Linear Algebra and its Applications 682 (2024) 191–237

Contents lists available at ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

Stability of the Lanczos algorithm on matrices with regular spectral distributions ☆

Tyler Chen [a,*], Thomas Trogdon [b]

[a] New York University, United States of America
[b] University of Washington, United States of America

ARTICLE INFO

Article history:
Received 25 April 2023
Received in revised form 6 September 2023
Accepted 9 November 2023
Available online 15 November 2023
Submitted by D. Kressner

MSC:
65F10
65F15
65G50

Keywords:
Lanczos
Finite precision arithmetic
Random matrix

ABSTRACT

We study the stability of the Lanczos algorithm run on problems whose eigenvector empirical spectral distribution is near to a reference measure with well-behaved orthogonal polynomials. We give a backwards stability result which can be upgraded to a forward stability result when the reference measure has a density supported on a single interval with square root behavior at the endpoints. Our analysis implies the Lanczos algorithm run on many large random matrix models is in fact forward stable, and hence nearly deterministic, even when computations are carried out in finite precision arithmetic. Since the Lanczos algorithm is not forward stable in general, this provides yet another example of the fact that random matrices are far from "any old matrix", and care must be taken when using them to test numerical algorithms.

☆ **Funding.** This material is based on work supported by the National Science Foundation under Grant Nos. DGE-1762114 (TC), DMS-1945652 (TT). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

* Corresponding author.
E-mail addresses: tyler.chen@nyu.edu (T. Chen), trogdon@uw.edu (T. Trogdon).

https://doi.org/10.1016/j.laa.2023.11.006
0024-3795/

1. Introduction

The Lanczos algorithm is unstable in the sense that, even on the simplest problems, the output of the algorithm in finite precision arithmetic may be very different than what would have been obtained in exact arithmetic. Despite this, the Lanczos algorithm is among the most important algorithms in numerical linear algebra and is commonly used for a wide variety of fundamental linear-algebraic tasks including approximating eigenvalues and eigenvectors, the product of a matrix function on a vector, and quadratic forms involving matrix functions. Understanding the behavior of the Lanczos algorithm in finite precision arithmetic has been of interest since the introduction of the algorithm some 70 years ago [4,17,25,26,38].

Algorithm 1 Lanczos algorithm.

```
procedure LANCZOS(A, b, k)
    q_0 = b/‖b‖, β_{-1} = 0, q_{-1} = 0
    for n = 0, 1, ..., k − 1 do
        q̃_{n+1} = A q_n − β_{n−1} q_{n−1}
        α_n = q̃_{n+1}^T q_n
        q̂_{n+1} = q̃_{n+1} − α_n q_i
        β_n = ‖q̂_{n+1}‖
        q_{n+1} = q̂_{n+1}/β_n
    end for
end procedure
```

Throughout, $\mathbf{A}$ will be an $N \times N$ real symmetric matrix and $\mathbf{b}$ a unit-norm vector of length N. From $(\mathbf{A}, \mathbf{b})$ we obtain the eigenvector empirical spectral distribution (VESD) defined by

$$\mu_N(\mathrm{d}x) = \mu_{\mathrm{VESD}}(\mathrm{d}x; \mathbf{A}, \mathbf{b}) := \sum_{n=1}^{N} (\mathbf{b}^{\mathsf{T}}\mathbf{u}_n)^2 \, \delta_{\lambda_n}(\mathrm{d}x), \tag{1.1}$$

where $(\lambda_n, \mathbf{u}_n)$ are the eigenvalue-vector pairs of $\mathbf{A}$ and δ_c is the Dirac delta distribution centered at c. We use the former notation when $\mathbf{A}$ and $\mathbf{b}$ are clear from context. When run on $(\mathbf{A}, \mathbf{b})$ for k iterations *in exact arithmetic*, the Lanczos algorithm (Algorithm 1) outputs an orthonormal basis $[\mathbf{q}_0, \ldots, \mathbf{q}_k]$ for the Krylov subspace

$$\mathrm{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^k\mathbf{b}\}$$

and coefficients $(\alpha_0, \ldots, \alpha_{k-1})$, $(\beta_0, \ldots, \beta_{k-1})$ for a three-term recurrence satisfied by the basis vectors. In matrix form, this recurrence can be written

$$\mathbf{A}\mathbf{Q}_k = \mathbf{Q}_k\mathbf{T}_k + \beta_{k-1}\mathbf{q}_k\mathbf{e}_{k-1}^{\mathsf{T}}, \tag{1.2}$$

where $\mathbf{e}_{k-1} = [0, \ldots, 0, 1]^{\mathsf{T}}$ and

$$\mathbf{Q}_k = \begin{bmatrix} | & | & & | \\ \mathbf{q}_0 & \mathbf{q}_1 & \cdots & \mathbf{q}_{k-1} \\ | & | & & | \end{bmatrix}, \quad \mathbf{T}_k = \operatorname{tridiag}\begin{pmatrix} & \beta_0 & \beta_1 & \cdots & \beta_{k-2} & \\ \alpha_0 & & \alpha_1 & \cdots & & \alpha_{k-1} \\ & \beta_0 & \beta_1 & \cdots & \beta_{k-2} & \end{pmatrix}. \tag{1.3}$$

The Lanczos algorithm run on $(\mathbf{A}, \mathbf{b})$ is mathematically equivalent to the Stieltjes procedure for computing the recurrence coefficients for the orthogonal polynomials of the VESD μ_N [15]. It is common to refer to the matrix $\mathbf{T}_k$ as the Jacobi matrix associated with μ_N, and from this point on, we will make no distinction between the Lanczos algorithm in exact arithmetic and Stieltjes procedure. The k-point Gaussian quadrature rule for μ_N will be written as μ_k, and is equal to the VESD for $(\mathbf{T}_k, \mathbf{e}_0)$, where $\mathbf{e}_0 = [1, 0, \ldots, 0]^\mathsf{T}$. That is,

$$\mu_k(\mathrm{d}x) = \mu_{\mathrm{VESD}}(\mathrm{d}x; \mathbf{T}_k, \mathbf{e}_0) := \sum_{n=1}^{k} (\mathbf{e}_0^\mathsf{T} \mathbf{s}_n)^2 \, \delta_{\theta_n}(\mathrm{d}x), \tag{1.4}$$

where $(\theta_n, \mathbf{s}_n)$ are the eigenvalue-vector pairs of $\mathbf{T}_k$. Note that (1.1) and (1.4) coincide once k is large enough that the dimension of the Krylov subspace stops growing. This occurs once k is equal to the number of points of support for μ_N. However, implicit in our analysis, is the assumption $k \ll N$.

When the Lanczos algorithm is run on $(\mathbf{A}, \mathbf{b})$ for k iterations in *finite precision arithmetic*, the vectors $[\overline{\mathbf{q}}_0, \ldots, \overline{\mathbf{q}}_k]$ and coefficients $(\overline{\alpha}_0, \ldots, \overline{\alpha}_{k-1})$, $(\overline{\beta}_0, \ldots, \overline{\beta}_{k-1})$ generated by the algorithm may be nothing like their exact arithmetic counterparts. Analogously to (1.4), we define the VESD $\overline{\mu}_k$ for $(\overline{\mathbf{T}}_k, \mathbf{e}_0)$ by

$$\overline{\mu}_k(\mathrm{d}x) = \mu_{\mathrm{VESD}}(\mathrm{d}x; \overline{\mathbf{T}}_k, \mathbf{e}_0) := \sum_{n=1}^{k} (\mathbf{e}_0^\mathsf{T} \overline{\mathbf{s}}_n)^2 \, \delta_{\overline{\theta}_n}(\mathrm{d}x), \tag{1.5}$$

where $(\overline{\theta}_n, \overline{\mathbf{s}}_n)$, $n = 1, \ldots, k$ are the eigenvalues-vectors pairs of $\overline{\mathbf{T}}_k$, the symmetric tridiagonal matrix with diagonal $(\overline{\alpha}_0, \ldots, \overline{\alpha}_{k-1})$ and sub/super-diagonals $(\overline{\beta}_0, \ldots, \overline{\beta}_{k-2})$.

In numerical analysis, there are a number of notions of stability. Arguably, the most common are forward stability and backward stability, which we now describe in the context of the Lanczos algorithm.

Definition 1.1. The Lanczos algorithm run for k iterations in finite precision arithmetic on an input $(\mathbf{A}, \mathbf{b})$ to obtain output $\overline{\mathbf{T}}_k$ is

- *forward stable* if $\overline{\mathbf{T}}_k$ is near $\mathbf{T}_k$, the output of exact Lanczos run on $(\mathbf{A}, \mathbf{b})$, and
- *backward stable* if $\overline{\mathbf{T}}_k$ is the Jacobi matrix for a nearby input $(\mathbf{A}_*, \mathbf{b}_*)$; that is, if exact Lanczos run on $(\mathbf{A}_*, \mathbf{b}_*)$ produces $\overline{\mathbf{T}}_k$.

For the purposes of this paper, *nearby* is understood to mean differing by an amount with a polynomial dependence on k and the machine precision ϵ_{mach} (in some reason-

able metric). Ideally, the dependence on ϵ_{mach} is linear, and when $\epsilon_{\text{mach}} = 0$, the exact arithmetic behavior is recovered. As with most stability analyses of the Lanczos algorithm, the value of our work is not in the numerical value of the bounds themselves, but rather in the intuition the bounds convey. For instance, situations in which our bounds depend exponentially on k provide insight into problems on which the Lanczos algorithm is potentially unstable. In line with this philosophy, we will not attempt to optimize polynomial dependencies in k; instead, we aim to minimize the complexity of the statements and proofs of our results.

As noted above, understanding the stability of the Lanczos algorithm in finite precision arithmetic has been an active area of the research for the past half century. Perhaps the most well-known work is that of Paige [34–36] (which we discuss further in Section 2.1) and Greenbaum [18]. In addition, a number of books and notes contain extensive writing on the topic [26,38].

Greenbaum's analysis, which is the preeminent backwards stability analysis of the Lanczos algorithm, proves the existence of a nearby problem $(\mathbf{A}_*, \mathbf{b}_*)$ such that, when Lanczos is run on $(\mathbf{A}_*, \mathbf{b}_*)$ for k iterations in exact arithmetic, $\overline{\mathbf{T}}_k$ is output. Here nearby roughly means (i) every eigenvalue of $\mathbf{A}_*$ is near an eigenvalue of $\mathbf{A}$, and (ii) $\mu_{\text{VESD}}(\cdot\,; \mathbf{A}_*, \mathbf{b}_*)$ is near to $\mu_N = \mu_{\text{VESD}}(\cdot\,; \mathbf{A}, \mathbf{b})$. This result is very strong in that it applies to any input $(\mathbf{A}, \mathbf{b})$. The main drawbacks are that the nearby problem $(\mathbf{A}_*, \mathbf{b}_*)$ is of a different dimension than the original problem, and the precise definition of nearby has a sub-linear dependence on the machine precision which is generally believed to be pessimistic. In addition, the proofs of the result are quite technical.

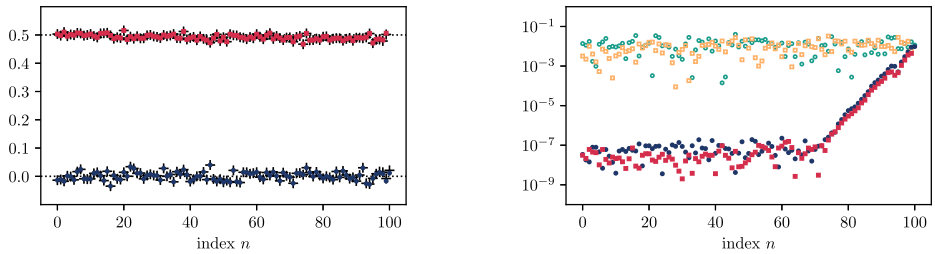
Another important stability result, which seems to have been mostly overlooked by the numerical analysis community, is Knizhnerman's analysis of the modified Chebyshev moments of $\overline{\mu}_k$ [21]. In particular, Knizhnerman shows that the modified Chebyshev moments of $\overline{\mu}_k$ are near those of μ_k. This paper extends Knizhnerman's work.

1.1. Motivation

Testing numerical algorithms on random matrices is a widespread practice. However, as noted by Edelman and Rao [11],

> It is a mistake to link psychologically a random matrix with the intuitive notion of a 'typical' matrix or the vague concept of 'any old matrix'.

In particular, numerical algorithms run on random matrices may fail to capture the typical behavior of the algorithm on an arbitrary matrix. The Lanczos algorithm is a clear example of this. While the algorithm is not forward stable in general, when run on a large random matrix, drawn from a suitable distribution, $\overline{\mathbf{T}}_k$ matches closely to $\mathbf{T}_k$, at least while number of iterations k is sufficiently small compared to the dimension N. We illustrate this phenomenon numerically in Fig. 1.



(a) *Recurrence coefficients* $\overline{\alpha}_n$ (•) *and* $\overline{\beta}_n$ (▪). *Exact arithmetic counterparts shown as pluses* (+) *and limiting values shown as dotted lines* (·····).

(b) *Forward error of recurrence coefficients* $|\alpha_n - \overline{\alpha}_n|$ (•) *and* $|\beta_n - \overline{\beta}_n|$ (▪) *and distance to limiting values* $|0 - \overline{\alpha}_n|$ (◦) *and* $|1/2 - \overline{\beta}_n|$ (▫).

Fig. 1. Here $(\mathbf{A}, \mathbf{b})$ corresponds to a 2000×2000 random matrix, drawn from the Gaussian orthogonal ensemble (see Section 5.1), and independent vector. In the large N limit, the VESD of matrices drawn from this ensemble converge to the semicircle distribution on $[-1, 1]$ (density $\propto \sqrt{1-x^2}$). Therefore the Lanczos coefficients α_i and β_i from the "exact" computation (with reorthogonalization in quadruple precision arithmetic) respectively converge to $1/2$ and 0; i.e. the Lanczos algorithm exhibits deterministic behavior. In our particular experiment we observe fluctuations on the order of 10^{-2} around the limiting values due to finite N effects. Remarkably, the coefficients $\overline{\alpha}_n$ and $\overline{\beta}_n$ output by the Lanczos algorithm run in *single precision* floating point arithmetic without reorthogonalization are within the unit roundoff ($\approx 10^{-7}$) of α_n and β_n, at least while n is sufficiently small; i.e. the algorithm is forward stable.

The aim of this paper is to provide an intuitive explanation for the observation that the Lanczos algorithm is stable on problems whose VESD are sufficiently regular. More specifically, our approach extends the work of Knizhnerman [21] to prove the existence of a measure μ_* near to μ_N whose moments agree with $\overline{\mu}_k$ through degree $2k-1$, at least when the VESD of $(\mathbf{A}, \mathbf{b})$ is sufficiently regular. In fact, under certain regularity conditions, we show there exists a vector $\mathbf{b}_*$ near to $\mathbf{b}$ such that Lanczos run on $(\mathbf{A}, \mathbf{b}_*)$ in exact arithmetic for k iterations outputs $\overline{\mathbf{T}}_k$. In other words, on a restricted set of inputs, we provide a simpler proof for a stronger version of Greenbaum's results. We then provide forward stability results by analyzing the orthogonal polynomials of slightly perturbed measures. This shows that, on many large random matrix models, the output of the Lanczos algorithm is nearly deterministic, even when computations are carried out in finite precision arithmetic. Our analysis is accompanied by numerical experiments and several explicit examples.

1.2. Notation

Throughout this work, we use $\Lambda(\mathbf{A})$ to refer to the spectrum of a matrix. For a function $f : U \to \mathbb{C}$ with $S \subseteq U$, we define $\|f\|_S := \sup_{x \in S} |f(x)|$. For a vector $\mathbf{b}$, $\|\mathbf{b}\|$ refers to the Euclidean 2-norm and $\|\mathbf{A}\|$ gives the associated induced operator norm for a matrix $\mathbf{A}$. The n-th canonical basis vector, indexed from 0, is $\mathbf{e}_n$. The Kolmogorov–Smirnov distance between two measures ν_1 and ν_2 is $d_{\mathrm{KS}}(\nu_1, \nu_2) := \sup_{x \in \mathbb{R}} |\nu_1((-\infty, x]) - \nu_2((-\infty, x])$. All measures we consider will be Borel measures. Indeed, all measures will be either fully discrete or have a continuous density.

2. Setup and background

Let μ be a unit-mass measure with support contained in $[a, b]$. We will refer to μ as the *reference measure*, and it will be helpful to think of μ as near to μ_N; for instance $\mu = \mu_N$ or μ being the limiting measure for the VESD of a large random matrix ensemble. In particular, we will typically have $[a, b] \approx [\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})]$. We denote by $p_n = p_n(\cdot\,; \mu)$, $n \geq 0$ the orthonormal polynomials for μ. That is, the $p_n(\cdot\,; \mu)$ satisfy[1]

$$\int p_n(x; \mu) p_i(x; \mu) \mu(\mathrm{d}x) = \mathbb{1}(n = i),$$

where $\mathbb{1}(\texttt{true}) = 1$ and $\mathbb{1}(\texttt{false}) = 0$.

The modified moments of a measure ν with respect to the orthogonal polynomials of μ are defined by

$$\mathfrak{m}_n(\nu; \mu) := \int p_n(x; \mu) \nu(\mathrm{d}x), \qquad n \geq 0. \tag{2.1}$$

Clearly $\mathfrak{m}_n(\mu; \mu) = \mathbb{1}(n = 0)$ and $\mathfrak{m}_0(\mu; \mu) = \mathfrak{m}_0(\mu_N; \mu) = \mathfrak{m}_0(\mu_k; \mu) = \mathfrak{m}_0(\overline{\mu}_k; \mu) = 1$.

As mentioned in the introduction, [21] shows that the modified moments of μ_N and $\overline{\mu}_k$ through degree $2k - 1$ are close when μ is a properly scaled and shifted version of the orthogonality measure for the Chebyshev polynomials of the first kind. A similar statement, with some polynomial losses in k, can therefore be expected to hold for any μ whose orthogonal polynomials have a Chebyshev series representation with reasonable coefficients.

The idea underlying our analysis is to construct a (potentially signed) measure μ_* as a perturbation to the reference measure μ:

$$\mu_*(\mathrm{d}x) := (1 + h(x))\, \mu(\mathrm{d}x), \quad h(x) := \sum_{n=0}^{2k-1} \big(\mathfrak{m}_n(\overline{\mu}_k; \mu) - \mathfrak{m}_n(\mu; \mu)\big) p_n(x; \mu). \tag{2.2}$$

This construction ensures μ_* has the same moments as $\overline{\mu}_k$ through degree $2k - 1$ and the same moments as μ for higher degrees. Indeed, by definition, the p_n are orthonormal with respect to μ, so

$$\begin{aligned} \mathfrak{m}_n(\mu_*; \mu) &= \int p_n(x; \mu) \mu_*(\mathrm{d}x) \\ &= \int p_n(x; \mu) \mu(\mathrm{d}x) + \int p_n(x; \mu) \sum_{i=0}^{2k-1} \big(\mathfrak{m}_n(\overline{\mu}_k; \mu) - \mathfrak{m}_n(\mu; \mu)\big) p_i(x; \mu) \mu(\mathrm{d}x) \end{aligned}$$

[1] These polynomials are constructed by performing Gram–Schmidt on the monomial basis in order of increasing degree and are normalized to have a positive leading coefficient.

$$= \mathfrak{m}_n(\mu;\mu) + \sum_{i=0}^{2k-1} \left(\mathfrak{m}_n(\overline{\mu}_k;\mu) - \mathfrak{m}_n(\mu;\mu)\right) \int p_n(x;\mu)p_i(x;\mu)\mu(\mathrm{d}x)$$

$$= \mathfrak{m}_n(\mu;\mu) + \sum_{i=0}^{2k-1} \left(\mathfrak{m}_n(\overline{\mu}_k;\mu) - \mathfrak{m}_i(\mu;\mu)\right)\mathbb{1}(i=n)$$

$$= \begin{cases} \mathfrak{m}_n(\overline{\mu}_k;\mu) & i = 0,1,\ldots,2k-1 \\ \mathfrak{m}_n(\mu;\mu) & i = 2k, 2k+1, \ldots. \end{cases}$$

Since the moments of μ_* match those of $\overline{\mu}_k$ through degree $2k-1$, when the Stieltjes procedure is run on μ_* for k iterations, $\overline{\mathbf{T}}_k$ is the output.

