jk}) \simeq \sqrt{2} \|T_k\|_2^{-1} = (\frac{4k^2-1}{2k^2-1})^{1/2}$ . Notice that there is no  $(1-z^2)^{-1/2}$  weight factor in the definition of  $\|T_k\|_2$  because  $e_j(x)T_k(\xi)$  is treated as a pseudodifferential symbol and has to be evaluated on the *uniform* grid. In practice, this approximation becomes very accurate with larger  $n$  and we see no need to be rigorous here. As  $k$  increases,  $\lambda(B_{jk})$  approaches  $\sqrt{2}$ . More importantly,  $\lambda(B_{jk}) \leq \lambda(B_{j1})$  for any  $j, k$ , so

$$\lambda(\mathcal{B}) = \sqrt{3}.$$

Applying the same technique to approximate the sum  $\langle g_k, g_{k'} \rangle$ , we find that  $\langle g_k, g_{k'} \rangle \propto (1 - (k+k')^2)^{-1} + (1 - (k-k')^2)^{-1}$  when  $k+k'$  is even and zero otherwise. We then compute  $N = \mathbb{E}M$  for various  $K$  and plot  $\kappa(\mathcal{B})$  versus  $K$ , the number of Chebyshev polynomials. As shown in Figure 3.1(a),

$$\kappa(\mathcal{B}) \simeq 1.3K.$$

This means that if we expect to recover  $p = \tilde{O}(n)$  coefficients, we must keep  $K$  fixed. Otherwise, if  $p = K^2$ , only  $p = \tilde{O}(n^{1/2})$  are guaranteed to be recovered by Theorem 1.3.

In two dimensions, a plausible expansion is

$$(3.7) \quad g_k(\xi) = e^{ik_1 \arg \xi} T_{k_2}(\varphi(\|\xi\|)) \text{ for } 1 \leq k_2 \leq K,$$

where  $k = (k_1, k_2)$  and  $\varphi(r) = (\sqrt{2}r/\xi_0) - 1$  maps  $\|\xi\|$  into  $[-1, 1]$ . We call this the ‘‘Chebyshev on a disk’’ expansion.

The quantity  $\lambda(B_{jk})$  is approximately  $2(\int_{x=-1}^1 \int_{y=-1}^1 T_k(\psi(x, y))^2 dx dy)^{-1/2}$ , where  $\psi(x, y) = (2x^2 + 2y^2)^{1/2} - 1$ . The integral is evaluated numerically and appears to converge<sup>6</sup> to  $\sqrt{2}$  for large  $k_2$ . Also,  $k_2 = 1$  again produces the worst  $\lambda(B_{jk})$  and

$$\lambda(\mathcal{B}) \leq 2.43.^7$$

<sup>6</sup>This is because when we truncate the disk of radius  $\xi_0\sqrt{2}$  to a square of length  $2\xi_0$ , most is lost along the vertical axis and away from the diagonals. However, for large  $k$ ,  $T_k$  oscillates very much and the truncation does not matter. If we pretend that the square is a disk, then we are back in the one-dimensional (1D) case where the answer approaches  $\sqrt{2}$  for large  $k$ .

<sup>7</sup>The exact value is  $2(4 - \frac{8}{3}\sqrt{2} \sinh^{-1}(1))^{-1/2}$ .

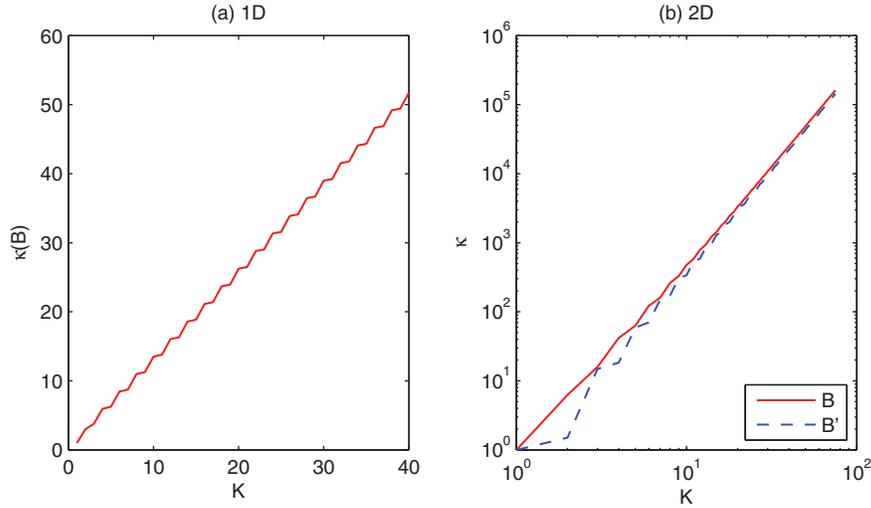


FIG. 3.1. Let  $K$  be the number of Chebyshev polynomials used in the expansion of the symbol; see (3.6) and (3.7). Observe that in one dimension,  $\kappa(\mathcal{B}) = O(K)$ , while in two dimensions,  $\kappa(\mathcal{B}) = O(K^3)$ . These condition numbers mean that we cannot expect to retrieve  $p = \tilde{O}(n)$  parameters unless  $K$  is fixed and independent of  $p, n$ .

As for  $\kappa(\mathcal{B})$ , observe that when  $k_1 \neq k'_1$ ,  $\langle g_{k_1 k_2}, g_{k'_1 k'_2} \rangle = \pm 1$  due to symmetry,<sup>8</sup> whereas when  $k_1 = k'_1$ , the inner product is proportional to  $n$  and is much larger. As a result, the  $g_k$ 's with different  $k_1$ 's hardly interact and in studying  $\kappa(\mathcal{B})$ , one may assume that  $k_1 = k'_1 = 0$ . To improve  $\kappa(\mathcal{B})$ , we can normalize  $g_k$  such that the diagonal entries of  $N$  are all ones, that is  $g'_k(\xi) = g_k(\xi) / \|g_k(\xi)\|$ .

This yields another set of basis matrices  $\mathcal{B}'$ . Figure 3.1(b) reveals that

$$\kappa(\mathcal{B}) = O(K^3) \text{ and } \kappa(\mathcal{B}') \simeq \kappa(\mathcal{B}).$$

The latter can be explained as follows: we saw earlier that  $\langle B_{jk}, B_{jk} \rangle$  converges as  $k_2$  increases, so the diagonal entries of  $N$  are about the same and the normalization is only a minor correction.

