SIAM J. SCI. COMPUT.
Vol. 46, No. 1, pp. A179–A203

WEIGHTED TRACE-PENALTY MINIMIZATION FOR FULL CONFIGURATION INTERACTION*

WEIGUO GAO†, YINGZHOU LI‡, AND HANXIANG SHEN§

Abstract. A novel unconstrained optimization model named weighted trace-penalty minimization (WTPM) is proposed to address the extreme eigenvalue problem arising from the full configuration interaction (FCI) method. Theoretical analysis shows that the global minimizers of the WTPM objective function are the desired eigenvectors, rather than the eigenspace. Analyzing the condition number of the Hessian operator in detail contributes to the determination of a near-optimal weight matrix. With the sparse feature of FCI matrices in mind, the coordinate descent (CD) method is adapted to WTPM and results in the WTPM-CD method. The reduction of computational and storage costs in each iteration shows the efficiency of the proposed algorithm. Finally, the numerical experiments demonstrate the capability to address large-scale FCI matrices.

Key words. eigensolver, weighted trace-penalty minimization, full configuration interaction, coordinate descent method

MSC code. 65F15

DOI. 10.1137/23M1547676

1. Introduction.

The time-independent, nonrelativistic Schrödinger equation is a linear Hermitian extreme eigenvalue problem,

$$H|\psi_i\rangle = E_i|\psi_i\rangle, \quad i = 1, 2, \ldots, p, \tag{1.1}$$

where H is a Hamiltonian operator, $(E_i, |\psi_i\rangle)$ denotes the ground-state and lowest few excited-state energies and their corresponding wavefunctions, and p is the number of desired eigenvalues. Efficiently solving the Schrödinger equation plays a fundamental role in the field of electronic structure calculation. The problem is of high-dimensionality, i.e., $|\psi\rangle = \psi(x_1, \ldots, x_{N_e})$ for $x_i \in \mathbb{R}^3$ being the position of electrons and N_e being the number of active electrons in the system. Further, identical electron wavefunctions admit an antisymmetry property, which corresponds to the Pauli exclusion principle. Full configuration interaction (FCI) is a variational method that solves (1.1) numerically exactly within the space of all Slater determinants [21, 22, 38]. By the nature of Slater determinants, the antisymmetry property is encoded into the many-body basis functions. Nevertheless, the high-dimensionality feature still leads to extremely large Hamiltonian matrices. More precisely, the size of

*Submitted to the journal's Numerical Algorithms for Scientific Computing section January 17, 2023; accepted for publication (in revised form) October 6, 2023; published electronically January 18, 2024.

https://doi.org/10.1137/23M1547676

Funding: The first author was partially supported by the National Key R&D Program of China under grant 2020YFA0711902 and by the National Natural Science Foundation of China under grant 71991471. The second and third authors were partially supported by the National Natural Science Foundation of China under grant 12271109 and by the Science and Technology Commission of Shanghai Municipality under grant 22TQ017.

†School of Mathematical Sciences and School of Data Science, Fudan University, Shanghai, 200433, China, and Shanghai Artificial Intelligence Laboratory, Shanghai, 200232, China (wggao@fudan.edu.cn).

‡School of Mathematical Sciences, Fudan University, Shanghai, 200433, China (yingzhouli@fudan.edu.cn).

§Shanghai Center for Mathematical Science, Fudan University, Shanghai, 200438, China (hxshen19@fudan.edu.cn).

Hamiltonian matrices grows factorially with respect to the number of electrons and orbitals in the system. For a system with N_{orb} orbitals and N_{e} electrons, the number of Slater determinants is $O(\binom{N_{\text{orb}}}{N_{\text{e}}})$ [23]. A single H_2O molecule with cc-pVDZ basis (24 orbitals) and 10 active electrons leads to a $4.53\cdot 10^8 \times 4.53\cdot 10^8$ Hamiltonian matrix [43]. The antisymmetry property of the problems leads to the notorious sign problem.

In electronic structure calculation, widely used eigensolvers can be classified into two groups: Krylov subspace methods and optimization methods. Krylov subspace methods include the Chebyshev–Davidson algorithm [47], the locally optimal block preconditioned conjugate gradient method [24], the block Krylov–Schur algorithm [48], the projected preconditioned conjugate gradient algorithm [42], etc. In all these methods, an explicit orthogonalization step is carried out every few iterations, which costs a significant amount of computational resources throughout the algorithm. The other group, optimization methods, transforms the eigenvalue problem into an optimization problem with or without the orthogonality constraint. The orthogonality constrained optimization problem is also known as the Stiefel manifold optimization. The corresponding optimization algorithms [1, 14, 19, 36] require either a projection or a retraction step to keep the iteration on the manifold. The computational costs for these projections and retractions remain the same as that of the orthogonalization. Solving the eigenvalue problem via an unconstrained optimization is popular recently, especially when the large-scale eigenvalue problems are considered. Such methods include the symmetric low-rank product model [27, 29, 43], the orbital minimization method [13, 30], trace-penalty minimization [45], etc. All these methods converge to the invariant subspace corresponding to the smallest eigenvalues and require a single Rayleigh–Ritz procedure to extract eigenvectors from the invariant subspace.

However, due to the large-scale matrix size, none of the eigensolvers mentioned above can be applied directly to the FCI eigenvalue problem. Almost all of them run into the memory bottleneck. Many nonstandard eigensolvers are designed particularly for FCI ground-state computation. The density matrix renormalization group (DMRG) [9, 33, 46] approximates the high-dimensional wavefunctions by a tensor train. The underlying eigenvalue problem is solved by the vanilla power method. FCI quantum Monte Carlo (FCIQMC) [6, 7, 31] adopts the quantum Monte Carlo method to overcome the sign problem and the curse of dimensionality. Related methods [11, 35] in this family further adjust the variance-bias trade-off to obtain efficient algorithms. Selected configuration interaction (SCI) [18, 20, 37, 41] employs perturbation analysis of eigenvalue problems and selects important configurations accordingly. Then the eigenvalue problem of the selected principal submatrix is solved via traditional eigensolvers. Various methods in this family differ from each other in the computationally efficient approximations to the perturbation result. Coordinate descent FCI (CDFCI) [27, 43] applies the coordinate descent method on the symmetric low-rank product model to select important configurations and to update the coefficients. A carefully designed compression scheme is incorporated to overcome the memory bottleneck.

Furthermore, the nonstandard eigensolvers mentioned above can be adapted to the low-lying excited-states computation. DMRG [2] and FCIQMC [5] adopt the deflation idea and compute excited states one by one. SCI [37, 41] needs to select many more configurations to capture the important configurations for excited states. CDFCI [44] could be naturally extended and converge to the invariant subspace formed by the ground state and excited states. The extra rotation matrix within the invariant subspace leads to less sparse iteration variables and increases the memory

cost. Specifically targeting the low-lying excited-states computation, the triangularized orthogonalization free method [15, 16] proposes a triangularized iterative scheme converging to eigenvectors directly without any projection or Rayleigh–Ritz step.

In this paper, inspired by the weighted subspace-search variational quantum eigensolver [32] from quantum computing and the trace-penalty model [45], we propose an unconstrained optimization model called weighted trace-penalty minimization (WTPM),

$$\min_{X\in\mathbb{R}^{n\times p}} f_{\mu,W}(X) = \frac{1}{2}\mathrm{tr}(X^\top AX) + \frac{\mu}{4}\|X^\top X - W\|_{\mathrm{F}}^2, \tag{1.2}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $\mu > 0$ is the penalty parameter, and W is a diagonal weight matrix with distinct diagonal entries. Our analysis shows that (1.2) does not have spurious local minima, and the 2^p isolated global minima of (1.2) are scaled eigenvectors corresponding to the smallest p eigenvalues of A. Moreover, we calculate the condition number of the W-dependent Hessian matrix of $f_{\mu,W}(X)$ at global minima, which leads to the local convergence rate for the first-order method. A near-optimal weight matrix W is then derived to achieve a fast local convergence rate.

With the size of the FCI problem size in mind, we focus on first-order methods to address (1.2). One choice is the gradient descent method with Barzilai–Borwein (BB) stepsize [3]. A more desirable choice is the coordinate descent method, which reveals the sparsity in FCI problems efficiently [43]. Hence, we tailor the coordinate descent method for (1.2) and obtain an efficient eigensolver for the FCI eigenvalue problem. Global convergence of the proposed eigensolver can be proved in the same way as that of CDFCI [27], whereas the local linear convergence is obtained with a rate related to the Hessian operator. We emphasize that both gradient descent and coordinate descent methods for (1.2) converge to the scaled eigenvectors directly. Hence the expensive and parallel inefficient Rayleigh–Ritz step is omitted entirely from both methods.

Numerically, we test and compare the performance of the original trace-penalty model [45] and our WTPM on FCI matrices from practice. The numerical results show that both models with first-order methods converge to desired global minima. Adding the extra weight matrix in WTPM does not destroy the efficiency of the original trace-penalty model. For the FCI matrices, the WTPM-CD method converges in far less cost of flops. WTPM-CD finds the sparse representations of the eigenvectors corresponding to the smallest p eigenvalues, whereas the trace-penalty model requires an extra Rayleigh–Ritz step.

The rest of this paper is organized as follows. In section 2, we give a theoretical analysis on (1.2) with a focus on the global minima and the condition number of the Hessian operator. Section 3 proposes algorithms for (1.2) and analyzes their performance and complexity. Section 4 reports the numerical results showing the efficiency of our method. Finally, we conclude this paper in section 5 with some discussion on future work.

2. Model analysis. This section focuses on the analysis of the energy landscape of the WTPM model. We will first analyze the stationary points and Hessian operator of (1.2) in sections 2.1 and 2.2, respectively. Based on the condition number of the Hessian operator at the global minimum, a near-optimal choice of the weight matrix W is discussed in section 2.3 to achieve a near-optimal local convergence rate for first-order methods. Finally in section 2.4, we discuss the extensions of all analysis results to Hermitian matrices.

We consider a real symmetric matrix A of size $n \times n$. The eigenvalue decomposition of A is denoted as

$$A = V\Lambda V^\top, \tag{2.1}$$

where $V = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^{n\times n}$ and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ such that

$$\lambda_1 < \lambda_2 < \cdots < \lambda_p < \lambda_{p+1} \leqslant \cdots \leqslant \lambda_n. \tag{2.2}$$

A pair (λ_i, v_i) is an eigenpair of A. Throughout this paper, we aim to compute the smallest p eigenpairs of A via WTPM. The real weight matrix in (1.2) satisfies the following assumption.

Assumption 1. The weight matrix is diagonal, $W = \operatorname{diag}(w_1, w_2, \ldots, w_p)$ such that

$$w_1 > w_2 > \cdots > w_p > \frac{\lambda_p}{\mu}. \tag{2.3}$$

2.1. Stationary points. Stationary points of (1.2) satisfy the first-order necessary optimal condition,

$$\nabla f_{\mu,W}(X) = AX + \mu X(X^\top X - W) = 0. \tag{2.4}$$

Left multiplying both sides of (2.4) by the transpose of X, we obtain

$$X^\top AX = \mu X^\top X(W - X^\top X), \tag{2.5}$$

where the left-hand side is symmetric. The symmetry property of (2.5) leads to the fact that $X^\top XW = WX^\top X$. Since W is a diagonal matrix and all entries are distinct as in (2.3), the equation $X^\top XW = WX^\top X$ implies that $X^\top X$ is diagonal, i.e., columns of stationary point X are either zero or mutually orthogonal. In Theorem 2.1, we give the explicit form of the stationary points of (1.2), where each column of X is either an eigenvector of A or the zero vector.

