THE CONVERGENCE AND ERROR ANALYSIS OF COORDINATE DESCENT METHODS WITH COMPRESSION FOR FULL CONFIGURATION INTERACTION

YINGZHOU LI* AND QIANG WU†

Abstract. We study the effect of compression in Coordinate Descent Full Configuration Interaction (CDFCI) within an unconstrained optimization formulation of the full configuration interaction ground-state problem. Under suitable local assumptions, we prove that the compressed iteration converges linearly to the solution of an associated restricted problem. We also characterize the convergence point of the compressed algorithm. Under an additional exponential decay assumption on the target eigenvector, we show that the resulting eigenvalue error is of order $O(\tau^2)$, where τ denotes the compression threshold. Numerical results support the analysis.

1. Introduction. Full Configuration Interaction (FCI) provides the exact solution of the electronic Schrödinger equation within a finite orbital basis. Under the Slater determinant basis, the FCI problem can be formulated as the eigenvalue problem

$$Hc = \lambda c, \tag{1.1}$$

where H is the Hamiltonian matrix and the smallest eigenvalue corresponds to the ground-state energy. The dimension of H grows combinatorially with the number of electrons and orbitals, making direct eigensolvers impractical except for very small systems. A number of approaches have therefore been developed for large-scale FCI calculations, including Davidson-type methods, selected configuration interaction methods, and stochastic projector methods [1, 2, 6, 9, 10, 14]

Among these methods, Coordinate Descent Full Configuration Interaction (CDFCI), proposed in [18], reformulates the eigenvalue problem as an unconstrained optimization problem and updates one coordinate at a time. Since each iteration only accesses a small portion of the Hamiltonian, CDFCI avoids the storage and computational costs associated with full-vector updates and is therefore suitable for very large FCI problems [18, 19, 22].

A compression step is commonly used in practical implementations of CDFCI. During the iteration, the support of the coefficient vector typically grows continuously. Without compression, both storage and computational cost increase rapidly as the iteration proceeds. To control the size of the iterate, entries whose magnitudes are below a prescribed threshold are discarded. The resulting vector remains sparse, allowing the algorithm to be applied to larger systems.

The compression step, however, makes the analysis substantially more difficult. Most convergence analyses of coordinate descent methods assume that the iteration is performed in the full coordinate space and do not account for compression or truncation of the iterates [15, 16, 20]. In compressed CDFCI, however, coordinates may be repeatedly removed and reintroduced during the iteration, and the resulting algorithm falls outside the standard coordinate descent framework. Consequently, the existing theory cannot be directly applied to compressed CDFCI.

Compression and truncation have also been used in eigenvalue computations. Representative examples include truncated power methods, sparse PCA algorithms,

*School of Mathematical Sciences, Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University and Key Laboratory of Computational Physical Sciences, Ministry of Education (yingzhouli@fudan.edu.cn)

†School of Mathematical Sciences, Fudan University (wuq23@m.fudan.edu.cn)

compressed subspace iterations, and Krylov-based eigensolvers [3,5,8,13,17,21]. Most existing analyses rely on the assumption that the target eigenvector is sparse or approximately sparse. In contrast, the exact FCI eigenvector is generally not sparse, and compression is introduced solely to control the size of the iterate. Therefore, these results do not directly apply to CDFCI with compression.

The effect of compression is also related to eigenvalue perturbation theory [4,7,11]. Classical perturbation results describe the change of eigenvalues and eigenvectors under matrix perturbations. In the present setting, however, the compression error is generated dynamically during the iteration and cannot be represented by a fixed perturbation matrix. This requires a different analysis.

The purpose of this paper is to study the effect of compression in CDFCI. We establish convergence results for the compressed iteration and quantify the error introduced by compression.

The main contributions of this work are summarized as follows.

1. We prove that the compressed CDFCI iteration converges linearly to the solution of an associated restricted optimization problem.
2. We characterize the limiting point of the compressed iteration and show its relation to the corresponding restricted eigenvalue problem.
3. We derive an eigenvalue error estimate caused by compression. In particular, we prove that the error between the ground-state eigenvalue λ of the full problem and the optimal eigenvalue $\lambda_{k\text{-opt}}$ of the compressed k-dimensional problem satisfies
$$|\lambda - \lambda_{k\text{-opt}}| = O(\tau^2),$$
where τ denotes the compression threshold.

The remainder of this paper is organized as follows. Section 2 reviews the coordinate descent algorithm with compression. Section 3 presents the main theoretical results, including the linear convergence rate and an upper bound on the energy error, and outlines the key ideas of the proofs. Section 4 is devoted to the rigorous proofs of these results. Section 5 reports numerical experiments that validate the theoretical findings and illustrate the practical performance of the algorithm. Finally, Section 6 concludes the paper and discusses possible directions for future research.

2. Preliminaries on CDM and Compression .

2.1. Dimensionality and Sparsity Structure. The dimension of the FCI Hamiltonian matrix is determined by the number of spin orbitals M and electrons N, scaling according to the binomial coefficient $\binom{M}{N}$. Consequently, the matrix dimension grows exponentially with respect to the system size. This combinatorial growth is illustrated in Table 1, which reports the dimensions of the FCI Hamiltonian for the water molecule (H_2O, $N = 10$) under various basis sets.

Despite this exponential growth in dimension, the FCI Hamiltonian matrix remains extremely sparse due to the structure of the second-quantized Hamiltonian: a matrix element is nonzero only when the corresponding pair of Slater determinants differs by at most two spin orbitals. As a consequence, the number of nonzero entries per row scales only polynomially with M and N. The last column of Table 1 reports an estimated upper bound on the maximum number of nonzero entries per row, which is derived from the counts of single and double excitations:
$$N(M-N) + \frac{N(N-1)}{2} \cdot \frac{(M-N)(M-N-1)}{2}.$$
Furthermore, the ground-state eigenvector of the FCI Hamiltonian exhibits exponen-

Basis set	Spin orbitals M	FCI dimension	Max nnz per row
STO-3G	24	2.0×10^6	4,235
cc-pVDZ	48	6.54×10^9	32,015
cc-pVTZ	82	2.13×10^{12}	115,740
cc-pVQZ	140	5.73×10^{14}	378,625

Table 1: FCI Hamiltonian dimension and estimated maximum number of nonzero elements per row for H_2O ($N = 10$) under different basis sets.

tial decay. Specifically, the magnitudes of its entries, when sorted in descending order, decrease exponentially. This property provides a justification for the truncation strategy, as the tail components make a negligible contribution to the wave function. This decay also plays a central role in establishing sharper error bounds for the compression scheme.

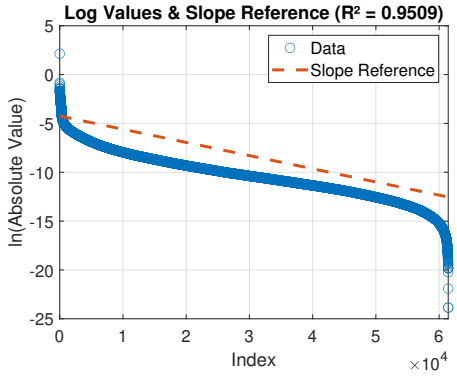


Fig. 1: Exponential decay of the ground-state eigenvector components for H_2O-STO3G basis.

2.2. Coordinate Descent with Compression. The FCI eigenvalue problem can be recast as the following unconstrained nonconvex optimization problem:

$$\min_{c \in \mathbb{R}^{N_{\mathrm{FCI}}}} f(c) = \left\| H + cc^\top \right\|_{\mathrm{F}}^2, \tag{2.1}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and N_{FCI} is the dimension of the FCI Hamiltonian matrix. The gradient of the objective function is given by

$$\nabla f(c) = 4Hc + 4\left(c^\top c\right) c. \tag{2.2}$$

As established in [12], let

$$\lambda = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots$$

be the eigenvalues of H, and let c_i denote the normalized eigenvector corresponding to λ_i. The stationary points of (2.1) are precisely 0 and

$$\pm\sqrt{-\lambda_i}\, c_i, \qquad i = 0, 1, \ldots.$$

Moreover, only the points $\pm\sqrt{-\lambda}\,c_0$ are local minimizers, and both attain the same global minimum value. All other stationary points are saddle points or local maximizer. Consequently, solving (2.1) yields both the ground-state energy λ and the corresponding ground-state eigenvector c_0.

To solve the optimization problem (2.1), traditional eigenvalue solvers and general-purpose optimization algorithms are computationally intractable due to the exponential dimensionality of the Hamiltonian matrix. Consequently, [12] proposed a coordinate descent framework for solving this optimization problem.

