切比雪夫滤波向量条件数的估计及其在 ChASE 库中的应用

Edoardo Di Napoli 和 Xinzhe Wu‡

摘要。 切比雪夫滤波子空间迭代是一种众所周知的算法，用于求解（对称/厄米）代数特征值问题，并已在多种应用程序代码中实现 [19, 24] ，也已用于独立库 [14]。该算法的一个关键部分是对经切比雪夫滤波器滤波后张成活动子空间的向量数组进行 QR 分解。通常，这种数组的条件数事先未知且可能很高，并会直接影响 QR 分解算法的选择。本文展示了如何通过精确且计算开销低的估计，为该条件数提供上界。随后，我们利用这些估计，在 ChASE 库中实现了一种 QR 分解算法选择机制。我们展示了该机制如何在不影响库准确性的前提下提升其性能。

关键词。 切比雪夫多项式、子空间迭代、滤波向量、条件数 QR 分解、数值精度、高性能计算。

arXiv:2603.10514v1 [math.NA] 11 Mar 2026

1. 引言。 采用切比雪夫加速的子空间迭代是最早发展起来、且最具影响力的迭代技术之一，用于计算对称（厄米）代数特征值问题中谱的选定部分，并且至今仍是现代大型特征值求解器的核心。首先，通过多项式滤波步骤增强试探子空间，放大与目标特征值相关的分量；随后，对所得向量进行正交归一化，并通过瑞利–里兹投影构造约化特征值问题。然后，采用标准的稠密对角化技术（例如分治算法）求解该约化问题，再将所得近似特征对投影回原始空间并测试其收敛性。

从计算角度来看，该工作流主要由两项关键计算构成：滤波步骤可高效表示为重复的矩阵乘法（通用矩阵乘法），因此能够很好地映射到现代硬件；正交归一化步骤通常采用豪斯霍尔德 QR 算法执行，是第二昂贵的运算。与滤波器不同，豪斯霍尔德 QR 高度依赖底层 BLAS 例程，因此不易实现高效的高性能计算。一个自然的替代方案是 CholeskyQR 算法，它包含大量 BLAS-3 操作，并且可以采用类似切比雪夫滤波器的方式实现。然而，CholeskyQR 对待正交归一化向量矩阵的条件数极为敏感，而多项式滤波往往会产生高度相关的子空间，这一困难因此进一步加剧。CholeskyQR2 [16] 和移位 CholeskyQR2 [15] 等近期变体提供了颇具前景的替代方案，前提是能够获得有关子空间条件数的可靠信息。在本文中，我们针对各种详尽场景，对滤波子空间的条件性进行了深入的数值分析，并展示了如何利用这些信息，在子空间迭代过程中安全地修改正交归一化步骤。我们在 ChASE 特征值求解器库中实现了所得到的策略，并证明该策略在现代架构上同时带来了数值稳健性和显著的性能提升。

1.1. 滤波子空间迭代：起源与现代复兴。科学文献中最早明确提及将子空间迭代用于求解对称代数特征值问题的文献之一，是 L. Bauer 在 1957 [4] 发表的工作。继这项工作之后，正如 Parlett [22], 所指出的，基于子空间迭代的特征值求解器 [.....是在20世纪60年代和20世纪70年代发展起来的，当时 Lanczos 算法正处于低潮期]。Rutishauser 在 1969 年至 1971 [26–28] 期间发表的一系列论文中的奠基性工作，塑造了子空间迭代的现代形态。该系列的第一篇论文 [26] 以 Bauer 的同时迭代法为基础，并引入了许多现代子空间迭代特征值求解器实现中所包含的关键思想。在这项开创性工作之后不久，Rutishauser 又发表了两篇内容相近、提供了更多算法细节的论文；该算法被命名为*ritzit*，这两篇论文奠定了其稳健性和有效性的基础 [27, 28]。

Rutishauser首次发表论文后不久，几位作者便着手改进他的初始算法。例如，Ritz迭代将特征值问题投影到搜索空间，并采用Jacobi步；Stewart将其改进为瑞利-里兹方法，以避免处理非正定矩阵 [30]。Stewart严格证明了，在正交迭代（子空间 + 正交归一化）与瑞利-里兹方法结合使用时，特征对的收敛性会得到增强。此

‡J¨ulich 超级计算中心，J¨ulich 研究中心。Wilhelm-Johnen大街，52425–J¨ulich，德国。 e.di.napoli@fz-juelich.de

与此同时，土木工程师提出了一种子空间迭代算法，其版本与Rutishauser的算法非常相似，但他们的论述不够清晰，所用语言和形式体系也与数值分析学家的惯常做法有所不同 [7]。感兴趣的读者可以在Parlett的著作中找到对改进后的$ritzit$ 算法的清晰综述 [22]。

在这些发展的同时，若干作者将子空间迭代推广到非厄米和非对称特征值问题（参见 [31] 中的一个例子）。在此后直至上一个十年的岁月里，针对厄米特特征值问题的迭代特征值求解器的发展因 Lanczos 算法及其变体的复兴而走向了不同方向 [8–11, 23, 25, 33, 36]。子空间迭代特征值求解器自2000年代中期起重新流行，并应用于电子结构理论，最初是在稀疏特征值问题的背景下 [40, 41], ，后来也用于一系列稠密特征值问题 [5, 20]。这一回归部分源于只需求解整个子空间而无需解析单个特征对的新需求 [41]，部分源于该特征值求解器能够以近似特征向量作为输入，从而显著缩短其收敛时间 [13]。

过去十年见证了滤波子空间迭代在电子结构计算新型先进方法中的日益广泛应用。切比雪夫滤波子空间迭代已被用于支持一种采用自适应局域基组 [2, 3]求解 Kohn-Sham 方程的间断伽辽金方法。类似地，ChASE 库 [34] 的早期版本被集成用于求解源自 Bethe-Salpeter 方程 [39]的 Tamm-Dancoff 近似（TDA）的激子哈密顿量。研究还表明，切比雪夫滤波子空间迭代在多核环境下能够带来高性能提升 [18]。最近，我们团队投入了大量精力，将 ChASE 库发展为功能完备的加速分布式并行库 [37, 38] ，并进一步扩展到源自超越 TDA 的 Bethe-Salpeter 方程的伪厄米特特征值问题 [21]。

1.2. 一种基于切比雪夫加速子空间迭代的现代特征值求解器。 设 A 为大小为 n的对称或复厄米特矩阵，具有 n 个实特征值，其排序使得关系$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ 成立。这种升序与文献中的传统做法相反，其灵感来源于电子结构计算中的应用，但在其他方面等价于更常见的降序描述。我们关注前 nev 个特征对——即特征值及其对应的特征向量 $(\lambda_1, x_1), (\lambda_2, x_2) \ldots, (\lambda_{\mathsf{nev}}, x_{\mathsf{nev}})$——其中 nev $\ll n$。对所需特征谱部分的选择是任意的，但与本文所给示例一致。我们仅需指出，带有多项式滤波器的子空间迭代同样可用于谱的最高部分，而不仅用于计算最小的特征对。

算法 1.1 ChASE：切比雪夫加速子空间迭代特征值求解器
要求： 矩阵 A，（可选）向量矩阵 $V_{(0)} \equiv \left[v_1^{(0)}, \ldots, v_{nev}^{(0)}\right]$

确保： nev 个极端特征对（Λ，Y），其中 $\Lambda = \mathsf{diag}(\lambda_1, \ldots, \lambda_{\mathsf{nev}})$ 且 $Y = [y_1, \ldots, y_{\mathsf{nev}}]$。

将搜索空间的大小设置为 $\ell >$nev
初始化 $V_{(0)} \leftarrow \mathsf{randn}(n, \ell)$
估计 λ_1、λ_ℓ 和 λ_n。 ▷ Lanczos
当 size(Y) $<$ nev **时**
  对向量进行滤波，$V^{(i)} \leftarrow \mathsf{p}_{m_i}(A)V_{(i-1)}$. ▷ Chebyshev filter
  正交归一化 $Q^{(i)} \leftarrow \left[Y\ V_{(i)}\right]\left(R^{(i)}\right)^{-1}$. ▷ QR algorithm
  计算瑞利商 $\tilde{A} = \left(Q^{(i)}\right)^H AQ^{(i)}$. ▷ Rayleigh-Ritz(开始)
  求解约化问题 $\tilde{A}W_{(i)} = W^{(i)}\tilde{\Lambda}^{(i)}$。
  计算 $V_{(i)} \leftarrow Q^{(i)}W^{(i)}$ ▷ Rayleigh-Ritz(结束)
  对于 $a =$ 已收敛的 $\rightarrow$nev **执行**
    如果 $\mathrm{Res}(V_{(i):a}, \tilde{\lambda}_a) <$ tol **则** ▷ Residuals
      $\Lambda \leftarrow \left[\Lambda\ \tilde{\lambda}_a\right]$
      $Y \leftarrow \left[Y\ V^{(i)}_{:,a}\right]$ ▷ Deflation & Locking
    结束如果
  结束循环
  $V_{(i)} \leftarrow \left[V_{(i):,\text{未收敛}:\ell}\right]$
结束 while

ChASE 算法 Alg. 1.1 是Rutishauser子空间迭代的一个更为精细的版本 .

初始输入仅需要矩阵 A，但也可以接受一组向量 $V^{(0)}$，它们张成所需的子空间。当作为输入指定时，$V^{(0)}$ 表示近似解（例如，在求解一系列特征值问题时）。若未指定，$V^{(0)}$ 则用正态随机向量初始化（第2行）。阶 ℓ（即 $V^{(0)}$ 的阶）使得搜索空间大于（但不会大很多）所需特征对的数目nev。

为了高效运行，切比雪夫滤波器需要以下量的估计值：λ_1、λ_ℓ 和 λ_n。只有最后一个量必须确实有上界；估计前两个量不会影响算法的有效性。三个值都可以作为输入提供，也可以通过运行几次兰索斯迭代来计算（第3行）。每次重复执行 while 循环体时，向量 $V^{(i)}$ 都会经过切比雪夫多项式滤波（第5行），多项式的次数 m_i可变。随后使用 QR 算法对所得向量进行正交归一化（第6行）：这是本文的重点步骤。正交归一化之后进行瑞利-里兹投影（第7行），并在投影结束时计算特征向量残差，将已收敛的特征对消去并锁定（第13行），同时将未收敛的向量再次送入滤波器，以重复整个过程（第16行）。这最后一步确保已满足停止准则的特征对不会继续接受滤波，从而避免不必要地增加 ChASE 执行的浮点运算次数（FLOPs）。只有剩余的未收敛向量才会被送入滤波步骤作进一步精化。

正交归一化阶段是算法过程中的关键步骤。在厄米情形下，ChASE 不仅对当前搜索子空间进行正交归一化，还确保其与所有先前已收敛特征向量保持正交性。该步骤通过去除空间中对应于已收敛方向的分量来增强收敛。因此，在 ChASE 中，无论已有多少特征对已经收敛并被锁定，QR 分解始终作用于大小为 $n \times \ell$ 的矩阵。出于数值稳健性考虑，该 QR 分解通常使用 LAPACK 或 ScaLAPACK 提供的 Householder QR 实现来完成。此步骤中的数值稳定性对 ChASE 的收敛行为至关重要，尤其是在初始迭代次数中，切比雪夫滤波向量可能高度相关时。尽管稳定性是重要因素，但由于使用了并不理想的算法（Householder QR），ChASE 库在性能上付出了代价。随着迭代过程中越来越多的特征对收敛并被锁定，QR 分解变得愈发重要，这一问题因此进一步加剧。当所计算的谱超过一小部分时，正交归一化步骤因而可能成为关键的性能瓶颈 [37]。

1.3. 用通信规避型 CholeskyQR 变体取代 Householder QR。在分布式内存架构上，尤其是在 GPU 上，传统的 Householder QR 分解往往会成为主要的性能瓶颈。尽管 Householder QR 具有出色的数值稳定性，但其性能主要受通信而非计算的限制。该算法需要频繁同步，并在进程之间进行大量小规模数据交换，尤其是在面板分解期间。这些同步点会引入延迟，且随着进程数增加，延迟表现会变差；在基于 GPU 的系统上，这种影响尤为显著，因为深层内存层次结构和内核启动开销会放大细粒度通信的成本。因此，随着 ChASE 中锁定特征对数量增加、待正交归一化的子空间扩大，Householder QR 的累积成本可能会主导整体运行时间 [37]。