THEOREM 2.1 (stationary points). *Assume A and W satisfy* (2.1) *and* (2.3), *respectively. Any stationary point $\widehat{X}$ of* (1.2) *has the form*

$$\widehat{X} = \widehat{U}_p \widehat{S}_p, \tag{2.6}$$

where $\widehat{U}_p = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_p)$ and $\widehat{S}_p = \operatorname{diag}(\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_p)$ such that

$$\begin{aligned} &A\hat{u}_i = \sigma_i \hat{u}_i, \quad \hat{u}_i^\top \hat{u}_j = \delta_{ij}, \text{ and} \\ &\hat{s}_i \in \left\{0, \sqrt{w_i - \frac{\sigma_i}{\mu}}\right\}. \end{aligned} \tag{2.7}$$

Proof. As we discussed above, $\widehat{X}^\top \widehat{X}$ is a diagonal matrix. Columns of (2.4), then, can be represented as

$$A\hat{x}_i = d_i \hat{x}_i, \quad i = 1, 2, \ldots, p, \tag{2.8}$$

where $\hat{x}_i$ is the ith column of $\widehat{X}$ and $d_i = \mu(w_i - \hat{x}_i^\top \hat{x}_i)$. Vector $\hat{x}_i$ is either a zero vector or an eigenvector of A. When $\hat{x}_i$ is a zero vector, it can be represented as the form in the theorem for $\hat{s}_i = 0$. When $\hat{x}_i$ is not a zero vector, we denote $\hat{x}_i = \hat{u}_i \hat{s}_i$ for

$\hat{u}_i$ being a unit length eigenvector of A associated with eigenvalue σ_i and $\hat{s}_i$ being a positive scalar. Then $\hat{s}_i$ satisfies

$$\sigma_i = d_i = \mu(w_i - \hat{s}_i^2) \Rightarrow \hat{s}_i = \sqrt{w_i - \frac{\sigma_i}{\mu}}.$$

Recall that $\widehat{X}^\top \widehat{X}$ is a diagonal matrix. For nonzero $\hat{s}_i$ and $\hat{s}_j$, it is required that $\hat{u}_i^\top \hat{u}_j = \delta_{ij}$. When $\hat{s}_i$ is zero, we could always find an extra orthogonal eigenvector of A, such that $\hat{u}_i^\top \hat{u}_j = \delta_{ij}$ still holds. Therefore, we proved that the stationary points of (1.2) must admit the form in the theorem and any point that admits the form in the theorem is a stationary point. ∎

Furthermore, we will distinguish the local and global minima from saddle points. Interestingly, it can be shown that the columns of global minimizer X^* are eigenvectors of A associated with the p smallest eigenvalues.

THEOREM 2.2 (global minima). *Assume A and W satisfy* (2.1) *and* (2.3). *Any global minimizer X^* of* (1.2) *has the form*

$$X^* = V_p S_p, \tag{2.9}$$

where $V_p = (v_1, v_2, \ldots, v_p)$ and $S_p = \mathrm{diag}(\pm s_1, \pm s_2, \ldots, \pm s_p)$ such that

$$s_i = \sqrt{w_i - \frac{\lambda_i}{\mu}}.$$

Proof. By Theorem 2.1, the stationary point has the form $\widehat{X} = \widehat{U}_p \widehat{S}_p$. Substituting it into the objective function leads to

$$2f_{\mu,W}(\widehat{X}) = \sum_{i=1}^{p} \left[\hat{s}_i^2 \sigma_i + \frac{\mu}{2}(\hat{s}_i^2 - w_i)^2\right] = \sum_{i=1}^{p} \frac{\mu}{2} w_i^2 - \sum_{i \in \{\hat{s}_i \neq 0\}} \frac{(\sigma_i - \mu w_i)^2}{2\mu},$$

where σ_i is one of the eigenvalues of A and the second equality is due to the expression of $\hat{s}_i$. If some $\hat{s}_i = 0$, then we have $\sigma_i - \mu w_i = 0$. Under Assumption 1, there are at least p eigenvalues to make $-\frac{(\lambda - \mu w_i)^2}{2\mu} < 0$ and at least one of them is not in $\{\sigma_i\}_{i=1}^p$. Replacing σ_i by one of the unselected eigenvalue with $-\frac{(\lambda - \mu w_i)^2}{2\mu} < 0$ would lead to a smaller objective function value. Hence if $\hat{s}_i = 0$ for some i, the stationary point is not a global minimizer. Notice that if $\hat{s}_i \neq 0$, then $\sigma_i < \mu w_i$. We only need to show that $\sigma_i = \lambda_i$.

First, we claim that $\sigma_1, \sigma_2, \ldots, \sigma_p$ must be a permutation of $\lambda_1, \lambda_2, \ldots, \lambda_p$. If not, for example, there exists σ_{i_0} such that $\mu w_{i_0} > \sigma_{i_0} > \lambda_p$. There also exists λ_j not used such that $\lambda_j \leqslant \lambda_p$. If $\sigma_{i_0} = \lambda_j$ instead, the value of $f_{\mu,W}(\widehat{X})$ will decrease. This contradicts the minimal property so our claim holds.

Next, since there are at least p eigenvalues smaller than any μw_i due to Assumption 1, the optimization (1.2) changes into the form as

$$\max_{\substack{\sigma_i \in \{\lambda_1, \ldots, \lambda_p\} \\ \sigma_i \neq \sigma_j \text{ if } i \neq j}} \sum_{i=1}^{p} (\sigma_i - \mu w_i)^2 = \sum_{i=1}^{p} -2\mu \sigma_i w_i + \text{const.} \tag{2.10}$$

According to the rearrangement inequality [17], we can conclude that the minimum value is reached if $\sigma_i = \lambda_i$, i.e.,

$$X^* = V_p S_p.$$

∎

Theorem 2.2 presents an interesting result that any global minimizer of the WTPM (1.2) is composed of desired eigenvectors directly rather than the underlying invariant subspace.

Another result is that there are no spurious local minima, i.e., the rest of the stationary points are all strict saddle points. To show this, we first introduce the Hessian operator of $f_{\mu,W}$,

$$\operatorname{Hess} f_{\mu,W}(X)[Z] = AZ + \mu(ZX^\top X + XZ^\top X + XX^\top Z - ZW), \tag{2.11}$$

where $Z \in \mathbb{R}^{n\times p}$ is an arbitrary matrix. A stationary point $\hat{X}$ is called a strict saddle point of $f_{\mu,W}$ if and only if the Hessian operator $\operatorname{Hess} f_{\mu,W}(\hat{X})$ has negative eigenvalues.[1]

In Theorem 2.3, we prove that (1.2) does not have any spurious local minimum.

THEOREM 2.3. *Assume A and W satisfy* (2.1) *and* (2.3). *There are no local minima other than the global minimizers X^*.*

Proof. First, we discuss the case that the stationary point $\widehat{X} = \widehat{U}_p\widehat{S}_p$ is of full column rank, i.e., $\hat{s}_i \neq 0\ \forall i$. We focus on the sign of

$$\begin{aligned}\operatorname{tr}\Big(Z^\top \operatorname{Hess} f_{\mu,W}(\hat{X})[Z]\Big)\\ = \operatorname{tr}\Big(Z^\top(A+\mu\hat{X}\hat{X}^\top)Z + \mu Z^\top Z(\hat{X}^\top\hat{X} - W) + \mu Z^\top\hat{X}Z^\top\hat{X}\Big)\end{aligned} \tag{2.12}$$

for $Z \in \mathbb{R}^{n\times p}$. If (2.12) at $\hat{X}$ is strictly negative for a particular Z, then we could conclude that the Hessian operator has negative eigenvalues and hence $\hat{X}$ is a strict saddle point.

Since $\hat{X}$ is not a global minimizer, there exists an eigenvector v_i in $\{v_1, v_2, \ldots, v_p\}$ satisfying $v_i^\top \hat{U}_p = 0$ and $\exists j$, $\lambda_i < \sigma_j$. Let Z be the matrix whose jth column is v_i and others are zero. Then we have

$$\operatorname{tr}\Big(Z^\top \operatorname{Hess} f_{\mu,W}(\hat{X})[Z]\Big) = \lambda_i - \sigma_j < 0, \tag{2.13}$$

implying that full-rank stationary points are all saddle points except global minima.

Next, we discuss the rank-deficient case. Without loss of generality we assume the jth column of $\hat{X}$ is zero, and there also exists an eigenvector v_i in $\{v_1, v_2, \ldots, v_p\}$ satisfying $v_i^\top \hat{X} = 0$. Let the jth column of Z be v_i and the others are zero. We can obtain

$$\operatorname{tr}\Big(Z^\top \operatorname{Hess} f_{\mu,W}(\hat{X})[Z]\Big) = \lambda_i - \mu w_j < 0. \tag{2.14}$$

Hence, rank-deficient stationary points are all saddle points. ∎

2.2. Hessian operator. For first-order optimization methods, the local convergence rate relies on the condition number of the Hessian operator [8]. Specifically, in the neighborhood of a global minimum X^*, the error of a gradient descent method with exact line search decreases linearly as

$$f_{\mu,W}(X^{(j+1)}) - f_{\mu,W}(X^*) \leqslant \big(1-\kappa^{-1}\big)\big(f_{\mu,W}(X^{(j)}) - f_{\mu,W}(X^*)\big), \tag{2.15}$$

[1] Here the strict saddle point includes the maximizer of the problem, i.e., a stationary point with negative semidefinite Hessian operator.

where $X^{(j)}$ denotes the iteration variable at the jth iteration and κ denotes the condition number of the Hessian operator at X^*.

From Theorem 2.2, we find that all global minima of (1.2) are isolated points and the Hessian operator at any global minimizer X^* is strictly positive definite. In order to give a depiction of the local convergence rate of (1.2), we present a tight estimation of the condition number of the Hessian operator in Theorem 2.4.

THEOREM 2.4. *Assume A and W satisfy* (2.1) *and* (2.3). *Let X^* be a global minimizer of $f_{\mu,W}$. Then*

$$\kappa\left(\operatorname{Hess} f_{\mu,W}(X^*)\right) \triangleq \frac{\max\limits_{\|Z\|_{\mathrm{F}}=1} \operatorname{tr}\left(Z^\top \operatorname{Hess} f_{\mu,W}(X^*)[Z]\right)}{\min\limits_{\|Z\|_{\mathrm{F}}=1} \operatorname{tr}\left(Z^\top \operatorname{Hess} f_{\mu,W}(X^*)[Z]\right)} \tag{2.16a}$$

$$= \frac{\max\left\{\lambda_n-\lambda_1, 2(\mu w_1-\lambda_1), \max\limits_{i<j}\lambda_{\max}(M_{ij})\right\}}{\min\left\{\lambda_{p+1}-\lambda_p, 2(\mu w_p-\lambda_p), \min\limits_{i<j}\lambda_{\min}(M_{ij})\right\}}, \tag{2.16b}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and the smallest eigenvalue of the matrix

$$M_{ij} = \begin{pmatrix} \mu w_i-\lambda_j & \sqrt{(\mu w_i-\lambda_i)(\mu w_j-\lambda_j)} \\ \sqrt{(\mu w_i-\lambda_i)(\mu w_j-\lambda_j)} & \mu w_j-\lambda_i \end{pmatrix},$$

respectively.