If  $a(x, \xi)$  is expanded using the same number of basis functions in each direction of  $x$  and  $\xi$ , i.e.,  $K = p^{1/4}$ , then Theorem 1.3 suggests that only  $p = \tilde{O}(n^{2/5})$  coefficients can be recovered.

To recap, for both one and two dimensions,  $\lambda(\mathcal{B})$  is a small number but  $\kappa(\mathcal{B})$  increases with  $K$ . Fortunately, if we know that the operator being probed is a second order derivative, for example, we can fix  $K = 2$ .

Numerically, we have observed that the Chebyshev expansion can produce dramatically better results than the Fourier expansion of the symbol. More details can be found in section 4.3.

**3.5. Order of an operator.** In standard texts,  $A$  is said to be a pseudodifferential operator of order  $w$  if its symbol  $a(x, \xi)$  is in  $C^\infty(\mathbb{R}^d \times \mathbb{R}^d)$  and for any multi-indices  $\alpha, \beta$  there exists a constant  $C_{\alpha\beta}$  such that

$$|\partial_\xi^\alpha \partial_x^\beta a(x, \xi)| \leq C_{\alpha\beta} \langle \xi \rangle^{w-|\alpha|} \text{ for all } \xi, \text{ where } \langle \xi \rangle = 1 + \|\xi\|.$$

<sup>8</sup>The  $\xi$  and  $-\xi$  terms cancel each other. Only  $\xi = 0$  contributes to the sum.

Letting  $\alpha = \beta = 0$ , we see that such operators have symbols that grow or decay as  $(1 + \|\xi\|)^w$ . As an example, the Laplacian is of order 2. The factor 1 prevents  $\langle \xi \rangle$  from blowing up when  $\xi = 0$ . There is nothing special about it and if we take extra care when evaluating the symbol at  $\xi = 0$ , we can use

$$\langle \xi \rangle = \|\xi\|.$$

For forward matrix probing, if it is known a priori that  $a(x, \xi)$  behaves like  $\langle \xi \rangle^w$ , it makes sense to expand  $a(x, \xi) \langle \xi \rangle^{-w}$  instead. Another way of viewing this is that the symbol of the operator  $B_{jk}$  is modified from  $e_j(x)g_k(\xi)$  to  $e_j(x)g_k(\xi) \langle \xi \rangle^w$  to suit  $A$  better.

For backward matrix probing, if  $A$  is of order  $z$ , then  $A^{-1}$  is of order  $-z$  and we should replace the symbol of  $B_{jk}$  with  $e_j(x)g_k(\xi) \langle \xi \rangle^{-w}$ . We believe that this small correction has an impact on the accuracy of matrix probing, as well as the condition numbers  $\kappa(\mathcal{B}_A)$  and  $\lambda(\mathcal{B}_A)$ .

Recall that an element of  $\mathcal{B}_A$  is  $B_{jk}A$ . If  $A$  is of order  $w$  and  $B_{jk}$  is of order 0, then  $B_{jk}A$  is of order  $w$  and  $\lambda(B_{jk}A)$  will grow with  $n^w$ , which will adversely affect the conditioning of matrix probing. However, by multiplying the symbol of  $B_{jk}$  by  $\langle \xi \rangle^{-w}$ , we can expect  $B_{jk}A$  to be order 0 and that  $\lambda(B_{jk}A)$  is independent of the size of the problem  $n$ . The argument is heuristical but we will support it with some numerical evidence in section 4.3.

**4. Numerical examples.** We carry out four different experiments. The first experiment suggests that Theorem 1.4 is not tight. The second experiment presents the output of backward probing in a visual way. In the third experiment, we explore the limitations of backward probing and also tests the Chebyshev expansion of symbols. The last experiment involves the forward probing of the foveation operator, which is related to human vision.

**4.1. 1D statistical study.** We are interested in whether the probability bound in Theorem 1.3 is tight with respect to  $p$  and  $n$ , but as the tail probabilities are small and hard to estimate, we opt to study the first moment instead. In particular, if Theorem 1.4 captures exactly the dependence of  $\mathbb{E}\|M - N\|/\|N\|$  on  $p$  and  $n$ , then we would need  $n$  to grow faster than  $p \log^2 n$  for  $\mathbb{E}\|M - N\|/\|N\|$  to vanish, assuming  $\lambda(\mathcal{B})$  is fixed.

For simplicity, we use the Fourier expansion of the symbol in one dimension so that  $\lambda(\mathcal{B}) = \kappa(\mathcal{B}) = 1$ . Let  $J$  be the number of basis functions in both  $x$  and  $\xi$  and  $p = J^2$ . Figure 4.1(a) suggests that  $\mathbb{E}\|M - N\|/\|N\|$  decays to zero when  $n = p \log^c p$  and  $c > 1$ . It follows from the previous paragraph that Theorem 1.4 cannot be tight.

Nevertheless, Theorem 1.4 is optimal in the following sense. Imagine a more general bound

$$(4.1) \quad \mathbb{E} \frac{\|M - N\|}{\|N\|} \leq (\log^\alpha n) \left(\frac{p}{n}\right)^\beta \text{ for some } \alpha, \beta > 0.$$

In Figure 4.2(a), we see that for various values of  $p/n$ ,  $\alpha = 1$  since the graphs are linear. On the other hand, if we fix  $p$  and vary  $n$ , the log-log graph of Figure 4.2(b) shows that  $\beta = 1/2$ . Therefore, any bound in the form of (4.1) is no better than Theorem 1.4.

Next, we fix  $p = 25, n = 51$  and sample  $\|M - N\|/\|N\|$  many times to estimate the tail probabilities. In Figure 4.1(b), we see that the tail probability of  $\mathbb{P}(\|M - N\|/\|N\| > t)$  decays as  $\exp(-c_1 t)$  when  $t$  is big and as  $\exp(-c_2 t^2)$  when  $t$

