SIAM J. MATRIX ANAL. APPL.
Vol. 43, No. 3, pp. 1263–1290
© 2022 Society for Industrial and Applied Mathematics

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

A THEORY OF QUANTUM SUBSPACE DIAGONALIZATION*

ETHAN N. EPPERLY†, LIN LIN‡, AND YUJI NAKATSUKASA§

Abstract. Quantum subspace diagonalization methods are an exciting new class of algorithms for solving large-scale eigenvalue problems using quantum computers. Unfortunately, these methods require the solution of an ill-conditioned generalized eigenvalue problem, with a matrix pencil corrupted by a nonnegligible amount of noise that is far above the machine precision. Despite pessimistic predictions from classical worst-case perturbation theories, these methods can perform reliably well if the generalized eigenvalue problem is solved using a standard truncation strategy. By leveraging and advancing classical results in matrix perturbation theory, we provide a theoretical analysis of this surprising phenomenon, proving that under certain natural conditions, a quantum subspace diagonalization algorithm can accurately compute the smallest eigenvalue of a large Hermitian matrix. We give numerical experiments demonstrating the effectiveness of the theory and providing practical guidance for the choice of truncation level. Our new results can also be of independent interest to solving eigenvalue problems outside the context of quantum computation.

Key words. quantum subspace diagonalization, quantum linear algebra, generalized eigenvalue problem, matrix perturbation theory

MSC codes. 68Q12, 65F15, 15A22, 15A45

DOI. 10.1137/21M145954X

1. Introduction.

Quantum computing is a fundamentally new computational paradigm, which has the potential to have a transformative impact on certain areas of computational science [27, 28]. One particularly compelling use case for quantum computers is to solve eigenvalue problems related to quantum many-body systems, for which the dimension of the discretized matrix grows exponentially with respect to the number of particles.

Quantum subspace diagonalization (QSD) methods [6, 13, 15, 20, 21, 26, 31, 32], also known as quantum Krylov methods, are an exciting class of quantum algorithms for solving large-scale Hermitian eigenvalue problems. One common key step of these algorithms is to solve a nearly singular generalized eigenvalue problem, where each entry of the associated matrix pair can be corrupted by Monte Carlo errors many orders of magnitude larger than the round-off error typically seen in classical computation. For such noisy generalized eigenvalue problems, classical perturbation theory fails to explain why such problems could be solved accurately. Despite this, QSD methods appear to work in practice, at least on some examples and with some procedure to compensate for the measurement error. This article is addressed squarely at

*Received by the editors November 15, 2021; accepted for publication (in revised form) April 22, 2022; published electronically August 1, 2022.
https://doi.org/10.1137/21M145954X
Funding: The work of the first author was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship grant DE-SC0021110. The second author was supported by the Department of Energy grant DE-SC0017867 and the NSF Quantum Leap Challenge Institute (QLCI) program through grant OMA-2016245. The second author is a Simons investigator.

†Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125 USA (eepperly@caltech.edu).
‡Department of Mathematics, and Challenge Institute of Quantum Computation, University of California Berkeley, Berkeley, CA 94720 USA, and Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA (linlin@math.berkeley.edu).
§Mathematical Institute, Oxford University, Oxford, UK (nakatsukasa@maths.ox.ac.uk).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

explaining why, in theory, QSD algorithms perform well and how, in practice, errors in the problem data can be effectively dealt with.

While our analysis is centered around the context of QSD, the underlying problem of solving an ill-conditioned generalized eigenvalue problem with noisy data is a fundamental linear algebra problem, which emerges in multiple application areas such as electronic structure theory [3, 14], control theory [9], and variational Monte Carlo optimization [30]. Our results may be of particular interest in filter diagonalization methods [18, 36] which were a classical antecedent to QSD methods [26] and also require the solution of an ill-conditioned generalized eigenvalue problem.

Notation. Vectors and matrices will be denoted by boldface lower- and uppercase letters, respectively. We denote the conjugate transpose by *. All matrices and vectors are assumed to be over the complex numbers. The unembellished norm $\|\cdot\|$ shall refer to the Euclidean norm of a vector $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^*\boldsymbol{x}}$ or the spectral norm (largest singular value) of a matrix. At times, we shall also make use of the Frobenius norm $\|\boldsymbol{B}\|_{\mathrm{F}} := \sqrt{\mathrm{tr}(\boldsymbol{B}^*\boldsymbol{B})}$. The absolute values of the generalized eigenvalues of a pair $(\boldsymbol{H}, \boldsymbol{S})$ are denoted $|\Lambda(\boldsymbol{H}, \boldsymbol{S})|$. Relations $\approx$, $\lessapprox$, $\ll$, etc., are informal, with no precise mathematical relation being implied.

1.1. Quantum subspace diagonalization and its numerical challenges. We begin by describing a simple QSD algorithm developed in parallel by Parrish and McMahon [26] and Stair, Huang, and Evangelista [32] with a focus on its linear algebraic (Krylov) structure. For readers unfamiliar with quantum computation and why it might be advantageous over classical computing for certain problems, we recommend the classic textbook [24] as well as a recent tutorial aimed at a mathematical audience [22].

Suppose we are interested in the ground-state energy[1] (least eigenvalue) E_0 of a Hamiltonian operator (Hermitian matrix) $\widehat{\boldsymbol{H}}$. Our goal shall be to compute an approximation $\widetilde{E}_0$ for which the (forward) error $|\widetilde{E}_0 - E_0|$ is small. For the problem sizes for which QSD is appropriate, vectors with the dimension of the operator $\widehat{\boldsymbol{H}}$ are too large to store classically, so we would like to represent them as states in a quantum computer. The QSD method rests on two assumptions:

1. We can efficiently prepare a state $\boldsymbol{\varphi}_0$ on the quantum device that has a nontrivial overlap $|\boldsymbol{\varphi}_0^*\boldsymbol{\psi}_0| \gg 0$ to the true ground-state eigenvector $\boldsymbol{\psi}_0$.
2. The time evolution $\boldsymbol{\varphi} \mapsto \mathrm{e}^{\mathrm{i}t\widehat{\boldsymbol{H}}}\boldsymbol{\varphi}$ can be efficiently simulated on the quantum device.

To improve on our initial guess $\boldsymbol{\varphi}_0$, we enlarge to a subspace spanned by vectors

$$\boldsymbol{\varphi}_j = \mathrm{e}^{\mathrm{i}t_j\widehat{\boldsymbol{H}}} \quad \text{for } j = 0, \dots, n-1, \tag{1.1}$$

where $t_j = j\Delta t$ are a time sequence with step size $\Delta t > 0$. The subspace $\mathrm{span}\{\boldsymbol{\varphi}_j\}$ forms a "unitary Krylov space" and plays a role analogous to the Krylov subspace in the Lanczos method. Eigenvalue estimates for the operator $\widehat{\boldsymbol{H}}$ can be computed from this unitary Krylov subspace by applying the Rayleigh–Ritz method [25, sect. 11].

In a classical Krylov method, one would usually orthogonalize the basis vectors (1.1). Unfortunately, this orthogonalization operation can be inherently difficult to perform on a quantum computer, so we instead work in the basis (1.1) as computed. In a nonorthogonal basis, the Rayleigh–Ritz eigenvalue estimates are obtained by solving a generalized eigenvalue problem

[1] The ground-state energy, by itself, is a useful quantity for applications in electronic structure, as it determines the energy landscape for dynamical simulation.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\boldsymbol{H}\boldsymbol{c} = E\,\boldsymbol{S}\boldsymbol{c}. \tag{1.2}$$

Here the *projected Hamiltonian* and *overlap* matrices $\boldsymbol{H}$ and $\boldsymbol{S}$ are Hermitian–Toeplitz matrices defined as

$$\boldsymbol{H}_{jk} = \boldsymbol{\varphi}_j^* \widehat{\boldsymbol{H}} \boldsymbol{\varphi}_k, \quad \boldsymbol{S}_{jk} = \boldsymbol{\varphi}_j^* \boldsymbol{\varphi}_k, \tag{1.3}$$

$\boldsymbol{c}$ is the reduced Ritz vector, and E is the Ritz value. Each matrix entry $\boldsymbol{H}_{jk}, \boldsymbol{S}_{jk}$ can be estimated via a Monte Carlo sampling procedure such as the Hadamard test on a quantum computer (see, e.g., [24, Chapt. 5], [6, App. D]). Once a sufficiently good estimate to $\boldsymbol{H}, \boldsymbol{S}$ is obtained, the generalized eigenvalue problem (1.2) is solved on a classical computer.

Remark 1.1 (other QSD methods). A number of QSD methods have been proposed which differ in how the basis states (1.1) are generated and how the eigenvalue estimates are obtained. Alternative methods for basis vector generation (1.1) from one or more initial guesses include multiplication by creation and annihilation operators [5, 20] and imaginary-time evolution $\boldsymbol{\varphi}_j := \mathrm{e}^{\mathrm{i}(\mathrm{i}t_j)\widehat{\boldsymbol{H}}}\boldsymbol{\varphi}_0$ [21] (with $t_j \geq 0$). Methods also differ in whether the Rayleigh–Ritz procedure is applied to $\widehat{\boldsymbol{H}}$ itself [26, 32] or the time evolution operator $\mathrm{e}^{\mathrm{i}\,\Delta t\,\widehat{\boldsymbol{H}}}$ [6, 15]. For concreteness, we shall focus in this article on the QSD method as discussed, though our analysis should have insights for understanding the broader class of QSD algorithms.

There are two main questions in the analysis of the (forward) error of the QSD method:

(A) How to analyze the Rayleigh–Ritz error due to the use of a finite-dimensional unitary Krylov subspace.

(B) How to analyze the error of the generalized eigenvalue problem (1.2) in the presence of the Monte Carlo noise for estimating the matrix entries in $\boldsymbol{H}, \boldsymbol{S}$.

Below we first discuss issues related to question (B), which are particularly challenging from the perspective of numerical linear algebra.

1.2. Numerical issues with QSD. Numerical results indicate that the size of n needed to obtain desired accuracy can be very modest (e.g., $10 \leq n \leq 100$), so we are free to use any algorithm [11, 25] to solve the dense generalized eigenvalue problem (1.2). However, it is frequently observed that the states $\boldsymbol{\varphi}_0, \ldots, \boldsymbol{\varphi}_{n-1}$ are very close to being linearly dependent, leading to the matrices $\boldsymbol{H}$ and $\boldsymbol{S}$ being nearly rank-deficient and the generalized eigenvalue problem (1.2) nearly singular. This ill-conditioning is an intrinsic feature of this method since the problem necessarily becomes ill-conditioned if the initial guess $\boldsymbol{\varphi}_0$ possesses the desirable property of approximately lying in a low-dimensional invariant subspace.

The near-singularity of the problem (1.2) becomes particularly alarming when taken in conjunction with the fact that the matrix elements (1.3) will be corrupted by several types of error when measured from a quantum computer. Some forms of error, such as discretization error in evaluating the time evolution $\mathrm{e}^{\mathrm{i}t_j\widehat{\boldsymbol{H}}}\boldsymbol{\varphi}_0$ by a Trotter formula [4] and gate errors, can in principle be systematically controlled on a fault-tolerant quantum device. However, even on a flawless quantum device, the matrix elements (1.3) still need to be computed via sampling, which incurs Monte Carlo–type $\approx \delta^{-2}$ samples to compute each entry to δ-accuracy. We shall refer to all of these errors collectively as "noise."

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Classical perturbation theory [34, sect. VI.3] as well as modern improvements [19] do not apply when the perturbation is large enough to make the problem (1.2) singular, which is almost always the case for the QSD algorithm because of sampling error and ill-conditioning. Indeed, the following variant of the classical example [39, eq. (4.10)] shows that a perturbation just a touch larger than the distance to singularity can send the eigenvalues (originally, 1 and 2) to any pair of numbers $\alpha, \beta \in \mathbb{C}$:

$$\boldsymbol{H} = \begin{bmatrix} 2 & 0 \\ 0 & \epsilon \end{bmatrix}, \boldsymbol{S} = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix} \xrightarrow{\mathcal{O}(\epsilon) \text{ error}} \widetilde{\boldsymbol{H}} = \begin{bmatrix} 2 & \alpha\epsilon \\ \beta\epsilon & 0 \end{bmatrix}, \widetilde{\boldsymbol{S}} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 0 \end{bmatrix}. \tag{1.4}$$

Even the well-conditioned eigenvalue 2 can be perturbed arbitrarily far if the noise is large enough to make the problem singular!

There is evidence that adversarially chosen perturbations such as (1.4) are pathologically unlikely to occur, with the well-conditioned eigenvalues of a pair $(\boldsymbol{H}, \boldsymbol{S})$ changing only modestly after perturbation. Indeed, Wilkinson showed that "most" $\mathcal{O}(\epsilon)$ perturbations to (1.4) have an eigenvalue near the well-conditioned eigenvalue of 2 [39], and recent analysis by Lotz and Noferini [16] showed that some eigenvalues of genuinely singular generalized eigenvalue problems can, in effect, be locally well-conditioned with high probability.

Even if a good approximation to the ground-state energy is among the eigenvalues of $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$, identifying it can be difficult. When the pair $(\boldsymbol{H}, \boldsymbol{S})$ is nearly singular, the perturbed problem is almost ensured to possess spurious eigenvalues. To see why this is the case, observe that if $(\boldsymbol{H}, \boldsymbol{S})$ is nearly singular, there exists an eigenpair $(\boldsymbol{c}, E)$ such that $\boldsymbol{H}\boldsymbol{c}, \boldsymbol{S}\boldsymbol{c} \approx \boldsymbol{0}$. Perturbations in $\boldsymbol{H}$ and $\boldsymbol{S}$ create large changes in both the numerator and the denominator of the Rayleigh quotient $E = \boldsymbol{c}^*\boldsymbol{H}\boldsymbol{c}/\boldsymbol{c}^*\boldsymbol{S}\boldsymbol{c}$; since eigenvalues of a pair with positive definite $\boldsymbol{S}$ extremize the Rayleigh quotient, noise can easily introduce fake eigenvalues much smaller than the genuine least eigenvalue of $(\boldsymbol{H}, \boldsymbol{S})$. Reliably distinguishing genuine eigenvalues from such fake eigenvalues is challenging; we tried many heuristics and all of them failed for a nontrivial fraction of random initializations of the measurement error (see section SM2).

To address these issues, we shall solve the eigenvalue problem (1.2) using the following truncation scheme: First, compute an eigendecomposition of the matrix $\boldsymbol{S}$ and discard all eigenvalues smaller than or equal to a threshold $\epsilon > 0$. Then, letting $\boldsymbol{V}_{>\epsilon}$ denote a matrix whose columns are the nondiscarded eigenvectors and $\boldsymbol{\Lambda}_{>\epsilon} := \boldsymbol{V}_{>\epsilon}^*\boldsymbol{S}\boldsymbol{V}_{>\epsilon}$, we solve the reduced generalized eigenvalue problem

$$\boldsymbol{V}_{>\epsilon}^*\boldsymbol{H}\boldsymbol{V}_{>\epsilon}\,\boldsymbol{c} = \widetilde{E}\,\boldsymbol{V}_{>\epsilon}^*\boldsymbol{S}\boldsymbol{V}_{>\epsilon}\boldsymbol{c},$$

or equivalently find the eigenvalues of $\boldsymbol{\Lambda}_{>\epsilon}^{-1/2}\boldsymbol{V}_{>\epsilon}^*\boldsymbol{S}\boldsymbol{V}_{>\epsilon}\boldsymbol{\Lambda}_{>\epsilon}^{-1/2}$. This procedure appears to have been first discovered in quantum physics by Löwdin in 1967 [17] (also rediscovered in [12]), where it is associated with the name *canonical orthogonalization*. We shall call this procedure *thresholding* and present it in Algorithm 1.1 for convenient reference in the rest of the document. A more careful variant of thresholding from the numerical analysis community was proposed by Fix and Heiberger in 1972 [10], though its authors expressly advise against using it in precisely the setting of the QSD algorithm where the pair $(\boldsymbol{H}, \boldsymbol{S})$ is nearly singular.