Introduce the quantities

$$M_k(\mu;[a,b]) := \max_{n\le 2k-1} \|p_n(\cdot\,;\mu)\|_{[a,b]}, \tag{2.3}$$

$$\mathfrak{m}_k^{\Delta}(\nu_1,\nu_2;\mu) := \max_{n\le 2k-1} |\mathfrak{m}_n(\nu_1;\mu) - \mathfrak{m}_n(\nu_2;\mu)|. \tag{2.4}$$

Clearly

$$\|h\|_{[a,b]} \le \sum_{i=0}^{2k-1} |\mathfrak{m}_n(\overline{\mu}_k;\mu) - \mathfrak{m}_n(\mu;\mu)| M_k(\mu;[a,b]) \le 2k\mathfrak{m}_k^{\Delta}(\overline{\mu}_k,\mu;\mu)M_k(\mu;[a,b]), \tag{2.5}$$

so if $\mathfrak{m}_k^{\Delta}(\overline{\mu}_k,\mu;\mu)$ is sufficiently small relative to the reciprocal of $kM_k(\mu;[a,b])$, then $\|h\|_{[a,b]} \le 1$ and μ_* is a well-defined non-negative measure. In this case, if $\mu \approx \mu_N$ then we also have $\mu_* \approx \mu_N$; i.e. backwards stability.

Remark 2.1. If we take $\mu = \mu_N$, then, assuming $\|h\|_{\Lambda(\mathbf{A})} \le 1$, μ_* is the VESD of $(\mathbf{A}, \mathbf{b}_*)$, where

$$\mathbf{b}_* = (\mathbf{I} + h(\mathbf{A}))^{1/2}\mathbf{b}.$$

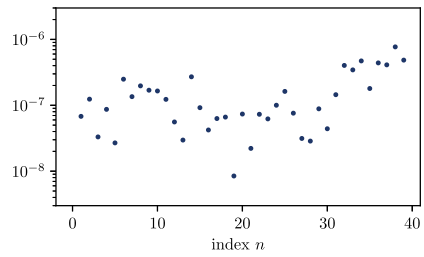
This is a perturbation of $\mathbf{b}$ in the sense that

$$\|\mathbf{b} - \mathbf{b}_*\| \le \|\mathbf{I} - (\mathbf{I} + h(\mathbf{A}))^{1/2}\|\|\mathbf{b}\| \le \|h(\mathbf{A})\|.$$

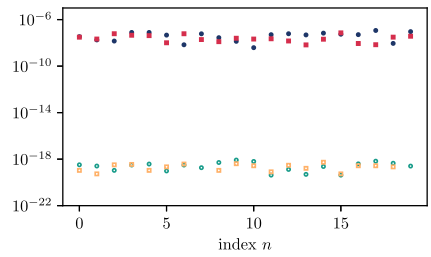
In Fig. 2 we illustrate this approach for $\mu = \mu_N$, where $\mathbf{A}$ is the same random matrix as used in Fig. 1. Bounds for μ_* are derived in Sections 3 and 4.

There is a fundamental equivalence between the Jacobi matrix $\mathbf{T}_k$ produced by the Stieltjes procedure and the modified moments of μ_N through degree $2k-1$ with respect to some fixed measure.[2] The conditioning of the map from moments to Jacobi matrix is very poor in general [14–16,29], so even if the modified moments of μ_N and $\overline{\mu}_k$ are close,

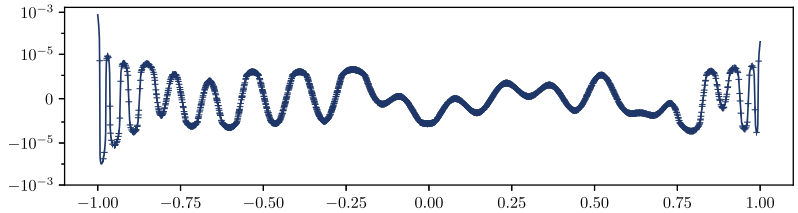
[2] These quantities are also equivalent to the k-point Gaussian quadrature rule μ_k for μ_N.



(a) *Forward error of modified moments* $|\mathfrak{m}_n(\mu_N;\mu_N) - \mathfrak{m}_n(\overline{\mu}_k;\mu_N)|$ (•). *Note that* $\mathfrak{m}_n(\mu_N;\mu_N) = \mathfrak{m}_n(\mu_k;\mu_N)$ *for* $n \le 2k-1$.



(b) *Forward error of recurrence coefficients* $|\alpha_n - \overline{\alpha}_n|$ (•) *and* $|\beta_n - \overline{\beta}_n|$ (▪) *and* $|\alpha_n^* - \overline{\alpha}_n|$ (◦) *and* $|\beta_n^* - \overline{\beta}_n|$ (▫), *where* α_n^*, β_n^* *correspond to an "exact" computation with* $(\mathbf{A}_*, \mathbf{b}_*)$.



(c) *Perturbation function* $h(x)$ (——) *and value at eigenvalues of* $\mathbf{A}$ (+).

Fig. 2. We use the same 2000×2000 random matrix $\mathbf{A}$ and fixed vector $\mathbf{b}$ from Fig. 1. After running the Lanczos algorithm on $(\mathbf{A}, \mathbf{b})$ in single precision finite precision arithmetic without reorganization, we use (2.2) and Remark 2.1 to construct a slightly perturbed $\mathbf{b}_*$. Lanczos run on $(\mathbf{A}, \mathbf{b}^*)$ "exactly" (with reorthogonization in quadruple precision floating point arithmetic) produces a nearly identical output as the original single precision finite precision computation. For reference, $\|\mathbf{b} - \mathbf{b}_*\| \approx 3.6 \cdot 10^{-5}$ is only a few orders of magnitude above the machine precision in which the original computation was carried out.

this does not generally imply the corresponding Jacobi matrices $\mathbf{T}_k$ and $\overline{\mathbf{T}}_k$ are close. However, in certain situations when μ_N is sufficiently regular, then the conditioning of the map from modified moments to Jacobi matrices is well-conditioned [14] and we can expect the Lanczos algorithm to be forward stable. Bounds for $\overline{\mathbf{T}}_k$ are derived in Section 4.

2.1. Perturbed Lanczos recurrences

To carry out our analysis, we require some understanding of the behavior of the Lanczos methods in finite precision arithmetic. Much is known about this topic [26,38], but we summarize only what is needed for our analysis.

The finite precision arithmetic outputs $\overline{\mathbf{Q}}_k$ and $\overline{\mathbf{T}}_k$ no longer satisfy the three-term Lanczos recurrence (1.2) exactly. Instead, they satisfy a perturbed recurrence

$$\mathbf{A}\overline{\mathbf{Q}}_k = \overline{\mathbf{Q}}_k\overline{\mathbf{T}}_k + \overline{\beta}_{k-1}\overline{\mathbf{q}}_k\mathbf{e}_{k-1}^{\mathsf{T}} + \mathbf{F}_k, \tag{2.6}$$

where the perturbation term $\mathbf{F}_k$ accounts for local rounding errors made by the algorithm. Since $\mathbf{F}_k$ does not involve accumulated errors, but rather errors made over a single iteration of the Lanczos algorithm, it can intuitively be expected to be small.

We denote by $\mathbf{R}_k$ and $\mathbf{D}_k$ the strictly upper triangular and diagonal parts of $\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{Q}}_k$ so that $\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{Q}}_k = \mathbf{R}_k + \mathbf{R}_k{}^{\mathsf{T}} + \mathbf{D}_k$; i.e.

$$\mathbf{R}_k := \text{strict-triu}(\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{Q}}_k), \qquad \mathbf{D}_k := \text{diag}(\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{Q}}_k). \tag{2.7}$$

Since $\mathbf{D}_k - \mathbf{I}$ accounts for errors made when normalizing the Lanczos basis vectors, we expect it to be small. On the other hand, since rounding errors accumulate in the columns of $\overline{\mathbf{Q}}_k$, the entries of $\mathbf{R}_k$ need not be small. In fact, in many situations $\mathbf{R}_k$ can have entries of size $O(1)$ indicating a complete loss of orthogonality in the Lanczos basis vectors.

The matrix $\mathbf{R}_k$ satisfies a perturbed three-term recurrence

$$\overline{\mathbf{T}}_k\mathbf{R}_k = \mathbf{R}_k\overline{\mathbf{T}}_k + \overline{\beta}_{k-1}\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{q}}_k\mathbf{e}_{k-1}^{\mathsf{T}} + \mathbf{H}_k, \tag{2.8}$$

with an upper triangular perturbation term $\mathbf{H}_k$. Straightforward algebraic manipulations of (2.6) using (2.7) show that $\mathbf{H}_k$ should be expected to be small as well.

Finally, we define $\eta_k \geq 0$ to be the smallest value such that

$$\Lambda(\overline{\mathbf{T}}_k) \subseteq [\lambda_{\min}(\mathbf{A}) - \eta_k, \lambda_{\max}(\mathbf{A}) + \eta_k]. \tag{2.9}$$

Definition 2.2. We say the Lanczos algorithm was run for k iterations with precision ϵ_{lan} if

$$\|\mathbf{F}_k\| \leq \|\mathbf{A}\|\epsilon_{\text{lan}}, \qquad \|\mathbf{D}_k - \mathbf{I}\| \leq \epsilon_{\text{lan}}, \qquad \|\mathbf{H}_k\| \leq \|\mathbf{A}\|\epsilon_{\text{lan}}, \qquad \eta_k \leq \|\mathbf{A}\|\epsilon_{\text{lan}}$$

where $\mathbf{F}_k$, $\mathbf{D}_k$, $\mathbf{H}_k$, and η_k are defined in (2.6) to (2.9).

Bounds for $\|\mathbf{F}_k\|$, $\|\mathbf{D}_k\|$, and $\|\mathbf{H}_k\|$ appear in [34,35] and the most well-known bound for η_k appears in [36]. More recently, Paige has shown a bound for η_k [32, Theorem A.1] which, when combined with [31, Theorem 3.1] improves the dependence on k in the bound for η_k over [36].

Proposition 2.3 *(informal; see [33,36]). When the Lanczos algorithm is run for k iterations on a computer with relative machine precision $\epsilon_{\text{mach}} < O(1/k)$, then*

$$\epsilon_{\text{lan}} = \max\left\{N, \frac{\||\mathbf{A}|_{\text{entry}}\|}{\|\mathbf{A}\|}\,\text{row-nnz}(\mathbf{A})\right\} O(\text{poly}(k)\epsilon_{\text{mach}}).$$

Here row-nnz$(\mathbf{A})$ *is the largest number of nonzero entries in a row of* $\mathbf{A}$ *and* $|\mathbf{A}|_{\text{entry}}$ *is the entry-wise absolute value of* $\mathbf{A}$*; i.e.* $[|\mathbf{A}|_{\text{entry}}]_{i,j} = |[\mathbf{A}]_{i,j}|$.

Paige's analysis is far more precise than Proposition 2.3. In particular the analyses result in explicit bounds on the powers of k and the constants in front of each of the quantities in Definition 2.2. In Paige's analyses, terms of order $(\epsilon_{\text{mach}})^2$ are typically

discarded for clarity, but the results are essentially the same if the higher order terms are accounted for.

Remark 2.4. It always holds that $\||\mathbf{A}|_{\text{entry}}\| \leq N^{1/2}\|\mathbf{A}\|$, so for uniformly sparse matrices with up to $O(N^{1/2})$ entries per row, the Lanczos algorithm is run with precision ϵ_{lan} if $\epsilon_{\text{mach}} = O(\epsilon_{\text{lan}}/(\text{poly}(k)N))$.

3. Backwards stability

Our first main result shows that constructing μ_* as in (2.2) gives a nearby problem to μ_N when the reference measure μ is chosen suitably.

Theorem 3.1 *(Backwards stability). There exist absolute constants C, D such that, for $(\mathbf{A}, \mathbf{b})$ with VESD μ_N and any unit-mass measure μ with support contained in $[a, b]$, the following statement holds:*

Suppose Lanczos is run on $(\mathbf{A}, \mathbf{b})$ for $k \geq 1$ iterations with precision $\epsilon_{\text{lan}} < 1/(\sigma C k^2)$, where $\sigma := \max\{1, 2\|\mathbf{A}\|/(b-a)\}$, to produce $\overline{\mu}_k$ and

$$\text{supp}(\mu_N) \subseteq [a - (b-a)/(32k^2), b + (b-a)/(32k^2)].$$

Then the (possibly-signed) measure μ_* constructed in (2.2) is close to μ in the sense that

(a) $\mathfrak{m}_k^{\Delta}(\mu_*, \mu_N; \mu) \leq D\sigma M_k(\mu; [a, b])k^3\epsilon_{\text{lan}}$, *and*
(b) $\|h\|_{[a,b]} \leq 2k\ M_k(\mu; [a, b])(\mathfrak{m}_k^{\Delta}(\mu_*, \mu_N; \mu) + \mathfrak{m}_k^{\Delta}(\mu_N, \mu; \mu))$

Furthermore, provided that $\|h\|_{[a,b]} < 1$, μ_* is non-negative measure whose moments through degree $2k-1$ exactly match those of $\overline{\mu}_k$.

The majority of the remainder of this section is devoted to proving Theorem 3.1.

As noted in Remark 2.1, if $\mu = \mu_N$, μ_* is the VESD of a nearby problem $(\mathbf{A}, \mathbf{b}_*)$, which is the *same dimension* as the original problem $(\mathbf{A}, \mathbf{b})$. In this case $\mathfrak{m}_k^{\Delta}(\mu_N, \mu_N; \mu_N) = 0$ and we have the following corollary:

Corollary 3.2. *Under the assumptions of Theorem 3.1 (with $\mu = \mu_N$), and assuming ϵ_{lan} is sufficiently small, there exists a nearby vector $\mathbf{b}^*$ satisfying*

$$\|\mathbf{b} - \mathbf{b}_*\| \leq 2\sigma D k^4 M_k(\mu_N; [a, b])^2 \epsilon_{\text{lan}},$$

such that the Lanczos algorithm run on $(\mathbf{A}, \mathbf{b}^*)$ for k iterations in exact arithmetic produces $\overline{\mu}_k$.

This implies backwards stability in the classical numerical linear algebra sense if $M_k(\mu_N; [a, b])$ has polynomial growth in k. If μ_N is sufficiently uniform relative to k

then the orthogonal polynomials through degree k are easily shown to have polynomial growth using standard techniques; see for instance [14, Lemma 6].

Lemma 3.3. *Suppose,* $\operatorname{supp}(\mu) \subseteq [a,b]$ *and, for some* $K > 0$ *and* $k \geq 1$,

$$\mu([x,y]) \geq K, \qquad \forall x, y \in [a,b] : |x-y| \geq (b-a)/(16k^2). \tag{3.1}$$

Then,

$$M_k(\mu; [a,b]) \leq \frac{2}{\sqrt{K}}.$$

In some situations, the condition (3.1) can be verified directly for $\mu = \mu_N$. However, it will typically be easier to assume the Kolmogorov–Smirnov distance $d_{\mathrm{KS}}(\mu_N, \mu_\infty)$ between μ_N and some sufficiently regular measure μ_∞ is small.

Assumption 3.4 *(regularity of* μ_N*). Suppose* μ_∞ *is a measure with support* $[a,b]$ *such that, for some* $L, \gamma > 0$,

$$\mu_\infty([x,y]) \geq L|x-y|^\gamma, \qquad \forall x, y \in [a,b], \tag{3.2}$$

that for some $k \geq 1$,

$$\operatorname{supp}(\mu_N) \subseteq [a - (b-a)/(32k^2), b + (b-a)/(32k^2)],$$

and that for some $\alpha > 0$,

$$d_{\mathrm{KS}}(\mu_N, \mu_\infty) \leq N^{-\alpha}. \tag{3.3}$$

In Section 5 we will discuss several common random matrix ensembles for which Assumption 3.4 is satisfied in a probabilistic sense.

When Assumption 3.4 is satisfied, the following result gives us a bound for $M_k(\mu_N; [a,b])$.

Corollary 3.5. *Given Assumption 3.4, suppose*

$$k \leq \sqrt{\frac{b-a}{32}} \left(\frac{LN^\alpha}{3} \right)^{1/(2\gamma)}.$$

Then, for $[a', b'] := [a - (b-a)/(32k^2), b + (b-a)/(32k^2)]$,

$$M_k(\mu_N; [a', b']) \leq \frac{4}{\sqrt{L}} \left(\frac{32}{b-a} \right)^{\gamma/2} k^\gamma.$$

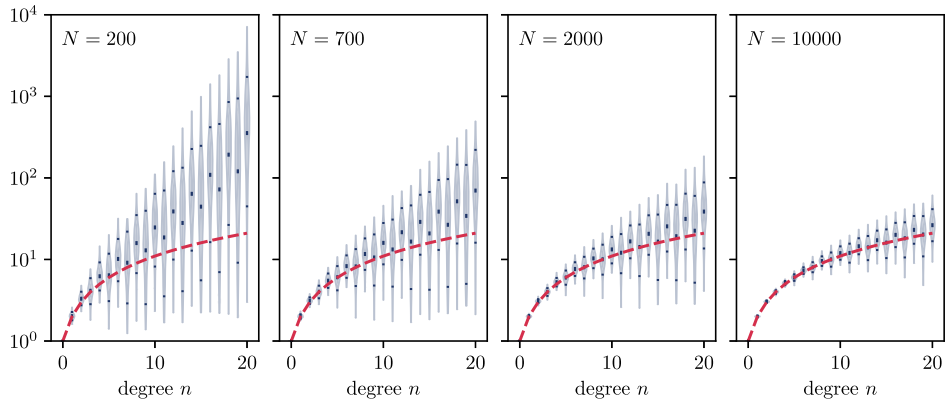


Fig. 3. Maximum value of orthogonal polynomial $p_n(\cdot\,; \mu_N)$ over $[-1, 1]$ and the maximum value in the $N \to \infty$ limit (- - -), where μ_N is drawn from the same random matrix model as in Figs. 1 and 2. For each n, N, the violin plot gives the distribution of $\|p_n(\cdot\,; \mu_N)\|_{[-1,1]}$, with the 5%, 50%, and 95% quantiles marked explicitly. Note that for k growing sufficiently slow with N, the maximum value of p_n has polynomial growth for all $n \le k$.

Fig. 3 shows the growth of the orthogonal polynomials $p_n(\cdot\,;\ \mu_N)$ with n corresponding to the same random matrix model used in other figures. As expected, as N increases, the degree n for which the orthogonal polynomials of μ_N grow like those of μ_∞ increases.

Our forward stability analysis in Section 4 is based on a perturbation to a measure with a sufficiently nice density. Assuming μ_∞ is sufficiently nice, we apply Theorem 3.1 with $\mu = \mu_\infty$. This requires bounding $M_k(\mu_\infty; [a, b])$ and $\mathfrak{m}_k^{\Delta}(\mu_\infty,\ \mu_N; \mu_N)$.

Corollary 3.6. *Given Assumption 3.4, suppose that for some $c > 0$*

$$k \le \left(\frac{b-a}{16}\right)^{\gamma/(4+2\gamma)} \left(\frac{c\sqrt{L}N^{\alpha}}{32}\right)^{1/(2+\gamma)}.$$

Then

$$M_k(\mu_\infty; [a, b]) \le \frac{2}{\sqrt{L}} \left(\frac{16}{b-a}\right)^{\gamma/2} k^{\gamma}, \qquad \mathfrak{m}_k^{\Delta}(\mu_N, \mu_\infty; \mu_\infty) \le c.$$

The proofs of Lemma 3.3 and Corollaries 3.5 and 3.6 are given in Section 7.1. Stronger bounds can be obtained in many situations. The stated bounds are simply meant to give a simple sufficient condition for the orthogonal polynomials to have polynomial growth with respect to k.

3.1. Bounding the modified Chebyshev moments

We will make frequent use of the well-known Chebyshev polynomials of the first and second kinds. These families of polynomials are respectively defined by the recurrences

$$T_0(x) = 1, \qquad T_1(x) = x, \qquad T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \qquad n \geq 2,$$
$$U_0(x) = 1, \qquad U_1(x) = 2x, \qquad U_n(x) = 2xU_{n-1}(x) - U_{n-2}(x), \qquad n \geq 2,$$

and are respectively orthogonal with respect to the measures μ_T and μ_U, each supported on $[-1, 1]$, defined by

$$\mu_T(\mathrm{d}x) := \frac{1}{\pi}\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x, \qquad \mu_U(\mathrm{d}x) := \frac{2}{\pi}\sqrt{1-x^2}\,\mathrm{d}x.$$

The Chebyshev polynomials of the first kind also satisfy the identities

$$T_{2n}(x) = 2T_n(x)^2 - 1, \qquad T_{2n+1}(x) = 2T_n(x)T_{n+1}(x) - x, \qquad n \geq 0. \tag{3.4}$$

To prove Theorem 3.1, it essentially suffices to show that the modified moments of the finite precision and exact arithmetic computations are near. We begin by providing a slightly modified version of [21] for $\mu = \mu_T$ to allow for eigenvalues of $\mathbf{A}$ (and therefore $\overline{\mathbf{T}}_k$) which may be just outside of $[-1, 1]$.

It is well known that $\|T_n\|_{[-1,1]} \leq 1$ and $\|U_n\|_{[-1,1]} \leq n + 1$. Similar bounds hold on a slight extension of $[-1, 1]$.

