# CONVERGENCE OF ADAPTIVE COMPRESSION METHODS FOR HARTREE-FOCK-LIKE EQUATIONS

LIN LIN* AND MICHAEL LINDSEY†

**Abstract.** The adaptively compressed exchange (ACE) method provides an efficient way for solving Hartree-Fock-like equations in quantum physics, chemistry, and materials science. The key step of the ACE method is to adaptively compress an operator that is possibly dense and full-rank. In this paper, we present a detailed study of the adaptive compression operation, and establish rigorous convergence properties of the adaptive compression method in the context of solving linear eigenvalue problems. Our analysis also elucidates the potential use of the adaptive compression method in a wide range of problems.

**Key words.** Adaptive compression; Global convergence; Eigenvalue problem; Orthogonal projector; Hartree-Fock; Quantum chemistry

**AMS subject classifications.** 65F15, 15A18, 47H10, 58C30, 81V55

## 1. Introduction.

The Fock exchange operator plays a fundamental role in many-body quantum physics. The Hartree-Fock equation (HF) [34] is the starting point of nearly all wavefunction based correlation methods in quantum chemistry. Hartree-Fock-like equations also appear in the widely used Kohn-Sham density functional theory (KSDFT) [17, 20] with hybrid exchange-correlation functionals [2, 16, 30] in quantum chemistry and materials science. As an example, the B3LYP functional [2], which is only one specific functional used by KSDFT, has generated more than $60,000$ citations.[1]

Hartree-Fock-like equations require the solution of a large number of eigenpairs of a nonlinear integro-differential operator. From a computational perspective, after linearization and a certain numerical discretization to be detailed later, we solve the following linear eigenvalue problem

$$(A + B)v_i = \lambda_i v_i, \quad i = 1, \ldots, n. \tag{1.1}$$

Here $A, B \in \mathbb{C}^{N\times N}$ are Hermitian matrices. The eigenvalues $\{\lambda_i\}$ are real and ordered non-decreasingly. Due to the Pauli exclusion principle we need to compute the eigenpairs $(\lambda_i, v_i)$ corresponding to the lowest $n$ eigenvalues, which are separated from the rest of the eigenvalues by a positive spectral gap $\lambda_g := \lambda_{n+1} - \lambda_n$. Here $n$ encodes the number of electrons in the system, and can range from tens to tens of thousands. This means that a potentially large number of eigenpairs need to be computed. We consider the case that $N$ is large enough so that it is only viable to use an iterative method to solve (1.1).

In Hartree-Fock-like equations, $A$ in (1.1) is obtained by discretizing a differential operator involving the Laplace operator. $B$ is obtained by discretizing the Fock exchange operator, which is an integral operator, and $B$ is negative definite. The discretized Fock exchange operator $B$ is in general a dense full-rank matrix, and it is

*Department of Mathematics, University of California, Berkeley, Berkeley, CA 94720 and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. Email: linlin@math.berkeley.edu

†Department of Mathematics, University of California, Berkeley, Berkeley, CA 94720. Email: lindsey@math.berkeley.edu

[1] Data from ISI Web of Science, February, 2017.

prohibitively expensive to compute or even to store $B$. As one refines the discretization, the spectral radius of $A$, denoted by $\|A\|_2$, can become unbounded, while $\|B\|_2$ remains bounded. In an iterative method, one needs to repeatedly apply $(A + B)$ to some vector $v$, i.e. the matrix-vector multiplication operations $Av$ and $Bv$ are coupled together. Due to the large spectral radius $\|A\|_2$, many matrix-vector multiplications may be needed to reach convergence. In practice, each matrix vector multiplication $Bv$ requires the solution of $n$ Poisson type equations [15], which is far more expensive than computing $Av$. Therefore the computational cost of most iterative solvers will be dominated by the number of matrix-vector multiplication operations involving $B$. It is common that the evaluation of $Bv$ alone takes 95% or more of the overall computational time, which severely limits the capability of solving Hartree-Fock-like equations for studying quantum systems of large sizes. In the past decades, there has been a large amount of work dedicated to reducing the cost of performing *each* matrix-vector multiplication $Bv$. This is often done by approximating the dense matrix $B$ by a sparse matrix, which is valid when the spectral gap $\lambda_g$ is large [19, 25, 14, 10, 27, 4, 37, 9, 3, 8, 7].

Recently we have developed an *adaptively compressed exchange operator* formulation (ACE) [22], which reduces the cost for solving Hartree-Fock-like equations from a different and yet more general perspective. The key observation is that we only need to find an effective operator $\underline{B}$ so that $Bv = \underline{B}v$ is satisfied for $v \in \text{span } V$, where $V = [v_1, \ldots, v_n]$. $\underline{B}$ can be constructed to be of strict rank $n$, and hence the computational cost of $\underline{B}v$ is much smaller than that of $Bv$. Note that the subspace span $V$ is precisely the solution for (1.1) and is not known *a priori*. Therefore $\underline{B}$ needs to be constructed in an adaptive manner. Starting from some initial guess $V^{(0)}$, we will obtain a sequence $V^{(k)}$ and corresponding compressed operators $\underline{B}[V^{(k)}]$. More specifically, our approach is a fixed-point iteration given by

$$(A + \underline{B}[V^{(k)}])v_i^{(k+1)} = \lambda_i^{(k+1)} v_i^{(k+1)}, \quad i = 1, \ldots, n. \tag{1.2}$$

Here the operator $\underline{B}$ depends nonlinearly on $V$. If the sequence of subspaces span $V^{(k)}$ converges to span $V$, then in the limit the compressed operator $\underline{B}[V^{(k)}]$ will agree with $B$ on span $V$, and the eigenvalue problem (1.1) is solved *without loss of accuracy*.

This paper aims to prove the convergence properties of this adaptive compression method. At first glance, the advantage of converting a linear eigenvalue problem (1.1) to a nonlinear eigenvalue problem (1.2) is not clear. We will see that the adaptive compression method decouples the matrix-vector multiplication operations $Av$ and $Bv$, and asymptotically the number of $Bv$ operations is independent of the spectral radius $\|A\|_2$.

We will demonstrate that $\underline{B}$ depends only on span $V$, so we can consider the fixed point iteration (1.2) to be a map $P^{(k)} \mapsto P^{(k+1)}$, where $P^{(k)}$ is the orthogonal projector $P^{(k)} = V^{(k)}(V^{(k)})^*$. Let $\mathbf{H}_N$ denote the set of Hermitian $N \times N$ matrices, and $\mathcal{D} \subset \mathbb{C}^{N\times N}$ denote the set of rank-$n$ orthogonal projectors on $\mathbb{C}^N$. The main results of the paper are as follows.

THEOREM 1.1 (Optimality). *For $B \prec 0$, the adaptive compression $\underline{B}[V]$ is the unique rank-n Hermitian matrix that agrees with $B$ on span $V$. Furthermore, $B \preceq \underline{B}[V] \preceq 0$.*

THEOREM 1.2 (Local convergence). *For every pair $(A, B) \in \mathbf{H}_N \times \mathbf{H}_N$ with*

*$B \prec 0$, the fixed point iteration* (1.2) *converges locally to $P = VV^*$. The number of matrix-vector multiplications $Bv$ needed for $k$ steps of fixed point iteration is $nk$. Starting from $P^{(0)} \in \mathcal{D}$, the asymptotic convergence rate is*

$$\|P - P^{(k)}\|_2 \lesssim \gamma^k \|P - P^{(0)}\|_2, \quad \text{where} \quad \gamma \leq \frac{\|B\|_2}{\|B\|_2 + \lambda_g}.$$

THEOREM 1.3 (Global convergence). *For almost every pair $(A, B) \in \mathbf{H}_N \times \mathbf{H}_N$ (with respect to the Lebesgue measure on $\mathbf{H}_N \times \mathbf{H}_N$) with $B \prec 0$, the fixed point iteration* (1.2) *converges globally to $P = VV^*$ for almost every initial guess $P^{(0)} \in \mathcal{D}$ (with respect to a natural measure on $\mathcal{D}$).*

REMARK 1.4. *With minor modification, the condition $B \prec 0$ can be relaxed, so that the adaptive compression method is applicable to all $B \in \mathbf{H}_N$. See Section 3.2.*

REMARK 1.5. *Let $\mathbf{S}_N$ denote the set of real-symmetric $N \times N$ matrices, and $\mathcal{D}_{\mathbb{R}} \subset \mathbb{R}^{N \times N}$ denote the set of rank-$n$ orthogonal projectors on $\mathbb{R}^N$. Then Theorems 1.2 and 1.3 hold if we replace $\mathbf{H}_N$ with $\mathbf{S}_N$ and $\mathcal{D}$ with $\mathcal{D}_{\mathbb{R}}$.*

In practice, Eq. (1.1) is only the linearized Hartree-Fock-like equation, and it is possible to employ the flexibility in the adaptive compression formulation by delaying the update of the compressed operator $\underline{B}$ to further reduce the number of $Bv$ operations. This strategy is undertaken in [22]. Numerical observation indicates that the ACE formulation can significantly reduce the number of iterations to solve Hartree-Fock-like equations, and may reduce the computational time by an order of magnitude [22]. The adaptive compression formulation has already been adopted by community software packages for electronic structure calculations such as Quantum ESPRESSO [11] for solving Hartree-Fock-like equations for real materials.

**1.1. Applicability to nearly degenerate eigenvalue problems.** Theorem 1.2 suggests that the adaptive compression method converges fast when the spectral gap $\lambda_g$ is large, which is the case for insulating systems in quantum physics. However, $\lambda_g$ is small for semiconducting systems, and can be virtually zero for metallic systems. In this case, one can compute $n$ eigenvectors, where $n$ is set to be larger than $m$, the number of eigenvectors needed in solving Hartree-Fock-like equations. Although the convergence of the rank-$n$ projector $P^{(k)}$ is expected to be slow, one is actually only interested in the convergence of the rank-$m$ "sub-projector" $P_m^{(k)}$ onto the span of the lowest $m$ eigenvectors. This procedure is rigorously justified in Theorem 1.6. We find that the asymptotic convergence rate of the sub-projector is governed by the gap $\lambda_{n+1} - \lambda_m$, rather than the gap $\lambda_{m+1} - \lambda_m > 0$, which is assumed to be positive only to ensure that the rank-$m$ orthogonal projector $P_m$ is unambiguously defined.

THEOREM 1.6 (Convergence of sub-projectors). *Let $P^{(k)}$ converge to $P$ (as broadly guaranteed by Theorem 1.3). Then $P_m^{(k)}$ converges to $P_m$ with asymptotic convergence rate given by*

$$\|P_m - P_m^{(k)}\|_2 \lesssim \gamma_m^k \|P - P^{(0)}\|_2, \quad \text{where} \quad \gamma_m \leq \frac{\|B\|_2}{\|B\|_2 + \Delta_m}.$$

*Here* $\Delta_m = \lambda_{n+1} - \lambda_m$.

**1.2. Applicability to more general problems.** Although we have Hartree-Fock-like equations in mind throughout the paper, it is easy to see that the adaptive compression method can be applied to a wider variety of problems. The case that $B$ is "small" and "costly" can occur when $B$ comes from the discretization of an integral operator or a more general nonlocal operator. For example, in linear response theories such as time-dependent density functional theory and Bethe-Salpeter equations [31, 26], generalized eigenvalue problems arise involving matrices of the form $A + B$, where $A$ is a diagonal matrix and $B$ is a discretized nonlocal operator with additional structure. Adaptive compression methods with structure-preserving properties could be applicable to these problems. The concept of adaptive compression can also be useful in solving linear equations, as recently demonstrated in the adaptively compressed polarizability operator formulation for first principle phonon spectrum calculations [23]. We are currently exploring these directions.

**1.3. Related work.** A Hartree-Fock-like equation, considered as in (1.1) after linearization and discretization, constitutes a standard linear eigenvalue problem, and the present work should be directly compared with existing iterative eigensolvers, such as the subspace iteration method [28], the shift-invert Lanczos method [29], the preconditioned steepest descent method [5], the preconditioned conjugate gradient method [18], the Jacobi-Davidson method [33], etc. In these approaches, the matrix-vector multiplication always takes the form $(A+B)v$, and the number of $Bv$ operations is $n$ times the number of iterations. In the absence of a good preconditioner, the number of iterations in these solvers typically depends on $\|A + B\|_2$, which is undesirable. Even when a good preconditioner is available, we still find that the adaptive compression method can be advantageous, thanks to the flexibility introduced by decoupling $Av$ and $Bv$ operations. Note that Eq. (1.2) is only a fixed point iteration, and the convergence rate of the adaptive compression method can be further enhanced by combining with existing acceleration techniques such as the usage of conjugate directions [18] and Broyden type methods [1]. We will report detailed numerical study of the adaptive compression methods in a forthcoming publication. We also note that the adaptive compression method is very simple to implement and only requires a "black-box" subroutine for the computation of $Bv$. Hence in the context of solving Hartree-Fock-like equations, it is compatible with any existing method that reduces the cost of the matrix-vector multiplication, such as those using linear scaling techniques and using fast solvers for elliptic equations.

**1.4. Outline of the paper.** The rest of the paper is organized as follows. After presenting a brief introduction to Hartree-Fock-like equations in Section 2, we introduce the adaptive compression method in Section 3. Section 4 discusses the properties and optimality of the compression map $V \mapsto \underline{B}[V]$. In Section 5 we establish the local convergence with an asymptotic rate, followed by the global convergence in Section 6. Finally, some technical calculations and proofs omitted in the main text are presented in the appendices.

**2. Hartree-Fock-like equations.** The Hartree-Fock-like equations are a set of nonlinear equations as follows [24]

$$\begin{aligned} H[P]\psi_i &= \left(-\frac{1}{2}\Delta + V_{\text{ion}} + V_{\text{Hxc}}[P] + V_X[P]\right)\psi_i = \varepsilon_i\psi_i, \\ \int \psi_i^*(\mathbf{r})\psi_j(\mathbf{r})\,\mathrm{d}\mathbf{r} &= \delta_{ij}, \quad P(\mathbf{r},\mathbf{r}') = \sum_{i=1}^{N_e}\psi_i(\mathbf{r})\psi_i^*(\mathbf{r}'). \end{aligned} \tag{2.1}$$

Here the eigenvalues $\{\varepsilon_i\}$ are ordered non-decreasingly, and $N_e$ is the number of electrons (spin degeneracy omitted). $P$ is the density matrix, which is an orthogonal projector with an exact rank $N_e$. The diagonal entries of the kernel of $P$ gives the electron density $\rho(\mathbf{r}) = P(\mathbf{r},\mathbf{r})$. $V_{\text{ion}}$ characterizes the electron-ion interaction in all-electron calculations. $V_{\text{Hxc}}$ is a local operator, and characterizes the Hartree contribution and the exchange-correlation contribution modeled at a local or semi-local level. It typically depends only on the electron density. The exchange operator $V_X$ is an integral operator with kernel

$$V_X[P](\mathbf{r},\mathbf{r}') = -P(\mathbf{r},\mathbf{r}')K(\mathbf{r},\mathbf{r}'). \tag{2.2}$$

Here $K(\mathbf{r},\mathbf{r}')$ is the kernel for the electron-electron interaction. For example, in the Hartree-Fock theory, $K(\mathbf{r},\mathbf{r}') = 1/|\mathbf{r}-\mathbf{r}'|$ is the Coulomb operator. In screened exchange theories [16], $K$ can be a screened Coulomb operator with kernel $K(\mathbf{r},\mathbf{r}') = \operatorname{erfc}(\mu|\mathbf{r}-\mathbf{r}'|)/|\mathbf{r}-\mathbf{r}'|$. $V_X$ is a negative semidefinite operator. The kernel of $V_X$ is not low rank due to the Hadamard product (i.e. element-wise product) between the kernels of $P$ and $K$. From a computational perspective, it is prohibitively expensive to explicitly construct $V_X[P]$, and it is only viable to apply it to a vector $v(\mathbf{r})$ as

$$\left(V_X[P]v\right)(\mathbf{r}) = -\sum_{i=1}^{N_e}\psi_i(\mathbf{r})\int K(\mathbf{r},\mathbf{r}')\psi_i^*(\mathbf{r}')v(\mathbf{r}')\,\mathrm{d}\mathbf{r}'. \tag{2.3}$$

This operation is much more expensive than computing $(H[P]-V_X[P])v$. In practical Hartree-Fock calculations, the application of $V_X[P]$ to vectors can often take more than 95% of the overall computational time.

The Hartree-Fock-like equations require the density matrix $P$ to be computed self-consistently. A common strategy is to solve the linearized Hartree-Fock equation by fixing the density matrix $P$ so that $H[P]$ becomes a fixed operator. Then ones solves a nonlinear fixed point problem to obtain the self-consistent $P$. The most time consuming step is to solve the linearized Hartree-Fock equation. After numerical discretization, this gives rise to the linear eigenvalue problem (1.1), where $B$ corresponds to the discretized Fock operator $V_X[P]$, and $A$ corresponds to the remaining part $H[P]-V_X[P]$. We also remark that after numerical discretization, $B$ is a negative definite matrix.

**3. Adaptive compression method.**

**3.1. Method description.** In order to reduce the number of matrix-vector multiplication operations $Bv$, the simplest idea is to fix $w_i := Bv_i$ at some stage, and to replace $Bv_i$ by $w_i$ for a number of iterations. This leads to the following sub-problem

$$Av_i + w_i = \lambda_i v_i, \quad i = 1,\ldots,n. \tag{3.1}$$

Note that Eq. (3.1) is not an eigenvalue problem: if $v_i$ is a solution to (3.1), then $v_i$ multiplied by a constant $c$ is typically not a solution. Eq. (3.1) could be solved using optimization based methods, but such a problem is typically more difficult than an Hermitian eigenvalue problem. In practice, software packages for solving Hartree-Fock-like equations are typically built around eigensolvers, which is another important factor that makes the sub-problem (3.1) undesirable.

The adaptive compression method reuses the information in $\{w_i = Bv_i\}$ in a different way, which retains the structure of the eigenvalue problem (1.1). Define $V = [v_1, \ldots, v_n]$, $W = [w_1, \ldots, w_n]$, so $V, W \in \mathbb{C}^{N\times n}$, and construct

$$\underline{B}[V] = W(W^*V)^{-1}W^*. \tag{3.2}$$

Since $B \prec 0$, $W^*V \equiv V^*BV$ has only negative eigenvalues and is invertible. $\underline{B}[V]$ is Hermitian of rank $n$, and agrees with $B$ when applied to $V$ as

$$\underline{B}[V]V = W(W^*V)^{-1}W^*V = W = BV. \tag{3.3}$$

We shall refer to the operation from $B$ to $\underline{B}[V]$ as an *adaptive compression*.

In an iterative scheme, denote by $V^{(k)} = [v_1^{(k)}, \ldots, v_n^{(k)}]$ the approximate eigenvectors at the $k$-th iteration of (1.2). Then the adaptive compression method proceeds as follows. After $\underline{B}[V^{(k)}]$ is constructed, (1.2) can be solved via *any* iterative eigensolver to obtain $V^{(k+1)}$. The iterative eigensolver only requires the application of $A$ and the low rank matrix $\underline{B}$ to vectors, and does not require any additional application of $B$ until $V^{(k+1)}$ is obtained. If span $V^{(k)}$ converges to span $V$, then the consistency condition $\underline{B}[V]V = BV$ is satisfied, and the adaptive compression method is numerically exact. The adaptive compression method for solving the linear eigenvalue problem (1.1) is given in Algorithm 1, where we initialize $V^{(0)}$ by solving the eigenvalue problem in the absence of $B$.

**Algorithm 1** The adaptive compression method for solving Eq. (1.1)

1: Initialize $V^{(0)}$ by solving $Av_i^{(0)} = \lambda_i^{(0)} v_i^{(0)}, \quad i = 1, \ldots, n.$
2: **while** convergence not reached **do**
3: Compute $W^{(k)} = BV^{(k)}$.
4: Evaluate $\left[(W^{(k)})^* V^{(k)}\right]^{-1}$ to construct $\underline{B}[V^{(k)}]$ implicitly.
5: Solve (1.2) to obtain $V^{(k+1)}$.
6: Set $k \leftarrow k + 1$.
7: **end while**

**3.2. Relaxing the definitiveness condition for $B$.** As will be seen later, the condition that $B \prec 0$ is important for the consistency of the adaptive compression method, but this constraint can be easily relaxed as follows for more general $B$. Note that replacing $B$ with $B_t := B - t$ (here $t$ as a matrix means the identity matrix scaled by a real number $t$) in the eigenvalue problem (1.1) yields an equivalent eigenvalue problem, where all eigenvalues are shifted down by $t$ and the corresponding eigenspaces are unchanged. Thus taking $t > \lambda_{\max}(B)$ ensures that $B_t$ is negative definite. We call this procedure a *$t$-shifted adaptive compression*. The spirit of this construction is related to the "level-shifting" method used in quantum chemistry [32]. Theorem 1.2 suggests that the convergence rate of Algorithm 1 can be optimized by minimizing $\|B_t\|_2$. This also opens up the interesting possibility of accelerating the convergence

of the adaptive compression method by taking $t$ to be negative when $B$ is already negative definite. In the discussion below, we will assume that $B$ is negative definite unless otherwise specified.