Despite appearing quite natural, there are examples where thresholding fails to work and is thus not appropriate for arbitrary Hermitian definite generalized eigenvalue problems. For instance, if one applies thresholding with parameter ϵ to the pair

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Algorithm 1.1 Thresholding procedure for solving a noise-corrupted or nearly singular generalized eigenvalue problem $\boldsymbol{H}\boldsymbol{c} = E\,\boldsymbol{H}\boldsymbol{c}$.

procedure THRESHOLDING($\boldsymbol{H}$, $\boldsymbol{S}$, ϵ)
  $(\boldsymbol{V}, \boldsymbol{D}) \leftarrow \texttt{eig}(\boldsymbol{S})$
  $I \leftarrow \{i : \boldsymbol{D}_{ii} > \epsilon\}$
  $\boldsymbol{V} \leftarrow \boldsymbol{V}(:, I)$
  return smallest eigenvalue of $(\boldsymbol{V}^*\boldsymbol{H}\boldsymbol{V}, \boldsymbol{V}^*\boldsymbol{S}\boldsymbol{V})$
end procedure

$$\boldsymbol{H} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & \epsilon^2 \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon^2 \end{bmatrix}, \tag{1.5}$$

one recovers an eigenvalue of 1, which is far from the genuine eigenvalues 0 and 2 of the pair. Even with the existence of bad examples like (1.5), thresholding appears to be quite reliable at filtering out noise and dealing with the ill-conditioning of the overlap matrix for QSD-derived pairs $(\boldsymbol{H}, \boldsymbol{S})$ in our experiments, with similar results being observed for an SVD-based truncation strategy in [15, sect. II.F]. Despite truncation strategies such as canonical orthogonalization and Fix–Heiberger having a 50-year history, we are unaware of any general theory of why these methods work.

1.3. Overview and main results. We aim to elucidate why the QSD algorithm works when combined with the thresholding procedure, undeterred by the presence of negative examples such as (1.4) and (1.5). Our main three results are as follows:

(i) In the absence of noise and with an appropriate choice of the time sequence, the QSD procedure with thresholding is accurate. This provides a positive answer to the question (A) in the error analysis of the QSD method due to the Krylov subspace approximation in the presence of thresholding.

(ii) The thresholded problem is stable under noise in the sense that the thresholded problem we solve from the noise-corrupted pair $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ is close to the unperturbed pair $(\boldsymbol{H}, \boldsymbol{S})$.

(iii) If the thresholded problem and its noisy perturbation are sufficiently close (as we establish with the previous result), the well-conditioned and well-separated eigenvalues of the thresholded problem are accurately computed in the presence of noise. Results (ii) and (iii) provide a positive answer to the question (B) in the error analysis of the QSD method due to the classical solution of the noisy generalized eigenvalue problem.

Together, these results paint a reasonably complete picture of why the QSD algorithm works in the presence of noise when thresholding is used. To our knowledge, this is the first work providing rigorous analysis of the theoretical efficacy of QSD-type algorithms, both in the noise-free and noisy settings.

Our first main result can be summarized informally as follows (a formal statement is presented as Theorem 2.7).

INFORMAL THEOREM 1.2. *Suppose the thresholding procedure (Algorithm* 1.1*) is applied to the perturbed pair* $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}}) = (\boldsymbol{H}+\boldsymbol{\Delta}_{\boldsymbol{H}}, \boldsymbol{S}+\boldsymbol{\Delta}_{\boldsymbol{S}})$*, which are Hermitian matrices of size* n*. Consider the thresholded matrix pair* $(\boldsymbol{A}, \boldsymbol{B}) := (\boldsymbol{V}_{>\epsilon}^*\boldsymbol{H}\boldsymbol{V}_{>\epsilon}, \boldsymbol{V}_{>\epsilon}^*\boldsymbol{S}\boldsymbol{V}_{>\epsilon})$ *and let* E_0 *be its least eigenvalue. Assume* E_0 *is sufficiently well-separated from other eigenvalues of* $(\boldsymbol{A}, \boldsymbol{B})$*, let* d_0^{-1} *denote the condition number of the* eigenangle $\tan^{-1} E_0$*, and suppose the perturbations* $\boldsymbol{\Delta}_{\boldsymbol{H}}$ *and* $\boldsymbol{\Delta}_{\boldsymbol{S}}$ *have spectral norm not exceeding* η*. There*

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

exists a constant $0 \le \alpha \le 1/2$ *such that the recovered eigenvalue* $\widetilde{E}_0$ *from the noise-perturbed pair* $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ *using threshold parameter* $\epsilon = \Theta(\eta^{\frac{1}{1+\alpha}})$ *satisfies the bound*

$$\left|\tan^{-1} \widetilde{E}_0^{\text{th}} - \tan^{-1} E_0^{\text{th}}\right| \le \mathcal{O}\left(d_0^{-1} \eta^{\frac{1}{1+\alpha}}\right). \tag{1.6}$$

The implicit constant in the $\mathcal{O}$*-notation depends on the eigenvalues of* $(\boldsymbol{H}, \boldsymbol{S})$*, the spectrum of* $\boldsymbol{S}$*, and* n*.*

Our result shows that, using the empirically observed value $\alpha = 1/4$ (see section SM6), we are able to recover the smallest eigenvalue of $(\boldsymbol{H}, \boldsymbol{S})$ (or more specifically its arctangent) with error proportional to $\eta^{4/5}$ times its condition number. Given examples (1.4) and (1.5) showing arbitrarily large errors for small perturbations of nearly singular generalized eigenvalue problems and thresholding, the fact that we are able to obtain any nontrivial error bounds for the QSD algorithm with thresholding may be regarded as surprising.

Our second main result provides an end-to-end bound for the QSD method.

INFORMAL THEOREM 1.3. *Let* ΔE_j *denote the difference between the* j*th smallest and the smallest eigenvalue* E_0 *of* $\widehat{\boldsymbol{H}}$ *and let* γ_0 *denote the inner product between the initial state* $\boldsymbol{\varphi}_0$ *and the true ground-state eigenvector of* $\widehat{\boldsymbol{H}}$*. Let* $(\boldsymbol{H}, \boldsymbol{S})$ *denote the noise-free output of the QSD algorithm for a particular choice of time step, and instate the notation and assumptions of Informal Theorem* 1.2*. Then*

$$\left|\tan^{-1} \widetilde{E}_0^{\text{th}} - \tan^{-1} E_0\right| \le \mathcal{O}\left(\frac{1 - |\gamma_0|^2}{|\gamma_0|^2} \mathrm{e}^{-n\mathcal{O}\left(\frac{\Delta E_1}{\Delta E_{N-1}}\right)} + \left[\frac{\Delta E_{N-1}}{|\gamma_0|^2} + d_0^{-1}\right] \eta^{\frac{1}{1+\alpha}}\right).$$

The remainder of this paper is organized as follows. For expository reasons, we present our main results in the reverse order outlined here. Section 2 discusses perturbation analysis for the thresholded problem, leading to a formalization, Theorem 2.7, of Informal Theorem 1.2 in section 2.3. Section 3 discusses Rayleigh–Ritz errors due to approximation by the finite-dimensional unitary Krylov subspace and thresholding procedure. We then present additional results in section 4 which are independent of the rest of the presentation. We draw particular attention to Theorem 4.2, which shows that thresholding applied to a general pair $(\boldsymbol{H}, \boldsymbol{S})$ recovers the least eigenvalue accurately if it is well-conditioned. This does not contradict the bad example (1.5) since both its eigenvalues are ill-conditioned with condition numbers $\Theta(\epsilon^{-1})$. We conclude with numerical experiments (section 5) and conclusions (section 6).

2. Perturbation analysis for the thresholded problem. In this section, we analyze the effects of noise on the solution of the generalized eigenvalue problem (1.2) using the thresholding algorithm, Algorithm 1.1. The main result of this section is that well-separated, well-conditioned eigenvalues of the thresholded problem can be recovered accurately in the presence of noise. Together with section 3, which analyzes both the Rayleigh–Ritz and thresholding errors, this makes up a fairly complete explanation for the success of the QSD algorithm when implemented with thresholding.

Let $\boldsymbol{H}$ and $\boldsymbol{S}$ denote the exact outputs of the QSD algorithm (1.3) and $\boldsymbol{V}$ the eigenvectors of $\boldsymbol{S}$ with eigenvalues greater than ϵ. Dependence of $\boldsymbol{V} := \boldsymbol{V}_{>\epsilon}$ on ϵ, as in the introduction, has been suppressed for conciseness. The thresholded problem is described by the pair $(\boldsymbol{A}, \boldsymbol{B}) := (\boldsymbol{V}^* \boldsymbol{H} \boldsymbol{V}, \boldsymbol{V}^* \boldsymbol{S} \boldsymbol{V})$.[2] When implemented on a

[2]To make the output of the thresholding procedure unambiguous, we assume eigenvectors are arranged left to right in decreasing order of the corresponding eigenvalues. The ordering convention does not effect the outputs of the thresholding procedure, Algorithm 1.1.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

quantum computer, $\boldsymbol{H}$ and $\boldsymbol{S}$ are corrupted by noise as $\widetilde{\boldsymbol{H}} := \boldsymbol{H} + \boldsymbol{\Delta_H}$ and $\widetilde{\boldsymbol{S}} := \boldsymbol{S} + \boldsymbol{\Delta_S}$. As a simple measure of the size of the perturbation, we introduce

$$\eta := \sqrt{\eta_{\boldsymbol{H}}^2 + \eta_{\boldsymbol{S}}^2} := \sqrt{\|\boldsymbol{\Delta_H}\|^2 + \|\boldsymbol{\Delta_S}\|^2}, \tag{2.1}$$

which represents the noise level. In principle, one could undertake a careful analysis of the different sources of error (e.g., discretization, gate, and sampling) to obtain probabilistic bounds on η. (We provide such a bound for the sampling error alone in section 4.1.) For now, we shall just assume η or a good bound for it is known, and the threshold level ϵ is chosen to be (at least) larger than η. With the perturbations $\widetilde{\boldsymbol{H}}$ and $\widetilde{\boldsymbol{S}}$ in hand, the practitioner computes the large-eigenvalue eigenvectors $\widetilde{\boldsymbol{V}}$ of the perturbation $\widetilde{\boldsymbol{S}}$, and constructs the perturbed thresholded problem $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}}) := (\widetilde{\boldsymbol{V}}^* \widetilde{\boldsymbol{H}} \widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{V}}^* \widetilde{\boldsymbol{S}} \widetilde{\boldsymbol{V}})$. We denote the dimension of $\widetilde{\boldsymbol{A}}$ and $\widetilde{\boldsymbol{B}}$ as q.

We hope to show that the smallest eigenvalue of the pair $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$—i.e., our computed approximation to the ground state energy—is close to the smallest eigenvalue of $(\boldsymbol{A}, \boldsymbol{B})$. Unfortunately, there are a number of reasons to worry this might not be the case. First, even if $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ is close to $(\boldsymbol{H}, \boldsymbol{S})$, it is still possible that $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ is not close to $(\boldsymbol{A}, \boldsymbol{B})$. A small perturbation in just $\boldsymbol{S}$ can lead to a large perturbation of $\boldsymbol{A}$: For a small parameter $\eta > 0$,

$$\begin{aligned} \boldsymbol{H} &= \begin{bmatrix} 20 & 0 \\ 0 & 1 \end{bmatrix}, \ \boldsymbol{S} = \begin{bmatrix} 1 & 0 \\ 0 & 1 - \frac{\eta}{2} \end{bmatrix} \quad \xrightarrow{\eta \text{ error}} \quad \widetilde{\boldsymbol{H}} = \begin{bmatrix} 20 & 0 \\ 0 & 1 \end{bmatrix}, \ \widetilde{\boldsymbol{S}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 + \frac{\eta}{2} \end{bmatrix}, \\ \boldsymbol{A} &= \begin{bmatrix} 20 & 0 \\ 0 & 1 \end{bmatrix}, \ \boldsymbol{B} = \begin{bmatrix} 1 & 0 \\ 0 & 1 - \frac{\eta}{2} \end{bmatrix} \quad \longrightarrow \quad \widetilde{\boldsymbol{A}} = \begin{bmatrix} 1 & 0 \\ 0 & 20 \end{bmatrix}, \ \widetilde{\boldsymbol{B}} = \begin{bmatrix} 1 + \frac{\eta}{2} & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{2.2}$$

Additionally, $\widetilde{\boldsymbol{A}}$ and $\boldsymbol{A}$ can even be of different sizes if the perturbation causes the number of eigenvalues larger than ϵ to change. Fortunately, (2.2) suggests that the potential for small errors in $\boldsymbol{S}$ to magnify into large errors in $\boldsymbol{A}$ might have a benign source, with the error in this example caused simply by a reordering of the eigenvectors. The eigenvalues of the pair $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ are indifferent to a symmetric reordering of its rows and columns or, more generally, a $*$-conjugation $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}}) \mapsto (\boldsymbol{W}^* \widetilde{\boldsymbol{A}} \boldsymbol{W}, \boldsymbol{W}^* \widetilde{\boldsymbol{B}} \boldsymbol{W})$. Thus, it is sufficient for purposes of analysis to show that $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ are close to $(\boldsymbol{A}, \boldsymbol{B})$ after $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ is replaced by an appropriate $*$-conjugation.

Assume this issue is addressed, and we obtain a $*$-conjugation of $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ that is close to $(\boldsymbol{A}, \boldsymbol{B})$. Classical worst-case perturbation theory still paints a grim portrait on the sensitivities of the eigenvalues of the thresholded pair $(\boldsymbol{A}, \boldsymbol{B})$. After possibly replacing $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ by a $*$-conjugation, let $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}}) = (\boldsymbol{A} + \boldsymbol{\Delta_A}, \boldsymbol{B} + \boldsymbol{\Delta_B})$. A measure of the difference between $(\boldsymbol{A}, \boldsymbol{B})$ and $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ is given by

$$\chi := \sqrt{\|\boldsymbol{\Delta_A}\|^2 + \|\boldsymbol{\Delta_B}\|^2}. \tag{2.3}$$

Let E be the least eigenvalue of $(\boldsymbol{A}, \boldsymbol{B})$ and $\widetilde{E}$ be the least eigenvalue of $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$. The classical perturbation theorem of Stewart [33, Thm. 3.2] can only show that

$$\left| \tan^{-1} \widetilde{E} - \tan^{-1} E \right| \leq \sin^{-1} \frac{\chi}{c(\boldsymbol{A}, \boldsymbol{B})} \leq \sin^{-1} \frac{\chi}{\epsilon}. \tag{2.4}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Here,

$$c(\boldsymbol{A}, \boldsymbol{B}) := \min_{\|\boldsymbol{x}\|=1} \sqrt{(\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x})^2 + (\boldsymbol{x}^* \boldsymbol{B} \boldsymbol{x})^2} \tag{2.5}$$

is the Crawford number, which is only guaranteed to be larger than ϵ under the standing assumptions. This leads to pessimistic error bounds on the order of χ/ϵ. We would much prefer a bound which scales like χ times the condition number of $\tan^{-1} E$, without an explicit ϵ dependence.

In the rest of this section, we address these challenges. The analysis has two parts. In the first part, we use the Davis–Kahan $\sin \boldsymbol{\Theta}$ theorem to show that, after replacing by a $*$-conjugation, $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ is $\chi \leq \mathcal{O}(\eta/\epsilon^\alpha)$-close to $(\boldsymbol{A}, \boldsymbol{B})$, where $0 \leq \alpha \leq 1/2$ is a constant. In particular, α is ensured to be no more than $1/2$, with $\alpha = 1/4$ appearing to be more representative in numerical experiments. For the second part, we use the perturbation theory of Mathias and Li [19] to improve on Stewart's bound (2.4) in the case when the eigenvalue of interest is well-conditioned, obtaining the desired dependence on χ rather than χ/ϵ. We present these pieces in reverse order.