Lemma 3.7. *For any polynomial p of degree n, with $\eta := 1/(2n^2)$,*

$$\|p\|_{[-1-\eta,1+\eta]} \leq 2\|p\|_{[-1,1]}.$$

This implies a bound for matrix Chebyshev polynomials of $\mathbf{A}$ and $\overline{\mathbf{T}}_k$.

Lemma 3.8. *Suppose that Lanczos is run on $(\mathbf{A}, \mathbf{b})$ for $k \geq 1$ iterations with precision $\epsilon_{\mathrm{lan}} < 1/(5k^2)$ and that $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$. Then, for all $n \leq k$,*

$$\|T_n(\mathbf{A})\|, \|T_n(\overline{\mathbf{T}}_k)\| \leq 2, \qquad \|U_n(\mathbf{A})\|, \|U_n(\overline{\mathbf{T}}_k)\| \leq 2(k+1).$$

Proof. For $k \geq 1$, we have that $(1 + 1/(4k^2))(1 + \epsilon_{\mathrm{lan}}) \leq 1 + 1/(2k^2)$. Thus, since $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$, our assumption on ϵ_{lan} and Definition 2.2 imply $\Lambda(\overline{\mathbf{T}}_k) \subseteq [-1 - 1/(2k^2), 1 + 1/(2k^2)]$. The result follows by applying Lemma 3.7 and the fact $\|U_n\|_{[-1,1]} \leq n+1$, since the operator norm of a matrix function of a symmetric matrix is simply the maximum value of the function's absolute value evaluated at the eigenvalues of that matrix. □

We will also use the following fact about perturbed Chebyshev recurrences. This is a special case of a more general formula involving the associated polynomials of some family of orthogonal polynomials.

Lemma 3.9. *Suppose that*

$$d_0 = 0, \qquad d_1 = f_0, \qquad d_n(x) = 2xd_{n-1}(x) - d_{n-2}(x) + 2f_{n-1}, \qquad n \geq 2.$$

Then, introducing the notation $U_{-1}(x) = 0$,

$$d_n(x) = U_{n-1}(x)f_0 + 2\sum_{i=2}^{n} U_{n-i}(x)f_{i-1}, \qquad n \geq 0.$$

Lemmas 3.7 and 3.9 are proved in Section 7.2.

The next several results and the accompanying proofs follow [21] closely. We include them so that our analysis is self-contained and in order to compute explicit constants. In our proofs, for notational brevity, we define and use the vectors:

$$\mathbf{t}_n := T_n(\mathbf{A})\mathbf{b}, \qquad \overline{\mathbf{t}}_n := T_n(\overline{\mathbf{T}}_k)\mathbf{e}_0, \qquad \mathbf{d}_n := \mathbf{t}_n - \overline{\mathbf{Q}}_k\overline{\mathbf{t}}_n, \qquad \mathbf{r}_n := \mathbf{R}_k\overline{\mathbf{t}}_n. \tag{3.5}$$

The first technical lemma we need is a bound on how well polynomials in $\mathbf{A}$ applied to $\mathbf{b}$ are approximated by the Lanczos quantities. To the best of our knowledge, a similar bound first appeared in [7] to analyze the behavior of the well-known Lanczos method for matrix function approximation; see also [27].

Lemma 3.10. *Suppose that Lanczos is run on* $(\mathbf{A}, \mathbf{b})$ *for* $k > 1$ *iterations with precision* $\epsilon_{\text{lan}} < 1/(5k^2)$ *and that* $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$. *Then, for all* $n \leq k-1$,

$$\|T_n(\mathbf{A})\mathbf{b} - \overline{\mathbf{Q}}_k T_n(\overline{\mathbf{T}}_k)\mathbf{e}_0\| \leq 9k^2\epsilon_{\text{lan}}.$$

Proof. Since $k > 1$, using the notation in (3.5) and recalling the perturbed recurrence (2.6), we have

$$\mathbf{d}_0 = \mathbf{b} - \overline{\mathbf{Q}}_k\mathbf{e}_0 = \mathbf{0}, \qquad \mathbf{d}_1 = \mathbf{A}\mathbf{b} - \overline{\mathbf{Q}}_k\overline{\mathbf{T}}_k\mathbf{e}_0 = (\beta_k\mathbf{q}_{k-1}\mathbf{e}_{k-1}^{\mathsf{T}} + \mathbf{F}_k)\mathbf{e}_0 = \mathbf{F}_k\overline{\mathbf{t}}_0.$$

For $n = 2, \ldots, k-1$, we can use the definitions of $\mathbf{t}_n$ and $\overline{\mathbf{t}}_n$, the definition of the Chebyshev polynomials, and the perturbed recurrence (2.6) to write

$$\begin{aligned}
\mathbf{d}_n &= (2\mathbf{A}\mathbf{t}_{n-1} - \mathbf{t}_{n-2}) - (2\overline{\mathbf{Q}}_k\overline{\mathbf{T}}_k\overline{\mathbf{t}}_{n-1} - \overline{\mathbf{Q}}_k\overline{\mathbf{t}}_{n-2}) \\
&= 2(\mathbf{A}\mathbf{t}_{n-1} - (\mathbf{A}\overline{\mathbf{Q}}_k - \beta_k\mathbf{q}_{k-1}\mathbf{e}_{k-1}^{\mathsf{T}} - \mathbf{F}_k)\overline{\mathbf{t}}_{n-1}) - (\mathbf{t}_{n-2} - \overline{\mathbf{Q}}_k\overline{\mathbf{t}}_{n-2}) \\
&= 2(\mathbf{A}\mathbf{t}_{n-1} - (\mathbf{A}\overline{\mathbf{Q}}_k\overline{\mathbf{t}}_{n-1} - \beta_k\mathbf{q}_{k-1}\mathbf{e}_{k-1}^{\mathsf{T}}\overline{\mathbf{t}}_{n-1} - \mathbf{F}_k\overline{\mathbf{t}}_{n-1})) - \mathbf{d}_{n-2}
\end{aligned}$$

Note that $(\overline{\mathbf{T}}_k)^i$ has half bandwidth i, so $(\overline{\mathbf{T}}_k)^i$ is zero in the bottom left entry provided $i < k-1$. Since T_i is a degree i polynomial, this implies that $\mathbf{e}_{k-1}^{\mathsf{T}}\overline{\mathbf{t}}_i = \mathbf{e}_{k-1}^{\mathsf{T}}T_i(\overline{\mathbf{T}}_k)\mathbf{e}_0 = 0$ for any $i < k-1$. Since $n < k$, applying this with $i = n-1$ we find

$$\mathbf{d}_n = 2\mathbf{A}\mathbf{d}_{n-1} - \mathbf{d}_{n-2} + 2\mathbf{F}_k\overline{\mathbf{t}}_{n-1}.$$

Lemma 3.9 with $x \to \mathbf{A}$, $d_n(x) \to \mathbf{d}_n$, and $f_n \to \mathbf{F}_k\overline{\mathbf{t}}_n$ allows us to obtain an explicit expression

$$\mathbf{d}_n = U_{n-1}(\mathbf{A})\mathbf{F}_k\bar{\mathbf{t}}_0 + 2\sum_{i=2}^{n} U_{n-i}(\mathbf{A})\mathbf{F}_k\bar{\mathbf{t}}_{i-1}. \tag{3.6}$$

Since $\epsilon_{\text{lan}} < 1/(5k^2)$ and $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$, if $\ell \leq k-1$, then Lemma 3.8 gives the bounds

$$\|U_\ell(\mathbf{A})\| \leq 2k, \qquad \|\bar{\mathbf{t}}_\ell\| = \|T_\ell(\overline{\mathbf{T}}_k)\| \leq 2.$$

Using Definition 2.2 and the assumption $k > 1$,

$$\|\mathbf{F}_k\| \leq \|\mathbf{A}\|\epsilon_{\text{lan}} \leq (1 + 1/(4k^2))\epsilon_{\text{lan}} \leq (17/16)\epsilon_{\text{lan}} < (9/8)\epsilon_{\text{lan}}.$$

Finally, we apply the triangle inequality to (3.6), double the first term for convenience, apply the above bounds, and use the fact $n \leq k$ to obtain the bound

$$\|\mathbf{d}_n\| \leq 2\sum_{i=1}^{n} \|U_{n-i}(\mathbf{A})\|\|\mathbf{F}_k\|\|\bar{\mathbf{t}}_{i-1}\| \leq 2n(2k)((9/8)\epsilon_{\text{lan}})(2) < 9k^2\epsilon_{\text{lan}}. \quad \square$$

If $\overline{\mathbf{Q}}_k$ had nearly orthonormal columns, we could use (3.4) to upgrade Lemma 3.10 to a bound on the modified moments produced by the Lanczos algorithm. However, since we do not have such a guarantee, we require a bit more work. We begin with a lemma akin to [21, Lemma 1].

Lemma 3.11. *Suppose that Lanczos is run on* $(\mathbf{A}, \mathbf{b})$ *for* $k > 1$ *iterations with precision* $\epsilon_{\text{lan}} < 1/(5k^2)$ *and that* $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$. *Then, for all* $n \leq k-1$,

$$\|\mathbf{R}_k T_n(\overline{\mathbf{T}}_k)\mathbf{e}_0\| \leq 9\|\mathbf{A}\|k^2\epsilon_{\text{lan}}.$$

Proof. Since $\mathbf{R}_k$ is strictly upper triangular, again using the notation in (3.5) and recalling the perturbed recurrence (2.8), we have

$$\mathbf{r}_0 = \mathbf{0}, \qquad \mathbf{r}_1 = \mathbf{R}_k\overline{\mathbf{T}}_k\mathbf{e}_0 = (\overline{\mathbf{T}}_k\mathbf{R}_k - \overline{\beta}_{k-1}\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{q}}_k\mathbf{e}_{k-1}^{\mathsf{T}} - \mathbf{H}_k)\mathbf{e}_0 = -\mathbf{H}_k\bar{\mathbf{t}}_0.$$

Analogous to the recurrence used in the previous proof, but now using the perturbed recurrence (2.8), for $n = 2, \ldots, k-1$ the $\mathbf{r}_n$ satisfy the perturbed three-term recurrence

$$\begin{aligned}
\mathbf{r}_n &= 2\mathbf{R}_k\overline{\mathbf{T}}_k\bar{\mathbf{t}}_{n-1} - \mathbf{R}_k\bar{\mathbf{t}}_{n-2} \\
&= 2(\overline{\mathbf{T}}_k\mathbf{R}_k - \beta_{k-1}\overline{\mathbf{Q}}_k^{\mathsf{T}}\mathbf{q}_k\mathbf{e}_{k-1}^{\mathsf{T}} - \mathbf{H}_k)\bar{\mathbf{t}}_{n-1} - \mathbf{r}_{n-2} \\
&= 2(\overline{\mathbf{T}}_k\mathbf{R}_k\bar{\mathbf{t}}_{n-1} - \beta_{k-1}\overline{\mathbf{Q}}_k^{\mathsf{T}}\mathbf{q}_k\mathbf{e}_{k-1}^{\mathsf{T}}\bar{\mathbf{t}}_{n-1} - \mathbf{H}_k\bar{\mathbf{t}}_{n-1}) - \mathbf{r}_{n-2} \\
&= 2\overline{\mathbf{T}}_k\mathbf{r}_{n-1} - \mathbf{r}_{n-2} - 2\mathbf{H}_k\bar{\mathbf{t}}_{n-1}.
\end{aligned}$$

As above, using Lemma 3.9 with $x \to \overline{\mathbf{T}}_k$, $d_n \to \mathbf{r}_n$, and $f_n \to -\mathbf{H}_k\bar{\mathbf{t}}_n$,

$$\mathbf{r}_n = -U_{n-1}(\overline{\mathbf{T}}_k)\mathbf{H}_k\overline{\mathbf{t}}_0 - 2\sum_{i=2}^{n} U_{n-i}(\overline{\mathbf{T}}_k)\mathbf{H}_k\overline{\mathbf{t}}_{i-1}.$$

As before, for $\ell \leq k-1$, Lemma 3.8 gives bounds

$$\|U_\ell(\overline{\mathbf{T}}_k)\| \leq 2k, \qquad \|\overline{\mathbf{t}}_\ell\| = \|T_\ell(\overline{\mathbf{T}}_k)\| \leq 2,$$

and Definition 2.2 and the assumption $k > 1$ give the bound

$$\|\mathbf{H}_k\| \leq \|\mathbf{A}\|\epsilon_{\text{lan}} \leq (1 + 1/(4k^2))\epsilon_{\text{lan}} \leq (17/16)\epsilon_{\text{lan}} < (9/8)\epsilon_{\text{lan}}.$$

We therefore obtain

$$\|\mathbf{r}_n\| \leq 2\sum_{i=1}^{n} \|U_{n-i}(\overline{\mathbf{T}}_k)\|\|\mathbf{H}_k\|\|\overline{\mathbf{t}}_{i-1}\| \leq 2n(2k)((9/8)\epsilon_{\text{lan}})(2) < 9k^2\epsilon_{\text{lan}}. \quad \square$$

We are now prepared to apply (3.4) to bound the modified Chebyshev moments. For clarity, and following [21, Lemmas 2 and 3], we split this into a few steps. Note that the maximal degree of the matrix-polynomials in the quadratic forms we analyze are $2k-2$. Owing to the fact that $\overline{\mathbf{T}}_k$ is like a Jacobi matrix, one might expect the maximal degree should be $2k-1$, and indeed, in [21] a similar results for polynomials up to degree $2k-1$ is proved. This is not needed for our analysis.

Lemma 3.12. *Suppose that Lanczos is run on* $(\mathbf{A}, \mathbf{b})$ *for* $k > 1$ *iterations with precision* $\epsilon_{\text{lan}} < 1/(5k^2)$ *and that* $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$. *Then, for all* $m, n \leq k-1$,

$$|\mathbf{b}^{\mathsf{T}} T_m(\overline{\mathbf{T}}_k)\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{Q}}_k T_n(\overline{\mathbf{T}}_k)\mathbf{b} - \mathbf{e}_0^{\mathsf{T}} T_m(\overline{\mathbf{T}}_k)T_n(\overline{\mathbf{T}}_k)\mathbf{e}_0| \leq 37k^2\epsilon_{\text{lan}}.$$

Proof. Using the notation in (3.5) and the definitions of $\mathbf{R}_k$ and $\mathbf{D}_k$,

$$\overline{\mathbf{t}}_m^{\mathsf{T}}\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{Q}}_k\overline{\mathbf{t}}_n = \overline{\mathbf{t}}_m^{\mathsf{T}}(\mathbf{R}_k + \mathbf{R}_k{}^{\mathsf{T}} + \mathbf{I} + (\mathbf{D}_k - \mathbf{I}))\overline{\mathbf{t}}_n.$$

From Definition 2.2 we have $\|\mathbf{D}_k - \mathbf{I}\| \leq \epsilon_{\text{lan}}$. By assumption, $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$ and $\epsilon_{\text{lan}} < 1/(5k^2)$, so for all $\ell \leq k-1$, Lemmas 3.8 and 3.11 respectively give bounds

$$\|\overline{\mathbf{t}}_\ell\| = \|T_\ell(\overline{\mathbf{T}}_k)\| \leq 2, \qquad \|\mathbf{R}_k\overline{\mathbf{t}}_\ell\| = \|\mathbf{R}_k T_\ell(\overline{\mathbf{T}}_k)\mathbf{e}_0\| \leq 9k^2\epsilon_{\text{lan}}.$$

Combining these, we find,

$$\begin{aligned}|\overline{\mathbf{t}}_m\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{Q}}_k\overline{\mathbf{t}}_n - \overline{\mathbf{t}}_m^{\mathsf{T}}\overline{\mathbf{t}}_n| &\leq \|\overline{\mathbf{t}}_m\|\|\mathbf{R}_k\overline{\mathbf{t}}_n\| + \|\overline{\mathbf{t}}_n\|\|\mathbf{R}_k\overline{\mathbf{t}}_m\| + \|\mathbf{D}_k - \mathbf{I}\|\|\overline{\mathbf{t}}_m\|\|\overline{\mathbf{t}}_n\| \\ &\leq 2(9k^2\epsilon_{\text{lan}}) + 2(9k^2\epsilon_{\text{lan}}) + \epsilon_{\text{lan}}(2)(2) \leq (36k^2+4)\epsilon_{\text{lan}}.\end{aligned}$$

Since $k > 1$, $36k^2 + 4 \leq 37k^2$. $\square$

Lemma 3.13. *Suppose that Lanczos is run on* $(\mathbf{A}, \mathbf{b})$ *for* $k > 1$ *iterations with precision* $\epsilon_{\text{lan}} < 1/(5k^2)$ *and that* $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$. *Then, for all* $m, n \leq k-1$,

$$|\mathbf{b}^{\mathsf{T}} T_m(\mathbf{A}) T_n(\mathbf{A}) \mathbf{b} - \mathbf{e}_0^{\mathsf{T}} T_m(\overline{\mathbf{T}}_k) T_n(\overline{\mathbf{T}}_k) \mathbf{e}_0| \leq 127k^2 \epsilon_{\text{lan}}.$$

Proof. Using the notation in (3.5),

$$\mathbf{t}_m^{\mathsf{T}} \mathbf{t}_n = (\mathbf{d}_m + \overline{\mathbf{Q}}_k \bar{\mathbf{t}}_m)^{\mathsf{T}} (\mathbf{d}_n + \overline{\mathbf{Q}}_k \bar{\mathbf{t}}_n) = \mathbf{d}_m^{\mathsf{T}} \mathbf{d}_n + \mathbf{d}_m^{\mathsf{T}} \overline{\mathbf{Q}}_k \bar{\mathbf{t}}_n + \bar{\mathbf{t}}_m^{\mathsf{T}} \overline{\mathbf{Q}}_k^{\mathsf{T}} \mathbf{d}_n + \bar{\mathbf{t}}_m^{\mathsf{T}} \overline{\mathbf{Q}}_k^{\mathsf{T}} \overline{\mathbf{Q}}_k \bar{\mathbf{t}}_n.$$

Thus, applying the triangle inequality and submultiplicativity of the operator norm,

$$|\mathbf{t}_m^{\mathsf{T}} \mathbf{t}_n - \bar{\mathbf{t}}_m^{\mathsf{T}} \bar{\mathbf{t}}_n| \leq |\bar{\mathbf{t}}_m \overline{\mathbf{Q}}_k^{\mathsf{T}} \overline{\mathbf{Q}}_k \bar{\mathbf{t}}_n - \bar{\mathbf{t}}_m^{\mathsf{T}} \bar{\mathbf{t}}_n| + \|\mathbf{d}_m\| \|\overline{\mathbf{Q}}_k \bar{\mathbf{t}}_n\| + \|\mathbf{d}_n\| \|\overline{\mathbf{Q}}_k \bar{\mathbf{t}}_m\| + \|\mathbf{d}_m\| \|\mathbf{d}_n\|. \tag{3.7}$$

By assumption, $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$ and $\epsilon_{\text{lan}} < 1/(5k^2)$, so for all $\ell \leq k-1$, Lemmas 3.8 and 3.10 respectively give bounds

$$\|\bar{\mathbf{t}}_\ell\| = \|T_\ell(\overline{\mathbf{T}}_k)\| \leq 2 \qquad \|\mathbf{d}_\ell\| = \|T_n(\mathbf{A})\mathbf{b} - \overline{\mathbf{Q}}_k T_n(\overline{\mathbf{T}}_k)\mathbf{e}_0\| \leq 9k^2 \epsilon_{\text{lan}}.$$

This implies $9k^2 \epsilon_{\text{lan}} < 2$, so we find that

$$\|\overline{\mathbf{Q}}_k \bar{\mathbf{t}}_\ell\| = \|\mathbf{d}_\ell + \mathbf{t}_\ell\| \leq \|\mathbf{d}_\ell\| + \|\mathbf{t}_\ell\| < 2 + 2 = 4.$$

Under these same assumptions, Lemma 3.12 gives a bound

$$|\bar{\mathbf{t}}_m \overline{\mathbf{Q}}_k^{\mathsf{T}} \overline{\mathbf{Q}}_k \bar{\mathbf{t}}_n - \bar{\mathbf{t}}_m^{\mathsf{T}} \bar{\mathbf{t}}_n| \leq 37k^2 \epsilon_{\text{lan}}.$$

Plugging the above bounds into (3.7) we find

$$|\mathbf{t}_m^{\mathsf{T}} \mathbf{t}_n - \bar{\mathbf{t}}_m^{\mathsf{T}} \bar{\mathbf{t}}_n| \leq 37k^2 \epsilon_{\text{lan}} + 9k^2 \epsilon_{\text{lan}}(4) + 9k^2 \epsilon_{\text{lan}}(4) + (9k^2 \epsilon_{\text{lan}})(2) \leq 127k^2 \epsilon_{\text{lan}}. \quad \square$$

A bound for the modified moments with respect to the Chebyshev polynomials, akin to [21, Theorem 1], is now immediate.