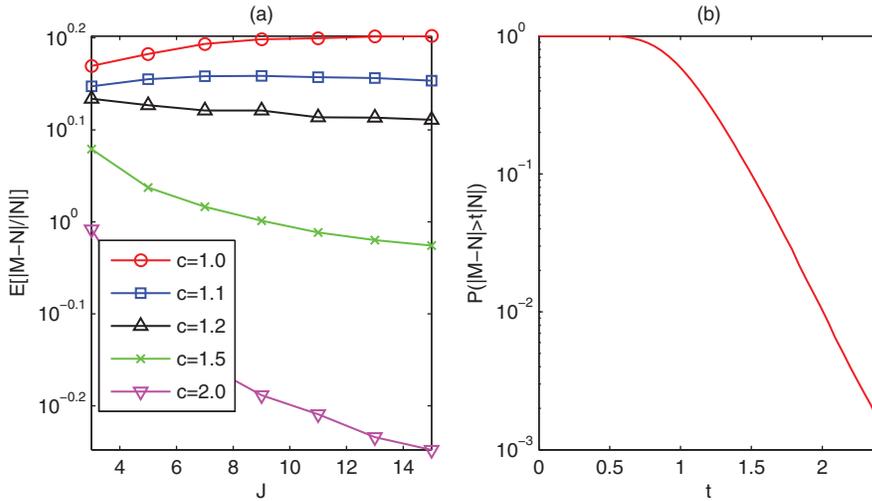


FIG. 4.1. Consider the Fourier expansion of the symbol.  $J$  is the number of basis functions in  $x$  and  $\xi$ , so  $p = J^2$ . Let  $n = p \log^c p$ . Figure (a) shows that the estimated  $\mathbb{E}\|M - N\|/\|N\|$  decays for  $c \geq 1.1$ , which suggests that Theorem 1.4 is not tight. In figure (b), we estimate  $\mathbb{P}(\|M - N\|/\|N\| > t)$  by sampling  $\|M - N\|/\|N\|$   $10^5$  times. The tail probability appears to be subgaussian for small  $t$  and subexponential for larger  $t$ .

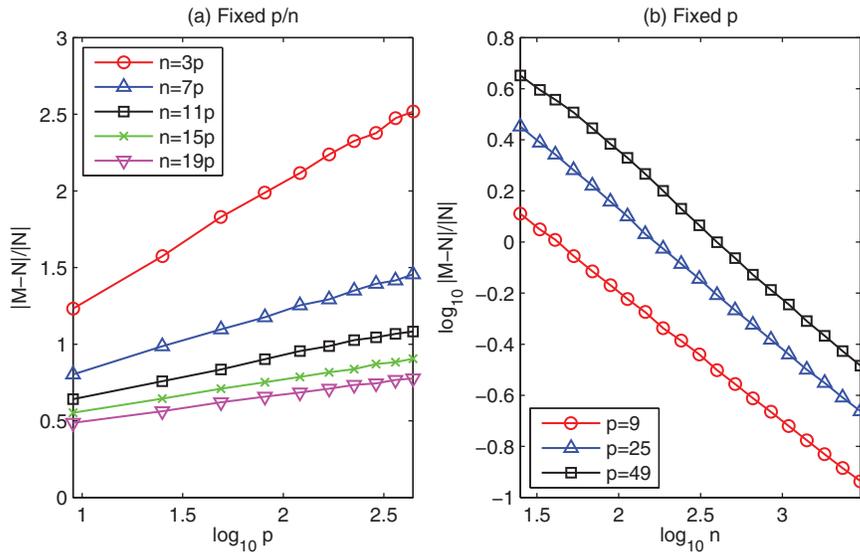


FIG. 4.2. Consider bounding  $\mathbb{E}\|M - N\|/\|N\|$  by  $(\log^\alpha n)(p/n)^\beta$ . There is little loss in replacing  $\log n$  with  $\log p$  in the simulation. In figure (a), the estimated  $\mathbb{E}\|M - N\|/\|N\|$  depends linearly on  $\log p$ , so  $\alpha \geq 1$ . In figure (b), we fix  $p$  and find that for large  $n$ ,  $\beta = 1/2$ . The conclusion is that the bound in Theorem 1.4 has the best  $\alpha, \beta$ .

is small for some positive numbers  $c_1, c_2$ . This behavior may be explained by Rauhut and Tropp's not yet published result.

**4.2. Elliptic equation in one dimension.** We find it instructive to consider a 1D example of matrix probing because it is easy to visualize the symbol  $a(x, \xi)$ .

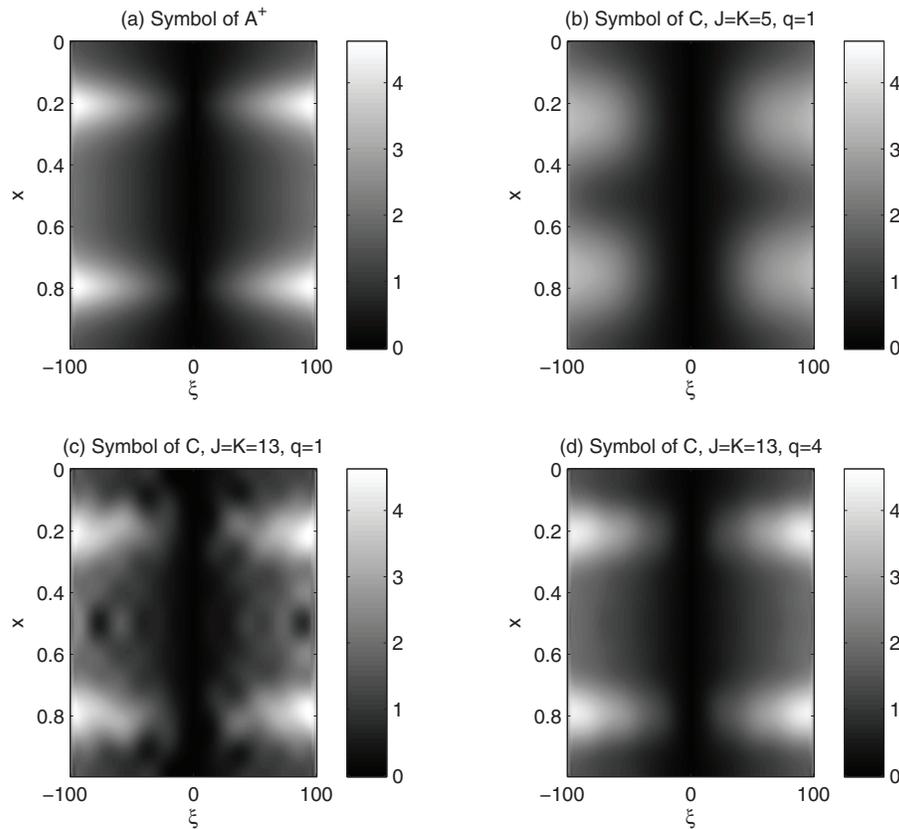


FIG. 4.3. Let  $A$  be the 1D elliptic operator in (4.2) and  $A^+$  be its pseudoinverse. Let  $C$  be the output of backward matrix probing with the following parameters:  $q$  is the number of random vectors applied to  $A^+$ ;  $J, K$  are the number of  $e_j$ 's and  $g_k$ 's used to expand the symbol of  $A^+$  in equation (3.5). Figure (a) is the symbol of  $A^+$ . Figure (b) is the symbol of  $C$  with  $J = K = 5$ . It lacks the sharp features of figure (a) because  $\mathcal{B}$  is too small to represent  $A^+$  well. With  $J = K = 13$ , probing with only one random vector leads to ill-conditioning and an inaccurate result in figure (b). In figure (c), four random vectors are used and a much better result is obtained. Note that the symbols are multiplied by  $\langle \xi \rangle^3$  for better visual contrast.