Proof. Given a $Z\in\mathbb{R}^{n\times p}$ such that $\|Z\|_{\mathrm{F}}=1$, it can be represented as

$$\begin{aligned} &Z = V_p Z_1 + \bar{V}_p Z_2, \\ &\operatorname{tr}(Z_1^\top Z_1) + \operatorname{tr}(Z_2^\top Z_2) = 1, \end{aligned}$$

where $\bar{V}_p=(v_{p+1},\dots,v_n)$, $Z_1\in\mathbb{R}^{p\times p}$, and $Z_2\in\mathbb{R}^{(n-p)\times p}$. Using the expressions of the global minimizer (2.9) and the Hessian operator (2.11), we obtain

$$\begin{aligned} \operatorname{tr}\left(Z^\top \operatorname{Hess} f_{\mu,W}(X^*)[Z]\right) &= \operatorname{tr}(Z_2^\top \bar{\Lambda}_p Z_2) - \operatorname{tr}(Z_2^\top Z_2\Lambda_p) \\ &\quad + \operatorname{tr}(\mu Z_1^\top W Z_1) - \operatorname{tr}(Z_1^\top Z_1\Lambda_p) + \operatorname{tr}(\mu Z_1^\top S_p Z_1^\top S_p), \end{aligned} \tag{2.17}$$

where $\Lambda_p=\operatorname{diag}(\lambda_1,\dots,\lambda_p)$ and $\bar{\Lambda}_p=\operatorname{diag}(\lambda_{p+1},\dots,\lambda_n)$. The first two terms in (2.17) can be bounded as

$$\operatorname{tr}(Z_2^\top Z_2)(\lambda_{p+1}-\lambda_p) \leqslant \operatorname{tr}(Z_2^\top \bar{\Lambda}_p Z_2) - \operatorname{tr}(Z_2^\top Z_2\Lambda_p) \leqslant \operatorname{tr}(Z_2^\top Z_2)(\lambda_n-\lambda_1), \tag{2.18}$$

where the upper and lower bounds are achieved when $Z_2 = c\cdot(e_{n-p},0,\dots,0)$ or $Z_2=c\cdot(0,\dots,0,e_1)$, where e_i denotes the ith column of the identity matrix and c is a scalar.

Next, we would like to bound the terms in (2.17) associated with Z_1. Let $Z_1=(z_{ij})_{p\times p}$. A short calculation shows that

$$\begin{aligned} &\operatorname{tr}(\mu Z_1^\top W Z_1) - \operatorname{tr}(Z_1^\top Z_1\Lambda_p) + \mu\operatorname{tr}(Z_1^\top S_p Z_1^\top S_p) \\ &\quad = 2\sum_{i=1}^{p}(\mu w_i-\lambda_i)z_{ii}^2 + \sum_{j>i}\begin{pmatrix} z_{ij} & z_{ji}\end{pmatrix} M_{ij}\begin{pmatrix} z_{ij} \\ z_{ji}\end{pmatrix}, \end{aligned} \tag{2.19}$$

where $M_{ij} \in \mathbb{R}^{2\times 2}$ as defined in the theorem. Eigenvalues of M_{ij} can be calculated explicitly,

$$\begin{aligned}\lambda(M_{ij}) = &\frac{1}{2}(\mu w_i + \mu w_j - \lambda_i - \lambda_j)\\ &\pm \sqrt{\frac{1}{4}(\mu w_i + \mu w_j - \lambda_i - \lambda_j)^2 - \mu(w_i - w_j)(\lambda_j - \lambda_i)},\end{aligned} \tag{2.20}$$

and both of them are positive. Thus, (2.19) has the lower and upper bounds,

$$(2.19) \geqslant \operatorname{tr}(Z_1^\top Z_1) \cdot \min\left\{2(\mu w_p - \lambda_p), \min_{i<j} \lambda_{\min}(M_{ij})\right\}, \tag{2.21a}$$

$$(2.19) \leqslant \operatorname{tr}(Z_1^\top Z_1) \cdot \max\left\{2(\mu w_1 - \lambda_1), \max_{i<j} \lambda_{\max}(M_{ij})\right\}. \tag{2.21b}$$

Again, both bounds in (2.21) are achievable. In (2.21a), the inequality is saturated if Z_1 is parallel to $(0,\dots,0,e_p)$, or if $(z_{ij}, z_{ji})^\top$ is the eigenvector corresponding to $\min_{i<j} \lambda_{\min}(M_{ij})$ and other entries of Z_1 are zero. Similarly in (2.21b), the equality is satisfied if Z_1 is parallel to $(e_1,0,\dots,0)$, or if $(z_{ij}, z_{ji})^\top$ is the eigenvector corresponding to $\max_{i<j} \lambda_{\max}(M_{ij})$ and other entries of Z_1 are zero. Putting all bounds in (2.18) and (2.21) together, we proved (2.16). ∎

Theorem 2.4 gives the exact condition number at global minimizers, which provides an estimation of the local convergence rate around the optima. Based on (2.16), we could estimate the local convergence if the weight matrix W is chosen. In order to minimize the condition number of the Hessian operator, we will provide an intuitive approach to select a near-optimal weight matrix in the next section.

Remark 1. We give a discussion for matrices with degenerate eigenvalues among $\lambda_1, \lambda_2, \dots, \lambda_p$, i.e., we relax the assumption (2.2) as

$$\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_p < \lambda_{p+1} \leqslant \cdots \leqslant \lambda_n. \tag{2.22}$$

Theorems 2.1 and 2.3 remain valid in their current forms. Theorem 2.2 is also valid up to some changes due to the nonuniqueness of the eigenvectors of A. We give an example to show the idea of the required changes. Assume $\lambda_i = \lambda_{i+1}$ for $i < p$. Then any vector $v \in \operatorname{span}\{v_i, v_{i+1}\}$ is an eigenvector of A corresponding to λ_i and λ_{i+1}. Therefore, the matrix V_p in Theorem 2.2 may not be in the current form, $V = (v_1, v_2, \dots, v_n)$. Instead, the matrix V_p could be changed to

$$V_p = (v_1, v_2, \dots, v_p) \cdot \begin{pmatrix} I_{i-1} & & \\ & Q_1 & \\ & & I_{n-i-1} \end{pmatrix},$$

where Q_1 is a 2×2 orthogonal matrix. More generally, if there are r distinct eigenvalues among $\{\lambda_1, \lambda_2, \dots, \lambda_p\}$, the matrix V_p in Theorem 2.2 admits the form

$$V_p = (v_1, v_2, \dots, v_p) \cdot \begin{pmatrix} Q_1 & & & \\ & Q_2 & & \\ & & \ddots & \\ & & & Q_r \end{pmatrix},$$

where Q_i is an orthogonal matrix with dimension being the degree of degeneracy of the ith distinct eigenvalues.

However, as for Theorem 2.4, since the global minimizers are not isolated, the Hessian matrix of $f_{\mu,W}$ at X^* cannot be positive definite. We could employ techniques in section 2.4 to remove the nonzero null space in the Hessian matrix.

2.3. Near-optimal weight matrix. According to the analysis above, the parameter μ and the weight matrix W could be considered as an ensemble μW. That means that the degree of freedom of the parameters in (1.2) is p instead of $p+1$. Therefore, we can always set that $\mu = 1$ in analysis. Given the exact expression of the condition number $\kappa(\operatorname{Hess} f_{\mu,W}(X^*))$, we would like to minimize the condition number with respect to the weight matrix W and obtain the optimal weight matrix W^*, i.e.,

$$W^* = \arg\min_{W} \frac{\max\left\{\lambda_n - \lambda_1, 2(w_1 - \lambda_1), \max\limits_{i<j} \lambda_{\max}(M_{ij})\right\}}{\min\left\{\lambda_{p+1} - \lambda_p, 2(w_p - \lambda_p), \min\limits_{i<j} \lambda_{\min}(M_{ij})\right\}}. \tag{2.23}$$

However, (2.23) relies on the eigenvalues of A, which is not known a priori. Hence solving (2.23) exactly is infeasible.

In the following, we derive a near-optimal solution to (2.23) with eigenvalues of A. The final near-optimal solution relies on the relative eigenvalue distribution of A, which is known a priori in many practical applications, e.g., FCI. In quantum chemistry, an FCI solver is usually applied after Hartree–Fock calculation, which offers a good estimation of the eigenvalues [40]. For general eigenvalue problems, we could estimate the eigenvalues of A at a lower cost than eigensolvers [28].

To simplify the later discussion, we assume that $\lambda_p + \lambda_{p+1} < \lambda_1 + \lambda_n$, which in almost all practical applications is satisfied if $n \gg p$. We choose the weight matrix W such that

$$\frac{\lambda_1 + \lambda_n}{2} \geqslant w_1 > \cdots > w_p \geqslant \frac{\lambda_p + \lambda_{p+1}}{2}. \tag{2.24}$$

Then the objective function in (2.23) can be simplified as

$$\kappa(\operatorname{Hess} f_{\mu,W}(X^*)) = \frac{\max\left\{\lambda_n - \lambda_1, \max\limits_{i<j} \lambda_{\max}(M_{ij})\right\}}{\min\left\{\lambda_{p+1} - \lambda_p, \min\limits_{i<j} \lambda_{\min}(M_{ij})\right\}}. \tag{2.25}$$

Recall that the eigenvalues of M_{ij} as in (2.20) admit

$$\begin{aligned}\lambda(M_{ij}) &= \frac{1}{2}(w_i + w_j - \lambda_i - \lambda_j)\left(1 \pm \sqrt{1 - \frac{4(w_i - w_j)(\lambda_j - \lambda_i)}{(w_i + w_j - \lambda_i - \lambda_j)^2}}\right) \\ &\leqslant w_i + w_j - \lambda_i - \lambda_j \leqslant \lambda_n - \lambda_1,\end{aligned}$$

where the second inequality follows from (2.24). Thus, to find the minimizer of (2.23) means to maximize the denominator, whose difficulty lies in solving

$$\max_{w_1 > \cdots > w_p} \min_{1 \leqslant i < j \leqslant p} \frac{1}{2}(w_i + w_j - \lambda_i - \lambda_j)\left(1 - \sqrt{1 - \frac{4(w_i - w_j)(\lambda_j - \lambda_i)}{(w_i + w_j - \lambda_i - \lambda_j)^2}}\right). \tag{2.26}$$

Exactly solving (2.26) remains complicated. Here we give an intuitive analysis. Notice that, due to (2.24), $(w_i + w_j - \lambda_i - \lambda_j)$ is lower bounded by $\lambda_{p+1} - \lambda_p$. When

$\frac{4(w_i-w_j)(\lambda_j-\lambda_i)}{(w_i+w_j-\lambda_i-\lambda_j)^2} > c \ \forall\ i<j$ and $c>0$ is a constant bounded away from zero, then we have $\kappa\left(\operatorname{Hess} f_{\mu,W}(X^*)\right) < \frac{\lambda_n-\lambda_1}{\lambda_{p+1}-\lambda_p}\frac{2}{1-\sqrt{1-c}}$. When some $\frac{4(w_i-w_j)(\lambda_j-\lambda_i)}{(w_i+w_j-\lambda_i-\lambda_j)^2}$ approaches zero, e.g., the eigengap $\lambda_j-\lambda_i$ is small, we apply the linear approximation to the square root term in (2.26) and obtain

$$\max_{w_1>\cdots>w_p}\min_{1\leqslant i<j\leqslant p}\frac{(w_i-w_j)(\lambda_j-\lambda_i)}{(w_i+w_j-\lambda_i-\lambda_j)}. \tag{2.27}$$