Theorem 3.14. *Suppose that Lanczos is run on* $(\mathbf{A}, \mathbf{b})$ *for* $k > 1$ *iterations with precision* $\epsilon_{\text{lan}} < 1/(5k^2)$ *and that* $\|\mathbf{A}\| \leq 1 + 1/(4k^2)$. *Then, for all* $n \leq 2k-2$,

$$\left| \int T_n(x) \mu_N(\mathrm{d}x) - \int T_n(x) \overline{\mu}_k(\mathrm{d}x) \right| \leq 381k^2 \epsilon_{\text{lan}}.$$

Proof. By definition,

$$\left| \int T_n(x) \mu_N(\mathrm{d}x) - \int T_n(x) \overline{\mu}_k(\mathrm{d}x) \right| = |\mathbf{b}^{\mathsf{T}} T_n(\mathbf{A}) \mathbf{b} - \mathbf{e}_0^{\mathsf{T}} T_n(\overline{\mathbf{T}}_k) \mathbf{e}_0|.$$

First, suppose $n = 2i$ for $i \leq k-1$. As noted in (3.4), $T_n(x) = 2T_i^2(x) - 1$. By assumption $\mathbf{b}^\mathsf{T}\mathbf{b} = \mathbf{e}_0^\mathsf{T}\mathbf{e}_0 = 1$. Therefore, we can apply Lemma 3.13 to bound

$$|\mathbf{b}^\mathsf{T} T_n(\mathbf{A})\mathbf{b} - \mathbf{e}_0^\mathsf{T} T_n(\overline{\mathbf{T}}_k)\mathbf{e}_0| = |2\mathbf{t}_i^\mathsf{T}\mathbf{t}_i - 2\overline{\mathbf{t}}_i^\mathsf{T}\overline{\mathbf{t}}_i| \leq 2(127k^2\epsilon_{\text{lan}}) = 254k^2\epsilon_{\text{lan}}.$$

Now, assume $n = 2i+1$, $i \leq k-2$. Then $T_n(x) = 2T_i(x)T_{i+1}(x) - x$. Then, since $T_1(x) = x$, again using Lemma 3.13,

$$\begin{aligned}|\mathbf{b}^\mathsf{T} T_n(\mathbf{A})\mathbf{b} - \mathbf{e}_0^\mathsf{T} T_n(\overline{\mathbf{T}}_k)\mathbf{e}_0| &= |(2\mathbf{t}_i^\mathsf{T}\mathbf{t}_{i+1} - \mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{b}) - (2\overline{\mathbf{t}}_i^\mathsf{T}\overline{\mathbf{t}}_{i+1} - \mathbf{e}_0^\mathsf{T}\overline{\mathbf{T}}_k\mathbf{e}_0)| \\ &\leq 2|\mathbf{t}_i^\mathsf{T}\mathbf{t}_{i+1} - \overline{\mathbf{t}}_i^\mathsf{T}\overline{\mathbf{t}}_{i+1}| + |\mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{b} - \mathbf{e}_0^\mathsf{T}\overline{\mathbf{T}}_k\mathbf{e}_0| \\ &\leq 2(127k^2\epsilon_{\text{lan}}) + 127k^2\epsilon_{\text{lan}} = 381k^2\epsilon_{\text{lan}}.\end{aligned}$$

The result follows. □

3.2. General modified moments via a change of basis

When the orthogonal polynomials $p_n(\cdot\,;\mu)$ have reasonable Chebyshev series, then a statement similar to Theorem 3.14 holds for the moments with respect to μ.

Corollary 3.15 *(Stability of moments wrt. μ). Let μ_N be the VESD for $(\mathbf{A}, \mathbf{b})$ and μ a unit-mass measure with support contained in $[-1, 1]$.*

Then, there exist absolute constants C, D such that, whenever Lanczos is run on $(\mathbf{A}, \mathbf{b})$ for $k \geq 1$ iterations with precision $\epsilon_{\text{lan}} < 1/(Ck^2)$ to produce $\overline{\mu}_k$ and $\operatorname{supp}(\mu_N) \subseteq [-1-\eta, 1+\eta]$ where $\eta < 1/(16k^2)$, then $\mathfrak{m}_k^{\Delta}(\mu_N, \overline{\mu}_k; \mu) \leq DM_k(\mu; [-1,1])k^3\epsilon_{\text{lan}}$.

Proof. Note that $\int T_n(x)^2\mu_T(\mathrm{d}x) = 1/2$ for $n \geq 1$. We can therefore decompose $p_n(x;\mu)$, $n \geq 0$ into Chebyshev polynomials of the first kind by

$$p_n(x;\mu) = c_{n,0}T_0(x) + c_{n,1}T_1(x) + \cdots + c_{n,n}T_n(x),$$

where the coefficients are obtained by

$$c_{n,0} := \int p_n(x;\mu)\mu_T(\mathrm{d}x), \qquad c_{n,i} := 2\int p_n(x;\mu)T_i(x)\mu_T(\mathrm{d}x), \qquad 1 \leq i \leq n.$$

Note that for all $n \leq 2k-1$ and $i \leq n$, since $\|T_i\|_{[-1,1]} \leq 1$,

$$|c_{n,i}| \leq 2\int |p_n(x;\mu)|\,|T_i(x)|\mu_T(\mathrm{d}x) \leq 2M_k(\mu;[-1,1]). \tag{3.8}$$

Assuming $\epsilon_{\text{lan}} < 1/(5(k+1)^2)$ and $\operatorname{supp}(\mu_N) \subseteq [-1-\eta, 1+\eta]$ where $\eta < 1/(16k^2) \leq 1/(4(k+1)^2)$ we can apply Theorem 3.14 and (3.8) to get the bound, for $n \leq 2k-1$,

$$\begin{aligned}|\mathfrak{m}_n(\mu_N;\mu)-\mathfrak{m}_n(\overline{\mu}_k;\mu)| &= \left|\sum_{i=1}^n c_{n,i}\left(\int T_i(x)\mu_N(\mathrm{d}x)-\int T_i(x)\overline{\mu}_k(\mathrm{d}x)\right)\right| \\ &\le \sum_{i=1}^n |c_{n,i}|\left|\int T_i(x)\mu_N(\mathrm{d}x)-\int T_i(x)\overline{\mu}_k(\mathrm{d}x)\right| \\ &\le 4kM_k(\mu;[-1,1])381(k+1)^2\epsilon_{\mathrm{lan}}.\end{aligned}$$

Finally, since $k\ge 1$, $1/(5(k+1)^2)\ge 1/(20k^2)\mathrm{W}$ and $4(381)k(k+1)^2\le 6096k^3$. Setting $C=20$ and $D=6096$ establishes the result. □

Clearly a better bound could be obtained directly from the coefficients[3] $c_{n,i}$ rather than in terms of the maximum value of the $p_n(\cdot;\mu)$. However, we are more interested in the existence of bounds which deteriorate with polynomials of k rather than the precise dependencies on k, and the present approach results in slightly simpler statements and proofs.

3.3. Proof of backwards stability

We are now prepared to prove Theorem 3.1. The approach is straightforward: transform $[a,b]$ to $[-1,1]$ and then apply Corollary 3.15 to get a bound for the moments. This will give us a bound on the size of h in (2.2).

Proof of Theorem 3.1. If $\|h\|_{[a,b]}<1$, then as described in Section 2, μ_* is a well-defined positive measure whose moments agree $\overline{\mu}_k$ through degree $2k-1$.

Define

$$\hat{\mathbf{A}}:=\frac{2}{b-a}\mathbf{A}-\frac{b+a}{b-a}\mathbf{I},\qquad \hat{\overline{\alpha}}_i:=\frac{2}{b-a}\overline{\alpha}_i-\frac{b+a}{b-a},$$
$$\hat{\overline{\beta}}_i:=\frac{2}{b-a}\overline{\beta}_i,\qquad \hat{\mathbf{F}}_k:=\frac{2}{b-a}\mathbf{F}_k,\qquad \hat{\mathbf{H}}_k:=\frac{2}{b-a}\mathbf{H}_k,\qquad \hat{\eta}_k:=\frac{2}{b-a}\eta_k.$$

Then,

$$\begin{aligned}\hat{\mathbf{A}}\overline{\mathbf{Q}}_k &= \overline{\mathbf{Q}}_k\hat{\overline{\mathbf{T}}}_k+\hat{\overline{\beta}}_{k-1}\overline{\mathbf{q}}_k\mathbf{e}_{k-1}^{\mathsf{T}}+\hat{\mathbf{F}}_k,\\ \hat{\overline{\mathbf{T}}}_k\mathbf{R}_k &= \mathbf{R}_k\hat{\overline{\mathbf{T}}}_k+\hat{\overline{\beta}}_{k-1}\overline{\mathbf{Q}}_k^{\mathsf{T}}\overline{\mathbf{q}}_k\mathbf{e}_{k-1}^{\mathsf{T}}+\hat{\mathbf{H}}_k,\\ \Lambda(\hat{\overline{\mathbf{T}}}_k) &\subseteq [\lambda_{\min}(\hat{\mathbf{A}})-\hat{\eta}_k,\lambda_{\max}(\hat{\mathbf{A}})+\hat{\eta}_k]\end{aligned}$$

where

$$\|\hat{\mathbf{F}}_k\|\le\frac{2\|\mathbf{A}\|}{b-a}\epsilon_{\mathrm{lan}},\qquad \|\mathbf{D}_k-\mathbf{I}\|\le\epsilon_{\mathrm{lan}},\qquad \|\hat{\mathbf{H}}_k\|\le\frac{2\|\mathbf{A}\|}{b-a}\epsilon_{\mathrm{lan}},\qquad \hat{\eta}_k\le\frac{2\|\mathbf{A}\|}{b-a}\epsilon_{\mathrm{lan}}.$$

[3] The coefficients $c_{n,i}$ define a so-called connection coefficient matrix [46]. Since we are always interested in an expansion in the Chebyshev polynomials of the first kind, we do not introduce this concept in generality.

Thus, $(\hat{\overline{\mathbf{T}}}_k, \overline{\mathbf{Q}}_k)$ can be viewed as the output of the Lanczos algorithm run on $(\hat{\mathbf{A}}, \mathbf{b})$ with precision

$$\hat{\epsilon}_{\text{lan}} := \max\left\{\frac{2\|\mathbf{A}\|}{b-a}, 1\right\}\epsilon_{\text{lan}} = \sigma\epsilon_{\text{lan}}.$$

Define

$$t(x) = \frac{2}{b-a}\left(x - \frac{b+a}{2}\right), \qquad t([a,b]) = [-1,1],$$

and let $\hat{\mu}_N$ and $\hat{\mu}$ be the pushforward measures of μ_N and μ, respectively, under t. That is, for any measurable function f,

$$\int f(x)\mathrm{d}\mu_N(x) = \int f(t(x))\mathrm{d}\hat{\mu}_N(x), \qquad \int f(x)\mathrm{d}\mu(x) = \int f(t(x))\mathrm{d}\hat{\mu}(x).$$

This implies modified moments of $\hat{\mu}_N$ with respect to the orthogonal polynomials of $\hat{\mu}$ are the same as those of μ_N with respect to μ. Indeed,

$$p_n(t(x); \hat{\mu}) = p_n(x; \mu)$$

so

$$\mathfrak{m}_n(\mu_N; \mu) = \int p_n(x;\mu)\mathrm{d}\mu_N = \int p_n(t(x);\hat{\mu})\mathrm{d}\mu_N = \int p_n(x;\hat{\mu})\mathrm{d}\hat{\mu}_N = \mathfrak{m}_n(\hat{\mu}_N;\hat{\mu}).$$

Moreover,

$$\operatorname{supp}(\hat{\mu}_N) \subseteq [-1 - 1/(16k^2), 1 + 1/(16k^2)].$$

The assumption on ϵ_{lan} ensures $\hat{\epsilon}_{\text{lan}} < 1/(Ck^2)$, so Corollary 3.15 gives a bound

$$\mathfrak{m}_n^{\Delta}(\overline{\mu}_k, \mu_N; \mu) = \mathfrak{m}_n^{\Delta}(\hat{\overline{\mu}}_k, \hat{\mu}_N; \hat{\mu}) \le DM_k(\mu; [a,b])k^3\hat{\epsilon}_{\text{lan}}.$$

Since the moments $\mathfrak{m}_n(\mu_*; \mu) = \mathfrak{m}_n(\overline{\mu}_k; \mu)$ for $n \le 2k-1$, we get Theorem 3.1(a). Using the triangle inequality we also have

$$\mathfrak{m}_n^{\Delta}(\overline{\mu}_k, \mu; \mu) \le \mathfrak{m}_n^{\Delta}(\overline{\mu}_k, \mu_N; \mu) + \mathfrak{m}_n^{\Delta}(\mu_N, \mu; \mu).$$

From (2.5) we have $\|h\|_{[a,b]} \le 2k\mathfrak{m}_k^{\Delta}(\overline{\mu}_k, \mu; \mu)M_k(\mu; [a,b])$. Thus,

$$\|h\|_{[a,b]} \le 2kM_k(\mu; [a,b])\big(\mathfrak{m}_n^{\Delta}(\overline{\mu}_k, \mu_N; \mu) + \mathfrak{m}_n^{\Delta}(\mu_N, \mu; \mu)\big).$$

Replacing $\hat{\epsilon}_{\text{lan}}$ with $\sigma\epsilon_{\text{lan}}$ gives Theorem 3.1(b). □

4. Perturbation theory for recurrence coefficients

We will now show that $\overline{\mathbf{T}}_k$ is close to $\mathbf{T}_k$, at least when μ_N is near a sufficiently nice measure μ_∞. Our forward stability results, which are stated at the end of the section, are derived by applying Theorem 3.1 with $\mu = \mu_\infty$ and then using a Riemann–Hilbert approach to analyze perturbations to the orthogonal polynomials of μ_∞.

The conditioning of the map from the modified moments of a measure to the corresponding recurrence coefficients was studied in [14]; see also [15, Section 2.1.6]. However, it is not immediately clear how to apply the formulas and bounds to the situation in which k grows as ϵ_{lan} tends to zero.

To study such a scaling, we use an alternate approach. As a matter of technical convenience, we assume that μ has a density with square root behavior at the edges.

Assumption 4.1. *Suppose that* $\operatorname{supp}(\mu) = [a, b]$ *and that* μ *has density*

$$\rho(z) = g(z)\sqrt{(b-z)(z-a)},$$

where g *is positive on* $[a, b]$ *and analytic on an open set that contains* $[a, b]$.

Note that through the obvious affine transformation, we now assume, without loss of generality, that $\operatorname{supp}(\mu) = [-1, 1]$. We show that when μ is perturbed slightly, the recurrence coefficients for the perturbed measure are near those of μ. This culminates in Theorem 4.7, a forward stability result for the Lanczos algorithm given at the end of the section.

Suppose μ satisfies Assumption 4.1, $[a, b] = [-1, 1]$ and let $\pi_n(z; \mu)$ denote the n-th monic orthogonal polynomial. Consider the matrix-valued function

$$\mathbf{Y}_n(z; \mu) = \begin{bmatrix} \pi_n(z; \mu) & c_n(z; \mu) \\ \gamma_{n-1}(\mu)\pi_{n-1}(z; \mu) & \gamma_{n-1}(\mu)c_{n-1}(z; \mu) \end{bmatrix}, \qquad n \geq 1,$$

where

$$c_n(z; \mu) := \frac{1}{2\pi \mathrm{i}} \int \frac{\pi_n(x; \mu)}{x - z} \mu(\mathrm{d}x), \qquad \gamma_{n-1}(\mu) := -2\pi \mathrm{i} \| \pi_n(\cdot\,; \mu)\|_{L^2(\mu)}^{-2}.$$

For a function y, analytic in $\mathbb{C} \setminus [-1, 1]$, define the boundary values

$$y^{\pm}(z) := \lim_{\epsilon \downarrow 0} y(z \pm \mathrm{i}\epsilon), \quad z \in [-1, 1],$$

provided this limit exists.

With this notation in mind, $\mathbf{Y}_n(z; \mu)$ has the following properties:

- $\mathbf{Y}_n^+(z; \mu) = \mathbf{Y}_n^-(z; \mu) \begin{bmatrix} 1 & \rho(z) \\ 0 & 1 \end{bmatrix}, \qquad z \in [-1, 1],$

- $\mathbf{Y}_n(z;\mu)z^{-n\boldsymbol{\sigma}_3} = \mathbf{I} + O(z^{-1}), \quad z \to \infty, \quad \boldsymbol{\sigma}_3 := \operatorname{diag}(1,-1),$
- $\det \mathbf{Y}_n(z;\mu) = 1$, and
- $\mathbf{Y}_n(z;\mu)$ is analytic in $\mathbb{C} \setminus [-1,1]$.

This is the so-called Fokas–Its–Kitaev Riemann–Hilbert problem [13]: The problem of finding the sectionally analytic function $\mathbf{Y}_n$ from the stated conditions. References [6, 23] provide a comprehensive introduction. It has been used in many contexts, both computationally and asymptotically, see [6,30,44], for example. But it can also be used for perturbation theory [9]. The Riemann–Hilbert representation play a role similar to the contour integral representation of classical orthogonal polynomials and allows one to estimate quantities related to the polynomials (e.g., recurrence coefficients) under consideration. And even in the classical cases, it can prove to be a more powerful tool. Below we will use Riemann–Hilbert theory to estimate $Y_n(z;\mu)$ which, via $Y_n^1(\mu)$ below, gives estimates on the recurrence coefficients for orthogonal polynomials, i.e., the output of the Lanczos iteration.

Note that the orthonormal polynomials p_n satisfy

$$p_n(z;\mu) = \frac{\pi_n(z;\mu)}{\|\pi_n(\cdot\,;\mu)\|_{L^2(\mu)}}, \qquad \|\pi_n(\cdot\,;\mu)\|^2_{L^2(\mu)} = \int \pi_n(x;\mu)^2 \mu(\mathrm{d}x).$$

The recurrence coefficients for the orthonormal polynomials are $(\alpha_j(\mu))_{j\geq 0}$, $(\beta_j(\mu))_{j\geq 0}$ such that, with $\beta_{-1}(\mu) := 0 =: p_{-1}(x;\mu)$,

$$xp_n(x;\mu) = \alpha_n(\mu)p_n(x;\mu) + \beta_n(\mu)p_{n+1}(x;\mu) + \beta_{n-1}(\mu)p_{n-1}(x;\mu).$$

These recurrence coefficients can be obtained directly from $\mathbf{Y}_n(z;\mu)$ via the formulae [6]

$$\alpha_n(\mu) = [\mathbf{Y}_n^{(1)}(\mu)]_{1,1} - [\mathbf{Y}_{n+1}^{(1)}(\mu)]_{1,1},$$
$$\beta_n^2(\mu) = \frac{[\mathbf{Y}_{n+1}^{(1)}(\mu)]_{1,2}}{[\mathbf{Y}_n^{(1)}(\mu)]_{1,2}} = [\mathbf{Y}_n^{(1)}(\mu)]_{1,2}[\mathbf{Y}_n^{(1)}(\mu)]_{2,1},$$

where $\mathbf{Y}_n^{(1)}$ is the unique, z-independent matrix such that

$$\mathbf{Y}_n(z;\mu)z^{-n\boldsymbol{\sigma}_3} = \mathbf{I} + \mathbf{Y}_n^{(1)}(\mu)z^{-1} + O(z^{-2}), \qquad z \to \infty.$$

Let[4]

$$\mathfrak{c} = \frac{1}{2}.$$

[4] In the general case where $[a,b] \neq [-1,1]$, one would set $\mathfrak{c} = \frac{b-a}{4}$.