The Coordinate Descent FCI (CDFCI) algorithm maintains two sparse vectors, $c^{(l)}$ and $b^{(l)}$, throughout the iterations. Here, $c^{(l)}$ denotes the coefficient vector computed at the l-th iteration, while $b^{(l)}$ serves as a compressed approximation of the matrix-vector product $Hc^{(l)}$.

In the $(l+1)$-th iteration, CDFCI first identifies the coordinate index $i^{(l+1)}$ corresponding to the gradient component with the largest magnitude:

$$i^{(l+1)} = \arg\max_j \left| 4b_j^{(l)} + 4\left((c^{(l)})^\top c^{(l)}\right) c_j^{(l)} \right|. \tag{2.3}$$

Subsequently, the algorithm performs an exact line search to determine the optimal update. To facilitate the compression of b, we formulate this step as finding the optimal value for the selected coordinate, rather than an additive increment. Specifically, the new value α is determined by

$$\alpha = \arg\min_{\widetilde{\alpha}\in\mathbb{R}} \; f\left(c^{(l)} - c^{(l)}_{i^{(l+1)}} e_{i^{(l+1)}} + \widetilde{\alpha} e_{i^{(l+1)}}\right), \tag{2.4}$$

where $e_{i^{(l+1)}}$ denotes the standard basis vector corresponding to index $i^{(l+1)}$. This update rule implies that the entry $c_{i^{(l+1)}}$ is implicitly reset to zero and then assigned the new value α that minimizes the objective function.

The vector $b^{(l)}$ plays a pivotal role in ensuring computational efficiency. Specifically, once the new coefficient vector $c^{(l+1)}$ is determined, $b^{(l)}$ is updated recursively to obtain $b^{(l+1)}$. The update rule is given by

$$b^{(l+1)} = b^{(l)} - c^{(l)}_{i^{(l+1)}} H_{:,i^{(l+1)}} + \alpha H_{:,i^{(l+1)}}. \tag{2.5}$$

This incremental update strategy avoids the expensive recomputation of the full matrix-vector product Hc. Instead, it efficiently adjusts $b^{(l)}$ by subtracting the contribution of the previous scalar component at coordinate $i^{(l+1)}$ and adding the contribution from the newly computed value α.

However, updating the vector b exactly is computationally expensive. To reduce the memory cost, [18] introduced a compression strategy that controls when new nonzero entries are added to b. Given a threshold τ, the following rules are applied to each coordinate j:

(i) **If** $b_j^{(l)} \neq 0$ (the coordinate is already selected): The value is always updated, regardless of the magnitude of the change.

(ii) **If** $b_j^{(l)} = 0$ (the coordinate is not selected): The coordinate is updated only if the new term is sufficiently large, that is, if

$$\left|\alpha H_{j,i^{(l+1)}}\right| > \tau.$$

Otherwise, the update is discarded and $b_j^{(l+1)}$ remains zero.

This strategy preserves all existing nonzero entries, while new coordinates are added to the active set only if their contribution is significant.

A coordinate i remains inactive if it is not selected by the greedy rule. For such a coordinate, c_i stays zero. The compression strategy ensures that b_i also remains zero unless it receives a large update. This controls the number of nonzero elements (the support size). Crucially, once a coordinate i is selected and updated, it enters the active set and is never compressed again. Thereafter, its value is recomputed at every iteration and is no longer affected by the compression threshold.

This implies that we only discard entries in b for coordinates where $c_i = 0$. Consequently, the Rayleigh quotient

$$\frac{(c^{(l)})^\top H c^{(l)}}{(c^{(l)})^\top c^{(l)}} \tag{2.6}$$

can be computed exactly using the auxiliary vector b:

$$\frac{(c^{(l)})^\top b^{(l)}}{(c^{(l)})^\top c^{(l)}}. \tag{2.7}$$

This equality holds because any compressed term in b corresponds to a zero coefficient in c, so it makes no contribution to the inner product (that is, $c_i b_i = 0$).

3. Main Theoretical Results. In this section, we present the theoretical analysis of the CDFCI algorithm with compression. Our analysis consists of two main parts. First, we prove the linear convergence of the algorithm by studying the energy decrease in each iteration. Second, we analyze the convergence point of the algorithm and show that the final result approximates the true ground state with a bounded energy error. Together, these results provide theoretical guarantees for both the convergence rate and the solution accuracy.

3.1. Linear Convergence. In this subsection, we analyze the convergence of the CDFCI algorithm with compression. We begin by stating the necessary assumptions. Throughout the analysis, we assume that the Hamiltonian matrix $H \in \mathbb{R}^{n\times n}$ is sparse, and let n_z denote the maximum number of nonzero entries in any row of H. This assumption is consistent with the structure of the FCI Hamiltonian, as discussed in Section 2.

We start by introducing the necessary definitions and lemmas. These results establish that the objective function is strongly convex in a neighborhood of the optimal solution, which is the key ingredient in the proof of the linear convergence result stated in Theorem 3.7.

DEFINITION 3.1. *A continuously differentiable function $g : \mathbb{S} \to \mathbb{R}$ is said to be strongly $\|\cdot\|_p$-convex if there exists a constant $\mu_p > 0$ such that*

$$g(y) \geq g(x) + \nabla g(x)^\top (y - x) + \frac{\mu_p}{2}\|y - x\|_p^2, \quad \forall x, y \in \mathbb{S}. \tag{3.1}$$

To analyze the local behavior of the objective function near the optimal solutions $\pm c$, we define the following neighborhoods, following the approach in [12]:

$$B^\pm = \left\{ y \;\middle|\; \|y \mp c\|_2 \leq \frac{1}{30}\frac{\min(-2\lambda, -\lambda + \lambda_2)}{\sqrt{-\lambda}} \right\},$$
$$D^\pm = \left\{ x \in B^\pm \;\middle|\; f(x) \leq \min_{y\in\partial B^\pm} f(y) \right\}.$$

Here, λ and λ_2 denote the smallest and second smallest eigenvalues of the Hamiltonian H, respectively. The following Lemma 3.2 shows that the objective function f is strongly convex within the neighborhoods $B^{\pm}$.

LEMMA 3.2. *The function f is strongly $\|\cdot\|_2$-convex on both B^+ and B^- with the constant $\mu_2 = 3\min(2|\lambda|, |\lambda - \lambda_2|)$. Furthermore, for any $p \geq 1$, there exists a constant $\mu_p > 0$ such that f is strongly $\|\cdot\|_p$-convex.*

We assume that the initial vector $c^{(0)}$ lies in $D^{\pm}$. This assumption is reasonable in quantum chemistry. Specifically, the Hartree-Fock method typically provides an initial guess that is close to the true ground state. Therefore, the initial coefficient vector is usually close to a minimizer of f, and thus lies inside the strongly convex neighborhood required for our analysis.

Let V_0 denote the set of nonzero coordinates (the support) of the initial vector $c^{(0)}$. Let V_l be the cumulative support set, which contains all nonzero coordinates appearing in the iterates up to the l-th step. We prove that the total number of active coordinates remains bounded.

DEFINITION 3.3. *We define the cumulative support V_l at iteration l as*

$$V_l := \bigcup_{k=0}^{l} \operatorname{supp}(c^{(k)}), \quad l \geq 0,$$

and the maximal support set V as

$$V := \bigcup_{l \geq 0} V_l.$$

LEMMA 3.4. *The maximal support set V defined in Definition 3.3 is finite.*

Proof. By construction, the sets satisfy the recurrence $V_l = V_{l-1} \cup \operatorname{supp}(c^{(l)})$. Consequently, $\{V_l\}$ forms a nested sequence:

$$V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots.$$

Since every V_l is a subset of the finite index set $\{1, 2, \ldots, n\}$, their union $V = \bigcup_{l \geq 0} V_l$ must also be a subset of $\{1, 2, \ldots, n\}$. Therefore, V is finite. Moreover, by definition, V contains the support of $c^{(l)}$ for all l, and thus serves as the maximal support set throughout the iterations. ∎

To establish the linear convergence of the CDFCI algorithm with compression, we proceed in three steps, each is supported by a corresponding lemma. The overall strategy is as follows:

1. Function value gap control: We first show that the objective gap $f(c^{(l)}) - f(c)$ is $O(\tau^2)$ when the step l is sufficiently large as in Lemma 3.5
2. Based on the error bound, we prove that the limit of the sequence lies inside the region where the function is strongly convex. This is shown in Lemma 3.6.
3. Convergence rate: Finally, by leveraging the strong convexity of the objective in this region, we establish the linear convergence of the algorithm as in Theorem 3.7.