为缓解这一瓶颈，我们用通信规避型（CA）CholeskyQR 变体 [15, 16]，取代传统的 Householder QR；这些变体更适用于分布式架构和基于 GPU 的架构。在 ChASE 的每次迭代中，都会对一个矩形矩阵$V^{(i)} \in \mathbb{C}^{n\times k}$ 执行 QR 分解，其中 $n > k$。在可用的通信规避型 QR 方法中，尽管 CholeskyQR 与高瘦 QR（TSQR） [12], 的渐近通信成本相近，我们仍更倾向于 CholeskyQR。CholeskyQR 的主要优势在于简单：它依赖加法归约操作，而 TSQR 在每个归约步骤都需要对一个小矩阵执行 QR 分解 [16]。这种差异简化了实现，并提升了其在层次化架构上的性能，尤其是在可使用高效集体操作（例如消息传递接口（MPI）_Allreduce）的情况下。此外，只有在 $n \gg k$ [1]时，TSQR 相较于 ScaLAPACK 的 Householder QR 才具有明显的性能优势；而这通常并非 ChASE 的应用场景。例如，在凝聚态物质模拟中，ChASE 通常针对约 10% 的谱，因此对应的是适中而非极端的 n/k 比率。

尽管 CholeskyQR 的性能和可扩展性更优，但当$V^{(i)}$ 的条件数增大时，它会出现正交性损失。将该过程执行两次可以减轻这一问题，从而得到 CholeskyQR2 算法 [16]（参见算法1.2）。然而，CholeskyQR2

仍受限于 Gram 矩阵 $R = X^H X$ 的 Cholesky 分解必须成功这一要求；只有当条件数 $\kappa_2(X)$ 不超过 $\mathcal{O}(\mathbf{u}^{-1/2})$ 时，这一要求才能满足，其中 $\mathbf{u}$ 为单位舍入误差，且 $\kappa_2(X) = {}^{\sigma_{\max}(X)}\,\sigma_{\min}(X)$。

算法 1.2 针对 $X = QR$ 的 CholeskyQR

函数 CholeskyQR(X,cholDeg)
  对于 $i = 1, \ldots,$cholDeg **执行**
    $R \leftarrow X^H X$
    $[R,$信息$] \leftarrow$ Cholesky(R)
    如果 信息 $\neq 0$ **则**
      返回 失败
    结束条件判断
    $X \leftarrow XR^{-1}$
  结束循环
  返回 X
结束函数

为克服这一限制，Fukaya 等人 [15] 提出了移位 CholeskyQR2（或 s-CholeskyQR2）算法，该算法引入了一个预条件步骤，通过将 Gram 矩阵 R 平移单位矩阵的一个小倍数，有效降低了 X的条件数 $\kappa_2(X)$，从而将 CholeskyQR2 的适用范围扩展到 $\mathcal{O}(\mathbf{u}^{-1})$ 的矩阵。表1.1 总结了各种 QR 变体对应的条件数范围。这一特征使得算法能够动态选择最合适的 QR 分解变体，以兼顾性能与数值精度。

表 1.1：最大　　的可接受条件数（对应不同的 QR 变体）。$\mathbf{u}$ 表示单位舍入误差　　.

	CholeskyQR	CholeskyQR2	s-CholeskyQR2	HHQR
$\kappa_2(X)$	$\mathcal{O}(1)$	$\mathcal{O}(\mathbf{u}^{-1/2})$	$\mathcal{O}(\mathbf{u}^{-1})$	Any

鉴于其优异的性能特点，在 ChASE 库中采用乔列斯基分解及其变体极具吸引力。同时，保持数值精度至关重要，不容妥协。要兼顾性能与稳定性，关键在于可靠地估计滤波向量 $\kappa_2(V^{(i)})$ 的条件数。直接计算这一数值的代价高昂，会抵消 CholeskyQR 带来的性能提升。在这项工作中，我们展示了一种替代方法，能够以较低开销可靠地估计条件数。我们首先介绍由此得出的策略，并说明如何利用该策略调整算法中的 QR 步骤。本文其余部分将严谨推导这些结果，并通过数值实验阐明它们对 ChASE 库稳定性和性能的影响。

1.3.1. 条件数估计与动态 CAQR。对于最简单的情形，当 ChASE 在单次迭代中收敛且不优化多项式次数（所有向量的次数均相同）时，我们对滤波向量 $V^{(i)}$ 的条件数进行界定，使其满足

$$\kappa_2\left(V^{(i)}\right) \leq \eta|\rho_1|^m, \tag{1.1}$$

其中 $|\rho_1|^m$ 是最小特征对的收敛比（参见定义 2.3）。启用多项式次数优化时，会针对数组 $V^{(i)}$中的每一列分别调整滤波次数。在这种情况下， $|\rho_1|^m$ 中的指数将替换为最高多项式次数，该次数通常对应于搜索空间的最后一个向量 m_ℓ。当向量被消去并锁定时，界限的表达式还会进一步修改为

$$\kappa_2(V^{(i)}) \lesssim \eta|\rho_{k+1}|^{m_{k+1}}|\rho_1|^{m_\ell - m_{k+1}}. \tag{1.2}$$

其中， m_k 是与索引为 k 的 Ritz 对相对应的切比雪夫多项式次数，而 $V^{(i)} \leftarrow \begin{bmatrix} Y & V_{(i)} \end{bmatrix}$是待分解的矩形矩阵，其中包含已收敛特征向量 Y 和新滤波得到的子空间。

这些估算仅依赖于 ChASE 中已有的数据，即谱界（c, e）、计算得到的 Ritz 值 Λ、优化后的滤波器次数 m_i以及锁定特征对的数量k。如前所述，这些估算基于对切比雪夫滤波子空间的详细谱分析，构成了本文的核心贡献，后续各节将对此进行阐述。根据估算得到的条件数（公式 1.1 和 1.2），我们按照算法 1.3中的启发式方法选择适当的乔列斯基分解变体。对于双精度计算，如果估算的条件数超过 10^8, ，则采用 s-CholeskyQR2 变体。对于较小的条件数（低于 20），标准 CholeskyQR 即可满足要求；而对于中等情况，则使用 CholeskyQR2。为确保稳健性，回退机制（算法 1.3，第 8 行）可确保在 s-CholeskyQR2 失败时，安全地回退到 ScaLAPACK 的 Householder QR 例程。

算法 1.3 ChASE 中的动态 CAQR 选择

1: **函数** DYNAMIC-CAQR(X,估计条件数)
2: **如果** 估计条件数 $> 10^8$ **则** ▷ 条件数较高：使用 s-CholeskyQR2
3: $R \leftarrow X^H X$
4: 范数 $\leftarrow \|X\|_F$
5: $s \leftarrow 11(mn + n(n+1))$ **u**范数
6: [R,信息] $\leftarrow$ CHOLESKY($R + sI$)
7: **如果** 信息 $\neq 0$ **则** ▷ 乔列斯基分解失败，回退到 Householder QR
8: $X \leftarrow$ HOUSEHOLDERQR(X)
9: **else**
10: $X \leftarrow XR^{-1}$
11: $X \leftarrow$ CHOLESKYQR(X，胆固醇度 $= 2$)
12: **结束 if**
13: **否则，如果** 估计条件数 < 20 **则** ▷ 条件良好：使用单次 CholeskyQR
14: $X \leftarrow$ CHOLESKYQR(X，Cholesky 阶数 $= 1$)
15: **else** ▷ 中等条件数：使用 CholeskyQR2
16: $X \leftarrow$ CHOLESKYQR(X，Cholesky 阶数 $= 2$)
17: **结束 if**
18: **返回** X
19: **结束函数**

在第 2 节中，我们描述滤波向量 $V^{(i)}$ 的谱性质，既包括多项式次数一般的情形，也包括针对每个滤波向量分别优化该次数的情形。第 3 节对所有可能情形下 $\kappa_2(V^{(i)})$ 的界进行数值分析。第 4 节专门介绍条件数界的数值测试、动态 CAQR 与传统 QR 分解的比较，以及若干缩放测试。第 5 节对全文进行总结并得出结论。

2. 切比雪夫多项式滤波子空间的谱分解。 本节将对经过滤波过程后由向量构成的子空间进行形式化刻画。我们首先回顾切比雪夫多项式的渐近性质，然后说明如何利用这些多项式对向量矩阵进行滤波。接下来，我们引入与每个特征对相关的收敛比概念，并说明这些比值如何自然地出现在滤波器的谱分解中。借助这一框架，我们可以通过矩阵分解来表示滤波后的向量，并清晰地区分与谱中目标部分相关的分量和与非目标部分相关的分量。

2.1. 切比雪夫多项式及其渐近性质。 为了讨论子空间迭代算法的切比雪夫加速的一般收敛性质，我们先简要回顾第一类切比雪夫多项式的定义，并简要介绍其渐近性质。

DEFINITION 2.1（切比雪夫多项式）。 区间 $[-1, 1]$ 上的 m 次切比雪夫多项式 $C_m(t)$ 定义为

$$C_m(t) = \cos\left(m \cos^{-1}(t)\right), \quad |t| \leq 1 \tag{2.1}$$

通过以下方式，可以将定义域扩展到整个实轴：

$$C_m(t) = \cosh\left(m \cosh^{-1}(t)\right), \quad t \in \mathbb{R}$$

通常，更实用的做法是通过三项递推关系计算切比雪夫多项式。该关系可由两个余弦和的积化和差公式（根据奥斯本法则推广至双曲余弦）轻松推导得出。

$$C_{m+1}(t) = 2tC_m(t) - C_{m-1}(t) \qquad \forall\, t \in \mathbb{R}. \tag{2.2}$$

这个递推关系，再加上根据式 (2.1) 直接计算 $C_0(t) = 1$ 和 $C_1(t) = t$，表明任意次数 m 的切比雪夫多项式确实是关于 t 的多项式，其最高次数为 m。在区间 $[-1,1]$ 上，次数为 m 的多项式能够重现具有 $m-1$个极值点的振荡函数；而对于 $|t| > 1$，即使 m取值不大，函数也会迅速发散。通过将 $C_m(t)$ 表示为 t的隐函数，可以很容易地量化这一行为。令 $\rho \doteq \exp(y)$，并设 $\cosh(y) = \frac{\rho+\rho^{-1}}{2} = t$，可得

$$\rho^2 - 2t\rho + 1 = 0 \quad \Rightarrow \quad \rho = t \pm \sqrt{t^2-1},$$

这意味着 ρ 和 ρ^{-1} 都是可接受的解。按照惯例，我们选择

$$|\rho| = \max_{\{\pm\}} \left| t \pm \sqrt{t^2-1} \right| \;, \quad |\rho|^{-1} = \min_{\{\pm\}} \left| t \pm \sqrt{t^2-1} \right| \quad \text{for} \quad |t| > 1$$

代回 ρ 的定义中，以便表示 y

$$y = \cosh^{-1}(t) = \ln|\rho| \,.$$

现在，多项式可以改写为

$$C_m(t) = \cosh\left(m \ln|\rho|\right) = \frac{|\rho|^m + |\rho|^{-m}}{2}. \tag{2.3}$$

由于按照定义，$|\rho|$ 始终大于 1，因此多项式关于 $|t| > 1$ 的主导渐近行为由 $C_m(t) \sim \frac{|\rho|^m}{2}$ 给出。换言之，在区间 $[-1,1]$之外，$C_m(t)$会像次数为 m的多项式一样发散。顺带一提，在此区间内，$C_m(t)$的完整表达式仍然成立。唯一的区别是，此时 ρ 取复数值，反双曲函数也如此

$$\rho = t \pm i\sqrt{1-t^2} \quad \Rightarrow \quad \cosh^{-1}(t) = \arg(t \pm i\sqrt{1-t^2}) = y,$$

其中 y 不过是复平面上单位圆上的一个点所对应的角度。因此，式（2.3）现在表示 $C_m(t)$ 在 $|t| < 1$上的最大值，即 1。以下定理 [29]为子空间迭代的切比雪夫加速奠定了基础。

THEOREM 2.2. 令 $|\gamma| > 1$ 和 $\mathbb{P}_m$ 表示次数小于或等于 m的多项式集合。那么极值

$$\min_{\{p \in \mathbb{P}_m, p(\gamma)=1\}} \max_{\{t\in[-1,1]\}} |p(t)|$$

由.....达到

$$\mathrm{p}_m(t) \doteq \frac{C_m(t)}{C_m(\gamma)}.$$

通俗地说，固定一点 γ，并考虑所有经过该点且在某个不包含 γ 的区间上达到最大值的多项式。这些多项式中较小的那个由 $\mathrm{p}_m(t)$ 唯一确定。在不影响定理表述的前提下，$\mathrm{p}_m(t)$ 的定义可以扩展到一般的 t，但就我们的目的而言，它最适用于 $|\gamma| \geq |t| \geq 1$。