One considers

$$\mathbf{S}_n(z;\mu) = \mathfrak{c}^{-n\boldsymbol{\sigma}_3}\mathbf{Y}_n(z;\mu)\phi(z)^{-n\boldsymbol{\sigma}_3},$$

where $\phi(z) = z + (z^2-1)^{1/2}$ and the branch of the square root is chosen such that it is analytic in $\mathbb{C}\setminus[-1,1]$ and positive for $z > 1$. Then it follows that [22]

$$\mathbf{S}_n(z;\mu) = \mathbf{I} + \mathbf{S}_n^{(1)}(\mu)z^{-1} + O(z^{-2}), \qquad z\to\infty,$$

where $\mathbf{S}_n^{(1)}(\mu) = O(1)$ as $n\to\infty$. Then, we see that $\phi(z) = 2z(1+O(z^{-2}))$. From this we obtain the expressions

$$\mathbf{S}_n^{(1)}(\mu) = \mathfrak{c}^{-n\boldsymbol{\sigma}_3}\mathbf{Y}_n^{(1)}(\mu)\mathfrak{c}^{n\boldsymbol{\sigma}_3}, \quad \mathbf{Y}_n^{(1)}(\mu) = \mathfrak{c}^{n\boldsymbol{\sigma}_3}\mathbf{S}_n^{(1)}(\mu)\mathfrak{c}^{-n\boldsymbol{\sigma}_3}.$$

We are concerned with the case where μ_* is a relative perturbation of μ. Suppose

$$\mu_*(\mathrm{d}x) = (1+h(x))\rho(x)\mathrm{d}x, \quad |h(x)|<1. \tag{4.1}$$

This encompasses our definition of μ_* in Theorem 3.1. Now, consider

$$\check{\mathbf{Y}}_n(z;\mu) = \mathfrak{c}^{-n\boldsymbol{\sigma}_3}\mathbf{Y}_n(z;\mu), \qquad \mathbf{X}_n(z;\mu,\mu_*) = \check{\mathbf{Y}}_n(z,\mu_*)\check{\mathbf{Y}}_n(z;\mu)^{-1}.$$

Then for $z\in[-1,1]$,

$$\begin{aligned}
\mathbf{X}_n^+(z;\mu,\mu_*) &= \check{\mathbf{Y}}_n^+(z,\mu_*)\check{\mathbf{Y}}_n^+(z;\mu)^{-1}\\
&= \check{\mathbf{Y}}_n^-(z,\mu_*)\begin{bmatrix}1 & (1+h(z))\rho(z)\\ 0 & 1\end{bmatrix}\check{\mathbf{Y}}_n^+(z;\mu)^{-1}\\
&= \check{\mathbf{Y}}_n^-(z,\mu_*)\begin{bmatrix}1 & (1+h(z))\rho(z)\\ 0 & 1\end{bmatrix}\begin{bmatrix}1 & -\rho(z)\\ 0 & 1\end{bmatrix}\check{\mathbf{Y}}_n^-(z;\mu)^{-1}\\
&= \check{\mathbf{Y}}_n^-(z,\mu_*)\begin{bmatrix}1 & h(z)\rho(z)\\ 0 & 1\end{bmatrix}\check{\mathbf{Y}}_n^-(z;\mu)^{-1}\\
&= \mathbf{X}_n^-(z;\mu,\mu_*)\check{\mathbf{Y}}_n^-(z;\mu)\begin{bmatrix}1 & h(z)\rho(z)\\ 0 & 1\end{bmatrix}\check{\mathbf{Y}}_n^-(z;\mu)^{-1}\\
&= \mathbf{X}_n^-(z;\mu,\mu_*)\left[\mathbf{I} + h(z)\rho(z)\check{\mathbf{Y}}_n^-(z;\mu)\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\check{\mathbf{Y}}_n^-(z;\mu)^{-1}\right].
\end{aligned}$$

Define

$$\begin{aligned}
\mathbf{M}_n(z;\mu) &:= \check{\mathbf{Y}}_n^-(z;\mu)\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\check{\mathbf{Y}}_n^-(z;\mu)^{-1} \qquad (4.2)\\
&= \begin{bmatrix}0 & \mathfrak{c}^{-n}\pi_n(z;\mu)\\ 0 & \mathfrak{c}^{n}\gamma_{n-1}(\mu)\pi_{n-1}(z;\mu)\end{bmatrix}\begin{bmatrix}\mathfrak{c}^{n}\gamma_{n-1}(\mu)c_{n-1}^-(z;\mu) & -\mathfrak{c}^{-n}c_n^-(z;\mu)\\ -\mathfrak{c}^{n}\gamma_{n-1}(\mu)\pi_{n-1}(z;\mu) & \mathfrak{c}^{-n}\pi_n(z;\mu)\end{bmatrix}
\end{aligned}$$

$$= \begin{bmatrix} -\gamma_n(\mu)\pi_n(z;\mu)\pi_{n-1}(z;\mu) & \mathfrak{c}^{-2n}\pi_n(z;\mu)^2 \\ -\mathfrak{c}^{2n}\gamma_{n-1}^2(\mu)\pi_{n-1}(z;\mu)^2 & \gamma_{n-1}(\mu)\pi_{n-1}(z;\mu)\pi_n(z;\mu) \end{bmatrix}.$$

For a piecewise-smooth contour $\Gamma \subset \mathbb{C}$, and matrix-valued functions $\mathbf{X} : \Gamma \to \mathbb{C}^{2\times 2}$ we use the norm

$$\|\mathbf{X}\|_{L^2(\Gamma)} := \left(\int_\Gamma \|\mathbf{X}(x)\|_{\mathsf{F}}^2 \, |\mathrm{d}x| \right)^{1/2},$$

where $\|\cdot\|_{\mathsf{F}}$ denotes the Frobenius (Hilbert–Schmidt) norm on $\mathbb{C}^{2\times 2}$ and $L^2(\Gamma)$ is used to denote the space of measurable functions such that this norm is finite. Define the Cauchy operator

$$\mathcal{C}^- u(z) = \lim_{\epsilon \downarrow 0} \mathcal{C}u(z - \mathrm{i}\epsilon), \quad \mathcal{C}u(z) = \frac{1}{2\pi \mathrm{i}} \int_{-1}^{1} \frac{u(z')}{z' - z} \, \mathrm{d}z'.$$

On $L^2([-1,1])$, $\mathcal{C}^-$ is bounded with norm one [1].

Lemma 4.2. *Suppose Assumption 4.1 holds and that h is as in (4.1). If*

$$\Delta(n) = \Delta(n;\mu,\mu_*) := \sup_{z\in[-1,1]} \|h(z)\rho(z)\mathbf{M}_n(z;\mu)\|_{\mathsf{F}} < 1$$

then there exists $\mathbf{u}_n \in L^2([-1,1])$ such that

$$\mathbf{X}_n(z;\mu,\mu_*) = \mathbf{I} + \mathcal{C}\mathbf{u}_n(z) = \mathbf{I} + O\left(\frac{\Delta(n)}{1+|z|}\right), \quad \|\mathbf{u}_n\|_{L^2([-1,1])} \le \frac{\Delta(n)\sqrt{2}}{1-\Delta(n)},$$

uniformly with respect to z on closed subsets of $\mathbb{C} \setminus [-1,1]$.

Proof. Due to the analyticity of $\mathbf{X}_n$, it follows from the theory of Hardy spaces [10] that there exists $\mathbf{u}_n \in L^2(\mathrm{supp}(\mu))$ such that $\mathbf{X}_n(z;\mu,\mu_*) = \mathbf{I} + \mathcal{C}\mathbf{u}_n(z)$. Then the Plemelj-Sokhotski lemma gives that $\mathbf{u}_n(z) = \mathbf{X}_n^+(z;\mu,\mu_*) - \mathbf{X}_n^-(z;\mu,\mu_*)$. This is a solution of the singular integral equation

$$\mathbf{u}_n - (\mathcal{C}^-\mathbf{u}_n)h\rho\mathbf{M}_n = h\rho\mathbf{M}_n,$$

where the operator

$$\mathbf{u} \mapsto \mathbf{u} - (\mathcal{C}^-\mathbf{u})h\rho\mathbf{M}_n, \tag{4.3}$$

is near identity if $\Delta(n) < 1$. Thus $\mathbf{u}_n$ is the unique solution of this integral equation and

$$\|\mathbf{u}_n\|_{L^2([-1,1])} \leq \frac{\Delta(n)\sqrt{2}}{1-\Delta(n)}.$$

The claim then follows from the expression

$$\mathbf{X}_n(z;\mu,\mu_*) = \mathbf{I} + \mathcal{C}\mathbf{u}_n(z). \quad \square$$

This lemma is useful when Theorem 3.1 applies. And, as we do in the following section, it can be used to compare both $\mathbf{T}_k$ and $\overline{\mathbf{T}}_k$ to a limiting three-term recurrence relation satisfied by limiting measure μ_∞. But Theorem 3.1 with $\epsilon_{\text{lan}} = 0$ and $\mu = \mu_\infty$ gives

$$\|h_\infty\|_{[-1,1]} \leq 2D\sigma M_k(\mu_\infty;[-1,1])k\mathfrak{m}_k^{\Delta}(\mu_N,\mu_\infty;\mu_\infty), \tag{4.4}$$

where h_∞ is such that $(1+h_\infty)\mu_\infty$ has its first $2k-1$ moments match with μ_N. And this allows one to estimate the behavior of the orthogonal polynomials with respect to μ_N. In the context of Lemma 4.2, recall that, for simplicity, that $[a,b] = [-1,1]$ and that Γ is a contour that encircles $[-1,1]$ that is a distance at least ν from $[-1,1]$. Then if $\mathbf{u}_n(s)_{ij}$ denotes the (i,j) entry of $\mathbf{u}_n(s)$,

$$\begin{aligned}\|\mathbf{u}_n(\cdot)_{ij}\|_{L^2([-1,1])} \leq \sqrt{2}\frac{\Delta(n,\mu_\infty,\mu_N)}{1-\Delta(n,\mu_\infty,\mu_N)} &\Rightarrow \|\mathcal{C}\mathbf{u}_n(\cdot)_{ij}\|_{L^\infty(\Gamma)} \\ &\leq \sqrt{2}\frac{\Delta(n,\mu_\infty,\mu_N)}{1-\Delta(n,\mu_\infty,\mu_N)}\nu^{-1}|\Gamma|^{1/2},\end{aligned}$$

and therefore

$$\|\mathbf{X}_n(\cdot;\mu_\infty,\mu_N)\|_{L^\infty(\Gamma)} \leq \sqrt{2} + \|\mathcal{C}\mathbf{u}_n\|_{L^\infty(\Gamma)} \leq \sqrt{2} + \frac{\sqrt{8|\Gamma|}}{\nu}\frac{\Delta(n,\mu_\infty,\mu_N)}{1-\Delta(n,\mu_\infty,\mu_N)}.$$

And since $\det \mathbf{X}_N = 1$, the same estimate holds for $\mathbf{X}_N^{-1}$. Now, consider Theorem 3.1 with $\mu = \mu_N$ and suppose that $\|h\|_{\Lambda(A)} \leq 1$. We wish to again use the Fokas–Its–Kitaev Riemann–Hilbert problem, but now it must be modified to handle discrete weights. The reformulation we use can be found in [9], and it involves a jump condition on a contour that encircles the support (with counter-clockwise orientation) of the measures under consideration. Instead of the jump matrix being of the form

$$\begin{bmatrix} 1 & \rho(z) \\ 0 & 1 \end{bmatrix}$$

for $\mu(\mathrm{d}x) = \rho(x)\mathrm{d}x$ it becomes, for $z \in \Gamma$,

$$\begin{bmatrix} 1 & -\frac{1}{2\pi \mathrm{i}}\int_{-1}^{1}\frac{\mu(\mathrm{d}x)}{x-z} \\ 0 & 1 \end{bmatrix}.$$

Define

$$r_N(z) = \frac{1}{2\pi \mathrm{i}} \sum_{n=1}^{N} h(\lambda_n) \frac{(\mathbf{b}^{\mathsf{T}}\mathbf{u}_n)^2}{\lambda_n - z},$$

and we note that $\mathfrak{c} = 1/2$. If μ_N is the VESD for $(\mathbf{A}, \mathbf{b})$ and μ_* is the VESD for $(\mathbf{A}, \mathbf{b}_*)$ as in Remark 2.1 we find

$$\mathbf{X}_n^+(z; \mu_N, \mu_*) = \mathbf{X}_n^-(z; \mu_N, \mu_*) \left[\mathbf{I} + r_N(z)\mathbf{M}_n(z; \mu_N)\right], \quad z \in \Gamma,$$

where the $^{\pm}$ superscripts denote the limit to Γ from the interior/exterior of Γ. Then, we can write

$$\check{\mathbf{Y}}_n^-(z; \mu_N) = \mathbf{X}_n(z; \mu_\infty, \mu_N)\check{\mathbf{Y}}_n(z; \mu_\infty).$$

Therefore

$$\mathbf{X}_n^+(z; \mu_N, \mu_*) = \mathbf{X}_n^-(z; \mu_N, \mu_*) \left[\mathbf{I} + r_N(z) \underbrace{\mathbf{X}_n(z; \mu_\infty, \mu_N)\mathbf{M}_n(z; \mu_\infty)\mathbf{X}_n(z; \mu_\infty, \mu_N)^{-1}}_{\mathbf{M}_n(z; \mu_N)}\right].$$

We then use Theorem 3.1 to estimate r_N on Γ by

$$\|r_N\|_{L^\infty(\Gamma)} \leq \frac{1}{2\pi\nu}\|h\|_{\Lambda(A)} \leq \frac{k\ M_k(\mu; [-1, 1])}{\pi\nu}(\mathfrak{m}_k^{\Delta}(\mu_*,\ \mu_N; \mu_N) + \mathfrak{m}_k^{\Delta}(\mu_N, \mu_N; \mu_N)).$$

And here h is such that $\mu_* = (1+h)\mu_N$ has the same moments as $\overline{\mu}_k$, the finite precision Lanczos measure for $n = 1, 2, \ldots, 2k-1$.

Assumption 4.3. *Suppose Assumption 4.1 holds, let $\hat{\mu}_\infty$ be the push-forward measure of μ_∞ under $x \mapsto \frac{2}{b-a}\left(x - \frac{b+a}{2}\right)$, and let $\hat{\rho}$ be the corresponding density. Suppose further that*

(a) there exist $E, \delta, K > 0$ such that for all $k > K$, $\max_{n \leq k+1} \sup_{z \in [a,b]} \|\hat{\rho}(z)\mathbf{M}_n(z; \hat{\mu}_\infty)\|_{\mathsf{F}} \leq Ek^\delta$, and
(b) Γ is chosen such that there exist $E', \delta', K' > 0$ such that for all $k > K'$,

$$\max_{n \leq k+1} \sup_{z \in \Gamma} \|\mathbf{M}_n(z; \hat{\mu}_\infty)\|_{\mathsf{F}} \leq E' k^{\delta'}, \quad \nu \geq \frac{1}{k^2}.$$

Remark 4.4. In this assumption, we want to be able to consider $\nu = O(k^{-2})$ because that is the scale on which orthogonal polynomials grow polynomially, see Lemma 3.7.

Lemma 4.5. *Fix μ_∞ with support $[-1,1]$ and constants $L, \alpha, E, E', \delta, \delta'$. Suppose Assumption 3.4 (with parameters $L, \gamma = 3/2, \alpha$) holds for μ_N and μ_∞ and that $\mu = \mu_\infty$ satisfies Assumption 4.3 (with parameters E, E', δ, δ'). Suppose further that*

$$k = o(N^{2\alpha/(8+\delta)}), \qquad \epsilon_{\text{lan}} = o(k^{-(9+\delta')}), \qquad N \to \infty.$$

Then, for μ_* as in (4.1),

$$\mathbf{X}_n(z; \mu_N, \mu_*) = \mathbf{I} + O\left(\frac{\epsilon_{\text{lan}} k^{8+\delta'}}{1+|z|}\right), \qquad N \to \infty,$$

uniformly on sets bounded away from Γ.

Proof. We point out that Assumption 4.1 forces $\gamma \geq 3/2$ and $\gamma = 3/2$ is always possible. Recall that $a = -1$ and $b = 1$. If we set $c = N^{-\beta}$ and assume

$$k \leq \left(\frac{1}{8}\right)^{\gamma/(4+2\gamma)} \left(\frac{\sqrt{L} N^{\alpha-\beta}}{32}\right)^{1/(2+\gamma)}, \tag{4.5}$$

then Corollary 3.6 gives the bounds

$$M_k(\mu_\infty; [-1,1]) \leq \frac{2}{\sqrt{L}} 8^{\gamma/2} k^\gamma, \qquad \mathfrak{m}_k^{\Delta}(\mu_N, \mu_\infty; \mu_\infty) \leq N^{-\beta}.$$

With $\epsilon_{\text{lan}} = 0$, $\mu = \mu_\infty$, Theorem 3.1 then implies

$$\|h_\infty\|_{[-1,1]} \leq 2k M_k(\mu_\infty; [-1,1]) \mathfrak{m}_k^{\Delta}(\mu_N, \mu_\infty; \mu_\infty).$$

Here $\sigma \leq 2$ since $\|\mathbf{A}_N\| \leq 2$ by Assumption 3.4. Using Assumption 4.3 we find a bound

$$\Delta(k; \mu_\infty, \mu_N) = \sup_{z \in [-1,1]} |h_\infty(z)| \|\rho(z) \mathbf{M}_k(z; \mu_\infty)\|_\infty = O(k^{1+\gamma+\delta} N^{-\beta}). \tag{4.6}$$

Assuming $\epsilon_{\text{lan}} < 1/(\sigma C k^2)$, $\mu = \mu_N$, we revisit Theorem 3.1 which then implies,

$$\|h\|_{[-1,1]} \leq 2D\sigma M_k(\mu_N; [-1,1])^2 k^4 \epsilon_{\text{lan}}.$$

Again $\sigma \leq 2$. Provided that $\Delta(n; \mu_\infty, \mu_N) < 1$, we estimate

$$\|r_N \mathbf{X}_n(\cdot; \mu_\infty, \mu_N) \mathbf{M}_n(\cdot; \mu_\infty) \mathbf{X}_n(\cdot; \mu_\infty, \mu_N)^{-1}\|_{L^\infty(\Gamma)} \leq \left(1 + \frac{2\sqrt{|\Gamma|}}{\nu} \frac{\Delta(n, \mu_\infty, \mu_N)}{1 - \Delta(n, \mu_\infty, \mu_n)}\right)^2 \frac{\|h\|_{[-1,1]}}{\pi\nu} \|\mathbf{M}_n(\cdot; \mu_\infty)\|_{L^\infty(\Gamma)}. \tag{4.7}$$

Since the Riemann–Hilbert problem has a jump condition on the contour Γ, (4.3) is replaced with

$$\mathbf{u} \mapsto \mathbf{u} - (\mathcal{C}_\Gamma^- \mathbf{u}) r_N \mathbf{M}_n(\cdot; \mu_N).$$

And to obtain a near-identity operator, we require that (4.7) is small. Here $\mathcal{C}_\Gamma$ is the Cauchy operator on the contour Γ and it, operating on $L^2(\Gamma)$, may have a norm larger than one [1] but we may choose Γ so that the norm is bounded by 2. Since $\gamma = 3/2$, condition (4.5) is stronger than

$$k \leq \frac{1}{4}\left(\frac{LN^\alpha}{3}\right)^{1/(2\gamma)},$$

which, with Corollary 3.5, gives the bound

$$M_k(\mu_N; [-1,1]) \leq \frac{4}{\sqrt{L}} 16^{\gamma/2} k^\gamma.$$

Suppose

$$k = o(N^{\beta/(3+\gamma+\delta)}), \qquad \epsilon_{\text{lan}} = o(k^{-(\delta'+2\gamma+6)}). \tag{4.8}$$

Then,

$$\Delta(n, \mu_\infty, \mu_N) = \sup_{z\in[-1,1]} |h(z)| \|\rho(z)\mathbf{M}_k(z;\ \mu_N)\|_\infty = O(k^{1+\gamma+\delta} N^{-\beta}) = o(1).$$

The assumption on k in (4.8) implies $k = o(N^\beta/(1+\gamma+\delta))$ so that, using (4.6),

$$\frac{1}{\nu}\frac{\Delta(n, \mu_\infty, \mu_N)}{1 - \Delta(n, \mu_\infty, \mu_N)} = o(k^{3+\gamma+\delta} N^{-\beta}) = o(1),$$

and therefore

$$(4.7) = O(\nu^{-1}\epsilon_{\text{lan}} k^{\delta'+2\gamma+4}) = o(\epsilon_{\text{lan}} k^{\delta'+2\gamma+6}).$$

We can conclude that

$$\mathbf{u}_n - (\mathcal{C}_\Gamma^- \mathbf{u}_n) r_N \mathbf{M}_n(\cdot; \mu_N) = r_N \mathbf{M}_n(\cdot; \mu_N)$$

has a unique solution which satisfies $\|\mathbf{u}_n\|_{L^2(\Gamma)} = O(\epsilon_{\text{lan}} k^{\delta'+2\gamma+6})$ and $\mathbf{X}_n = \mathbf{I} + \mathcal{C}_\Gamma \mathbf{u}_n$.