LEMMA 3.5. *Let $c^{(0)} \in D^{\pm}$ be the initial point. When l is sufficiently large, we can obtain*

$$f(c^{(l)}) - f(c) < \frac{8n_z^2\tau^2}{\mu_1}.$$

LEMMA 3.6. *Define the maximal support set V of the initial point $c^{(0)}$ as in Definition 3.3. Consider the principal submatrix of H restricted to the index set $V \times V$. Let $c^{k\text{-}opt}$ be the eigenvector corresponding to its smallest eigenvalue $\lambda^{k\text{-}opt}$, normalized such that $\|c^{k\text{-}opt}\|_2 = \sqrt{-\lambda^{k\text{-}opt}}$. When τ is sufficiently small, $c^{k\text{-}opt}$ lies within $D^{\pm}$.*

THEOREM 3.7. *From Lemma 3.6, it follows that the algorithm with compression converges to $c^{k\text{-}opt}$ or $-c^{k\text{-}opt}$ and still exhibits linear convergence, that is:*

$$f(c^{(l+1)}) - f(c^{k\text{-}opt}) \leq (1 - \frac{\mu_1}{L})(f(c^{(l)}) - f(c^{k\text{-}opt})).$$

The proofs of the above results are deferred to Section 4.

3.2. Analysis of the convergence points. We now analyze the convergence points. Without loss of generality, we assume that the first k coordinates are selected during the iteration process. Let $c_1^{\text{k-opt}} \in \mathbb{R}^k$ denote the sub-vector of $c^{\text{k-opt}}$ corresponding to these selected coordinates.

In addition, since the Hamiltonian matrix H is extremely sparse and each row contains at most n_z nonzero entries, we consider the regime $k > n_z$, which is the regime considered in the subsequent analysis. Accordingly, we partition the matrix $H \in \mathbb{R}^{n\times n}$ as

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix},$$

where $H_{11} \in \mathbb{R}^{k\times k}$ corresponds to the block of selected coordinates. Let $\lambda^{\text{k-opt}}$ denote the smallest eigenvalue of H_{11}.

It follows from the proof of Theorem 3.7 that the sub-vector $c_1^{\text{k-opt}}$ is an eigenvector of H_{11} associated with $\lambda^{\text{k-opt}}$, and satisfies the condition $\|c_1^{\text{k-opt}}\|_2 = \sqrt{-\lambda^{\text{k-opt}}}$.

Since the support of $c^{\text{k-opt}}$ is fixed to the first k indices, it follows that for all $i > k$, the coordinate i is never selected. This implies that during the CDFCI process, we have the following bound:

$$|(Hc^{\text{k-opt}})_i| = \left| \sum_{j=1}^{k} H_{ij} (c_1^{\text{k-opt}})_j \right| < n_z \tau. \tag{3.2}$$

This inequality holds because whenever the auxiliary vector b is updated, the value b_i is compressed to zero due to the thresholding rule. Therefore, the coordinate i is excluded from selection.

To analyze the convergence, we introduce structural assumptions on the Hamiltonian and its ground-state eigenvector. These assumptions allow us to quantify the effect of the compression strategy.

ASSUMPTION 3.8. *We assume that the matrix $H \in \mathbb{R}^{n\times n}$ has a positive spectral gap. Specifically, let λ be the smallest eigenvalue of H. We define the gap as*

$$\text{gap}(H) := \min_{\lambda_H \in \sigma(H),\, \lambda_H \neq \lambda} |\lambda - \lambda_H| > 0.$$

ASSUMPTION 3.9. *Let $c \in \mathbb{R}^n$ be the eigenvector corresponding to the smallest eigenvalue λ. We assume that c exhibits exponential decay. Specifically, if we rearrange the indices such that the entries of c are sorted by magnitude (i.e., $|c_1| \geq |c_2| \geq \cdots \geq |c_n|$), then there exist constants $N > 0$ and $P > 0$ such that*

$$|c_j| \leq |c_N| e^{-P(j-N)}, \quad \forall j \geq N.$$

As a direct consequence of Theorem 3.9, we establish a technical lemma that provides norm estimates for exponentially decaying vectors. This result will be repeatedly used to control tail contributions in the analysis.

LEMMA 3.10. *Let $x \in \mathbb{R}^n$ be a vector with exponential decay, meaning that there exist constants $N > 0$ and $P > 0$ such that*

$$|(x)_j| < |(x)_N| e^{-P(j-N)}, \quad \forall j \geq N.$$

Then there exists a constant $Q > 0$, depending only on the N and P and independent of the dimension n, such that

$$\|x\|_s < Q\|x\|_\infty, \quad \text{for } s = 1 \text{ or } 2.$$

From Lemma 3.10 we can get $\|c\|_1^2 < Q^2\|c\|_\infty^2 \leq Q^2\|c\|_2^2 = -\lambda Q^2$.

Next, we derive a preliminary bound on the eigenvalue error $\Delta\lambda$.

LEMMA 3.11. *Let $\Delta\lambda = |\lambda - \lambda^{k\text{-}opt}|$. Recall from the previous section that for all unselected coordinates i (where $k+1 \leq i \leq n$), the matrix-vector product satisfies $|H_{ij}(c_1^{k\text{-}opt})_j| < \tau$. Under this condition, we have*

$$\Delta\lambda < \frac{n_z\|c\|_1\tau}{-2\lambda} < \frac{n_z Q\tau}{2\sqrt{-\lambda}}. \tag{3.3}$$

We now analyze the structure of the true eigenvector c. Consistent with the partition of H, we split c into two parts: $c_1 \in \mathbb{R}^k$, which contains the first k components, and $c_2 \in \mathbb{R}^{n-k}$, which contains the remaining components. This leads to the following lemma regarding the magnitude of the tail component c_2.

LEMMA 3.12. *Under Theorem 3.8 and Theorem 3.9, we have*

$$\|c_2\|_\infty < \frac{2mn_z\sqrt{k}\tau}{\text{gap}(H)}, \tag{3.4}$$

where the constant m is defined as

$$m = \sqrt{\frac{-\lambda^{k\text{-}opt}\|c\|_1^2}{4k\lambda^2} + 1}.$$

Note that m is bounded by $\sqrt{\frac{-\lambda^{k\text{-}opt}Q^2}{4k\lambda} + 1}$ and is independent of the dimension n.

Building on the analysis of c_2, we now establish a refined estimate for the eigenvalue error. We show that the energy error is bounded by a quantity proportional to τ^2.

THEOREM 3.13. *Under Assumptions 3.8 and 3.9, the eigenvalue error satisfies the following bound:*

$$\Delta\lambda = |\lambda - \lambda^{k\text{-}opt}| < \frac{4mn_z^2\sqrt{k}Q\tau^2}{-\lambda\,\text{gap}(H)}. \tag{3.5}$$

4. Technical proof. Throughout this section, we assume that the support of $c^{(l)}$ (and hence the selected coordinates) is contained in the first k indices. Since all arguments are invariant under permutations of coordinates, we fix this ordering convention throughout the remainder of this section. Furthermore, for notational simplicity, we assume that $k > n_z$, where n_z denotes the maximum number of nonzero entries in any row of H. The general case can be handled by replacing n_z with $\min(k, n_z)$ throughout the analysis.

4.1. Lemma 3.5's proof.

Proof. Firstly, at iteration l, after selecting the coordinate j_l, the compression algorithm behaves identically to the original (uncompressed) algorithm. Therefore, Lemma 5.5 in [12] still holds:

$$f(c^{(l+1)}) \leq f(c^{(l)}) - \frac{1}{2L}\left(\nabla_{j_l} f(c^{(l)})\right)^2. \tag{4.1}$$

Furthermore, the fact that $c^{(l+1)} \in D^{\pm}$ also follows from the theoretical analysis of the original algorithm. The function $f\left(c^{(l)}\right)$ is monotonically decreasing and has a lower bound $f(c)$, converging to a value $\eta \geq f(c)$. Fixing $0 < \varepsilon < \frac{n_z^2\tau^2}{2L}$, when l is sufficiently large, $f\left(c^{(l)}\right) - f\left(c^{(l+1)}\right) < \varepsilon$. That is $\frac{1}{2L}\left(\nabla_{j_l} f(c^{(l)})\right)^2 < \varepsilon$, $\left|\nabla_{j_l} f\left(c^{(l)}\right)\right| < \sqrt{2L\varepsilon}$. Because $\left|\nabla_{j_l} f\left(c^{(l)}\right)\right|$ attains the maximum absolute value among the first k components (i.e., the selected coordinates) of $\nabla f\left(c^{(l)}\right)$, we have:

$$|\nabla_i f(c^{(l)})| < \sqrt{2L\varepsilon}, \quad \forall 1 \leq i \leq k. \tag{4.2}$$