这个多项式可以直接推广到一般区间 $x \in [\alpha,\beta] \subset \mathbb{R}$。下面的线性变换 $t(x) = \frac{x-c}{e}$ 将区间 $[\alpha,\beta]$ 映射到 $[-1,1]$，其中 $c = \frac{\alpha+\beta}{2}$ 表示区间中心，$e = \frac{\beta-\alpha}{2}$ 表示其半宽。对于较大的 m，$p_m(x)$ 的首阶项为

$$\lim_{m\to\infty} \mathrm{p}_m(x_t) \doteq \frac{C_m(\frac{x_t-c}{e})}{C_m(\frac{x_\gamma-c}{e})} \sim \frac{|\rho_{x_t}|^m}{|\rho_{x_\gamma}|^m} \quad \text{with} \quad |\rho_{x_\gamma}| = \max_{\{\pm:\ x_\gamma \notin [\alpha,\beta]\}} \left| \frac{x_\gamma - c}{e} \pm \sqrt{\left(\frac{x_\gamma - c}{e}\right)^2 - 1} \right|. \tag{2.4}$$

当 $x \in [\alpha,\beta]$ 渐近极限简化为 $\mathrm{p}_m(x) \sim \frac{1}{|\rho_{x_\gamma}|^m}$，因为 $\sup_{x\in[\alpha,\beta]} |\rho_x| = 1$。这种渐近行为对于将此类多项式用作子空间迭代的加速器至关重要。

2.2. 使用切比雪夫多项式对向量进行滤波。 定理 2.2 为构建切比雪夫多项式滤波器奠定了基础，该滤波器可加快标准子空间迭代特征值求解器的收敛。初始输入向量 V 通过利用三项递推关系进行滤波，该关系隐式地表示为 A 对向量的作用 V

$$C_m(V) = 2\ A\ C_{m-1}(V) - C_{m-2}(V) \quad ; \quad C_m(V) \stackrel{\text{def}}{=} C_m(A) \cdot V. \tag{2.5}$$

多项式滤波器通过操纵初始子空间来改进子空间迭代，使得相对于目标谱部分而言，非目标谱部分中的分量被大幅抑制。图 2.1 以图形方式展示了这一点，其中红色的特征谱部分被映射到区间 $[-1, 1]$，并对应于非目标谱部分。

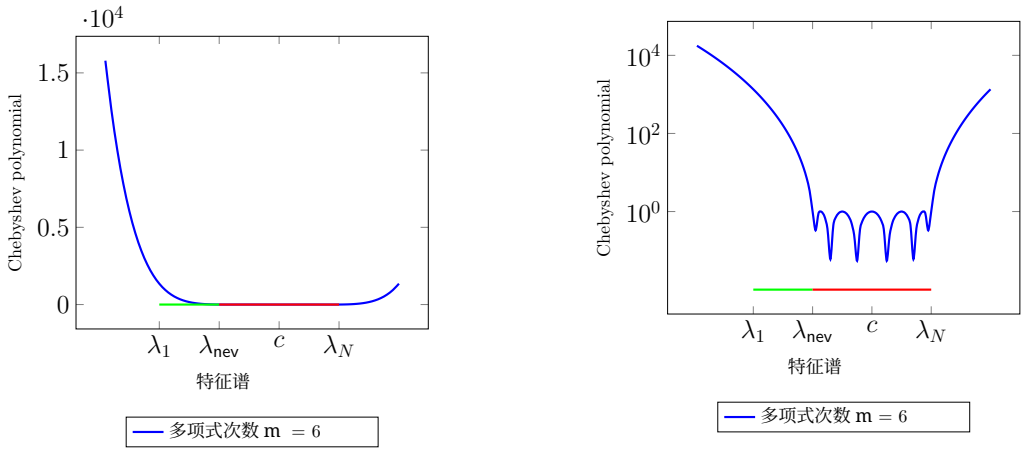


图 2.1：示意图展示了如何使用切比雪夫多项式 $C_m(\lambda)$，增强与厄米矩阵 A 的谱中所需部分 $[\lambda_1, \lambda_{\text{nev}}]$（绿色部分）对应的特征向量对齐的分量 V。相反，与区间 $[\lambda_{\text{nev}}, \lambda_N]$（红色部分）内特征对对应的分量 V 会受到抑制。这里，λ_N 表示 A 的最大特征值。

衡量切比雪夫多项式所执行滤波程度的一个量是收敛比，它通过变量 $|\rho_a|$ 定义，其中 a 是区间 $[\alpha, \beta]$ 之外特征值的索引。

Definition 2.3. 与特征值 λ_a 对应的特征向量的**收敛比** 定义为

$$\tau(\lambda_a) = |\rho_a|^{-1} = \min_{\{\pm : \lambda_a \notin [\alpha,\beta]\}} \left| \frac{\lambda_a - c}{e} \pm \sqrt{\left(\frac{\lambda_a - c}{e}\right)^2 - 1} \right|.$$

λ_a 离区间 $[\alpha, \beta]$ 越远，$|\rho_a|$ 就越大，任何滤波向量 $v \in V$ 收敛到与特征值 λ_a 对应的特征向量 x_a 的速度也就越快。实际上，认为 $|\rho_a|^{-1}$ 代表 v 的收敛比，与其说是一个精确的论断，不如说是一条经验法则：严格证明需要付出相当大的努力，不在本文的讨论范围之内，而且我们的结果并不依赖于此。需要牢记的是，$\forall \lambda_a \notin [\alpha, \beta]$，$|\rho_a| > 1$。

在处理切比雪夫加速的数值实现时，$\mathrm{p}_m(A)$ 对 V 的作用。常数分母 $\eta_m \doteq C_m(\frac{\lambda_1 - c}{e})$ 的计算与矩阵分子分开进行。通过命名 $\sigma_{m+1} \doteq \frac{\eta_m}{\eta_{m+1}}$，可以递归地定义 σ_m

$$\sigma_m = \frac{1}{\frac{2}{\sigma_1} - \sigma_{m-1}} \quad \text{with} \quad \sigma_1 = \frac{\eta_0}{\eta_1} = \frac{e}{\lambda_1 - c}. \tag{2.6}$$

根据此定义　　已知 n，$\mathrm{p}_m(A)$ 对 V 的作用可表示为一种特化的三项递推

关系

$$\begin{aligned}\mathrm{p}_m(A)V &= \frac{C_m\left(\frac{A-cI}{e}\right)}{\eta_m}V \\ &= \frac{2\,\frac{A-cI}{e}\,C_{m-1}\left(\frac{A-cI}{e}\right) - C_{m-2}\left(\frac{A-cI}{e}\right)}{\eta_m}V \\ &= 2\,\frac{A-c}{e}\,\frac{\eta_{m-1}}{\eta_m}\,\frac{C_{m-1}\left(\frac{A-cI}{e}\right)}{\eta_{m-1}}V - \frac{\eta_{m-1}}{\eta_m}\,\frac{\eta_{m-2}}{\eta_{m-1}}\,\frac{C_{m-2}\left(\frac{A-cI}{e}\right)}{\eta_{m-2}}V \\ &= 2\,\frac{A-cI}{e}\,\sigma_m\,p_{m-1}(A)V - \sigma_m\,\sigma_{m-1}\,p_{m-2}(A)V.\end{aligned}$$

在实际　　意义上，上述多项式滤波器以迭代方式实现，如算法 2.　　1.

算法 2.1 切比雪夫多项式滤波器

~　　~**要求：** 矩阵 A、随机向量 V 以及特征值估计 λ_1 和 λ_{nev}。**确保：** 滤波向量 $V^{(m)}$。

$c = \frac{\tilde{\lambda}_1+\tilde{\lambda}_{\mathsf{nev}}}{2}$, $e = \frac{\tilde{\lambda}_{\mathsf{nev}}-\tilde{\lambda}_1}{2}$, $\sigma_1 = \frac{e}{\tilde{\lambda}_1-c}$;

$V^{(1)} = \frac{\sigma_1}{e}\,(A - c\,I_n)\,V$;

对 $i = 2 \to m$ **执行**

$\sigma_i = \frac{1}{\frac{2}{\sigma_1}-\sigma_{i-1}}$;

$V^{(i)} = \frac{2\sigma_i}{e}\,(A - c\,I_n)\,V^{(i-1)} - \sigma_{i-1}\sigma_i V^{(i-2)}$;

结束 for

如上所述，可以证明，随机向量数组 V 的分量会收敛到由特征向量 x_a 所张成的子空间，$\lambda_a \notin [\alpha, \beta]$ 其具有 τ_a 收敛比。粗略地说，这一观察意味着，每个向量 $v_a \in V$ 在理论上都可以通过滤波器进行滤波，而多项式次数只需足够高，以确保其收敛到 x_a 达到特定阈值。在 ChASE 库中，该机制通过多项式次数 m 实现；它是一个与 V 维度相同的有序次数数组，而不是对所有向量仅使用一个值。算法 2.1 `for`循环随后通过 $m \to m_a$ 以及添加一条 if 语句进行了略微修改，该语句会在滤波向量 v_a 一旦 $m_a < i$时立即停止滤波。所得算法优化了达到收敛所需的最少运算次数。

2.3. 滤波向量的谱分解。 本节将给出 $\mathrm{p}_m(A)V$ 的一个简单谱分解，随后将在接下来的章节中利用它推导关于 $\mathrm{p}_m(A)V$ 条件数估计的一些结果。利用 $A = X\Lambda\ X^H$的谱分解，我们将初始向量集合 $V^{(0)}$ 表示为在精确特征向量 $X = (X_\ell X_3)$上的投影，其中这些向量的阶数分别为 ℓ和 $n-\ell$

$$V^{(0)} \equiv X\ X^H V^{(0)} = X_\ell S_\ell + X_3 S_3$$

其中 $S_\ell \in \mathbb{C}^{\ell\times\ell}$，且 $S_3 \in \mathbb{C}^{(n-\ell)\times\ell}$ （见 [32]）。这里我们想强调，大小为 ℓ 的子空间可以进一步划分为两个子空间 $X_\ell = （X_1 X_2）$，其对应的系数矩阵分别为 S_1 和 S_2，其中 $S_1 \in \mathbb{C}^{k\times\ell}$，且 $S_2 \in \mathbb{C}^{(\ell-k)\times\ell}$。下面的谱表示中不会用到这一步细分，但它有助于强调对应前 nev 个特征值的子空间与 ChASE 中其余 nextra 个特征值所对应子空间之间的区别。本文中，当考虑在部分向量已锁定并消去时对条件数界的估计（见第 3.3 节），这一步细分也很有用。暂时省略 V 的上标，并考虑在常数次数 m 的情况下，滤波器在一般的第 i 次迭代中输出的向量（见算法 1.1 第 5 行）的这种分解

$$\mathrm{p}_m(A)V = X_\ell \Gamma_\ell^m S_\ell + X_3 \Gamma_3^m S_3 \tag{2.7}$$

其中对角矩阵 $\mathrm{p}_m(\Lambda_\ell) \doteq \Gamma_\ell^m \in \mathbb{C}^{\ell\times\ell}$，且 $\mathrm{p}_m(\Lambda_3) \doteq \Gamma_3^m \in \mathbb{C}^{(n-\ell)\times(n-\ell)}$。即使 m 的值适度偏小，这些矩阵仍为正定矩阵，因此可逆。除非向量 V[1] 的选择极为特殊，我们也可以有把握地假设 S_ℓ 满秩，因此可逆。

表达式 (2.7) 中的 $\mathrm{p}_m(A)V$可以进一步变形，以便清楚地区分由 X_ℓ 张成的目标子空间与其余部分，

$$\mathrm{p}_m(A)V = \left(X_\ell + X_3\Gamma_3^m S_3 \left(S_\ell\right)^{-1}\left(\Gamma_\ell^m\right)^{-1}\right)\Gamma_\ell^m S_\ell. \tag{2.8}$$

Γ 矩阵是对角矩阵；即使 m 的值相当小，也已有简单的表达式：

$$\begin{aligned}\Gamma_\ell^m &= |\rho_1|^{-m}\times\mathrm{diag}(|\rho_1|^m,|\rho_2|^m,\ldots,|\rho_\ell|^m)\\ (\Gamma_\ell^m)^{-1} &= |\rho_1|^{m}\times\mathrm{diag}(|\rho_1|^{-m},|\rho_2|^{-m},\ldots,|\rho_\ell|^{-m})\\ \Gamma_3^m &= |\rho_1|^{-m}\times I.\end{aligned} \tag{2.9}$$

如第 2.2 节末尾所示，在次数优化的情况下，多项式次数 m不再是一个整数，而是由整数构成的数组 $m = (m_1, m_2, \ldots, m_\ell)$，其中每个整数对应 V 的一个列向量，且 $m_1 \le m_2 \le \cdots \le m_\ell$。此时，表达式 $\mathrm{p}_m(A)V$则推广为表示

$$\mathrm{p}_m(A)V \equiv \left(\mathrm{p}_{m_1}(A)v_1, \mathrm{p}_{m_2}(A)v_2, \ldots, \mathrm{p}_{m_\ell}(A)v_\ell\right),$$