2.1. Eigenvalue perturbation bounds. Suppose that, potentially after a redefining $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ to a $*$-conjugation, $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ and $(\boldsymbol{A}, \boldsymbol{B})$ are separated by a small distance χ, as defined in (2.3). We shall show that the perturbation theory developed by Mathias and Li [19] allows us to significantly improve on the bound (2.4) furnished by Stewart's theory. The beauty of the Mathias–Li theory is that the Crawford number in (2.4) can be replaced by a quantity related to the conditioning of E, provided a spectral gap condition is satisfied. We begin with a mildly specialized version of [19, Thm. 3.3].

FACT 2.1. *Let $(\boldsymbol{A}, \boldsymbol{B})$ be a pair of $q \times q$ Hermitian matrices with all eigenvalues of $\boldsymbol{B}$ larger than ϵ. Consider perturbations $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}}) := (\boldsymbol{A} + \boldsymbol{\Delta_A}, \boldsymbol{B} + \boldsymbol{\Delta_B})$ where $\boldsymbol{\Delta_A}$ and $\boldsymbol{\Delta_B}$ are Hermitian matrices and χ is defined in* (2.3) *and satisfies $q\chi \leq \epsilon$. Let $E_0, \ldots, E_{q-1}$ be the eigenvalues of $(\boldsymbol{A}, \boldsymbol{B})$ with unit-norm eigenvectors $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{q-1}$. Define*

$$\ell_j = \tan^{-1} E_j - \sin^{-1} \frac{q\chi}{d_j}, \quad u_j = \tan^{-1} E_j + \sin^{-1} \frac{q\chi}{d_j}, \tag{2.6}$$

where

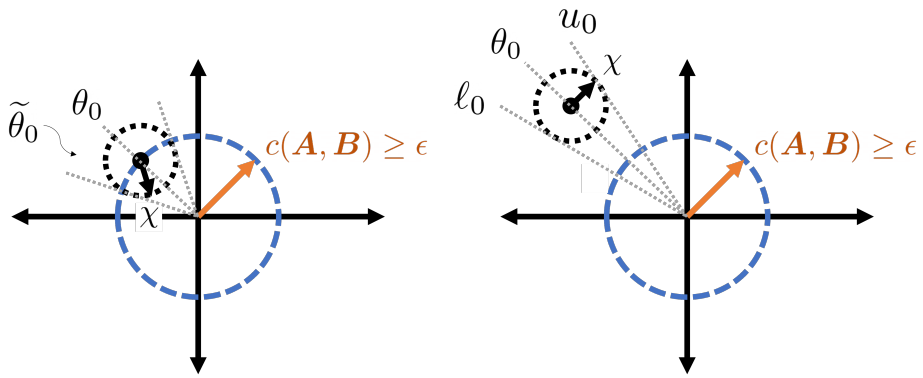
$$d_j := |\boldsymbol{x}_j^* (\boldsymbol{A} + \mathrm{i}\boldsymbol{B}) \boldsymbol{x}_j|. \tag{2.7}$$

Let $\{\ell_j^\uparrow\}_{j=0}^{q-1}$ and $\{u_j^\uparrow\}_{j=0}^{q-1}$ denote the increasing rearrangements of the bounds $\{\ell_j\}_{j=0}^{q-1}$ and $\{u_j\}_{j=0}^{q-1}$. Then with $\widetilde{E}_0 \leq \widetilde{E}_1 \leq \cdots \leq \widetilde{E}_{q-1}$ the eigenvalues of $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$, the following bound holds:

$$\ell_j^\uparrow \leq \tan^{-1} \widetilde{E}_j \leq u_j^\uparrow.$$

We show a pictorial comparison of the proof of Stewart's bound (2.4) and the Mathias–Li bound (2.9) in Figure 1. The perturbation theory of the definite generalized eigenvalues is naturally phrased in terms of the eigenangle $\theta_j := \tan^{-1} E_j$, which represents the angle of the ray through the complex number $\boldsymbol{x}_j^*(\boldsymbol{A} + \mathrm{i}\boldsymbol{B})\boldsymbol{x}_j$ and the positive imaginary axis. In Stewart's theory, one argues that, for a unit vector $\boldsymbol{x}$, the complex number $z := \boldsymbol{x}^*(\boldsymbol{A} + \mathrm{i}\boldsymbol{B})\boldsymbol{x}$ must be a distance at least $c(\boldsymbol{A}, \boldsymbol{B})$ from the origin

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.



(a) Stewart (b) Mathias–Li

FIG. 1. *Pictorial comparison of the proofs of Stewart's bound* (2.4) *and Mathias and Li's bound* (2.9).

and the perturbed point $\widetilde{z} := \boldsymbol{x}^*(\widetilde{\boldsymbol{A}} + \mathrm{i}\widetilde{\boldsymbol{B}})\boldsymbol{x}$ is a distance at most χ from z. In view of a variational characterization Stewart proved for the eigenangles [33, Thm. 3.1], it follows that the eigenangles change by at most $\sin^{-1}(\chi/c(\boldsymbol{A},\boldsymbol{B}))$. Mathias and Li instead consider the points $z_j := \boldsymbol{x}_j^*(\boldsymbol{A} + \mathrm{i}\boldsymbol{B})\boldsymbol{x}_j$ for the unit-norm eigenvectors $\boldsymbol{x}_j$. A disk of radius χ centered at z_j is enclosed by rays with angles $\{\theta_j \pm \sin^{-1}(\chi/d_j)\} = \{\ell_j, u_j\}$. Using Stewart's variational principle, Mathias and Li prove that, while the perturbed eigenangle $\widetilde{\theta}_j := \tan^{-1}\widetilde{E}_j$ need not lie within $[\ell_j, u_j]$, it must lie in $[\ell_j^{\uparrow}, u_j^{\uparrow}]$.

A consequence of Mathias and Li's analysis (see [19, eq. (3.6)]) is that $\ell_j \leq \widetilde{\theta}_j \leq u_j$ does hold if there is a large enough gap between θ_j and other eigenangles relative to the size of the perturbation. This bound $\ell_j \leq \widetilde{\theta}_j \leq u_j$ is nearly as good a bound as one could hope since d_j^{-1} is the condition number of θ_j [34, Thm. VI.2.2].

COROLLARY 2.2. *Instate the notation and assumptions of* Fact 2.1. *Suppose that* E_j *satisfies the gap condition*

$$\min\bigl(\tan^{-1} E_j - \tan^{-1} E_{j-1},\, \tan^{-1} E_{j+1} - \tan^{-1} E_j\bigr) \geq \sin^{-1}\frac{q\chi}{\epsilon} - \sin^{-1}\frac{q\chi}{d_j}, \tag{2.8}$$

where the first or second term of the minimum can be ignored if $j = 0$ *or* $j = q-1$, *respectively. Then*

$$\bigl|\tan^{-1}\widetilde{E}_j - \tan^{-1} E_j\bigr| \leq \sin^{-1}\frac{q\chi}{d_j}. \tag{2.9}$$

A couple comments are in order before we proceed with the proof of Corollary 2.2. We have purged the suboptimal factor χ/ϵ from the eigenangle bound (2.9) and replaced it with the often much smaller quantity χ/d_j (tempered by a dimensional factor). However, χ/ϵ remains, just in the gap condition (2.8) (and the hypothesis $q\chi \leq \epsilon$). If the eigenangle gap on the left-hand side of (2.8) is reasonably large, then ϵ must merely be a modest multiple of χ for (2.8) to be satisfied.

Proof of Corollary 2.2. Denote $\widetilde{\theta}_i := \tan^{-1}\widetilde{E}_i$ and $\theta_i := \tan^{-1} E_i$ for every $i = 0, 1, \ldots, q-1$. We shall prove the upper bound $\widetilde{\theta}_j - \theta_j \leq \sin^{-1}(q\chi/d_j)$ with the corresponding lower bound being proven in exactly the same way. We shall do this by showing $u_j^{\uparrow} \leq u_j$ under the gap assumption (2.8). To do this, we shall show that

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$u_i \le \theta_j + \sin^{-1}(q\chi/d_j)$ for every $0 \le i < j$. If $j = 0$, then no such i exists and the upper bound is automatically true. We thus continue in the case $j > 0$.

Fix $0 \le i < j$. Since every eigenvalue of $\boldsymbol{B}$ is at least ϵ, we have that $d_i = |\boldsymbol{x}_i^*(\boldsymbol{A}+\mathrm{i}\boldsymbol{B})\boldsymbol{x}_i| \ge \boldsymbol{x}_i^*\boldsymbol{B}\boldsymbol{x}_i > \epsilon$. Thus, we have $u_i \le \theta_i + \sin^{-1}(q\chi/\epsilon)$. Since $E_i \le E_{j-1}$, we have $\theta_i = \tan^{-1} E_i \le \tan^{-1} E_{j-1}$. Thus,

$$u_i \le \tan^{-1} E_i + \sin^{-1}\frac{q\chi}{\epsilon} \le \tan^{-1} E_{j-1} + \sin^{-1}\frac{q\chi}{\epsilon} \le \tan^{-1} E_j + \sin^{-1}\frac{q\chi}{d_j}$$

by (2.8). From this we conclude $u_j^{\uparrow} \le u_j$ so, by Fact 2.1, $\widetilde{\theta}_j \le u_j^{\uparrow} \le u_j = \theta_j + \sin^{-1}(q\chi/d_j)$. ∎

2.2. How do perturbations affect the thresholded problem? In this section, we seek to understand how perturbations of the pair $(\boldsymbol{H}, \boldsymbol{S})$ affect the thresholded problem $(\boldsymbol{A}, \boldsymbol{B})$. As the example (2.2) shows, it need not be the case that $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ is close to $(\boldsymbol{A}, \boldsymbol{B})$ if $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ is close to $(\boldsymbol{H}, \boldsymbol{S})$. Thus, in view of the fact that $*$-conjugations do not change the eigenvalues of a matrix pair [34, Thm. VI.1.8], our goal will be to show there exists a nonsingular matrix $\boldsymbol{W}$ such that $(\boldsymbol{W}^*\widetilde{\boldsymbol{A}}\boldsymbol{W}, \boldsymbol{W}^*\widetilde{\boldsymbol{B}}\boldsymbol{W})$ is close to $(\boldsymbol{A}, \boldsymbol{B})$. An instrumental tool in this goal will be the Davis–Kahan $\sin\boldsymbol{\Theta}$ theorem [7], which we state a somewhat less general version here for reference.

FACT 2.3 (Davis–Kahan $\sin\boldsymbol{\Theta}$ Theorem; see also [1, Thm. VII.3.1]). *Consider Hermitian matrices $\boldsymbol{M}$ and $\widetilde{\boldsymbol{M}}$ and let $\boldsymbol{\Pi}$ and $\widetilde{\boldsymbol{\Pi}}$ be the spectral projectors of $\boldsymbol{M}$ and $\widetilde{\boldsymbol{M}}$ associated with collections of eigenvalues of $\boldsymbol{M}$ within an interval $[a, b]$ and of $\widetilde{\boldsymbol{M}}$ outside the interval $[a-\delta, b+\delta]$, respectively. Then $\|\boldsymbol{\Pi}\widetilde{\boldsymbol{\Pi}}\| \le \|\widetilde{\boldsymbol{M}} - \boldsymbol{M}\|/\delta$.*

We begin with some notation. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ denote the eigenvectors of $\boldsymbol{S}$ with associated eigenvalues $\lambda_1, \dots, \lambda_n$ and $\boldsymbol{\Pi}$ denote the spectral projector associated with eigenvalues of $\boldsymbol{S}$ which are larger than ϵ. We denote by m the critical index for which $\lambda_m > \epsilon \ge \lambda_{m+1}$. All quantities with tildes shall denote quantities defined as above for the perturbed problem.

We begin with the following simple bound.

PROPOSITION 2.4. *Suppose that*

$$\|\boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{H}\widetilde{\boldsymbol{\Pi}}\| \le \chi_{\boldsymbol{H}}, \tag{2.10}$$

$$\|\boldsymbol{\Pi}\boldsymbol{S}\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\| \le \chi_{\boldsymbol{S}}, \tag{2.11}$$

and suppose that $\widetilde{\boldsymbol{S}}$ and $\boldsymbol{S}$ have the same number of eigenvalues larger than ϵ. Then for the nonsingular matrix $\boldsymbol{W} := \widetilde{\boldsymbol{V}}^*\boldsymbol{V}$,

$$\sqrt{\|\boldsymbol{W}^*\widetilde{\boldsymbol{A}}\boldsymbol{W} - \boldsymbol{A}\|^2 + \|\boldsymbol{W}^*\widetilde{\boldsymbol{B}}\boldsymbol{W} - \boldsymbol{B}\|^2} \le \sqrt{(\chi_{\boldsymbol{H}} + \eta_{\boldsymbol{H}})^2 + (\chi_{\boldsymbol{S}} + \eta_{\boldsymbol{S}})^2}.$$

This result gives a bound on χ as defined in (2.3) if one redefines $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ to its $*$-conjugation by $\boldsymbol{W}$.

Proof. First, note that $\boldsymbol{W}$ is nonsingular because it is the product of a matrix with full row rank and full column rank. Note also that $\|\boldsymbol{V}\|, \|\widetilde{\boldsymbol{V}}\|, \|\boldsymbol{\Pi}\|, \|\widetilde{\boldsymbol{\Pi}}\| = 1$ since $\boldsymbol{V}$ and $\widetilde{\boldsymbol{V}}$ have orthonormal columns. The result then follows immediately by the bound

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\begin{aligned}\|\boldsymbol{W}^*\widetilde{\boldsymbol{A}}\boldsymbol{W}-\boldsymbol{A}\| &= \|\boldsymbol{V}^*\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^*\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^*\boldsymbol{V}-\boldsymbol{V}^*\boldsymbol{H}\boldsymbol{V}\| = \|\boldsymbol{V}^*(\widetilde{\boldsymbol{\Pi}}\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{\Pi}}-\boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi})\boldsymbol{V}\| \\ &\le \|\boldsymbol{V}^*(\widetilde{\boldsymbol{\Pi}}\boldsymbol{H}\widetilde{\boldsymbol{\Pi}}-\boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi})\boldsymbol{V}\| + \|\boldsymbol{V}^*\widetilde{\boldsymbol{\Pi}}(\widetilde{\boldsymbol{H}}-\boldsymbol{H})\widetilde{\boldsymbol{\Pi}}\boldsymbol{V}\| \\ &\le \|\widetilde{\boldsymbol{\Pi}}\boldsymbol{H}\widetilde{\boldsymbol{\Pi}}-\boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi}\| + \|\widetilde{\boldsymbol{H}}-\boldsymbol{H}\| \le \chi_{\boldsymbol{H}}+\eta_{\boldsymbol{H}}\end{aligned}$$

and similarly for $\|\boldsymbol{W}^*\widetilde{\boldsymbol{B}}\boldsymbol{W}-\boldsymbol{B}\|$. ∎

Note that a necessary condition for the hypotheses of our error bound Corollary 2.2 is for the distance χ (2.3) between $(\widetilde{\boldsymbol{A}},\widetilde{\boldsymbol{B}})$ and $(\boldsymbol{A},\boldsymbol{B})$ to be strictly smaller than ϵ. This forecloses our ability to use simple bounds for, e.g., (2.11) such as

$$\|\boldsymbol{\Pi}\boldsymbol{S}\boldsymbol{\Pi}-\widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\| \le \|\boldsymbol{\Pi}\boldsymbol{S}\boldsymbol{\Pi}-\boldsymbol{S}\| + \|\boldsymbol{S}-\widetilde{\boldsymbol{S}}\| + \|\widetilde{\boldsymbol{S}}-\widetilde{\boldsymbol{\Pi}}\widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{\Pi}}\| \le 2(\epsilon+\eta).$$

We thus seek bounds of the form (2.10) and (2.11) without an additive $\mathcal{O}(\epsilon)$ term.