Consider the operator

$$(4.2) \quad Au(x) = -\frac{d}{dx}\alpha(x)\frac{du(x)}{dx}, \text{ where } \alpha(x) = 1 + 0.4\cos(4\pi x) + 0.2\cos(6\pi x).$$

Note that we use periodic boundaries and  $A$  is positive semidefinite with a 1D nullspace consisting of constant functions.

We probe for  $A^+$  using Algorithm 1 and the Fourier expansion of its symbol or (3.5). Since  $A$  is of order 2, we premultiply  $g_k(\xi)$  by  $\langle \xi \rangle^{-2}$  as explained in section 3.5.

In the experiment,  $n = 201$  and there are two other parameters  $J, K$  which are the number of  $e_j$ 's and  $g_k$ 's used in (3.5). To be clear,  $-\frac{J-1}{2} \leq j \leq \frac{J-1}{2}$  and  $-\frac{K-1}{2} \leq k \leq \frac{K-1}{2}$ .

Let  $C$  be the output of matrix probing. In Figure 4.3(b), we see that  $J = K = 5$  is not enough to represent  $A^+$  properly. This is expected because our media  $\alpha(x)$  has a bandwidth of 7. We expect  $J = K = 13$  to do better, but the much larger  $p$  leads to overfitting and a poor result, as is evident from the wobbles in the symbol of  $C$  in

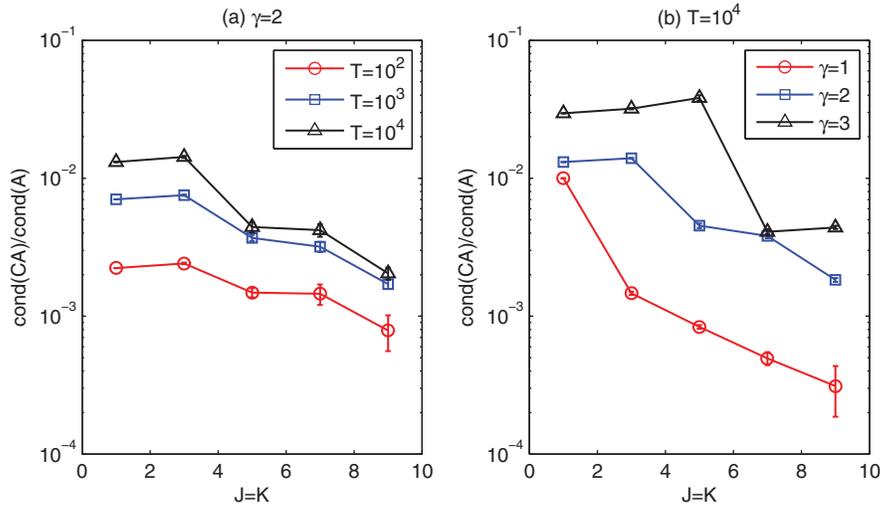


FIG. 4.4. Let  $A$  be the operator defined in (4.3) and  $C$  be the output of backward probing. In figure (b), we fix  $T = 10^4$  and find that as  $J$  goes from  $2\gamma - 1$  to  $2\gamma + 1$ , the bandwidth of the media, the quality of the preconditioner  $C$  improves by a factor between  $10^{0.5}$  and 10. In figure (a), we fix  $\gamma = 2$  and find that increasing the contrast worsens  $\text{cond}(CA)/\text{cond}(A)$ . Nevertheless, the improvement between  $J = 3$  and  $J = 5$  becomes more distinct. The error bars correspond to  $\hat{\sigma}$ , where  $\hat{\sigma}^2$  is the estimated variance. They indicate that  $C$  is not just good on average but good with high probability.

Figure 4.3(c). Probing with four random vectors, we obtain a much better result, as shown in Figure 4.3(d).

**4.3. Elliptic equation in two dimensions.** In this section, we extend the previous setup to two dimensions and address a different set of questions. Consider the operator  $A$  defined as

$$(4.3) \quad Au(x) = -\nabla \cdot \alpha(x)\nabla u(x) \text{ where } \alpha(x) = \frac{1}{T} + \cos^2(\pi\gamma x_1) \sin^2(\pi\gamma x_2).$$

The positive value  $T$  is called the contrast while the positive integer  $\gamma$  is the roughness of the media, since the bandwidth of  $\alpha(x)$  is  $2\gamma + 1$ . Again, we assume periodic boundary conditions such that  $A$ 's nullspace is exactly the set of constant functions.

Let  $C$  be the output of the backward probing of  $A$ . As we shall see, the quality of  $C$  drops as we increase the contrast  $T$  or the roughness  $\gamma$ .

Fix  $n = 101^2$  and expand the symbol using (3.5). Let  $J = K$  be the number of basis functions used to expand the symbol in each of its four dimensions, that is,  $p = J^4$ .

In Figure 4.4(b), we see that between  $J = 2\gamma - 1$  and  $J = 2\gamma + 1$ , the bandwidth of the media, there is a marked improvement in the preconditioner, as measured by the ratio  $\text{cond}(CA)/\text{cond}(A)$ .<sup>9</sup>

On the other hand, Figure 4.4(a) shows that as the contrast increases, the preconditioner  $C$  degrades in performance, but the improvement between  $J = 2\gamma - 1$  and  $2\gamma + 1$  becomes more pronounced.

<sup>9</sup>Since  $A$  has one zero singular value,  $\text{cond}(A)$  actually refers to the ratio between its largest singular value and its second smallest singular value. The same applies to  $CA$ .