其中 $V \equiv (v_1, v_2, \ldots, v_\ell)$，且 $v_a \in \mathbb{C}^n$。尽管表面上看起来更为复杂，$\mathrm{p}_m(A)V$仍可通过 Γ 矩阵表示，唯一的区别是，它们用收敛比的逆表示时，表达式变为

$$\begin{aligned}\Gamma_\ell^m &= \mathrm{diag}(1, \frac{|\rho_2|^{m_2}}{|\rho_1|^{m_2}}, \ldots, \frac{|\rho_\ell|^{m_\ell}}{|\rho_1|^{m_\ell}})\\ (\Gamma_\ell^m)^{-1} &= \mathrm{diag}(1, \frac{|\rho_1|^{m_2}}{|\rho_2|^{m_2}}, \ldots, \frac{|\rho_1|^{m_\ell}}{|\rho_\ell|^{m_\ell}})\\ \Gamma_3^m &= |\rho_1|^{-m_\ell}\times I.\end{aligned} \tag{2.10}$$

2.3.1. 锁定和消去向量的情况。在 ChASE 库中，与任何子空间迭代算法的实现一样，从滤波到收敛检查的所有步骤都在循环中重复执行。检查收敛性时，残差低于阈值的每个向量都会被锁定，并从 V 中消去（见算法1.1）。在这种情况下，需要重新正交化的向量的谱分解具有不同的结构，有待进一步分析。

不失一般性，我们先假设在某次迭代中，k 个向量已锁定在数组 $W \in \mathbb{C}^{n\times k}$ 中。在下一次迭代中，滤波步骤之后，QR 算法将对矩阵 $Z \equiv [W\ \mathrm{p}_m(A)V]$进行正交化，其中 $V \in \mathbb{C}^{n\times(\ell-k)}$。与之前一样，$\mathrm{p}_m(A)V$ 满足式 2.8的谱分解，唯一的区别在于矩阵 $S_{\ell,3}$现在只有 $\ell - k$列。对于 W，我们可以定义如下分解

$$W = X_\ell Q_\ell + X_3 Q_3 \quad \text{with } Q_\ell \in \mathbb{C}^{\ell\times k} \text{ and } Q_3 \in \mathbb{C}^{(n-\ell)\times k}. \tag{2.11}$$

尽管存在相似之处，而且由于收敛检查，这些矩阵$Q_{\ell,\ 3}$ 与 $S_{\ell,\ 3}$具有截然不同的性质。这是因为每个列向量 $w_i \in W$ 都满足残差方程

$$\|r_i\|_2 = \frac{\|Aw_i - \tilde{\lambda}_i w_i\|_2}{\|w_i\|_2} < \mathrm{TOL}. \tag{2.12}$$

此外，如果特征值满足一致分离条件，则可以保证 [22][Sec-4.5] 每个特征值 λ_i 都位于由计算得到的˜特征值 λ_i 和残差所定义的区间内。换句话说，˜ ˜ ˜如果所有区间 $[\lambda_i - \|r_i\|_2, \lambda_i + \|r_i\|_2]$ 互不相交，则可以保证 λ_i 为 λ_i[2]提供唯一近似。

[1]在 ChASE 中，初始向量 $V^{(0)}$ 是随机生成并正交归一的；向量 $V_{(i)}$ 则根据构造正交归一，且通常与滤波子空间[2]有很大的重叠。处理聚集特征值的情况需要多做一些工作，并了解聚集特征值的残差矩阵（参见 [22][第 11.5] 节）

PROPOSITION 2.4. 假设每个特征对 $(w_i, \lambda_i)\, i = 1, \ldots, k$ 都满足一致分离条件，且其残差 $\|r_i\| < \text{TOL}$，$\exists\ \zeta_i = \min_{j \neq i} |\lambda_j - \lambda_i|$ 使得 $\forall j \neq i$

$$\|(Q_\ell)^2_{:\neq i,i}\|_2^2,\ \|(Q_3)^2_{:,i}\|_2^2 \leq \left(\frac{\text{TOL}}{\zeta_i}\right)^2$$

证明。 我们将 A 按其谱分解写成 $A = X\Lambda X^H$，并利用 $w_i = X_\ell (Q_\ell)_{:,\, i} + X_3 (Q_3)_{:,\, i}$ 的分解，则方程（2.12）的分子变为

$$\|Aw_i - \tilde{\lambda}_i w_i\|_2 = \left[\sum_{\substack{j=1 \\ j\neq i}}^{\ell} (\lambda_j - \tilde{\lambda}_i)^2 |Q_\ell|^2_{j,i} + (\lambda_i - \tilde{\lambda}_i)^2 |Q_\ell|^2_{i,i} + \sum_{j=1}^{n-\ell} (\lambda_j - \tilde{\lambda}_i)^2 |Q_3|^2_{j,i}\right]^{\frac{1}{2}}$$

$$\geq \left[\sum_{\substack{j=1 \\ j\neq i}}^{\ell} (\lambda_j - \tilde{\lambda}_i)^2 |Q_\ell|^2_{j,i} + \sum_{j=1}^{n-\ell} (\lambda_j - \tilde{\lambda}_i)^2 |Q_3|^2_{j,i}\right]^{\frac{1}{2}} \geq \zeta_i \left[\sum_{\substack{j=1 \\ j\neq i}}^{\ell} |Q_\ell|^2_{j,i} + \sum_{j=1}^{n-\ell} |Q_3|^2_{j,i}\right]^{\frac{1}{2}}.$$

假设 w_i 已归一化为单位，则最后一个不等式可直接转化为

$$\left(\frac{\text{TOL}}{\zeta_i}\right)^2 \geq \|(Q_\ell)^2_{:\neq i,i}\|_2^2 + \|(Q_3)^2_{:,i}\|_2^2$$

从而证明了该命题的陈述。 □

就一切实际目的而言，当 $k < \ell$ 时 $|\lambda_{\ell+j} - \tilde{\lambda}_i| \gg \zeta_i$，这意味着 $\left(\frac{\text{TOL}}{\zeta_i}\right)^2 \gg \|(Q_3)^2_{:,\, i}\|_2^2$。于是可以合理地假设 $W \approx X_\ell Q_\ell$。我们将大小为 $\mathcal{X}_\ell = \mathcal{R}(X_\ell)$ 的子空间分成两部分：一部分由前 k 个特征向量 X_1 张成，其个数对应于锁定向量的数目；另一部分由其余的 $\ell - k$ 个特征向量 X_2 张成。于是，$Q_\ell = (Q_1^\top\ Q_2^\top)^\top$，类似地 $S_\ell = (S_1^\top\ S_2^\top)^\top$，其中 $S_1 \in \mathbb{C}^{k\times(\ell-k)}$ 且 $S_2 \in \mathbb{C}^{(\ell-k)\times(\ell-k)}$。$W$ 与 V 的分解可简单地改写为

$$W = X_1 Q_1 + X_2 Q_2 \quad \text{and} \quad V = X_1 S_1 + X_2 S_2 + X_3 S_3.$$

REMARK: 在 *ChASE*算法 *1.1*中，一旦有若干向量 W 被消去并锁定，它们在构造时也会与新的向量数组正交 V，该数组是切比雪夫滤波器的新输出，也是下一次子空间迭代的输入。这意味着 $\|W^H V\|_2 = \|Q_1^H S_1 + Q_2^H S_2\|_2 \leq \mathcal{O}\left(\epsilon^\perp\right)$ 与 $\epsilon^\perp$瑞利-里兹步中达到的正交精度一致。根据构造，Q_1 和 S_2 都应为满秩，这意味着 $\|Q_2\|_2,\ \|S_1\|_2 \sim \mathcal{O}\left(\epsilon^\perp\right)$。这是一个操作层面的结果，取决于瑞利-里兹步中所用直接/稠密特征值求解器达到的正交性。
假设上述操作性注释有效，我们最终可以按以下方式写出 Z中的向量集合

$$Z = [W\ \ \mathrm{p}_m(A)V] = \left[X_1 Q_1 \quad \left(X_2 + X_3 \Gamma_3^m S_3 \left(S_2\right)^{-1} \left(\Gamma_2^m\right)^{-1}\right) \Gamma_2^m S_2\right]. \tag{2.13}$$

3. 滤波向量数组的条件数界限。 本节基于第 2 节引入的谱分解，推导一系列定理和命题，为 $\mathrm{p}_m(A)V$的条件数给出上界。我们首先考察一个较简单的情形，即对 m 中的所有向量统一应用相同的多项式次数 V。随后，我们将分析推广到更一般的情形，即针对每个滤波向量分别优化多项式次数。最后，我们给出一个通用表达式，用于在实际情形中估算条件数：此时部分向量已经收敛，因而已被消去并锁定。

3.1. 单一多项式次数且无锁定机制。

LEMMA 3.1. 根据方程(2.8) 中谱分解的矩阵元定义，可得以下表达式：

1. $\kappa_2(X_\ell \Gamma_\ell^m) = \kappa_2(\Gamma_\ell^m) = \frac{|\rho_1|^m}{|\rho_\ell|^m}$
2. $\kappa_2(S_\ell) = 1$

3. $\|\Gamma_\ell^m S_\ell\|_2 \le 1$4. $\|S_\ell^+ (\Gamma_\ell^m)^{-1}\|_2 \le \frac{|\rho_1|^m}{|\rho_\ell|^m}$ ——

证明。 一般矩形矩阵 M 的条件数通过伪逆M^+定义为 $\mathrm{cond}_p(M) \doteq \|M\|_p\|M^+\|_p$。具体而言，使用 2-范数（$p=2$）时， $\kappa_2(M) = \frac{\sigma_{\max}(M)}{\sigma_{\min}(M)}$，其中$\sigma_{\max}(M)$ 和 $\sigma_{\min}(M)$ 分别是 M的最大和最小非零奇异值。

有了上述定义，引理中的每个结论都可以通过几步推导得出。

1. 根据定义， X_ℓ 是酉矩阵，因此 $\|X_\ell\Gamma_\ell^m\|_2 = \|\Gamma_\ell^m\|_2$ 并且

$$\kappa_2(X_\ell\Gamma_\ell^m) = \kappa_2(\Gamma_\ell^m) = \frac{\sigma_1(\Gamma_\ell^m)}{\sigma_\ell(\Gamma_\ell^m)} = \frac{|\rho_1|^m}{|\rho_\ell|^m}.$$

2. 根据 S_ℓ的定义， $\kappa_2(S_\ell) = \|X_\ell^H V\|_2\|(X_\ell^H V)^{-1}\|$。由于按构造， V 要么是独立同分布的（在子空间迭代 0 时），要么是正交归一的（在后续子空间迭代中），因此 $\|X_\ell V\|_2 = \|V\|_2 =$1. 对 S_ℓ的逆的范数也有相同结论。

3. 第三个陈述可由一组简单的关系式推出

$$\|\Gamma_\ell^m S_\ell\|_2 \le \|\Gamma_\ell^m\|_2\|S_\ell\|_2 = \|\Gamma_\ell^m\|_2 = \sigma_{\max}(\Gamma_\ell^m) = 1,$$

其中使用了施瓦茨不等式、陈述 2. 中的中间结果，以及下式的定义 Γ不等式(2.9，其值从左到右按降序排列。

4. 与前一个陈述类似，最后一个陈述也可由一组简单的不等式推出

$$\|S_\ell^+ (\Gamma_\ell^m)^{-1}\|_2 \le \|S_\ell^+\|_2\|(\Gamma_\ell^m)^{-1}\|_2 = \|(\Gamma_\ell^m)^{-1}\|_2 = \frac{|\rho_1|^m}{|\rho_\ell|^m}.$$ □

虽然上述引理的结果很简单，但在 $p=2$ 范数的情况下尤其重要，这归功于以下适用于两个 A 且维度相符的矩形矩阵 B 的不等式

$$\kappa_2(AB) \le \kappa_2(A)\,\kappa_2(B). \tag{3.1}$$

例如，根据第 3. 和第 4. 个陈述以及条件数的定义，可以自动得到 $\kappa_2(\Gamma_\ell^m S_\ell) \le \frac{|\rho_1|^m}{|\rho_\ell|^m}$。此外，现在我们可以陈述并证明下面的命题。

Proposition 3.2. 给定一个方形厄米矩阵 A，我们使用由 $p_m(A)$定义的 m 次切比雪夫多项式，对正交归一向量数组 V进行滤波。对于给定的 m，可以构造由该多项式滤波得到的子空间，使得 $1 < |\rho_\ell|^m < 1+\sqrt{2}$，则对于常数$\eta > 1$，以下不等式成立

$$\kappa_2\left(p_m(A)V\right) \le \eta|\rho_1|^m \tag{3.2}$$

证明。 为简化起见，定义 $E^m \doteq \Gamma_3^m S_3\ (S_\ell)^{-1}\ (\Gamma_\ell^m)^{-1}$，并将方程（2.8）再次写成

$$\mathrm{p}_m(A)V = (X_\ell + X_3E^m)\Gamma_\ell^m S_\ell$$

在大型 m 极限下，我们预期 $X_\ell + X_3E^m \longrightarrow X_\ell$。对于某些有限但足够大的 m 值，可以将表达式 X_3E^m 视为 X_ℓ 的扰动，并写成 $X_3E^m = -\delta X_\ell$，使得

$$\mathrm{p}_m(A)V = (\mathbb{1} - \delta X_\ell X_\ell^H)X_\ell\Gamma_\ell^m S_\ell = (\mathbb{1} - D)X_\ell\Gamma_\ell^m S_\ell.$$