We begin with the more challenging of the two bounds, namely (2.10). Certainly, we should not expect a meaningful bound (2.10) if $\boldsymbol{H}$ and $\boldsymbol{S}$ have no relation to each other. For the sake of generality, we shall perform analysis under the assumption that $(\boldsymbol{H},\boldsymbol{S})$ obey a weighted geometric mean inequality of the form

$$|\boldsymbol{v}_i^*\boldsymbol{H}\boldsymbol{v}_j| \le \mu\min(\lambda_i,\lambda_j)^{1-\alpha}\max(\lambda_i,\lambda_j)^{\alpha} \quad \text{for all } 1\le i,j\le n, \tag{2.12}$$

where $0\le\alpha\le 1/2$ and $\mu>0$ are constants. While this may appear strange, (2.12) necessarily holds with $\mu=\max|\Lambda(\boldsymbol{H},\boldsymbol{S})|$ and $\alpha=1/2$ by a direct application of the Courant–Fischer principle for generalized eigenvalue problems [34, Cor. VI.1.16]. Numerical experiments suggest that $\alpha=1/4$ and $\mu\approx\max|\Lambda(\boldsymbol{H},\boldsymbol{S})|$ appear to be more revealing of the empirically observed values of $|\boldsymbol{v}_i^*\boldsymbol{H}\boldsymbol{v}_j|$ for $(\boldsymbol{H},\boldsymbol{S})$ computed by QSD for the physical models we tried; see section SM6.

Our most challenging technical result of this section will be a bound on the projection difference (2.10) under the assumption (2.12).

THEOREM 2.5. *Instate the prevailing notation and assume the bound* (2.12) *holds for* $0\le\alpha\le 1/2$. *Assume the eigenvalue gap condition*

$$\lambda_{m+1}+\eta_{\boldsymbol{S}} \le \epsilon < (1+\rho)\epsilon \le \lambda_m \tag{2.13}$$

for some $\rho>0$. *Suppose in addition that* $\eta_{\boldsymbol{S}}$ *is sufficiently small that* $(1+\rho^{-1})\eta_{\boldsymbol{S}}/\epsilon\le 1$. *Then the projection error* (2.10) *obeys the following bound:*

$$\|\boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi}-\widetilde{\boldsymbol{\Pi}}\boldsymbol{H}\widetilde{\boldsymbol{\Pi}}\| \le 3\mu n^3(1+\rho^{-1})\left(\frac{\|\boldsymbol{S}\|}{\epsilon}\right)^{\alpha}\eta_{\boldsymbol{S}}, \tag{2.14}$$

where $\eta_{\boldsymbol{S}}$ *is defined in* (2.1).

There are some unappealing features of this result, namely the cubic dependence on the problem size and the $\mathcal{O}(\eta_{\boldsymbol{S}}/\epsilon^{\alpha})$ scaling. The first of these, the cubic dependence on n, we believe to likely be an artifact of our proof technique, applying Davis–Kahan "entry by entry." Fortunately, numerical experiments do not suggest a dramatic dependence of the error on n. The second effect—the $\eta_{\boldsymbol{S}}/\epsilon^{\alpha}$ dependence rather than a more desirable $\eta_{\boldsymbol{S}}$ dependence—appears to be a genuine feature of this problem, at least without additional assumptions; see section SM5.[3] Fortunately, we have yet to

[3] Our evidence for this, presented in section SM5, is a synthetically generated pair $(\boldsymbol{H},\boldsymbol{S})$ obeying the geometric mean condition (2.12); we did not obtain this pair from QSD.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

find an instance of a pair $(\boldsymbol{H}, \boldsymbol{S})$ generated by the QSD algorithm for which the ϵ^α factor appears to be necessary to understand the true error.

Finally, we note that the separation hypothesis (2.13) is relatively mild. The first inequality of (2.13) is necessary just to ensure that $\boldsymbol{A}$ and $\widetilde{\boldsymbol{A}}$ have the same size. If we assume just a little more of a spectral gap around the thresholding level, quantified by the requirement that ρ is bounded away from zero, then we get a nice bound. A less careful application of Davis–Kahan would require that *all* the eigenvalues of $\boldsymbol{S}$ are well-separated, so we consider a modest gap at the thresholding level to be a fairly mild requirement.

Proof of Theorem 2.5. The proof shall be an enthusiastic exercise in applying the Davis–Kahan $\sin\boldsymbol{\Theta}$ theorem, Fact 2.3. We begin by bounding

$$\begin{aligned}\|\boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{H}\widetilde{\boldsymbol{\Pi}}\|^2 &\le \|\boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{H}\widetilde{\boldsymbol{\Pi}}\|_{\mathrm{F}}^2 \\ &= \sum_{i,j=1}^n \left(\boldsymbol{v}_i^*\left(\boldsymbol{\Pi}\boldsymbol{H}\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{H}\widetilde{\boldsymbol{\Pi}}\right)\boldsymbol{v}_j\right)^2 =: \sum_{i,j=1}^n I_{ij}^2.\end{aligned} \tag{2.15}$$

Our strategy will be to bound each of the terms I_{ij}.

For each i, we can expand $\widetilde{\boldsymbol{\Pi}}\boldsymbol{v}_i = \sum_{k=1}^n c_{ik}\boldsymbol{v}_k$. Multiplying through by $\boldsymbol{v}_i^*$ then gives that $c_{ik} = \boldsymbol{v}_i^*\widetilde{\boldsymbol{\Pi}}\boldsymbol{v}_k$, which shows in particular that $c_{ik} = c_{ki}$. Our first goal will be to bound $|c_{ik}|$. We break into two cases, $k \le m$ and $k > m$.

For case one, assume that $k \le m$. By Weyl's inequality [34, Cor. IV.4.9], the $(m+1)$st largest eigenvalue of $\widetilde{\boldsymbol{S}}$ satisfies $\widetilde{\lambda}_{m+1} \le \lambda_{m+1} + \eta_{\boldsymbol{S}} \le \epsilon$. The Davis–Kahan theorem shows the difference $\boldsymbol{\delta}_k := \widetilde{\boldsymbol{\Pi}}\boldsymbol{v}_k - \boldsymbol{v}_k$ satisfies

$$\|\boldsymbol{\delta}_k\| \le \frac{\eta_{\boldsymbol{S}}}{\lambda_k - \widetilde{\lambda}_{m+1}} \le \frac{\eta_{\boldsymbol{S}}}{\lambda_k - \epsilon} \le \frac{(1+\rho^{-1})\eta_{\boldsymbol{S}}}{\lambda_k}.$$

This gives a bound on the coefficients c_{ik} for $i \ne k$:

$$|c_{ik}| = \left|\boldsymbol{v}_i^*\widetilde{\boldsymbol{\Pi}}\boldsymbol{v}_k\right| = |\boldsymbol{v}_i^*\boldsymbol{v}_k + \boldsymbol{v}_i^*\boldsymbol{\delta}_k| = |\boldsymbol{v}_i^*\boldsymbol{\delta}_k| \le \|\boldsymbol{\delta}_k\| \le \frac{(1+\rho^{-1})\eta_{\boldsymbol{S}}}{\lambda_k}. \tag{2.16}$$

For $i = k$, we have

$$|1 - c_{ii}| = |1 - \boldsymbol{v}_i^*(\boldsymbol{v}_i + \boldsymbol{\delta}_i)| \le \|\boldsymbol{\delta}_i\| \le \frac{(1+\rho^{-1})\eta_{\boldsymbol{S}}}{\lambda_i}. \tag{2.17}$$

Now consider case two where $k > m$. Since $\widetilde{\lambda}_{m+1} \le \epsilon$ as argued above, $\boldsymbol{\Pi}$ and $\widetilde{\boldsymbol{\Pi}}$ are projections onto subspaces of the same dimensions so $\|\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\| = \|\boldsymbol{\Pi}(\boldsymbol{I} - \widetilde{\boldsymbol{\Pi}})\|$ by [34, Thm. I.5.5]. Applying Davis–Kahan then gives

$$\|\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\| = \|\boldsymbol{\Pi}(\boldsymbol{I} - \widetilde{\boldsymbol{\Pi}})\| \le \frac{\eta_{\boldsymbol{S}}}{\rho\epsilon}.$$

Then

$$|c_{ik}| = \left|\boldsymbol{v}_i^*\widetilde{\boldsymbol{\Pi}}\boldsymbol{v}_k\right| \le \left|\boldsymbol{v}_i^*(\widetilde{\boldsymbol{\Pi}} - \boldsymbol{\Pi})\boldsymbol{v}_k\right| + |\boldsymbol{v}_i^*\boldsymbol{\Pi}\boldsymbol{v}_k| \le \|\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\| \le \frac{\eta_{\boldsymbol{S}}}{\rho\epsilon}. \tag{2.18}$$

With these bounds in hand, we return to bounding I_{ij} as defined in (2.15). Let us introduce shorthand notation $a \wedge b$ and $a \vee b$ for the minimum and maximum of a and b, respectively. Expanding $\widetilde{\boldsymbol{\Pi}}\boldsymbol{v}_i$ and $\widetilde{\boldsymbol{\Pi}}\boldsymbol{v}_j$ and using the bound (2.12), we obtain

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

(2.19)

$$I_{ij} = \boldsymbol{v}_i^\dagger(\boldsymbol{\Pi H \Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{H}\widetilde{\boldsymbol{\Pi}})\boldsymbol{v}_j = (1 - \overline{c_{ii}}c_{jj}) \cdot \boldsymbol{v}_i^\dagger \boldsymbol{H}\boldsymbol{v}_j - \sum_{\substack{k,\ell=1 \\ k\neq i \text{ or } \ell\neq j}}^{n} \overline{c_{ik}}c_{j\ell}\boldsymbol{v}_k^\dagger \boldsymbol{H}\boldsymbol{v}_\ell$$

$$\leq \mu \left[|1 - \overline{c_{ii}}c_{jj}|(\lambda_i \wedge \lambda_j)^{1-\alpha}(\lambda_i \vee \lambda_j)^\alpha + \sum_{\substack{k,\ell=1 \\ k\neq i \text{ or } \ell\neq j}}^{n} |c_{ik}c_{j\ell}|(\lambda_k \vee \lambda_\ell)^{1-\alpha}(\lambda_k \wedge \lambda_\ell)^\alpha \right].$$

We thus turn our attention to bounding the summands in the second term of (2.19) for $k \neq i$ or $j \neq \ell$. First, assume that $k \neq i$ and suppose that $k \leq m$; we bound using (2.16)

$$|c_{ik}c_{j\ell}|(\lambda_k \wedge \lambda_\ell)^{1-\alpha}(\lambda_k \vee \lambda_\ell)^\alpha \leq |c_{ik}|\lambda_k^{1-\alpha}\|\boldsymbol{S}\|^\alpha \leq \frac{(1+\rho^{-1})\eta_{\boldsymbol{S}}\lambda_k^{1-\alpha}\|\boldsymbol{S}\|^\alpha}{\lambda_k}$$
$$\leq (1+\rho^{-1})\eta_{\boldsymbol{S}}(\|\boldsymbol{S}\|/\epsilon)^\alpha.$$

Next assuming $k > m$, (2.18) yields

$$|c_{ik}c_{j\ell}|(\lambda_k \wedge \lambda_\ell)^{1-\alpha}(\lambda_k \vee \lambda_\ell)^\alpha \leq |c_{ik}|\epsilon^{1-\alpha}\|\boldsymbol{S}\|^\alpha \leq \eta_{\boldsymbol{S}}\rho^{-1}(\|\boldsymbol{S}\|/\epsilon)^\alpha.$$

Turning our attention to the first term of the final bound in (2.19), (2.17) and the assumption $(1+\rho^{-1})\eta_{\boldsymbol{S}}/\epsilon \leq 1$ give

$$\begin{aligned} &|1 - \overline{c_{ii}}c_{jj}|(\lambda_i \wedge \lambda_j)^{1-\alpha}(\lambda_i \vee \lambda_j)^\alpha \\ &\quad \leq (|1 - c_{ii}| + |1 - c_{jj}| + |1 - c_{ii}||1 - c_{jj}|)(\lambda_i \wedge \lambda_j)^{1-\alpha}\|\boldsymbol{S}\|^\alpha \\ &\quad \leq 3\eta_{\boldsymbol{S}}(1+\rho^{-1})(\|\boldsymbol{S}\|/\epsilon)^\alpha. \end{aligned}$$

Using the three previous displays to bound each of the n^2 summands in (2.19), we obtain

$$I_{ij} \leq 3\mu n^2(1+\rho^{-1})(\|\boldsymbol{S}\|/\epsilon)^\alpha \eta_{\boldsymbol{S}},$$

which then leads to the stated bound. ∎

A bound for (2.11) is entirely analogous. The analysis is made significantly easier by the fact that the spectral projector $\boldsymbol{\Pi}$ is defined in terms of the matrix $\boldsymbol{S}$ itself.

THEOREM 2.6. *Instate the prevailing notation. Assume that* (2.13) *holds for some* $\rho > 0$. *The projection error* (2.11) *satisfies the bound*

$$\|\boldsymbol{\Pi S \Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\| \leq \|\boldsymbol{\Pi S \Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\|_{\mathrm{F}} \leq 2(1+\rho^{-1})\eta_{\boldsymbol{S}}n + \epsilon^{-1}\big[(1+\rho^{-1})\eta_{\boldsymbol{S}}n\big]^2.$$

In particular if $(1+\rho^{-1})\eta_{\boldsymbol{S}}n/\epsilon \leq 1$, *we have*

$$\|\boldsymbol{\Pi S \Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\| \leq \|\boldsymbol{\Pi S \Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\|_{\mathrm{F}} \leq 3(1+\rho^{-1})\eta_{\boldsymbol{S}}n.$$

Proof. The proof is quite similar to Theorem 2.5 and we shall thus proceed more quickly. First, we bound

$$\|\boldsymbol{\Pi S \Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\|^2 \leq \|\boldsymbol{\Pi S \Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\|_{\mathrm{F}}^2 = \sum_{i,j=1}^{n} \Big(\boldsymbol{v}_i^* \Big(\boldsymbol{\Pi S \Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}}\Big)\boldsymbol{v}_j\Big)^2 =: \sum_{i,j=1}^{n} I_{ij}^2.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Consider the expansion $\widetilde{\boldsymbol{\Pi}}\boldsymbol{v}_i = \sum_{k=1}^n c_{ik}\boldsymbol{v}_k$ as in the proof of Theorem 2.5. By the same arguments, the bounds (2.16), (2.17), and (2.18) hold under the respective hypotheses that $i \le m$ and $i \ne k$, $i = k$, and $i > m$.

We now compute I_{ij} using the fact that $\boldsymbol{v}_i^*\boldsymbol{S}\boldsymbol{v}_j = \lambda_i\delta_{ij}$ where δ_{ij} denotes the Kronecker delta:

$$I_{ij} = \boldsymbol{v}_i^*(\boldsymbol{\Pi}\boldsymbol{S}\boldsymbol{\Pi} - \widetilde{\boldsymbol{\Pi}}\boldsymbol{S}\widetilde{\boldsymbol{\Pi}})\boldsymbol{v}_j = (1 - \overline{c_{ii}}c_{jj}) \cdot \boldsymbol{v}_i^*\boldsymbol{S}\boldsymbol{v}_j - \sum_{\substack{k,\ell=1 \\ k\ne i \text{ or } \ell\ne j}}^n \overline{c_{ik}}c_{j\ell}\boldsymbol{v}_k^*\boldsymbol{H}\boldsymbol{v}_\ell$$

$$= \lambda_i\delta_{ij} - \sum_{k=1}^n \overline{c_{ik}}c_{jk}\lambda_k.$$

We now distinguish two cases. First suppose $i \ne j$. Then we bound using (2.16) and (2.18):

$$|I_{ij}| \le \sum_{k=1}^n |c_{ik}c_{jk}|\lambda_k \le |c_{ij}|(\lambda_i + \lambda_j) + \sum_{\substack{k=1 \\ k\notin\{i,j\}}}^m |c_{ik}c_{jk}|\lambda_k + \sum_{m+1}^n |c_{ik}c_{jk}|\lambda_k$$

$$\le 2\eta_{\boldsymbol{S}}(1+\rho^{-1}) + \frac{n(1+\rho^{-1})^2\eta_{\boldsymbol{S}}^2}{\epsilon}.$$

Next suppose $i = j$. Then applying (2.16), (2.17), and (2.18) gives the same bound

$$|I_{ij}| \le 2\eta_{\boldsymbol{S}}(1+\rho^{-1}) + \frac{n(1+\rho^{-1})^2\eta_{\boldsymbol{S}}^2}{\epsilon}.$$

This entrywise bound immediately gives the desired result. ∎

2.3. Main result. We conclude this section by combining Corollary 2.2, Proposition 2.4, and Theorems 2.5 and 2.6 into our main result, which provides a formal statement of Informal Theorem 1.2 from the introduction.