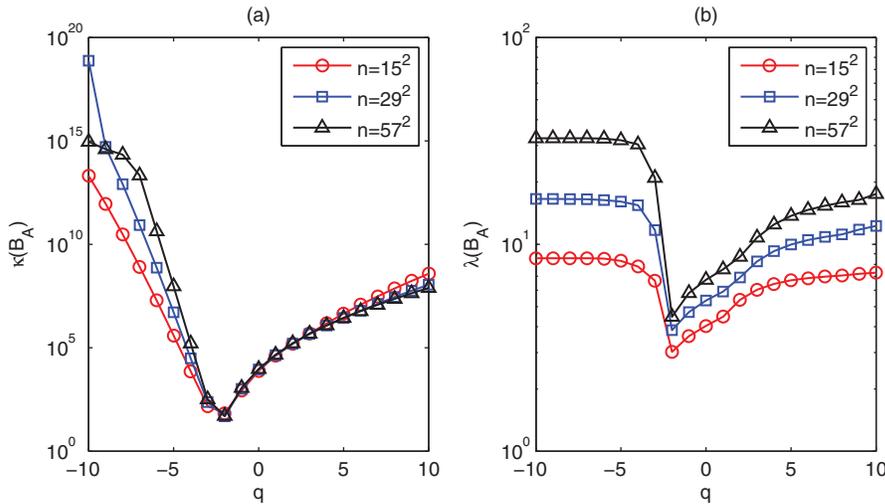


FIG. 4.5. Consider the backward probing of  $A$  in (4.3), a pseudodifferential operator of order 2. Perform order correction by multiplying  $g_k(\xi)$  by  $\langle \xi \rangle^q$  in the expansion of the symbol. See section 3.5. Observe that at  $q = -2$ , the condition numbers  $\lambda(\mathcal{B}_A)$  and  $\kappa(\mathcal{B}_A)$  are minimized and hardly grow with  $n$ .

The error bars in Figure 4.4 are not error margins but  $\hat{\sigma}$ , where  $\hat{\sigma}^2$  is the unbiased estimator of the variance. They indicate that  $\text{cond}(CA)/\text{cond}(A)$  is tightly concentrated around its mean, provided  $J$  is not too much larger than is necessary. For instance, for  $\gamma = 1$ ,  $J = 3$  already works well, but pushing to  $J = 9$  leads to greater uncertainty.

Next, we consider *forward probing* of  $A$  using the ‘‘Chebyshev on a disk’’ expansion or (3.7). Let  $m$  be the order correction, that is, we multiply  $g_k(\xi)$  by  $\langle \xi \rangle^m = \|\xi\|^m$ . Let  $C$  be the output of the probing and  $K$  be the number of Chebyshev polynomials used.

Fix  $n = 55^2$ ,  $T = 10$ ,  $\gamma = 2$ , and  $J = 5$ . For  $m = 0$  and  $K = 3$ , i.e., no order correction and using up to quadratic polynomials in  $\xi$ , we obtain a relative error  $\|C - A\|/\|A\|$  that is less than  $10^{-14}$ . On the other hand, using Fourier expansion, with  $K = 5$  in the sense that  $-\frac{K-1}{2} \leq k_1, k_2 \leq \frac{K-1}{2}$ , the relative error is on the order of  $10^{-1}$ . The point is that in this case,  $A$  has an exact ‘‘Chebyshev on a disk’’ representation and probing using the correct  $\mathcal{B}$  enables us to retrieve the coefficients with negligible errors.

Finally, we consider backward probing with the Chebyshev expansion. We use  $J = 5$ ,  $\gamma = 2$ , and  $T = 10$ . Figure 4.5 shows that when  $m = -2$ , the condition numbers  $\lambda(\mathcal{B}_A)$  and  $\kappa(\mathcal{B}_A)$  are minimized and hardly increase with  $n$ . This emphasizes the importance of knowing the order of the operator being probed.

**4.4. Foveation.** In this section, we forward probe for the foveation operator, a space-variant imaging operator [7], which is particularly interesting as a model for human vision. Formally, we may treat the foveation operator  $A$  as a Gaussian blur with a width or standard deviation that varies over space, that is,

$$(4.4) \quad Au(x) = \int_{\mathbb{R}^2} K(x, y)u(y)dy \text{ where } K(x, y) = \frac{1}{w(x)\sqrt{2\pi}} \exp\left(\frac{-\|x - y\|^2}{2w^2(x)}\right),$$

where  $w(x)$  is the width function which returns only positive real numbers.

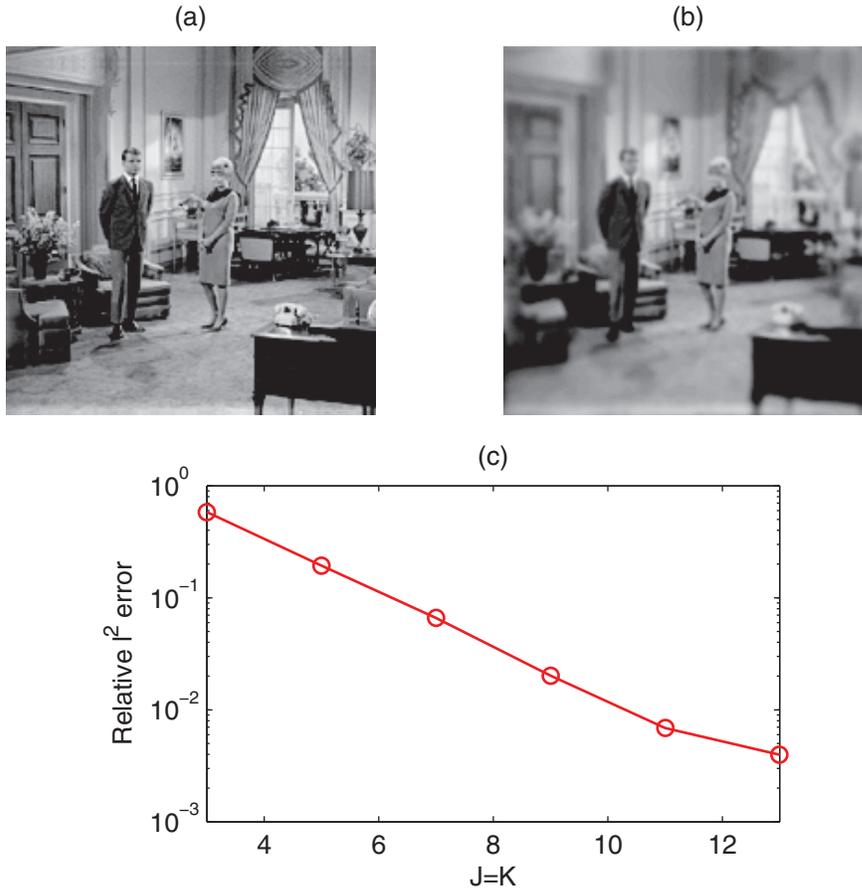


FIG. 4.6. Let  $A$  be the foveation operator in (4.4) and  $C$  be the output of the forward probing of  $A$ . Figure (a) is the test image  $z$ . Figure (b) is  $Cz$  and shows that  $C$  behaves like the foveation operator as expected. Figure (c) shows that the relative  $\ell^2$  error (see text) decreases exponentially with the number of parameters  $p = J^4$ .