**4. Optimality of the adaptive compression.** In (3.2), we have specified how to compress an Hermitian negative definite matrix $B$ into a rank $n$ matrix with the same behavior on span $V$. Since $V$ has orthonormal columns, the orthogonal projector onto span $V$ is

$$P = VV^*. \tag{4.1}$$

In the context of Hartree-Fock-like equations, $P$ in Eq. (4.1) is the discretized density matrix. In the discussion below, we use the terminology *density matrix* in a slightly more general sense:

DEFINITION 4.1. *For $H \in \mathbf{H}_N$ with eigenvalues $\{\lambda_i\}_{i=1}^N$ ordered non-decreasingly and a given number $1 \le n \le N$, if the spectral gap $\lambda_g := \lambda_{n+1} - \lambda_n$ is positive, the density matrix associated with $H$ and $n$ is defined to be the orthogonal projector onto the span of the first $n$ eigenvectors of $H$.*

REMARK 4.2. *In this paper, all density matrices are idempotent, i.e. $P^2 = P$. When the context is clear, we may drop the dependence on $H$ and $n$ and simply refer to an orthogonal projector $P$ as a density matrix. We also let $\mathcal{D} = \mathcal{D}_{\mathbb{C}} \subset \mathbb{C}^{N\times N}$ denote the set of rank-$n$ density matrices.*

Using the density matrix, the compressed matrix $\underline{B}[V]$ can be expressed as

$$\underline{B}[V] = BV(V^*BV)^{-1}V^*B = B(PBP)^\dagger B, \tag{4.2}$$

where $(PBP)^\dagger$ is the Moore-Penrose pseudoinverse [12] of the rank-$n$ matrix $PBP$.

In order to prove the second equality of (4.2), we denote by $\{v_i\}_{i=1}^N$ a completion of $\{v_i\}_{i=1}^n$ to an orthonormal basis of $\mathbb{C}^N$. For any $1 \le m \le N$, define

$$V_m = [v_1, \ldots, v_m]. \tag{4.3}$$

In particular, $V_N = [v_1, \ldots, v_N]$ consists of all eigenvectors, and $V \equiv V_n = [v_1, \ldots, v_n]$ consists of the eigenvectors to be computed. The matrix representation of $PBP$ with respect to the basis $V_N$ is given in the block form by

$$[PBP]_{V_N} = \begin{pmatrix} V^*BV & 0 \\ 0 & 0 \end{pmatrix}, \tag{4.4}$$

where the size of the upper-left block is $n \times n$. Thus the matrix representation of the pseudoinverse $(PBP)^\dagger$ is

$$[(PBP)^\dagger]_{V_N} = \begin{pmatrix} (V^*BV)^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{4.5}$$

Hence $(PBP)^\dagger = V(V^*BV)^{-1}V^*$, which implies the second equality of (4.2). Eq. (4.2) suggests that $\underline{B}$ is a matrix function of the density matrix $P$, or equivalently, a function of the subspace $\mathrm{Im}(P) = \mathrm{span}\ V$. With some abuse of notation, we will not distinguish between $\underline{B}[V]$ and $\underline{B}[P]$, and we will mostly use the projector formulation $\underline{B}[P]$ in the discussion below.

Denote by

$$[B]_{V_N} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^* & B_{22} \end{pmatrix} \tag{4.6}$$

the matrix representation of $B$, and $B_{11} = V^*BV$. Then (4.2) and (4.5) give the matrix representation of $\underline{B}[P]$ as

$$[\underline{B}[P]]_{V_N} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^* & B_{12}^* B_{11}^{-1} B_{12} \end{pmatrix}. \tag{4.7}$$

Note that only the lower-right matrix block is changed in the adaptive compression.

REMARK 4.3 (Smoothness of adaptive compression). *We can rewrite* (4.2) *as*

$$\underline{B}[P] = B\left[PBP + (I-P)\right]^{-1} B - B(I-P)B,$$

*from which it is clear that* $\underline{B}$ *is smooth (in particular, continuous) as a function* $P \mapsto \underline{B}[P]$ *on the set of density matrices.*

Our consideration of adaptive compression is motivated by the following fact:

PROPOSITION 4.4 (Axiomatic characterization of adaptive compression, I). *Let* $B \in \mathbf{H}_N$ *be negative definite, and let* $P$ *be a rank-*$n$ *orthogonal projector. Then* $\underline{B}[P]$ *is the unique Hermitian matrix* $B'$ *satisfying* $B'|_{\mathrm{Im}(P)} \equiv B|_{\mathrm{Im}(P)}$ *and* $\operatorname{rank} B' \leq n$. *(In fact,* $\operatorname{rank}(\underline{B}[P]) = n$.*)*

*Proof.* We have already established that $\underline{B}[P]$ satisfies the stated properties, so we need only prove uniqueness.

To this end, suppose that $B'$ is a matrix satisfying the stated properties, so $B'$ is Hermitian, has rank at most $n$, and agrees with $B$ on $\mathrm{Im}(P)$. As in the preceding discussion, let $v_1, \ldots, v_N$ be an orthonormal basis for $\mathbb{C}^N$, with $v_1, \ldots, v_n$ forming an orthonormal basis for $\mathrm{Im}(P)$. With $V_N$ as in (4.3), write the matrix of $B'$ in this basis:

$$[B']_{V_N} = \begin{pmatrix} B'_{11} & B'_{12} \\ B'^*_{12} & B'_{22} \end{pmatrix}.$$

where the upper-left block is $n \times n$. Since $B'$ must agree with $B$ on $v_1, \ldots, v_n$, we must have $B'_{11} = B_{11}$ and $B'_{12} = B_{12}$, where the $B_{ij}$ are as in (4.6). In summary,

$$[B']_{V_N} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^* & * \end{pmatrix}. \tag{4.8}$$

Since $B_{11} = V^*BV$ is invertible (where $V = [v_1, \ldots, v_n]$), the first $n$ columns of $[B']_{V_N}$ must be linearly independent. This means that the rank of $B'$ is at least $n$, hence equal to $n$. Then for any $j = 1, \ldots, N-n$, the $(n+j)$-th column of $[B']_{V_N}$ must be a linear combination of the first $n$ columns. However, the coefficients of this linear combination are completely determined by $B$, since by (4.8) the $j$-th column of $B_{12}$ is a linear combination of the columns of $B_{11}$ with *these same coefficients*. By the linear independence of the columns of $B_{11}$, there is exactly one way to write each column of $B_{12}$ as a linear combination of columns of $B_{11}$, i.e. Eq. (4.7). ∎

REMARK 4.5. *For Proposition 4.4 (and indeed for the entire discussion of Section 4 thus far), it is not necessary to assume that* $B \prec 0$. *In fact, it is sufficient to assume that* $B$ *is Hermitian and* $V^*BV$ *is invertible. (Note that there exist invertible Hermitian matrices such that* $V^*BV$ *is not invertible, though this cannot happen if* $B$ *is definite.) However, the case of definite* $B$ *affords adaptive compression with additional properties (see Lemma 4.8) that are crucial for the utility of adaptive compression in solving eigenvalue problems. As discussed in Section 3.2, when* $B$ *is indefinite, the appropriate generalization of adaptive compression for the purpose of*

*solving the eigenvalue problem* (1.1) *does not involve performing adaptive compression on $B$ directly, but rather reduces to the case of definite $B$ by subtracting a multiple of the identity.*

Before proceeding, we state a linear-algebraic result on Schur complements that will be useful for understanding the adaptive compression.

LEMMA 4.6. *The positive semidefiniteness* [*resp., definiteness*] *of a Hermitian matrix $M := \begin{pmatrix} X & Y \\ Y^* & Z \end{pmatrix}$ (where $X$ is invertible) is equivalent to having both $X \succeq 0$ and $S := Z - Y^*X^{-1}Y \succeq 0$ [resp., $\succ 0$]. In fact, if $M \succeq t \geq 0$, then $S \succeq t$ as well.*

REMARK 4.7. *Note that $S$ is a Schur complement. Excluding the last sentence, Lemma 4.6 is a standard result in linear algebra (see, e.g., Theorem 1.12 of [38]).*

*Proof.* We only prove the last statement. Assume that $M \succeq t$. Define

$$F(u,v) := \begin{pmatrix} u \\ v \end{pmatrix}^* M \begin{pmatrix} u \\ v \end{pmatrix} = u^*Xu + u^*Yv + v^*Y^*u + v^*Zv.$$

Observe that

$$F(-X^{-1}Yv, v) = v^*Y^*X^{-1}Yv - v^*Y^*X^{-1}Yv - v^*Y^*X^{-1}Yv + v^*Zv = v^*Sv.$$

Using the previous two equalities and the fact that $M \succeq t \geq 0$, observe that for any $v$,

$$v^*Sv = \begin{pmatrix} -X^{-1}Yv \\ v \end{pmatrix}^* M \begin{pmatrix} -X^{-1}Yv \\ v \end{pmatrix} \geq t \left\| \begin{pmatrix} -X^{-1}Yv \\ v \end{pmatrix} \right\|_2^2 \geq t\|v\|_2^2.$$

This completes the proof via the Courant-Fischer minimax theorem [12]. ∎

Taking $B$ to be negative definite, it follows from (4.2) that $\underline{B}[P]$ is negative semidefinite, i.e., $\underline{B}[P] \preceq 0$. Since $\underline{B}[P]$ is a low-rank substitute for the negative definite matrix $B$, one might additionally hope that the compression does not make $B$ "more negative" in any direction, i.e. $\underline{B}[V] \succeq B$. Lemma 4.8 shows that this is indeed the case.

LEMMA 4.8. *Let $B \in \mathbb{C}^{N\times N}$ be a negative definite matrix. For any rank-$n$ projector $P$, the matrix $\underline{B}[P] - B$ is positive semidefinite. Therefore $B \preceq \underline{B}[P] \preceq 0$.*

*Proof.* Note from (4.6) and (4.7) that

$$[\underline{B}[V] - B]_{V_N} = \begin{pmatrix} 0 & 0 \\ 0 & B_{12}^*B_{11}^{-1}B_{12} - B_{22} \end{pmatrix}, \tag{4.9}$$

so Lemma 4.8 is equivalent to the statement that $B_{12}^*B_{11}^{-1}B_{12} - B_{22} \succeq 0$, i.e., that the Schur complement $B_{22} - B_{12}^*B_{11}^{-1}B_{12}$ is negative semidefinite. But this follows from Lemma 4.6, together with the fact that $B$ is negative definite. ∎

REMARK 4.9. *Observe that Theorem 1.1 follows directly from Proposition 4.4 and 4.8. Lemma 4.8 will also be key for proving the convergence of Algorithm 1.*

The Schur complement perspective on adaptive compression yields further insights. Note that the stipulation that $\underline{B}[V]$ agrees with $B$ on span $V$ determines the upper-left and upper-right blocks of $\underline{B}[V]$ as in (4.7), and the stipulation that $\underline{B}[V]$ is Hermitian then fixes the lower-left block. The only thing that then remains to be specified is the lower-right block, which is identified as in $\underline{B}[V]$. This suggests the following characterization of adaptive compression:

PROPOSITION 4.10 (Axiomatic characterization of adaptive compression, II). *Let $B \in \mathbb{C}^{N\times N}$ be a negative definite matrix, and let $E$ be an $n$-dimensional subspace of*

$\mathbb{C}^N$. *Then* $\underline{B}[E]$ *is the maximal Hermitian negative semidefinite matrix* $B'$ *satisfying* $B'|_E \equiv B|_E$, *in the sense that for any other such* $B'$, *we have* $B' \preceq \underline{B}[E]$.

*Proof.* Suppose that $B' \preceq 0$ with $B' \preceq \underline{B}[E]$. Let $V_N = [v_1, \ldots, v_N]$ be an orthonormal basis for $\mathbb{C}^N$, with $v_1, \ldots, v_n$ forming an orthonormal basis for $E$. As in the proof of Proposition 4.4, the matrix of $B'$ in this basis is of the form

$$[B']_{V_N} = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^* & Z \end{pmatrix},$$

where the $B_{ij}$ are as in (4.6). Since $B'$ is negative semidefinite, by Lemma 4.6 the Schur complement $Z - B_{12}^* B_{11}^{-1} B_{12}$ is also negative semidefinite, i.e., $Z \preceq B_{12}^* B_{11}^{-1} B_{12}$. But by (4.7) this implies that $B' \preceq \underline{B}[E]$. ∎

Similar results hold for the $t$-shifted adaptive compression. For $B$ Hermitian, $t > \lambda_{\max}(B)$, and a rank-$n$ projector $P$, define $\underline{B}[P,t] = \underline{B_t}[P] + t$, where $B_t = B - t$. Then we have

1. $B$ and $\underline{B}[P,t]$ agree on the image of $P$ denoted by $\mathrm{Im}(P)$.
2. $B \preceq \underline{B}[P,t] \preceq t$.
3. $\underline{B}[P,t]$ is not of rank $n$, but $\underline{B}[P,t]$ is the sum of a rank-$n$ matrix and a multiple of the identity, and hence is computationally efficient to apply.

**5. Local convergence analysis.** Since each step of (1.2) is an Hermitian eigenvalue problem, we can require $V^{(k)}$ to be columns of a unitary matrix and let $P^{(k)} = V^{(k)}(V^{(k)})^*$. Then let $P^{(k+1)}$ be the density matrix associated with $A + \underline{B}[P^{(k)}]$. The convergence of the adaptive compression method for the linear problem (1.1) can be stated in terms of the convergence of the density matrix $P^{(k)} \to P$. For any $H \in \mathbf{H}_N$, let $\lambda_i\{H\}$ denote its $i$-th smallest eigenvalue (counting multiplicity). In this notation, the true eigenvalues of $(A+B)$ are $\lambda_i \equiv \lambda_i\{A+B\}$.

We now formally define the *fixed point iteration map* $\mathfrak{F}(\cdot)$ as follows.

Definition 5.1. *For a density matrix* $Q$, *let* $\mathfrak{F}(Q)$ *be the orthogonal projector* $\sum_{i=1}^n u_i u_i^*$, *where the* $u_i$ *are orthonormal eigenfunctions of*

$$\big(A + \underline{B}[Q]\big)u_i = \lambda_i\big\{A + \underline{B}[Q]\big\}u_i.$$

$\mathfrak{F}(Q)$ *is canonically defined if* $A + \underline{B}[Q]$ *has a positive spectral gap, such a projector is unique, and it is the density matrix associated with* $A + \underline{B}[Q]$. *Otherwise, we make an arbitrary choice in the eigenspace associated with* $\lambda_n\big\{A + \underline{B}[Q]\big\}$ *so that* $\mathfrak{F}(Q)$ *is of rank* $n$.

Using the fixed point iteration map, we can rephrase Algorithm 1 compactly as

$$P^{(k+1)} := \mathfrak{F}(P^{(k)}). \tag{5.1}$$

We will see below that for $Q$ sufficiently close to $P$ (the true density matrix of $A+B$), $A+\underline{B}[Q]$ has a positive spectral gap, and hence its density matrix is indeed canonically defined. Thus, for all $Q$ in a neighborhood of $P$, $\mathfrak{F}(Q)$ is the density matrix associated with $A+\underline{B}[Q]$. The local convergence of Algorithm 1 can be studied via the properties of the map $\mathfrak{F}$ near the true density matrix $P$. A necessary requirement for Algorithm 1 to converge is the consistency condition

$$\mathfrak{F}(P) = P. \tag{5.2}$$

In order to guarantee local linear convergence, the spectral radius for the Jacobian of $\mathfrak{F}$ must also be bounded by unity, so that the fixed point $P$ is attractive with respect to the iteration (5.1). This leads to a sharp estimate of the local convergence rate, which is upper-bounded by the rate provided in Theorem 1.2.

**5.1. Consistency.** For a general Hermitian $B \in \mathbf{H}_N$, even if $\underline{B}$ is constructed from the true density matrix $P$, the true eigenvectors $\{v_1, \ldots, v_n\}$ of $A+B$ may not correspond to the *lowest* $n$ eigenvalues of the modified operator $A + \underline{B}[P]$, despite the fact that they are still eigenvectors of the modified operator. In such case, the consistency requirement (5.2) is violated.

However, the consistency condition of the fixed point iteration will be satisfied when $B \prec 0$. Lemma 4.8 implies that $\underline{B}[P] - B \succeq 0$. Thus replacing $B$ by $\underline{B}[P]$ just means adding a positive semidefinite operator that is zero on span $V$. This keeps the bottom $n$ eigenvalues intact and shift the rest of the eigenvalues upwards. Lemma 5.2 verifies this statement, which implies Eq. (5.2).

LEMMA 5.2. *Let $P$ be the density matrix associated with $A+B$, and $\underline{P}$ be the density matrix associated with $A + \underline{B}[P]$. Then*

$$\lambda_i\big\{A + \underline{B}[P]\big\} \geq \lambda_i = \lambda_i\left\{A + B\right\}$$

*for $i = 1, \ldots, N$, with equality if $i \leq n$. Moreover, $\underline{P} = P$.*

*Proof.* Eq. (3.3) implies that $\{(\lambda_i, v_i)\}_{i=1}^n$ are eigenpairs of $A + \underline{B}[P]$. Hence it is sufficient to show that $\{\lambda_i\}_{i=1}^n$ are also the lowest $n$ eigenvalues.

The Courant-Fischer minimax theorem [12] and Lemma 4.8 give

$$\begin{aligned}
\lambda_i\left\{A + B\right\} &= \min_{\dim S = i} \max_{0 \neq u \in S} \frac{u^*(A + \underline{B}[P])u}{u^* u} \\
&= \min_{\dim S = i} \max_{0 \neq u \in S} \left( \frac{u^*(A+B)u}{u^* u} + \frac{u^*(\underline{B}[P] - B)u}{u^* u} \right) \qquad (5.3) \\
&\geq \min_{\dim S = i} \max_{0 \neq u \in S} \frac{u^*(A+B)u}{u^* u} = \lambda_i.
\end{aligned}$$

Since $\{\lambda_i\}_{i=1}^n$ are already eigenvalues, the only possibility is that $\lambda_i = \lambda_i\left\{A+B\right\}$ for $1 \leq i \leq n$, and hence $P = \underline{P}$. ∎

We now verify that $\mathfrak{F}$ is canonically defined for density matrices in a neighborhood of $P$. This amounts to proving that $A + \underline{B}[Q]$ has a spectral gap for density matrices $Q$ sufficiently close to $P$. By Lemma 5.2, the spectral gap of $A + \underline{B}[P]$ is at least as large as the spectral gap of $A + B$ denoted by $\lambda_g$. In particular, the spectral gap of $A + \underline{B}[P]$ is positive. Then since the $k$-th eigenvalue of a Hermitian matrix $M$ is a Lipschitz function of $M$ (see e.g. [12]), and since $\underline{B}[Q]$ is continuous in the density matrix $Q$ (see Remark 4.3), $A + \underline{B}[Q]$ has a positive spectral gap for density matrices $Q$ sufficiently close to $P$, as claimed.

**5.2. Linearization.** We study the response of $\mathfrak{F}$ to a small perturbation of $P$ in two steps. First we determine the change in the density matrix induced by a small perturbation of the matrix $H = A + \underline{B}[P]$. This gives a Jacobian denoted by $DP_H$. Then we describe how $\underline{B}$ (and hence also the matrix $A + \underline{B}$ of the eigenvalue problem in question in each iteration) responds to the small perturbation of $P$. This gives a Jacobian $D\underline{B}_P$. The composition of these Jacobian operators yields the Jacobian of $\mathfrak{F}$ at $P$, denoted by $D\mathfrak{F}_P$. In the physics literature for solving Hartree-Fock-like equations, $DP_H$ is called the *irreducible polarizability matrix*.

For any orthogonal projector $Q$, let $Q^\perp := I - Q$ denote the orthogonal projector onto $\mathrm{Im}(Q)^\perp$. We first give explicit expressions for $DP_H$ and $D\underline{B}_P$ in Lemma 5.3 and 5.4 respectively, for which the proofs are given in Appendix A.