THEOREM 2.7. *Let $(\boldsymbol{H}, \boldsymbol{S})$ be a pair of $n \times n$ Hermitian matrices perturbed to a pair $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ by perturbations $\boldsymbol{\Delta}_{\boldsymbol{H}}$ and $\boldsymbol{\Delta}_{\boldsymbol{S}}$ of spectral norms $\eta_{\boldsymbol{H}}$ and $\eta_{\boldsymbol{S}}$. Assume the following:*

- *The pair $(\boldsymbol{H}, \boldsymbol{S})$ satisfies the geometric mean bound (2.12) for some parameters $\mu > 0$ and $0 \le \alpha \le 1/2$.*
- *There exists an index m for which (2.13) holds for some $\rho > 0$.*
- *The noise $\eta_{\boldsymbol{S}}$ is sufficiently small so that $(1+\rho^{-1})\eta_{\boldsymbol{S}}/\epsilon \le 1$.*

Let $(\boldsymbol{A}, \boldsymbol{B}) = (\boldsymbol{V}^*\boldsymbol{H}\boldsymbol{V}, \boldsymbol{V}^*\boldsymbol{S}\boldsymbol{V})$ denote the thresholded matrix pair. The eigenvalues recovered by the thresholding procedure applied to the noise-perturbed pair $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ are the same as the eigenvalues of a pair $(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}})$ satisfying

$$\sqrt{\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|^2 + \|\widetilde{\boldsymbol{B}} - \boldsymbol{B}\|^2} \le 3(2+\mu)n^3(1+\rho^{-1})\left(\frac{\|\boldsymbol{S}\|}{\epsilon}\right)^\alpha \eta_{\boldsymbol{S}} + \eta_{\boldsymbol{H}} =: \chi.$$

Let E_0 and E_1 denote the least and second-to-least eigenvalues of $(\boldsymbol{A}, \boldsymbol{B})$. Suppose further that

- *the error bound χ is sufficiently small: $n\chi \le \epsilon$;*
- *the gap condition $\tan^{-1} E_1 - \tan^{-1} E_0 \ge \sin^{-1}(n\chi/\epsilon)$ holds.*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Then with d_0^{-1} the condition number of the eigenangle $\tan^{-1} E_0$ *and* $\widetilde{E}_0$ *the eigenvalue recovered by thresholding applied to* $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$,

$$\left|\tan^{-1} E_0 - \tan^{-1} \widetilde{E}_0\right| \leq \sin^{-1} \frac{n\chi}{d_0}.$$

In particular, for the theorem to hold, the threshold parameter must be chosen,

$$\epsilon > n\chi = 3(2+\mu)n^4(1+\rho^{-1})\left(\frac{\|\boldsymbol{S}\|}{\epsilon}\right)^{\alpha} \eta_{\boldsymbol{S}} + n\eta_{\boldsymbol{H}}.$$

This leads to the claimed value $\epsilon = \Theta(\eta^{\frac{1}{1+\alpha}})$ and bound $|\tan^{-1} E_0 - \tan^{-1} \widetilde{E}_0| \leq \mathcal{O}(d_0^{-1}\eta^{\frac{1}{1+\alpha}})$ in Informal Theorem 1.2.

3. Analysis of unitary Krylov subspace approximation with thresholding. Having studied the errors due to noise in the previous section, we now turn to analyzing the Rayleigh–Ritz errors in computing the ground-state eigenvalue E_0 of $\widehat{\boldsymbol{H}}$ using the unitary Krylov space (1.1) and the thresholding procedure Algorithm 1.1. Specifically, we shall bound the difference between E_0 and the result $\widetilde{E}_0$ of applying thresholding to the pair $(\boldsymbol{H}, \boldsymbol{S})$ (1.3) computed on an error-free quantum computer.

If one considers our analysis with the threshold parameter ϵ set to zero, one obtains an analysis of the QSD method with no truncation. To our knowledge, this is the first quantitative error analysis of the QSD method even in the noise-free setting. This builds on two earlier explanations of the success of QSD. The first, by Stair, Huang, and Evangelista [32], argues based on Taylor series that the QSD subspaces approximately coincide with the classical polynomial Krylov space. This explanation has two drawbacks: (1) small time steps make the ill-conditioning of $\boldsymbol{H}$ and $\boldsymbol{S}$ worse [32, sect. 2.1] and (2) QSD often performs better with larger time steps [15, sect. II.B]. An alternate analysis based on filter diagonalization is provided by Klymko et al. [15], which provides an overcomplete set of *phase cancellation conditions* under which QSD computes the eigenvalues of interest with zero error. They then argue that these conditions hold approximately in the long-time limit for a randomly chosen time step.

Our analysis is more direct than the two previous, emulating the classical analysis of the Lanczos method by Saad [29]. This leads to a quantitative error bound in terms of the distribution of the spectra which decreases exponentially in the number of time steps used. The basic idea will be to use a linear combination of the QSD basis states $\boldsymbol{\varphi}_j$ which has the effect of applying a trigonometric polynomial to the eigenvalues $E_0, \ldots, E_{N-1}$ of $\widehat{\boldsymbol{H}}$. If $\boldsymbol{\varphi}_j$ approximately lies in the span of the first $M+1$ eigenstates, we shall choose this trigonometric polynomial to be large at the eigenvalue of interest and exponentially small (in n) at eigenvalues $E_1, \ldots, E_M$. The trigonometric polynomial will be bounded so it won't amplify any components of the eigenvector in the direction of any of the remaining eigenvectors. As an additional feature of this analysis, we are able to directly analyze thresholding "for free" in the noiseless setting, where thresholding has the effect of perturbing this trigonometric polynomial. The main theorem of this section is as follows.

THEOREM 3.1. *Let $\boldsymbol{\psi}_0, \ldots, \boldsymbol{\psi}_{N-1}$ be the eigenvectors of a Hermitian operator $\widehat{\boldsymbol{H}}$ with eigenvalues $E_0, \ldots, E_{N-1}$. Suppose the initial vector is expanded as*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$$\boldsymbol{\varphi}_0 = \sum_{j=0}^{N-1} \gamma_i \boldsymbol{\psi}_i. \tag{3.1}$$

Let $\Delta E_j := E_j - E_0$ *and choose an index* $0 \le M \le N-1$. *Suppose the QSD algorithm is implemented with time sequence* $\{t_j\}_{j=-k}^{k}$ *for* $t_j = \pi j/\Delta E_M$. *Suppose the generalized eigenvalue problem* (1.2) *is solved with thresholding parameter* ϵ *and let* ϵ_{total} *be the sum of the eigenvalues of* $\boldsymbol{S}$ *discarded by thresholding. Then*

(3.2)
$$0 \le \widetilde{E}_0 - E_0 \le \frac{2\left[\Delta E_{N-1}\epsilon_{\text{total}} + 4(1+\frac{\pi\Delta E_1}{\Delta E_M})^{-2k}\sum_{i=1}^{M}\Delta E_i|\gamma_i|^2 + \sum_{i=M+1}^{N-1}\Delta E_i|\gamma_i|^2\right]}{|\gamma_0|^2 - 2|\gamma_0|\sqrt{(2k+1)\epsilon}}.$$

Remark 3.2. The value $|\gamma_0|^2 = |\langle \varphi_0 \,|\, \psi_0\rangle|^2$ is referred to as the initial overlap. In order for the QSD algorithm to succeed with a relatively small number of Krylov steps, $|\gamma_0|^2$ must be sufficiently large (for instance, ≥ 0.1). This is qualitatively different from the assumption of classical Krylov subspace methods for solving eigenvalue problems, where the initial overlap $|\gamma_0|^2$ should also be nonzero but can be very small.

Let us note some salient features of this result. First, if we consider the noise-free case without thresholding, we obtain the bound

$$0 \le \widetilde{E}_0 - E_0 \le 8\left(1+\frac{\pi\Delta E_1}{\Delta E_M}\right)^{-2k}\sum_{i=1}^{M}\Delta E_i\frac{|\gamma_i|^2}{|\gamma_0|^2} + 2\sum_{i=M+1}^{N-1}\Delta E_i\frac{|\gamma_i|^2}{|\gamma_0|^2}.$$

The error bound has two terms, a term concerning the eigenvalues $E_1, \dots, E_M$ which is damped exponentially fast with rate $\Delta E_1/\Delta E_M$ and an undamped (but also unamplified) term with the eigenvalues $E_{M+1}, \dots, E_{N-1}$. If the components of $\boldsymbol{\varphi}_0$ in the directions of the $(M+1)$st to $(N-1)$st eigenvectors are small in the sense that $\Delta E_i|\gamma_i|^2 \ll 1$, this second term will be small. This bound thus has a tradeoff: If M is chosen larger (i.e., the time step becomes smaller), more terms will be exponentially damped but at a slower rate as $\Delta E_1/\Delta E_M$ will decrease. One can obtain the simplest bound by choosing $M = N-1$ and bounding $\Delta E_i \le \Delta E_{N-1}$, leading to

$$0 \le \widetilde{E}_0 - E_0 \le 8\Delta E_{N-1}\frac{1-|\gamma_0|^2}{|\gamma_0|^2}\left(1+\frac{\pi\Delta E_1}{\Delta E_{N-1}}\right)^{-2k}. \tag{3.3}$$

Unfortunately, this simplified bound often grossly overestimates the error.

Our analysis also indicates that thresholding has only a mild effect on the accuracy. We pick up a $1+\mathcal{O}(\sqrt{k\epsilon})$ prefactor and an additional term proportional to the sum of the discarded eigenvalues of $\boldsymbol{S}$, which in turn can be bounded $\epsilon_{\text{total}} \le (2k+1)\epsilon$. In practice, we usually have $\epsilon_{\text{total}} \approx \epsilon$ due to rapid spectral decay.

Combining Theorem 3.1 with Informal Theorem 1.2 (formalized as Theorem 2.7) leads directly to Informal Theorem 1.3. More precise but also more complex error bounds can be obtained by using the full power (3.2) of Theorem 3.1 directly with Theorem 2.7.

3.1. Proof of Theorem 3.1. Our proof is based on the observation that thresholding is equivalent to applying the Rayleigh–Ritz procedure with a subspace spanned

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

by the dominant left singular vectors of the Krylov matrix. Consider the Krylov matrix $\boldsymbol{K}$ defined and factorized as

$$\begin{aligned}\boldsymbol{K} &:= \begin{bmatrix}\boldsymbol{\varphi_{-k}} & \cdots & \boldsymbol{\varphi_k}\end{bmatrix} \\ &= \underbrace{\begin{bmatrix}\boldsymbol{\psi}_0 & \cdots & \boldsymbol{\psi}_{N-1}\end{bmatrix}}_{:=\boldsymbol{\Psi}} \underbrace{\begin{bmatrix}\gamma_0 & & \\ & \ddots & \\ & & \gamma_{N-1}\end{bmatrix}}_{:=\boldsymbol{\Gamma}} \underbrace{\begin{bmatrix}\mathrm{e}^{-\mathrm{i}kE_0} & \cdots & \mathrm{e}^{\mathrm{i}kE_0} \\ \vdots & \ddots & \vdots \\ \mathrm{e}^{-\mathrm{i}kE_{N-1}} & \cdots & \mathrm{e}^{\mathrm{i}kE_{N-1}}\end{bmatrix}}_{:=\boldsymbol{F}}.\end{aligned} \tag{3.4}$$

Then we easily see that $\boldsymbol{H} = \boldsymbol{K}^*\widehat{\boldsymbol{H}}\boldsymbol{K}$ and $\boldsymbol{S} = \boldsymbol{K}^*\boldsymbol{K}$. Since the eigenvalues of $\boldsymbol{K}^*\boldsymbol{K}$ are the squares of singular values of $\boldsymbol{K}$ with eigenvectors equal to the right singular vectors of $\boldsymbol{K}$, it follows that the thresholded problem is precisely the Rayleigh–Ritz procedure applied to the left singular subspace of $\boldsymbol{K}$ with singular values larger than $\sqrt{\epsilon}$. From these left singular vectors, we are able to reconstruct the matrix $\boldsymbol{K}$ up to a Frobenius norm error $\sqrt{\epsilon_{\text{total}}}$. We thus can analyze QSD with thresholding in much the same way as Saad's analysis of the Lanczos method [29] with two twists. First, we have trigonometric polynomials in place of polynomials owing to the QSD algorithm's use of the unitary time-evolution operator. Second, our trigonometric basis functions are perturbed as a result of the truncated singular value decomposition.

With this roadmap in mind, we begin with a trigonometric version of a classic result in polynomial approximation theory.

LEMMA 3.3. *Let $0 < a < \pi$ and denote by $\mathcal{T}_k$ the space of degree $\le k$ trigonometric polynomials. The trigonometric polynomial minimax approximation problem*

$$\beta(a,k) = \min_{\substack{p\in\mathcal{T}_k \\ p(0)=1}} \max_{t\in(-\pi,\pi)\setminus(-a,a)} |p(t)|$$

is solved by

$$p^*(\theta) = \frac{T_k(1 + 2\frac{\cos\theta - \cos a}{\cos a + 1})}{T_k(1 + 2\frac{1-\cos a}{\cos a+1})}, \tag{3.5}$$

where T_k denotes the kth Chebyshev polynomial. The optimum value is

$$\beta(a,k) = \left(T_k\left(1 + 2\tfrac{1-\cos a}{\cos a+1}\right)\right)^{-1} \le 2\left(1 + 2\sqrt{\frac{1-\cos a}{\cos a + 1}}\right)^{-k} \le 2(1+a)^{-k}. \tag{3.6}$$

Proof. Once one convinces oneself that an optimal solution can be taken to be real and even, this result follows immediately from the analogous result for polynomial approximation [2, Thm. 4.1.11] together with the standard reparametrization $f(x) \mapsto f^\circ(\theta) := f(\cos\theta)$ which puts into bijection algebraic and even trigonometric polynomials. ∎

We shall need a bound for the optimal trigonometric polynomial (3.5).

PROPOSITION 3.4. *The trigonometric polynomial p^* defined in* (3.5) *is bounded in absolute value by* 1 *and satisfies the L^2 bound*

$$\int_{-\pi}^{\pi} |p^*(\theta)|^2 \, d\theta \le 2a + (2\pi - 2a)(\beta(a,k))^2 \le 2\pi. \tag{3.7}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Proof. First, we show that p^* is monotone on $[0, a]$ by showing its derivative can't have a zero on $(0, a)$. Up to a scaling factor and a change of variables on the input, p^* coincides with T_k on $[a, \pi]$. Thus, p^* has $k-1$ local extrema on (a, π) and symmetrically $k-1$ on $(-\pi, -a)$. Since p^* is even and 2π-periodic, p^* has local extrema at 0 and π. Since $(p^*)'$ is a degree-k trigonometric polynomial, it has at most $2k$ zeros, all of which have already been accounted for. Thus, p^* is monotone on $[0, a]$.

Since $p^*(a) \leq \beta(a, k) < 1$, p^* is monotone decreasing on $[0, a]$ and thus achieves its maximum value on $[-a, a]$ of 1 at 0. On $(-\pi, \pi) \setminus [-a, a]$, $|p^*| \leq \beta(a, k)$, from which the bound (3.7) follows. ∎

We shall be content with using the looser upper bound 2π in (3.7) in our subsequent analysis. With these approximation results in hand, we prove Theorem 3.1.