The resolution of the output image is highest at the point where  $w(x)$  is minimal. Call this point  $x_0$ . It is the point of fixation, corresponding to the center of the fovea. For our experiment, the width function takes the form of  $w(x) = (\alpha\|x - x_0\|^2 + \beta)^{1/2}$ . Our images are  $201 \times 201$  and treated as functions on the unit square. We choose  $x_0 = (0.5, 0.5)$  and  $\alpha, \beta > 0$  such that  $w(x_0) = 0.003$  and  $w(1, 1) = 0.012$ .

The symbol of  $A$  is  $a(x, \xi) = \exp(-2\pi^2 w(x)^2 \|\xi\|^2)$ , and we choose to use a Fourier series or (3.5) for expanding it. Let  $C$  be the output of matrix probing and  $z$  be a standard test image. Figure 4.6(c) shows that the relative  $\ell^2$  error  $\|Cz - Az\|_{\ell^2} / \|Az\|_{\ell^2}$  decreases exponentially as  $p$  increases. In general, forward probing yields great results like this because we know its symbol well and can choose an appropriate  $\mathcal{B}$ .

**4.5. Inverting the wave equation Hessian.** In seismology, it is common to recover the model parameters  $m$ , which describe the subsurface, by minimizing the least squares misfit between the observed data and  $F(m)$ , where  $F$ , the forward model, predicts data from  $m$ .

Methods to solve this problem can be broadly categorized into two classes: steepest descent or Newton’s method. The former takes more iterations to converge but each iteration is computationally cheaper. The latter requires the inversion of the Hessian of the objective function but achieves quadratic convergence near the optimal point.

In another paper, we use matrix probing to precondition the inversion of the Hessian. Removing the nullspace component from the noise vector is more tricky (see Algorithm 1) and involves checking whether “a curvelet is visible to any receiver” via raytracing. For details on this more elaborate application, see [9].

**5. Conclusion and future work.** When a matrix  $A$  with  $n$  columns belongs to a specified  $p$ -dimensional subspace, say,  $A = \sum_{i=1}^p c_i B_i$ , we can probe it with a few random vectors to recover the coefficient vector  $c$ .

Let  $q$  be the number of random vectors used,  $\kappa$  be the condition number of the Gram matrix of  $B_1, \dots, B_p$ , and  $\lambda$  be the “weak condition number” of each  $B_i$  (cf. Definition 1.2) which is related to the numerical rank. From Theorem 1.3 and section 1.3, we learn that when  $nq \propto p(\kappa\lambda \log n)^2$ , the linear system that has to be solved to recover  $c$  (cf. (1.1)) will be well-conditioned with high probability. Consequently, the reconstruction error is small by Proposition 1.5.

The same technique can be used to compute an approximate  $A^{-1}$  or a preconditioner for inverting  $A$ . In [9], we used it to invert the wave equation Hessian—here we demonstrate that it can also be used to invert elliptic operators in smooth media (cf. sections 4.2 and 4.3).

Some possible future work includes the following:

1. Extend the work of Pfander, Rauhut, and others [21, 20, 22]. These papers are concerned with sparse signal recovery. They consider the special case where  $\mathcal{B}$  contains  $n^2$  matrices each representing a time-frequency shift, but  $A$  is an unknown linear combination of only  $p$  of them. The task is to identify these  $p$  matrices and the associated coefficients by applying  $A$  to noise vectors. Our proofs may be used to establish similar recovery results for a more general  $\mathcal{B}$ . However, note that in [20], Pfander and Rauhut show that  $n \propto p \log n$  suffices, whereas our main result requires an additional log factor.
2. Build a framework for probing  $f(A)$  interpreted as a Cauchy integral

$$f(A) = \frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI - A)^{-1} dz,$$

where  $\Gamma$  is a closed curve enclosing the eigenvalues of  $A$ . For more on approximating matrix functions, see [13, 15].

3. Consider expansion schemes for symbols that highly oscillate or have singularities that are well-understood.

**Appendix A. Linear algebra.** Recall the definitions of  $\kappa(\mathcal{B})$  and  $\lambda(\mathcal{B})$  at the beginning of the paper. The following concerns probing with multiple vectors (cf. section 1.3).

**PROPOSITION A.1.** *Let  $I_q \in \mathbb{C}^{q \times q}$  be the identity. Let  $\mathcal{B} = \{B_1, \dots, B_p\}$ . Let  $B'_j = I_q \otimes B_j$  and  $\mathcal{B}' = \{B'_1, \dots, B'_p\}$ . Then  $\kappa(\mathcal{B}) = \kappa(\mathcal{B}')$  and  $\lambda(\mathcal{B}) = \lambda(\mathcal{B}')$ .*

*Proof.* Define  $N \in \mathbb{C}^{p \times p}$  such that  $N_{jk} = \langle B_j, B_k \rangle$ . Define  $N' \in \mathbb{C}^{p \times p}$  such that  $N'_{jk} = \langle B'_j, B'_k \rangle$ . Clearly,  $N' = qN$ , so their condition numbers are the same and  $\kappa(\mathcal{B}) = \kappa(\mathcal{B}')$ .

For any  $A = B_j \in \mathbb{C}^{m \times n}$  and  $A' = B'_j$ , we have  $\frac{\|A'\|(nq)^{1/2}}{\|A'\|_F} = \frac{\|A\|(nq)^{1/2}}{\|A\|_F q^{1/2}} = \frac{\|A\|n^{1/2}}{\|A\|_F}$ . Hence,  $\lambda(\mathcal{B}) = \lambda(\mathcal{B}')$ .  $\square$

**Appendix B. Probabilistic tools.** In this section, we present some probabilistic results used in our proofs. The first theorem is used to decouple homogeneous Rademacher chaos of order 2 and can be found in [8, 23], for example.

**THEOREM B.1.** *Let  $(u_i)$  and  $(\tilde{u}_i)$  be two iid sequences of real-valued random variables and  $A_{ij}$  be in a Banach space where  $1 \leq i, j \leq n$ . There exists universal constants  $C_1, C_2 > 0$  such that for any  $s \geq 1$ ,*

$$(B.1) \quad \left( \mathbb{E} \left\| \sum_{1 \leq i \neq j \leq n} u_i u_j A_{ij} \right\|^s \right)^{1/s} \leq C_1 C_2^{1/s} \left( \mathbb{E} \left\| \sum_{1 \leq i, j \leq n} u_i \tilde{u}_j A_{ij} \right\|^s \right)^{1/s}.$$

A homogeneous *Gaussian* chaos is one that involves only products of Hermite polynomials with the same total degree. For instance, a homogeneous Gaussian chaos of order 2 takes the form  $\sum_{1 \leq i \neq j \leq n} g_i g_j A_{ij} + \sum_{i=1}^n (g_i^2 - 1) A_{ii}$ . It can be decoupled according to Arcones and Giné [2].