Solving the max-min problem (2.26) exactly is difficult. Since (2.24) give lower and upper bounds of the denominator in (2.27), we only focus on optimizing the numerator part,

$$F(W)=\max_{w_1>\cdots>w_p}\min_{1\leqslant i<j\leqslant p}(w_i-w_j)(\lambda_j-\lambda_i),$$

which has the analytical solution $\widehat{W}$ satisfying

$$\begin{aligned}
\hat{w}_1&=\frac{\lambda_1+\lambda_n}{2},\\
\hat{w}_p&=\frac{\lambda_p+\lambda_{p+1}}{2},\\
\hat{w}_i-\hat{w}_{i+1}&=\left(\sum_{j=1}^{p-1}(\lambda_{j+1}-\lambda_j)^{-1}\right)^{-1}\frac{\hat{w}_1-\hat{w}_p}{\lambda_{i+1}-\lambda_i},\\
F(\widehat{W})&=\left(\sum_{j=1}^{p-1}(\lambda_{j+1}-\lambda_j)^{-1}\right)^{-1}(\hat{w}_1-\hat{w}_p).
\end{aligned}$$

Furthermore, through a simple calculation, we obtain that

$$\frac{F(\widehat{W})}{p-1}\leqslant F(\widetilde{W})\leqslant F(\widehat{W}),$$

where the weight matrix $\widetilde{W}$ is evenly distributed between $\hat{w}_p$ and $\hat{w}_1$. Such an inequality indicates that the uniform weight matrix $\widetilde{W}$ is a simple but effective choice for small p because it could use a few eigenvalues known a priori to determine a weight matrix with a controlled condition number. Later in section 4, all the numerical experiments use the evenly distributed weight matrix $\widetilde{W}$ since in practice it needs no extra cost.

2.4. Generalization for Hermitian matrices. In this section, we will discuss the extension of (1.2) for the eigenvalue problem of complex Hermitian matrices. Given that A is a Hermitian matrix and (Λ, V) is the eigenpairs such that

$$A=V\Lambda V^{\mathrm{H}}, \tag{2.28}$$

where $V=(v_1,v_2,\ldots,v_n)\in\mathbb{C}^{n\times n}$ and $\Lambda=\operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$ such that

$$\lambda_1<\lambda_2<\cdots<\lambda_p<\lambda_{p+1}\leqslant\cdots\leqslant\lambda_n,$$

we generalize the weighted trace-penalty model (1.2) for complex Hermitian matrices as

$$\min_{X\in\mathbb{C}^{n\times p}} f_{\mu,W}(X)=\frac{1}{2}\operatorname{tr}(X^{\mathrm{H}}AX)+\frac{\mu}{4}\|X^{\mathrm{H}}X-W\|_{\mathrm{F}}^2, \tag{2.29}$$

where the conditions on μ and W remain unchanged, i.e., μ is a positive scalar and W is a real diagonal matrix that satisfies Assumption 1. The first-order optimal condition is

$$\nabla f_{\mu,W}(X) = AX + \mu X(X^{\mathrm{H}}X - W) = 0. \tag{2.30}$$

Theorems 2.1 and 2.2 could be generalized to complex matrices, which are detailed in Theorems 2.5 and 2.6. The proofs of these theorems remain similar to the cases of real matrices.

THEOREM 2.5. *Assume A and W satisfy* (2.28) *and Assumption* 1*, respectively. Any stationary point $\widehat{X}$ of* (2.29) *has the form*

$$\widehat{X} = \widehat{U}_p \widehat{S}_p, \tag{2.31}$$

where $\widehat{U}_p = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_p) \in \mathbb{C}^{n\times p}$ and $\widehat{S}_p = \mathrm{diag}(\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_p) \in \mathbb{R}^{p\times p}$ such that

$$\begin{aligned} &A\hat{u}_i = \sigma_i \hat{u}_i, \quad \hat{u}_i^{\mathrm{H}}\hat{u}_j = \delta_{ij}, \text{ and} \\ &\hat{s}_i \in \left\{0, \sqrt{w_i - \frac{\sigma_i}{\mu}}\right\}. \end{aligned} \tag{2.32}$$

THEOREM 2.6. *Assume A and W satisfy* (2.28) *and Assumption* 1*, respectively. The global minimizer X^* of* (2.29) *has the form*

$$X^* = V_p S_p,$$

where $V_p = (v_1 e^{\imath\theta_1}, v_2 e^{\imath\theta_2}, \ldots, v_p e^{\imath\theta_p})$ for every $\theta_i \in \mathbb{R}$, and $S_p = \mathrm{diag}(s_1, s_2, \ldots, s_p) \in \mathbb{R}^{p\times p}$ such that

$$s_i^2 = w_i - \frac{\lambda_i}{\mu}.$$

However, the properties of the Hessian operator for (2.29),

$$\mathrm{Hess}\, f_{\mu,W}(X)[C] = AC + \mu(CX^{\mathrm{H}}X + XC^{\mathrm{H}}X + XX^{\mathrm{H}}C - CW), \tag{2.33}$$

change dramatically. The major difference is that the Hessian operator (2.33) is no longer a positive definite operator; instead, it is positive semidefinite. From Theorem 2.6, we find that any global minimizer X^* multiplied by a phase rotation $e^{\imath\theta}$ remains a global minimizer. Hence, unlike the symmetric matrix case where global minimizers are isolated, for Hermitian matrices, the global minimizers are located on a circle of the np-dimensional complex space and the Hessian operator at these global minimizers is positive semidefinite but not positive definite. In the following, we update our previous results and extend them for Hermitian matrices.

Define the inner product of $X \in \mathbb{C}^{n\times p}$ and $Y \in \mathbb{C}^{n\times p}$ as

$$\langle X, Y\rangle \triangleq \mathrm{Re}[\mathrm{tr}(X^{\mathrm{H}}Y)],$$

where $\mathrm{Re}[\cdot]$ denotes the real part of a complex number. For any global minimizer X^*, each element in

$$\left\{X \in \mathbb{C}^{n\times p} \mid X = X^* e^{\imath\Theta}, \Theta = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_p), \theta_i \in \mathbb{R}\right\} \tag{2.34}$$

is still a global minimizer. The set of tangent vectors of the manifold (2.34) is

$$T \triangleq \left\{ \imath X^* \Gamma \in \mathbb{C}^{n\times p} \,|\, \Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_p) \in \mathbb{C}^{p\times p} \right\}.$$

Then, we constrain the condition number of the Hessian operator on the perpendicular manifold

$$T^{\perp} = \left\{ C \in \mathbb{C}^{n\times p} \mid \mathrm{Im}[C^{\mathrm{H}} X^*]_{ii} = 0 \ \forall i \right\}, \tag{2.35}$$

where the subscript ii denotes the ith diagonal element and $\mathrm{Im}[\cdot]$ denotes the imaginary part. Finally, we would like to estimate the lower and upper bounds of the inner product,

$$\begin{aligned} \langle C, \mathrm{Hess}\, f_{\mu,W}(X^*)[C] \rangle \Big|_{C\in T^{\perp}} &= \mathrm{tr}(C^{\mathrm{H}} A C) - \mathrm{tr}(C^{\mathrm{H}} C \Lambda_p) \\ &+ \mathrm{tr}(\mu C^{\mathrm{H}} X^* (X^*)^{\mathrm{H}} C) + \mathrm{Re}\left[\mathrm{tr}(\mu C^{\mathrm{H}} X^* C^{\mathrm{H}} X^*)\right]. \end{aligned} \tag{2.36}$$

Similar to the proof of Theorem 2.4, (2.36) is upper and lower bounded by the numerator and the denominator in (2.16b), respectively.

Now, we are going to explain the reason behind splitting the space into T and $T^{\perp}$. Considering the gradient $\nabla f_{\mu,W}(X)$ in (2.30), it can be represented near the global minimizer as

$$\nabla f_{\mu,W}(X) = G_X + \Delta G_X, \tag{2.37}$$

where $G_X \in \{C \in \mathbb{C}^{n\times p} : \mathrm{Im}[C^{\mathrm{H}} X^*]_{ii} = 0 \ \forall i\}$ and $\|\Delta G_X\|$ is $o(\|G_X\|)$. Let $X = X^* + \Delta X$. Consequently, without loss of generality let $\mu = 1$ and (2.37) can be shown by

$$\begin{aligned} \nabla f_{\mu,W}(X) &= A(X^* + \Delta X) + (X^* + \Delta X)\big((X^* + \Delta X)^{\mathrm{H}}(X^* + \Delta X) - W\big) \\ &= A\cdot \Delta X + X^*\big((X^*)^{\mathrm{H}} \Delta X + \Delta X^{\mathrm{H}} X^*\big) + \Delta X\big((X^*)^{\mathrm{H}} X^* - W\big) \\ &\quad + o(\|\Delta X\|), \end{aligned} \tag{2.38}$$

$$\triangleq G_X + \Delta G_X, \tag{2.39}$$

where the second equality adopts (2.30). Thus,

$$\begin{aligned} (X^*)^{\mathrm{H}} G_X &= (X^*)^{\mathrm{H}} A \Delta X + (X^*)^{\mathrm{H}} X^* \big((X^*)^{\mathrm{H}} \Delta X + \Delta X^{\mathrm{H}} X^*\big) \\ &\quad + (X^*)^{\mathrm{H}} \Delta X \big((X^*)^{\mathrm{H}} X^* - W\big) \\ &= (W - |S_p|^2)(X^*)^{\mathrm{H}} \Delta X + (X^*)^{\mathrm{H}} \Delta X (|S_p|^2 - W) \\ &\quad + |S_p|^2 \big((X^*)^{\mathrm{H}} \Delta X + \Delta X^{\mathrm{H}} X^*\big), \end{aligned} \tag{2.40}$$

where $|S_p|^2 = \mathrm{diag}\big(|s_1|^2, \ldots, |s_p|^2\big)$. It is revealed that the diagonal elements of the third term are real, and the diagonal parts of the first two terms are opposite, which means the diagonal of (2.40) is real and $G_X \in T^{\perp}$.

That indicates the gradient in the neighborhood of the global minimizer is located on the manifold (2.35) dominantly, and when we use a gradient descent method to solve the optimization, the convergence rate is mainly dependent on the constrained condition number. Though the Hermitian matrix is interesting in some applications, in our target applications the matrices are real symmetric. Hence, we omit the detail for the analysis of Hermitian matrices.

3. Algorithms. In this section, we introduce several algorithms to address the unconstrained nonconvex minimization problem (1.2) for large-scaled matrices A.