由方程（3.1）可得

$$\kappa_2\left(\mathrm{p}_m(A)V\right) \le \kappa_2(\mathbb{1}-D)\,\kappa_2(X_\ell\Gamma_\ell^m S_\ell) \le \kappa_2(\mathbb{1}-D)\frac{|\rho_1|^m}{|\rho_\ell|^m}.$$

根据条件数 $\kappa_2(\mathbb{1}-D) = \|(\mathbb{1}-D)\|_2\|(\mathbb{1}-D)^{-1}\|_2$ 的定义，我们需要求出右侧各项 2-范数的界。由三角不等式 $\|\mathbb{1}-D\|_p \le 1+\|D\|_p$，再由 Golub Van-Loan（第三版）的引理 2.3.3 [17] 可知， $\|(\mathbb{1}-D)^{-1}\|_p \le (1-\|D\|_p)^{-1}$ 若 $\|D\|_p < 1$。接下来只需推导范数 D 的界。将其写成显式形式会很有用

$$-D = X_3\Gamma_3^m S_3 S_\ell^{-1}(\Gamma_\ell^m)^{-1}X_\ell^H = \left(X_3\Gamma_3^m X_3^H\right)VV^H\left(X_\ell\Gamma_\ell^m X_\ell^H\right)^{-1},$$

其中在最后一个等式中，我们使用了 S 矩阵的定义。对 D 取范数，并依次使用三角不等式和引理 3.1中的结果，可得以下不等式

$$\|D\|_2 \leq \|X_3\Gamma_3^m X_3^H\|_2 \|VV^H\|_2 \| \left(X_\ell \Gamma_\ell^m X_\ell^H\right)^{-1} \|_2 \leq \|\Gamma_3^m\|_2 \| \left(\Gamma_\ell^m\right)^{-1} \|_2 = |\rho_\ell|^{-m}.$$

根据假设， $|\rho_\ell|^m$ 始终是一个大于 1 的数，因此我们可以将最后这个结果代入 $\kappa_2(\mathbb{1} - D)$ 的定义，并得到

$$\kappa_2(\mathrm{p}_m(A)V) \leq \frac{|\rho_1|^m}{|\rho_\ell|^m} \cdot \frac{|\rho_\ell|^m + 1}{|\rho_\ell|^m - 1}. \tag{3.3}$$

在命题 $\eta \doteq \frac{|\rho_\ell|^m+1}{|\rho_\ell|^m}(|\rho_\ell|^m-1) > 1$ 的条件下，证明得证。 □

REMARK: 在实际情况中， $|\rho_\ell|^m$ 始终是一个大于 *1* 的数，但几乎从不大于 *2*。这是因为子空间的选取使得特征值 λ_ℓ 非常接近谱的滤波区间端点，但始终包含在该区间内。后一条件保证了 $|\rho_\ell|^m > 1$，而特征值接近端点则确保其不会过大，且始终远小于 $|\rho_1|^m$。因此，参数 η 大于 *1*，但不会大太多。我们将在数值实验中看到，选择 $\eta = 1$ 并使用简化版的界是相当稳妥的。

3.2 优化多项式次数。 正如我们在第2.2节末尾所见，当多项式次数被优化时， m 便成为一个大小为 ℓ 的向量，其中的整数值随向量索引增大而越来越大。在这种情况下，引理3.1 可以通过增加一个简单条件而容易地加以推广。

LEMMA 3.3. 给定方程(2.8) 的谱分解中矩阵元的定义，其中各 Γ由方程*2.10*表示，并且附加条件为 $|\rho_i|^{m_i}|\rho_1|^{(m_{i+1}-m_i)} \geq |\rho_{i+1}|^{m_{i+1}}$ 则下列表达式成立：

1. $\kappa_2(X_\ell\Gamma_\ell^m) = \kappa_2(\Gamma_\ell^m) = \frac{|\rho_1|^{m_\ell}}{|\rho_\ell|^{m_\ell}}$
2. $\kappa_2(S_\ell) = 1$
3. $\|\Gamma_\ell^m S_\ell\|_2 \leq 1$
4. $\|S_\ell^+ \left(\Gamma_\ell^m\right)^{-1}\|_2 \leq \frac{|\rho_1|^{m_\ell}}{|\rho_\ell|^{m_\ell}}$

证明。 命题 2 和 3 的证明与引理 3.1完全相同。其余命题唯一的区别在于 Γ的结构。只要满足引理的附加条件， Γ的对角元素值仍按降序排列，因此证明中相同的步骤便可得出所述结果。

□

对于命题 3.2， 也可得到类似的结果，此时 $m \to m_\ell$。最终结果是，在 $1 < |\rho_\ell|^{m_\ell} < 1 + \sqrt{2}$的条件下，

$$\kappa_2\left(\mathrm{p}_m(A)V\right) \leq \eta|\rho_1|^{m_\ell} \tag{3.4}$$

其中 $\eta \doteq \frac{|\rho_\ell|^{m_\ell}+1}{|\rho_\ell|^{m_\ell}}(|\rho_\ell|^{m_\ell}-1) > 1$。该命题的证明与命题 3.2 的证明步骤完全相同，此处不再赘述。关键在于认识到矩阵 $\|D\| \leq |\rho_\ell|^{m_\ell}$的范数。与注 3.1类似，只要 m_ℓ 的值为 $\mathcal{O}(10)$， $|\rho_\ell|^{m_\ell}$ 就是一个大于 1 但仍具有 $\mathcal{O}(1)$的数，实际应用中可以忽略。

3.3. 向量的锁定与消去。 在这种情况下，对 V的条件数进行上界估计会复杂一些，并且需要理解公式（2.13）中所示谱分解的条件数如何受到其各个组成部分的影响。在尝试进行严格分析之前，先进行一个简单的直观考察是有益的。数组 Z 的前 k 列由一个性质非常良好的子数组 X_1Q_1构成，其条件数按构造很可能已经尽可能接近于 1。事实上，该数组表示 Z 中已经收敛并被消去的部分。相反，子数组（$X_2\Gamma_2^m S_2 + X_3\Gamma_3^m S_3$）包含了 Z 中可能存在问题的部分。因此，可以预期只有这部分会对 Z的条件数产生贡献。为了更严格地说明这一点，我们陈述并证明以下命题。

THEOREM 3.4. 给定两个矩形矩阵 A 和 B，满足 $A \in \mathbb{C}^{n\times q}$ 和 $B \in \mathbb{C}^{n\times p}$， $n > q + p$和 $A^HB = 0$（或等价地 $B^HA = 0$）， $M \doteq [A\ B] \in \mathbb{C}^{n\times\ell}$的条件数为

$$\kappa_2(M) \leq \frac{\sigma_1(A) + \sigma_1(B)}{\inf\{\sigma_q(A), \sigma_p(B)\}} \tag{3.5}$$

证明。 根据定义， M 的条件数为

$$\kappa_2(M) = \|M\|_2\|M^+\|_2 = \frac{\sigma_1(M)}{\sigma_\ell(M)}$$

其中，与之前一样， M^+ 表示 M 和 $\ell = q + p$的伪逆。下面一系列不等式成立

$$\sigma_1(M) = \|M\|_2 = \sqrt{\lambda_1(MM^H)} = \|AA^H + BB^H\|_2 \le \|AA^H\|_2 + \|BB^H\|_2 = \sigma_1(A) + \sigma_1(B)$$

其中 λ_1 和 σ_1 分别是最大特征值和最大奇异值。伪逆的范数则稍显棘手。由于 M 具有 ℓ 个线性无关列，伪逆的定义为$M^+ =(M^HM)^{-1}M^H$。以分块形式表示，这变为

$$M^+ = \left[\begin{pmatrix} A^H \\ B^H \end{pmatrix} \begin{pmatrix} A & B \end{pmatrix}\right]^{-1} \begin{pmatrix} A^H \\ B^H \end{pmatrix} = \begin{pmatrix} A^HA & 0 \\ 0 & B^HB \end{pmatrix}^{-1} \begin{pmatrix} A^H \\ B^H \end{pmatrix}$$

$$= \begin{pmatrix} (A^HA)^{-1}A^H & 0 \\ 0 & (B^HB)^{-1}B^H \end{pmatrix} = \begin{pmatrix} A^+ & 0 \\ 0 & B^+ \end{pmatrix}$$

其中，第一步使用了命题的假设，第二步利用了分块对角矩阵的逆等于各分块的逆这一事实。借助该表达式，我们现在可以容易地证明

$$[\sigma_\ell(M)]^{-1} = \sigma_1(M^+) = \left\| \begin{pmatrix} A^+ & 0 \\ 0 & B^+ \end{pmatrix} \right\|_2 \le \sup\{\|A^+\|_2, \|B^+\|_2\} = \sup\{\sigma_1(A^+), \sigma_1(B^+)\}$$

$$= \sup\{\sigma_q(A)^{-1}, \sigma_p(B)^{-1}\} = [\inf\{\sigma_q(A), \sigma_p(B)\}]^{-1}$$

中间的不等式可通过矩阵范数的定义 $\|M^+\|_2 =\sup_{\|v\|=1} \|M^+v\|_2$，并取如下定义的向量 v 来推导

$$v = \begin{pmatrix} ax \\ by \end{pmatrix} \text{ with } \quad a^2 + b^2 = 1 \text{ and } \quad \|x\| = \|y\| = 1$$

使得

$$\|M^+\|_2^2 = a^2 \sup \|A^+x\|_2^2 + b^2 \sup \|B^+y\|_2^2 \le (a^2 + b^2) \sup\{\|A^+\|_2^2, \|B^+\|_2^2\} = \sup\{\|A^+\|_2, \|B^+\|_2\}^2 \quad \square$$

定理 3.4 的陈述可直接应用于 Z （2.13）的矩阵，其中 $M = Z$， $A = X_1Q_1$和 $B = X_2\Gamma_2^mS_2 + X_3\Gamma_3^mS_3$ ，相应地 $q = k$ 和 $p = \ell - k$。由于 $A = X_1Q_1$ 表示消胀并锁定的向量数组，因此按构造它是一个酉矩阵。因此， σ_k（X_1Q_1） $\sim\sigma_1$（X_1Q_1） $= \|X_1Q_1\|_2 = 1$。反之

$$\sigma_{\ell-k}^2(B) = \lambda_{\ell-k}\left(S_2^H(\Gamma_2^m)^H\Gamma_2^mS_2 + S_3^H(\Gamma_3^m)^H\Gamma_3^mS_3\right) = \inf_{\|v\|_2=1} v^H\left(S_2^H(\Gamma_2^m)^H\Gamma_2^mS_2 + S_3^H(\Gamma_3^m)^H\Gamma_3^mS_3\right)v$$

$$= \inf_{\|v\|_2=1}(v^HS_2^H(\Gamma_2^m)^H\Gamma_2^mS_2v) + \inf_{\|v\|_2=1}(v^HS_3^H(\Gamma_3^m)^H\Gamma_3^mS_3v)$$

$$= \inf_{\|z\|_2=1}(z^H(\Gamma_2^m)^H\Gamma_2^mz) + \inf_{\|u\|_2=1}(u^HS_3^H(\Gamma_3^m)^H\Gamma_3^mS_3u) = \sigma_{\ell-k}^2(\Gamma_2^m) + \sigma_{n-\ell}^2(\Gamma_3^m)$$

由于 Γ_2^m和 Γ_3^m是正定的，且 $\sigma_{\ell-k}(\Gamma_2^m) = \frac{|\rho_\ell|^{m_\ell}}{|\rho_1|^{m_\ell}} \ll 1$ 和 $\sigma_{n-\ell}(\Gamma_3^m) = \frac{1}{|\rho_1|^{m_\ell}} \ll 1$ ，因此

$$\sigma_{\ell-k}(\Gamma_2^mS_2 + \Gamma_3^mS_3) \le \sigma_{\ell-k}(\Gamma_2^m) + \sigma_{n-\ell}(\Gamma_3^m) \ll 1.$$

同样，可以论证 $\sigma_1((\Gamma_2^mS_2 + \Gamma_3^mS_3) \gg 1$。综合所有部分，可得

$$\kappa_2(Z) \le \frac{1 + \sigma_1((\Gamma_2^mS_2 + \Gamma_3^mS_3)}{\sigma_{\ell-k}(\Gamma_2^mS_2 + \Gamma_3^mS_3)} \approxeq \kappa_2(\Gamma_2^mS_2 + \Gamma_3^mS_3)$$