When $i > k$, $b_i^{(l)}$ and $c_i = 0$, and $|\nabla_i f(c^{(l)})| = 4(Hc^{(l)})_i + ({c^{(l)}}^\top c^{(l)})(c^{(l)})_i = 4(Hc^{(l)})_i$. $b_i^{(l)} = 0$ indicates that b_i was compressed during the update. Each time we update $c_{j_l}^{(l)}$ using exact line search, it is equivalent to updating from 0, and we can consider it as having performed k compressions. $b_i^{(l)}$ has been compressed at most k times and the j-th compression introduces an error less than $|4H_{ij}c_j^{(l)}|$. Thus,

$$\begin{aligned} |\nabla_i f(c^{(l)})| &= 4|(Hc^{(l)})_i| \\ &\leq |4b_i^{(l)}| + \sum_{j=1}^{k} |4H_{ij}c_j^{(l)}| \\ &= \sum_{j=1}^{k} |4H_{ij}c_j^{(l)}| < 4\min\{n_z, k\}\tau = 4n_z\tau, \qquad \forall\, i > k, \end{aligned} \tag{4.3}$$

where the last equality follows from the convention $k > n_z$ adopted at the beginning of this section. Since $\varepsilon < \frac{n_z^2\tau^2}{2L}$ and $\sqrt{2L\varepsilon} < n_z\tau$, we have

$$\|\nabla f(c^{(l)})\|_\infty < 4n_z\tau. \tag{4.4}$$

Moreover, as $c^{(l)}$ is updated, this inequality continues to hold. Due to the strong convexity of f on $B^{\pm}$:

$$\begin{aligned} f(c) &\geq f(c^{(l)}) - (-\nabla f(c^{(l)})^\top (c - c^{(l)}) - \frac{\mu_1}{2}\|c - c^{(l)}\|_1^2) \\ &\geq f(c^{(l)}) - (\|\nabla f(c^{(l)})\|_\infty \|c - c^{(l)}\|_1 - \frac{\mu_1}{2}\|c - c^{(l)}\|_1^2)) \\ &\geq f(c^{(l)}) - \frac{1}{2\mu_1}\|\nabla f(c^{(l)})\|_\infty^2 \\ &= f(c^{(l)}) - \frac{8n_z^2\tau^2}{\mu_1}, \end{aligned} \tag{4.5}$$

where the third inequality is obtained by regarding the expression within the parentheses as a quadratic function of $\left\|c - c^{(l)}\right\|_1$, and finding the maximum value of this quadratic function. ∎

Having established that the when l is large the point of the iteration yields a function value close to the global minimum, we next check the limit point location. In particular, we show that the limiting point must lie within the region $D^{\pm}$.

4.2. Lemma 3.6's proof.

Proof. Let $C_V = \{x \in \mathbb{R}^n : \mathrm{supp}(x) \subseteq V\}$. By construction, both $c^{\text{k-opt}}$ and $c^{(l)}$ lie in C_V. Recall that our objective function is

$$f(x) = \|H + xx^\top\|_{\mathrm{F}}^2.$$

For any $x \in C_V$, the matrix $xx^\top$ has support contained in $V \times V$; hence the value of $f(x)$ depends only on the principal submatrix H_{VV} of H. Therefore, minimizing $f(x)$ over C_V is equivalent to solving the reduced problem

$$\min_{z \in \mathbb{R}^{|V|}} \|H_{VV} + zz^\top\|_{\mathrm{F}}^2.$$

In [12] it is shown that, for an objective of the form $\|H + xx^\top\|_{\mathrm{F}}^2$, the local minimum is achieved precisely at the eigenvector corresponding to the smallest eigenvalue.

Applying this result to the reduced problem with $A = H_{VV}$, we conclude that the minimizer of

$$\min_z \|H_{VV} + zz^\top\|_{\mathrm{F}}^2$$

is given by the eigenvector associated with the smallest eigenvalue of H_{VV}. By definition, this minimizer is exactly $c^{\text{k-opt}}$, and hence $c^{\text{k-opt}}$ is a local minimum of $f(x)$ over C_V. thus $f(c^{\text{k-opt}}) < f(c^{(l)})$ and using Lemma 3.5 we have:

$$f\left(c^{\text{k-opt}}\right) - f(c) < \frac{8n_z^2\tau^2}{\mu_1}. \tag{4.6}$$

Because $f(c)$ attains the global minimum of f, the quantity

$$\min_{y \in \partial B_{\pm}} f(y) - f(c)$$

is a positive constant. If the compression threshold τ satisfies

$$\tau < \frac{\sqrt{\mu_1\left(\min_{y \in \partial B_{\pm}} f(y) - f(c)\right)}}{2\sqrt{2}n_z},$$

then

$$f(c^{\text{k-opt}}) - f(c) < \min_{y \in \partial B_{\pm}} f(y) - f(c), \tag{4.7}$$

and $f(c^{\text{k-opt}}) < \min_{y \in \partial B_{\pm}} f(y)$, which implies that $c^{\text{k-opt}}$ must lie inside $D^{\pm}$.

To see this, suppose by contradiction that $c^{\text{k-opt}} \notin D^{\pm}$. Then $c^{\text{k-opt}}$ lies in $\mathbb{R}^n \setminus B_{\pm}$, where the minimum of f can only occur either on the boundary $\partial B_{\pm}$ or at a stationary point of f. The stationary points outside $B_{\pm}$ are $\pm\sqrt{-\lambda'}x$ (with $\lambda' > \lambda$ and x the corresponding eigenvector) and the zero vector. However, if

$$\begin{aligned} \tau &< \min\left\{ \frac{\sqrt{\mu_1(f(\sqrt{-\lambda'}x) - f(c))}}{2\sqrt{2}n_z}, \frac{\sqrt{\mu_1(f(\mathbf{0}) - f(c))}}{2\sqrt{2}n_z} \right\} \\ &= \min\left\{ \frac{\sqrt{\mu_1(\lambda^2 - \lambda'^2)}}{2\sqrt{2}n_z}, \frac{\sqrt{\mu_1\lambda^2}}{2\sqrt{2}n_z} \right\}, \end{aligned} \tag{4.8}$$

then we have

$$f(c^{\text{k-opt}}) < f(c) + \frac{8n_z^2\tau^2}{\mu_1} < \min\Big\{f(\pm\sqrt{-\lambda'}x),\, f(\mathbf{0})\Big\}.$$

Hence, $f(c^{\text{k-opt}})$ is smaller than the function values at all points in $\mathbb{R}^n \setminus B_{\pm}$, which contradicts the assumption that $c^{\text{k-opt}}$ lies outside $D^{\pm}$. Therefore, $c^{\text{k-opt}} \in D^{\pm}$. ∎

Since the iterates lie in a strongly convex region, we can now proceed to prove the linear convergence of the iteration.

4.3. Theorem 3.7's proof.

Proof. We use $\Delta\bar{c}$ to represent the truncated vector obtained by taking the first k components from $c^{\text{k-opt}} - c^{(l)}$, and $\nabla\bar{f}(c^{(l)})$ to represent the truncated vector obtained by taking the first k components of $\nabla f(c^{(l)})$. Since the nonzero elements of $c^{\text{k-opt}}$ and $c^{(l)}$ are all in first k positions, we get $\nabla f(c^{(l)})^\top(c^{\text{k-opt}} - c^{(l)}) = \nabla\bar{f}(c^{(l)})^\top\Delta\bar{c}$ and $\|c^{\text{k-opt}} - c^{(l)}\|_1 = \|\Delta\bar{c}\|_1$. Applying the strong convexity of f in $D^{\pm}$ again:

$$\begin{aligned}
f(c^{\text{k-opt}}) &\geq f(c^{(l)}) - (-\nabla f(c^{(l)})^\top(c^{\text{k-opt}} - c^{(l)}) - \frac{\mu_1}{2}\|c^{\text{k-opt}} - c^{(l)}\|_1^2) \\
&= f(c^{(l)}) - (-\nabla\bar{f}(c^{(l)})^\top\Delta\bar{c} - \frac{\mu_1}{2}\|\Delta\bar{c}\|_1^2) \\
&\geq f(c^{(l)}) - (\|\nabla\bar{f}(c^{(l)})\|_\infty\|\Delta\bar{c}\|_1 - \frac{\mu_1}{2}\|\Delta\bar{c}\|_1^2) \\
&\geq f(c^{(l)}) - \frac{1}{2\mu_1}\|\nabla\bar{f}(c^{(l)})\|_\infty^2 \\
&= f(c^{(l)}) - \frac{1}{2\mu_1}\max_{1\leq i\leq k}|\nabla_i f(c^{(l)})|^2 \\
&= f(c^{(l)}) - \frac{1}{2\mu_1}(\nabla_{j_l} f(c^{(l)}))^2.
\end{aligned} \tag{4.9}$$