Proof of Theorem 3.1. Let $\boldsymbol{\varphi}_0$ be expanded as (3.1) and consider the Krylov matrix and its factorization defined in (3.4). Let $\widetilde{\boldsymbol{K}}$ represent the truncation of $\boldsymbol{K}$ by settings its singular values which are at most $\sqrt{\epsilon}$ to zero and factor it as $\widetilde{\boldsymbol{K}} = \boldsymbol{\Psi}\boldsymbol{\Gamma}\widetilde{\boldsymbol{F}}$ so that

$$\epsilon_{\text{total}} = \|\widetilde{\boldsymbol{K}} - \boldsymbol{K}\|_{\text{F}}^2 = \|\boldsymbol{\Gamma}(\widetilde{\boldsymbol{F}} - \boldsymbol{F})\|_{\text{F}}^2 = \sum_{i=0}^{N-1} \sum_{j=-k}^{k} |\gamma_i|^2 \underbrace{\left|\widetilde{f}_{ij} - \mathrm{e}^{\mathrm{i}jE_i}\right|^2}_{:=\alpha_{ij}^2}, \tag{3.8}$$

where $\widetilde{f}_{ij}$ denotes the ij entry of $\widetilde{\boldsymbol{F}}$.

Let $\mathcal{R}_\epsilon$ denote the range of $\widetilde{\boldsymbol{K}}$. By the Courant–Fischer theorem [34, Cor. IV.4.7],

$$\widetilde{E}_0 - E_0 = \min_{|\xi\rangle \in \mathcal{R}_\epsilon \setminus \{\boldsymbol{0}\}} \frac{\boldsymbol{x}\boldsymbol{i}^*(\widehat{\boldsymbol{H}} - E_0\boldsymbol{I})\boldsymbol{\xi}}{\boldsymbol{\xi}^*\boldsymbol{\xi}}, \tag{3.9}$$

where $\boldsymbol{I}$ denotes the identity. Since $\widehat{\boldsymbol{H}} - E_0\boldsymbol{I}$ is positive semidefinite, we have $\widetilde{E}_0 - E_0 \geq 0$. The remainder of our effort will be dedicated to establishing an upper bound.

We shall use the minimax optimal trigonometric polynomial p^* (3.5) to construct an ansatz $\boldsymbol{\xi}_\star$ to plug into (3.9). Let $p^*(\theta - \pi E_0/\Delta E_{M-1}) = \sum_{j=-k}^{k} c_j \mathrm{e}^{\mathrm{i}j\theta}$ be the Fourier series of the $\pi E_0/\Delta E_{M-1}$-translate of p^*. Choose as ansatz

$$\boldsymbol{\xi}_\star := \sum_{i=0}^{N-1} \sum_{j=-k}^{k} \gamma_i \widetilde{f}_{ij} c_j \boldsymbol{\psi}_i \in \mathcal{R}_\epsilon.$$

Plugging this into (3.9), we obtain an upper bound

$$\widetilde{E}_0 - E_0 \leq \frac{\boldsymbol{\xi}_\star^*(\widehat{\boldsymbol{H}} - E_0\boldsymbol{I})\boldsymbol{\xi}_\star}{\boldsymbol{\xi}_\star^*\boldsymbol{\xi}_\star} = \frac{\sum_{i=1}^{N-1} \Delta E_i |\gamma_i|^2 \left|\sum_{j=-k}^{k} \widetilde{f}_{ij} c_j\right|^2}{\sum_{i=0}^{N-1} |\gamma_i|^2 \left|\sum_{j=-k}^{k} \widetilde{f}_{ij} c_j\right|^2}. \tag{3.10}$$

First, focus on the numerator of the final bound in (3.10). We bound

$$\left|\sum_{j=-k}^{k} \widetilde{f}_{ij} c_j\right|^2 \leq \left|\sum_{j=-k}^{k} (\widetilde{f}_{ij} - \mathrm{e}^{\mathrm{i}jE_i}) c_j + \sum_{j=-k}^{k} \mathrm{e}^{\mathrm{i}jE_i} c_j\right|^2 \leq \left(\sum_{|j|\leq k} \alpha_{ij} c_j + p^*(E_i - E_0)\right)^2,$$

where $\alpha_{ij} \geq 0$ is defined in (3.8).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

First suppose that $1 \le i \le M$. Then by the fact that $|p^*(E_i - E_0)| \le \beta(a,k)$ as defined in (3.6) with $a := \pi\Delta E_1/\Delta E_M$, the Parseval theorem (i.e., $2\pi\sum_{j=-k}^{k}|c_j|^2 = \int_{-\pi}^{\pi}|p^*(\theta)|^2\,d\theta$), and the bound (3.7) we obtain

$$\left|\sum_{j=-k}^{k}\widetilde{f}_{ij}c_j\right|^2 \le 2\left(\sum_{j=-k}^{k}|c_j|^2\right)\left(\sum_{j=-k}^{k}\alpha_{ij}^2\right) + 2(\beta(a,k))^2 \le 2\sum_{j=-k}^{k}\alpha_{ij}^2 + 2(\beta(a,k))^2. \tag{3.11}$$

For $M < i < N$, we get the same bound except with 1 in place of $\beta(a,k)$ since $|p^*| \le 1$ by Proposition 3.4.

For the denominator of the final bound in (3.10), we bound

$$\left|\sum_{j=-k}^{k}\widetilde{f}_{0j}c_j\right| \ge 1 - \sum_{j=-k}^{k}\alpha_{0j} \ge 1 - \sqrt{(2k+1)\sum_{j=-k}^{k}\alpha_{0j}^2} \ge 1 - \frac{1}{|\gamma_0|}\sqrt{(2k+1)\epsilon}, \tag{3.12}$$

where we used a spectral norm bound similar to (3.8):

$$\epsilon \ge \|\widetilde{\boldsymbol{K}} - \boldsymbol{K}\| = \|\boldsymbol{\Gamma}(\widetilde{\boldsymbol{F}} - \boldsymbol{F})\| \ge \sqrt{|\gamma_0|^2\sum_{j=-k}^{k}\left|\widetilde{f}_{ij} - \mathrm{e}^{\mathrm{i}jE_i}\right|^2} = |\gamma_0|\sqrt{\sum_{j=-k}^{k}\alpha_{0j}^2}.$$

Plugging (3.11) and (3.12) into (3.10),

$$\begin{aligned}\widetilde{E}_0 - E_0 &\le \frac{2\left[\sum_{i=1}^{M}\Delta E_i\left(\sum_{j=-k}^{k}\alpha_{ij}^2 + (\beta(a,k))^2\right) + \sum_{i=M+1}^{N-1}\Delta E_i\left(\sum_{j=-k}^{k}\alpha_{ij}^2 + 1\right)\right]}{\left(|\gamma_0| - \sqrt{(2k+1)\epsilon}\right)^2} \\ &\le \frac{2\left[\Delta E_{N-1}\epsilon_{\text{total}} + (\beta(a,k))^2\sum_{i=1}^{M}\Delta E_i|\gamma_i|^2 + \sum_{i=M+1}^{N-1}\Delta E_i|\gamma_i|^2\right]}{|\gamma_0|^2 - 2|\gamma_0|\sqrt{(2k+1)\epsilon}}.\end{aligned}$$

Using the bound (3.6) with $a = \pi\Delta E_1/\Delta E_M$ leads precisely to (3.2). ∎

4. Additional results and discussions. In this section, we include some additional results and discussions which follow from our analysis but are not directly germane to the main analysis of the QSD algorithm comprising Theorems 2.7 and 3.1.

4.1. On the Toeplitz structure of $\boldsymbol{H}, \boldsymbol{S}$. With the choice of the basis vectors $\boldsymbol{\varphi}_j = \mathrm{e}^{\mathrm{i}t_j\widehat{\boldsymbol{H}}}\boldsymbol{\varphi}_0$ as in the Parrish–McMahon QSD procedure, the matrix elements of the projected matrices satisfy

$$\boldsymbol{H}_{jk} = \boldsymbol{\varphi}_j^*\widehat{\boldsymbol{H}}\boldsymbol{\varphi}_k = \boldsymbol{\varphi}_0^*\widehat{\boldsymbol{H}}\mathrm{e}^{\mathrm{i}(t_k-t_j)\widehat{\boldsymbol{H}}}\boldsymbol{\varphi}_0 \quad \text{and} \quad \boldsymbol{S}_{jk} = \boldsymbol{\varphi}_j^*\boldsymbol{\varphi}_k = \boldsymbol{\varphi}_0^*\mathrm{e}^{\mathrm{i}(t_k-t_j)\widehat{\boldsymbol{H}}}\boldsymbol{\varphi}_0.$$

Therefore, both $\boldsymbol{H}, \boldsymbol{S}$ are Hermitian–Toeplitz matrices. Unfortunately, the Toeplitz structure relies on the assumption that the Hamiltonian simulation problem (i.e., $\mathrm{e}^{\mathrm{i}t_j\widehat{\boldsymbol{H}}}\boldsymbol{\varphi}_0$) is computed exactly. In practice, the Hamiltonian simulation problem is often performed with approximate techniques (such as Trotter splitting), and the resulting projected matrices $\widetilde{\boldsymbol{H}}$ and $\widetilde{\boldsymbol{S}}$ may not have the Toeplitz structure. If this is

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

the case, all n^2 entries of $\widetilde{\boldsymbol{H}}$ (and perhaps $\widetilde{\boldsymbol{S}}$ as well) need to be computed to apply the Rayleigh–Ritz procedure to the computed basis states $\boldsymbol{\varphi}_0, \dots, \boldsymbol{\varphi}_{n-1}$ in earnest. However, if one measures only the first row of $\boldsymbol{H}$ and $\boldsymbol{S}$ and imputes the remaining entries from the Hermitian–Toeplitz structure, then resulting recovered matrices $\widetilde{\widetilde{\boldsymbol{H}}}$ and $\widetilde{\widetilde{\boldsymbol{S}}}$ represent the true $\boldsymbol{H}$ and $\boldsymbol{S}$ corrupted by both Monte Carlo and discretization errors. Our main analysis makes no use of the Toeplitz structure.

An important question for the QSD procedure is how entrywise errors in the entries $\boldsymbol{H}$ and $\boldsymbol{S}$ correspond to errors in the spectral norm. For the standard QSD estimation procedure, the entries of $\boldsymbol{M} \in \{\boldsymbol{H}, \boldsymbol{S}\}$ are approximated by averaging m unbiased estimators each with maximum error B ($B = \mathcal{O}(1)$ for $\boldsymbol{S}$ and $B = \mathcal{O}(\|\boldsymbol{H}\|)$ for $\boldsymbol{H}$). Consider the case where the Hamiltonian simulation problem is solved exactly and we compute estimates for $\boldsymbol{M} \in \{\boldsymbol{H}, \boldsymbol{S}\}$ by measuring the first row of $\boldsymbol{M}$ and computing the remaining entries from the Hermitian–Toeplitz structure. Straightforward application of matrix concentration inequalities then shows that the approximation is $\mathcal{O}(B\sqrt{(n \log n)/m})$-close to $\boldsymbol{M}$ (see, e.g., [35, Thms. 3.6.1 and 4.6.1]).

4.2. Stability of best low-rank approximation. Theorem 2.6 constitutes a stability result for the Eckart–Young best rank-m approximation $[\![\boldsymbol{S}]\!]_m = \boldsymbol{\Pi}\boldsymbol{S}\boldsymbol{\Pi}$ to $\boldsymbol{S}$. Since this result may be of independent interest, we state it here unburdened by the particularities of the QSD context.

THEOREM 4.1. *Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ be a positive semidefinite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. Let $[\![\boldsymbol{A}]\!]_m$ denote the Eckart–Young best rank-m approximation to $\boldsymbol{A}$. Then, for any quadratic unitarily invariant norm $\|\cdot\|_{\mathrm{QUI}}$ (such as the spectral or Frobenius norms; see* [1, Def. IV.2.9]*),*

$$\|[\![\boldsymbol{A}+\boldsymbol{\Delta}]\!]_m - [\![\boldsymbol{A}]\!]_m\|_{\mathrm{QUI}} \leq \|\boldsymbol{\Delta}\|_{\mathrm{QUI}} + \frac{2n\lambda_m\|\boldsymbol{\Delta}\|}{\lambda_m - \lambda_{m+1} - \|\boldsymbol{\Delta}\|}\left(1 + \frac{0.5\,n\|\boldsymbol{\Delta}\|}{\lambda_m - \lambda_{m+1} - \|\boldsymbol{\Delta}\|}\right).$$

Proof. This follows immediately from Theorem 2.6 and the maximality of the Frobenius norm among quadratic unitarily invariant norms [1, eq. (IV.38)]. ∎

This bound is interesting because, in the setting where the gap $\lambda_m - \lambda_{m+1} - \|\boldsymbol{\Delta}\|$ is comparable in size to λ_m, the best rank-m approximation changes only by $\approx n\|\boldsymbol{\Delta}\|_{\mathrm{QUI}}$ independent of the approximation error $\|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_m\|_{\mathrm{QUI}}$. This can be a large improvement over the general-purpose bound (similar to [8, Cor. 2.4]) which holds for general unitarily invariant norms $\|\cdot\|_{\mathrm{UI}}$,

$$\|[\![\boldsymbol{A}+\boldsymbol{\Delta}]\!]_m - [\![\boldsymbol{A}]\!]_m\|_{\mathrm{UI}} \leq 2(\|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_m\|_{\mathrm{UI}} + \|\boldsymbol{\Delta}\|_{\mathrm{UI}}),$$

which does depend on the approximation error $\|\boldsymbol{A} - [\![\boldsymbol{A}]\!]_m\|_{\mathrm{UI}}$. We believe the dimensional factor n is likely quite pessimistic and can hopefully be replaced by a small constant (at least) if the matrix enjoys a rapidly decaying spectrum.

4.3. General analysis of thresholding. We have presented an analysis in Theorem 3.1 of thresholding for the QSD problem, which simultaneously treats the Rayleigh–Ritz error from approximation from the titular quantum subspace and the thresholding procedure together. This is quite natural as the total error is what is most important to the practitioner, and one should in principle be able to obtain a more precise error bound by not decoupling these pieces. However, it remains a question of mathematical interest and of interest to the broader uses of thresholding beyond QSD to provide an analysis of thresholding for general matrices.

As the bad example (1.5) shows, thresholding may not work for general matrices, with thresholding with parameter ϵ able to introduce errors $\gg \epsilon$. This behavior

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

is less surprising, however, when one notes that neither eigenvalue of (1.5) is well-conditioned, with both having a condition number $\Theta(\epsilon^{-1})$. In fact, this is the only obstruction to thresholding working (at least for recovering the smallest eigenvalue), as we shall show that the least eigenvalue is recovered accurately by thresholding if it is well-conditioned. (We recall [34, Thm. VI.2.2] that the condition number of the eigenangle $\tan^{-1} E_0$ associated with the least eigenvalue E_0 of $(\boldsymbol{H}, \boldsymbol{S})$ is $\|\boldsymbol{c}_0\|^2\sqrt{1+E_0^2}$ where $\boldsymbol{c}_0$ is the $\boldsymbol{S}$-normalized eigenvector, $\boldsymbol{c}_0^*\boldsymbol{S}\boldsymbol{c}_0 = 1$, associated with E_0.)

THEOREM 4.2. *Let $\boldsymbol{c}_0$ be the $\boldsymbol{S}$-normalized eigenvector associated with the least eigenvalue E_0 of the generalized eigenvalue problem* (1.2) *for a Hermitian and Hermitian positive matrix $\boldsymbol{H}$ and $\boldsymbol{S}$. Then, for $\widetilde{E}_0$ the least eigenvalue recovered by thresholding with parameter ϵ and provided $2\sqrt{\epsilon}\|\boldsymbol{c}_0\| < 1$,*

$$0 \le \widetilde{E}_0 - E_0 \le \frac{\Delta E\, \epsilon \|\boldsymbol{c}_0\|^2}{1 - 2\sqrt{\epsilon}\|\boldsymbol{c}_0\|},$$

where ΔE is the difference between the largest and smallest eigenvalue of $(\boldsymbol{H}, \boldsymbol{S})$.