COROLLARY 3.5. 给定由方阵 A生成的次数为 m 的多项式滤波器 $p_m(A)$，以及由消去子空间和锁定子空间 W 与滤波子空间 $p_m(A)V$ 的并集构成的子空间 Z，对于 $1 < |\rho_\ell|^{m_\ell} < 1 + \sqrt{2}$ 和常数 $\eta > 1$，以下不等式成立

$$\kappa_2(Z) \lesssim \eta |\rho_{k+1}|^{m_{k+1}} |\rho_1|^{m_\ell - m_{k+1}} \tag{3.6}$$

证明。 推论的证明完全沿用命题 3.2中的步骤。我们从式（3.3）出发，将条件数写成若干条件数的乘积

$$\kappa_2\left(\Gamma_2^m S_2 + \Gamma_3^m S_3\right)) \leq \kappa_2(\mathbb{1} - D)\, \kappa_2(X_2 \Gamma_2^m S_2)$$

唯一的区别在于现在 $D = X_3 \Gamma_3^m S_3 S_2^{-1} (\Gamma_2^m)^{-1} X_2^H$。由于存在锁定与消去机制， Γ_2^m的渐近形式与 Γ_ℓ^m不同，即

$$\Gamma_2^m = \operatorname{diag}\left(\frac{|\rho_{k+1}|^{m_{k+1}}}{|\rho_1|^{m_{k+1}}}, \ldots, \frac{|\rho_\ell|^{m_\ell}}{|\rho_1|^{m_\ell}}\right).$$

因此， $\|D\|_2$ 可以用 $|\rho_\ell|^{-m_\ell}$ 进行上界估计，如3.2 所示，而

$$\kappa_2(X_2 \Gamma_2^m S_2) \leq \frac{|\rho_{k+1}|^{m_{k+1}} |\rho_1|^{m_\ell - m_{k+1}}}{|\rho_\ell|^{m_\ell}}$$

综合以上各项可得

$$\kappa_2(Z) \lesssim \kappa_2(\Gamma_2^m S_2 + \Gamma_3^m S_3) \leq \frac{|\rho_\ell|^m + 1}{|\rho_\ell|^m (|\rho_\ell|^m - 1)} \cdot |\rho_{k+1}|^{m_{k+1}} |\rho_1|^{m_\ell - m_{k+1}}. \tag{3.7}$$

再次，在推论的条件下，可以定义 $\eta \doteq \frac{|\rho_\ell|^m + 1}{|\rho_\ell|^m \, |\rho_\ell|^m - 1}() > 1$，从而完成证明。

□

4. 数值实验。 ChASE 的数值行为和基准测试是在德国于利希超级计算中心的 JURECA-DC 集群上进行的，该集群拥有 480 个标准同构计算节点。每个节点配备两颗主频为 2.25 GHz 的 64 核 AMD EPYC 7742 处理器（16×32 GB DDR4 内存），互连网络采用 InfiniBand HDR100 Mellanox Connect-X6。ChASE 使用 JURECA-DC 上的默认软件栈进行编译。C/C++ 编译器为 GCC 13.3.0，MPI 库为 OpenMPI 5.0.5，BLAS/LAPACK 库为 Intel MKL 2024.2.0。

对于 JURECA-DC 上的测试，每个节点的 MPI 进程数和每个进程的 OpenMP 线程数分别为 4 和 32。残差的阈值容差 tol 固定为 10^{-10}，除非另有说明，否则始终启用 ChASE 滤波器的次数优化。本文中的所有测试均采用双精度进行。

4.1. ChASE 中的并行方案。 ChASE 采用混合并行化策略，旨在高效处理分布式内存系统上的大规模特征值问题。切比雪夫滤波器在二维 MPI 进程网格上应用；在可能的情况下，进程数会被选为构成近似方形的网格，以平衡通信与计算。滤波过程利用目标矩阵的厄米或对称结构，优化矩阵–矩阵乘法内核。根据滤波器的多项式次数，该运算可能会沿二维网格的行通信器或列通信器执行集体全归约操作。

相比之下，ChASE 中的 QR 分解是在从二维网格派生的一维列通信器上执行的。每个列通信器都包含与矩阵垂直分区对应的一组 MPI 进程。在每个通信器内，QR 分解以并行方式执行，而不同列通信器上的分解结果彼此冗余。

这种并行性的划分——滤波采用二维并行，QR 分解采用一维并行——确保计算量占主导的滤波器和数值上至关重要的 QR 分解都能实现高效率。所选方案也能自然地适配基于矩阵结构优化的滤波器配置。关于并行设计、实现方案选择以及采用列冗余 QR 分解的原因，详情请参阅 [37]。

4.2. 测试矩阵套件：DFT 和贝特-萨尔彼得方程矩阵。在数值测试和基准测试中，我们使用了一组来自领域应用的特征值问题。表4.1列出了这些问题的基本信息，包括简称、规模、所需特征对的数量、ChASE 中使用的额外搜索空间大小、用于提取这些特征对的应用软件、各问题的类型、nev——实际应用所要求计算的特征对数量，以及nex——外部搜索空间的大小。

FLEUR 问题由 FLEUR 代码 [35], 生成，该代码实现了一种全电子全势线性增强平面波（FLAPW）方法。在这种方法中，波函数在缀加球内用增强平面波展开，在间隙区内用平面波展开，从而形成精度很高但计算量也很大的基组。相比之下，FHI-aims 的 DFT 计算 [6] 采用数值原子中心轨道（NAO）作为基函数。这些基函数严格局域于原子周围，能够高效地计算积分。因此，与基于 FLAPW 的 FLEUR 矩阵相比，其哈密顿矩阵和重叠矩阵具有不同的谱性质。具体而言，由于轨道具有局域性，FHI-aims 矩阵的带宽往往较小，特征值也更为聚集；而 FLEUR 矩阵通常更稠密，特征值在整个谱中的分布也更均匀。

BSE UIUC 问题取自伊利诺伊大学厄巴纳-香槟分校开发的耶拿 BSE 代码分支 [39], 。该代码以底层 DFT 哈密顿量为基础，计算电子-空穴激发。这些问题通常呈现簇集的特征值谱，对应于低能激发，这给迭代特征值求解器带来了独特的挑战。

通过纳入来自两种采用不同离散化方案的应用代码的 DFT 计算特征值问题，以及 BSE 模拟中的特征值问题，该测试套件涵盖了多种谱特征、矩阵规模和数值复杂性，可全面基准测试 ChASE 的性能与稳健性。由于 FHI-aims 和耶拿 BSE 的矩阵谱通常更稠密、聚集程度更高，因此需要相对于对应的 nev使用相对更大的搜索空间 nex（见表 4.1）。

4.3. 精确条件数与估计值。 本节将展示第 3 节中针对切比雪夫滤波器滤波后的矩形矩阵条件数所引入上界的有效性[3]。此处考虑的测试问题列于表 4.1。结果见图 4.1，其中比较了每次 ChASE 迭代中的条件数估计值 κ_2^{est}与滤波矩阵条件数的精确计算值 κ_2 ，直至收敛。

我们在 JURECA-DC 集群上使用 4 个节点进行了测试。所有测试均在启用（*opt*）或禁用（*no-opt*）切比雪夫多项式次数优化的情况下进行，以突出条件数估计对优化机制的依赖关系。对于*no-opt*，每次迭代的切比雪夫多项式次数均固定为 20。对于 *opt*，首次迭代的初始次数为 20，之后的迭代中则针对每个滤波向量动态调整次数，并将最大次数限制为 36，以避免过度的病态性。

如前几节所定义，给定一个矩形矩阵 M ，其条件数 κ_2 等于 $\sigma_{max}(M)\sigma_{min}(M)$，其中 $\sigma_{max}(M)$ 和 $\sigma_{min}(M)$分别表示M的最大和最小奇异值。本测试中，矩阵 M 沿列通信器分布，其奇异值通过 ScaLAPACK 并行奇异值分解例程直接计算，具体使用PDGESVD 和PZGESVD 分别处理实数和复数双精度情形。估计得到的条件数 κ_2^{est}采用最一般的形式，由公式（1.2）给出；当禁用次数优化（*no-opt*）时，该式可简化为 $|\rho_{k+1}|^m$ 。无消去和锁定的情形（公式 1.1）会在最初的 1–2 次迭代中自动恢复，此时尚无任何特征对收敛。

图 4.1 表明，无论是否进行次数优化，估算的条件数始终大于等于计算得到的条件数，因此是一种有效且可靠的估算。对于许多情况， κ_2^{est}与 κ_2 之比始终小于 2。对于某些情况（例如 AuAg 13k 或 TiO_2 12k 的 $\mathcal{O}$（10^4）opt 情况），在最初的几次迭代中，由于尚未锁定任何特征对，该比值可能大于 $\mathcal{O}$（10^4）。这可能是因为 ChASE 在前 2 到 3 次迭代中对参数 e 和 c 的估算可能相当不准确。无论其严格程度如何，κ_2^{est}始终为 κ_2 提供上界，因此它是一个可靠的参数，可供 ChASE 用于从较稳定的 QR 分解（如 s-乔列斯基 QR 分解）切换到稳定性较差但效率更高的 CholeksyQR2，甚至在最后一两次迭代中切换到 CholeskyQR1，此时估算值为 $\mathcal{O}$（1）。

[3]如果某些特征 对已锁定的特征对而言，这应是已锁定的特征向量与经过滤波的未收敛特征向量的并集

表 4.1：数值测试用的 DFT 矩阵列表。

Name	N	nev	nex	问题类型	代码来源	Type
NaCl 9k	9273	256	60	DFT	**FLEUR**	厄米
AuAg 13k	13379	972	100	DFT	**FLEUR**	厄米
TiO2 12k	12455	1076	100	DFT	**FLEUR**	厄米
TiO2 29k	29528	2560	400	DFT	**FLEUR**	厄米
HfO2 62k	62681	100	100	BSE	**BSE UIUC**	厄米
In2O3 76k	76887	120	60	BSE	**BSE UIUC**	厄米
In2O3 115k	115459	2400	1600	BSE	**BSE UIUC**	厄米
Si Graphene 50k	51476	2400	1600	DFT	**FHI-aims**	对称
Cu_2BaSnS_4 80k	80250	3600	2400	DFT	**FHI-aims**	对称

在$no\text{-}opt$ 情况下，最高的条件数出现在第一次迭代。具体而言，在图4.1中，步骤 0 对应于对随机生成的初始猜测向量进行 QR 分解，这发生在首次应用切比雪夫滤波器之前。因此，如果第一次迭代时 M 的条件数低于某个阈值，那么后续迭代中就可以避免使用 s-乔列斯基 QR2。相反，在opt 情况下，早期阶段 M 的条件数可能远高于第一次迭代时的条件数。这是由于次数优化过程中允许的最大次数较高所致，用户可以通过为次数优化设置相对较小的最大次数来控制这一影响。不过请注意，采用次数优化的 ChASE 始终比不采用次数优化的 ChASE 收敛快得多（例如 In_2O_3 115k）。总之，我们对滤波矩阵条件数的上界估算确保了所提出的 QR 变体切换启发式方法在 ChASE 的整个执行周期中都可靠。

4.4. ChASE 与 CholeskyQR 和 Householder QR 的比较。本节比较 ChASE 在每次迭代中采用 Householder QR（HHQR）分解，或根据第1.3节介绍的启发式方法自动选择 QR 变体时的数值表现。此处，HHQR 指 ScaLAPACK 提供的 Householder QR 实现，该实现基于一维 MPI 网格，在每个列通信器上独立执行。ScaLAPACK 块循环分布的块大小设为 64。本节所有测试均在 JURECA-DC 集群的 4 个节点上运行，并使用所有可用的 CPU 核心。报告的每个数值均为 5 次重复运行的平均值。

测试问题如表4.1所列。所有情况下，待分解矩阵的维度均为$N \times (\mathsf{nev} + \mathsf{nex})$。表4.2 报告了矩阵-向量乘法（MatVecs）的总次数[4] , 、收敛所需的 ChASE 迭代次数、总求解时间以及 QR 分解耗时。结果表明，采用 CholeskyQR 的 ChASE 与 HHQR 相比，收敛数值结果相同：所有测试问题的迭代次数和 MatVec 运算次数实际上均保持不变。主要差异在于性能：CholeskyQR 始终能减少 QR 分解耗时，对于中等规模问题（$\mathsf{nev} \lesssim 1000$），加速比为 2-6×；对于较大型问题（$\mathsf{nev} \gtrsim 2000$），加速比最高可达 4-，从而根据问题不同，将总求解时间缩短 10–20%。