Recalling that $\gamma = 3/2$, to balance our constraints on k, we set

$$\beta = \frac{\alpha}{7}\left(\frac{16+2\delta}{7(9+2\delta)}\right)^{-1} = \alpha\frac{9+2\delta}{16+2\delta} \quad \Longrightarrow \quad \frac{\alpha-\beta}{2+\gamma} = \frac{2(\alpha-\beta)}{7} = \frac{2\beta}{9+2\delta} = \frac{\beta}{3+\gamma+\delta}.$$

Thus, both conditions relating k and N are satisfied $k = o(N^{2\alpha/(16+2\delta)}) = o(N^{\alpha/(8+\delta)})$. □

4.1. Recurrence coefficients

The residue at infinity of $\mathbf{S}_n$ can be computed by first defining $\mathbf{X}_n^{(1)}$ uniquely by

$$\mathbf{X}_n(z;\mu,\mu_*) = \mathbf{I} + \mathbf{X}_n^{(1)}(\mu,\mu_*)z^{-1} + O(z^{-2}), \quad z \to \infty,$$

and using

$$\begin{aligned}
\mathfrak{c}^{n\boldsymbol{\sigma}_3}\check{\mathbf{Y}}_n(z;\mu_*)z^{-n\boldsymbol{\sigma}_3} &= \mathfrak{c}^{n\boldsymbol{\sigma}_3}\mathbf{X}_n(z;\mu,\mu_*)\check{\mathbf{Y}}_n(z;\mu)z^{-n\boldsymbol{\sigma}_3} \\
&= \mathfrak{c}^{n\boldsymbol{\sigma}_3}\mathbf{X}_n(z;\mu,\mu_*)\mathfrak{c}^{-n\boldsymbol{\sigma}_3}\mathfrak{c}^{n\boldsymbol{\sigma}_3}\check{\mathbf{Y}}_n(z;\mu)z^{-n\boldsymbol{\sigma}_3} \\
&= \left(\mathbf{I} + \mathfrak{c}^{n\boldsymbol{\sigma}_3}\mathbf{X}_n^{(1)}(\mu,\mu_*)\mathfrak{c}^{-n\boldsymbol{\sigma}_3}z^{-1} + O(z^{-2})\right) \\
&\qquad \left(\mathbf{I} + z^{-1}\left[\mathfrak{c}^{n\boldsymbol{\sigma}_3}\mathbf{S}_n^{(1)}(\mu)\mathfrak{c}^{-n\boldsymbol{\sigma}_3} - \frac{b+a}{2}n\boldsymbol{\sigma}_3\right] + O(z^{-2})\right),
\end{aligned}$$

and therefore

$$\mathbf{Y}_n^{(1)}(\mu_*) = \mathfrak{c}^{n\boldsymbol{\sigma}_3}\left(\mathbf{X}_n^{(1)}(\mu,\mu_*) + \mathbf{S}_n^{(1)}(\mu) - \frac{b+a}{2}n\boldsymbol{\sigma}_3\right)\mathfrak{c}^{-n\boldsymbol{\sigma}_3}.$$

Then, we may express the recurrence coefficients for μ_* in terms of those for μ, and $\mathbf{X}_n^{(1)}$, via

$$\begin{aligned}
\alpha_n(\mu_*) &= \alpha_n(\mu) + [\mathbf{X}_n^{(1)}(\mu,\mu_*)]_{1,1} - [\mathbf{X}_{n+1}^{(1)}(\mu,\mu_*)]_{1,1}, \\
\beta_n(\mu_*)^2 &= [\mathbf{X}_n^{(1)}(\mu,\mu_*) + \mathbf{S}_n^{(1)}(\mu)]_{1,2}[\mathbf{X}_n^{(1)}(\mu,\mu_*) + \mathbf{S}_n^{(1)}(\mu)]_{2,1} \\
&= \beta_n(\mu)^2 + [\mathbf{X}_n^{(1)}(\mu,\mu_*)]_{1,2}[\mathbf{S}_n^{(1)}(\mu)]_{2,1} + [\mathbf{X}_n^{(1)}(\mu,\mu_*)]_{2,1}[\mathbf{S}_n^{(1)}(\mu)]_{1,2} \\
&\qquad + [\mathbf{X}_n^{(1)}(\mu,\mu_*)]_{1,2}[\mathbf{X}_n^{(1)}(\mu,\mu_*)]_{2,1}.
\end{aligned}$$

To obtain the optimal scaling of quantities with respect to $\operatorname{supp}(\mu) = [a,b]$, it is convenient to rescale first so that the support of μ is transformed to $[-1,1]$, and then undo the scaling after estimates are obtained. This gives a perturbation result.

Corollary 4.6. *Given the assumptions of Lemma 4.2, let $\hat{\mu}$ and $\hat{\mu}_*$ be the push-forward measures of μ and μ_* under $x \mapsto \frac{2}{b-a}\left(x - \frac{b+a}{2}\right)$, respectively. Suppose that $\Delta(n) = \Delta(n;\hat{\mu},\hat{\mu}_*)$ satisfies $\Delta(n), \Delta(n+1) \leq 1/2$. Then*

$$\begin{aligned}
|\alpha_n(\mu_*) - \alpha_n(\mu)| &\leq \pi^{-1}(\Delta(n) + \Delta(n+1))(b-a), \\
|\beta_n(\mu_*)^2 - \beta_n(\mu)^2| &\leq 2\pi^{-1}\left[\max_n \|\mathbf{S}_n^{(1)}(\hat{\mu})\|\right]\Delta(n)(b-a)^2 + \pi^{-2}\Delta(n)^2(b-a)^2.
\end{aligned}$$

Proof. Lemma 4.2 applies. We work with the expression

$$\mathbf{X}_n^{(1)}(\hat{\mu},\hat{\mu}_*) = -\frac{1}{2\pi \mathrm{i}}\int_{-1}^{1} \mathbf{u}_n(s)\,\mathrm{d}s.$$

Then if $\mathbf{u}_n(s)_{ij}$ denotes the (i,j) entry of $\mathbf{u}_n(s)$,

$$\|\mathbf{u}_n(\cdot)_{ij}\|_{L^1([-1,1])} \leq \sqrt{2}\|\mathbf{u}_n(\cdot)_{ij}\|_{L^2([-1,1])}, \tag{4.9}$$

and therefore

$$\|\mathbf{X}_n^{(1)}(\hat{\mu},\hat{\mu}_*)\|_{\max} \leq \frac{1}{\pi}\frac{\Delta(n)}{1-\Delta(n)}.$$

The same estimates hold with n replaced with $n+1$. The claim follows by recalling that $\max_n \|\mathbf{S}_n^{(1)}(\hat{\mu})\| < \infty$ and that

$$\alpha_n(\mu) = \frac{b-a}{2}\alpha_n(\hat{\mu}) + \frac{b+a}{2}, \qquad \alpha_n(\mu_*) = \frac{b-a}{2}\alpha_n(\hat{\mu}_*) + \frac{b+a}{2},$$
$$\beta_n(\mu) = \frac{b-a}{2}\beta_n(\hat{\mu}), \qquad \beta_n(\mu_*) = \frac{b-a}{2}\beta_n(\hat{\mu}_*). \quad \square$$

We now establish our forward stability result.

Theorem 4.7 *(Forward stability). Suppose* $\mathrm{supp}(\mu_\infty) = [a,b]$ *and let*

$$\Gamma = \{z \in \mathbb{C} : 1 + k^{-1} = |z + \sqrt{z-1}\sqrt{z+1}|\}.$$

Suppose the assumptions[5] *of Lemma 4.5 hold for* Γ *and* $\hat{\mu}_\infty, \hat{\mu}_N$*, the push-forward measures of* μ_∞, μ_N*, respectively, under* $x \mapsto \frac{2}{b-a}\left(x - \frac{b+a}{2}\right)$*. Suppose further that*

$$k = o(N^{\alpha/(8+\delta)}), \qquad \epsilon_{\mathrm{lan}} = o(k^{-(9+\delta')}), \qquad N \to \infty.$$

Then

$$\max_{n\leq k+1} |\alpha_n(\mu_N) - \alpha_n(\overline{\mu}_k)| = O\left((b-a)\epsilon_{\mathrm{lan}} k^{\delta'+9}\right), \quad \textit{and}$$
$$\max_{n\leq k+1} |\beta_n(\mu_N)^2 - \beta_n(\overline{\mu}_k)^2| = O\left((b-a)^2\epsilon_{\mathrm{lan}} k^{\delta'+9}\right), \quad N \to \infty.$$

Therefore, if $((\mathbf{A}_N, \mathbf{b}_N))_{N\geq 1}$ *is a sequence of problems such that* $\hat{\mu}_N$ *(*$\mu_N = \mu_{\mathrm{VESD}}(\mathbf{A}_N, \mathbf{b}_N)$*) satisfies Assumption 3.4,*

$$\|\mathbf{T}_k - \overline{\mathbf{T}}_k\|_{\max} = O((b-a)\epsilon_{\mathrm{lan}} k^{9+\delta'}).$$

[5] Note that within the assumptions of Lemma 4.5 is Assumption 4.3 which involves Γ.

Proof. Note that

$$\alpha_n(\mu_N) = \frac{b-a}{2}\alpha_n(\hat{\mu}_N) + \frac{b+a}{2}, \qquad \alpha_n(\overline{\mu}_k) = \frac{b-a}{2}\alpha_n(\hat{\overline{\mu}}_k) + \frac{b+a}{2},$$
$$\beta_n(\mu_N) = \frac{b-a}{2}\beta_n(\hat{\mu}_N), \qquad \beta_n(\overline{\mu}_k) = \frac{b-a}{2}\beta_n(\hat{\overline{\mu}}_k).$$

The only portion of this result that does not follow immediately from Lemma 4.5 is the last claim because one needs to show

$$|\beta_n(\mu_N) - \beta_n(\overline{\mu}_k)| = (b-a)\frac{|\beta_n(\hat{\mu}_N)^2 - \beta_n(\hat{\overline{\mu}}_k)^2|}{|\beta_n(\hat{\mu}_N) + \beta_n(\hat{\overline{\mu}}_k)|} = O((b-a)|\beta_n(\hat{\mu}_N)^2 - \beta_n(\hat{\overline{\mu}}_k)^2|).$$

Let $\hat{\mu}_\infty$ be the pushforward measure of μ_∞ under the same mapping as above. Corollary 4.6 gives a lower bound on $|\beta_n(\hat{\mu}_N) + \beta_n(\hat{\overline{\mu}}_k)|$ using the fact that $\beta_n(\hat{\mu}_N) \to \beta_n(\hat{\mu}_\infty)$ where the positivity of $\beta_n(\hat{\mu}_\infty)$ is crucial. The restrictions on k for this fact are milder than the assumptions of the theorem. And then $\mathbf{S}_n^{(1)}(\hat{\mu}_N)$ is estimated using the relation

$$\mathbf{S}_n^{(1)}(\hat{\mu}_N) = \mathbf{S}_n^{(1)}(\hat{\mu}_\infty) + \mathbf{X}_n^{(1)}(\mu_\infty, \hat{\mu}_N),$$

and using Lemma 4.2. □

Remark 4.8. Our approach yields a forwards stability result for the Lanczos algorithm in the case that μ_N is near to a measure μ satisfying Assumption 4.1. This is the case in many situations, for instance, when μ_N is the VESD associated to many large random matrices and μ is the limiting measure. In particular, in the context of Fig. 1 our approach explains the observation that $\overline{\alpha}_i$ and $\overline{\beta}_i$ are near 0 and 1/2 respectively.

5. Random matrices

A natural setting in which Assumption 3.4 holds is when $\mu_N = \mu_{\mathrm{VESD}}(\cdot\,; \mathbf{A}, \mathbf{b})$ is the VESD of a random matrix and (random or deterministic) vector and μ_∞ is the limiting spectral distribution. In what follows $\mathbb{E}$ and $\mathbb{P}$ will denote the expectation and the probability of an event with respect to a probability distribution that will be clear from context.

We first define the notion of a *local law*. This holds for a wide class of random matrices [12,24] and we suppose the limiting measure μ_∞ has bounded support. We discuss two classical examples of such random matrices below.

Define two N-dependent regions in the complex plane by

$$\mathcal{D}(N,\tau) = \{z \in \mathbb{C} : N^{-1+\tau} \le \operatorname{Im} z \le \tau^{-1}, \quad \operatorname{dist}(\operatorname{Re} z, \operatorname{supp} \mu_\infty) < \tau^{-1}\},$$
$$\mathcal{D}_o(N,\tau) = \{z \in \mathbb{C} : 0 < \operatorname{Im} z \le \tau^{-1}, \quad N^{-2/3+\tau} \le \operatorname{dist}(\operatorname{Re} z, \operatorname{supp} \mu_\infty) < \tau^{-1}\}.$$

The first region is useful for estimating quantities near the interior of the support of μ_∞ and the latter is used to estimate quantities near the edges of the support of μ_∞.

Definition 5.1. Suppose $\mathbf{A}_N$ is an $N \times N$ random matrix and $\mathbf{b}_N$ is an N-dimensional vector. The sequence $((\mathbf{A}_N, \mathbf{b}_N))_{N\geq 1}$, $\|\mathbf{b}_N\| = 1$ is said to satisfy a local law with limit μ_∞ if (1) for every fixed $\tau > 0$, $\epsilon > 0$ and $D > 0$ there exists C such that

$$\begin{aligned}\sup_{z\in\mathcal{D}(N,\tau)} \mathbb{P}\left(\left|\mathbf{b}_N^{\mathsf{T}}(\mathbf{A}_N - z\mathbf{I})^{-1}\mathbf{b}_N - \mathcal{S}(z;\mu_\infty)\right| \geq N^{\epsilon}\left(\sqrt{\frac{\operatorname{Im}\mathcal{S}(z;\mu_\infty)}{N\operatorname{Im} z}} + \frac{1}{N\operatorname{Im} z}\right)\right) \\ \leq CN^{-D},\\ \sup_{z\in\mathcal{D}_o(N,\tau)} \mathbb{P}\left(\left|\mathbf{b}_N^{\mathsf{T}}(\mathbf{A}_N - z\mathbf{I})^{-1}\mathbf{b}_N - \mathcal{S}(z;\mu_\infty)\right| \geq N^{\epsilon}\sqrt{\frac{\operatorname{Im}\mathcal{S}(z;\mu_\infty)}{N\operatorname{Im} z}}\right) \leq CN^{-D}\end{aligned} \tag{5.1}$$

where

$$\mathcal{S}(z;\mu_\infty) = \int \frac{\mu_\infty(\mathrm{d}x)}{x - z},$$

and (2) there exists $L > 0$ such that

$$\mathbb{P}(\|\mathbf{A}_N\| > L) \leq CN^{-D}.$$

We pause to note that, importantly,

$$\mathbf{b}_N^{\mathsf{T}}(\mathbf{A}_N - z\mathbf{I})^{-1}\mathbf{b}_N = \mathcal{S}(z;\mu_{\mathrm{VESD}}(\mathbf{A}_N, \mathbf{b}_N)).$$

For a function F of bounded variation we also use the notation

$$\mathcal{S}(z;F) := \int \frac{\mathrm{d}F(x)}{x - z}, \qquad \operatorname{Im} z > 0,$$

to denote the Riemann–Stieltjes integral.

To turn a local law into an estimate on the KS distance we use the following.

Corollary 5.2 *([2], Corollary B.15). Let F be a distribution function and let G be a function of bounded variation satisfying $\int |F(x) - G(x)|\mathrm{d}x < \infty$. Assume that, for some constants $A > B > 0$,*

$$\int_{-B}^{B} \mathrm{d}F(x) = 1, \qquad \int_{-\infty}^{-B} |\mathrm{d}G(x)| = 0 = \int_{B}^{\infty} |\mathrm{d}G(x)|.$$

Then

$$d_{\mathrm{KS}}(F,G) := \sup_x |F(x)-G(x)| \leq \frac{1}{\pi(1-\kappa)(2\gamma-1)}\left[\int_{-A}^{A} |\mathcal{S}(z;\mathrm{d}F)-\mathcal{S}(z;\mathrm{d}G)|\mathrm{d}u + \frac{1}{v}\sup_x \int_{|y|\leq 2va} |G(x+y)-G(x)|\,\mathrm{d}y\right],$$

where κ *satisfies*

$$\kappa = \frac{4B}{\pi(A-B)(2\gamma-1)} < 1,$$

$z = u + \mathrm{i}v$, *and* γ *and* a *are related via*

$$\gamma = \frac{1}{\pi}\int_{-a}^{a} \frac{\mathrm{d}u}{1+u^2} > \frac{1}{2}.$$

Without loss of generality, to estimate a KS distance, we can suppose that $\mathrm{supp}(\mu_\infty) \subseteq [-b,b]$. For $0<\tau<1$, set $A = b+\tau^{-1}$, $B = b+\tau$. Then

$$\kappa = \frac{4(b+\tau)}{\pi(\tau^{-1}-\tau)(2\gamma-1)} \overset{\tau\to 0}{\longrightarrow} 0.$$

And we choose τ sufficiently small so that this quantity κ is less than 1. We assume τ, and hence A, are chosen in this way for the forthcoming results.

Lemma 5.3. *Suppose* $\mathrm{supp}(\mu_\infty) \subseteq [-b,b]$, $b>0$, *where* μ_∞ *has a Hölder continuous, bounded density on* $\mathbb{R}$. *Suppose the sequence* $((\mathbf{A}_N,\mathbf{b}_N))_{N\geq 1}$ *satisfies a local law with limit* μ_∞. *Then for every* $\epsilon>0$, $D>0$, $M>1$, $z=u+\mathrm{i}v \in D(N,\tau)$,

$$\mathbb{P}\left(\int_{-A}^{A} |\mathcal{S}(z;\mu_N)-\mathcal{S}(z;\mu_\infty)|\mathrm{d}u \geq N^{\epsilon}\sqrt{\frac{1}{N\operatorname{Im} z}+\frac{8A^2}{(\operatorname{Im} z)^2 M}}\right) \leq CMN^{-D},$$

where $\mu_N(\mathrm{d}x) = \mu_{\mathrm{VESD}}(\mathrm{d}x;\mathbf{A}_N,\mathbf{b}_N)$. *Furthermore, if* G *is the cumulative distribution function for* μ_∞ *then*

$$\frac{1}{v}\sup_x \int_{|y|\leq 2va} |G(x+y)-G(x)|\,\mathrm{d}y \leq \sup_x |G'(x)| 2va^2.$$

Proof. The last statement follows immediately after using the mean-value theorem. To establish the first claim, we must discretize the integral that is involved and show that we need only use polynomially many discretization points to approximate it to any desired

accuracy. To do this, we show that the Lipschitz constant of the integrand depends on $\operatorname{Im} z$ in a sufficiently benign way. For $\operatorname{Im} z > 0$,

$$\left|\frac{\mathrm{d}}{\mathrm{d}z}\mathcal{S}(\mu_\infty; z)\right| = \int \frac{\mu_\infty(\mathrm{d}x)}{|x-z|^2} \leq \frac{1}{(\operatorname{Im} z)^2}.$$

From this, it follows for $f(z) := \mathcal{S}(z; \mu_{\mathrm{VESD}}(\mathbf{A}_N, \mathbf{b}_N)) - \mathcal{S}(z; \mu_\infty)$ we have

$$\big||f(u+\mathrm{i}v)| - |f(u'+\mathrm{i}v)|\big| \leq |f(u+\mathrm{i}v) - f(u'+\mathrm{i}v)| \leq \frac{2}{v^2}|u-u'|, \quad v > 0.$$

This implies that if one discretizes the integral

$$\int_{-A}^{A} |f(u+\mathrm{i}v)|\,\mathrm{d}u,$$

using $M+1$ equally-spaced points $-A = x_0, x_1, \ldots, x_M = A$, then

$$\left|\frac{2A}{M}\sum_{j=0}^{M-1} |f(x_j+\mathrm{i}v)| - \int_{-A}^{A} |f(u+\mathrm{i}v)|\,\mathrm{d}u\right| \leq \frac{8A^2}{v^2 M},$$

because

$$\left|\int_{x_j}^{x_{j+1}} (|f(u+\mathrm{i}v)| - |f(x_j+\mathrm{i}v)|)\mathrm{d}u\right| \leq \frac{2A}{M}\frac{2}{v^2}|x_j - x_{j+1}| = \frac{2}{v^2}\left(\frac{2A}{M}\right)^2.$$

Now set $v = N^{-\alpha}$ for $0 \leq \alpha < 1/2$ and fix $\epsilon > 0$. And let $E_{N,M}$ be the event on which

$$\big|\mathbf{b}_N^{\mathsf{T}}(\mathbf{A}_N - z_j\mathbf{I})^{-1}\mathbf{b}_N - \mathcal{S}(z_j;\mu_\infty)\big| \geq N^{\epsilon}\left(\sqrt{\frac{\operatorname{Im}\mathcal{S}(z_j;\mu_\infty)}{N\operatorname{Im} z_j}} + \frac{1}{N\operatorname{Im} z_j}\right),$$

$$z_j = x_j + \mathrm{i}N^{-\alpha},$$

for some $j = 0, 1, 2, \ldots, M-1$. Using the estimates for $\mathcal{D}(N, \tau)$, we have that $\mathbb{P}(E_{N,M}) \leq CMN^{-D}$ where D is as large as we like. The result then follows by simply using that $\operatorname{Im}\mathcal{S}(z;\mu_\infty)$ is bounded in the upper-half plane and then bounding

$$\left(\sqrt{\frac{\operatorname{Im}\mathcal{S}(z_j;\mu_\infty)}{N\operatorname{Im} z_j}} + \frac{1}{N\operatorname{Im} z_j}\right) \leq K\sqrt{\frac{1}{Nv}},$$

for a constant K. □

Theorem 5.4. *Suppose the sequence* $((\mathbf{A}_N, \mathbf{b}_N))_{N\geq 1}$ *satisfies a local law with limit* μ_∞ *and that* μ_∞ *has a Hölder continuous, bounded density on* $\mathbb{R}$. *Then for any* $D > 0$, $\epsilon > 0$, *there exists* $C > 0$ *such that*

$$\mathbb{P}\left(d_{\mathrm{KS}}(\mu_{\mathrm{VESD}}(\mathbf{A}_N, \mathbf{b}_N), \mu_\infty) \geq N^{\epsilon-1/3}\right) \leq CN^{-D},$$

for N *sufficiently large.*

Proof. So, using both Corollary 5.2 and Lemma 5.3 if μ_∞ has a Hölder continuous, bounded density on $\mathbb{R}$ and if the sequence $((\mathbf{A}_N, \mathbf{b}_N))_{N\geq 1}$ satisfies a local law with limit μ_∞, we can conclude that

$$\mathbb{P}\left(d_{\mathrm{KS}}(\mu_{\mathrm{VESD}}(\mathbf{A}_N, \mathbf{b}_N), \mu_\infty) \geq N^{\epsilon}\sqrt{\frac{1}{N\eta} + \frac{8A^2}{\eta^2 M} + 2\|G'\|_\infty a^2\eta}\right) \leq CMN^{-D},$$

where all the constants have the same meaning as in Corollary 5.2 and Lemma 5.3. To optimize the error here, we see that one should take $\eta = N^{-1/3}$ and $M = N$, establishing the claim. □

Remark 5.5. The portion of the local law that applies to $\mathcal{D}_o$ can be used to show that for any $\epsilon > 0$ the support points of the VESD for $(\mathbf{A}_N, \mathbf{b}_N)$ must lie within a distance $N^{-2/3+\epsilon}$ of $\operatorname{supp}\mu$ with overwhelming probability [12].