The fourth inequality is obtained similarly by regarding the expression within the parentheses as a quadratic function of $\|\Delta\bar{c}\|_1$.Substituting $f(c^{(l+1)}) \leq f(c^{(l)}) - \frac{1}{2L}(\nabla_{j_l} f(c^{(l)}))^2$ obtained in Lemma 3.5 into the inequality, we get:

$$f(c^{(l+1)}) - f(c^{\text{k-opt}}) \leq (1 - \frac{\mu_1}{L})(f(c^{(l)}) - f(c^{\text{k-opt}})). \tag{4.10}$$

We have got the linear convergence of function value. Since $D^{\pm}$ can be divided into two symmetric regions D^+ and D^-, the limiting point of the iteration must lie in one of them. Without loss of generality, we assume $c^{(l)} \in D^+$, and hence we may choose $c^{\text{k-opt}} \in D^+$ as the convergence target. By the strong convexity of $f(x)$ in D^+, we obtain

$$\begin{aligned}
\|c^{(l)} - c^{\text{k-opt}}\|_2^2 &\leq \frac{2}{\mu_2}\Big(f(c^{(l)}) - f(c^{\text{k-opt}}) - \nabla f(c^{\text{k-opt}})^\top(c^{(l)} - c^{\text{k-opt}})\Big) \\
&= \frac{2}{\mu_2}\big(f(c^{(l)}) - f(c^{\text{k-opt}})\big) \\
&\leq \frac{2}{\mu_2}\Big(1 - \frac{\mu_1}{L}\Big)^l\big(f(c^{(0)}) - f(c^{\text{k-opt}})\big) \;\longrightarrow\; 0.
\end{aligned} \tag{4.11}$$

The second equality holds because $c^{\text{k-opt}}$ is a local minimizer of f restricted to its support set V (i.e., the first k indices). In particular, on the coordinates in V, the gradient vanishes, i.e.,

$$\nabla_i f(c^{\text{k-opt}}) = 0 \quad \text{for all } i \in V.$$

Moreover, all nonzero entries of $(c^{(l)} - c^{\text{k-opt}})$ lie within this support set V. Therefore, the inner product

$$\nabla f(c^{\text{k-opt}})^T (c^{(l)} - c^{\text{k-opt}}) = 0,$$

and it follows that the iteration points converge to $c^{\text{k-opt}}$.

On the other hand, if $c^{(l)} \in D^-$, then by symmetry we may take $-c^{\text{k-opt}} \in D^-$ as the reference point. The same strong convexity argument applied in D^- leads to an identical bound, and hence $\|c^{(l)} \mp c^{\text{k-opt}}\|_2 \to 0$ depending on which region the iterates belong to.

Having presented the linear convergence result for the iterative sequence toward its stationary point $c^{\text{k-opt}}$, we next quantify the error in the estimated eigenvalue obtained from $c^{\text{k-opt}}$ relative to the true eigenvalue.

5. Lemma 3.10's proof.

Proof. We only prove the case for $s = 1$, and the case for $s = 2$ follows similarly.

$$\begin{aligned} \|x\|_1 = \sum_{i=1}^{n} |(x)_i| &\le N\|x\|_\infty + \sum_{j=N+1}^{n} \|x\|_\infty e^{-P(j-N)} \\ &= N\|x\|_\infty + \left(\frac{e^{-P} - e^{-P(n-N+1)}}{1 - e^{-P}} \right) \|x\|_\infty \\ &\le \left(N + \frac{e^{-P}}{1 - e^{-P}} \right) \|x\|_\infty. \end{aligned} \tag{5.1}$$

We let $Q_1 = N + \frac{e^{-P}}{1-e^{-P}}$. For case $s = 2$, we can also get a constant Q_2 similarly. We let $Q = \max\{Q_1, Q_2\}$. ∎

5.1. Lemma 3.11's proof.

Proof. As shown in the proof of Lemma 3.5, the convergence analysis implies that for sufficiently large l,

$$\|\nabla f(c^{(l)})\|_\infty = 4 \left\| Hc^{(l)} + \left(c^{(l)\top} c^{(l)}\right) c^{(l)} \right\|_\infty < 4 n_z \tau.$$

So,

$$\left\| Hc^{(l)} + \left(c^{(l)T} c^{(l)}\right) c^{(l)} \right\|_\infty < n_z \tau.$$

Since for every iterate $c^{(l)}$, we have

$$\left\| Hc^{(l)} + (c^{(l)\top} c^{(l)}) c^{(l)} \right\|_\infty < n_z \tau,$$

and $c^{(l)} \to c^{\text{k-opt}}$ as $l \to \infty$ (as in (4.11)), the same bound holds at the limit. Indeed, the mapping

$$\Phi(x) := Hx + (x^\top x)\, x$$

is continuous on $\mathbb{R}^n$. Therefore

$$\Phi(c^{\text{k-opt}}) = \lim_{l \to \infty} \Phi(c^{(l)}),$$

and taking the ℓ_∞-norm and passing to the limit yields

$$\left\| Hc^{\text{k-opt}} + \left(c^{\text{k-opt}^\top} c^{\text{k-opt}}\right) c^{\text{k-opt}} \right\|_\infty \le n_z \tau.$$

From above we can get:

(5.2)
$$\left|c^\top\left(Hc^{\text{k-opt}} + (c^{\text{k-opt}^\top}c^{\text{k-opt}})c^{\text{k-opt}}\right)\right| \le \|c\|_1\left\|Hc^{\text{k-opt}} + (c^{\text{k-opt}^\top}c^{\text{k-opt}})c^{\text{k-opt}}\right\|_\infty \le n_z\|c\|_1\tau.$$

Because c is the eigenvector of H corresponding to λ, $c^{\text{k-opt}}$ is the eigenvector of H_{11} and $\|c^{\text{k-opt}}\|_2 = -\lambda^{\text{k-opt}}$, we have:

(5.3)
$$\left|\lambda c^\top c^{\text{k-opt}} - \lambda^{\text{k-opt}}c^\top c^{\text{k-opt}}\right| = \left|c^\top Hc^{\text{k-opt}} + (c^{\text{k-opt}^\top}c^{\text{k-opt}})c^\top c^{\text{k-opt}}\right| \le n_z\|c\|_1\tau.$$

By Lemma 3.6 $c^{\text{k-opt}} \in D^{\pm}$, and let us denote

$$d = \frac{1}{30}\frac{\min\left(-2\lambda, -\lambda+\lambda_2\right)}{\sqrt{-\lambda}}. \tag{5.4}$$

We can get $d < \frac{1}{2}\sqrt{-\lambda} = \frac{1}{2}\|c\|_2$ and

$$c^\top c^{\text{k-opt}} \ge c^\top(c - d\frac{c}{\|c\|_2}) = \|c\|_2^2 - d\|c\|_2 > \frac{1}{2}\|c\|_2^2 = -\frac{\lambda}{2}. \tag{5.5}$$

So,

$$|\Delta\lambda| < -\frac{2n_z\|c\|_1\tau}{\lambda} < \frac{2n_zQ\tau}{\sqrt{-\lambda}}. \tag{5.6}$$

Based on the $O(\tau)$ estimate of the eigenvalue error, we further analyze the bound of truncated components of the eigenvector. By controlling their contribution, we are able to derive a sharper $O(\tau^2)$ bound on the eigenvalue approximation error.