3.1. Gradient descent methods. A common choice for our unconstrained minimization problem (1.2) is the gradient descent method. A general gradient descent method with various stepsize strategies admits the form

$$X^{(j+1)} = X^{(j)} - \alpha^{(j)} \nabla f_{\mu,W}(X^{(j)}), \tag{3.1}$$

where the superscript (j) denotes the iteration index and $\alpha^{(j)}$ is the stepsize at the jth iteration. Different stepsize strategies lead to different convergence properties. We first consider a fixed stepsize that is sufficiently small. As shown in Theorem 2.3, the weighted trace penalty model (1.2) does not have spurious local minima. We could then first adopt the idea in [15] to guarantee that the iteration never escapes from a big area such that the Lipshitz constant is bounded. Then by the discrete stable manifold theorem [16, 26], we could show that the gradient descent method for (1.2) converges to global minima for all initial points besides a set of measure zero. Consider another choice of stepsize strategy, i.e., random perturbation of a fixed stepsize. Instead of using the discrete stable manifold theorem, we could apply ideas from the stable manifold theorem for random dynamical systems [10] to show global convergence almost surely. Wen et al. [45] showed that if the stepsize is sufficiently small, not necessarily constant, the iteration variable of the gradient descent method for the original trace-penalty model stays full-rank. Such a result could be extended to our weighted trace-penalty model as well, though the aforementioned stepsize strategies, in theory, work well on global convergence. In practice, these stepsize strategies are too conservative to be numerically efficient.

For most applications, especially the FCI eigenvalue problem we consider in this paper, a good initialization is available, and, hence, more aggressive stepsize strategies are adopted in practice. Such stepsize strategies include but are not limited to exact line search, BB stepsize, etc. For the weighted trace-penalty model, the stepsize $\alpha^{(j)}$ can be computed by exact line search to make $X^{(j+1)}$ attain local directional optima in each iteration, which leads to a cubic polynomial of α as the subproblem. Numerically, we find that BB stepsize works better in the gradient descent method for our weighted trace-penalty model. Therefore, we mainly focus on the BB stepsize. Let $\delta_X^{(j)} \triangleq X^{(j)} - X^{(j-1)}$ and $\delta_G^{(j)} \triangleq \nabla f_{\mu,W}(X^{(j)}) - \nabla f_{\mu,W}(X^{(j-1)})$. The BB stepsize is defined as

$$\alpha_{\text{odd}}^{(j)} = \operatorname{tr}\left(\left(\delta_X^{(j)}\right)^\top \delta_G^{(j)}\right) \Big/ \|\delta_G^{(j)}\|_{\text{F}}^2,$$
$$\alpha_{\text{even}}^{(j)} = \|\delta_X^{(j)}\|_{\text{F}}^2 \Big/ \operatorname{tr}\left(\left(\delta_X^{(j)}\right)^\top \delta_G^{(j)}\right),$$

where the subscripts "odd" and "even" mean the iteration number (j) is odd or even. The BB stepsize requires some extra storage to store intermediate matrices $\delta_X^{(j)}$ and $\delta_G^{(j)}$ (or their variants), and the computational cost for the stepsize is $O(np)$. As a comparison, in the gradient descent method, the dominant per-iteration computational cost is to compute the matrix-matrix product of AX, which costs about $O(\text{nnz}(A) \cdot p)$ for $\text{nnz}(A)$ denoting the number of nonzero entries in A.

Compared with the original trace-penalty optimization [45], the only change is the weight matrix. Introducing such a weight matrix reduces the cardinality of the global minima set from infinite to finite and makes all global minima isolated from

each other, while searching for the global minima becomes more difficult. This could be seen from the theoretical condition numbers of the Hessian operator in (2.16) and that in [45],

$$\kappa\left(\operatorname{Hess} f_{\mu,W}\right) \geqslant \kappa\left(\operatorname{Hess} f_{\mu,I}\Big|_{V_p^{\perp}}\right) \triangleq \frac{\lambda_n - \lambda_1}{\lambda_{p+1} - \lambda_p}. \tag{3.2}$$

Viewing both optimization methods as eigensolvers, the original trace-penalty optimization requires an extra Rayleigh–Ritz process, whereas the weighted trace-penalty model converges to desired eigenpairs directly.

3.2. Coordinate descent for FCI. The gradient descent method for (1.2) works well for problems of small to moderate size, while for FCI matrices, the gradient descent method becomes less efficient and, in many cases, infeasible. In this section, we will introduce the coordinate descent method to optimize (1.2).

Recall that an FCI matrix has the following properties:

- Extremely large-scale: in practice, the dimension of the FCI matrix could easily exceed $O(10^{14})$. This makes the eigenvectors impossible to store in memory. The FCI matrix itself has to be generated on the fly, and we cannot keep the whole matrix in memory but calculate one column or row when we use it.
- Sparsity: the Hamiltonian operator under the second-quantization [4] admits

$$\hat{H} = \sum_{p,q} t_{pq} \hat{a}_p^{\dagger} \hat{a}_q + \frac{1}{2} \sum_{p,q,r,s} u_{pqrs} \hat{a}_p^{\dagger} \hat{a}_q^{\dagger} \hat{a}_s \hat{a}_r,$$

  where $\hat{a}_p^{\dagger}$ and $\hat{a}_p$ denote the creation and annihilation operators of an electron with spin-orbital index p, and t_{pq} and u_{pqrs} are one- and two-electron integrals, and the (i,j)th element of FCI matrices is nonzero if and only if the ith basis wavefunction and the jth basis wavefunction differ in at most two occupied spin-orbitals [12, 39]. Thus, the number of nonzero elements grows polynomially with respect to the number of particles, whereas the FCI matrix size grows factorially.
- Approximately sparse eigenvectors: the eigenvectors associated with low-lying eigenvalues of FCI matrices usually are sparse. The magnitudes of different entries vary widely, ranging from 10^{-16} to 10^{-1} in normalized eigenvectors. Only a few dominant entries account for nearly all the norms of eigenvectors. In practice, we approximate these eigenvectors by sparse vectors and focus on those dominant entries.

Taking all these properties into account, for symmetric FCI matrix A, the gradient descent method is not feasible: the matrix A and iteration variable X cannot be hosted in memory; and computing AX each iteration is not affordable. CDFCI [43], solving the leading eigenpair of the FCI problem, inspires us that the coordinate descent method would be an efficient algorithm to locate the sparse entries and calculate the values.

When a coordinate descent method is considered, the per-iteration updating scheme for (1.2) is of the form

$$\text{Pick a coordinate from } X^{(j)}, \text{ i.e., } (k^{(j)}, \ell), \tag{3.3a}$$

$$\alpha^{(j)} = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}} f_{\mu,W}\left(X^{(j)} + \alpha E_{k^{(j)}\ell}\right), \tag{3.3b}$$

$$X^{(j+1)} = X^{(j)} + \alpha^{(j)} E_{k^{(j)}\ell}, \tag{3.3c}$$

where $E_{k^{(j)}\ell} \in \mathbb{R}^{n\times p}$ denotes the matrix whose $(k^{(j)},\ell)$th element is one and zero elsewhere. The optimization problem in (3.3b) is a fourth-order polynomial of α, whose minimizers can be obtained via solving a cubic polynomial directly. The cubic polynomial is of the form

$$(\alpha + x_{k\ell})^3 + c_1(\alpha + x_{k\ell}) + c_0 = 0,$$

where the coefficients are

$$\begin{aligned} c_1 &= \frac{1}{\mu}a_{kk} - w_\ell + \sum_{m=1}^{n}(x_{m\ell})^2 + \sum_{m=1}^{p}(x_{km})^2 - 2(x_{k\ell})^2, \\ c_0 &= \frac{1}{\mu}\sum_{m=1}^{n} a_{km}x_{m\ell} - \frac{a_{kk}x_{k\ell}}{\mu} + \sum_{m=1}^{n}\sum_{s=1}^{p} x_{ks}x_{ms}x_{m\ell} \\ &\quad + x_{k\ell}^3 - x_{k\ell}\left(\sum_{s=1}^{p}(x_{ks})^2 + \sum_{m=1}^{n}(x_{m\ell})^2\right). \end{aligned} \tag{3.4}$$

Since all variables in the above two equations are at the jth iteration, we drop the iteration index superscript for all variables. As we shall see later in Algorithm 3.1, we maintain $Y = AX$ and $S = X^\top X$ throughout iterations. Hence both coefficients can be computed in $O(p)$ operations and then the exact line search for stepsize $\alpha^{(j)}$ can be calculated efficiently.

Our coordinate picking strategy, (3.3a), is inspired by CDFCI [43], which depends on the gradient and the nonzero pattern of A. In the (j)th iteration, we focus on the ℓth column for $\ell \equiv j (\mathrm{mod}\, p)$. We search for the entry with largest magnitude of the ℓth column of $\nabla f_{\mu,W}(X^{(j)})$ among the nonzero pattern of the $k^{(j-p)}$th column of A, where $k^{(j-p)}$ is the row coordinate updated in the $(j-p)$th iteration. That is,

$$k^{(j)} = \underset{i\in\mathcal{N}\left(A_{:,k^{(j-p)}}\right)}{\arg\max} \left|\left(\nabla f_{\mu,W}(X^{(j)})\right)_{i\ell}\right|, \tag{3.5}$$

where $\mathcal{N}(\cdot)$ denotes the nonzero pattern, $A_{:,k^{(j-p)}}$ denotes the $k^{(j-p)}$th column of A, and $(\nabla f_{\mu,W}(X^{(j)}))_{i\ell}$ denotes the (i,ℓ)th element of $\nabla f_{\mu,W}(X^{(j)})$.

According to the expression of $\nabla f_{\mu,W}$ as in (2.4), though only a small set of entries is needed, computing them at every iteration is not affordable. Thanks to an important feature of the coordinate descent method, i.e., a single entry is updated per iteration, we could efficiently maintain two important quantities: $Y^{(j)} \approx AX^{(j)}$ and $S^{(j)} = (X^{(j)})^\top X^{(j)}$. Similar to CDFCI [43], we maintain $Y^{(j)}$ as a compressed approximation of $AX^{(j)}$. The updating combined with compression formula is

$$Y_{i\ell}^{(j+1)} = \begin{cases} Y_{i\ell}^{(j)} + \alpha^{(j)}A_{ik^{(j)}} & \text{if } |\alpha^{(j)}A_{ik^{(j)}}| > \varepsilon \text{ or } Y_{i\ell}^{(j)} \neq 0, \\ Y_{i\ell}^{(j)} & \text{otherwise} \end{cases} \tag{3.6}$$

for $i \in \mathcal{N}(A_{:,k^{(j)}})$. Besides, in order to get a stable estimation of corresponding eigenvalues with high accuracy, we need to recalculate the element $Y_{k^{(j)}\ell}^{(j+1)}$ exactly by $Y_{k^{(j)}\ell}^{(j+1)} = (A_{:,k^{(j)}})^\top X_{:,\ell}$, since the Rayleigh quotient has the updating scheme

$$\left(X^\top AX\right)_{\ell\ell}^{(j+1)} = \left(X^\top AX\right)_{\ell\ell}^{(j)} + 2\alpha^{(j)}Y_{k^{(j)}\ell}^{(j+1)} - (\alpha^{(j)})^2 A_{k^{(j)}k^{(j)}}. \tag{3.7}$$

Algorithm 3.1 WTPM by coordinate descent (WTPM-CD).