Proof. Let $\mathcal{R}_\epsilon$ be the span of the eigenvectors of $\boldsymbol{S}$ with eigenvalue greater than ϵ. Using the same observation which motivated the proof of Theorem 3.1, we have

$$\widetilde{E}_0 - E_0 = \min_{\boldsymbol{c} \in \mathcal{R}_\epsilon \setminus \{\boldsymbol{0}\}} \frac{\boldsymbol{c}^*(\boldsymbol{H} - E_0\,\boldsymbol{S})\boldsymbol{c}}{\boldsymbol{c}^*\boldsymbol{S}\boldsymbol{c}}. \tag{4.1}$$

In particular, since $\boldsymbol{H} - E_0\boldsymbol{S}$ is positive semidefinite, this implies $\widetilde{E}_0 - E_0 \ge 0$ so we just need to concern ourselves with obtaining an upper bound. Letting $\widetilde{\boldsymbol{S}}$ be the $\boldsymbol{S}$ matrix with all its eigenvalues at most ϵ set to zero, we shall evaluate (4.1) at $\widetilde{\boldsymbol{c}}_0 := \boldsymbol{S}^{-1/2}\widetilde{\boldsymbol{S}}^{1/2}\boldsymbol{c}_0$, obtaining an upper bound. Defining the error $\boldsymbol{\delta} := \widetilde{\boldsymbol{c}}_0 - \boldsymbol{c}_0$, we have that $\boldsymbol{\delta}$ satisfies the bound $\left\|\boldsymbol{S}^{1/2}\boldsymbol{\delta}\right\| \le \left\|\boldsymbol{S}^{1/2} - \widetilde{\boldsymbol{S}}^{1/2}\right\|\|\boldsymbol{c}_0\| \le \sqrt{\epsilon}\|\boldsymbol{c}_0\|$. Thus

$$\widetilde{\boldsymbol{c}}_0^*\boldsymbol{S}\widetilde{\boldsymbol{c}}_0 \ge \boldsymbol{c}_0^*\boldsymbol{S}\boldsymbol{c}_0 - 2|\boldsymbol{c}_0^*\boldsymbol{S}\boldsymbol{\delta}| \ge 1 - 2(\boldsymbol{c}_0^*\boldsymbol{S}\boldsymbol{c}_0)^{1/2}(\boldsymbol{\delta}^*\boldsymbol{S}\boldsymbol{\delta})^{1/2} \ge 1 - 2\sqrt{\epsilon}\|\boldsymbol{c}_0\|$$

and

$$\widetilde{\boldsymbol{c}}_0^*(\boldsymbol{H} - E_0\boldsymbol{S})\widetilde{\boldsymbol{c}}_0 = \boldsymbol{\delta}^*(\boldsymbol{H} - E_0\boldsymbol{S})\boldsymbol{\delta} \le \left\|\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{S}^{-1/2} - E_0\boldsymbol{I}\right\| \left\|\boldsymbol{S}^{1/2}\boldsymbol{\delta}\right\|^2 \le \Delta E\, \epsilon \|\boldsymbol{c}_0\|^2.$$

Plugging $\widetilde{\boldsymbol{c}}_0$ into (4.1) and applying the previous two displays leads immediately to the stated result. ∎

Theorem 4.2 yields an alternative analysis of the QSD algorithm with thresholding: Combine Theorem 3.1 with $\epsilon = 0$ (to measure the Rayleigh–Ritz error in isolation) together with Theorem 4.2 (to measure the thresholding error). Compared to using Theorem 3.1 alone, this alternative analysis has the advantage that the two types of error (Rayleigh–Ritz and thresholding) can be bounded independently of each other. However, to use Theorems 3.1 and 4.2 in this way, an additional piece of information is needed beyond what is required by Theorem 3.1 alone, namely the norm $\|\boldsymbol{c}_0\|$.[4] See SM4 for a comparison of these two approaches on some numerical examples.

[4]The norm of the smallest eigenvector of the perturbed thresholded pair $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ can provide a good a posteriori estimate of this quantity in the limit of a small perturbation (see [16, sect. 6.2]).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

5. Numerical experiments. In this section, we present some numerical experiments demonstrating the success of the theory in explaining the performance of QSD and other features of the numerical performance of the QSD method. We consider two sets of examples: the one-dimensional transverse field Ising model (TFIM) and the one-dimensional Hubbard model.

The Hamiltonian of the TFIM with L spins is

$$\widehat{\boldsymbol{H}} = -\sum_{i=1}^{L} \widehat{\boldsymbol{Z}}_i \widehat{\boldsymbol{Z}}_{i+1} - g \sum_{i=1}^{L} \widehat{\boldsymbol{X}}_i, \tag{5.1}$$

where $\widehat{\boldsymbol{X}}_i, \widehat{\boldsymbol{Z}}_i$ are the Pauli X, Z operators acting on the ith spin, respectively. We use the periodic boundary condition, and therefore $\widehat{\boldsymbol{Z}}_{L+1}$ is identified with $\widehat{\boldsymbol{Z}}_1$.

The Hamiltonian for the (spinful) Hubbard model of L sites is

$$\widehat{\boldsymbol{H}} = -\sum_{i=1}^{L} \sum_{\sigma \in \{\uparrow,\downarrow\}} \widehat{\boldsymbol{a}}_{i\sigma}^* \widehat{\boldsymbol{a}}_{i+1,\sigma} + U \sum_{i=1}^{L} \widehat{\boldsymbol{a}}_{i\uparrow}^* \widehat{\boldsymbol{a}}_{i\uparrow} \widehat{\boldsymbol{a}}_{i\downarrow}^* \widehat{\boldsymbol{a}}_{i\downarrow}. \tag{5.2}$$

Here $\widehat{\boldsymbol{a}}_{i\sigma}^*, \widehat{\boldsymbol{a}}_{i\sigma}$ are the fermionic creation and annihilation operators at site i with spin σ, which can be expressed in terms of spin operators following the Jordan–Wigner transformation (see, e.g., [23]). Similarly due to periodic boundary conditions, $\widehat{\boldsymbol{a}}_{L+1,\sigma}^*, \widehat{\boldsymbol{a}}_{L+1,\sigma}$ are identified with $\widehat{\boldsymbol{a}}_{1\sigma}^*, \widehat{\boldsymbol{a}}_{1\sigma}$, respectively. We are interested in finding the ground state energy of $\widehat{\boldsymbol{H}}$. The dimension of $\widehat{\boldsymbol{H}}$ of the TFIM is 2^L and that for the Hubbard model is 2^{2L} (due to spin degrees of freedom). Due to the high dimensionality, these models are generally intractable to be solved directly when L is large.[5] In all examples below, we set $L = 10$, $g = -\sqrt{2}$ for the TFIM model and $L = 10$, $U = 8$ for the Hubbard model.[6] The Hubbard model also has an extra parameter called the total number of fermions denoted by N_e, which constrains $\widehat{\boldsymbol{H}}$ to a smaller diagonal block, and its value is set to the half filling with $N_e = L = 10$. We find that the numerical results do not depend sensitively on these parameters. Additional numerical results with other values of L, U, etc., can be found in the accompanying supplemental file supplement.pdf [local/web 2.10MB].

The number of time steps used shall be denoted by n, and the time grid is $t_j = j\Delta t$ where $j = 0, \ldots, n-1$. The initial vector $\boldsymbol{\varphi}_0$ of the TFIM model is taken to be a product state (an eigenstate with $g = 0$), and that of the Hubbard model is taken to be a Slater determinant state (an eigenstate with $U = 0$), respectively. We find that such a setup leads to a sufficiently large initial overlap $|\gamma_0|^2 = |\langle \varphi_0 \,|\, \psi_0 \rangle|^2$. This ensures that in the noiseless setting, the ground state energy of $\widehat{\boldsymbol{H}}$ can be estimated to high accuracy with a very modest value of n (see Table 1).

The Hamiltonian simulation and the computation of the projected matrix elements are performed using the QuSpin package [37, 38] in Python on a classical computer. All further experiments are performed in MATLAB, with the noise matrices $\boldsymbol{\Delta}_{\boldsymbol{H}}$ and $\boldsymbol{\Delta}_{\boldsymbol{S}}$ modeled as complex Gaussian Hermitian–Toeplitz matrices with the entries in the first rows independent with specified variances.

[5] There exist special techniques that are particularly efficient for handling one-dimensional quantum systems. The main point of our numerical results is to demonstrate the performance of QSD algorithms, which does not rely on such special properties.

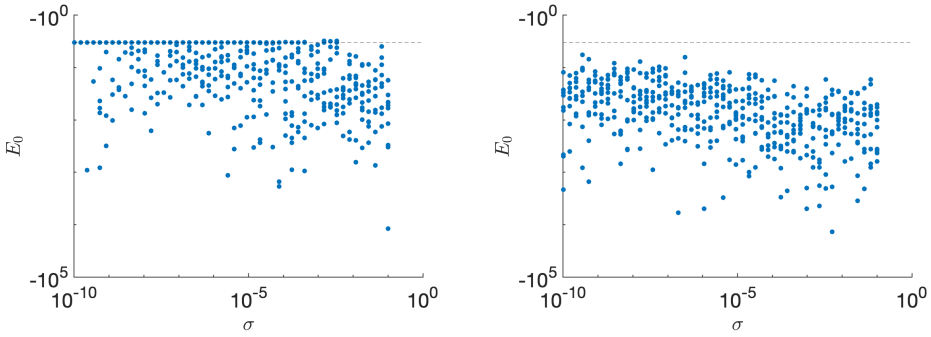
[6] In particular, $U = 8.0$ for the Hubbard model corresponds to a strongly correlated quantum system and is typically considered to be difficult.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

TABLE 1

Some parameters for the TFIM and the Hubbard models. Exact E_0 is obtained by diagonalizing $\widehat{\boldsymbol{H}}$. Noiseless E_0 is obtained by solving the projected generalized eigenvalue problem (of size $n \times n$) using the QSD algorithm in the noiseless setting.

	TFIM ($L = 10$)	Hubbard ($L = 10$)
Time step Δt	1.0	0.1
Initial overlap $\lvert\gamma_0\rvert^2$	0.079	0.122
Exact E_0	$-15.979975\mathbf{0}$	$-3.3149967\mathbf{3}$
Noiseless E_0 with $n = 40$	$-15.979974\mathbf{8}$	$-3.3149967\mathbf{0}$



(a) $n = 20$ (b) $n = 80$

FIG. 2. *Least eigenvalues computed from the perturbed pair $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ without any procedure to ameliorate the affects of noise. Shown are 10 random initializations of the noise for several random noise levels σ for the Hubbard example with $n = 20$ (left) and $n = 80$ (right). The true eigenvalue is shown as a horizontal dashed line.*

5.1. The need for thresholding. We first demonstrate why we advocate the use of thresholding by showing the potential pitfalls of some other strategies. Naturally, the first strategy one might attempt would be to do nothing at all: just solve the generalized eigenvalue problem

$$\widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{c}} = \widetilde{E}\,\widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{c}} \tag{5.3}$$

and return the least computed eigenvalue. The futility of this strategy is shown in Figure 2. Even for extremely low noise levels ($\sigma \approx 10^{-10}$), we see that the recovered least eigenvalue can deviate quite far from the genuine least eigenvalue with high probability. For nicer problem instances (see, e.g., Figure 2(a)), characterized by an only modestly ill-conditioned (e.g., $\kappa(\boldsymbol{S}) = \|\boldsymbol{S}\|\|\boldsymbol{S}^{-1}\| \lesssim 10^{12}$) $\boldsymbol{S}$ matrix, the eigenvalue of interest could perhaps be reliably recovered by taking a median over multiple trials (each requiring the quantum computation to be rerun). However, for problem instances with a more ill-conditioned $\boldsymbol{S}$ matrix, such as shown in Figure 2(b), the probability of finding an eigenvalue close to the genuine smallest eigenvalue appears to occur with vanishingly small probability.

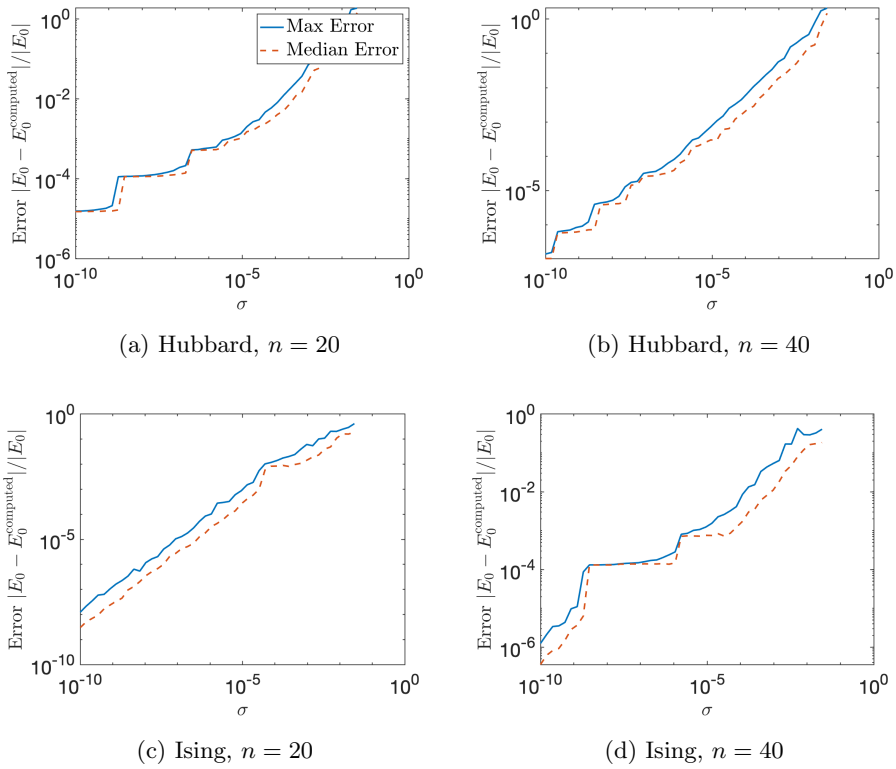
Alternately, one might try to apply "just a bit of thresholding" by setting the threshold parameter at a small constant value, independent of the noise level. This can be modestly effective for a well-conditioned $\boldsymbol{S}$ matrix (particularly if combined with a median of multiple trials), but it falls down as soon as $\sigma \gtrapprox \epsilon$ in general. See Figure SM6 for a demonstration of this. If one is to rely on thresholding alone to deal with the noise, then the threshold parameter must be chosen large enough.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

As another alternative to thresholding, one might attempt to solve the problem without explicitly filtering out the noise by thresholding (or using only a tiny threshold much smaller than the noise level) and attempting to systematically determine which eigenvalues are "real." We investigate such strategies in section SM2 and ultimately conclude that these are less robust and less accurate than thresholding.

5.2. Choice of the thresholding parameter. Now that we have demonstrated why we prefer thresholding for solving the noise-perturbed QSD generalized eigenvalue problem (5.3), let us demonstrate the success of thresholding. In Figure 3, we demonstrate the error for the thresholding procedure with a threshold parameter ϵ proportional to the noise level.[7] (See also Figure SM7 for more examples.) As the plots show, thresholding is robust on these examples in the sense that the maximum error over multiple trials is similar to the median, showing that thresholding is reliably able to filter out the noise over different random initializations.[8] Since the norm of the noise matrix is a random quantity prone to occasionally being appreciably larger



(a) Hubbard, $n = 20$ (b) Hubbard, $n = 40$

(c) Ising, $n = 20$ (d) Ising, $n = 40$

FIG. 3. *Maximum (blue solid) and median (red dashed) error over* 100 *initializations for eigenvalues computed from the noise-perturbed pair* $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ *using thresholding with threshold parameter* $25\sigma\|\widetilde{\boldsymbol{S}}\|$ *for Hubbard model (top) and Ising model (bottom) with* $n = 20$ *(left) and* 40 *(right).*

[7] This is smaller than the theory (Theorem 2.5 and Corollary 2.2) predicts ϵ should be taken as, which suggests that one should choose $\epsilon \propto \sigma^{1/(1+\alpha)}$ ($\alpha \leq 1/2$ is the value for which (2.12) holds).