5.2. Lemma 3.12's proof.

Proof. Because c is the eigenvector, we have

$$\begin{aligned} H_{11}c_1 + H_{12}c_2 &= \lambda c_1, \\ H_{21}c_1 + H_{22}c_2 &= \lambda c_2. \end{aligned} \tag{5.7}$$

We denote $c - c^{\text{k-opt}} = \Delta c = \begin{pmatrix} \Delta c_1 \\ c_2 \end{pmatrix}$. Using the compression conditions, we have:

$$\begin{aligned} &H_{11}c_1^{\text{k-opt}} = \lambda^{\text{k-opt}}c_1^{\text{k-opt}}, \\ &|(H_{21})_{ij}(c_1^{\text{k-opt}})_j| < \tau; \quad \forall i > k, j \le k. \end{aligned} \tag{5.8}$$

Combining (5.7) and (5.8) we have:

$$\begin{aligned} H_{11}\Delta c_1 + H_{12}c_2 &= \lambda c_1 - \lambda^{\text{k-opt}}c_1^{\text{k-opt}} \\ &= \lambda c_1 - \lambda c_1^{\text{k-opt}} + \lambda c_1^{\text{k-opt}} - \lambda^{\text{k-opt}}c_1^{\text{k-opt}} \\ &= \lambda\Delta c_1 + \Delta\lambda c_1^{\text{k-opt}}. \end{aligned} \tag{5.9}$$

We also have

$$H_{21}\Delta c_1 + H_{21}c_1^{\text{k-opt}} + H_{22}c_2 = H_{21}c_1 + H_{22}c_2 = \lambda c_2. \tag{5.10}$$

These equation can be rearranged as follows:

$$\begin{aligned} H_{11}\Delta c_1 + H_{12}c_2 - \Delta\lambda c^{\text{k-opt}} &= \lambda \Delta c_1, \\ H_{21}\Delta c_1 + H_{22}c_2 - \lambda c_2 &= -H_{21}c_1^{\text{k-opt}}. \end{aligned} \tag{5.11}$$

Suppose that $\Delta c + x$ is an eigenvector of H associated with the eigenvalue λ. Then x satisfies

$$H\Delta c + Hx = \lambda \Delta c + \lambda x. \tag{5.12}$$

Combining (5.11) and (5.12) we have

$$\begin{aligned} (\lambda I - H)x = H\Delta c - \lambda \Delta c &= \begin{pmatrix} H_{11}\Delta c_1 + H_{12}c_2 - \lambda \Delta c_1 \\ H_{21}\Delta c_1 + H_{22}c_2 - \lambda c_2 \end{pmatrix} \\ &= \begin{pmatrix} \Delta\lambda c_1^{\text{k-opt}} \\ -H_{21}c_1^{\text{k-opt}} \end{pmatrix}. \end{aligned} \tag{5.13}$$

The least squares solution of this equation is

$$x = (\lambda I - H)^\dagger \begin{pmatrix} \Delta\lambda c_1^{\text{k-opt}} \\ -H_{21}c_1^{\text{k-opt}} \end{pmatrix}, \tag{5.14}$$

where $(\lambda I - H)^\dagger$ is the pseudo-inverse of $\lambda I - H$. Now we analyze the infinity norm of x. By spanning $H_{21}c_1^{\text{k-opt}}$ we can get

$$\left\| H_{21}c_1^{\text{k-opt}} \right\|_2^2 = \sum_{i=k}^{n} \left(\sum_{j=1}^{k} H_{ij}c_j \right)^2 = \sum_{i=k}^{n} \left(\sum_{j_1,j_2=1}^{k} H_{ij_1}c_{j_1}H_{ij_2}c_{j_2} \right). \tag{5.15}$$

We already have $|H_{ij_1}c_{j_1}|$ and $|H_{ij_2}c_{j_2}| < \tau$, and we need to count there are how many nonzero elements in $H_{ij_1}H_{ij_2}$. First, choose i and j_1, which gives most kn_z possible choices. Then, j_2 has most n_z options because we have fixed row index i, resulting in a total of kn_z^2 choices. So $\|H_{21}\tilde{c}_1\|_2^2 \le kn_z^2\tau^2$. And we already know $\|c^{\text{k-opt}}\|_2^2 = -\lambda^{\text{k-opt}}$, and from Lemma 3.11 we know that $|\Delta\lambda| < -\frac{2n_z\|c\|_1\tau}{\lambda}$. Therefore,

$$\left\| \begin{pmatrix} \Delta\lambda\tilde{c}_1 \\ -H_{21}c_1^{\text{k-opt}} \end{pmatrix} \right\|_2 < m\sqrt{k}\, n_z\tau, \qquad m := \sqrt{\frac{-4\lambda^{\text{k-opt}}\|c\|_1^2}{k\lambda^2} + 1}. \tag{5.16}$$

Consequently,

$$\begin{aligned} \|x\|_\infty &\le \|x\|_2 \\ &\le \left\|(\lambda I - H)^\dagger\right\|_2 \left\| \begin{pmatrix} \Delta\lambda\tilde{c}_1 \\ -H_{21}\tilde{c}_1 \end{pmatrix} \right\|_2 \\ &< \left\|(\lambda I - H)^\dagger\right\|_2 mn_z\sqrt{k}\,\tau \\ &= \frac{mn_z\sqrt{k}\,\tau}{\text{gap}(H)}. \end{aligned} \tag{5.17}$$

We have already assumed that $\Delta c + x$ is an eigenvector of H associated with the eigenvalue λ. Therefore, there exists a scalar $\alpha \in \mathbb{R}$ such that

$$\Delta c + x = \alpha c.$$

We proceed by contradiction. Assume that $\|c_2\|_\infty \geq 2\|x\|_\infty$. Then there exists an index j in the support of c_2 such that

$$|(c_2)_j| = \|c_2\|_\infty.$$

For this index, we have

$$|(c_2)_j + x_j| \geq |(c_2)_j| - |x_j| \geq \frac{1}{2}|(c_2)_j|.$$

On the other hand, since $\Delta c + x = \alpha c$, we also have

$$|(c_2)_j + x_j| = |\alpha|\,|(c_2)_j|,$$

which implies $|\alpha| \geq \frac{1}{2}$. Moreover, since $-\Delta c$ is also a solution of (5.13), we obtain

$$\|x\|_2 \leq \|\Delta c\|_2 < d.$$

Consequently,

$$\|\Delta c + x\|_2 \leq \|\Delta c\|_2 + \|x\|_2 < 2d.$$

However, noting that $\|c\|_2 = \sqrt{-\lambda}$, the relation $\Delta c + x = \alpha c$ together with $|\alpha| \geq \frac{1}{2}$ yields

$$\|\Delta c + x\|_2 = |\alpha|\,\|c\|_2 \geq \frac{1}{2}\sqrt{-\lambda},$$

which contradicts the assumption $2d < \frac{1}{2}\sqrt{-\lambda}$. Therefore, the assumption $\|c_2\|_\infty \geq 2\|x\|_\infty$ is false, and we conclude that

$$\|c_2\|_\infty < 2\|x\|_\infty = \frac{2mn_z\sqrt{k}\,\tau}{\mathrm{gap}(H)}.$$

□

5.3. Theorem 3.13's proof.

Proof. Recall that

$$\lambda c_1 = H_{11}c_1 + H_{12}c_2, \quad \lambda^{\text{k-opt}}c_1^{\text{k-opt}} = H_{11}c_1^{\text{k-opt}}. \tag{5.18}$$

Subtracting the two equations, we have:

$$\lambda c_1 - \lambda^{\text{k-opt}}c_1^{\text{k-opt}} - H_{11}(c_1 - c_1^{\text{k-opt}}) = H_{12}c_2. \tag{5.19}$$

Left-multiplying by ${c^{\text{k-opt}}}^\top$, we get:

$$\lambda {c^{\text{k-opt}}}^\top c_1 - \lambda^{\text{k-opt}}\|c_1^{\text{k-opt}}\|_2^2 - \lambda^{\text{k-opt}}{c_1^{\text{k-opt}}}^\top c_1 + \lambda^{\text{k-opt}}\|c_1^{\text{k-opt}}\|_2^2 = {c_1^{\text{k-opt}}}^\top H_{12}c_2.$$

Rearranging this equation, we obtain

$$\Delta\lambda {c_1^{\text{k-opt}}}^\top c_1 = (\lambda - \lambda^{\text{k-opt}}){c_1^{\text{k-opt}}}^\top c_1 = {c_1^{\text{k-opt}}}^\top H_{12}c_2. \tag{5.20}$$

Now we should analyze the upper bound of ${c_1^{\text{k-opt}}}^\top H_{12}c_2$. Since

$${c_1^{\text{k-opt}}}^\top H_{12}c_2 = \sum_{i=1}^{k}\sum_{j=k}^{n} H_{ij}(c_1^{\text{k-opt}})_i(c_2)_j, \tag{5.21}$$

and

$$|H_{ij}(c_1^{\text{k-opt}})_i| < \tau, \quad \sum_{i=1}^{k} |H_{ij}(c_1^{\text{k-opt}})_i| < n_z\tau,$$

we have

$$|{c_1^{\text{k-opt}}}^\top H_{12} c_2| < \sum_{j=k}^{n} |n_z\tau||(c_2)_i|. \tag{5.22}$$