LEMMA 5.3. *For $H \in \mathbf{H}_N$ with a positive spectral gap, $\Delta H \in \mathbf{H}_N$, $1 \leq n \leq N$ and $\epsilon > 0$ sufficiently small, let $P, P_\epsilon$ be the rank-$n$ density matrices associated with*

*$H$ and $H + \epsilon\Delta H$, respectively. Then*

$$\begin{aligned} DP_H[\Delta H] &= \sum_{i=1}^{n} \sum_{a=n+1}^{N} \frac{1}{\mu_i - \mu_a} u_a (u_a^* \Delta H u_i) u_i^* + \text{h.c.} \\ &= \sum_{i=1}^{n} \left[ P^\perp (\mu_i - H) P^\perp \right]^\dagger \Delta H u_i u_i^* + \text{h.c.}, \end{aligned}$$

*where* h.c. *stands for the Hermitian conjugate of the term that precedes it and $u_1, \ldots, u_N$ are orthonormal eigenvectors of $H$ with corresponding eigenvalues $\mu_1 \leq \cdots \leq \mu_N$. (Note that $\mu_n < \mu_{n+1}$ by assumption.)*

LEMMA 5.4. *For $\epsilon > 0$ sufficiently small, let $P, P_\epsilon$ be density matrices with $\Delta P = \lim_{\epsilon \to 0} (P_\epsilon - P)/\epsilon$. Then*

$$D\underline{B}_P[\Delta P] := \lim_{\epsilon \to 0} \frac{\underline{B}[P_\epsilon] - \underline{B}[P]}{\epsilon} = (B - \underline{B}[P]) (\Delta P)(PBP)^\dagger B + \text{h.c.} \tag{5.4}$$

The composition of Lemma 5.3 with Lemma 5.4 gives an explicit expression for $D\mathfrak{F}_P$:

LEMMA 5.5. *For $\epsilon > 0$ sufficiently small, let $P, P_\epsilon$ be density matrices with $\Delta P = \lim_{\epsilon \to 0} (P_\epsilon - P)/\epsilon$. Then*

$$\begin{aligned} D\mathfrak{F}_P[\Delta P] &:= \lim_{\epsilon \to 0} \frac{\mathfrak{F}(P_\epsilon) - \mathfrak{F}(P)}{\epsilon} \\ &= \sum_{i=1}^{n} \left( P^\perp + (\underline{B}[P] - B)^\dagger (A + B - \lambda_i) P^\perp \right)^\dagger (\Delta P) v_i v_i^* + \text{h.c.} \end{aligned}$$

*Proof.* Applying Lemma 5.3 (with $H = A + \underline{B}[P]$ and $\Delta H = D\underline{B}_P[\Delta P]$) and Lemma 5.4, we have

$$D\mathfrak{F}_P[\Delta P] = \sum_{i=1}^{n} \left[ P^\perp (\lambda_i - H) P^\perp \right]^\dagger \left[ (B - \underline{B}[P]) (\Delta P)(PBP)^\dagger B + \text{h.c.} \right] v_i v_i^* + \text{h.c.}$$

For $i = 1, \ldots, n$, $\big(B - \underline{B}[P]\big) v_i = 0$ and $(PBP)^\dagger B v_i = B^{-1} \underline{B}[P] v_i = v_i$, so our expression for $D\mathfrak{F}_P[\Delta P]$ simplifies to

$$\begin{aligned} D\mathfrak{F}_P[\Delta P] &= \sum_{i=1}^{n} \left[ P^\perp (\lambda_i - A - \underline{B}[P]) P^\perp \right]^\dagger (B - \underline{B}[P]) (\Delta P) v_i v_i^* + \text{h.c.} \\ &= \sum_{i=1}^{n} \left[ P^\perp (\underline{B}[P] - B + A + B - \lambda_i) P^\perp \right]^\dagger (\underline{B}[P] - B) (\Delta P) v_i v_i^* + \text{h.c.} \end{aligned}$$

Now for any $i = 1, \ldots, n$, $\mathrm{Im}(P)$ is an invariant subspace for the self-adjoint operator $A + \underline{B}[P] - \lambda_i$, and $\mathrm{Im}(P)^\perp$ is an invariant subspace as well. As an operator $\mathrm{Im}(P)^\perp \to \mathrm{Im}(P)^\perp$, $A + \underline{B}[P] - \lambda_i$ is positive definite and hence invertible. Thus the pseudoinverse in the preceding expression is effectively taking a matrix inverse on the lower-right block the matrix representation as in (4.6), while all other blocks are zero.

Similarly, $\mathrm{Im}(P)^\perp$ is invariant for $\underline{B}[P] - B$, which is only nonzero in its lower-right block. By Lemma 4.6, $\underline{B}[P] - B$ is positive definite (hence invertible) as an operator

$\mathrm{Im}(P)^\perp \to \mathrm{Im}(P)^\perp$. By taking the factor of $\underline{B}[P] - B$ inside of the pseudoinverse we obtain the desired equality. □

For $i = 1, \ldots, n$, define

$$Z_i := \left( P^\perp + (\underline{B}[P] - B)^\dagger \left( A + B - \lambda_i \right) P^\perp \right)^\dagger .$$

then Lemma 5.5 can be equivalently expressed as

$$D\mathfrak{F}_P[\Delta P] = \sum_{i=1}^{n} Z_i(\Delta P) v_i v_i^* + \text{h.c.}$$

REMARK 5.6. *The matrix of the linear transformation $Z_i$ in Lemma 5.5 is given by*

$$[Z_i]_{V_N} = \begin{pmatrix} 0 & 0 \\ 0 & J_i \end{pmatrix},$$

*where*

$$J_i := \left[ I_{N-n} - (S_{22})^{-1} (\Lambda_2 - \lambda_i) \right]^{-1}.$$

*Here* $\Lambda_2 := \mathrm{diag}(\lambda_{n+1}, \ldots, \lambda_N)$, *and*

$$S_{22} = B_{22} - B_{12}^* B_{11}^{-1} B_{12}$$

*is the Schur complement with* $S_{22} \prec 0$.

We can view the Jacobian $D\mathfrak{F}_P$ as a linear operator on the tangent space at $P$ of the manifold of all rank-$n$ density matrices. We will see later that the set of eigenvalues of $D\mathfrak{F}_P$ is the union of the set of eigenvalues of $\{J_i\}$. We find an upper bound for all eigenvalues of $\{J_i\}$ in Lemma 5.7:

LEMMA 5.7. *For $i = 1, \ldots, n$, $J_i$ is diagonalizable with $\sigma(J_i) \subset (0, 1)$ and*

$$\gamma := \max_{i=1,\ldots,n} \lambda_{\max}(J_i) \leq \frac{\|S_{22}\|_2}{\lambda_g + \|S_{22}\|_2} \leq \frac{\|P^\perp B P^\perp\|_2}{\lambda_g + \|P^\perp B P^\perp\|_2} \leq \frac{\|B\|_2}{\lambda_g + \|B\|_2} < 1.$$

*Proof.* We adopt the notation used in Remark 5.6. Since the eigenvalues of a matrix are invariant under conjugation (i.e. similarity transformation), conjugating $J_i$ by $(\Lambda_2 - \lambda_i)^{1/2}$ yields the equality of spectra

$$\sigma(J_i) = \sigma\left( \left[ I_{N-n} + (\Lambda_2 - \lambda_i)^{1/2} (-S_{22})^{-1} (\Lambda_2 - \lambda_i)^{1/2} \right]^{-1} \right).$$

Here the equality is defined in the sense of sets. The matrix on the right-hand side is positive definite, so $\sigma(Z_i) \subset (0, 1)$ as claimed. In fact, the matrix $(\Lambda_2 - \lambda_i)^{1/2} (-S_{22})^{-1} (\Lambda_2 - \lambda_i)^{1/2}$ is positive definite and we have

$$\sigma(J_i) = \frac{1}{1 + \sigma\left[ (\Lambda_2 - \lambda_i)^{1/2} (-S_{22})^{-1} (\Lambda_2 - \lambda_i)^{1/2} \right]}.$$

Now observe

$$\lambda_{\min}\left[ (\Lambda_2 - \lambda_i)^{1/2} (-S_{22})^{-1} (\Lambda_2 - \lambda_i)^{1/2} \right]$$

$$
\begin{aligned}
&= \left(\lambda_{\max}\left[(\Lambda_2-\lambda_i)^{-1/2}\,(-S_{22})\,(\Lambda_2-\lambda_i)^{-1/2}\right]\right)^{-1} \\
&\geq \left(\|(\Lambda_2-\lambda_i)^{-1/2}\|_2^2\cdot\|S_{22}\|_2\right)^{-1} \\
&\geq \frac{\lambda_g}{\|S_{22}\|_2}.
\end{aligned}
$$

This establishes the first claimed inequality. Recall that $S_{22} \preceq 0$, but also $S_{22} = B_{22} - B_{12}^* B_{11}^{-1} B_{12}$, so $S_{22} \succeq B_{22}$. Thus

$$\|S_{22}\|_2 \leq \|B_{22}\|_2 = \|P^\perp B P^\perp\|_2 \leq \|B\|_2.$$

Since $x \mapsto x/(1+x)$ is increasing for $x \geq 0$, this proves the rest of the inequalities.

The diagonalizability of $J_i$ is implied by the similarity transformation. □

**5.3. Dynamical systems perspective on adaptive compression.** In order to study the local convergence properties of the fixed point iteration map $\mathfrak{F}$, we first note that the set of all density matrices $\mathcal{D}$ is not a subspace, but a smooth submanifold of $\mathbb{C}^{N\times N} \simeq \mathbb{R}^{2N^2}$. $\mathcal{D}$ can be identified with the Grassmannian $\mathbf{Gr}(n, \mathbb{C}^N)$, which is the set of all complex $n$-dimensional subspaces of $\mathbb{C}^N$. Since the fixed point iteration map $\mathfrak{F}$ is a map from $\mathcal{D}$ to itself and is smooth on a neighborhood of $P$, we consider the linearization of $\mathfrak{F}$ about the fixed point $P$ is the tangent space $T_P\mathcal{D} \subset \mathbb{C}^{N\times N} \simeq \mathbb{R}^{2N^2}$. This tangent space can be characterized as follows.

First note that any smooth path of rank-$n$ density matrices, denoted by $\gamma(t)$ with $\gamma(0) = P$, can be expressed as

$$\gamma(t) = V_N U(t) \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} U(t)^* V_N^*,$$

where $U(t)$ is a smooth path of unitary matrices with $U(0) = I$ and $V_N = [v_1, \ldots, v_N]$. Since the Lie algebra of the unitary group (i.e. the tangent space at the identity element) is the set of skew-Hermitian matrices, we have

$$U'(0) = \begin{pmatrix} Y & -X^* \\ X & Z \end{pmatrix},$$

where $Y^* = -Y$ and $Z^* = -Z$. Then

$$\gamma'(0) = V_N \begin{pmatrix} Y & -X^* \\ X & Z \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} V_N^* + \text{h.c.} = \widetilde{V} \begin{pmatrix} 0 & X^* \\ X & 0 \end{pmatrix} V_N^*.$$

Hence the tangent space

$$T_P\mathcal{D} = \left\{ V_N \begin{pmatrix} 0 & X^* \\ X & 0 \end{pmatrix} V_N^* \,:\, X \in \mathbb{C}^{(N-n)\times n} \right\}, \tag{5.5}$$

and we can make the identification $T_P\mathcal{D} \simeq \mathbb{C}^{(N-n)\times n}$. Observe that the map $\Phi : \mathbb{C}^{(N-n)\times n} \to \mathcal{D}$ defined by

$$X \mapsto V_N \left[\exp\begin{pmatrix} 0 & -X^* \\ X & 0 \end{pmatrix}\right] \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \left[\exp\begin{pmatrix} 0 & -X^* \\ X & 0 \end{pmatrix}\right]^* V_N^* \tag{5.6}$$

is a local diffeomorphism near the origin. Then for $Q \in \mathcal{D}$ sufficiently close to $P$, we can identify $Q$ with $X := \Phi^{-1}(Q) \in \mathbb{C}^{(N-n)\times n}$. Then we can identify $\mathfrak{F}$ with a map $\mathfrak{H}$ defined on a neighborhood $\mathcal{U}$ of the origin in $\mathbb{C}^{(N-n)\times n}$

REMARK 5.8. *Adopting this perspective, Remark 5.6 translates to*

$$D\mathfrak{H}_0[X] = (J_1X_1, \ldots, J_nX_n),$$

*for any $X = (X_1, \ldots, X_n) \in \mathbb{C}^{(N-n)\times n}$, where $D\mathfrak{H}_0$ is the usual Jacobian of the map $\mathcal{U} \to \mathbb{C}^{(N-n)\times n}$ at the origin, naturally viewed as a tensor in $\mathbb{C}^{(N-n)\times n\times(N-n)\times n}$. Identifying tangent vector $X$ with its vectorization in $\mathbb{C}^{(N-n)n}$, the matricized representation of $D\mathfrak{H}_0$ in $\mathbb{C}^{(N-n)n\times(N-n)n}$ yields*

$$D\mathfrak{H}_0 = \begin{pmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_n \end{pmatrix}.$$

Near the fixed point $P$, we can view Algorithm 1 as a discrete-time dynamical system on $\mathbb{C}^{(N-n)n}$. The stability of the fixed point $P$ is then determined by the spectrum $\sigma(D\mathfrak{H}_0)$ of the Jacobian $D\mathfrak{H}_0$, which is the union of the spectra $\sigma(J_i)$ over $i = 1, \ldots, n$.

**5.4. Asymptotic convergence rate.** We will make use the following Lemma to show the local convergence.

LEMMA 5.9. *Let $F : \mathbb{R}^p \cap B_\delta(0) \to \mathbb{R}^p$ be a smooth map such that $F(0) = 0$, $DF(0)$ is diagonalizable, and the spectral radius $\gamma := \sup|\sigma(DF(0))|$ of $DF(0)$ is strictly less than 1. Then for any $\epsilon > 0$ for which $\gamma + \epsilon < 1$, there exist constants $C, c > 0$ such that if $\|x\|_2 < c$, then $\|F^k(x)\|_2 \le C(\gamma+\epsilon)^k\|x\|_2$ for all $k \ge 0$.*

*Proof.* First note that we can assume that in fact $DF(0)$ is diagonal by replacing $F$ with $\phi^{-1} \circ F \circ \phi$ for a suitable change of basis $\phi$. Then $\|DF(0)\|_2 = \gamma$, and there exists $c$ such that $\|y\|_2 < c$ implies $\|DF(y)\|_2 < \gamma + \epsilon$. Thus if $\|x\|_2 < c$, then

$$\begin{aligned} \|F(x)\|_2 &= \|F(x) - F(0)\|_2 \\ &= \left\| \int_0^1 DF(tx) \cdot x \, \mathrm{d}t \right\|_2 \\ &\le \int_0^1 \|DF(tx)\|_2\|x\|_2 \, \mathrm{d}t \le (\gamma+\epsilon)\|x\|_2. \end{aligned}$$

Repeated application of this inequality yields the result. ∎

REMARK 5.10. *The reader familiar with dynamical systems should note that Lemma 5.9 is almost a recapitulation of the stable manifold theorem in the case that the local stable manifold has full dimension.*

Now we are ready to prove Theorem 1.2, which is stated more precisely in Theorem 5.11.

THEOREM 5.11. *Let $\epsilon > 0$ be small enough so that $\gamma + \epsilon < 1$, where $\gamma$ is as in Lemma 5.7. Then there exist constants $C, c > 0$ such that if $\|P^{(0)} - P\|_2 \le c$, then*

$$\|P^{(k)} - P\|_2 \le C(\gamma+\epsilon)^k\|P^{(0)} - P\|_2$$

*for all $k \ge 0$.*

*Proof.* Fix $\epsilon$ as in the statement of the theorem. We can identify $\mathbb{C}^{(N-n)n}$ with $\mathbb{R}^{2(N-n)n}$, and the corresponding realification of $D\mathfrak{H}_0$ has all of its eigenvalues in $(0,\gamma]$. (It has two copies of each of the eigenvalues of $D\mathfrak{H}_0$ as an operator $\mathbb{C}^{(N-n)n} \to \mathbb{C}^{(N-n)n}$.) By Lemma 5.9, there exists a neighborhood $\mathcal{V}$ of 0 within $\mathcal{U} \subset \mathbb{C}^{(N-n)n}$ and a constant $C$ such that $\mathfrak{H}^k(X^{(0)}) \in \mathcal{V}$ and moreover $\|\mathfrak{H}^k(X^{(0)})\|_2 \le C(\gamma+\epsilon)^k \|X^{(0)}\|_2$ for all $k \ge 0$. From (5.6) we have

$$P^{(k)} = \widetilde{V} \left[ \exp \begin{pmatrix} 0 & -(X^{(k)})^* \\ X^{(k)} & 0 \end{pmatrix} \right] \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \left[ \exp \begin{pmatrix} 0 & -(X^{(k)})^* \\ X^{(k)} & 0 \end{pmatrix} \right]^* \widetilde{V}^*.$$

Since

$$P = \widetilde{V} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \widetilde{V}^*,$$

it follows (for a possibly enlarged constant $C$) that $\|P^{(k)} - P\|_2 \le C(\gamma+\epsilon)^k \|P^{(0)} - P\|_2$, as was to be shown. ∎

REMARK 5.12. *Recall from Lemma 5.7 that* $\gamma \le \frac{\|B\|_2}{\lambda_g + \|B\|_2} < 1$, *so we have a linear rate of convergence that depends only on the ratio* $\|B\|_2/\lambda_g$. *If this ratio is smaller, then the convergence is faster, and vice-versa.*

**5.5. Convergence of sub-projectors.** Now we prove Theorem 1.6 regarding the rate of convergence of the rank-$m$ sub-projectors $P_m^{(k)}$ to $P_m$. In this section we use $C$ to denote a constant that possibly changes across usages and is understood to be sufficiently large in each context.

The important observation is that $P_m$ can be identified with an invariant submanifold for the dynamics, to which the dynamics are attracted via a (relatively) rapid transient.

Consider

$$\mathcal{D}_m := \{Q \in \mathcal{D} \,:\, Q|_{\mathrm{Im}(P_m)} = \mathrm{Id}_{\mathrm{Im}(P_m)}\},$$

which is a submanifold of $\mathcal{D}$, and can be identified as the submanifold of $\mathbf{Gr}(n, \mathbb{C}^N)$ consisting of the $n$-dimensional subspaces of $\mathbb{C}^N$ that contain $\mathrm{Im}(P_m)$. This is in turn isomorphic to $\mathbf{Gr}(n-m, \mathbb{C}^N/\mathrm{Im}(P_m)) \simeq \mathbf{Gr}(n-m, \mathbb{C}^{N-m})$.

We assume that $\lambda_{m+1} - \lambda_m > 0$, and we allow this gap to be small in practice. Then

$$\lambda_i\{A + \underline{B}[Q]\} = \lambda_i\{A + \underline{B}[P]\} + \mathcal{O}(\|Q-P\|_2^2).$$

In particular, for each $i = m+1, \ldots, N$, we have $\lambda_i\{A + \underline{B}[Q]\} > \lambda_m$ for all $Q$ sufficiently close to $P$. If $Q \in \mathcal{D}_m$, then $(\lambda_i, v_i)$ is an eigenpair for $A + \underline{B}[Q]$ for $i = 1, \ldots, m$, and these eigenvalues are the lowest $m$ eigenvalues of $A + \underline{B}[Q]$. It follows that $\mathrm{Im}(\mathfrak{F}(Q)) \supset \mathrm{Im}(P_m)$. Hence near the fixed point $P$, $\mathcal{D}_m$ is invariant under the fixed point iteration map $\mathfrak{F}$.

For $Q \in \mathcal{D}$, define $\mathfrak{F}_m(Q)$ to be the rank-$m$ projector onto the span of the lowest $m$ eigenvectors of $A + \underline{B}[Q]$. The assumption $\lambda_{m+1} - \lambda_m > 0$ guarantees that this map is canonically defined and smooth near $P$, and $\mathfrak{F}_m(Q) = P_m$ for all $Q \in \mathcal{D}_m$ sufficiently close to $P$. Then there is a neighborhood $\mathcal{N}$ of $P$ in $\mathcal{D}$ such that $\mathfrak{F}(\mathcal{D}_m \cap \mathcal{N}) \subset \mathcal{D}_m$ and such that $\mathfrak{F}_m(\mathcal{D}_m \cap \mathcal{N}) = P_m$. In particular, we have constructed a local invariant manifold $\mathcal{D}_m$ for the dynamics due to the fixed point iteration.

We would like to prove that the dynamics converge rapidly to this invariant manifold locally, in the sense that

$$\operatorname{dist}(P^{(k)}, \mathcal{D}^m \cap \mathcal{N}) \leq C\gamma_m^k \cdot \operatorname{dist}(P^{(0)}, \mathcal{D}^m \cap \mathcal{N}), \tag{5.7}$$

where we can take $\gamma_m := \|B\|_2/(\|B\|_2 + \Delta)$ and where 'dist' indicates the distance between sets induced by the norm $\|\cdot\|_2$. We claim that in fact Theorem 1.6 would follow from (5.7), together with the preceding remarks. We will justify the choice of constant $\gamma_m$ later, but for now we map out the rest of the argument.