Initialize $X^{(0)} \in \mathbb{R}^{n\times p}$, penalty parameter μ, and weight matrix W.
Store matrices $Y^{(0)} = AX^{(0)}$ and $S^{(0)} = (X^{(0)})^\top X^{(0)}$.
Store p-dimensional vector $d^{(0)} =$ the diagonal of $(X^{(0)})^\top Y^{(0)}$.
Construct $\nabla f_{\mu,W}(X^{(0)})$. Set $j = 0$.
while stopping criterion not achieved **do**
  Pick the coordinate $(k^{(j)}, \ell)$ to be updated in $(j+1)$th iteration by **picking rule** (3.5).
  Compute the coefficients c_0, c_1 by (3.4) and obtain the increment $\alpha^{(j)}$.
  $X^{(j+1)} = X^{(j)} + \alpha^{(j)} E_{k^{(j)}\ell}$.
  Update $Y^{(j+1)}$ by (3.6).
  Update $d^{(j+1)}$ by (3.7).
  Update $S^{(j+1)}$ by (3.8).
  Construct the searching domain in $\nabla f_{\mu,W}(X^{(j+1)})$ dependent on $\mathcal{N}(A_{:,k^{(j)}})$.
  $j \leftarrow j+1$.
end while

Due to the symmetry of $S^{(j)}$, only the upper-triangular part is stored and updated, and the updating expression for $S^{(j)}$ is

$$S_{im}^{(j+1)} = \begin{cases} S_{im}^{(j)} + \alpha^{(j)} X_{k^{(j)}i}^{(j)} & \text{if } i < \ell, m = \ell, \\ S_{im}^{(j)} + 2\alpha^{(j)} X_{k^{(j)}i}^{(j)} + (\alpha^{(j)})^2 & \text{if } i = \ell, m = \ell, \\ S_{im}^{(j)} + \alpha^{(j)} X_{k^{(j)}m}^{(j)} & \text{if } i = \ell, m > \ell, \\ S_{im}^{(j)} & \text{otherwise.} \end{cases} \tag{3.8}$$

The compression strategy of $Y^{(j)}$ restricts the increase of the number of nonzero elements of both Y and X. Compressing coordinates is very much desired, which saves a significant amount of memory. Algorithm 3.1 illustrates the framework of the algorithm.

A good choice of initial point would make iterative methods efficient. For FCI problems, Hartree–Fock provides excellent initial values for ground states and a few low-lying excited states. We lack a systematic way of choosing good initial vectors for other excited states. In principle, the initial $X^{(0)}$ must be an extremely sparse matrix so that $Y^{(0)} = AX^{(0)}$ could be calculated in a reasonable amount of time. In particular, we adopt the following initialization for our numerical results. We find the p smallest elements and corresponding indices $\{i_1, i_2, \ldots, i_p\}$ in the diagonal of the FCI matrix. The initial point $X^{(0)}$ is set to be $(e_{i_1}, e_{i_2}, \ldots, e_{i_p})$, where e_i denotes the ith column of the identity matrix.

As for the stopping criterion, in general, we use the residual norm $\|AX - X\Lambda\|_{\mathrm{F}}$, where Λ consists of the Ritz values or Rayleigh quotients, or use the gradient norm $\|\nabla f_{\mu,W}\|_{\mathrm{F}}$ to compare with the tolerance *tol*. However, in these methods, high computational cost arises from the extremely large size of A. The matrices AX and $\nabla f_{\mu,W}$ gathered during the iteration are not exact due to inadequate update of $Y^{(j)}$ and $\nabla f_{\mu,W}(X^{(j)})$. Accordingly, we introduce the summation of historical absolute increments as the stopping criterion at jth iteration, i.e.,

$$\sum_{i=0}^{h} \gamma^i |\alpha^{(j-i)}| < tol, \tag{3.9}$$

where h is a positive integer and $\gamma \in (0,1)$. Usually, we choose $h = 100$ and $\gamma = 0.99$.

In the jth iteration, the cost of updating $X_{k^{(j)}\ell}$ dominantly includes selecting the index of largest elements $(k^{(j)}, \ell)$ with $O(\mathrm{nnz}(A_{:,k^{(j-p)}}))$ flops, updating $Y^{(j)}$ with $O(\mathrm{nnz}(A_{:,k^{(j)}}))$ flops, and inadequately constructing $\nabla f_{\mu,W}(X^{(j)})$ with $O(\mathrm{nnz}(A_{:,k^{(j)}}))$ flops. Thus the coordinate descent method makes the computational cost affordable in each iteration. Another advantage of the algorithm is the utilization of memory. It only requires us to store some sparse matrices like $X^{(j)}$, $Y^{(j)}$, a $p \times p$ matrix $S^{(j)}$, and incomplete $\nabla f_{\mu,W}(X^{(j)})$ in memory.

In theory, the coordinate descent method converges faster than the full gradient descent method. However, due to the fact that modern computer architecture prefers batch operations, i.e., contiguous memory operations, the full gradient descent method often outperforms the coordinate descent method in runtime. However, the intrinsic structure of the FCI matrix benefits most from the coordinatewise method. The gradient descent method has to access the sparse matrix and the related entries in X, which destroys the contiguous memory access. On the other hand, the coordinate descent method allows us to compress coordinates and restrict the cost of memory. Furthermore, the updating strategy provides more chances for dominant elements to achieve their optimal values. Detailed numerical results are provided in the next section.

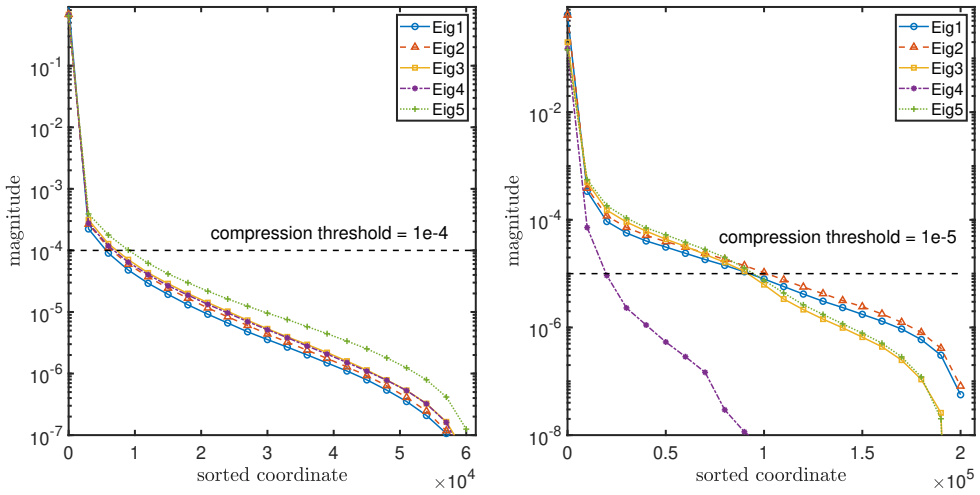
4. Numerical experiments. In this section, we will test the performance of our algorithms for computing a set of smallest eigenpairs of Hamiltonian matrices.

4.1. Performance in small systems. This section will discuss the performance of applying WTPM to some small systems and compare it with other eigensolvers. The FCI matrices are illustrated in Table 4.1. There are two matrices: "ham448" and "h2o." The "h2o" matrix is generated from one H_2O molecule system with STO-3G basis set and the "ham448" matrix is generated by the Hubbard model on a 4×4 grid with 8 fermions. The dimension n ranges from 6×10^4 to 2×10^5, which is much smaller than the dimension of the systems of practical interest. That is because we want to reveal the feature of WTPM compared with the other classical solvers. The average $\mathrm{nnz}(A_{:,j})$ shows the number of nonzero elements of each column in average, which roughly estimates the computational expense of $O(\mathrm{nnz}(A_{:,j}))$ flops in each iteration by WTPM-CD. These numerical experiments on testing matrices are performed in MATLAB R2021b.

There still exists a problem of how the μ and W are determined in practice. We present a feasible approach to choosing the proper parameters based on roughly estimated eigenvalues. Since we usually could get a good initial $X^{(0)}$ due to Hartree–Fock theory and this initial $X^{(0)}$ has only one nonzero element in each column, matrix $(X^{(0)})^\top A X^{(0)}$ with corresponding Rayleigh quotients $r_1 \leqslant r_2 \leqslant \cdots \leqslant r_p$ can be computed cheaply. Just let $\mu = 1$, and W is distributed evenly in the interval $[w_p, w_1]$ such that

TABLE 4.1
Testing FCI matrices.

Name	n	$\mathrm{nnz}(A)$	Average $\mathrm{nnz}(A_{:,j})$	$\mathrm{nnz}(A)/n^2$
ham448	207168	32040806	155	7.47e-4
h2o	61441	25060625	408	6.64e-3



(a) h2o (b) ham448

FIG. 4.1. *Absolute value of each coordinate in the sorted eigenvectors. "Eigk" denotes the kth smallest eigenvector.*

$$w_p = r_p + \varepsilon, \qquad w_1 = 2w_p - r_1, \tag{4.1}$$

where ε is a positive parameter depending on both the magnitude and the gap of initial Rayleigh quotients.

First, we demonstrate that the smallest eigenvectors are approximately sparse in Figure 4.1. In the accurate eigenvectors, the magnitude of coordinates decreases quickly and the vector norm is dominantly distributed on a few coordinates. It implies that in the process of updating we could reserve the coordinates whose magnitude is larger than a threshold and cut out others despite the loss of accuracy. Figure 4.1 also shows the compression threshold we used in the experiments.

We apply different eigensolvers to both FCI matrices, including LOBPCG in BLOPEX toolbox [24, 25], EigPen-B introduced in trace-penalty minimization [45], and our WTPM by both gradient descent (WTPM-GD) and coordinate descent (WTPM-CD) methods. The numerical results for the case $p = 5$ are illustrated in Table 4.2. All these solvers use the same initial matrix $X^{(0)}$, which is provided by Hartree–Fock theory, and terminate when the error decreases to around 10^{-3}. Let

$$err^{(j)} = \max_{\ell=1,2,\ldots,p} \left| \lambda_\ell - d_\ell^{(j)} \right|, \tag{4.2}$$

where λ_ℓ, $d_\ell^{(j)}$ respectively denote the exact eigenvalue and the corresponding Ritz value at the jth iteration. As for the individual settings of each solver, in LOBPCG, the preconditioner is not used. The penalty parameter in EigPen-B is set as

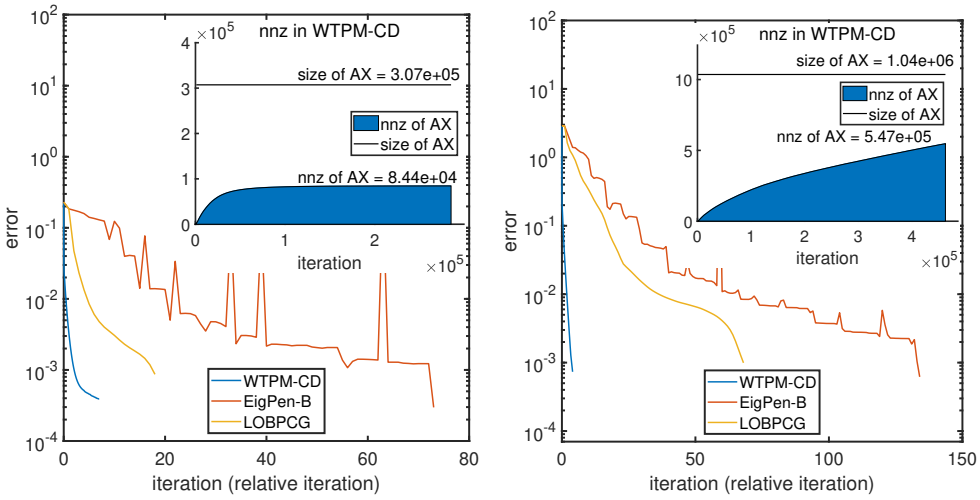
$$\mu = \max(r_p, 1). \tag{4.3}$$

The weight matrix W in WTPM is set as the above statement in (4.1) and $\mu = 1$.