Since c_2 is a subvector of c, it inherits the exponential decay property of c and has the same exponential decay coefficient. Using the method similar to Lemma 3.10, we can derive

$$\sum_{j=k}^{n} |(c_2)_i| < Q\|c_2\|_\infty = \frac{2mn_z\sqrt{k}Q\tau}{\text{gap}(H)}, \tag{5.23}$$

and

$$|{c_1^{\text{k-opt}}}^\top H_{12} c_2| < \frac{2mn_z^2\sqrt{k}Q\tau^2}{\text{gap}(H)}. \tag{5.24}$$

Moreover, from the estimate $d < \frac{1}{2}\|c\|_2$ established in the proof of Lemma 3.11, it follows that

$${c^{\text{k-opt}}}^\top c \geq \|c\|_2^2 - d\|c\|_2 > \tfrac{1}{2}\|c\|_2^2 = -\tfrac{\lambda}{2}.$$

From (5.20), we obtain:

$$\Delta\lambda < \frac{4mn_z^2\sqrt{k}Q\tau^2}{-\lambda\,\text{gap}(H)}. \tag{5.25}$$

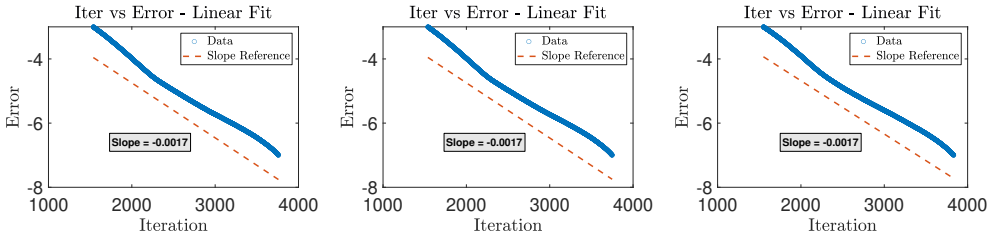
□

6. Numerical Results. In this section, we present numerical experiments to support our theoretical findings and demonstrate the practical performance of the compressed CDFCI algorithm. The experiments are organized into three parts. First, we verify the linear convergence using synthetic sparse matrices with exponentially decaying eigenvectors. We demonstrate that the convergence rate of the compressed algorithm matches that of the original uncompressed method. Second, we validate the error bound derived in the previous section. The results confirm that the numerical error remains consistently below the theoretical bound. Third, we apply the algorithm to realistic quantum chemistry systems. Exploiting the $O(\tau^2)$ scaling of the error, we use extrapolation techniques on systems such as H_2O and C_2 (with the cc-pVDZ basis set) to obtain improved estimates of the ground-state energy.

6.1. Verification of Linear Convergence. To construct test matrices with controlled spectral and sparsity properties, we first generate a set of orthonormal eigenvectors, ensuring that the first eigenvector (associated with the smallest eigenvalue) exhibits exponential decay. Random eigenvalues are then assigned to these eigenvectors, and the matrix is constructed as $A = V\Lambda V^\top$. To enforce sparsity, we truncate entries with small magnitudes in the resulting matrix. After sparsification, we verify whether the first eigenvector still maintains exponential decay. If not, we re-impose this structure by element-wise multiplication with an exponential decay factor and repeat the construction procedure until the desired properties are satisfied.

For all experiments, the initial iterate is chosen as the target eigenvector perturbed by a random noise vector. This initialization ensures that the iterates start in a neighborhood of the desired solution while avoiding the trivial case of starting from the exact eigenvector.

Figure 2 displays the convergence behavior of CDFCI under three different compression thresholds: $\tau = 0$ (uncompressed), 10^{-5}, and 10^{-7}. The horizontal axis represents the iteration count, and the vertical axis plots the logarithm of the objective gap, $\log(f(c^{(l)}) - f(c^{\text{k-opt}}))$. In each subplot, the red line represents a linear fit, highlighting the linear convergence rate.



(a) $\tau = 0$ (b) $\tau = 10^{-5}$ (c) $\tau = 10^{-7}$

Fig. 2: Convergence behavior under different compression thresholds. The horizontal axis represents the iteration number, and the vertical axis plots the logarithmic objective gap $\log(f(c^{(l)}) - f(c^{\text{k-opt}}))$. The red lines indicate the linear fit.

As shown in Figure 2, the algorithm maintains linear convergence even when compression is applied. Moreover, the observed convergence rates for the compressed cases are nearly identical to the uncompressed version, indicating that the compression strategy does not compromise the convergence speed.

6.2. Verification of the Error Bound. In this subsection, we investigate the tightness of the theoretical error bound derived in Section 3. We first consider the same synthetic matrix class used in the previous subsection, referring to it as **Type 1**. To further evaluate the error bound under different sparsity patterns, we introduce a second class of matrices, referred to as **Type 2**.

A Type 2 matrix consists of two main blocks: a diagonal block in the top-left and a tridiagonal block in the bottom-right, coupled by a small number of off-diagonal entries. This structure preserves the exponential decay of the first eigenvector while significantly increasing the overall sparsity of the matrix. Figure 3 illustrates the sparsity pattern of a representative Type 2 matrix, where the block structure and coupling entries are clearly visible.

Figure 4 presents a comparison between the actual Rayleigh quotient error and the theoretical bound for both matrix types. In both cases, the actual error remains consistently below the predicted bound and exhibits a similar decay pattern as τ decreases. Notably, the Type 2 matrix yields a significantly tighter bound: the typical ratio between the bound and the true error ranges from 10^1 to 10^2, whereas for the Type 1 matrix, this ratio ranges from 10^2 to 10^3.

Next, we analyze the compression error and the theoretical error bound for the H_2O molecule using the cc-pVDZ basis set. The behavior observed in Figure 5 aligns with the results from our synthetic test matrices: the actual error consistently satisfies

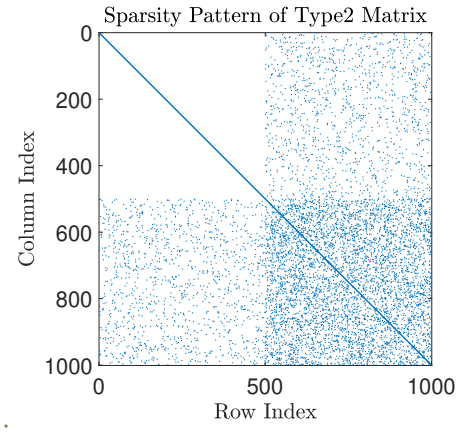
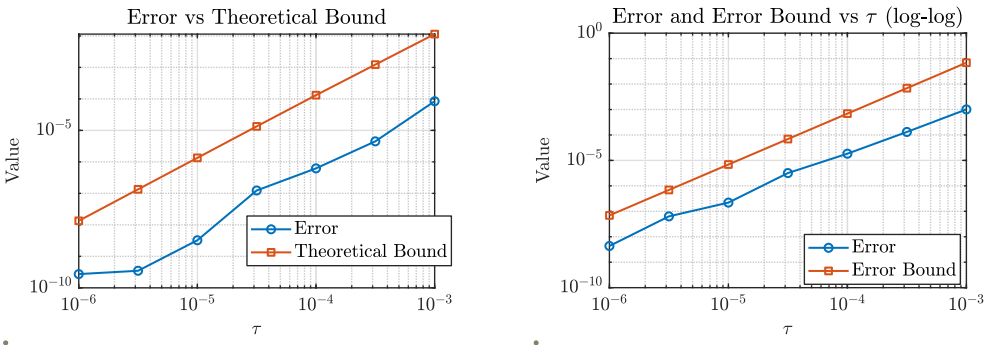


Fig. 3: Sparsity pattern of a representative Type 2 matrix.



(a) Type 1 matrix

(b) Type 2 matrix

Fig. 4: Comparison of the actual error and the theoretical bound under varying τ for the two types of synthetic matrices.

the theoretical bound and exhibits a similar decay pattern as τ decreases. However, the ratio between the bound and the true error is larger, typically ranging from 10^3 to 10^4. This looseness in the bound may be attributed to the extremely high sparsity characteristic of FCI Hamiltonians in quantum chemistry problems.

6.3. Extrapolation of Ground-State Energy via Compression Error. In this subsection, we demonstrate how to leverage the $O(\tau^2)$ quadratic dependence of the energy error on the compression threshold to obtain improved estimates of the ground-state energy. Even when the exact Full Configuration Interaction (FCI) solution is computationally inaccessible, this scaling law enables us to extrapolate the ground-state energy by fitting results at different values of τ.