To see the claim, note that since $\mathfrak{F}_m$ is smooth near $P$ (hence in particular locally Lipschitz), there exists $L$ such that, for all $\epsilon$ sufficiently small, if $Q$ satisfies $\operatorname{dist}(Q, \mathcal{D}_m \cap \mathcal{N}) \leq \epsilon$, then

$$L\epsilon \geq \operatorname{dist}(\mathfrak{F}_m(Q), \mathfrak{F}_m(\mathcal{D}_m \cap \mathcal{N})) = \operatorname{dist}(\mathfrak{F}_m(Q), P_m) = \|\mathfrak{F}_m(Q) - P_m\|_2.$$

Thus if we can establish (5.7), then substituting $Q = P^{(k)}$ yields

$$\|P_m^{(k+1)} - P_m\| \leq C\gamma_m^k \cdot \operatorname{dist}(P^{(0)}, \mathcal{D}^m \cap \mathcal{N}) \leq C\gamma_m^k \|P - P^{(0)}\|_2,$$

establishing Theorem 1.6.

We have then reduced Theorem 1.6 to the following lemma.

LEMMA 5.13. *There is a neighborhood $\mathcal{W}$ of $P$ in $\mathcal{D}$ such that if $P^{(0)} \in \mathcal{W}$, then* $\operatorname{dist}(P^{(k)}, \mathcal{D}_m \cap \mathcal{N}) \leq C\gamma_m^k \cdot \operatorname{dist}(P^{(0)}, \mathcal{D}^m \cap \mathcal{N})$.

In order to motivate the constant $\gamma_m$, note that $T_P\mathcal{D}_m$, considered as a subspace of $\mathbb{C}^{(N-n)n} \simeq T_P\mathcal{D}$, is given by

$$T_2 := \{(X_1, \ldots, X_n) \in \mathbb{C}^{(N-n)n} \,:\, X_1 = \cdots = X_m = 0\},$$

and we have locally the splitting $T_P\mathcal{D} \simeq \mathbb{C}^{(N-n)n} = T_1 \oplus T_2$, where

$$T_1 := \{(X_1, \ldots, X_n) \in \mathbb{C}^{(N-n)n} \,:\, X_{m+1} = \cdots = X_N = 0\}.$$

Observe that the eigenvalues of $D\mathfrak{H}_0|_{T_1}$ are the eigenvalues of $J_1, \ldots, J_m$. By the proof of Lemma 5.7, all of these eigenvalues are in $(0, \gamma_m)$, so the spectrum of $D\mathfrak{H}_0|_{T_1}$ is contained in $(0, \gamma_m)$. (The eigenvalues of $D\mathfrak{H}_0|_{T_2}$ are the eigenvalues of $J_{m+1}, \ldots, J_N$, which are all in $(0, 1)$.) At least formally, this discussion motivates the statement of Lemma 5.13. By considering a smooth change of coordinates near $P$ that straightens the invariant submanifold $\mathcal{D}_m$ and then diagonalizes the Jacobian, we can replace Lemma 5.13 with the following:

LEMMA 5.14. *Let $F : \mathbb{R}^p \cap B_\delta(0) \to \mathbb{R}^p$ be a smooth map such that $F(0) = 0$, $DF(0)$ is diagonal, $0 \prec DF(0) \prec 1$, and $DF(0)|_{E_1} \preceq \alpha < 1$, where $E_1 := \operatorname{span}\{e_1, \ldots, e_r\}$ for $r \leq p$. Further suppose that $E_2 := E_1^\perp$ is invariant under $F$, i.e., $F(E_2 \cap B_\delta(0)) \subset E_2$. Then there exists $\delta' \in (0, \delta)$ such that $F$ maps $B_{\delta'}(0)$ into itself and such that for any $x \in B_{\delta'}(0)$,*

$$\operatorname{dist}(F^k(x), E_2 \cap B_\delta(0)) \leq \alpha^k \cdot \operatorname{dist}(x, E_2 \cap B_\delta(0)).$$

*Proof.* See Appendix B. □

REMARK 5.15. *Note carefully that we* do not *consider a change of coordinates that produces a linear dynamical system, i.e., we do not assume $F$ is linear in Lemma 5.14. In general, such a change of coordinates does exist near a hyperbolic fixed point by the Hartman-Grobman theorem (see, e.g., Theorem 10.4 of [36]), but it is only guaranteed to be a homeomorphism (not necessarily Lipschitz). We need the change of coordinates to be Lipschitz in order to compare distances up to a constant.*

**6. Global convergence analysis.** Before providing a roadmap for the proof of the global convergence properties in Theorem 1.3, we first show that the adaptive compression method *cannot* be expected to converge globally to the solution of (1.1) for *every* initial guess $P^{(0)}$.

Consider taking $N = 2$, $n = 1$, $A$ is a zero matrix, and

$$B = \begin{pmatrix} -2 & 0 \\ 0 & -1 \end{pmatrix}.$$

Note that the true density matrix is $P = e_1 e_1^*$, where $e_1 = (1,0)^T, e_2 = (0,1)^T$. However $e_2 e_2^*$ is also a fixed point of $\mathfrak{F}$. Thus if we take $P^{(0)} = e_2 e_2^*$, we get convergence to the wrong fixed point.

A slightly more sophisticated example demonstrates that it is possible for Algorithm 1 to stall on some incorrect fixed point, even if not initialized there. Take $N = 3$ and $n = 1$ with

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -4 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

The true density matrix is $P = e_1 e_1^*$. However, suppose that $P^{(0)} = e_3 e_3^*$. Then

$$A + \underline{B}[P^{(0)}] = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

so $P^{(1)} = e_2 e_2^*$. Now

$$A + \underline{B}[P^{(1)}] = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

so $P^{(1)}$ is a fixed point, and Algorithm 1 fails to converge.

In the sequel we will see that such incorrect fixed points are unstable. Before embarking on the global convergence analysis, we pause to provide a brief outline of Section 6.

In Section 6.1 we introduce a key property of Algorithm 1: each of the bottom $n$ eigenvalues $\lambda_1^{(k)}, \ldots, \lambda_n^{(k)}$ of $A + \underline{B}[P^{(k-1)}]$ is monotonically non-increasing in $k$. We call this property *eigenvalue monotonicity*. Eigenvalue monotonicity implies that $\sum_{i=1}^n \lambda_i^{(k)}$ is convergent in $k$. In particular, when $k$ is large, $\sum_{i=1}^n \lambda_i^{(k)}$ does not change much across iterations. Lemma 6.4 shows that the change of $P^{(k)}$ across one iteration can be controlled by the change of $\sum_{i=1}^n \lambda_i^{(k)}$. So when $k$ is large, $\mathfrak{F}(P^{(k)}) \approx P^{(k)}$, i.e., the point $P^{(k)}$ is almost fixed by the mapping $\mathfrak{F}$.

Unfortunately, this is not yet enough to directly imply that the sequence $P^{(k)}$ is convergent, but one might hope that a point that is close to being fixed is close to some fixed point! Notice that a fixed point $P_f$ of $\mathfrak{F}$ must satisfy the condition that $\mathrm{Im}(P_f)$ is an invariant subspace for $A + B$, i.e., must satisfy $P_f = \sum_{i=1}^n u_i u_i^*$, where $u_i$ are eigenvectors of $A + B$. We will show (see Lemma 6.6) that a point that is almost fixed is indeed almost a point of this form.

Under suitable conditions that hold generically in the sense to be discussed in Section 6.2, $\mathfrak{F}$ always has finitely many points $\{P_\tau\}$ of such form. These points must

be mutually isolated since they are only finite in number. Since (1) $P^{(k)}$ must be close to at least one of these points $P_\tau$ for any $k$ large, (2) $P^{(k)}$ changes by a vanishingly small amount as $k$ becomes large, and (3) the points $P_\tau$ are mutually isolated, it follows that $P^{(k)}$ converges to one such point $P_\tau$ as $k \to \infty$. One might expect that such a limit point must actually be a fixed point $P_f$, and indeed this is true.

In summary, these arguments establish that we have global convergence to a fixed point. We have already demonstrated with toy counterexamples that this limit point may differ from the true density matrix $P$. However, the conditions of Section 6.2 also ensure that $A + \underline{B}[P_f]$ has a positive spectral gap for any fixed point $P_f$, which in turn ensures that $\mathfrak{F}$ is smooth near each of the fixed points $P_f$. A linearization-based analysis similar to that of Section 5.2 will reveal that all such pathological points are unstable. Together with careful measure-theoretic arguments (see Lemma 6.21), we establish Theorem 1.3.

**6.1. Eigenvalue monotonicity.** We now highlight a significant feature of Algorithm 1, which is the key for the proof of global convergence properties.

LEMMA 6.1 (Eigenvalue monotonicity). *For $i = 1, \dots, n$,*

$$\lambda_i^{(k)} := \lambda_i\big\{A + \underline{B}[P^{(k-1)}]\big\}$$

*is non-increasing in $k$.*

*Proof.* Let $v_i^{(k)} \in \mathrm{Im}(P^{(k)})$ be orthonormal eigenvectors of $A + \underline{B}[P^{(k-1)}]$ corresponding to the eigenvalues $\lambda_i^{(k)}$ for $i = 1, \dots, n$, and let $S_i^{(k)} = \mathrm{span}\{v_1^{(k)}, \dots, v_i^{(k)}\}$. Then we compute, for $i = 1, \dots, n$,

$$\begin{aligned}
\lambda_i^{(k+1)} &= \min_{\dim S = i} \max_{u \in S \setminus \{0\}} \frac{u^*\big(A + \underline{B}[P^{(k)}]\big)u}{u^*u} \\
&\leq \max_{u \in S_i^{(k)} \setminus \{0\}} \frac{u^*\big(A + \underline{B}[P^{(k)}]\big)u}{u^*u} \\
&\overset{\text{(i)}}{=} \max_{u \in S_i^{(k)} \setminus \{0\}} \frac{u^*\,(A + B)\,u}{u^*u} \\
&\overset{\text{(ii)}}{\leq} \max_{u \in S_i^{(k)} \setminus \{0\}} \frac{u^*\big(A + \underline{B}[P^{(k-1)}]\big)u}{u^*u} \\
&= \lambda_i^{(k)},
\end{aligned}$$

where (i) follows from the fact that $B \equiv \underline{B}[P^{(k)}]$ on $\mathrm{Im}(P^{(k)}) \supset S_i^{(k)}$ and (ii) follows from Lemma 4.8. This completes the proof. □

Thus we may think of Algorithm 1 as performing a descent on the bottom $n$ eigenvalues of $A + \underline{B}[Q]$ as $Q = P^{(k)}$ is updated iteratively. In order to achieve global convergence, we would need that these eigenvalues are globally minimized at $Q = P$. Indeed, this is the case:

LEMMA 6.2 (Global eigenvalue minimality). *For $i = 1, \dots, n$ and all density matrices $Q$,*

$$\lambda_i = \lambda_i\big\{A + \underline{B}[P]\big\} \leq \lambda_i\big\{A + \underline{B}[Q]\big\}.$$

*Proof.* Let $v_i^Q \in \mathrm{Im}(\mathfrak{F}(Q))$ be orthonormal eigenvectors of $A+\underline{B}[Q]$ corresponding to the eigenvalues $\lambda_i\{A+\underline{B}[Q]\}$ for $i=1,\dots,n$, and let $S_i^Q = \mathrm{span}\{v_1^Q,\dots,v_i^Q\}$. Again we compute, for $i=1,\dots,n$,

$$\begin{aligned}\lambda_i &= \min_{\dim S=i}\max_{u\in S\setminus\{0\}}\frac{u^*(A+B)u}{u^*u}\\ &\le \max_{u\in S_i^Q\setminus\{0\}}\frac{u^*(A+B)u}{u^*u}\\ &\le \max_{u\in S_i^Q\setminus\{0\}}\frac{u^*\big(A+\underline{B}[Q]\big)u}{u^*u}\\ &= \lambda_i\big\{A+\underline{B}[Q]\big\}.\end{aligned}$$

□

We now examine some consequences of eigenvalue monotonicity with a view toward establishing a global convergence result. First, from Lemma 6.1 we have the immediate corollary.

COROLLARY 6.3 (Eigenvalue convergence). $\lim_{k\to\infty}\lambda_i^{(k)}$ *exists for* $i=1,\dots,n$.

From this corollary and a refinement of earlier arguments, we derive the following result, which will be instrumental in establishing global convergence. The main idea of this result is that a small change in eigenvalues across one iteration is only possible if the density matrix changes by a correspondingly small amount.

LEMMA 6.4. *There exists a constant* $C>0$ *(depending only on* $B,n$*) such that*

$$\|P^{(k)}-P^{(k-1)}\|_2 \le C\sqrt{\delta^{(k)}}$$

*for all* $k$*, where*

$$\delta^{(k)} := \sum_{i=1}^n\left(\lambda_i^{(k)}-\lambda_i^{(k+1)}\right).$$

*It follows (by Corollary 6.3) that* $\|P^{(k)}-P^{(k-1)}\|_2\to 0$ *as* $k\to\infty$.

*Proof.* As in the proof of Lemma 6.1, let $v_i^{(k)}\in\mathrm{Im}(P^{(k)})$ be orthonormal eigenvectors of $A+\underline{B}[P^{(k-1)}]$ corresponding to the eigenvalues $\lambda_i^{(k)}$ for $i=1,\dots,n$, and let $S_i^{(k)}=\mathrm{span}\{v_1^{(k)},\dots,v_i^{(k)}\}$.

$$\begin{aligned}\sum_{i=1}^n\lambda_i^{(k+1)} &= \inf_{u_1,\dots,u_n\ \text{orthonormal}}\left\{\sum_{i=1}^n u_i^*\big(A+\underline{B}[P^{(k)}]\big)u_i\right\}\\ &\le \sum_{i=1}^n\big(v_i^{(k)}\big)^*\big(A+\underline{B}[P^{(k)}]\big)v_i^{(k)}\\ &= \sum_{i=1}^n\big(v_i^{(k)}\big)^*\big(A+B\big)v_i^{(k)}\\ &= \sum_{i=1}^n\big(v_i^{(k)}\big)^*\big(A+\underline{B}[P^{(k-1)}]\big)v_i^{(k)}-\sum_{i=1}^n\big(v_i^{(k)}\big)^*\big(\underline{B}[P^{(k-1)}]-B\big)v_i^{(k)}\end{aligned}$$

$$= \sum_{i=1}^{n} \lambda_i^{(k)} - \sum_{i=1}^{N} \big(v_i^{(k)}\big)^* P^{(k)} \big(\underline{B}[P^{(k-1)}] - B\big) v_i^{(k)}$$
$$= \sum_{i=1}^{n} \lambda_i^{(k)} - \operatorname{Tr}\Big[P^{(k)}\big(\underline{B}[P^{(k-1)}] - B\big)\Big].$$

We have

$$\operatorname{Tr}\Big[P^{(k)}\big(\underline{B}[P^{(k-1)}] - B\big)P^{(k)}\Big] \leq \sum_{i=1}^{n} \Big(\lambda_i^{(k)} - \lambda_i^{(k+1)}\Big) = \delta^{(k)}. \tag{6.1}$$

At this point we should hope that the left-hand side of (6.1) provides an upper bound for some measure of the distance between $P^{(k)}$ and $P^{(k-1)}$, and indeed this will be the case.

We first prove the following lemma.

LEMMA 6.5. *There exists $t > 0$ depending only on $B, n$ such that*

$$\operatorname{Tr}\big[R\big(\underline{B}[Q] - B\big)\big] \geq t \cdot \operatorname{Tr}\big[R\big(I - Q\big)\big]$$

*for all density matrices $Q$ and $R$.*

*Proof.* Note that

$$\lambda_{\min}\big\{Q + \big(\underline{B}[Q] - B\big)\big\} > 0,$$

for all density matrices $Q$. By the continuity of $\underline{B}$ on density matrices and $\lambda_{\min}$ on Hermitian matrices, as well as the compactness of the space of density matrices, it follows that there exists $t > 0$ such that

$$Q + \big(\underline{B}[Q] - B\big) \succeq t$$

for all density matrices $Q$. Furthermore, we can write

$$\underline{B}[Q] - B = (I - Q)\big[Q + \big(\underline{B}[Q] - B\big)\big](I - Q) \succeq t(I - Q).$$

Now the trace of a product of positive semidefinite matrices is nonnegative, so

$$\operatorname{Tr}\big(R\big[(\underline{B}[Q] - B) - t(I - Q)\big]\big) \geq 0,$$

for all density matrices $Q, R$, which yields the lemma. □

We now resume the proof of Lemma 6.4. Let $R = P^{(k)}$ and $Q = P^{(k-1)}$ in Lemma 6.5 to obtain

$$\operatorname{Tr}\Big[P^{(k)}\big(\underline{B}[P^{(k-1)}] - B\big)\Big] \geq t \cdot \operatorname{Tr}\Big[P^{(k)}\big(I - P^{(k-1)}\big)\Big],$$

and combine with (6.1) to give

$$\operatorname{Tr}\Big[P^{(k)}\big(I - P^{(k-1)}\big)\Big] \leq \alpha\delta^{(k)},$$

where $\alpha := t^{-1} > 0$. Now

$$\operatorname{Tr}\Big[P^{(k)}\big(I - P^{(k-1)}\big)\Big] = \operatorname{Tr}\Big[P^{(k)} - P^{(k)}P^{(k-1)}\Big] = n - \operatorname{Tr}\Big[P^{(k)}P^{(k-1)}\Big],$$

so in fact we have

$$\mathrm{Tr}\left[P^{(k)}P^{(k-1)}\right] \geq n - \alpha\delta^{(k)}. \tag{6.2}$$

To conclude the proof, observe

$$\begin{aligned}\|P^{(k)} - P^{(k-1)}\|_2^2 &\leq \|P^{(k)} - P^{(k-1)}\|_F^2 \\ &= \mathrm{Tr}\left[(P^{(k)} - P^{(k-1)})(P^{(k)} - P^{(k-1)})\right] \\ &= 2n - 2\cdot\mathrm{Tr}\left[P^{(k)}P^{(k-1)}\right] \\ &\leq 2\alpha\delta^{(k)},\end{aligned}$$

where we have used (6.2) in the last line. □

Note carefully that Lemma 6.4 *does not* imply that the sequence $P^{(k)}$ is convergent. In particular, we do not yet see that $P^{(k)}$ is Cauchy; we are only able to bound the change in density matrix over a single iteration. However, Lemma 6.4 does establish that for $k$ large, the density matrix $P^{(k)}$ is almost fixed by $\mathfrak{F}$. Note that any fixed point $P_f$ is a projector of the form $P_f = \sum_{i=1}^n u_i u_i^*$, where the $u_i$'s are eigenvectors of $A+B$. This motivates the following lemma, which implies that for $k$ large, $P^{(k)}$ is close to some point of this form.

LEMMA 6.6. *There exists a constant $C > 0$ depending only on $A, B, n$ such that if $\|\mathfrak{F}(Q) - Q\|_2 \leq \epsilon$ for any density matrix $Q$, then $Q = \sum_{i=1}^n u_i u_i^* + M$, where the $u_i$'s are orthonormal eigenvectors of $A+B$ and $\|M\|_2 \leq C\epsilon$.*

*Proof.* Write $\mathfrak{F}(Q) = \sum_{i=1}^n w_i w_i^*$, where $w_1, \ldots, w_n$ are orthonormal eigenvectors of $A + \underline{B}[Q]$ with corresponding eigenvalues $\mu_1 \leq \cdots \leq \mu_n$. Let $z_i = Qw_i$ for $i = 1, \ldots, n$. Observe that

$$\left[(A+B) + (\underline{B}[Q] - B)(\mathfrak{F}(Q) - Q)\right] w_i = \mu_i w_i$$

for $i = 1, \ldots, n$, since $(\underline{B}[Q] - B)Q = 0$ and $\mathfrak{F}(Q)w_i = w_i$. Therefore

$$\begin{aligned}\|(A+B)w_i - \mu_i w_i\| &= \|(\underline{B}[Q] - B)(\mathfrak{F}(Q) - Q)w_i\| \\ &\leq \|\underline{B}[Q] - B\|_2 \|\mathfrak{F}(Q) - Q\|_2.\end{aligned}$$

We assume $\|\mathfrak{F}(Q) - Q\|_2 \leq \epsilon$ as in the statement of the theorem, and recall $\|\underline{B}[Q] - B\|_2 \leq \|B\|_2$, so we have shown that

$$\|(A+B)w_i - \mu_i w_i\| \leq C\epsilon, \tag{6.3}$$

where $C = \|B\|_2$. In other words, if $\epsilon$ is small, then $w_i$ nearly satisfies the condition of being eigenvectors of $A+B$ with the corresponding eigenvalue $\mu_i$. We now aim to show that this implies that each $w_i$ is in fact close to some eigenvector of $A+B$. We remark that the discussion below is related to the "sin $\theta$ theorem" of Davis and Kahan [6], which characterizes the relation between the error of an approximate eigenvector and its residual.