总之，这些结果证实，在第3节引入的条件数界指导下，用CholeskyQR替换稳定的HHQR，能够保持ChASE的数值行为，同时显著减少计算量，尤其是对于所需特征对数量较多的问题。详细的并行扩展性与性能分析将在下一节讨论。

4.5 强扩展性。 我们通过以下方式评估 ChASE 的强扩展行为：$_2O_3$ 115k 特征值问题，定义见表 4.1 ，采用三种特征对配置：1）**小型**($\mathsf{nev} = 1200$，$\mathsf{nex} = 600$）；2）**中等**($\mathsf{nev} = 2400$，$\mathsf{nex} = 1200$）；3）**大型** ($\mathsf{nev} = 3600$，$\mathsf{nex} = 2400$）。本节始终使用这些标签来区分问题规模。所有运行均在 JURECA-DC 集群上进行，计算节点数从 16 扩展至 100 个，每个节点配备 4 个消息传递接口进程，每个消息传递接口进程包含 32 个 OpenMP 线程。

[4]在每次ChASE 迭代中，切比雪夫滤波器作用于矩形矩阵 M ，其中每一列可使用不同的多项式次数。矩阵-向量乘法（矩阵-向量乘法）的次数在所有未收敛特征向量和所有迭代次数上累加，从而衡量滤波器中的计算量，并揭示各个滤波向量的收敛行为。

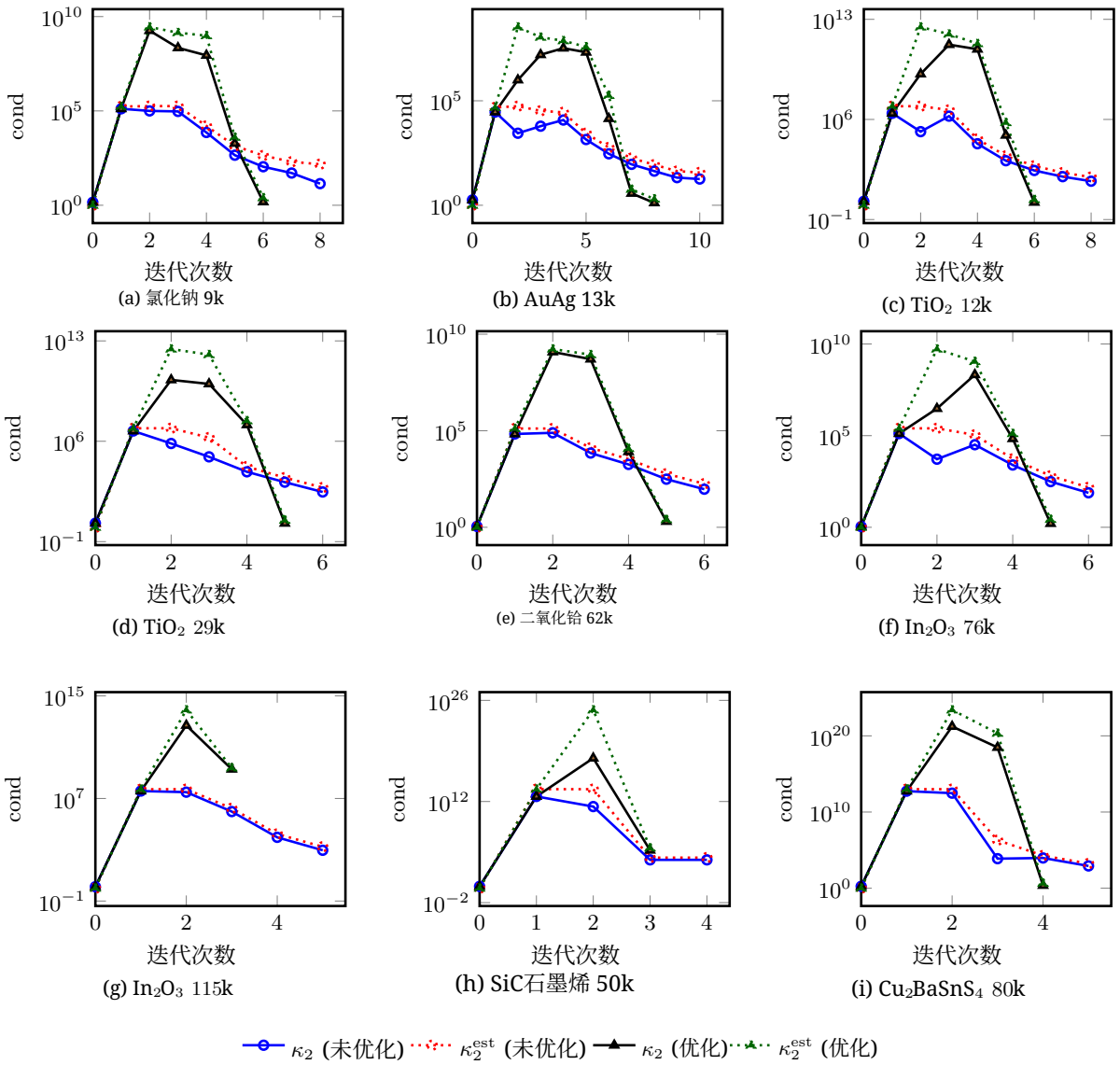


图 4.1：各种材料在优化和未优化情况下每次迭代的条件估计。子图 (a)–(i) 显示了不同材料体系的结果。

每项基准测试所用的节点数都经过选择，以便在 ChASE 中构成方形MPI进程网格，这通常有利于提升性能。

图 4.2 同时报告了总求解时间与并行效率，以及专门针对使用 CholeskyQR 和 Householder QR 进行 QR 分解的时间与效率。在 ChASE 中，QR分解仅在列通信器内执行，即全局二维MPI网格的一个子集。例如，在一次采用 64 个节点、共计 256 个 MPI进程的运行中，列通信器通常仅覆盖 16 个 MPI进程；因此，QR分解的计时与效率反映了这一较小通信器的扩展行为。

在所有配置中，随着消息传递接口进程数增加，总运行时间持续缩短，表明具有良好的强扩展。当使用 CholeskyQR 时，对于最小规模的情况（nev + nex =1800），总时间从 16 个节点上的约 534 秒降至 100 个节点上的 124 秒。对于最大规模的配置（nev+nex = 6000），运行时间从 16 个节点上的约 1890 秒降至 100 个节点上的 702 秒。相对于最小节点数的并行效率会随着通信开销增加而适度下降，这符合预期：在 100 个节点上，它从最小配置的 0.62–0.72 降至最大配置的 0.40- 0.43。请注意，中等和大型 nev 的并行效率下降得更快

表 4.2：ChASE-CPU 与 HHQR 和 CholeksyQR 的比较

Type	QR 实现	矩阵-向量乘法	迭代次数	总计（秒）	QR分解（秒）
NaCl 9k	HHQR	34,206	6	8.64	1.34
	CholeskyQR	34,212	6	8.14	0.51
AuAg 13k	HHQR	106,542	8	40.13	8.72
	CholeskyQR	106,618	8	32.96	1.72
TiO2 12k	HHQR	94,724	6	32.47	7.62
	CholeskyQR	93,904	6	26.05	1.92
TiO2 29k	HHQR	212,024	5	323.21	81.56
	CholeskyQR	211,984	5	278.11	16.23
HfO2 62k	HHQR	19,614	5	201.41	5.91
	CholeskyQR	19,614	5	195.33	1.95
In2O3 76k	HHQR	15,980	5	284.62	6.97
	CholeskyQR	15,976	5	278.25	1.81
In2O3 115k	HHQR	229,714	3	2533.05	184.65
	CholeskyQR	229,716	3	2439.24	74.91
Si Graphene 50k	HHQR	240,554	3	263.33	53.98
	CholeskyQR	240,536	3	208.91	8.90
Cu_2BaSnS_4 80k	HHQR	413,084	4	911.62	139.72
	CholeskyQR	429,620	4	747.72	30.11

，因为 QR 分解和 Rayleigh-Ritz过程整体上变得更加重要，而且其扩展性不如多项式滤波器。

就 QR 分解而言，CholeskyQR 版本在所有问题规模上始终优于 Householder QR。在最大配置（nev +nex $= 6000$）中，当在列通信器内的 20 个 MPI进程上执行时，CholeskyQR 在 100 秒内完成，而 Householder QR 需要 146 秒，加速比约为 $1.46\times$。对于中等配置，加速比甚至可达 2-3×，而对于最小的情形，该加速比可超过 $4\times$。由于 QR分解是每次 ChASE 迭代的主要计算组成部分之一，QR分解时间的这些缩减直接促成总运行时间的显著下降，尤其是对于大型特征值问题。

QR分解的并行效率也体现了这些改进。在所有节点数量和问题规模上，CholeskyQR 的效率均高于 Householder QR。例如，在最大配置中，CholeskyQR 在列通信器内的 8-20 个 MPI进程上保持约 0.79-0.83 的效率，而 Householder QR 的效率下降更为急剧，约为 0.53–0.61。这一差异反映了 CholeskyQR 较低的通信开销和更简单的计算结构，使其即使在节点数量增加时也能在列通信器内更有效地扩展。

重要的是，QR 变体的选择基本不会影响 ChASE 的收敛行为：迭代总次数和矩阵-向量乘法次数几乎完全相同。这些结果证实，采用启发式选择的 CholeskyQR 变体可以安全地取代 Householder QR，而不损害数值稳健性，同时还能提高并行效率、缩短 QR 执行时间并减少总求解时间。

5. 总结与结论。 在本研究中，我们探讨如何估计一组向量 V 经过多项式滤波器 $\mathrm{p}_m(A)$变换后的条件数。我们的研究缘于这样一种可能性：用计算效率更高的 CholeskyQR 算法取代传统的 Householder QR 分解。尽管 CholeskyQR 性能更优且能发挥更高的并行性，但众所周知，当向量矩阵 $\mathrm{p}_m(A)V$ 的条件数过大时，它会出现正交性损失。为缓解这一问题，可以采用更稳健的变体，例如 CholeskyQR2；对于病态性更严重的情况，则可采用移位 CholeskyQR2算法。然而，这些方法各自仅能在 $\mathrm{p}_m(A)V$的条件数处于特定范围内时保持可靠。因此，要有效使用这些方法，需要事先了解或估计这一数值。

直接计算 计算 $\mathrm{p}_m(A)V$ 的精确条件数将极其昂贵，并且

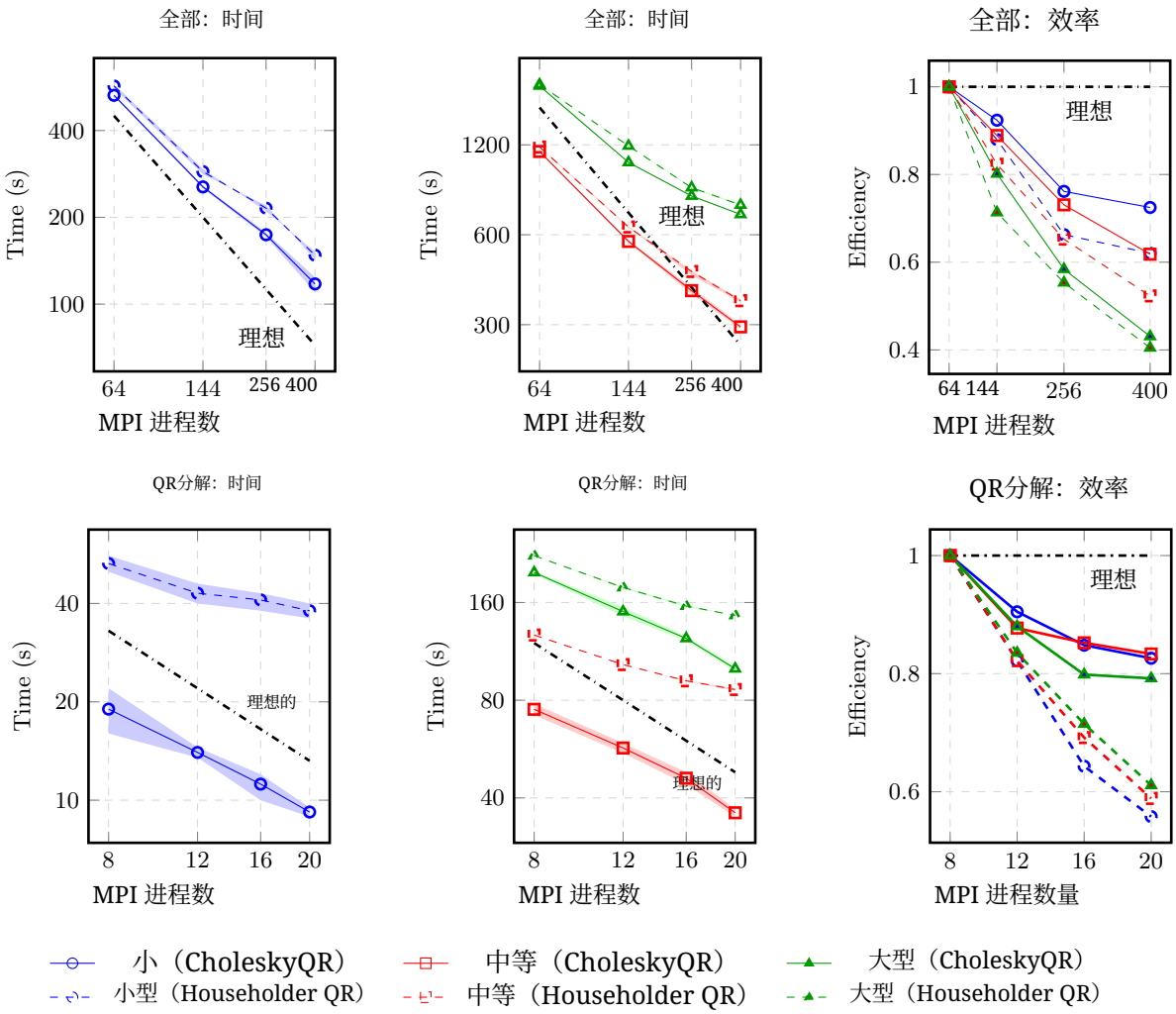


图4.2：不同配置下 CholeskyQR 与 Householder QR 的强扩展性能和并行效率对比。每项测量重复进行五次。运行时间图显示平均执行时间，阴影带表示各次运行中的最小值至最大值范围。并行效率根据平均运行时间计算。