**THEOREM B.2.** *Let  $(u_i)$  and  $(\tilde{u}_i)$  be two iid Gaussian sequences and  $A_{ij}$  be in a Banach space where  $1 \leq i, j \leq n$ . There exists universal constants  $C_1, C_2 > 0$  such that for any  $s \geq 1$ ,*

$$\left( \mathbb{E} \left\| \sum_{1 \leq i \neq j \leq n} u_i u_j A_{ij} + \sum_{i=1}^n (u_i^2 - 1) A_{ii} \right\|^s \right)^{1/s} \leq C_1 C_2^{1/s} \left( \mathbb{E} \left\| \sum_{1 \leq i, j \leq n} u_i \tilde{u}_j A_{ij} \right\|^s \right)^{1/s}.$$

*Remark B.1.* For Rademacher chaos,  $C_1 = 4$  and  $C_2 = 1$ . For Gaussian chaos, we can integrate equation (2.6) of [2] (with  $m = 2$ ) to obtain  $C_1 = 2^{1/2}$  and  $C_2 = 2^{14}$ . Better constants may be available.

We now proceed to the Khintchine inequalities. Let  $\|\cdot\|_{C_s}$  denote the  $s$ -Schatten norm. Recall that  $\|A\|_{C_s} = (\sum_i |\sigma_i|^s)^{1/s}$ , where  $\sigma_i$  is a singular value of  $A$ . The following is due to Lust-Piquard and Pisier [18, 19].

**THEOREM B.3.** *Let  $s \geq 2$  and  $(u_i)$  be a Rademacher or Gaussian sequence. Then for any set of matrices  $\{A_i\}_{1 \leq i \leq n}$ ,*

$$\left( \mathbb{E} \left\| \sum_{i=1}^n u_i A_i \right\|_{C_s}^s \right)^{1/s} \leq s^{1/2} \max \left( \left\| \left( \sum_{i=1}^n A_i^* A_i \right)^{1/2} \right\|_{C_s}, \left\| \left( \sum_{i=1}^n A_i A_i^* \right)^{1/2} \right\|_{C_s} \right).$$

The factor  $s^{1/2}$  above is not optimal. See, for example, [3] or [24, 26].

In [23], Theorem B.3 is applied twice in a clever way to obtain a Khintchine inequality for a decoupled chaos of order 2.

**THEOREM B.4.** *Let  $s \geq 2$  and  $(u_i)$  and  $(\tilde{u}_i)$  be two independent Rademacher or Gaussian sequences. For any set of matrices  $\{A_{ij}\}_{1 \leq i, j \leq n}$ ,*

$$\left( \mathbb{E} \left\| \sum_{1 \leq i, j \leq n} u_i \tilde{u}_j A_{ij} \right\|_{C_s}^s \right)^{1/s} \leq 2^{1/s} s \max \left( \|Q^{1/2}\|_{C_s}, \|R^{1/2}\|_{C_s}, \|F\|_{C_s}, \|G\|_{C_s} \right),$$

where  $Q = \sum_{1 \leq i, j \leq n} A_{ij}^* A_{ij}$  and  $R = \sum_{1 \leq i, j \leq n} A_{ij} A_{ij}^*$  and  $F, G$  are the block matrices  $(A_{ij})_{1 \leq i, j \leq n}$ ,  $(A_{ij}^*)_{1 \leq i, j \leq n}$ , respectively.

For Rademacher and Gaussian chaos, higher moments are controlled by lower moments, a property known as “hypercontractivity” [2, 8]. This leads to exponential tail bounds by Markov’s inequality as we illustrate below. The same result appears as Proposition 6.5 in [24].

PROPOSITION B.5. *Let  $X$  be a nonnegative random variable. Let  $\sigma, c, \alpha > 0$ . Suppose  $(\mathbb{E}X^s)^{1/s} \leq \sigma c^{1/s} s^{1/\alpha}$  for all  $s_0 \leq s < \infty$ . Then for any  $k > 0$  and  $u \geq s_0^{1/\alpha}$ ,*

$$\mathbb{P}(X \geq e^k \sigma u) \leq c \exp(-ku^\alpha).$$

*Proof.* By Markov’s inequality, for any  $s > 0$ ,  $\mathbb{P}(X \geq e^k \sigma u) \leq \frac{\mathbb{E}X^s}{(e^k \sigma u)^s} \leq c \left(\frac{\sigma s^{1/\alpha}}{e^k \sigma u}\right)^s$ . Pick  $s = u^\alpha \geq s_0$  to complete the proof.  $\square$

PROPOSITION B.6. *Let  $(u_i)$  be a Rademacher or Gaussian sequence and  $C_1, C_2$  be constants obtained from Theorem B.1 or B.2. Let  $\{A_{ij}\}_{1 \leq i, j \leq n}$  be a set of  $p$  by  $p$  matrices, and assume that the diagonal entries  $A_{ii}$  are positive semidefinite. Define  $M = \sum_i u_i u_j A_{ij}$  and  $\sigma = C_1 \max(\|Q\|^{1/2}, \|R\|^{1/2}, \|F\|, \|G\|)$ , where  $Q, R, F, G$  are as defined in Theorem B.4. Then*

$$\mathbb{P}(\|M - \mathbb{E}M\| \geq e\sigma u) \leq (2C_2 np) \exp(-u).$$

*Proof.* We will prove the Gaussian case first. Recall that the  $s$ -Schatten and spectral norms are equivalent: for any  $A \in \mathbb{C}^{r \times r}$ ,  $\|A\| \leq \|A\|_{C_s} \leq r^{1/s} \|A\|$ . Apply the decoupling inequality, that is, Theorem B.2, and deduce that for any  $s \geq 2$ ,

$$(\mathbb{E}\|M - N\|^s)^{1/s} \leq C_1 C_2^{1/s} \left( \mathbb{E} \left\| \sum_{1 \leq i, j \leq n} u_i \tilde{u}_j A_{ij} \right\|_{C_s}^s \right)^{1/s}.$$

Invoke Khintchine’s inequality, that is, Theorem B.4, and obtain

$$\begin{aligned} (\mathbb{E}\|M - N\|^s)^{1/s} &\leq C_1 (2C_2)^{1/s} s \max \left( \left\| Q^{1/2} \right\|_{C_s}, \left\| R^{1/2} \right\|_{C_s}, \|F\|_{C_s}, \|G\|_{C_s} \right) \\ &\leq C_1 (2C_2 np)^{1/s} s \max \left( \|Q\|^{1/2}, \|R\|^{1/2}, \|F\|, \|G\| \right) \\ &\leq \sigma (2C_2 np)^{1/s} s. \end{aligned}$$