To this end, let $v_1, \ldots, v_N$ be orthonormal eigenvectors of $A+B$ with corresponding eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$, and write $w_i = \sum_{j=1}^N c_{ij} v_j$. Then

$$\|(A+B)w_i - \mu_i w_i\|^2 = \left\|\sum_{j=1}^N c_{ij}(\lambda_j - \mu_i)v_j\right\|^2 = \sum_{j=1}^N |c_{ij}|^2 |\lambda_j - \mu_i|^2.$$

Combining with (6.3) yields

$$|c_{ij}|^2|\lambda_j - \mu_i|^2 \leq C^2\epsilon^2. \tag{6.4}$$

Let $\delta > 0$ be smaller than the gap between any pair of *distinct* eigenvalues of $A+B$. (Note carefully that this is still possible even if $A+B$ has repeated eigenvalues.) Fix $i$ for the moment, and decompose

$$w_i = \sum_{\{j\,:\,|\lambda_j-\mu_i|>\delta\}} c_{ij}v_j + \underbrace{\sum_{\{j\,:\,|\lambda_j-\mu_i|\leq\delta\}} c_{ij}v_j}_{=:\tilde{u}_i}.$$

Notice that if $|\lambda_j - \mu_i| > \delta$, then $|c_{ij}|^2 \leq C^2\epsilon^2/\delta^2$ by (6.4). Thus

$$\|w_i - \tilde{u}_i\|^2 = \sum_{\{j\,:\,|\lambda_j-\mu_i|>\delta\}} |c_{ij}|^2 \leq \frac{NC^2}{\delta^2}\epsilon^2.$$

In particular, for $\epsilon$ sufficiently small, $\|w_i - \tilde{u}_i\| < 1$, which implies that $\tilde{u}_i \neq 0$.

By the definition of $\delta$, there is at most one element in the set $\{\lambda_j\,:\,|\lambda_j - \mu_i| \leq \delta\}$. But since $\tilde{u}_i \neq 0$, there must also be at least one element. We denote this element by $\lambda[i]$. Observe that $\tilde{u}_i$ is in the $\lambda[i]$-eigenspace of $A+B$.

We have established (for a possibly enlarged constant $C$ depending only on $A, B$) that if $\epsilon$ is sufficiently small, then

$$\|w_i - \tilde{u}_i\| \leq C\epsilon.$$

Then the $\tilde{u}_i$ must be linearly independent for $\epsilon$ sufficiently small. Moreoever, since the $w_i$ are orthonormal, this implies (possibly enlarging $C$ once again) that

$$\left\|\mathfrak{F}(Q) - \tilde{U}(\tilde{U}^*\tilde{U})^{-1}\tilde{U}^*\right\|_2 = \left\|W(W^*W)^{-1}W^* - \tilde{U}(\tilde{U}^*\tilde{U})^{-1}\tilde{U}^*\right\|_2 \leq C\epsilon$$

for $\epsilon$ sufficiently small, where $\tilde{U} := [\tilde{u}_1, \ldots, \tilde{u}_n]$, so $\tilde{U}(\tilde{U}^*\tilde{U})^{-1}\tilde{U}^*$ is the orthogonal projector onto the span of the $\tilde{u}_i$, and likewise $W := [w_1, \ldots, w_n]$. Now the $\tilde{u}_i$'s are unnormalized eigenvectors of $A+B$ with possibly repeated eigenvalues, hence possibly not orthonormal or even orthogonal. However, $\mathrm{span}\{\tilde{u}_1, \ldots, \tilde{u}_n\}$ is invariant under $A+B$, hence can also be endowed with an orthonormal basis of eigenvectors $u_1, \ldots, u_n$ of $A+B$. This yields the equivalent orthogonal projector $\sum_{i=1}^n u_i u_i^*$. Now since $\|\mathfrak{F}(Q) - Q\|_2 \leq \epsilon$, this means that (enlarging $C$ again)

$$\left\|Q - \sum_{i=1}^n u_i u_i^*\right\|_2 \leq C\epsilon$$

for $\epsilon$ sufficiently small.

This establishes the statement of the lemma under the condition that $\epsilon$ is assumed sufficiently small. But since the space of density matrices is compact, there exists $K > 0$ such that $\|Q - \sum_{i=1}^n v_i v_i^*\|_2 \leq K$ for any density matrix $Q$. By enlarging $C$ sufficiently the lemma is proved. □

**6.2. Genericity assumptions.** We will impose some assumptions that will ensure that $\mathfrak{F}$ has finitely many fixed points and that at each fixed point $P_f$, $A + \underline{B}[P_f]$ has a spectral gap, so that $\mathfrak{F}(P_f)$ can be defined canonically. We will argue that these assumptions hold generically, i.e., can be made to hold by an arbitrarily small perturbation of the eigenvalue problem (1.1). Our genericity assumptions will allow us (1) to prove that $P^{(k)}$ converges to a fixed point and (2) to perform a first-order analysis of $P_f$ near each fixed point.

GENERICITY ASSUMPTION 1. *Assume that $A + B$ has distinct eigenvalues $\lambda_1 < \cdots < \lambda_N$ corresponding to orthonormal eigenvectors $v_1, \ldots, v_N$.*

This can be guaranteed by replacing $A$ or $B$ with a suitable arbitrarily small random perturbation of $A$ or $B$ (see, e.g., Section 1.3 of [35]).

GENERICITY ASSUMPTION 2. *For $\tau : \{1, \ldots, n\} \to \{1, \ldots, N\}$ increasing, let $P_\tau = \sum_{i=1}^n v_{\tau(i)} v_{\tau(i)}^*$, and let $S_\tau = \mathrm{Im}(P_\tau)$. Assume that for all such $\tau$,*

$$\lambda_{\tau(n)} \neq \lambda_{\min}\left\{ \left(A + \underline{B}[P_\tau]\right)\big|_{S_\tau^\perp} \right\},$$

*or, equivalently,*

$$\lambda_{\max}\left\{ \left(A + \underline{B}[P_\tau]\right)\big|_{S_\tau} \right\} \neq \lambda_{\min}\left\{ \left(A + \underline{B}[P_\tau]\right)\big|_{S_\tau^\perp} \right\}.$$

We now provide some interpretation for Genericity Assumption 2. If

$$\lambda_{\tau(n)} < \lambda_{\min}\left\{ \left(A + \underline{B}[P_\tau]\right)\big|_{S_\tau^\perp} \right\},$$

then $P_\tau$ is a fixed point of $\mathfrak{F}$. Moreover, $A + \underline{B}[P_\tau]$ has a positive spectral gap, so $\mathfrak{F}(P_\tau)$ is canonically defined. Meanwhile, if

$$\lambda_{\tau(n)} > \lambda_{\min}\left\{ \left(A + \underline{B}[P_\tau]\right)\big|_{S_\tau^\perp} \right\},$$

then $P_\tau$ is *definitely not* a fixed point of $\mathfrak{F}$ (though it does not necessarily follow that $A + \underline{B}[P_\tau]$ has a positive spectral gap). Lastly, if

$$\lambda_{\tau(n)} = \lambda_{\min}\left\{ \left(A + \underline{B}[P_\tau]\right)\big|_{S_\tau^\perp} \right\},$$

then $A + \underline{B}[P_\tau]$ has zero spectral gap, and $P_\tau$ *may or may not* be a fixed point, depending on the arbitrary choice made for $\mathfrak{F}(P_\tau)$. This is precisely the scenario that Genericity Assumption 2 rules out.

We will argue that Genericity Assumption 2 can be guaranteed by replacing (if necessary) $B$ with $B - t$ for all but finitely many $t \geq 0$. Note that this does not change the eigenspaces of $A + B$ and only affects the eigenvalues by shifting them all downward by $t$. We first provide a characterization of fixed points of $\mathfrak{F}$.

LEMMA 6.7 (Characterization of fixed points). *Suppose that $P_f$ is a fixed point of $\mathfrak{F}$. Then we can write*

$$A + \underline{B}[P_f] = \sum_{i=1}^N \mu_i z_i z_i^*, \tag{6.5}$$

*where $z_1, \ldots, z_N$ are orthonormal eigenvectors of $A + \underline{B}[P_f]$ with corresponding eigenvalues $\mu_1 \leq \cdots \leq \mu_N$. Moreover $z_1, \ldots, z_n$ are eigenvectors of $A + B$ forming an*

*orthonormal basis of* $\mathrm{Im}(P_f)$. *Consequently* $\mu_i = \lambda_{\tau(i)}$, *where* $\tau : \{1,\dots,n\} \to \{1,\dots,N\}$ *is increasing, and* $P_f = P$ *if and only if* $\mu_i = \lambda_i$ *for* $i = 1,\dots,n$. *Otherwise* $\mu_n \geq \lambda_n + \lambda_g$.

*Proof.* Let $P_f$ be a fixed point of $\mathfrak{F}$. Referring to Definition 5.1, we see that then $A + \underline{B}[P_f]$ maps $\mathrm{Im}(P_f)$ into itself. But $A + \underline{B}[Q] \equiv A + B$ on $\mathrm{Im}(P_f)$, so $A + B$ maps $\mathrm{Im}(P_f)$ into itself. $A + B$ can then be considered (via restriction) as a self-adjoint operator $\mathrm{Im}(P_f) \to \mathrm{Im}(P_f)$, so $\mathrm{Im}(P_f)$ has an orthonormal basis of eigenvectors $z_1,\dots,z_n$ of $A + B$ with corresponding eigenvalues $\mu_1 \leq \dots \leq \mu_n$.

Since $A + \underline{B}[P_f]$ is self-adjoint, we also have that $A + \underline{B}[P_f]$ maps $\mathrm{Im}(P_f)^\perp$ into itself, so $\mathrm{Im}(P_f)^\perp$ has an orthonormal basis $z_{n+1},\dots,z_N$ of eigenvectors of $A + \underline{B}[P_f]$ with corresponding eigenvalues $\mu_{n+1} \leq \dots \leq \mu_N$. The decomposition of (6.5) follows, provided we can show that $\mu_n \leq \mu_{n+1}$.

We will establish this now. First observe the general fact that for any density matrix $Q$, if $u \in \mathrm{Im}(\mathfrak{F}(Q))$ is a unit vector and

$$u^*\big(A + \underline{B}[Q]\big)u > \hat{u}^*\big(A + \underline{B}[Q]\big)\hat{u}$$

for some unit vector $\hat{u}$, then $\hat{u} \in \mathrm{Im}(\mathfrak{F}(Q))$ as well. Now suppose for contradiction that $\mu_n > \mu_{n+1}$. Then considering $z_n$, $z_{n+1}$, and $P_f$ in the places of $u$, $\hat{u}$, and $Q$, respectively, we conclude that $z_{n+1} \in \mathrm{Im}(\mathfrak{F}(P_f))$. But since $P_f$ is a fixed point of $\mathfrak{F}$, this means that $z_{n+1} \in \mathrm{Im}(P_f)$, which is impossible since $0 \neq z_{n+1} \in \mathrm{Im}(P_f)^\perp$.

Now if $\mu_i \neq \lambda_i$ for some $i \in \{1,\dots,n\}$, we must have $\mu_n = \lambda_m$ for some $m > n$, so $\mu_n \geq \lambda_{n+1} = \lambda_n + \lambda_g$. In this case, we cannot have $P_f = P$, for if this were true then $\mathrm{Im}(P)$ would contain an eigenvector of $A + B$ with eigenvalue greater than $\lambda_n$.

Lastly, suppose that $\mu_i = \lambda_i$ for $i = 1,\dots,n$. Then $(A + B)|_{\mathrm{Im}(P_f)} \preceq \lambda_n$. Since $A + B$ has a spectral gap, we must have that $\mathrm{Im}(P_f) = \mathrm{Im}(P)$, i.e., $P_f = P$. ∎

Recall that we would like to establish that Genericity Assumption 2 holds generically by replacing $B$ with $B - t$.

By Genericity Assumption 1, $A + B$ has only finitely many distinct eigenvectors (up to scaling). Then by Lemma 6.7, $\mathfrak{F}$ can only have finitely many fixed points. More precisely, this is the case because by Lemma 6.7 the candidates for fixed points are limited to projectors of the form $P_\tau = \sum_{i=1}^n v_{\tau(i)} v_{\tau(i)}^*$, where $\tau : \{1,\dots,n\} \to \{1,\dots,N\}$ is increasing.

For such $\tau$, note that $S_\tau = \mathrm{span}\{v_{\tau(1)},\dots,v_{\tau(n)}\}$ is an invariant subspace for $A + \underline{B}[P_\tau]$, and hence so is $S_\tau^\perp$. Let

$$\mu_\tau := \lambda_{\min}\left\{\big(A + \underline{B}[P_\tau]\big)\big|_{S_\tau^\perp}\right\}.$$

If $\mu_\tau < \lambda_{\tau(n)}$, then by Lemma 6.7, $P_\tau$ is not a fixed point. If $\mu_\tau > \lambda_{\tau(n)}$, then evidently $P_\tau$ is a fixed point. If $\mu_\tau > \lambda_{\tau(n)}$, then $P_\tau$ may or may not be a fixed point, since the spectral gap of $A + \underline{B}[P_\tau]$ is zero and the choice of $\mathfrak{F}(P_\tau)$ is not canonical. This last event is precisely what we would like to rule out.

More precisely, we would like to guarantee that for all of the (finitely many) increasing functions $\tau : \{1,\dots,n\} \to \{1,\dots,N\}$, we have that $\mu_\tau \neq \lambda_{\tau(n)}$.

Define $B_t := B - t$ for $t \geq 0$, and consider replacing $B$ with $B_t$ in the eigenvalue problem (1.1). Accordingly, define $\lambda_i(t)$ and $\mu_\tau(t)$ now as functions of $t \geq 0$. Evidently $\lambda_i(t) = \lambda_i - t$.

We would like to get a handle on $\mu_\tau(t)$. Extend $\tau$ to a permutation on all of $\{1,\dots,N\}$ (so $v_{\tau(n+1)},\dots,v_{\tau(N)}$ forms a basis for $S_\tau^\perp$), and recall from (4.7) that we

can write

$$\left[\underline{B_t}[P_\tau]\right]_{V_\tau} = \begin{pmatrix} B_{11} - t & B_{12} \\ B_{12}^* & B_{12}^*(B_{11} - t)^{-1} B_{12} \end{pmatrix},$$

for suitable blocks $B_{ij}$, where $V_\tau := [v_{\tau(1)}, \ldots, v_{\tau(N)}]$. Since $B_{11}$ is negative definite, we have that

$$B_{12}^*(B_{11} - t)^{-1} B_{12} \succeq B_{12}^* B_{11}^{-1} B_{12}$$

for all $t \geq 0$. It follows that $\mu_\tau(t) \geq \mu_\tau$.

Thus for every $\tau$, $f_\tau(t) := \mu_\tau(t) - \lambda_{\tau(n)}(t)$ is a strictly increasing function on $t \geq 0$, so $f_\tau$ can have at most one zero. Since there are only finitely many $\tau$ of interest, there can only be finitely many points at which $\mu_\tau(t) = \lambda_{\tau(n)}(t)$ for some $\tau$. This means that by replacing $B$ with $B - t$ for any $t \geq 0$ outside of a finite set, Genericity Assumption 2 holds.

REMARK 6.8. *In summary, Genericity Assumptions 1 and 2 can be made to hold by perturbing $A + B$ to have distinct eigenvalues, then in turn replacing $B$ with $B - t$ for any $t \geq 0$ outside of a finite set (the latter step yielding an equivalent eigenproblem). We have shown in particular that these assumptions imply that $\mathfrak{F}$ has only finitely many fixed points and that, for any fixed point $P_f$ of $\mathfrak{F}$, $A + \underline{B}[P_f]$ has a spectral gap. We keep these assumptions for the remainder of Section 6.*

In particular—recalling that $\mathbf{H}_N$ and $\mathbf{S}_N$ denote the sets of $N \times N$ Hermitian and $N \times N$ real-symmetric matrices, respectively—we have the following:

LEMMA 6.9. *Fix any $A \in \mathbf{H}_N$. Then Genericity Assumptions 1 and 2 hold both (1) for all $B \in \mathbf{H}_N$ outside of a set of zero measure with respect to the Lebesgue measure on $\mathbf{H}_N$ and (2) for all $B \in \mathbf{S}_N$ outside of a set of zero measure with respect to the Lebesgue measure on $\mathbf{S}_N$.*

REMARK 6.10. *Note that statement (1) does not imply statement (2). It is desirable to have both of these statements at our disposal for the following reason. If we are solving an eigenvalue problem where $B$ is real-symmetric, we would like to be able to guarantee that a small random* real-symmetric *perturbation of $B$ will satisfy the Genericity Assumptions. With only the first statement, we could only guarantee that this would work for a random* Hermitian *perturbation, which would almost surely introduce imaginary parts to all the entries of $B$. This would not be desirable from a computational perspective.*

*Proof.* To see that the statements (1) and (2) hold for Genericity Assumption 1 alone, refer to Section 1.3 of [35]. Now the set $\mathrm{Sc} := \{t \cdot I_N : t \in \mathbb{R}\}$ of scalar matrices is a one-dimensional subspace of the both of the real vector spaces $\mathbf{H}_N$ and $\mathbf{S}_N$. We have already argued in the preceding discussion that for any $X \in \mathbf{H}_N$ (hence also for any $X \in \mathbf{S}_N$), Genericity Assumption 2 holds for a.e. choice of $B$ in the one-dimensional space $X + \mathrm{Sc}$ (with respect to the one-dimensional Lebesgue measure). By Fubini's theorem (considering the product decompositions $\mathbf{H}_N = \mathrm{Sc} + \mathrm{Sc}^\perp$ and $\mathbf{S}_N = \mathrm{Sc} + \mathrm{Sc}^\perp$, where the orthogonal complements are taken within $\mathbf{H}_N$ and $\mathbf{S}_N$, respectively), Genericity Assumption 2 holds for a.e. choice of $B$ in $\mathbf{H}_N$ with respect to the Lebesgue measure on $\mathbf{H}_N$ and a.e. choice of $B$ in $\mathbf{S}_N$ with respect to the Lebesgue measure on $\mathbf{S}_N$. ∎

COROLLARY 6.11. *Genericity Assumptions 1 and 2 hold for almost every pair $(A, B)$ in $\mathbf{H}_N \times \mathbf{H}_N$ (with respect to the Lebesgue measure on $\mathbf{H}_N \times \mathbf{H}_N$) and for almost every pair $(A, B)$ in $\mathbf{S}_N \times \mathbf{S}_N$ (with respect to the Lebesgue measure on $\mathbf{S}_N \times \mathbf{S}_N$).*

*Proof.* This follows from Lemma 6.9 and Fubini's theorem. ∎

**6.3. Global convergence to a fixed point, local convergence revisited.** We are now ready to prove that the adaptive compression method converges globally to a fixed point (though we do not yet address whether the fixed point is the true density matrix $P$).

PROPOSITION 6.12. *$P^{(k)} \to P_f$ as $k \to \infty$ for some fixed point $P_f$ of $\mathfrak{F}$.*

*Proof.* As above let $v_1, \dots, v_N$ be an orthonormal basis of eigenvectors of $A + B$ with corresponding eigenvalues $\lambda_1 \le \cdots \le \lambda_N$. Let $\mathcal{T}$ be the set of all density matrices $P_\tau := \sum_{i=1}^n v_{\tau(i)} v_{\tau(i)}^*$ where $\tau : \{1, \dots, n\} \to \{1, \dots, N\}$ is increasing. Then Lemma 6.4 and Lemma 6.6 together imply that $\operatorname{dist}\left(P^{(k)}, \mathcal{T}\right) \to 0$ as $k \to \infty$. However, since (by Lemma 6.4) $\|P^{(k)} - P^{(k-1)}\| \to 0$, and since $\mathcal{T}$ consists of only finitely many (hence mutually isolated) points, it must be the case that $P^{(k)} \to P_\tau$ for some $\tau$. Below we show that $P_\tau$ must also be a fixed point of $\mathfrak{F}$.