Another thing we should emphasize is the computational complexity of each solver in one iteration. In LOBPCG, EigPen-B, and WTPM-GD, the dominant cost is $2p \cdot \mathrm{nnz}(A)$ flops to obtain matrix AX. But WTPM-CD updates each iteration mainly

TABLE 4.2
A comparison of updating iterations between eigensolvers for $p = 5$. The numbers in the "WTPM-CD" row represent relative iteration *outside the bracket and* actual iteration *in the bracket.*

	h2o		ham448	
	err	Iteration	*err*	Iteration
LOBPCG	8.676e-4	18	9.929e-4	68
EigPen-B	2.971e-4	73	6.19e-4	134
WTPM-GD	4.879e-4	185	3.68e-4	191
WTPM-CD	3.901e-4	**7** (283111)	7.29e-4	**4** (462000)



(a) h2o (b) ham448

FIG. 4.2. *Convergence of the smallest eigenvalues and number of nonzero elements (nnz) of the corresponding eigenvectors against iteration for $p = 5$. The nnz of AX in LOBPCG and EigPen-B is not shown here since it quickly increases to the maximum size of AX.*

at the expense of $2(p+2) \cdot \mathrm{nnz}(A_{:,\ell})$ flops, consisting of selecting the largest elements index, updating Y, and inadequately constructing $\nabla f_{\mu,W}$. That means, on average, the complexity of updating $\frac{np}{p+2}$ times in WTPM-CD equals that of updating once in the other three solvers. Hence, we introduce "relative iteration" in WTPM-CD, which means $\frac{np}{p+2}$ iterations, and the actual iteration number is shown in brackets in Table 4.2.

The results in Table 4.2 tell us that under the requirement of 10^{-3} accuracy, WTPM-CD shows the efficiency applied to FCI matrices, since its theoretical complexity is much lower than the others, though the coordinate method cannot take advantage of level-3 BLAS operations as the others and its actual runtime is longer on small systems. Instead, WTPM-GD seems not an optimal choice due to the slow convergence compared with EigPen-B and no compression of the coordinates. That is because WTPM's theoretical condition number is larger than the original trace-penalty minimization model if the elements of weight matrix W differ from each other.

Furthermore, Figure 4.2 shows the convergence of the eigenvalues in different eigensolvers and the number of nonzero elements of the matrix $Y^{(j)}$ in WTPM-CD varying against iteration. In LOBPCG and WTPM-CD, the error monotonically

decreases to the specified tolerance, but there exist spikes on the curve of error varying in EigPen-B. That is because the BB stepsize cannot guarantee monotonicity during minimization. The increasing tendency of $\mathrm{nnz}(Y^{(j)})$ in Figure 4.2 shows the effects of the coordinate descent method and compression update. The $\mathrm{nnz}(Y^{(j)})$ is restricted at a low level and so is the $\mathrm{nnz}(X^{(j)})$ because of the inequality $\mathrm{nnz}(Y^{(j)}) \geqslant \mathrm{nnz}(X^{(j)})$ in WTPM-CD, which could significantly reduce the burden of the memory source. This is what the other eigensolvers cannot do since the matrix multiplication without compression will result in a dense matrix AX.

Besides, we apply WTPM-CD with different weight matrices W to the "h2o" case to show the impact of W on the convergence as analyzed in section 2.3. We use the exact eigenvalues of A to generate three different weight matrices. Let $w_1 = \frac{\lambda_1+\lambda_n}{2}$ and $w_p = \frac{\lambda_p+\lambda_{p+1}}{2}$. We set

$$\widetilde{W}: \quad w_i = w_1 - (w_1 - w_p)\cdot\frac{i-1}{p-1},$$

$$\widehat{W}: \quad w_i - w_{i+1} = \left(\sum_{j=1}^{p-1}(\lambda_{j+1}-\lambda_j)^{-1}\right)^{-1}\frac{w_1-w_p}{\lambda_{i+1}-\lambda_i},$$

$$\text{Random}: \quad w_i \text{ is uniformly distributed in the interval } (w_p, w_1),$$

for $i = 2, 3, \ldots, p-1$. Figure 4.3 shows the convergence results for different weight matrices. It follows our conclusion in section 2.3 that $\widehat{W}$ is the best choice among the three weight matrices and $\widetilde{W}$ makes the convergence at least faster than the random choice does. In practice, if $\widehat{W}$ cannot be obtained a priori due to the cost of the construction, $\widetilde{W}$ is often found efficient enough for WTPM-CD.

In summary, these examples indicate that WTPM-CD could efficiently solve the eigenvalue problem (1.1) for FCI systems and provide an outstanding result for practical systems. In the next section, we will illustrate the performance of WTPM-CD in some larger FCI matrices of more practical interest.

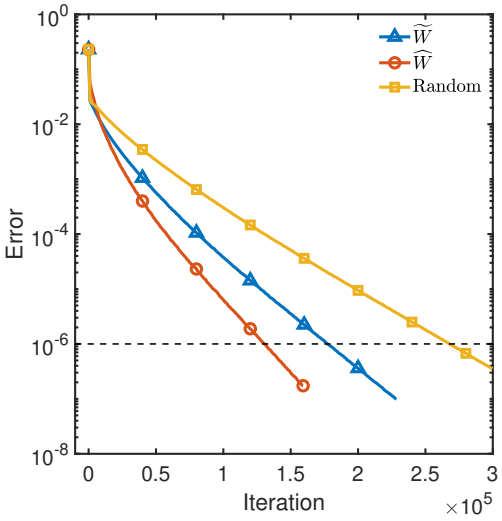


FIG. 4.3. *Convergence of the smallest eigenvalues against iteration for $p = 5$ and WTPM-CD applied to the "*h2o*" matrix with different weight matrices.*

4.2. Performance in large systems. This section provides some tests of larger FCI matrices by WTPM-CD. All programs are implemented in C++14 and compiled by Intel compiler 2021.5.0 with `-O3` option. MPI and OpenMP support are disabled for all programs in this section. All of the tests in this section are produced on a machine with an Intel Xeon Gold 6226R CPU at 2.90 GHz and 1 TB memory. The basic properties of matrices are illustrated in Table 4.3. Two matrices correspond to H_2O and C_2 molecules generated via restricted Hartree–Fock in the PSI4 package [34] with a ccpVDZ atomic orbital basis set.

In the experiments, we use the $err^{(j)}$ defined in (4.2) to measure the accuracy. Since we do not have the exact energies of excited states, we use the energies obtained by our algorithm without the compression threshold after a long enough time until the first seven digits to the right of the decimal remain unchanged as the benchmark. The penalty parameter is simply set as $\mu = 1$, and the weight matrix W follows the rules in (4.1). Thus we obtain

$$W_{h2o} = \mathrm{diag}(-75.0, -75.175, -75.35), \tag{4.4}$$

$$W_{c2} = \mathrm{diag}(-75.138, -75.238, -75.338). \tag{4.5}$$

In the "h2o_ccpvdz" case, the compression threshold is 1×10^{-6}, and that in the "c2_ccpvdz" case is 3×10^{-8}. Table 4.4 and Figure 4.4 show the convergence of WTPM-CD. The "GS," "1st ES," and "2nd ES" respectively denote the computed energies of the ground state, first excited state, and second excited state. The "Time" column denotes the time of the computed energies first reaching the appointed precision, i.e., the runtime when $err^{(j)} \leqslant tol$. This practical runtime shows the efficiency of the WTPM-CD on such molecules discretized by FCI.

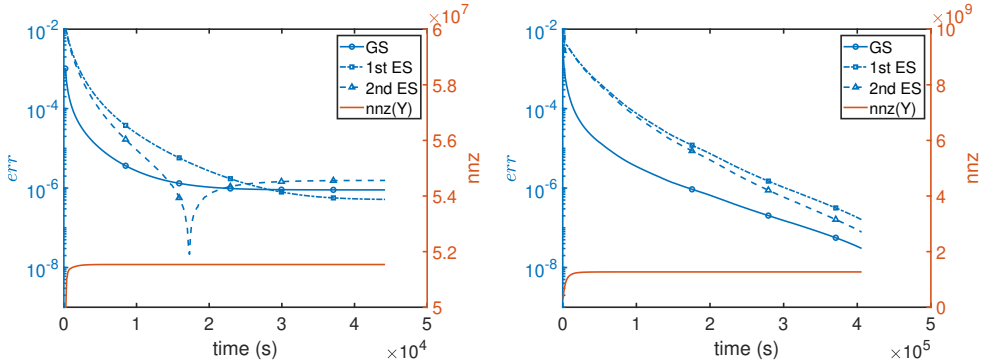
From Table 4.3, we can see that storing the dense matrices X (or AX) in double type needs at least 10 GB memory for "h2o_ccpvdz" and 400 GB memory for "c2_ccpvdz." This results in the memory bottleneck of classical eigensolvers due to the unavoidable gradient calculation or orthogonalization. The nnz(Y) in Table 4.4 tells us that under the 10^{-4} accuracy, the cost of memory by WTPM-CD decreases to 0.4 GB for "h2o_ccpvdz" and 10 GB for "c2_ccpvdz," which improves our capabil-

TABLE 4.3
Properties of testing molecule systems.

Name	Number of electrons	Number of orbitals	Matrix dimension
h2o_ccpvdz	10	24	4.53×10^8
c2_ccpvdz	12	28	1.77×10^{10}

TABLE 4.4
Convergence of WTPM-CD.

			Energy				
Matrix	p	tol	GS	1st ES	2nd ES	Time (s)	nnz(Y)
h2o_ccpvdz	3	1.0e-2	-76.24141	-75.88563	-75.86844	378	5.10e07
		1.0e-3	-76.24182	-75.89341	-75.86120	2179	5.16e07
		1.0e-4	-76.24186	-75.89430	-75.86050	8653	5.16e07
c2_ccpvdz	3	1.0e-2	-75.72888	-75.63648	-75.63609	560	3.03e08
		1.0e-3	-75.73193	-75.64174	-75.63398	34248	1.26e09
		1.0e-4	-75.73196	-75.64250	-75.63327	102503	1.27e09



(a) h2o_ccpvdz (b) c2_ccpvdz

FIG. 4.4. *Convergence of the smallest eigenvalues and number of nonzero elements (nnz) of the corresponding eigenvectors against runtime for* $p = 3$.

ity of solving larger FCI matrices generated by more complicated particle systems. Figure 4.4 shows the details of the convergence, from which we can see the nnz(Y) increases quickly at the beginning and is fixed at a small number in comparison with the dimension. Besides, the curve of "2nd ES" in Figure 4.4(a) has a different tendency from the others. That is because the energy of the second excited state in "h2o_ccpvdz" is monotonically increasing during the iteration, and it converges to a value larger than the benchmark. It indicates that if we set a compression threshold, the converged result may lose some accuracy.