5.1. Wigner matrices

Consider a random matrix

$$\mathbf{A}_N = \frac{1}{2\sqrt{N}}(a_{ij})_{1\leq i,j\leq N} \tag{5.2}$$

where the real-valued random variables a_{ij} are jointly independent for $i \leq j$ and satisfy

$$\mathbb{E}[a_{ij}] = 0, \qquad \mathbb{E}[a_{ij}^2] = \begin{cases} 1 & i \neq j \\ c & i = j \end{cases}, \qquad a_{ij} = a_{ji},$$

and

$$\mathbb{E}[|a_{ij}|^k] = C_k < \infty,$$

for all $k > 2$. Such a matrix is called a Wigner matrix and the distribution is referred to as a Wigner ensemble. In [12], for example, it is shown that for any sequence of vectors

$(\mathbf{b}_N)_{N\geq 1}$ that are independent[6] of a_{ij} for all i, j, the pairs $((\mathbf{A}_N, \mathbf{b}_N))_{N\geq 1}$ satisfy a local law supported on $[-1, 1]$ with

$$\mu_\infty(\mathrm{d}x) = \mu_U(\mathrm{d}x) = \frac{2}{\pi}\sqrt{1-x^2}\mathrm{d}x. \tag{5.3}$$

The most widely known case of a Wigner matrix is the so-called Gaussian Orthogonal Ensemble[7] (GOE):

$$\mathbf{A}_N = \frac{1}{2\sqrt{2N}}(\mathbf{X} + \mathbf{X}^\mathsf{T}), \tag{5.4}$$

where X is an $N \times N$ Gaussian matrix with independent and identically distributed (iid) standard normal entries. In the work of Trotter [43] (see also [5]) a full distributional characterization of the Householder tridiagonalization of GOE is given. From this, one can see that the upper-left subblocks of this tridiagonalization tend to

$$\operatorname{tridiag}\begin{pmatrix} 1/2 & 1/2 & \cdots & 1/2 \\ 0 & 0 & \cdots & 0 \\ 1/2 & 1/2 & \cdots & 1/2 \end{pmatrix},$$

which correctly reflects the fact that the local law has (5.3) as its limit.[8]

Then, for $\rho(x) = \frac{2}{\pi}\sqrt{1-x^2}$, $U_n(x) = 2^n\pi_n(x; \mu_U)$,

$$1 = \|U_n\|_{L^2(\mu_U)} = 2^n\|\pi_n(\cdot\,; \mu_U)\|_{L^2(\mu_U)}, \qquad \gamma_{n-1}(\mu_U) = -\mathrm{i}\pi 2^{2n-1}, \qquad \mathfrak{c} = \frac{1}{2},$$

and

$$\mathbf{M}_n(z; \mu_U) = \begin{bmatrix} -\pi\mathrm{i}U_n(z)U_{n-1}(z) & U_n(z)^2 \\ \frac{\pi^2}{4}U_{n-1}(z)^2 & -\pi\mathrm{i}U_n(z)U_{n-1}(z) \end{bmatrix}.$$

It follows that

$$|U_n(z)\sqrt{1-z^2}| \leq 1, \qquad z \in [-1, 1],$$

so

$$|\rho(z)\mathbf{M}_n(z; \mu_U)| \underset{\text{entrywise}}{\leq} \begin{bmatrix} 2|U_{n-1}(z)| & \frac{2}{\pi}|U_n(z)| \\ \frac{\pi}{2}|U_{n-1}(z)| & 2|U_{n-1}(z)| \end{bmatrix}, \quad z \in [-1, 1].$$

[6] This reference actually show that the local law holds for any fixed deterministic sequence of vectors, but the estimates the authors give are uniform in the choice of vectors, and the result can be extended to hold for random vectors that are independent of the matrix entries.

[7] Note that the scaling of the matrix here is chosen so that the eigenvalues typically lie within the interval $[-1, 1]$.

[8] This is the Jacobi matrix associated with Chebyshev second-kind polynomials.

Since $\|U_n\|_{[-1,1]} \leq n+1$, we see that we may set $\delta = 1, \delta' = 2$ in Theorem 4.7. Then for Assumption 3.4 to hold with high probability, we take $\gamma = 3/2$, $\beta < 1/3$. This gives forward stability, with high probability, provided

$$k = o(N^{\frac{2}{54}-\epsilon}), \quad \epsilon_{\text{lan}} = o(k^{-11}), \tag{5.5}$$

with forward error $O(\epsilon_{\text{lan}} k^{11})$. Better bounds could be obtained by deriving a version of Theorem 4.7 which uses an explicit bound on the growth of the orthogonal polynomials of μ_U instead of Corollary 3.6.

5.2. Sample covariance matrices

Consider the random matrix

$$\mathbf{A}_N = \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathsf{T}}, \quad \mathbf{X} = (x_{ij})_{\substack{1\leq i\leq N\\ 1\leq j\leq M}}, \qquad M \geq N, \tag{5.6}$$

where the real-valued random variables x_{ij} are jointly independent for all i, j and satisfy

$$\mathbb{E}[x_{ij}] = 0, \qquad \mathbb{E}[x_{ij}^2] = 1,$$

and

$$\mathbb{E}[|x_{ij}|^k] \leq C_k < \infty,$$

for all $k > 2$. Such a matrix is called a sample covariance matrix. In [24], for example, it is shown that for any sequence of vectors $(\mathbf{b}_N)_{N\geq 1}$ that are independent[9] of x_{ij} for all i, j, the pairs $((\mathbf{A}_N, \mathbf{b}_N))_{N\geq 1}$ satisfy a local law supported on $[\lambda_-, \lambda_+]$ with

$$\mu_\infty(\mathrm{d}x) = \mu_{\text{MP}}(\mathrm{d}x) = \frac{1}{2\pi dx}\sqrt{(\lambda_+ - x)(x-\lambda_-)}\,\mathrm{d}x, \qquad \lambda_\pm = (1 \pm \sqrt{d})^2,$$

if $N/M \to d \in (1, \infty)$. This is the well-known Marchenko–Pastur law.

The most widely studied example of a sample covariance matrix is the so-called Wishart distribution [45] where x_{ij} are iid standard normal random variables and in this case the Golub-Kahan bidiagonalization procedure can be carried out in a distributional sense [42] (see also [5]). From this, one sees that the upper-left subblocks of this tridiagonalization tend to

$$\text{tridiag}\begin{pmatrix} & \sqrt{d} & & \sqrt{d} & & \cdots & & \sqrt{d} & \\ 1 & & 1+d & & \cdots & & \cdots & & 1+d \\ & \sqrt{d} & & \sqrt{d} & & \cdots & & \sqrt{d} & \end{pmatrix}.$$

[9] As noted above, the reference here again establishes this result for a deterministic sequence of vectors.

This leads one to conjecture that this gives three-term recurrence coefficients for the polynomials orthogonal to the Marchenko–Pastur distribution [28] (in fact, this provides an alternate proof of this). Consider $q_n(x)$, $n = 0, 1, 2$ that satisfy, $q_0(x) = 1$,

$$\begin{aligned} xq_0(x) &= \sqrt{d}\, q_1(x) + q_0(x), \\ xq_n(x) &= \sqrt{d}\, q_{n+1}(x) + (1+d)q_n(x) + \sqrt{d}\, q_{n-1}(x), \qquad n \geq 1. \end{aligned}$$

Next, note that

$$y \mapsto 2y\sqrt{d} + 1 + d$$

maps $[-1, 1]$ to $[\lambda_-, \lambda_+]$. Set $x = 2y\sqrt{d} + 1 + d$, $\check{q}_n(y) = q_n(2y\sqrt{d} + 1 + d)$ and we find

$$\check{q}_1(y) = (2y + \sqrt{d})\check{q}_0(y), \qquad \check{q}_{n+1}(y) + \check{q}_{n-1}(y) = 2y\check{q}_n(y), \qquad n \geq 1.$$

From this it follows that

$$\check{q}_0(y) = U_0(y), \quad \check{q}_1(y) = U_1(y) + \sqrt{d}\, U_0(y),$$

and therefore, if we used the convention that $U_{-1}(y) = 0$

$$\check{q}_n(y) = U_n(y) + \sqrt{d}\, U_{n-1}(y), \qquad n \geq 0.$$

By explicitly calculating inner products, it was shown in [8] that

$$p_n(x; \mu_{\mathrm{MP}}) = \check{q}_n\left(\frac{x - 1 - d}{2\sqrt{d}}\right) = q_n(x).$$

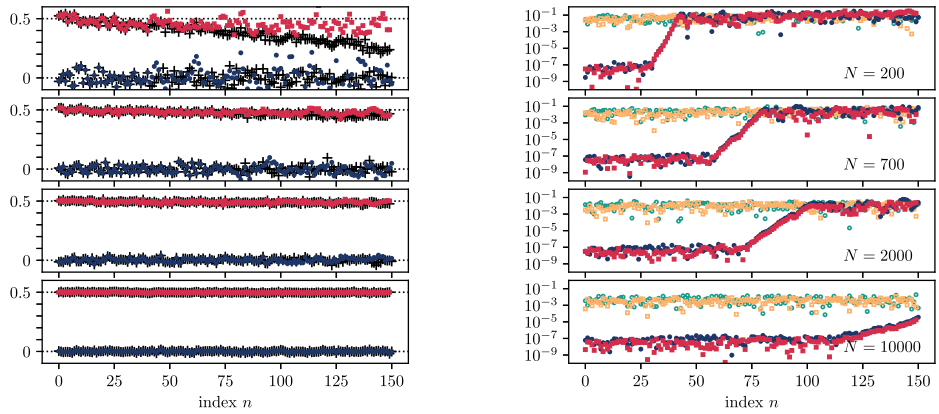
It can also be shown that $\mathfrak{c}(\lambda_+, \lambda_-) = \sqrt{d}$ is such that $\mathfrak{c}^n \pi_n(x; \mu_{\mathrm{MP}}) = p_n(x; \mu_{\mathrm{MP}})$. Therefore we obtain similar bounds on $\mathbf{M}_n(x, \mu_{\mathrm{MP}})$ as we did for $\mathbf{M}_n(x, \mu_U)$. Then (5.5) holds in the same way as for Wigner matrices.

6. Examples

In this section, we provide numerical experiments for several examples to which our analysis can be applied.

6.1. Wigner matrices

In Fig. 4, we show plots akin to those in Fig. 1 for several values of N. In particular, we take $\mathbf{A}_N$ as a random matrix from the Gaussian Orthogonal Ensemble (5.4) and $\mathbf{b}_N$ as a vector independent of $\mathbf{A}_N$. As expected, as N increases so that μ_N becomes nicer, the Lanczos algorithm remains forward stable for more iterations. The number of iterations



(a) *Recurrence coefficients* $\overline{\alpha}_n$ (•) *and* $\overline{\beta}_n$ (▪). *Exact arithmetic counterparts shown as pluses* (+) *and limiting values shown as dotted lines* (·····).

(b) *Forward error of recurrence coefficients* $|\alpha_n - \overline{\alpha}_n|$ (•) *and* $|\beta_n - \overline{\beta}_n|$ (▪) *and distance to limiting values* $|0 - \overline{\alpha}_n|$ (◦) *and* $|1/2 - \overline{\beta}_n|$ (▫).

Fig. 4. Output of Lanczos run on $(\mathbf{A}_N, \mathbf{b}_N)$ in single precision arithmetic, where $\mathbf{A}_N$ is a GOE matrix of size N and $\mathbf{b}_N$ is an independent vector; see Fig. 1 for more details.

for which it remains forward stable grows sublinearly with respect to N. When n is small enough relative to N, the finite precision coefficients $\overline{\alpha}_n$ and $\overline{\beta}_n$ are much closer to their exact arithmetic counterparts than to the limiting values. This behavior is suggested by Theorem 4.7.

6.2. Solving random linear systems

The mathematical behavior of a number of Krylov subspace methods used to solve systems involving random matrices have been studied rigorously. Such algorithms include conjugate gradient and MINRES [8,9,39], (accelerated) gradient descent [37], Neumann series iteration [47], and GMRES [3]. The most basic result of these analyses is that the macroscopic behavior of the algorithms, such as the error at step k, often becomes nearly deterministic when the random matrix is sufficiently large. That is, the error at step k converges to some fixed deterministic value when the random matrix becomes large. Since Lanczos-based methods such as conjugate gradient and MINRES are, in general, very susceptible to the impacts of floating point arithmetic [19] one may wonder the extent to which analyses such as [8,9,39] hold in finite precision arithmetic.

In exact arithmetic, assuming $\mathbf{A}$ is positive definite, the iterate $\mathbf{x}_k := \mathbf{Q}_k(\mathbf{T}_k)^{-1}\mathbf{e}_0$ is mathematically equivalent to the iterate produced by the conjugate gradient algorithm [20] used to solve $\mathbf{Ax} = \mathbf{b}$. This implies $\mathbf{x}_k$ is the optimal Krylov subspace approximation to $\mathbf{A}^{-1}\mathbf{b}$ in the $\mathbf{A}$-norm. When $\mathbf{A}$ is a sample covariance matrix of the form described in Section 5.2, [39] shows that (under certain moment conditions),

$$\|\mathbf{A}^{-1}\mathbf{b} - \mathbf{x}_k\|_{\mathbf{A}} \to \frac{d^{k/2}}{1-d} \text{ in probability as } N \to \infty \text{ and } N/M \to d. \tag{6.1}$$

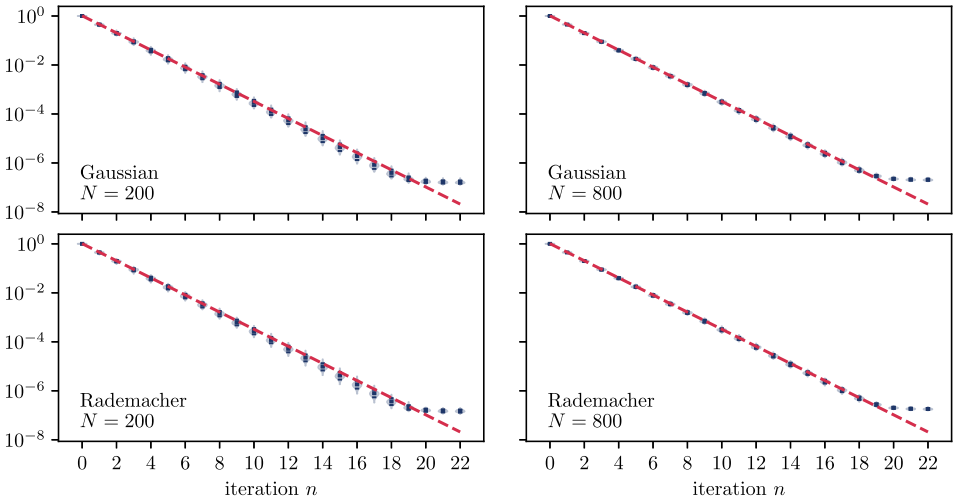


Fig. 5. Error of Lanczos used to solve the system $\mathbf{Ax} = \mathbf{b}$ in single precision floating point arithmetic. Here $\mathbf{b}_N$ is proportional to the all ones vector and $\mathbf{A}_N = N^{-1}\mathbf{X}\mathbf{X}^\mathsf{T}$ where the entries of $\mathbf{X}$ are either iid standard normal random variables or iid Rademacher random variables (± 1 with equal probability). For each n, N, the violin plot gives the distribution of error and the 5%, 50%, and 95% quantiles are marked. Notice the convergence to the "deterministic" behavior (- - -) as N increases, at least until the maximal accuracy is reached.

In finite precision arithmetic, the iterate

$$\overline{\mathbf{x}}_k := \overline{\mathbf{Q}}_k(\overline{\mathbf{T}}_k)^{-1}\mathbf{e}_0 \tag{6.2}$$

is, in general, no longer optimal. However, the analysis in this paper can be applied to the iterate $\overline{\mathbf{x}}_k$, at least assuming it is computed exactly from the quantities $\overline{\mathbf{Q}}_k$ and $\overline{\mathbf{T}}_k$.[10]

We perform a numerical experiment with sample covariance matrices $\mathbf{A}_N = N^{-1}\mathbf{X}\mathbf{X}^\mathsf{T}$, where the entries of $\mathbf{X}$ are either iid standard normal random variables or iid Rademacher random variables. In particular, we generate matrices for each of these distributions at $d = 0.2$ and $N = 200$ or $N = 800$ and set $\mathbf{b}_N$ proportional to the all ones vector. We then run Lanczos on $(\mathbf{A}_N, \mathbf{b}_N)$ in single precision arithmetic to get $\overline{\mathbf{Q}}_k$ and $\overline{\mathbf{T}}_k$. We compute $\overline{\mathbf{x}}_k$ using a standard linear system solver from numpy in double precision arithmetic.

The results of 1000 repetitions of each of these experiments are reported in Fig. 5. As expected, at least until convergence stagnates, the error of the algorithm concentrates around the estimate (6.1) as N increases. The error $\|\mathbf{A}^{-1}\mathbf{b} - \overline{\mathbf{x}}_k\|_\mathbf{A}$ stagnates around the machine precision, which is essentially all we could hope for given that $\overline{\mathbf{T}}_k$ is computed in single precision floating point arithmetic.

[10] A full analysis of the standard CG implementation [20] is well beyond the scope of this paper. In fact, to the best of our knowledge, it is not even known rigorously whether Paige's analysis extends to such an implementation. In practice the finite precision arithmetic behavior of (6.2) and the standard CG implementation are quite similar, so it is common to analyze (6.2) [18,41].