Observe that, for all $k$,

$$\lambda_{\max}\left\{\left(A + \underline{B}[P^{(k-1)}]\right)\big|_{\operatorname{Im}(P^{(k)})}\right\} \le \lambda_{\min}\left\{\left(A + \underline{B}[P^{(k-1)}]\right)\big|_{\operatorname{Im}(P^{(k)})^\perp}\right\}. \tag{6.6}$$

We will rewrite this inequality in a way that makes it clear that we can take a limit as $k \to \infty$. To this end, let $C \gg \|A\|_2 + \|B\|_2$, noting that $\|A\|_2 + \|B\|_2$ provides an upper bound on $\|A + \underline{B}[Q]\|_2$ for all density matrices $Q$, hence also an upper bound on the absolute value of the eigenvalues of $A + \underline{B}[Q]$. Then (6.6) is the same as

$$\begin{aligned}\lambda_{\max}&\left\{P^{(k)}\left(A + \underline{B}[P^{(k-1)}]\right)P^{(k)} - C \cdot \left(I - P^{(k)}\right)\right\} \\ &\le \lambda_{\min}\left\{\left(I - P^{(k)}\right)\left(A + \underline{B}[P^{(k-1)}]\right)\left(I - P^{(k)}\right) + C \cdot P^{(k)}\right\}.\end{aligned}$$

Then by continuity and the convergence $P^{(k)} \to P_\tau$ we have

$$\begin{aligned}\lambda_{\max}&\left\{P_\tau\left(A + \underline{B}[P_\tau]\right)P_\tau - C \cdot \left(I - P_\tau\right)\right\} \\ &\le \lambda_{\min}\left\{\left(I - P_\tau\right)\left(A + \underline{B}[P_\tau]\right)\left(I - P_\tau\right) + C \cdot P_\tau\right\},\end{aligned}$$

i.e.,

$$\lambda_{\tau(n)} = \lambda_{\max}\left\{\left(A + \underline{B}[P_\tau]\right)\big|_{\operatorname{Im}(P_\tau)}\right\} \le \lambda_{\min}\left\{\left(A + \underline{B}[P_\tau]\right)\big|_{\operatorname{Im}(P_\tau)^\perp}\right\}.$$

We have successfully passed (6.6) to the limit as $k \to \infty$.

If $P_\tau$ is not a fixed point, then Genericity Assumption 2 implies that

$$\lambda_{\tau(n)} > \lambda_{\min}\left\{\left(A + \underline{B}[P_\tau]\right)\big|_{\operatorname{Im}(P_\tau)^\perp}\right\},$$

yielding a contradiction. □

Next we see how the preceding results imply local convergence. Though we have already provided a more refined local convergence result (complete with a linear rate of convergence), it is noteworthy that local convergence can be proved "non-perturbatively". For this proof, we will *not* consider a linearization of $\mathfrak{F}$ about $P$, instead relying only on eigenvalue monotonicity as the tool.

PROPOSITION 6.13 (Local convergence via eigenvalue monotonicity). *If $P^{(0)}$ is sufficiently close (in the sense of any given norm on $\mathbb{C}^{N \times N}$) to the true density matrix $P$, then $P = \lim_{k \to \infty} P^{(k)}$.*

*Proof.* By the continuity of $\lambda_i\{\,\cdot\,\}$ and $\underline{B}$, for all choices of $P^{(0)}$ sufficiently close to $P$ we have that

$$\lambda_i\big\{A + \underline{B}[P^{(0)}]\big\} \leq \lambda_i + \lambda_g/2,$$

for all $i = 1, \ldots, n$. Then for such $P^{(0)}$, eigenvalue monotonicity (Lemma 6.1) implies that

$$\lambda_i\big\{A + \underline{B}[P^{(k)}]\big\} \leq \lambda_i + \lambda_g/2,$$

for all $k$, $i = 1, \ldots, n$.

At a fixed point $P_f$, the bottom $n$ eigenvalues of $A+\underline{B}[P_f]$ are eigenvalues of $A+B$. If $P_f \neq P$, then one of these eigenvalues must be at least as large as $\lambda_{n+1} = \lambda_n + \lambda_g$.

But Proposition 6.12 says that $P^{(k)} \to P_f$ for some fixed point $P_f$. By continuity, this is impossible if $P^{(0)}$ is taken as above, unless $P_f = P$. □

**6.4. Linearization around fixed points.** In this section we repeat the first-order analysis of Section 5.2 about an arbitrary fixed point of $\mathfrak{F}$. We will see that all fixed points except the true density matrix are repulsive in a certain sense.

Let $P_f$ be a fixed point of $\mathfrak{F}$. Then we can write $A + B = \sum_{i=1}^{N} \mu_i u_i u_i^*$, where $u_i$ are orthonormal eigenvectors of $A+B$ with eigenvalues $\mu_i$, and $P_f = \sum_{i=1}^{n} u_i u_i^*$. The picture is almost exactly the same as in Section 5.2, with the important difference that $\{\mu_1, \ldots, \mu_n\}$ might *not* be the same as $\{\lambda_1, \ldots, \lambda_n\}$. Though $\mu_1, \ldots, \mu_n$ may not be the bottom eigenvalues of $A + B$, they *are* the bottom eigenvalues of $A + \underline{B}[P_f]$, and in fact our genericity assumptions have guaranteed that $A + \underline{B}[P_f]$ enjoys a spectral gap. Thus $\mathfrak{F}$ is smooth near $P_f$, and the same reasoning that yielded 5.5 also yields the following.

LEMMA 6.14. *With notation as in the preceding discussion, for $\epsilon > 0$ sufficiently small, let $P_f(\epsilon)$ be density matrices with $\Delta P = \lim_{\epsilon\to 0}(P_f(\epsilon) - P_f)/\epsilon$. Then*

$$\begin{aligned} D\mathfrak{F}_{P_f}[\Delta P] &:= \lim_{\epsilon\to 0} \frac{\mathfrak{F}(P_f(\epsilon)) - \mathfrak{F}(P_f)}{\epsilon} \\ &= \sum_{i=1}^{n} \left( P_f^{\perp} + (\underline{B}[P_f] - B)^{\dagger} (A + B - \mu_i) P_f^{\perp} \right)^{\dagger} (\Delta P) u_i u_i^* + \text{h.c.} \end{aligned}$$

REMARK 6.15. *Write $B$ in the $U_N := [u_1, \ldots, u_N]$ basis as*

$$[B]_{U_N} = \begin{pmatrix} B_{11}^{P_f} & B_{12}^{P_f} \\ \left(B_{11}^{P_f}\right)^* & B_{22}^{P_f} \end{pmatrix}.$$

*Then in this basis, the matrix of the linear transformation*

$$Z_i^{P_f} := \left( P_f^{\perp} + (\underline{B}[P_f] - B)^{\dagger} (A + B - \mu_i) P_f^{\perp} \right)^{\dagger}$$

*appearing in Lemma 6.14 is given by*

$$\left[Z_i^{P_f}\right]_{U_N} = \begin{pmatrix} 0 & 0 \\ 0 & J_i^{P_f} \end{pmatrix},$$

*where*

$$J_i^{P_f} := \left[ I_{N-n} + \left( -S_{22}^{P_f} \right)^{-1} (M_2 - \mu_i) \right]^{-1} \tag{6.7}$$

*and $S_{22}^{P_f}$ is shorthand for the Schur complement and is negative definite, and $M_2 := \mathrm{diag}(\mu_{n+1}, \ldots, \mu_N)$. Note that if $P_f \neq P$, then for some $i \in \{1, \ldots, n\}$, $M_2 - \mu_i$ is diagonal with a strictly negative entry.*

We have, in contrast with Lemma 5.7:

LEMMA 6.16. *For $i = 1, \ldots, n$, $J_i^{P_f}$ is diagonalizable with $\sigma\big(J_i^{P_f}\big) \subset (0, \infty)\backslash\{1\}$. Moreover, if $P_f \neq P$, then $\lambda_{\max}\big(J_i^{P_f}\big) > 1$ for some $i \in \{1, \ldots, n\}$.*

*Proof.* In the proof we adopt notation from Remark 6.15. Note that $\big(-S_{22}^{P_f}\big) = (\underline{B}[P_f] - B)_{22}$ and we can alternatively write

$$Z_i^{P_f} = \big[(A + \underline{B}[P_f])_{22} - \mu_i)\big]^{-1} (\underline{B}[P_f] - B)_{22}, \tag{6.8}$$

where $(\,\cdot\,)_{22}$ denotes the lower-right block in the $u_i$ basis. As the product of two positive definite matrices, $Z_i^{P_f}$ has positive eigenvalues. This can be verified by conjugating by $(\underline{B}[P_f] - B)_{22}^{1/2}$.

By (6.7), we have the set equality

$$\begin{aligned}\sigma\left(Z_i^{P_f}\right) &= \frac{1}{1 + \sigma\Big[(\underline{B}[P_f] - B)_{22}^{-1}(M_2 - \mu_i)\Big]} \\ &= \frac{1}{1 + \sigma\Big[(\underline{B}[P_f] - B)_{22}^{-1/2}(M_2 - \mu_i)(\underline{B}[P_f] - B)_{22}^{-1/2}\Big]}.\end{aligned} \tag{6.9}$$

Now the signs of the eigenvalues of

$$(\underline{B}[P_f] - B)_{22}^{-1/2}(M_2 - \mu_i)(\underline{B}[P_f] - B)_{22}^{-1/2}$$

are the same as those of $M_2 - \mu_i$. Since we have assumed (Genericity Assumption 1) that the eigenvalues of $A+B$ are distinct, $M_2 - \mu_i$ is diagonal with nonzero eigenvalues. Thus $1 \notin \sigma\big(Z_i^{P_f}\big)$. This establishes the first statement of the lemma.

Now assume that $P_f \neq P$, and choose $i \in \{1, \ldots, n\}$ such that $M_2 - \mu_i$ has a strictly negative entry. By (6.9), $Z_i^{P_f}$ must then have an eigenvalue that is either negative or larger than 1, but we have already established that $\sigma(Z_i^{P_f}) \subset (0, \infty)$, so the latter possibility must be true. ∎

Let us identify the tangent space $T_{P_f}\mathcal{D}$ of the space of density matrices at $P_f$ with $\mathbb{C}^{(N-n)n}$ as we did in Section 5.2 at the fixed point $P$. Furthermore, let us identify $\mathfrak{F}$ with a map $\mathfrak{H}^{P_f}$ into $\mathbb{C}^{(N-n)n}$ defined on a neighborhood $\mathcal{U}$ of the origin in $\mathbb{C}^{(N-n)n}$ via a local diffeomorphism $\Phi : \mathcal{U} \to \mathcal{D}$, defined as in (5.6) but with the $u_i$ now in the places of the $v_i$. Then like before we have that

$$D\mathfrak{H}_0^{P_f} = \begin{pmatrix} J_1^{P_f} & 0 & \cdots & 0 \\ 0 & J_2^{P_f} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_n^{P_f} \end{pmatrix}.$$

By Lemma 6.16, $D\mathfrak{H}_0^{P_f}$ is invertible and has no eigenvalues of modulus 1. Thus in the language of dynamical systems, 0 is a hyperbolic fixed point of the dynamical system defined locally by $\mathfrak{H}$ near the origin of $\mathbb{C}^{(N-n)n}$.

Assume that $P_f \neq P$, then $D\mathfrak{H}_0^{P_f}$ has at least one eigenvalue larger than 1. If we identify $\mathbb{C}^{(N-n)n}$ with $\mathbb{R}^{2(N-n)n}$, then the corresponding realification of $D\mathfrak{H}_0^{P_f}$ (an

operator $\mathbb{R}^{2(N-n)n} \to \mathbb{R}^{2(N-n)n}$) has the same eigenvalues (though now two copies of each), hence two eigenvalues larger than 1. By the stable manifold theorem, the local stable manifold near the origin has real codimension at least 2, and there exists a neighborhood $\mathcal{V} \subset \mathcal{U}$ of the origin such that if $X \in \mathcal{V}$ is not in the local stable manifold, then $\left(\mathfrak{H}^{P_f}\right)^m (X) \notin \mathcal{V}$ for some $m \geq 1$ (refer to Theorems 10.6 and 10.7 of [36]). In particular, this implies the following.

PROPOSITION 6.17. *For any fixed point $P_f \neq P$, there is a neighborhood $\mathcal{P}_f$ of $P_f$ in the space $\mathcal{D}$ of density matrices and a subset $\widetilde{\mathcal{P}}_f \subset \mathcal{P}_f$ such that $\mathcal{P}_f \backslash \widetilde{\mathcal{P}}_f$ has measure zero in $\mathcal{D}$. If $P^{(k)} \in \widetilde{\mathcal{P}}_f$ for some $k$, then $P^{(m)} \notin \mathcal{P}_f$ for some $m > k$.*

REMARK 6.18. *Note that a notion of measure zero can be defined on any smooth manifold without actually choosing a measure or a Riemannian structure. One way to produce such a notion is to pick any Riemannian metric on $\mathcal{D}$ and consider the corresponding volume measure. The measure zero sets with respect to this volume measure will be the same regardless of the choice of metric.*

Proposition 6.17 is roughly saying that generically near a fixed point $P_f \neq P$, points are repelled from $P_f$. However, in the important special case that we are considering $A, B$ to be real-symmetric, note that $\mathfrak{F}$ can be interpreted as a map from $\mathcal{D}_{\mathbb{R}}$ into itself. If we initialize with a real-symmetric guess, then we never leave the submanifold of real-symmetric projectors. The notion of full measure does not project from $\mathcal{D}$ to the submanifold $\mathcal{D}_{\mathbb{R}}$ of *real-symmetric* orthogonal projectors, so Proposition 6.17 *does not* imply that generically (within $\mathcal{D}_{\mathbb{R}}$) points are repelled from $P_f$, and we must state this result separately.

PROPOSITION 6.19. *Suppose that $A$ and $B$ are real-symmetric, so $\mathfrak{F}$ maps $\mathcal{D}_{\mathbb{R}}$ into itself. For any fixed point $P_f \neq P$ in $\mathcal{D}_{\mathbb{R}}$, there is a neighborhood $\mathcal{P}_f$ of $P_f$ in the space $\mathcal{D}_{\mathbb{R}}$ a subset $\widetilde{\mathcal{P}}_f \subset \mathcal{P}_f$ such that $\mathcal{P}_f \backslash \widetilde{\mathcal{P}}_f$ has measure zero in $\mathcal{D}_{\mathbb{R}}$. If $P^{(k)} \in \widetilde{\mathcal{P}}_f$ for some $k$, then $P^{(m)} \notin \mathcal{P}_f$ for some $m > k$.*

*Proof.* By exactly the same proof is above (with $\mathbb{R}$ in place of $\mathbb{C}$), the restriction of the dynamical system defined by $\mathfrak{F}$ to the submanifold $\mathcal{D}_{\mathbb{R}}$ has a hyperbolic fixed point at $P_f$ with invertible Jacobian and the same eigenvalues as before (though only one copy of each now, instead of two). In particular, the stable manifold within $\mathcal{D}_{\mathbb{R}}$ has real codimension at least one, and by the same reasoning as before, this implies the statement. □

Propositions 6.17 and 6.19 formalize the notion that any fixed point $P_f \neq P$ is repulsive. More quantitatively speaking, based on Lemma 6.16 we expect to see "linear divergence" from any fixed point $P_f \neq P$ with rate $\max_{i=1,\ldots,n} \lambda_{\max}\left(Z_i^{P_f}\right) > 1$, but we do not formalize this notion.

**6.5. Fixed points are saddle points.** Consider the functional

$$F(Q) = \operatorname{Tr}\left[\mathfrak{F}(Q)\left(A + \underline{B}[Q]\right)\mathfrak{F}(Q)\right] = \sum_{i=1}^{n} \lambda_i \left\{A + \underline{B}[Q]\right\}. \tag{6.10}$$

By eigenvalue monotonicity, $F(P^{(k)})$ is non-increasing in $k$. We claim that fixed points of $\mathfrak{F}$ are critical points of $F$ and that a fixed point $P_f \neq P$ is not a local minimum. In fact, a fixed point $P_f \neq P$ is a strict saddle point of $F$ in that it is a strict local maximum of $F$ along some direction.

We state a more detailed version this fact formally in Proposition 6.20 below. This result informs our understanding of the behavior of the iteration near fixed points (see the discussion at the beginning of Section 6.6). However, we will not use it directly

to establish global convergence, and its proof is largely computational, so we relegate this proof to an appendix.

PROPOSITION 6.20. *Let $Q = Q(t)$ be a twice-differentiable density matrix-valued function of a single variable with $Q(0) = P_f$ a fixed point. Then $(F(Q))'(0) = 0$. Moreover, if $P_f \neq P$, then there exists such a function $Q(t)$ which additionally satisfies $(F(Q))''(0) < 0$. In fact, if we take $Q(t) = \Phi(tX)$, where $X \in \mathbb{C}^{(N-n)n}$ is an eigenvector of $D\mathfrak{H}_0^{P_f}$ with eigenvalue larger than [resp., smaller than] 1, then $(F(Q))''(0) < 0$ [resp., $> 0$].*

*Proof.* See Appendix C. □

**6.6. Global convergence.** We already have a fairly complete picture of the global behavior of Algorithm 1. In summary, we know that the fixed point iteration converges to a fixed point, and we know that fixed points $P_f \neq P$ are repulsive in the sense of Proposition 6.17. We also know that such fixed points are strict saddle points of the functional $F$ in Eq. (6.10). Moreover, along the repulsive directions at $P_f$ this functional has a strict local maximum. With a bit more work, it is possible to show that for almost all $Q$ in a sufficiently small neighborhood of $P_f$, there exists $m = m(Q)$ such that $F(\mathfrak{F}^m(Q)) < F(P_f)$. (We already know that generically such $Q$ must escape the neighborhood, but when they do so, they should align with the repulsive directions, so the value of $F$ must fall below $F(P_f)$. We omit a formal proof of this fact.) Thus by eigenvalue monotonicity, if we have converged sufficiently close to a fixed point $P_f \neq P$, and if we apply a small random perturbation and then restart Algorithm 1 from this point, then we will converge to another fixed point $\widetilde{P}_f$ with $F(\widetilde{P}_f) < F(P_f)$. Repeating this process finitely many times will bring us to the desired fixed point $P$. This suggests a satisfactory notion of the global convergence up to perturbation.

Nonetheless, it is still desirable to show that for almost every choice of initialization $P^{(0)} \in \mathcal{D}$, Algorithm 1 converges to $P$. (Similarly in the special case of real-symmetric $A$ and $B$, Algorithm 1 converges to $P$ for a.e. choice $P^{(0)} \in \mathcal{D}_{\mathbb{R}}$.) To use an analogy, a fixed point $P_f \neq P$ is like an egg resting on top of a barn. We know that if we apply a slight random perturbation to the egg, it will fall off the barn and never return to the top. But we would like to show that it is impossible for the egg to get stuck on top of the barn in the first place!

This is indeed true, and the key lemma is the following.

LEMMA 6.21 (Egg on barn lemma). *If we fix $B \in \mathbf{H}_N$, then for almost any $A \in \mathbf{H}_N$ (with respect to the Lebesgue measure on $\mathbf{H}_N$), we have the following: if $S$ has zero measure in the space $\mathcal{D}$ of density matrices, then $\mathfrak{F}^{-1}(S)$ also has zero measure in $\mathcal{D}$, where $\mathfrak{F}$ is considered as a map $\mathcal{D} \to \mathcal{D}$.*

*Similarly, if we fix any $B \in \mathbf{S}_N$, then for almost any $A \in \mathbf{S}_N$ (with respect to the Lebesgue measure on $\mathbf{S}_N$), we have the following: if $S$ has zero measure in the space $\mathcal{D}_{\mathbb{R}}$ of density matrices, then $\mathfrak{F}^{-1}(S)$ also has zero measure in $\mathcal{D}_{\mathbb{R}}$, where $\mathfrak{F}$ is considered as a map $\mathcal{D}_{\mathbb{R}} \to \mathcal{D}_{\mathbb{R}}$.*

The proof of Lemma 6.21 is technical. The main difficulty is that $\mathfrak{F}$ is not a diffeomorphism, and indeed is not even continuous. However, it is real-analytic on an open, connected subset of full measure, and this characterization allows us to rule out pathological behavior. We postpone the proof of Lemma 6.21 to Appendix D. Let us now use this lemma to prove the global convergence property.

Fix $K \in \{\mathbb{C}, \mathbb{R}\}$, and assume that $B$ is such that Lemma 6.21 applies. For a fixed point $P_f$, let $S_{P_f} = \{Q \in \mathcal{D}_K : \mathfrak{F}^k(Q) \to P_f\}$, so $\mathcal{D}_K = \bigcup_{P_f \text{ fixed}} S_{P_f}$. If $P_f \neq P$, then by Proposition 6.17 or Proposition 6.19, for any $Q \in S_{P_f}$, we must have that

$\mathfrak{F}^k(Q) \in \mathcal{P}_f \backslash \widetilde{\mathcal{P}}_f$ for some $k = k(Q) \geq 0$. This implies that $S_{P_f} \subset \bigcup_{k\geq 0} \mathfrak{F}^{-k}(\mathcal{P}_f \backslash \widetilde{\mathcal{P}}_f)$. But by Proposition 6.17, $\mathcal{P}_f \backslash \widetilde{\mathcal{P}}_f$ has measure zero. By Lemma 6.21 (and induction), $\mathfrak{F}^{-k}(\mathcal{P}_f \backslash \widetilde{\mathcal{P}}_f)$ has measure zero for all $k \geq 0$. Consequently, $S_{P_f}$ has measure zero for all $P_f \neq P$. Hence $S_P$ has full measure, as desired. This completes the proof of Theorem 1.3.