5. Conclusion. In this paper, we propose an eigensolver WTPM-CD, which is an efficient algorithm for FCI eigenvalue problems of quantum many-body systems. We first propose a novel unconstrained minimization objective, namely WTPM, for Hermitian eigenvalue problems. The theoretical analysis of the minimization model tells us that the global minimizers of WTPM are exactly the eigenvectors we expect instead of the invariant subspace, so the orthogonalization process required by other methods is not needed in WTPM. Moreover, we calculate the exact condition number of the Hessian operator, and use it to give a near-optimal weight matrix W. For the algorithm framework, the coordinate descent method with compression threshold reduces the number of nonzeros and the cost of storage. In numerical experiments, compared with LOBPCG and EigPen-B solvers, WTPM-CD shows a better computational complexity on small systems. On large-scale systems, WTPM-CD guarantees its efficiency while the other two solvers suffer from the bottleneck of memory.

There is still some interesting work to be explored in the future. First, we do not give theoretical proof on the convergence of the WTPM-CD algorithm. The performance of WTPM-CD converging varies while the picking rule changes. Some in-depth research on the convergence theory is necessary to explain such a phenomenon. And some cheaper picking rules, such as stochastic search, may be introduced to the coordinate descent algorithm. The parallelization also attracts us to dive into it, since at present we can only update one element in each iteration, and the utilization efficiency of multicore is at a low level. We believe it is possible to modify the coordinate descent method to update elements in a batch and increase the parallel efficiency on shared memory systems.

REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2008, https://doi.org/10.1515/9781400830244.

[2] A. Baiardi, C. J. Stein, V. Barone, and M. Reiher, *Vibrational density matrix renormalization group*, J. Chem. Theory Comput., 13 (2017), pp. 3764–3777, https://doi.org/10.1021/acs.jctc.7b00329.

[3] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148, https://doi.org/10.1093/imanum/8.1.141.

[4] F. A. Berezin, *The Method of Second Quantization*, Academic Press, New York, 1966.

[5] N. S. Blunt, S. D. Smart, G. H. Booth, and A. Alavi, *An excited-state approach within full configuration interaction quantum Monte Carlo*, J. Chem. Phys., 143 (2015), 134117, https://doi.org/10.1063/1.4932595.

[6] 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, https://doi.org/10.1038/nature11770.

[7] G. H. Booth, A. J. W. Thom, and A. Alavi, *Fermion Monte Carlo without fixed nodes: A game of life, death, and annihilation in Slater determinant space*, J. Chem. Phys., 131 (2009), 054106, https://doi.org/10.1063/1.3193710.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[9] G. K.-L. Chan and S. Sharma, *The density matrix renormalization group in quantum chemistry*, Annu. Rev. Phys. Chem., 62 (2011), pp. 465–481, https://doi.org/10.1146/annurev-physchem-032210-103338.

[10] Z. Chen, Y. Li, and J. Lu, *On the Global Convergence of Randomized Coordinate Gradient Descent for Non-convex Optimization*, https://doi.org/10.48550/ARXIV.2101.01323, 2021.

[11] D. Cleland, G. H. Booth, and A. Alavi, *Communications: Survival of the fittest: Accelerating convergence in full configuration-interaction quantum Monte Carlo*, J. Chem. Phys., 132 (2010), 041103, https://doi.org/10.1063/1.3302277.

[12] E. U. Condon, *The theory of complex spectra*, Phys. Rev., 36 (1930), pp. 1121–1133, https://doi.org/10.1103/PhysRev.36.1121.

[13] F. Corsetti, *The orbital minimization method for electronic structure calculations with finite-range atomic basis sets*, Comput. Phys. Commun., 185 (2014), pp. 873–883, https://doi.org/10.1016/j.cpc.2013.12.008.

[14] B. Gao, X. Liu, X. Chen, and Y.-x. Yuan, *A new first-order algorithmic framework for optimization problems with orthogonality constraints*, SIAM J. Optim., 28 (2018), pp. 302–332, https://doi.org/10.1137/16M1098759.

[15] W. Gao, Y. Li, and B. Lu, *Triangularized Orthogonalization-free Method for Solving Extreme Eigenvalue Problems*, https://doi.org/10.48550/ARXIV.2005.12161, 2020.

[16] W. Gao, Y. Li, and B. Lu, *Global Convergence of Triangularized Orthogonalization-free Method*, https://doi.org/10.48550/ARXIV.2110.06212, 2021.

[17] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed., Cambridge Mathematical Library, Cambridge University Press, Cambridge, UK, 1988.

[18] 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*, J. Chem. Theory Comput., 12 (2016), pp. 3674–3680, https://doi.org/10.1021/acs.jctc.6b00407.

[19] J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan, *A brief introduction to manifold optimization*, J. Oper. Res. Soc. China, 8 (2020), pp. 199–248, https://doi.org/10.1007/s40305-020-00295-9.

[20] B. Huron, J. P. Malrieu, and P. Rancurel, *Iterative perturbation calculations of ground and excited state energies from multiconfigurational zeroth-order wavefunctions*, J. Chem. Phys., 58 (1973), pp. 5745–5759, https://doi.org/10.1063/1.1679199.

[21] P. J. Knowles and N. C. Handy, *A new determinant-based full configuration interaction method*, Chem. Phys. Lett., 111 (1984), pp. 315–321, https://doi.org/10.1016/0009-2614(84)85513-X.

[22] P. J. Knowles and N. C. Handy, *A determinant based full configuration interaction program*, Comput. Phys. Commun., 54 (1989), pp. 75–83, https://doi.org/10.1016/0010-4655(89)90033-7.

[23] P. J. Knowles and N. C. Handy, *Unlimited full configuration interaction calculations*, J. Chem. Phys., 91 (1989), pp. 2396–2398, https://doi.org/10.1063/1.456997.

[24] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541, https://doi.org/10.1137/S1064827500366124.

[25] A. V. Knyazev, M. E. Argentati, I. Lashuk, and E. E. Ovtchinnikov, *Block locally optimal preconditioned eigenvalue xolvers (BLOPEX) in Hypre and PETSc*, SIAM J. Sci. Comput., 29 (2007), pp. 2224–2239, https://doi.org/10.1137/060661624.

[26] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, *First-order methods almost always avoid strict saddle points*, Math. Program., 176 (2019), pp. 311–337, https://doi.org/10.1007/s10107-019-01374-3.

[27] Y. Li, J. Lu, and Z. Wang, *Coordinatewise descent methods for leading eigenvalue problem*, SIAM J. Sci. Comput., 41 (2019), pp. A2681–A2716, https://doi.org/10.1137/18M1202505.

[28] L. Lin, Y. Saad, and C. Yang, *Approximating spectral densities of large matrices*, SIAM Rev., 58 (2016), pp. 34–65, https://doi.org/10.1137/130934283.

[29] X. Liu, Z. Wen, and Y. Zhang, *An efficient Gauss–Newton algorithm for symmetric low-rank product matrix approximations*, SIAM J. Optim., 25 (2015), pp. 1571–1608, https://doi.org/10.1137/140971464.

[30] J. Lu and K. Thicke, *Orbital minimization method with ℓ^1 regularization*, J. Comput. Phys., 336 (2017), pp. 87–103, https://doi.org/10.1016/j.jcp.2017.02.005.

[31] J. Lu and Z. Wang, *The full configuration interaction quantum Monte Carlo method through the lens of inexact power iteration*, SIAM J. Sci. Comput., 42 (2020), pp. B1–B29, https://doi.org/10.1137/18M1166626.

[32] K. M. Nakanishi, K. Mitarai, and K. Fujii, *Subspace-search variational quantum eigensolver for excited states*, Phys. Rev. Res., 1 (2019), 033062, https://doi.org/10.1103/PhysRevResearch.1.033062.

[33] R. Olivares-Amaya, W. Hu, N. Nakatani, S. Sharma, J. Yang, and G. K.-L. Chan, *The ab-initio density matrix renormalization group in practice*, J. Chem. Phys., 142 (2015), 034102, https://doi.org/10.1063/1.4905329.

[34] R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. I. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. I. Schaefer, K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, *Psi*4 1.1*: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability*, J. Chem. Theory Comput., 13 (2017), pp. 3185–3197, https://doi.org/10.1021/acs.jctc.7b00174.

[35] F. R. Petruzielo, A. A. Holmes, H. J. Changlani, M. P. Nightingale, and C. J. Umrigar, *Semistochastic projector Monte Carlo method*, Phys. Rev. Lett., 109 (2012), 230201, https://doi.org/10.1103/PhysRevLett.109.230201.

[36] A. Sameh and Z. Tong, *The trace minimization method for the symmetric generalized eigenvalue problem*, J. Comput. Appl. Math., 123 (2000), pp. 155–175, https://doi.org/10.1016/S0377-0427(00)00391-5.

[37] J. B. Schriber and F. A. Evangelista, *Adaptive configuration interaction for computing challenging electronic excited states with tunable accuracy*, J. Chem. Theory Comput., 13 (2017), pp. 5354–5366, https://doi.org/10.1021/acs.jctc.7b00725.

[38] C. D. Sherrill and H. F. Schaefer, *The configuration interaction method: Advances in highly correlated approaches*, Adv. Quantum Chem., 34 (1999), pp. 143–269, https://doi.org/10.1016/S0065-3276(08)60532-8.

[39] J. C. Slater, *The theory of complex spectra*, Phys. Rev., 34 (1929), pp. 1293–1322, https://doi.org/10.1103/PhysRev.34.1293.

[40] J. C. Slater, *A simplification of the Hartree-Fock method*, Phys. Rev., 81 (1951), pp. 385–390, https://doi.org/10.1103/PhysRev.81.385.

[41] N. M. Tubman, J. Lee, T. Y. Takeshita, M. Head-Gordon, and K. B. Whaley, *A deterministic alternative to the full configuration interaction quantum Monte Carlo method*, J. Chem. Phys., 145 (2016), 044112, https://doi.org/10.1063/1.4955109.

[42] E. Vecharynski, C. Yang, and J. E. Pask, *A projected preconditioned conjugate gradient algorithm for computing many extreme eigenpairs of a Hermitian matrix*, J. Comput. Phys., 290 (2015), pp. 73–89, https://doi.org/10.1016/j.jcp.2015.02.030.

[43] Z. Wang, Y. Li, and J. Lu, *Coordinate descent full configuration interaction*, J. Chem. Theory Comput., 15 (2019), pp. 3558–3569, https://doi.org/10.1021/acs.jctc.9b00138.

[44] Z. Wang, Z. Zhang, J. Lu, and Y. Li, *Coordinate Descent Full Configuration Interaction for Excited States*, https://doi.org/10.48550/ARXIV.2304.13380, 2023.

[45] Z. Wen, C. Yang, X. Liu, and Y. Zhang, *Trace-penalty minimization for large-scale eigenspace computation*, J. Sci. Comput., 66 (2016), pp. 1175–1203, https://doi.org/10.1007/s10915-015-0061-0.

[46] S. R. White and R. L. Martin, *Ab initio quantum chemistry using the density matrix renormalization group*, J. Chem. Phys., 110 (1999), pp. 4127–4130, https://doi.org/10.1063/1.478295.

[47] Y. Zhou and Y. Saad, *A Chebyshev–Davidson algorithm for large symmetric eigenproblems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 954–971, https://doi.org/10.1137/050630404.

[48] Y. Zhou and Y. Saad, *Block Krylov–Schur method for large symmetric eigenvalue problems*, Numer. Algorithms, 47 (2008), pp. 341–359, https://doi.org/10.1007/s11075-008-9192-9.