[8] This is seen to be not true for one example (Figure SM7d) in the Supplementary Materials. This shows that thresholding is not infallible but is still better than alternate strategies within our knowledge (section SM2).

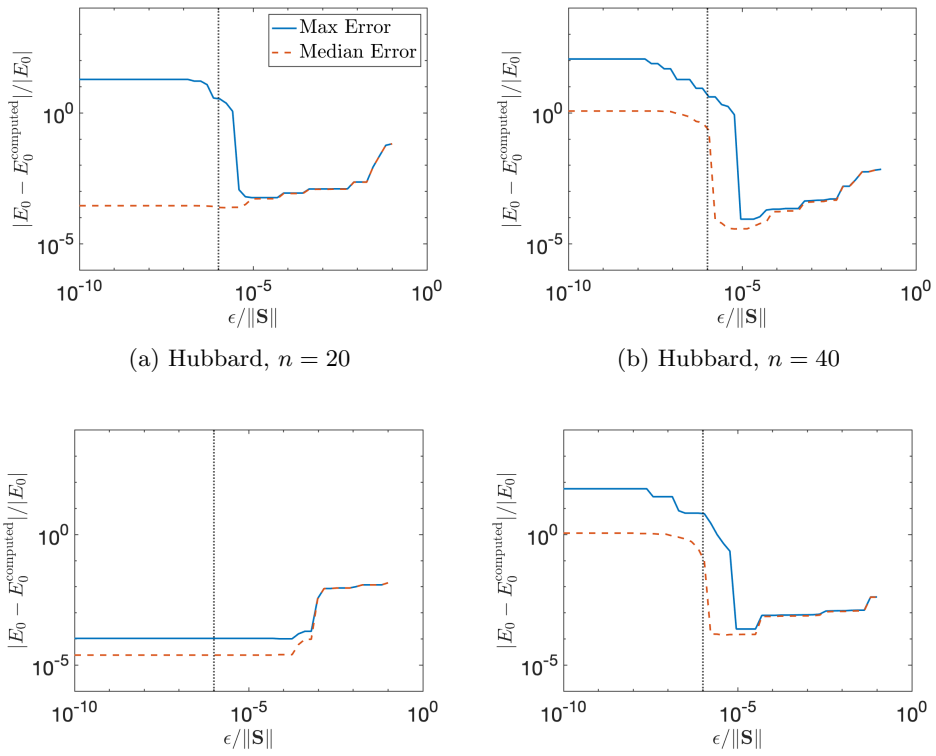
Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

than its average value, the threshold parameter must be somewhat larger than the expected noise level to achieve good performance with high probability. These plots demonstrate that, for these examples at least, this multiple can be quite modest—just 25 is enough. In the error plots in Figure 3 we see two types of behavior: sometimes the error increases relatively continuously with the thresholding parameter (e.g., most of Figure 3(c)) whereas other times it increases in a more stepwise fashion (e.g., low noise levels in Figure 3(a)). The first behavior is indicative of the error being dominated by the noise level (with the slope in log-log space being ≈ 1 demonstrating a linear dependence of the error on the noise level) with the second exemplifying the thresholding error being the dominant contribution. Behavior of the second type might suggest that the threshold parameter is being chosen conservatively and lower error could be achieved with a smaller threshold value.

The choice of the thresholding parameter is critical to the success of the method, as is demonstrated in Figure 4. In these plots, we show the median and maximum error for the computed eigenvalue over 100 random initializations of the noise (fixed to $\sigma = 10^{-6}$) for different thresholding levels. In the normative case, the error decreases as the threshold parameter is decreased up until the threshold parameter reaches a modest multiple of the noise level, after which the error sharply rises. In the Supplementary Materials (section SM3), we discuss an automatically tuned variant of the thresholding procedure which can help in picking a good threshold level ϵ.



(a) Hubbard, $n = 20$

(b) Hubbard, $n = 40$

(c) Ising, $n = 20$

(d) Ising, $n = 40$

FIG. 4. *Maximum (blue solid) and median (red dashed) error over* 100 *initializations for eigenvalues computed from the noise-perturbed pair* $(\widetilde{\boldsymbol{H}}, \widetilde{\boldsymbol{S}})$ *using thresholding for various values of the threshold* ϵ *for a fixed noise level* $\sigma = 10^{-6}$ *(dotted black line) for Hubbard model (top) and Ising model (bottom) with* $n = 20$ *(left) and* 40 *(right).*

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

6. Conclusions. In this article, we have presented the first theoretical analysis of the accuracy of the quantum subspace diagonalization method with a thresholding procedure. Our explanation has two parts:

1. With an appropriate choice of time sequence $\{t_j\}$, QSD with thresholding (Algorithm 1.1) can compute an accurate approximation to the ground-state energy (Theorem 3.1).
2. Under appropriate conditions, the thresholding procedure can robustly determine the smallest eigenvalue in the presence of perturbations to the pair $(\boldsymbol{H}, \boldsymbol{S})$ (Theorem 2.7).

These two pieces combine to give a bound on the total error of the QSD procedure (comprising Rayleigh–Ritz, thresholding, and perturbation errors) in Informal Theorem 1.3. The conditions of our theory are natural, and many of the parameters in our bounds can be estimated in the presence of noise, allowing our bounds to be able to give approximate bounds on the error a posteriori. Our numerical experiments (including additional experiments in the accompanying supplemental file supplement.pdf [local/web 2.10MB]) support the conclusion that QSD is accurate when implemented with thresholding (and not accurate when implemented without).

Our theoretical estimates can still be significantly improved, such as the bound $\chi_{\boldsymbol{H}} \leq \mathcal{O}(\eta_{\boldsymbol{S}}/\epsilon^{\alpha} + \eta_{\boldsymbol{H}})$ (see Theorem 2.5) for the discrepancy between the thresholded $\boldsymbol{H}$ and $\widetilde{\boldsymbol{H}}$ matrices, where $0 \leq \alpha \leq 1/2$. This suggests that we need to take $\epsilon = \Omega(\eta_{\boldsymbol{S}}^{1/(1+\alpha)})$ for accurate recovery to be guaranteed by Corollary 2.2, which has hypothesis $\chi \leq \epsilon/q$. This is in contradiction to our numerical experiments, where ϵ can be chosen to be a small multiple of η. Synthetically generated worst-case examples (see section SM5) suggest the bound $\chi_{\boldsymbol{H}} \leq \mathcal{O}(\eta_{\boldsymbol{S}}/\epsilon^{\alpha}+\eta_{\boldsymbol{H}})$ is tight, but it remains possible a better bound can be derived for pairs $(\boldsymbol{H}, \boldsymbol{S})$ generated by QSD. An interesting open question is to give a convincing explanation for why we appear to have $\alpha = 1/4$ and $\mu \approx \max|\Lambda(\boldsymbol{H}, \boldsymbol{S})|$ in (2.12) for many QSD instances. Despite the modest size of n in practice, the polynomial dependence on n in Corollary 2.2 and Theorems 2.5, 2.6, and 4.1 can lead to significant overestimates of the error. Therefore, another natural question is whether these dimensional factors can be improved.

Natural extensions of this work are to generalize our analysis to excited states (interior eigenvalues) and to develop bounds on the accuracy of the computed eigenvectors. Although the QSD algorithm cannot be used to coherently prepare an eigenstate on a quantum computer, we may still compute other physical observables from the approximate eigenstate. Mathias and Li's theory [19, sect. 6] suggests eigenvectors might be more sensitive to the noise than the eigenvalues, and we plan to study the accuracy of the computed eigenvectors in future work.

Disclaimer. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Acknowledgments. The first author thanks the Lawrence Berkeley Laboratory summer student program for providing a welcoming environment to perform this work. We thank Yulong Dong, Yu Tong, Norman Tubman, and Robert Webber for helpful discussions.

REFERENCES

[1] R. Bhatia, *Matrix Analysis*, Grad. Texts in Math. 169, Springer-Verlag, New York, 1997, https://doi.org/10.1007/978-1-4612-0653-8.

[2] Å. Björck, *Numerical Methods in Matrix Computations*, Texts in Appl. Math. 59, Springer International, New York, 2015, https://doi.org/10.1007/978-3-319-05089-8.

[3] Y. Cai, Z. Bai, J. E. Pask, and N. Sukumar, *Hybrid preconditioning for iterative diagonalization of ill-conditioned generalized eigenvalue problems in electronic structure calculations*, J. Comput. Phys., 255 (2013), pp. 16–30.

[4] A. M. Childs, Y. Su, M. C. Tran, N. Wiebe, and S. Zhu, *Theory of Trotter error with commutator scaling*, Phys. Rev. X, 11 (2021), 011020, https://doi.org/10.1103/PhysRevX.11.011020.

[5] J. I. Colless, V. V. Ramasesh, D. Dahlen, M. S. Blok, M. E. Kimchi-Schwartz, J. R. McClean, J. Carter, W. A. de Jong, and I. Siddiqi, *Computation of molecular spectra on a quantum processor with an error-resilient algorithm*, Phys. Rev. X, 8 (2018), 011021, https://doi.org/10.1103/PhysRevX.8.011021.

[6] C. L. Cortes and S. K. Gray, *Quantum Krylov Subspace Algorithms for Ground and Excited State Energy Estimation*, https://arxiv.org/abs/2109.06868, 2021.

[7] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46, https://doi.org/10.1137/0707001.

[8] P. Drineas and I. C. F. Ipsen, *Low-rank matrix approximations do not need a singular value gap*, SIAM J. Matrix Anal. Appl., 40 (2019), pp. 299–319, https://doi.org/10.1137/18M1163658.

[9] A. Emami-Naeini and P. Van Dooren, *Computation of zeros of linear multivariable systems*, Automatica, 18 (1982), pp. 415–430, https://doi.org/10.1016/0005-1098(82)90070-X.

[10] G. Fix and R. Heiberger, *An algorithm for the ill-conditioned generalized eigenvalue problem*, SIAM J. Numer. Anal., 9 (1972), pp. 78–88, https://doi.org/10.1137/0709009.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, 2013.

[12] L. A. Gribov and B. K. Novosadov, *Use of overcomplete basis sets in quantum-chemical calculations*, J. Molecular Structure THEOCHEM, 136 (1986), pp. 387–389, https://doi.org/10.1016/0166-1280(86)80152-X.

[13] W. J. Huggins, J. Lee, U. Baek, B. O'Gorman, and K. B. Whaley, *A non-orthogonal variational quantum eigensolver*, New J. Phys., 22 (2020), 073009, https://doi.org/10.1088/1367-2630/ab867b.

[14] M. Jungen and K. Kaufmann, *The Fix–Heiberger procedure for solving the generalized ill-conditioned symmetric eigenvalue problem*, Int. J. Quantum Chem., 41 (1992), pp. 387–397.

[15] K. Klymko, C. Mejuto-Zaera, S. J. Cotton, F. Wudarski, M. Urbanek, D. Hait, M. Head-Gordon, K. B. Whaley, J. Moussa, N. Wiebe, W. A. de Jong, and N. M. Tubman, *Real Time Evolution for Ultracompact Hamiltonian Eigenstates on Quantum Hardware*, https://arxiv.org/abs/2103.08563, 2021.

[16] M. Lotz and V. Noferini, *Wilkinson's bus: Weak condition numbers, with an application to singular polynomial eigenproblems*, Found. Comput. Math., 20 (2020), pp. 1439–1473, https://doi.org/10.1007/s10208-020-09455-y.

[17] P.-O. Löwdin, *Group algebra, convolution algebra, and applications to quantum mechanics*, Rev. Modern Phys., 39 (1967), pp. 259–287, https://doi.org/10.1103/RevModPhys.39.259.

[18] V. A. Mandelshtam and H. S. Taylor, *Harmonic inversion of time signals and its applications*, J. Chem. Phys., 107 (1997), pp. 6756–6769, https://doi.org/10.1063/1.475324.

[19] R. Mathias and C.-K. Li, *The Definite Generalized Eigenvalue Problem: A New Perturbation Theory*, T-NAREP 457, Manchester Centre for Computational Mathematics, 2004.

[20] J. R. McClean, M. E. Kimchi-Schwartz, J. Carter, and W. A. de Jong, *Hybrid quantum-classical hierarchy for mitigation of decoherence and determination of excited states*, Phys. Rev. A, 95 (2017), 042308, https://doi.org/10.1103/PhysRevA.95.042308.

[21] M. Motta, C. Sun, A. T. K. Tan, M. J. O'Rourke, E. Ye, A. J. Minnich, F. G. S. L. Brandão, and G. K.-L. Chan, *Determining eigenstates and thermal states on a quantum computer using quantum imaginary time evolution*, Nature Physics, 16 (2020), pp. 205–210, https://doi.org/10.1038/s41567-019-0704-4.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

[22] G. Nannicini, *An introduction to quantum computing, without the physics*, SIAM Rev., 62 (2020), pp. 936–981.
[23] J. W. Negele and H. Orland, *Quantum Many-Particle Systems*, Westview, Boulder, CO, 1988.
[24] M. A. Nielsen and I. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge, UK, 2000.
[25] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics in Appl. Math. 20, SIAM, Philadelphia, 1998, https://doi.org/10.1137/1.9781611971163.
[26] R. M. Parrish and P. L. McMahon, *Quantum Filter Diagonalization: Quantum Eigendecomposition Without Full Quantum Phase Estimation*, https://arxiv.org/abs/1909.08925, 2019.
[27] J. Preskill, *Quantum computing in the NISQ era and beyond*, Quantum, 2 (2018), p. 79.
[28] J. Preskill, *Quantum Computing* 40 *Years Later*, https://arxiv.org/abs/2106.10522, 2021.
[29] Y. Saad, *On the rates of convergence of the Lanczos and the block-Lanczos methods*, SIAM J. Numer. Anal., 17 (1980), pp. 687–706, https://doi.org/10.1137/0717059.
[30] I. Sabzevari, A. Mahajan, and S. Sharma, *An accelerated linear method for optimizing non-linear wavefunctions in variational Monte Carlo*, J. Chem. Phys., 152 (2020), 024111, https://doi.org/10.1063/1.5125803.
[31] K. Seki and S. Yunoki, *Quantum power method by a superposition of time-evolved states*, Phys. Rev. X Quantum, 2 (2021), 010333.
[32] N. H. Stair, R. Huang, and F. A. Evangelista, *A multireference quantum Krylov algorithm for strongly correlated electrons*, J. Chem. Theory Comput., 16 (2020), pp. 2236–2245, https://doi.org/10.1021/acs.jctc.9b01125.
[33] G. W. Stewart, *Perturbation bounds for the definite generalized eigenvalue problem*, Linear Algebra Appl., 23 (1979), pp. 69–85, https://doi.org/10.1016/0024-3795(79)90094-6.
[34] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Computer Science and Scientific Computing, Academic Press, New York, 1990.
[35] J. A. Tropp, *An Introduction to Matrix Concentration Inequalities*, Found. Trends Machine Learning 8, Now Publishers, Hanover, 2015, pp. 1–230.
[36] M. R. Wall and D. Neuhauser, *Extraction, through filter-diagonalization, of general quantum eigenvalues or classical normal mode frequencies from a small number of residues or a short-time segment of a signal.* I. *Theory and application to a quantum-dynamics model*, J. Chem. Phys., 102 (1995), pp. 8011–8022, https://doi.org/10.1063/1.468999.
[37] P. Weinberg and M. Bukov, *QuSpin: A Python package for dynamics and exact diagonalisation of quantum many body systems Part* I*: Spin chains*, SciPost Phys., 2 (2017).
[38] P. Weinberg and M. Bukov, *QuSpin: A Python package for dynamics and exact diagonalisation of quantum many body systems. Part* II*: Bosons, fermions and higher spins*, SciPost Phys., 7 (2019).
[39] J. H. Wilkinson, *Kronecker's canonical form and the QZ algorithm*, Linear Algebra Appl., 28 (1979), pp. 285–303, https://doi.org/10.1016/0024-3795(79)90140-X.

Downloaded 08/05/22 to 176.252.113.237 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.