**Acknowledgment.** The work of L. L. is partially supported by the National Science Foundation under grant DMS-1652330, the Alfred P. Sloan fellowship, and the DOE Center for Applied Mathematics for Energy Research Applications (CAMERA) program. The work of M.L. is partially supported by the National Science Foundation Graduate Research Fellowship Program under grant DGE-1106400.

**Appendix A. Derivative calculations in linearization.**

*Proof of Lemma 5.3.* The assumption of the positive spectral gap guarantees the existence of a simple contour $\mathcal{C}$ in the complex plane surrounding only the lowest $n$ eigenvalues of $H$. Using the contour integral representation of the density matrix $P$, we have

$$P = \frac{1}{2\pi\imath} \oint_{\mathcal{C}} (z - H)^{-1} \, \mathrm{d}z. \tag{A.1}$$

Assume that $\epsilon$ is small enough so that the contour $\mathcal{C}$ only surrounds the lowest $n$ eigenvalues of $H + \epsilon \Delta H$ as well. Then

$$P_\epsilon = \frac{1}{2\pi\imath} \oint_{\mathcal{C}} (z - H - \epsilon\Delta H)^{-1} \, \mathrm{d}z.$$

Since

$$(z - H - \epsilon\Delta H)^{-1} = (z - H)^{-1} + \epsilon (z - H)^{-1} \Delta H (z - H)^{-1} + O(\epsilon^2),$$

we have

$$\lim_{\epsilon \to 0} \frac{P_\epsilon - P}{\epsilon} = \frac{1}{2\pi\imath} \oint_{\mathcal{C}} (z - H)^{-1} \Delta H (z - H)^{-1} \, \mathrm{d}z. \tag{A.2}$$

Next, apply the spectral decomposition of $H$ to (A.2) to obtain

$$\begin{aligned} DP_H[\Delta H] =& \frac{1}{2\pi\imath} \oint_{\mathcal{C}} (z - H)^{-1} \Delta H (z - H)^{-1} \, \mathrm{d}z \\ =& \sum_{i,a=1}^{N} \frac{1}{2\pi\imath} \oint_{\mathcal{C}} (z - \mu_a)^{-1} u_a (u_a^* \Delta H u_i) u_i^* (z - \mu_i)^{-1} \, \mathrm{d}z \\ =& \sum_{i,a=1}^{N} \frac{1}{\mu_i - \mu_a} \left[ \frac{1}{2\pi\imath} \oint_{\mathcal{C}} \left( \frac{1}{z - \mu_i} - \frac{1}{z - \mu_a} \right) \mathrm{d}z \right] u_a (u_a^* \Delta H u_i) u_i^* \\ =& \sum_{i=1}^{n} \sum_{a=n+1}^{N} \frac{1}{\mu_i - \mu_a} u_a (u_a^* \Delta H u_i) u_i^* + \text{h.c.} \end{aligned} \tag{A.3}$$

In the last equation of (A.3), we have used the Cauchy integral formula. This establishes the first desired equality. For the second equality, simply collapse the inner sum over $a$. $\square$

*Proof of Lemma 5.4.* Recall that $\underline{B}[P_\epsilon] = B(P_\epsilon BP_\epsilon)^\dagger B = BG(\epsilon)B$, where $G(\epsilon) := (P_\epsilon BP_\epsilon)^\dagger$. We want to evaluate the derivative in $\epsilon$ of $G(\epsilon)$ at $\epsilon = 0$. To do so, we treat the pseudoinverse as follows. Note that we can alternatively write

$$G(\epsilon) = [P_\epsilon BP_\epsilon + \lambda(I - P_\epsilon)]^{-1} - \lambda^{-1}(I - P_\epsilon)$$

for any $\lambda > 0$. Then

$$\begin{aligned} G'(0) = &- [PBP + \lambda(I-P)]^{-1} [(\Delta P)BP + PB(\Delta P) - \lambda \Delta P] [PBP + \lambda(I-P)]^{-1} \\ &+ \lambda^{-1}(\Delta P) \\ = &- \left[(PBP)^\dagger + \lambda^{-1}(I-P)\right] [PB(\Delta P) + \text{h.c.}] \left[(PBP)^\dagger + \lambda^{-1}(I-P)\right] \\ &+ \lambda \left[(PBP)^\dagger + \lambda^{-1}(I-P)\right] [\Delta P] \left[(PBP)^\dagger + \lambda^{-1}(I-P)\right] \\ &+ \lambda^{-1}(\Delta P). \end{aligned} \tag{A.4}$$

Note that $(P_\epsilon)^2 = P_\epsilon$, and evaluating the derivative of this equality at $\epsilon = 0$ yields $P(\Delta P) + (\Delta P)P = (\Delta P)$. Then left- and right-multiplying both sides of this equality by $P$ yields $2P(\Delta P)P = P(\Delta P)P$, so

$$P(\Delta P)P = 0.$$

Observe that $(PBP)^\dagger = P(PBP)^\dagger P$, so there is significant cancellation in the second term of (A.4), which becomes

$$(\Delta P)(PBP)^\dagger + \text{h.c.}$$

Then substituting into (A.4), we obtain

$$\begin{aligned} G'(0) = &- \left[(PBP)^\dagger + \lambda^{-1}(I-P)\right] [PB(\Delta P) + \text{h.c.}] \left[(PBP)^\dagger + \lambda^{-1}(I-P)\right] \\ &+ [(\Delta P)(PBP)^\dagger + \text{h.c.}] + \lambda^{-1}(\Delta P). \end{aligned}$$

Now the preceding equality holds for *any* $\lambda > 0$. Thus taking the limit as $\lambda \to \infty$ establishes

$$\begin{aligned} G'(0) &= \left[-(PBP)^\dagger PB(\Delta P)(PBP)^\dagger + (\Delta P)(PBP)^\dagger\right] + \text{h.c.} \\ &= \left(I - (PBP)^\dagger B\right)(\Delta P)(PBP)^\dagger + \text{h.c.} \end{aligned}$$

Then

$$D\underline{B}_P[\Delta P] = BG'(0)B = (B - \underline{B}[P])(\Delta P)(PBP)^\dagger B + \text{h.c.},$$

as was to be shown. □

**Appendix B. Proof of Lemma 5.14.**

*Proof of Lemma 5.14.* To ease the notation, let $B_\delta = B_\delta(0)$ and $T = DF(0) \in \mathbb{R}^{k\times k}$. Let $Q_i$ be the orthogonal projector onto $E_i$, and define $\rho(x) = \|Q_1 x\|_2$. Throughout the proof we will use the shorthand notation $x = (x_1, x_2)$, e.g. $Q_1 x = (x_1, 0)$ and $\rho(x) = \|x_1\|_2$.

Fix $\epsilon > 0$ small enough such that $0 \prec T + \epsilon \prec 1$ and $0 \prec T|_{E_1} + \epsilon \prec \alpha$. Choose $\delta' \in (0, \delta)$ small enough so that $\|DF - T\|_2 \le \epsilon$ on $B_{\delta'}$. Then $\|DF\|_2 \le 1$ on $B_{\delta'}$, from which it follows that $F$ is non-expansive on $B_{\delta'}$, hence maps $B_{\delta'}$ into itself.

Then for the proof, we want to show that $\rho(F^k(x)) \leq \alpha^k \rho(x)$ for all $x \in B_{\delta'}$. For this it suffices to show that $\rho(F(x)) \leq \alpha\rho(x)$ for all $x \in B_{\delta'}$.

Define $F_1 : B_\delta \to \mathbb{R}^r$ and $F_2 : B_\delta \to \mathbb{R}^{p-r}$ by $F = (F_1, F_2)$. Define $T_1, T_2$ similarly. Our choice of $\delta'$ guarantees that $\|DF_1 - T_1\|_2 \leq \epsilon$ on $B_{\delta'}$. Now since $T$ is diagonal, $\|T_1\|_2 = \|Q_1 T Q_1\|_2 \leq \alpha - \epsilon$, so in fact we have $\|DF_1\|_2 \leq \alpha$ on $B_{\delta'}$.

Next observe that since $E_2$ is invariant, we have that $F_1(E_2 \cap B_\delta) = 0$. Then for $x \in B_{\delta'}$,

$$\begin{aligned}\rho(F(x)) = \|F_1(x)\|_2 &= \|F_1(x_1, x_2) - F_1(0, x_2)\|_2 \\ &= \left\| \int_0^1 DF_1(tx_1, x_2) \cdot x_1 \, \mathrm{d}t \right\|_2 \\ &\leq \int_0^1 \|DF_1(tx_1, x_2)\|_2 \|x_1\|_2 \, \mathrm{d}t \leq \alpha\rho(x),\end{aligned}$$

as desired. □

**Appendix C. Proof of Proposition 6.20.**

First we state a helpful lemma.

LEMMA C.1. *Let $Q = Q(t)$ be a differentiable density matrix-valued function of a single variable. Then (omitting dependence on $t$ from the notation)*

$$(\underline{B}[Q])' = (B - \underline{B}[Q])Q'B^{-1}\underline{B}[Q] + \text{h.c.}$$

*Proof.* The proof is just a recapitulation of the argument in Lemma 5.4. □

Now we prove Proposition 6.20.

*Proof of Proposition 6.20.* Compute (omitting dependence on $t$ from the notation):

$$\begin{aligned}(F(Q))' &= \operatorname{Tr}\left[(\mathfrak{F}(Q))' \, (A + \underline{B}[Q]) \, \mathfrak{F}(Q) + \text{h.c.}\right] + \operatorname{Tr}\left[\mathfrak{F}(Q) \, (\underline{B}[Q])' \, \mathfrak{F}(Q)\right] \\ &= 2 \cdot \operatorname{Tr}\left[(\mathfrak{F}(Q))' \, (A + \underline{B}[Q]) \, \mathfrak{F}(Q)\right] + 2 \cdot \operatorname{Tr}\left[\mathfrak{F}(Q)(B - \underline{B}[Q])Q'\mathfrak{F}(Q)\right].\end{aligned}$$

We have used Lemma C.1, together with the fact that $\underline{B}[Q]$ agrees with $B$ on $\operatorname{Im}(\mathfrak{F}(Q))$. The first term in the last expression turns out to be zero. (This is essentially the content of the Hellmann-Feynman theorem.) We verify this presently.

For $i = 1, \ldots, N$, let $u_i = u_i(t)$ be orthonormal such that $u_1, \ldots, u_n$ forms a basis for $\operatorname{Im}(Q(t))$. Then

$$\begin{aligned}\operatorname{Tr}\left[(\mathfrak{F}(Q))' \, (A + \underline{B}[Q]) \, \mathfrak{F}(Q)\right] &= \sum_{i=1}^N u_i^* (\mathfrak{F}(Q))' \, (A + \underline{B}[Q]) \, \mathfrak{F}(Q) u_i \\ &= \sum_{i=1}^n u_i^* (\mathfrak{F}(Q))' \, (A + \underline{B}[Q]) \, u_i.\end{aligned}$$

Now $\operatorname{Im}(\mathfrak{F}(Q))$ is an invariant subspace for $(A + \underline{B}[Q])$, so for $i = 1, \ldots, n$, $(A + \underline{B}[Q]) \, u_i$ is an element of $\operatorname{Im}(\mathfrak{F}(Q))$. Therefore

$$\operatorname{Tr}\left[(\mathfrak{F}(Q))' \, (A + \underline{B}[Q]) \, \mathfrak{F}(Q)\right] = \sum_{i=1}^n u_i^* \, \mathfrak{F}(Q)(\mathfrak{F}(Q))'\mathfrak{F}(Q) \, (A + \underline{B}[Q]) \, u_i.$$

Note that

$$(\mathfrak{F}(Q))' = (\mathfrak{F}(Q)\mathfrak{F}(Q))' = \mathfrak{F}(Q)(\mathfrak{F}(Q))' + (\mathfrak{F}(Q))'\mathfrak{F}(Q).$$

Multiply both sides by $\mathfrak{F}(Q)$ and rearrange the terms, we have

$$\mathfrak{F}(Q)(\mathfrak{F}(Q))'\mathfrak{F}(Q) = 0.$$

Therefore

$$\mathrm{Tr}\big[(\mathfrak{F}(Q))'\,(A + \underline{B}[Q])\,\mathfrak{F}(Q)\big] = 0$$

as claimed and

$$(F(Q))' = 2\cdot\mathrm{Tr}\big[\mathfrak{F}(Q)\,(B - \underline{B}[Q])\,Q'\,\mathfrak{F}(Q)\big]. \tag{C.1}$$

Define $\Delta P := Q'(0)$ and evaluate at $t = 0$. Note that $\mathfrak{F}(Q(0)) = P_f$, we obtain

$$(F(Q))'(0) = 2\cdot\mathrm{Tr}\big[P_f\,(B - \underline{B}[P_f])\,(\Delta P)\,P_f\big].$$

But $P_f(B - \underline{B}[P_f]) = 0$, so $(F(Q))'(0) = 0$, as desired.

Next take another derivative of (C.1) and evaluate at $t = 0$ to find

$$\begin{aligned}(F(Q))''(0) = {}& 2\cdot\mathrm{Tr}\big[D\mathfrak{F}_{P_f}[\Delta P]\,(B - \underline{B}[P_f])\,(\Delta P)\,P_f\big]\\ &+\ 2\cdot\mathrm{Tr}\big[P_f\,(B - \underline{B}[P_f])\,(\Delta P)\,D\mathfrak{F}_{P_f}[\Delta P]\big]\\ &-\ 2\cdot\mathrm{Tr}\big[P_f\,(\underline{B}[Q])'(0)\,(\Delta P)\,P_f\big] + 2\cdot\mathrm{Tr}\big[P_f\,(B - \underline{B}[P_f])\,Q''(0)\,P_f\big].\end{aligned}$$

Since $P_f(B - \underline{B}[P_f]) = 0$, the second and final terms vanish. Substituting in for $(\underline{B}[Q])'(0)$ via Lemma C.1 (and again using the facts $P_f(B - \underline{B}[P_f]) = 0$ and $B^{-1}\underline{B}[P_f]P_f = P_f$), we obtain

$$\begin{aligned}(F(Q))''(0) = {}& 2\cdot\mathrm{Tr}\big[D\mathfrak{F}_{P_f}[\Delta P]\,(B - \underline{B}[P_f])\,(\Delta P)\,P_f\big]\\ &+\ 2\cdot\mathrm{Tr}\big[P_f\,(\Delta P)\,(\underline{B}[P_f] - B)\,(\Delta P)\,P_f\big].\end{aligned}$$

For the rest of the proof, $u_i$ will always indicate $u_i(0)$. Now we substitute in for $D\mathfrak{F}_{P_f}[\Delta P]$ via Lemma 6.14. Since $(B - \underline{B}[P_f])u_i = 0$ for $i = 1,\dots,n$, only the "h.c." term survives, yielding

$$\begin{aligned}&(F(Q))''(0) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (C.2)\\ &= 2\sum_{i=1}^{n}\mathrm{Tr}\left[u_i u_i^*\,(\Delta P)\left(Z_i^{P_f}\right)^*(B - \underline{B}[P_f])\,(\Delta P)\,P_f\right]\\ &\quad +\ 2\cdot\mathrm{Tr}\big[P_f\,(\Delta P)\,(\underline{B}[P_f] - B)\,(\Delta P)\,P_f\big]\\ &= 2\sum_{i=1}^{n}u_i^*\,(\Delta P)\,(\underline{B}[P_f] - B)\,\left[P_f^{\perp}\,(A + \underline{B}[P_f] - \mu_i)\,P_f^{\perp}\right]^{\dagger}\,(B - \underline{B}[P_f])\,(\Delta P)\,u_i\\ &\quad +\ 2\sum_{i=1}^{n}u_i^*\,(\Delta P)\,(\underline{B}[P_f] - B)\,(\Delta P)\,u_i.\end{aligned}$$

Let $X$ be an eigenvector of $D\mathfrak{H}_0^{P_f}$ with corresponding eigenvalue $\sigma$, viewed as an element of $\mathbb{C}^{(N-n)\times n}$, so $X = (0,\dots,0,X_j,0,\dots,0)$, where $X_j$ is an eigenvector of $J_j^{P_f}$. Fix the path $Q(t) = \Phi(tX)$, then

$$(\Delta P)u_i = \delta_{ij}\begin{pmatrix}0\\ X_j\end{pmatrix},$$

and

$$\begin{aligned}(F(Q))''(0) = &\ -2X_j^*\,(\underline{B}[P_f]-B)_{22}\,J_j^{P_f}\,X_j \ + \ 2X_j^*\,(\underline{B}[P_f]-B)_{22}\,X_j \\ = &\ 2(1-\sigma)X_j^*\,(\underline{B}[P_f]-B)_{22}\,X_j.\end{aligned}$$

Since $(\underline{B}[P_f]-B)_{22} \succ 0$, we have $X_j^*\,(\underline{B}[P_f]-B)_{22}\,X_j > 0$, and therefore the sign of $(F(Q))''(0)$ is the sign of $1-\sigma$. The proposition is proved by recalling Lemma 6.16. □

**Appendix D. Proof of the egg on barn lemma.**

This section is devoted to the proof of Lemma 6.21, which we break into several pieces.

First we outline some notation that will allow us to treat the Hermitian and real-symmetric cases jointly. Fix $K \in \{\mathbb{C}, \mathbb{R}\}$. Let $\mathbf{K}_N$ denote $\mathbf{H}_N$ if $K = \mathbb{C}$ and $\mathbf{S}_N$ if $K = \mathbb{R}$. Let $\mathbf{E}_N \subset \mathbf{K}_N$ denote the elements of $\mathbf{K}_N$ with no repeated eigenvalues.

Fix some $B \in \mathbf{K}_N$ for the remainder of the section. We equip $\mathbf{K}_N$ with the Lebesgue measure, so statements about, e.g., 'almost every' $A$ in $\mathbf{K}_N$ should be understood with respect to this measure. Meanwhile, we equip $\mathcal{D}_K$ with the natural notion of 'measure zero' inherited from the Lebesgue measure on charts, which coincides with that of its volume measure induced by any choice Riemannian metric.

Let $\Phi : \mathcal{D}_K \to \mathbf{K}_N$ denote the map $Q \mapsto A + \underline{B}[Q]$, and let $\Psi : \mathbf{E}_N \to \mathcal{D}_K$ denote the map that sends a matrix in $\mathbf{E}_N$ to its density matrix in $\mathcal{D}_K$. (Note that the choice of density matrix is unambiguous when there are no repeated eigenvalues.) We would like to say that $\Phi^{-1}(\mathbf{E}_N)$ is a large (i.e., full-measure) subset of $\mathcal{D}_K$, so that we can define $\Psi \circ \Phi$ (which coincides with $\mathfrak{F}$) on this set. This is quite essential to the argument. Indeed, if this were not the case, then there would be a set of positive measure in $\mathcal{D}_K$ on which the behavior of $\mathfrak{F}$ was not canonically determined, much less differentiable. Fortunately, we have the following lemma, which says even more.

LEMMA D.1. *For almost every choice of $A$ in $\mathbf{K}_N$, $\mathcal{W} := \Phi^{-1}(\mathbf{E}_N)$ is a connected open subset of full measure in $\mathcal{D}_K$.*

*Proof.* The openness of $\mathcal{W}$ follows from the fact that $\Phi$ is a continuous map $\mathcal{D}_K \to \mathbf{K}_N$ and that $\mathbf{E}_N$ is open in $\mathbf{K}_N$.

Next note that a Hermitian (in particular, real-symmetric) matrix $X$ has repeated eigenvalues if and only if the discriminant of the characteristic polynomial of $X$ is zero. This is a real-algebraic condition on the entries of $X$ (with the real and complex parts treated separately in the case $K = \mathbb{C}$), so $\mathbf{K}_N \backslash \mathbf{E}_N$ is a real algebraic subset of the real vector space $\mathbf{K}_N$. In fact (see Section 1.3 of [35]), $\mathbf{K}_N \backslash \mathbf{E}_N$ has real codimension 3 in $\mathbf{K}_N$ if $K = \mathbb{C}$ and real codimension 2 in $\mathbf{K}_N$ if $K = \mathbb{R}$. Thus (since $\mathbf{K}_N \backslash \mathbf{E}_N$ is a real algebraic set), in either case $\mathbf{K}_N \backslash \mathbf{E}_N$ can be written as a (disjoint) union of finitely many smooth submanifolds $M_1, \ldots, M_k$ of $\mathbf{K}_N$, each of real codimension at least 2 in $\mathbf{K}_N$.