Apply Proposition B.5 with  $c = 2C_2 np$  and  $k = \alpha = 1$  to complete the proof for the Gaussian case. For the Rademacher case, we take similar steps. First, decouple  $(\mathbb{E}\|M - N\|^s)^{1/s}$  using Theorem B.1. This leaves us a sum that excludes the  $A_{ii}$ ’s. Apply Khintchine’s inequality with the  $A_{ii}$ ’s zeroed. Of course,  $Q, R, F, G$  in Proposition B.4 will not contain any  $A_{ii}$ ’s, but this does not matter because  $A_{ii}^* A_{ii}$  and  $A_{ii} A_{ii}^*$  and  $A_{ii}$  are all positive semidefinite for any  $1 \leq i \leq n$  and we can add them back. For example,  $\|(A_{ij})_{1 \leq i \neq j \leq n}\| \leq \|(A_{ij})_{1 \leq i, j \leq n}\|$  as block matrices.  $\square$

We conclude the paper by outlining how better constants (see Remark B.1) can be obtained for the Gaussian case. At the start of the proof of Proposition B.6, we can split  $(\mathbb{E}\|M - N\|^s)^{1/s}$  into two parts,  $(\mathbb{E}\|\sum_{1 \leq i \neq j \leq n} u_i u_j A_{ij}\|^s)^{1/s}$  and  $(\mathbb{E}\|\sum_{i=1}^n (u_i^2 - 1) A_{ii}\|^s)^{1/s}$ . For the first part, decouple using Theorem B.1 with  $C_2 = 1$ , then apply Theorem B.4. For the second part, note that every  $u_i^2 - 1$  is symmetrically distributed and has zero mean. Thus, we can introduce Rademacher variables, condition on the Gaussians, apply Theorem B.3, and pull out  $(\mathbb{E} \max_i |u_i^2 - 1|^s)^{1/s}$ . Although this log factor is in practice smaller than the constants we have, we prefer to avoid it by decoupling the Gaussian sum right away using [2].

**Acknowledgments.** We thank Holger Rauhut for interesting discussions and the anonymous referees for suggesting several improvements to the paper.

## REFERENCES

- [1] C. ANDRIEU, N. DE FREITAS, A. DOUCET, AND M.I. JORDAN, *An introduction to MCMC for machine learning*, Machine Learning, 50 (2003), pp. 5–43.
- [2] M.A. ARCONES AND E. GINÉ, *On decoupling, series expansions, and tail behavior of chaos processes*, J. Theoret. Probab., 6 (1993), pp. 101–122.
- [3] A. BUCHHOLZ, *Operator khintchine inequality in non-commutative probability*, Math. Ann., 319 (2001), pp. 1–16.
- [4] E.J. CANDÈS, J.K. ROMBERG, AND T. TAO, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223.
- [5] T.F.C. CHAN AND T.P. MATHEW, *The interface probing technique in domain decomposition*, SIAM J. Matrix Anal. Appl., 13 (1992), p. 212.
- [6] T.F. CHAN AND D.C. RESASCO, *A Survey of Preconditioners for Domain Decomposition*, Technical report, DTIC Document, Yale University, New Haven, CT, 1985.
- [7] E.C. CHANG, S. MALLAT, AND C. YAP, *Wavelet foveation*, Appl Comput. Harmon. Anal., 9 (2000), pp. 312–335.
- [8] V. DE LA PEÑA AND E. GINÉ, *Decoupling: From Dependence to Independence*, Springer-Verlag, New York, 1999.
- [9] L. DEMANET, P.D. LÉTOURNEAU, N. BOUMAL, H. CALANDRA, J. CHIU, AND S. SNELSON, *Matrix probing: A randomized preconditioner for the wave-equation Hessian*, Appl. Comput. Harmon. Anal., 32 (2012), pp. 155–168.
- [10] L. DEMANET AND L. YING, *Discrete symbol calculus*, SIAM Rev., 53 (2011), pp. 71–104.
- [11] A. EDELMAN, *Eigenvalues and Condition Numbers of Random Matrices*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1989.
- [12] G.B. FOLLAND, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1995.
- [13] N. HALE, N.J. HIGHAM, L.N. TREFETHEN, ET AL., *Computing  $\alpha$ ,  $\log(a)$ , and related matrix functions by contour integrals*, SIAM J. Numer. Anal., 46 (2008), pp. 2505–2523.
- [14] N. HALKO, P. MARTINSSON, AND J. TROPP, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), p. 217.
- [15] N.J. HIGHAM, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.
- [16] D.R. KARGER AND C. STEIN, *A new approach to the minimum cut problem*, J. ACM, 43 (1996), pp. 601–640.
- [17] R.M. KARP AND M.O. RABIN, *Efficient randomized pattern-matching algorithms*, IBM J. Res. Development, 31 (1987), pp. 249–260.
- [18] F. LUST-PIQUARD, *Inégalités de Khintchine dans  $C_p$* , CR Acad. Sci. Paris, 303 (1986), pp. 289–292.
- [19] F. LUST-PIQUARD AND G. PISIER, *Non-commutative Khintchine and Paley inequalities*, Ark. Mat., 29 (1991), pp. 241–260.
- [20] G.E. PFANDER AND H. RAUHUT, *Sparsity in time-frequency representations*, J. Fourier Anal. Appl., 16 (2010), pp. 233–260.
- [21] G.E. PFANDER, H. RAUHUT, AND J. TANNER, *Identification of matrices having a sparse representation*, IEEE Trans. Signal Process., 56 (2008), pp. 5376–5388.
- [22] G.E. PFANDER, H. RAUHUT, AND J.A. TROPP, *The restricted isometry property for time-frequency structured random matrices*, arXiv:1106.3184, 2011.
- [23] H. RAUHUT, *Circulant and Toeplitz matrices in compressed sensing*, in Proceedings of SPARS’09, Saint-Malo, France, 2009.
- [24] H. RAUHUT, *Compressive sensing and structured random matrices*, Theoret. Found. Numer. Methods Sparse Recovery, 9 (2010), pp. 1–92.
- [25] M.A. SHUBIN, *Pseudodifferential Operators and Spectral Theory*, Springer-Verlag, New York, 2001.
- [26] J.A. TROPP, *On the conditioning of random subdictionaries*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 1–24.