会损害基于 CholeskyQR 的方法所带来的性能优势。因此，本文通过对滤波向量矩阵的谱性质进行细致的数值分析来解决该问题。该分析促使我们提出一种简单且计算开销低的条件数估计。所提出的估计仅依赖于 $\mathrm{p}_m(A)V$ 中所选列向量的收敛比，以及施加于它们的多项式滤波器的次数 m_i。基于该估计，我们推导出一个实用模型，用于动态选择最合适的 QR 分解变体，从而在保持数值精度的同时最大化性能与可扩展性。所得策略已纳入 ChASE 库的 1.6.0 版本，并通过数值实验得到了充分验证。

致谢。本研究部分由 Gauss 超级计算中心（www.gauss-centre.eu）通过 SiVeGCS 项目以及虚拟材料设计联合实验室（JL-VMD）提供支持。该实验室是亥姆霍兹协会设立于亥姆霍兹研究领域“信息”中的跨中心平台。我们感谢于利希超级计算中心提供计算时间，以便开展数值测试。特别感谢量子材料模拟与数据实验室的团队成员提供富有启发性的讨论，也感谢 FLEUR、BSE UIUC 和 FHI-aims 代码的维护者生成并提供数值测试所用特征值问题的矩阵。

参考文献

[1] Grey Ballard, James Demmel, Laura Grigori, Mathias Jacquelin, Nicholas Knight, and Hong Diep Nguyen, *Reconstructing householder vectors from tall-skinny qr*, 《并行与分布式计算杂志》, 85（2015）, 第 3–31 页。[2] Amartya S. Banerjee, Lin Lin, Wei Hu, Chao Yang, and John E. Pask, *Chebyshev polynomial filtered subspace iteration in the discontinuous galerkin method for large-scale electronic structure calculations*, 《化学物理杂志》, 145（2016）, 第 154101 页。[3] Amartya S. Banerjee, Lin Lin, Phanish Suryanarayana, Chao Yang, and John E. Pask, *Two-level chebyshev filter based complementary subspace method: Pushing the envelope of large-scale electronic structure calculations*, 《化学理论与计算杂志》, 14（2018）, 第 2930–2946 页。[4] Friedrich L. Bauer, *Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme* 8 1957 214 235 , 《应用数学与物理杂志 ZAMP》, （）, 第 – 页。[5] Mario Berljafa, Daniel Wortmann and Edoardo Di Napoli *An optimized and scalable eigensolver for sequences* , 《并发与计算：实践与经验》, *of eigenvalue problems* 27 2015 905 922 （）, 第 – 页。[6] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler, *Ab initio molecular simulations with numeric atom-centered orbitals*, 《Computer Physics Communications》, 180（2009）, 第 2175–2196 页。[7] Maurice Clint and A. Jennings *The evaluation of eigenvalues and eigenvectors of real symmetric matrices by simultaneous iteration* 13 1970 76 80 , 《计算机杂志》, （）, 第 – 页。[8] Michel Crouzeix, Bernard Philippe, and Miloud Sadkane, *The Davidson method*, 《SIAM 科学计算杂志》, 15（1994）, 第 62–76 页。[9] Jane Cullum and W Donath, *A block Lanczos algorithm for computing the q algebraically largest eigenvalues and a corresponding eigenspace of large, sparse, real symmetric matrices*, 载于 1974 年 IEEE 决策与控制会议（含第 13 届自适应过程研讨会）, IEEE, 1974 年, 第 505–509 页。[10] Jane K Cullum and Ralph A Willoughby *Lanczos Algorithms for Large Symmetric Eigenvalue Computations Volume 1* 1985 , 《Theory》, Birkhauser, 。[11] Ernest R Davidson,*The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices* 17 1975 87 94 , 《计算物理学杂志》, （）, 第 – 页。[12] James Demmel, Laura Grigori Mark Hoemmen and Julien Langou *Communication-avoiding parallel and sequential qr factorizations* 0806 2159, 2008 , CoRR abs/。（）。[13] Edoardo Di Napoli and Mario Berljafa, *Block Iterative Eigensolvers for Sequences of Correlated Eigenvalue Problems* 2012 , arXiv.org, （）。[14] Edoardo Di Napoli, Xinzhe Wu, and Clement Richefort, *The ChASE library*。https://github.com/ ChASE-library/ChASE, 2026 年。访问日期：2026-03-10。[15] Takeshi Fukaya, Ramaseshan Kannan, Yuji Nakatsukasa, Yusaku Yamamoto, and Yuka Yanagisawa, *Shiftedcholesky qr for computing the qr factorization of ill-conditioned matrices*, 《SIAM 科学计算杂志》, 42（2020）, 第 A477–A503 页。[16] Takeshi Fukaya, Yuji Nakatsukasa, Yuka Yanagisawa, and Yusaku Yamamoto, *Choleskyqr2: a simple and communication-avoiding algorithm for computing a tall-skinny qr factorization on a large-scale parallel system*, 载于 2014 年第五届大规模系统可扩展算法最新进展研讨会, IEEE, 2014 年, 第 31–38 页。[17] Gene H. Golub and C. F. Van Loan, *Matrix Computations*, 约翰斯·霍普金斯大学（巴尔的摩和伦敦）, 第三版, 2012 年。[18] Moritz Kreutzer, Dominik Ernst, Alan R. Bishop, Holger Fehske, Georg Hager, Kengo Nakajima, andGerhard Wellein, *Chebyshev filter diagonalization on modern manycore processors and gpgpus*, 载于《高性能计算》, Rio Yokota、Mich`ele Weiland、David Keyes 和 Carsten Trinitis 编, 瑞士沙姆, 2018 年, Springer 国际出版公司, 第 329–349 页。[19] Leeor Kronik, Adi Makmal, Murilo L Tiago, M M G Alemany, Manish Jain, Xiangyang Huang, YousefSaad, and James R Chelikowsky, *PARSEC – the pseudopotential algorithm for real-space electronic structure calculations: recent advances and novel applications to nano-structures*, 《固体物理学现状（b）》, 243（2006）, 第 1063–1079 页。[20] Antoine Levitt and Marc Torrent, *Parallel eigensolvers in plane-wave Density Functional Theory*, 《Computer Physics Communications》, 187（2015）, 第 98–105 页。[21] Edoardo Di Napoli, Clément Richefort and Xinzhe Wu *Chebyshev accelerated subspsace eigensolver for pseudo-* 。[22] Beresford N. Parlett, *hermitian hamiltonians* 2026 *The Symmetric Eigenvalue Problem*, SIAM, 费城, 宾夕法尼亚州, 1998 年 1 月。[23] B N Parlett and D S Scott, *The Lanczos algorithm with selective orthogonalization*, 《计算数学》, 33（1979）, 第 217–238 页。[24] Aldo H. Romero, Douglas C. Allan, Bernard Amadon, Gabriel Antonius, Thomas Applencourt, LucasBaguet, Jordan Bieder, François Bottin, Johann Bouchet, Eric Bousquet, Fabien Bruneval, Guillaume Brunin Damien Caliste Michel Côté Jules Denier Cyrus Dreyer Philippe Ghosez Matteo Giantomassi Yannick Gillet Olivier Gingras Donald R. Hamann Geoffroy Hautier , François Jollet, Gérald Jomard Alexandre Martin Francesco Naccarato Guido Petretto , Henrique P. C. Miranda, Nicholas A. Pike Valentin Planes Sergei Prokhorenko Tonatiuh Rangel Fabio Ricci Gian-Marco , Rignanese Miquel Royo Massimiliano Stengel Marc Torrent Michiel J. van Setten , Benoit Van Troeye Matthieu J. Verstraete Julia Wiktor Josef W. Zwanziger and Xavier Gonze , *Abinit: Overview and focus on selected capabilities* 152 2020 124102 , 《化学物理杂志》, （）, 第 页。[25] Axel Ruhe, *Rational Krylov sequence methods for eigenvalue computation*, 《线性代数及其应用》, 58（1984）, 第 391–405 页。[26] Heinz Rutishauser, *Computational aspects of FL Bauer's simultaneous iteration method*, Numerische Mathematik, 13（1969）, 第 4–13 页。

[27] , *~~Simult~~aneous iteration method for symmetric matrices*, Numerische Mathematik, 16 (1970), 第 205–223 页。[28] H Rutishauser, *Simultaneous iteration method for symmetric matrices*,《自动计算手册》, (1971年)。[29] Yousef Saad, *Numerical methods for large eigenvalue problems*, SIAM, 费城, 宾夕法尼亚州, 2011 年。[30] G. W. Stewart, *Accelerating the orthogonal iteration for the eigenvectors of a Hermitian matrix*, Numerische Math- ematik, 13 (1969年), 第 362–376 页。[31] , *Simultaneous iteration for computing invariant subspaces of non-Hermitian matrices*, Numerische Mathematik, 25 (1976), 第 123–136 页。[32] , *Matrix algorithms. Vol. II*, SIAM, 费城, 宾夕法尼亚州, 2001 年。[33] , *A Krylov–Schur Algorithm for Large Eigenproblems*, SIAM 矩阵分析与应用杂志, 23 (2002), 第 601–614 页。[34] J. Winkelmann, P. S~~pringer~~, and E. Di Napoli, *ChASE: Chebyshev Accelerated Subspace Iteration Eigensolver for Sequ~~ences of~~ Hermitian Eigenvalue Problems* 45 2019 1 34 , ACM 数学软件汇刊, ( ), 第 – 页。[35] Daniel Wortmann, Gregor Michalicek Nadjib Baadji Wejdan Beida Markus Betzinger Gustav , , , , , , , , , , , Zenodo, Bihlmayer Thomas Bornhake Jens Bröder Tobias Burnus Jussi Enkovaara et al. *Fleur* (2023)。[36] K Wu and H Simon, *Thick-restart Lanczos method for large symmetric eigenvalue problems*, SIAM 矩阵分析与应用杂志, (2000)。[37] X. Wu, D. Davidović, S Achilles, and E. Di Napoli, *Chase: a distributed hybrid cpu-gpu eigensolver for large-scale hermitian eigenvalue problems*, 载于《高级科学计算平台会议论文集》, 2022 年, 第 1–12 页。[38] Xinzhe Wu and Edoardo Di Napoli, *Advancing the distributed multi-gpu chase library through algorithm optimiza-tion and nccl library*, 载于 SC-W '23: 《国际高性能计算、网络、存储与分析会议 SC '23 研讨会论文集》, SC-W '23, 美国纽约, 2023 年, 计算机协会, 第 1688–1696 页。[39] X. Zhang, S. Achilles, J. Winkelmann, R. Haas, A. Schleife, and E. Di Napoli, *Solving the bethe-salpeter equation on massively parallel architectures* 267 2021 108081 , Computer Physics Communications, ( ), 第 页。[40] Yunkai Zhou, James R. Chelikowsky, and Yousef Saad *Chebyshev-filtered subspace iteration method free of sparse* , , 《计算物理学杂志》, *diagonalization for solving the Kohn–Sham equation* 274 2014 770 782 ( ), 第 – 页。[41] Yunkai Zhou, Yousef Saad *Self-consistent-field calculations using* , Murilo L. Tiago, *Chebyshev-filtered subspace iteration* 219 2006 172 184 and James R. Chelikowsky, ,《计算物理学杂志》, ( ), 第 – 页。