The extrapolation scheme is based on the theoretical finding that the compression error scales quadratically with τ. Specifically, we adopt the following model (ansatz):

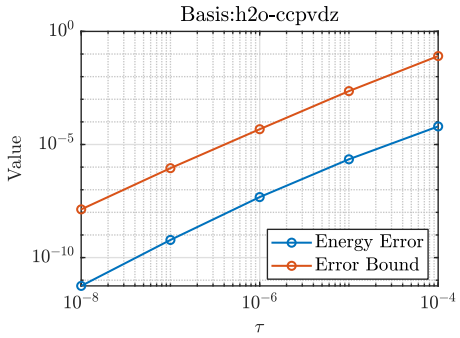


Fig. 5: Compression error and theoretical error bound for H_2O (cc-pVDZ).

$$E(\tau) = E_0 + a\tau^2, \tag{6.1}$$

where $E(\tau)$ denotes the computed ground-state energy at threshold τ, E_0 represents the extrapolated true energy (corresponding to the limit $\tau \to 0$), and a is a fitting coefficient.

To reduce the impact of ill-conditioning, we rewrite the model in logarithmic form:

$$\log(E(\tau) - E_0) = \log(a) + 2\log(\tau). \tag{6.2}$$

Given this relation, we employ an iterative procedure: we first select a preliminary estimate for E_0, and then solve a linear least-squares problem to find the best-fit parameter a. Subsequently, we refine E_0 via a bisection search to minimize the fitting residual in the logarithmic domain. This process is repeated until convergence, yielding an accurate extrapolated estimate of the true ground-state energy.

We apply this extrapolation strategy to quantum chemistry systems using standard basis sets, specifically H_2O and C_2 with cc-pVDZ. To evaluate the robustness of the method, we consider two distinct threshold regimes: a high-precision range (10^{-8}–10^{-7}) and a low-precision range (10^{-5}–10^{-4}). In each range, we uniformly sample 10 values of τ. This allows us to examine the extrapolation accuracy in both high- and low-precision regimes.

The results demonstrate that extrapolation yields significant accuracy improvements when the original data is in the high-precision regime (small τ). However, for data with lower precision (large τ), the improvement is less pronounced. This is expected, as the error asymptotically approaches the quadratic scaling $O(\tau^2)$ only as $\tau \to 0$. Consequently, the quadratic model provides a much better fit in the small τ regime. Developing extrapolation techniques that remain effective for larger τ values represents an important direction for future research.

7. Conclusion. In this work, we established a theoretical framework for the CDFCI with compression algorithm. Under mild assumptions on the spectral gap and exponential decay of the ground-state eigenvector, we proved that the algorithm maintains linear convergence toward the compressed stationary point and derived an error bound of order $O(\tau^2)$ for the Rayleigh quotient. Numerical experiments on both

Table 2: Comparison of extrapolation performance under different compression threshold regimes.

System	Min. Error (Computed)	Error (Extrapolated)
Small τ regime (10^{-8} to 10^{-7})		
H_2O (cc-pVDZ)	5.6×10^{-12}	1.1×10^{-12}
C_2 (cc-pVDZ)	2.0×10^{-10}	2.6×10^{-11}
Large τ regime (10^{-5} to 10^{-4})		
H_2O (cc-pVDZ)	2.2×10^{-6}	7.0×10^{-7}
C_2 (cc-pVDZ)	6.5×10^{-5}	3.4×10^{-5}

synthetic sparse matrices and realistic quantum chemistry systems confirm the theoretical predictions: the algorithm exhibits consistent linear convergence and achieves energy errors in agreement with the $O(\tau^2)$ bound. Furthermore, by exploiting the theoretically characterized dependence of the error on τ^2, we proposed an extrapolation scheme that yields significantly improved energy estimates, demonstrating the practical value of the derived error analysis.

Acknowledgments. The authors acknowledge Jianfeng Lu and Zhe Wang for their contributions to the original development of the CDFCI method. This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 12271109 and 12526211; the Shanghai Pilot Program for Basic Research–Fudan University under Grant No. 21TQ1400100 (22TQ017); the Scientific Research Innovation Capability Support Project for Young Faculty under Grant No. SRICSPYF-ZY2025159; and the Xuemin Institute of Advanced Studies, Fudan University.

REFERENCES

[1] G. H. Booth, A. Grüneis, G. Kresse, and A. Alavi, *Towards an exact description of electronic wavefunctions in real solids*, Nature, 493 (2013), pp. 365–370.

[2] G. H. Booth, A. J. Thom, and A. Alavi, *Fermion monte carlo without fixed nodes: A game of life, death, and annihilation in slater determinant space*, The Journal of chemical physics, 131 (2009).

[3] A. A. Casulli, D. Kressner, and N. Shao, *Lanczos with compression for symmetric eigenvalue problems*, arXiv preprint arXiv:2602.20948, (2026).

[4] X. S. Chen, W. Li, and W. W. Xu, *PERTURBATION ANALYSIS OF THE EIGENVECTOR MATRIX AND SINGULAR VECTOR MATRICES*, Taiwanese Journal of Mathematics, 16 (2012), https://doi.org/10.11650/twjm/1500406535.

[5] A. d'Aspremont, L. Ghaoui, M. Jordan, and G. Lanckriet, *A direct formulation for sparse pca using semidefinite programming*, Advances in neural information processing systems, 17 (2004).

[6] E. R. Davidson, *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices*, Journal of Computational Physics, 17 (1975), pp. 87–94, https://doi.org/https://doi.org/10.1016/0021-9991(75)90065-0, https://www.sciencedirect.com/science/article/pii/0021999175900650.

[7] J. Fan, W. Wang, and Y. Zhong, *An ∞ Eigenvector Perturbation Bound and Its Application to Robust Covariance Estimation*.

[8] S. M. Greene, R. J. Webber, T. C. Berkelbach, and J. Weare, *Approximating matrix eigenvalues by subspace iteration with repeated random sparsification*, SIAM Journal on Scientific Computing, 44 (2022), pp. A3067–A3097.

[9] A. A. Holmes, N. M. Tubman, and C. J. Umrigar, *Heat-bath configuration interaction:*

An efficient selected configuration interaction algorithm inspired by heat-bath sampling, Journal of chemical theory and computation, 12 (2016), pp. 3674–3680.

[10] B. Huron, J. Malrieu, and P. Rancurel, *Iterative perturbation calculations of ground and excited state energies from multiconfigurational zeroth-order wavefunctions*, The Journal of Chemical Physics, 58 (1973), pp. 5745–5759.

[11] Z. Jia and Z. Wang, *A convergence analysis of the inexact Rayleigh quotient iteration and simplified Jacobi-Davidson method for the large Hermitian matrix eigenproblem*, Science in China Series A: Mathematics, 51 (2008), https://doi.org/10.1007/s11425-008-0050-y.

[12] Y. Li, J. Lu, and Z. Wang, *Coordinatewise descent methods for leading eigenvalue problem*, SIAM Journal on Scientific Computing, 41 (2019), pp. A2681–A2716.

[13] H. Liu and A. Aravkin, *Analysis of Truncated Orthogonal Iteration for Sparse Eigenvector Problems*, Mar. 2021, https://doi.org/10.48550/arXiv.2103.13523, https://arxiv.org/abs/2103.13523.

[14] J. Lu and Z. Wang, *The full configuration interaction quantum monte carlo method through the lens of inexact power iteration*, SIAM Journal on Scientific Computing, 42 (2020), pp. B1–B29.

[15] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization, 22 (2012), pp. 341–362.

[16] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming, 144 (2014), pp. 1–38.

[17] A. Saad-Falcon, B. Ancelin, and J. Romberg, *Global convergence of adaptive sensing for principal eigenvector estimation*, arXiv preprint arXiv:2505.10882, (2025).

[18] Z. Wang, Y. Li, and J. Lu, *Coordinate descent full configuration interaction*, Journal of chemical theory and computation, 15 (2019), pp. 3558–3569.

[19] Z. Wang, Z. Zhang, J. Lu, and Y. Li, *Coordinate descent full configuration interaction for excited states*, Journal of Chemical Theory and Computation, 19 (2023), pp. 7731–7739.

[20] S. J. Wright, *Coordinate descent algorithms*, Mathematical programming, 151 (2015), pp. 3–34.

[21] X.-T. Yuan and T. Zhang, *Truncated power method for sparse eigenvalue problems.*, Journal of Machine Learning Research, 14 (2013).

[22] Y. Zhang, W. Gao, and Y. Li, *Parallel multicoordinate descent methods for full configuration interaction*, Journal of Chemical Theory and Computation, 21 (2025), pp. 2325–2337.