Ideally, this should indicate that $\Phi^{-1}(\mathbf{K}_N \backslash \mathbf{E}_N)$ is a union of finitely many smooth submanifolds of $\mathcal{D}_K$, each of real codimension at least 2. Indeed, we have by the Transversality Theorem (see, e.g., Section 2.3 of [13]) that for almost every $A \in \mathbf{K}_N$, the map $\Phi$ is transversal to $M_i$ for each $i = 1, \ldots, k$. Then by the preimage theorem for transversal maps (see, e.g., Section 1.4 of [13]), $\Phi^{-1}(M_i)$ is a submanifold of $\mathcal{D}_K$ with (real) codimension in $\mathcal{D}_K$ equal to the codimension of $M_i$ in $\mathbf{K}_N$, which is at least 2.

Thus $\mathcal{W} = \Phi^{-1}(\mathbf{E}_N)$ is equal to $\mathcal{D}_\mathcal{K}$ minus a finite number of submanifolds of codimension at least 2. These submanifolds have zero measure in $\mathcal{D}_K$ (this follows from Sard's theorem; refer, e.g., to [13]), so $\mathcal{W}$ has full measure in $\mathcal{D}_K$.

It only remains to show that $\mathcal{W}$ is connected. Since $\mathcal{D}_{\mathcal{K}}$ is connected, this follows from the general fact that if $Y$ is a connected (hence smoothly path-connected) manifold and $Y_1, \ldots, Y_k$ are submanifolds with codimension at least 2 in $Y$, then $Y \setminus \bigcup_i Y_i$ is connected.

This general fact also follows from a transversality argument, which we now provide for completeness. Let $x, y \in Y$, and let $\gamma : [0,1] \to Y$ be a smooth path with $\gamma(0) = x$ and $\gamma(1) = y$. But there is a homotopy of maps $\gamma_\epsilon$ (with $\gamma_0 = \gamma$) such that $\gamma_\epsilon$ is transversal to each of the $Y_i$ for a.e. $\epsilon$ (see, e.g., the proof of the "transversality homotopy theorem" of Section 2.3 of [13]). Since the $Y_i$ have codimension 2, this implies that $\gamma_\epsilon$ does not intersect any of the $Y_i$ (for a.e. $\epsilon$). Taking $\epsilon$ sufficiently small so that $x = \gamma(0)$ and $y = \gamma(1)$ are connected to $\gamma_\epsilon(0)$ and $\gamma_\epsilon(1)$, respectively, by paths within $Y \setminus \bigcup_i Y_i$, we see that $x$ and $y$ are connected by a path within $Y \setminus \bigcup_i Y_i$.) Also, we know that $\Phi^{-1}(\mathbf{E}_N)$ is open in $\mathcal{D}_K$ because $\mathbf{E}_N$ is open in $\mathbf{K}_N$. □

We now outline the main pieces remaining in the proof of Lemma 6.21. Recall that $\mathcal{D}_K$ is a real-analytic submanifold[2] of $\mathbb{R}^m$ for some $m$. We claim that $\Phi$ is real-analytic on $\mathcal{D}_K$ and that $\Psi$ is real-analytic on $\mathbf{E}_N$. This would imply that $\mathfrak{F}$ is a real-analytic map $\mathcal{W} \to \mathcal{D}_K$. In particular, by an analytic continuation argument (Lemma D.3), the set $\mathcal{W}'$ of points in $\mathcal{W}$ at which the Jacobian of $\mathfrak{F}$ fails to be invertible must either be all of $\mathcal{W}$ or have zero measure in $\mathcal{D}_K$. The former possibility can be ruled out.

Then in words, $\mathfrak{F}$ is a local diffeomorphism on an open set of full measure in $\mathcal{D}_K$. Diffeomorphisms preserve measure zero sets, and by covering $\mathcal{W}'$ with countably many small open sets on which $\mathfrak{F}$ is a diffeomorphism, we will see that the preimage of a measure zero set under $\mathfrak{F}$ must have measure zero.

First we turn to establishing the claimed real-analyticity.

LEMMA D.2. *$\mathfrak{F}|_{\mathcal{W}} : \mathcal{W} \to \mathcal{D}_K$ is a real-analytic map between real-analytic manifolds.*

*Proof.* For $Q \in \mathcal{D}_K$ (so in particular $Q = Q^*$), we can write

$$\begin{aligned}\Phi(Q) &= A + B(QBQ)^\dagger B \\ &= A + \frac{1}{2} B \left[ \left[ QBQ + (I - Q) \right]^{-1} - (I - Q) \right] B \\ &\qquad + \frac{1}{2} B \left[ \left[ Q^* B Q^* + (I - Q^*) \right]^{-1} - (I - Q^*) \right] B.\end{aligned}$$

Written in the latter form, it is clear that $\Phi$ extends to a real-analytic map to $\mathbf{E}_N$, defined on a neighborhood of $\mathcal{D}_K$ in $K^{N \times N}$ (considered, in either case for $K$, as a real coordinate space $\mathbb{R}^q$ for some $q$).

Consider $X_0 \in \mathbf{E}_N$, and let $\mathcal{C}$ be a simple contour in the complex plane surrounding only the lowest $n$ eigenvalues of $X$. The for all $X$ in a sufficiently small neighborhood of $X_0$ in $\mathbf{E}_N$, we have

$$\Psi(X) = \frac{1}{2\pi\imath} \oint_{\mathcal{C}} (z - X)^{-1} \, \mathrm{d}z,$$

where $\mathcal{C}$ is a simple contour in the complex plane surrounding only the lowest $n$ eigenvalues of $X$. In particular, we can choose $\mathcal{C}$ to be a circle of some radius $R > 0$,

[2] $\mathcal{D}_K$ can be identified with the Grassmannian $\mathbf{Gr}(n, K^N)$, i.e., the set of all $n$-dimensional subspaces of $K^N$, which is an algebraic variety of $K$-dimension $(N-n)n$. The space $\mathcal{D}_K$ itself is cut out by the conditions $Q^2 = Q$, $Q^* = Q$, and $\mathrm{Tr}(Q) = n$ on $Q \in K^{N \times N}$. These are real algebraic conditions on $K^{N \times N} \simeq \mathbb{R}^m$ (for some $m$), so $\mathcal{D}_K$ is a (smooth) real algebraic subvariety of $\mathbb{R}^m$. In particular, $\mathcal{D}_K$ has the structure of a real-analytic manifold.

so taking the parametrization $z(t) = R\cos(t) + iR\sin(t)$ yields

$$\Psi(X) = \int_0^{2\pi} \underbrace{\frac{R}{2\pi i}(R\cos(t) + i\sin(t) - X)^{-1}}_{=: G(X,t)} \, \mathrm{d}t.$$

Identifying the real vector space $\mathbf{K}_N$ with $\mathbb{R}^p$ for some $p$, we have that $G(X,t)$ is a rational function $\mathbb{R}^{m+1} \to \mathbb{C}^{N\times N}$, well-defined for all $X$ in a neighborhood of $X_0$, hence real-analytic (if we identify the target space with $\mathbb{R}^{2N^2}$). Since an integral of a real-analytic function with respect to one of its arguments is real-analytic (see Proposition 2.2.3 of [21]), we have established that $\Psi$ is a real-analytic function $\mathbf{E}_N \to \mathcal{D}_K$, where we can interpret the domain as sitting inside some $\mathbb{R}^p$ and the target as sitting inside of $\mathbb{R}^m$, as mentioned above.

Since the composition of real-analytic functions is real-analytic (see Proposition 2.2.8 of [21]), we have established that $\mathfrak{F} = \Psi \circ \Phi$ is real-analytic on a neighborhood of $\mathcal{W} \subset \mathcal{D}_K$ in $K^{N\times N}$. Since $\mathcal{W}$ is open in $\mathcal{D}_K$, $\mathcal{W}$ is a real-analytic submanifold of $\mathcal{D}_K$, and we can view $\mathfrak{F} : \mathcal{W} \to \mathcal{D}_K$ as a real-analytic map between real-analytic manifolds (without thinking of their ambient spaces). □

Next we prove a general fact about real-analytic maps between real-analytic manifolds. This is essentially an analytic continuation result.

Lemma D.3. *Suppose that $F : \mathcal{M} \to \mathcal{N}$ is a real-analytic map between real-analytic manifolds of equal dimension $k$, and $\mathcal{M}$ is connected. Let $\mathcal{M}'$ be the closed subset of points $x \in \mathcal{M}$ at which the Jacobian $DF_x : T_x\mathcal{M} \to T_{F(x)}\mathcal{N}$ is singular. Then either $\mathcal{M} = \mathcal{M}'$ or $\mathcal{M}'$ has zero measure in $\mathcal{M}$.*

*Proof.* In a local coordinate chart, the defining condition for $\mathcal{M}'$ is precisely that the determinant of the $k \times k$ Jacobian matrix in local coordinates (whose entries are real-analytic functions of the local coordinates) is zero. This set is a real-analytic function of local coordinates. The zero set of a real-analytic function on a connected open subset of $\mathbb{R}^k$ is either the whole set or a set of measure zero (in fact, by a much deeper result of Lojasiewicz, a finite union of analytic submanifolds of codimension at least 1—see Theorem 6.3.3 of [21]).

Suppose that the measure of $\mathcal{M}'$ is not zero, so $\mathcal{M}'$ must have positive measure in some coordinate chart, and by the preceding $\mathcal{M}'$ must contain some open set in this chart. Note that the set $\mathcal{A} := \{x \in \mathcal{M} : DF_x \text{ is singular on a neighborhood of } x\}$ is both open and closed in $\mathcal{M}$. The openness follows immediately from the definition, while the closedness follows from the real-analyticity of $F$. To see the latter point, let $y$ be a limit point of $\mathcal{A}$, and let $(\mathcal{U}, \varphi)$ be a coordinate chart near $y$ with $\varphi(y) = 0$. Let the determinant of the $k \times k$ Jacobian matrix $DF$ in local coordinates be denoted by $f$, so $f$ is real-analytic, and moreover $f \equiv 0$ on $\varphi(\mathcal{U})$. Then all of the derivatives of $f$ are uniformly zero on $\varphi(\mathcal{U})$, hence also at the limit point $0 = \varphi(y)$. Since $f$ is real-analytic at 0, this implies that $f \equiv 0$ on a neighborhood of 0, so $DF_x$ is singular on a neighborhood of $y$, i.e., $y \in \mathcal{A}$. This establishes that $\mathcal{A}$ is closed, as desired.

Since $\mathcal{M}$ is connected and $\mathcal{A}$ is both open and closed in $\mathcal{M}$, we must have either $\mathcal{A} = \emptyset$ or $\mathcal{A} = \mathcal{M}$. Since $\mathcal{M}'$ contains an open set, $\mathcal{A}$ cannot be empty. Consequently when the measure of $\mathcal{M}'$ is not zero, $\mathcal{A} = \mathcal{M}$ and $DF_x$ is singular for all $x$. □

In particular, Lemma D.3 implies (together with Lemma D.1) that

$$\mathcal{W}' := \{Q \in \mathcal{W} : D\mathfrak{F}_Q \text{ is singular}\}$$

is either equal to $\mathcal{W}$ or has zero measure in $\mathcal{W}$. The next lemma says that we can

rule out the former possibility.

LEMMA D.4. *For almost every choice of $A$ in $\mathbf{K}_N$, $\mathcal{W}'$ has zero measure in $\mathcal{W}$, hence also (by Lemma D.1) zero measure in $\mathcal{D}_K$. It follows that $\mathcal{W}\backslash\mathcal{W}'$ is an open subset of full measure in $\mathcal{D}_K$.*

*Proof.* We only need to rule out the possibility that $\mathcal{W}' = \mathcal{W}$. We will do so by considering a point near the true density matrix $P$.

Recall from our proof of local convergence that $D\mathfrak{F}_P$ has positive eigenvalues (with $\mathfrak{F}$ considered, depending on the case for $K$, as either a map $\mathcal{D}_{\mathbb{C}} \to \mathcal{D}_{\mathbb{C}}$ or $\mathcal{D}_{\mathbb{R}} \to \mathcal{D}_{\mathbb{R}}$), hence is nonsingular. This means that $P \notin \mathcal{W}'$.

If $P \in \mathcal{W}$, then $\mathcal{W}' \neq \mathcal{W}$, and we are done. More generally, even if $P \notin \mathcal{W}$, observe that since $\mathcal{W}$ is of full measure in $\mathcal{D}_K$ (hence dense in $\mathcal{D}_K$), there is a sequence of density matrices $Q_j \to P$ with $Q_j \in \mathcal{W}$. Since $D\mathfrak{F}_P$ is nonsingular, it follows that $D\mathfrak{F}_{Q_j}$ is nonsingular for $j$ sufficiently large. But then $Q_j \in \mathcal{W}$ and $Q_j \notin \mathcal{W}'$, so $\mathcal{W}' \neq \mathcal{W}$, as desired. □

Now we finish the proof of Lemma 6.21 by the lemma below.

LEMMA D.5. *Let $F : \mathcal{M} \to \mathcal{N}$ be a map between smooth manifolds of equal dimension, and let $\mathcal{V}$ be an open subset of full measure in $\mathcal{M}$ on which $F$ is smooth and $DF$ is nonsingular. Then for any set $S$ of measure zero in $\mathcal{N}$, $F^{-1}(S)$ has measure zero in $\mathcal{M}$.*

*Proof.* Now for every point in $x \in \mathcal{V}$, by the inverse function theorem we can find a neighborhood $\mathcal{U}_x \ni x$ in $\mathcal{M}$ such that $F : \mathcal{U}_x \to \mathcal{N}$ is a diffeomorphism onto its image. Moreover, the size of the neighborhood can be taken to depend only on the derivatives of $F$ near $x$. In particular, we can assume that the size of the neighborhood $\mathcal{U}_x$ is locally bounded away from zero. (By this we mean, fixing some arbitrary Riemannian metric, that for every $x \in \mathcal{V}$, we can take $\mathcal{U}_x$ to contain a Riemannian ball of radius $r(x)$ about $x$, where $x \mapsto r(x)$ is bounded away from zero on every compact subset of $\mathcal{V}$). By fixing a set of coordinate charts on the submanifold $\mathcal{V}$ and taking $\mathcal{X}$ to consist of all the $x$ that are rational points in any of these coordinate charts, we see that $\{\mathcal{U}_x : x \in \mathcal{X}\}$ forms a countable open cover of $\mathcal{V}$. We remark that the details of this construction are made quite explicit in order to avoid invoking the axiom of choice.

Let $S$ be a set with measure zero in $\mathcal{N}$. We can write

$$\begin{aligned} F^{-1}(S) &\subset (\mathcal{M}\backslash\mathcal{V}) \cup \left(F^{-1}(S) \cap \mathcal{V}\right) \\ &= (\mathcal{M}\backslash\mathcal{V}) \cup \bigcup_{x\in\mathcal{X}} \left(F^{-1}(S) \cap \mathcal{U}_x\right) \\ &= (\mathcal{M}\backslash\mathcal{V}) \cup \bigcup_{x\in\mathcal{X}} F^{-1}\left(S \cap F(\mathcal{U}_x)\right). \end{aligned}$$

Now the restriction of $F^{-1}$ to $F(\mathcal{U}_x)$ is a diffeomorphism, so $F^{-1}\left(S \cap F(\mathcal{U}_x)\right)$ is the diffeomorphic image of a measure zero set, hence has measure zero. As a countable union of measure zero sets, $F^{-1}(S)$ has measure zero. □

## REFERENCES


[1] D. G. ANDERSON, *Iterative procedures for nonlinear integral equations*, J. Assoc. Comput. Mach., 12 (1965), pp. 547–560.

[2] A. D. BECKE, *Density functional thermochemistry. iii. the role of exact exchange*, J. Chem. Phys., 98 (1993), p. 5648.

[3] W. CHEN, X. WU, AND R. CAR, *X-ray absorption signatures of the molecular environment in water and ice*, Phys. Rev. Lett., 105 (2010), p. 017802.

[4] A. Damle, L. Lin, and L. Ying, *Compressed representation of Kohn–Sham orbitals via selected columns of the density matrix*, J. Chem. Theory Comput., 11 (2015), pp. 1463–1469.

[5] E.R. Davidson, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices*, J. Comput. Phys., 17 (1975), pp. 87–94.

[6] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. iii.*, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

[7] W. Dawson and F. Gygi, *Performance and accuracy of recursive subspace bisection for hybrid DFT calculations in inhomogeneous systems*, J. Chem. Theory Comput., 11 (2015), pp. 4655–4663.

[8] R. A. DiStasio, B. Santra, Z. Li, X. Wu, and R. Car, *The individual and collective effects of exact exchange and dispersion interactions on the ab initio structure of liquid water*, J. Chem. Phys., 141 (2014), p. 084502.

[9] I. Duchemin and F. Gygi, *A scalable and accurate algorithm for the computation of Hartree–Fock exchange*, Comput. Phys. Commun., 181 (2010), pp. 855–860.

[10] W. E, T. Li, and J. Lu, *Localized bases of eigensubspaces and operator compression*, Proc. Natl. Acad. Sci., 107 (2010), pp. 1273–1278.

[11] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo Pasquarello, Lorenzo Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M Wentzcovitch, *QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials*, J. Phys.: Condens. Matter, 21 (2009), pp. 395502–395520.

[12] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins Univ. Press, Baltimore, fourth ed., 2013.

[13] V. Guillemin and A. Pollack, *Differential Topology*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[14] F. Gygi, *Compact representations of Kohn–Sham invariant subspaces*, Phys. Rev. Lett., 102 (2009), p. 166406.

[15] F. Gygi and A. Baldereschi, *Self-consistent Hartree-Fock and screened-exchange calculations in solids: Application to silicon*, Phys. Rev. B, 34 (1986), pp. 4405–4408.

[16] J. Heyd, G. E. Scuseria, and M. Ernzerhof, *Hybrid functionals based on a screened coulomb potential*, J. Chem. Phys., 118 (2003), pp. 8207–8215.

[17] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas*, Phys. Rev., 136 (1964), pp. B864–B871.

[18] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comp., 23 (2001), pp. 517–541.

[19] W. Kohn, *Density functional and density matrix method scaling linearly with the number of atoms*, Phys. Rev. Lett., 76 (1996), pp. 3168–3171.

[20] W. Kohn and L. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140 (1965), pp. A1133–A1138.

[21] S. G. Krantz and H. R. Parks, *A Primer of Real-Analytic Functions*, Springer, 2002.

[22] L. Lin, *Adaptively compressed exchange operator*, J. Chem. Theory Comput., 12 (2016), p. 2242.

[23] L. Lin, Z. Xu, and L. Ying, *Adaptively compressed polarizability operator for accelerating large scale ab initio phonon calculations*, Multiscale Model. Simul., 15 (2017), pp. 29–55.

[24] R. Martin, *Electronic Structure – Basic Theory and Practical Methods*, Cambridge Univ. Pr., West Nyack, NY, 2004.

[25] N. Marzari and D. Vanderbilt, *Maximally localized generalized Wannier functions for composite energy bands*, Phys. Rev. B, 56 (1997), p. 12847.

[26] G. Onida, L. Reining, and A. Rubio, *Electronic excitations: density-functional versus many-body Green's-function approaches*, Rev. Mod. Phys., 74 (2002), p. 601.

[27] V. Ozoliņš, R. Lai, R. Caflisch, and S. Osher, *Compressed modes for variational problems in mathematics and physics*, Proc. Natl. Acad. Sci., 110 (2013), pp. 18368–18373.

[28] B. N. Parlett, *The symmetric eigenvalue problem*, vol. 7, SIAM, 1980.

[29] B. N. Parlett and Y. Saad, *Complex shift and invert strategies for real matrices*, Linear Algebra Appl., 88 (1987), pp. 575–595.

[30] J. P. Perdew, M. Ernzerhof, and K. Burke, *Rationale for mixing exact exchange with density functional approximations*, J. Chem. Phys., 105 (1996), pp. 9982–9985.

[31] E. Runge and E. K .U. Gross, *Density-functional theory for time-dependent systems*, Phys.

Rev. Lett., 52 (1984), p. 997.
[32] V. R. Saunders and I. H. Hillier, *A "level–shifting" method for converging closed shell Hartree–Fock wave functions*, Int. J. Quantum Chem., 7 (1973), pp. 699–705.
[33] G. Sleijpen and H. A. Van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM Rev., 42 (2000), pp. 267–293.
[34] A. Szabo and N.S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, McGraw-Hill, New York, 1989.
[35] T. Tao, *Topics in Random Matrix Theory*, American Mathematical Society, 2012.
[36] G. Teschl, *Ordinary Differential Equations and Dynamical Systems*, American Mathematical Society, 2012.
[37] X. Wu, A. Selloni, and R. Car, *Order-N implementation of exact exchange in extended insulating systems*, Phys. Rev. B, 79 (2009), p. 085102.
[38] F. Zhang, ed., *The Schur Complement and its Applications*, Springer, 2005.