7. Deferred proofs

7.1. Proofs of bounds for regular measures

Proof of Lemma 3.3. The proof follows [14, Lemma 6]. Fix $n \leq 2k-1$. As a special case of the Markov brothers' inequality for polynomials [40, Theorem 1.10], the derivative $p_n'(\cdot\,;\mu)$ of $p_n(\cdot\,;\mu)$ satisfies

$$\|p_n'(\cdot\,;\mu)\|_{[a,b]} \leq \frac{2n^2}{b-a}\|p_n(\cdot\,;\mu)\|_{[a,b]}.$$

Let x^* be such that $|\,p_n(x^*;\mu)| = \|p_n(\cdot\,;\mu)\|_{[a,b]}$ and define

$$\mathcal{B} := \left\{x \in [a,b] : |x-x^*| \leq \frac{b-a}{4n^2}\right\}.$$

Using the triangle inequality, for any $x \in \mathcal{B}$,

$$\begin{aligned}
|\,p_n(x;\mu)| &= \left|p_n(x^*;\mu) + \int_{x^*}^{x} p_n'(y;\mu)\,\mathrm{d}y\right| \\
&\geq |\,p_n(x^*;\mu)| - |x-x^*|\|p_n'(\cdot\,;\mu)\|_{[a,b]} \\
&\geq |\,p_n(x^*;\mu)| - \frac{b-a}{4n^2}\frac{2n^2}{b-a}|p_n(x^*;\mu)| \\
&\geq \frac{1}{2}|\,p_n(x^*;\mu)|.
\end{aligned}$$

Both endpoints of $\mathcal{B}$ must be in $[a,b]$, so $\max(\mathcal{B})-\min(\mathcal{B}) \geq (b-a)/(4n^2) \geq (b-a)/(16k^2)$ and hence $\mu(\mathcal{B}) \geq K$. Using this and the fact that $p_n(\cdot\,;\mu)$ is normalized,

$$1 = \int_a^b p_n(x;\mu)^2\mu(\mathrm{d}x) \geq \int_{\mathcal{B}} p_n(x;\mu)^2\mu(\mathrm{d}x) \geq \frac{1}{4}p_n(x^*;\mu)^2\mu(\mathcal{B}) \geq \frac{1}{4}p_n(x^*;\mu)^2K.$$

Solving for $|\,p_n(x^*;\mu)|$ we find

$$\|p_n(\cdot\,;\mu)\|_{[a,b]} = |\,p_n(x^*;\mu)| \leq \frac{2}{\sqrt{K}}.$$

Since this holds for all $n \leq 2k-1$, the lemma follows. □

Proof of Corollary 3.5. Using the triangle inequality, for any x, y,

$$\begin{aligned}
&|\mu_N([x,y]) - \mu_\infty([x,y])| \\
&\quad = |\mu_N((-\infty,y]) - \mu_N((-\infty,x)) - \mu_\infty((-\infty,y]) + \mu_\infty((-\infty,x))|
\end{aligned}$$

$$\leq |\mu_N((-\infty, y]) - \mu_\infty((-\infty, y])| + |\mu_N((-\infty, x)) - \mu_\infty((-\infty, x))|$$
$$\leq 2d_{\mathrm{KS}}(\mu_N, \mu_\infty).$$

Thus, one easily verifies that

$$\mu_N([x, y]) \geq \mu_\infty([x, y]) - 2d_{\mathrm{KS}}(\mu_N, \mu_\infty).$$

Suppose $x, y \in [a', b']$ and $|x - y| \geq (b - a)/(16k^2)$. Then the length of $[x, y] \cap [a, b]$ is at least $(b - a)/(32k^2)$, so using (3.2),

$$\mu_\infty([x, y]) \geq L\left(\frac{b - a}{32k^2}\right)^\gamma.$$

Using this, our assumption on k, (3.3), and the fact $b' - a' \geq b - a$, we obtain a bound

$$\mu_N([x, y]) \geq L\left(\frac{b - a}{32k^2}\right)^\gamma - 2N^{-\alpha} \geq N^{-\alpha}, \qquad \forall x, y \in [a', b'] : |x - y| \geq (b' - a')/(16k^2).$$

Lemma 3.3 with $K = N^{-\alpha}$ then gives a bound for the orthogonal polynomials p_n for $\mu = \mu_N$:

$$M_k(\mu_N; [a', b']) \leq \frac{2}{\sqrt{N^{-\alpha}}} \leq 2\sqrt{\frac{3}{L}}\left(\frac{32}{b - a}\right)^{\gamma/2} k^\gamma.$$

Clearly $2\sqrt{3} < 4$, so the result is established. □

Proof of Corollary 3.6. Fix $n \leq 2k - 1$. Given (3.2), we can apply Lemma 3.3 with $K = L((b - a)/(16k^2))^\gamma$ to get a bound

$$M_k(\mu_\infty; [a, b]) \leq \frac{2}{\sqrt{L}}\left(\frac{16k^2}{b - a}\right)^{\gamma/2}.$$

This gives the first part of the lemma.

The Markov brothers' inequality for polynomials implies

$$\|p_n'(\cdot\,; \mu_\infty)\|_{[a', b']} \leq \frac{2n^2}{b' - a'}\|p_n(\cdot\,; \mu_\infty)\|_{[a', b']}.$$

Thus, since the supports of μ_N and μ_∞ are contained in $[a', b']$, integrating by parts,

$$|\,\mathfrak{m}_n(\mu_N; \mu_\infty) - \mathfrak{m}_n(\mu_\infty; \mu_\infty)| = \left|\int_{a'}^{b'} p_n(x; \mu_\infty)\big(\mu_N(\mathrm{d}x) - \mu_\infty(\mathrm{d}x)\big)\right|$$

$$\begin{aligned}
&\leq \int_{a'}^{b'} |\, p_n'(x;\mu_\infty)| \big|\mu_N((-\infty,x]) - \mu_\infty((-\infty,x])\big|\,\mathrm{d}x \\
&\leq (b'-a')\|p_n'(\cdot\,;\mu_\infty)\|_{[a',b']} d_{\mathrm{KS}}(\mu_N,\mu_\infty) \\
&\leq 2n^2\|p_n(\cdot\,;\mu_\infty)\|_{[a',b']} d_{\mathrm{KS}}(\mu_N,\mu_\infty).
\end{aligned}$$

Since $n \leq 2k-1$, with $\eta = 1/(16k^2) \leq 1/(2(2k)^2) \leq 1/(2n^2)$, Lemma 3.7 (which is independent of this result) yields the bound

$$\begin{aligned}
\|p_n(\cdot\,;\mu_\infty)\|_{[a',b']} &= \max_{x\in[-1-\eta,1+\eta]} \left| p_n\left(\frac{a+b}{2} + \frac{b-a}{2}x;\mu_\infty\right)\right| \\
&\leq 2 \max_{x\in[-1,1]} \left| p_n\left(\frac{a+b}{2} + \frac{b-a}{2}x;\mu_\infty\right)\right| \\
&= 2\|p_n(\cdot\,;\mu_\infty)\|_{[a,b]} \leq 2M_k(\mu_\infty;[a,b]).
\end{aligned}$$

Since $n \leq 2k$, using (3.3) and our bound on $M_k(\mu_\infty;[a,b])$, we obtain a bound for the modified moments

$$\begin{aligned}
|\, \mathfrak{m}_n(\mu_N;\mu_\infty) - \mathfrak{m}_n(\mu_\infty;\mu_\infty)| &\leq 2(2k)^2 2M_k(\mu_\infty;[a,b]) d_{\mathrm{KS}}(\mu_N,\mu_\infty) \\
&\leq 16k^2 \frac{2}{\sqrt{L}} \left(\frac{16k^2}{b-a}\right)^{\gamma/2} N^{-\alpha}.
\end{aligned}$$

Thus, using our assumption on k and that $n \leq 2k-1$,

$$\mathfrak{m}_n^{\triangle}(\mu_N,\mu_\infty;\mu_\infty) = \max_{n\leq 2k-1} |\mathfrak{m}_n(\mu_N;\mu_\infty) - \mathfrak{m}_n(\mu_\infty;\mu_\infty)| \leq c.$$

The lemma is established. □

7.2. Other proofs

Proof of Lemma 3.7. It is well-known that for any $x \in \mathbb{R}\setminus[-1,1]$,

$$|p(x)| \leq |T_n(x)|, \qquad \forall p : \deg(p) \leq n, \|p\|_{[-1,1]} \leq 1.$$

Thus, it suffices to show

$$|T_n(x)| \leq 2, \qquad x \in [-1-1/(2n^2), 1+1/(2n^2)].$$

We will in fact show $|T_n(z)| \leq 2$ for all $z \in E$, where $E := \{(w+w^{-1})/2 : 1 \leq |w| \leq r+\sqrt{r^2-1}\}$ is the Bernstein ellipse with rightmost point $r = 1+1/(2n^2)$.

Suppose $J(w) = \frac{1}{2}(w + w^{-1})$. Then, it is well known that, for any $n \geq 0$,

$$T_n(J(w)) = \frac{1}{2}\left(w^n + w^{-n}\right).$$

Let $z = J(w)$ be an arbitrary point on the boundary of the Bernstein ellipse E. Set $\rho = |w|$ so

$$|T_n(z)| \leq \frac{1}{2}\left(\rho^n + \rho^{-n}\right).$$

Suppose $n \geq 2$ and let $\rho = \rho(n) = 1 + \ln(2 + \sqrt{3})/n$. Then,

$$\lim_{n\to\infty} \frac{1}{2}\left(\rho(n)^n + \rho(n)^{-n}\right) = 2.$$

By direct computation, one verifies that

$$\frac{\mathrm{d}}{\mathrm{d}n}\left(\rho(n)^n + \rho(n)^{-n}\right) = \left(\rho(n)^n - \rho(n)^{-n}\right)\left(\ln(\rho(n)) - 1 + \rho(n)^{-1}\right).$$

We always have

$$\ln(\rho(n)) - 1 + \rho(n)^{-1} > 0, \qquad \rho(n) > 1.$$

Since $\rho(n)^n$ is monotonically increasing with n and $\rho(n)^{-n}$ is monotonically decreasing with n,

$$\rho(n)^n - \rho(n)^{-n} \geq \left(1 + \frac{\ln(2+\sqrt{3})}{2}\right)^2 - \left(1 + \frac{\ln(2+\sqrt{3})}{2}\right)^{-2} > 0.$$

Thus, the convergence of $(\rho(n)^n + \rho(n)^{-n})/2$ to 2 is monotonic from below. This implies $|T_n(z)| \leq 2$ for $z \in E_\rho$ and $n \geq 2$.

Now, note that

$$\begin{aligned}\frac{1}{2}\left(\rho(n) + \rho(n)^{-1}\right) &= 1 + \frac{1}{2n^2}\left(\ln(2+\sqrt{3})^2 - \frac{\ln(2+\sqrt{3})^3}{\ln(2+\sqrt{3}) + n}\right) \\ &\geq 1 + \frac{1}{2n^2}\left(\ln(2+\sqrt{3})^2 - \frac{\ln(2+\sqrt{3})^3}{\ln(2+\sqrt{3}) + 2}\right) \geq 1 + \frac{1}{2n^2}.\end{aligned}$$

This implies $E \subseteq E_\rho$ for $n \geq 2$. Clearly $T_0(z) = 1 \leq 2$ and $|T_1(z)| = |z| \leq 2$ for all $z \in E$. Thus, for all $n \geq 0$,

$$|T_n(z)| \leq 2, \qquad z \in E.$$

The result follows since $[-1 - 1/(2n^2), 1 + 1/(2n^2)] \subseteq E$. $\square$

Proof of Lemma 3.9. Suppose the lemma holds for $i < n$. Then,

$$\begin{aligned}
d_n(x) &= 2xd_{n-1}(x) - d_{n-2}(x) + 2f_{n-1} \\
&= 2x\left(U_{n-2}(x)f_0 + 2\sum_{i=2}^{n-1} U_{n-1-i}(x)f_{i-1}\right) - \left(U_{n-3}(x)f_0 + 2\sum_{i=2}^{n-2} U_{n-2-i}(x)f_{i-1}\right) \\
&\quad + 2f_{n-1} \\
&= \big(2xU_{n-2}(x) - U_{n-3}(x)\big)f_0 + 2\left(\sum_{i=2}^{n-2}\big(2xU_{n-1-i}(x) - U_{n-2-i}(x)\big)f_{i-1}\right) \\
&\quad + 4xf_{n-2} + 2f_{n-1} \\
&= U_{n-1}(x)f_0 + 2\left(\sum_{i=2}^{n-2} U_{n-i}(x)f_{i-1}\right) + 2U_1(x)f_{n-2} + 2U_0(x)f_{n-1} \\
&= U_{n-1}(x)f_0 + 2\sum_{i=2}^{n} U_{n-i}(x)f_{i-1}.
\end{aligned}$$

The result follows as the base case is assumed. □

Declaration of competing interest

We declare no competing interests.

Data availability

No data was used for the research described in the article.

References

[1] A. Böttcher, Y.I. Karlovich, Carleson Curves, Muckenhoupt Weights, and Toeplitz Operators, Birkhäuser Basel, Basel, 1997, ISBN: 978-3-0348-9828-7.

[2] Z.D. Bai, J.W. Silverstein, No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices, Ann. Probab. 26 (1) (Jan. 1998), https://doi.org/10.1214/aop/1022855421.

[3] T. Chen, A. Greenbaum, T. Trogdon, GMRES, pseudospectra, and Crouzeix's conjecture for shifted and scaled Ginibre matrices, https://arxiv.org/abs/2303.02042, in preparation, 2023.

[4] E. Carson, J. Liesen, Z. Strakoš, 70 years of Krylov subspace methods: the journey continues, arXiv:2211.00953 [math.NA], 2022.

[5] I. Dumitriu, A. Edelman, Matrix models for beta ensembles, J. Math. Phys. 43 (11) (Oct. 2002) 5830, ISSN: 00222488, https://doi.org/10.1063/1.1507823.

[6] P. Deift, Orthogonal Polynomials and Random Matrices: a Riemann-Hilbert Approach, Amer. Math. Soc., Providence, RI, 2000, p. 257.

[7] V.L. Druskin, L.A. Knizhnerman, Error bounds in the simple Lanczos procedure for computing functions of symmetric matrices and eigenvalues, Comput. Math. Math. Phys. 31 (7) (July 1991) 20–30, ISSN: 0965-5425.

[8] P. Deift, T. Trogdon, The conjugate gradient algorithm on well-conditioned Wishart matrices is almost deterministic, Q. Appl. Math. 79 (1) (July 2020) 125–161, https://doi.org/10.1090/qam/1574.

[9] X. Ding, T. Trogdon, A Riemann–Hilbert approach to the perturbation theory for orthogonal polynomials: applications to numerical linear algebra and random matrix theory, arXiv:2112.12354 [math.PR], 2021.
[10] P. Duren, Theory of H^p Spaces, Academic Press, 1970.
[11] A. Edelman, N.R. Rao, Random matrix theory, Acta Numer. 14 (Apr. 2005) 233–297, https://doi.org/10.1017/s0962492904000236.
[12] L. Erdős, H.-T. Yau, Dynamical Approach to Random Matrix Theory, Amer. Math. Soc., Providence, RI, 2017, pp. 1–226, ISBN: 0049-1748.
[13] A.S. Fokas, A.R. Its, A.V. Kitaev, The isomonodromy approach to matrix models in 2D quantum gravity, Commun. Math. Phys. 147 (2) (1992) 395–430, ISSN: 1432-0916.
[14] H.-J. Fischer, On the condition of orthogonal polynomials via modified moments, Z. Anal. Anwend. 15 (1) (1996) 223–244, https://doi.org/10.4171/zaa/696.
[15] W. Gautschi, Orthogonal Polynomials: Computation and Approximation. Numerical Mathematics and Scientific Computation, Oxford University Press, ISBN 9780198506720, 2004.
[16] W. Gautschi, On generating orthogonal polynomials, SIAM J. Sci. Stat. Comput. 3 (3) (Sept. 1982) 289–317, https://doi.org/10.1137/0903018.
[17] G.H. Golub, D.P. O'Leary, Some history of the conjugate gradient and Lanczos algorithms: 1948–1976, SIAM Rev. 31 (1) (Mar. 1989) 50–102, https://doi.org/10.1137/1031003.
[18] A. Greenbaum, Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences, Linear Algebra Appl. 113 (1989) 7–63, https://doi.org/10.1016/0024-3795(89)90285-1, ISSN: 0024-3795.
[19] A. Greenbaum, Iterative Methods for Solving Linear Systems, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, ISBN 0-89871-396-X, 1997.
[20] M.R. Hestenes, E. Stiefel, Methods of Conjugate Gradients for Solving Linear Systems, vol. 49, NBS, Washington, DC, 1952.
[21] L.A. Knizhnerman, The simple Lanczos procedure: estimates of the error of the Gauss quadrature formula and their applications, Comput. Math. Math. Phys. 36 (11) (Jan. 1996) 1481–1492, ISSN: 0965-5425.
[22] A.B.J. Kuijlaars, K.T.-R. McLaughlin, W. Van Assche, M. Vanlessen, The Riemann–Hilbert approach to strong asymptotics for orthogonal polynomials on $[-1, 1]$, Adv. Math. 188 (2) (2004) 337–398.
[23] A.B.J. Kuijlaars, Riemann-Hilbert Analysis for Orthogonal Polynomials, 2003, pp. 167–210.
[24] A. Knowles, J. Yin, Anisotropic local laws for random matrices, Probab. Theory Relat. Fields 169 (1–2) (Oct. 2017) 257–352, https://doi.org/10.1007/s00440-016-0730-4, ISSN: 0178-8051.
[25] C. Lanczos, An iteration method for the solution of the eigenvalue problem of linear differential and integral operators, J. Res. Natl. Bur. Stand. 45 (1950) 255–282.
[26] G. Meurant, The Lanczos and Conjugate Gradient Algorithms, Society for Industrial and Applied Mathematics, 2006, eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9780898718140.
[27] C. Musco, C. Musco, A. Sidford, Stability of the Lanczos method for matrix function approximation, Soc. Ind. Appl. Math. (Jan. 2018) 1605–1624, https://doi.org/10.1137/1.9781611975031.105.
[28] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR Sb. 1 (4) (1967) 457–483, https://doi.org/10.1070/SM1967v001n04ABEH001994, ISSN: 0025-5734.
[29] D.P. O'Leary, Z. Strakoš, P. Tichý, On sensitivity of Gauss–Christoffel quadrature, Numer. Math. 107 (1) (Apr. 2007) 147–174, https://doi.org/10.1007/s00211-007-0078-x.
[30] S. Olver, T. Trogdon, Numerical solution of Riemann–Hilbert problems: random matrix theory and orthogonal polynomials, Constr. Approx. 39 (1) (Dec. 2013) 101–149, https://doi.org/10.1007/s00365-013-9221-3, ISSN: 0176-4276.
[31] C.C. Paige, An augmented stability result for the Lanczos Hermitian matrix tridiagonalization process, SIAM J. Matrix Anal. Appl. 31 (5) (2010) 2347–2359, https://doi.org/10.1137/090761343.
[32] C.C. Paige, Accuracy of the Lanczos process for the eigenproblem and solution of equations, SIAM J. Matrix Anal. Appl. 40 (4) (2019) 1371–1398, https://doi.org/10.1137/17M1133725.
[33] C.C. Paige, Practical use of the symmetric Lanczos process with re-orthogonalization, BIT Numer. Math. 10 (2) (June 1970) 183–195, https://doi.org/10.1007/bf01936866.
[34] C.C. Paige, The computation of eigenvalues and eigenvectors of very large sparse matrices, PhD thesis, University of London, 1971, eprint: https://www.cs.mcgill.ca/~chris/pubClassic/PaigeThesis.pdf.
[35] C.C. Paige, Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix, IMA J. Appl. Math. 18 (3) (Dec. 1976) 341–349, https://doi.org/10.1093/imamat/18.3.341, ISSN: 0272-4960.

[36] C.C. Paige, Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem, Linear Algebra Appl. 34 (1980) 235–258, https://doi.org/10.1016/0024-3795(80)90167-6, ISSN: 0024-3795.
[37] C. Paquette, B. van Merriënboer, E. Paquette, F. Pedregosa, Halting time is predictable for large models: a universality property and average-case analysis, Found. Comput. Math. (Feb. 2022), https://doi.org/10.1007/s10208-022-09554-y.
[38] B.N. Parlett, The Symmetric Eigenvalue Problem, Society for Industrial and Applied Mathematics, Jan. 1998.
[39] E. Paquette, T. Trogdon, Universality for the conjugate gradient and MINRES algorithms on sample covariance matrices, Commun. Pure Appl. Math. (Sept. 2022), https://doi.org/10.1002/cpa.22081.
[40] T.J. Rivlin, An Introduction to the Approximation of Functions, Unabridged and corr. republication of the 1969 ed. Dover books on advanced mathematics, ISBN 9780486640693, 1981.
[41] Z. Strakos, A. Greenbaum, Open Questions in the Convergence Analysis of the Lanczos Process for the Real Symmetric Eigenvalue Problem, University of Minnesota, 1992, eprint: https://conservancy.umn.edu/handle/11299/1838.
[42] J.W. Silverstein, Eigenvalues and eigenvectors of large dimensional sample covariance matrices, Contemp. Math. 50 (1986).
[43] H.F. Trotter, Eigenvalue distributions of large Hermitian matrices; Wigner's semi-circle law and a theorem of Kac, Murdock, and Szegő, Adv. Math. 54 (1) (1984) 67–82, ISSN: 10902082, https://doi.org/10.1016/0001-8708(84)90037-9.
[44] A. Townsend, T. Trogdon, S. Olver, Fast computation of Gauss quadrature nodes and weights on the whole real line, IMA J. Numer. Anal. 36 (1) (Oct. 2014) 337–358, https://doi.org/10.1093/imanum/drv002, ISSN: 14643642, arXiv:1410.5286.
[45] J. Wishart, The generalised product moment distribution in samples from a normal multivariate population, Biometrika 20A (1–2) (1928) 32–52, https://doi.org/10.1093/biomet/20A.1-2.32, ISSN: 0006-3444.
[46] M. Webb, S. Olver, Spectra of Jacobi operators via connection coefficient matrices, Commun. Math. Phys. 382 (2) (Feb. 2021) 657–707, https://doi.org/10.1007/s00220-021-03939-w.
[47] Y. Zhang, T. Trogdon, A probabilistic analysis of the Neumann series iteration, Minn. J. Undergrad. Math. 7 (1) (May 2022), arXiv:1909.07506 [math.PR].