Lanczos Method for Eigensystems

Aiichiro Nakano

Collaboratory for Advanced Computing & Simulations
Department of Computer Science
Department of Physics & Astronomy
Department of Quantitative & Computational Biology
University of Southern California

Email: anakano@usc.edu

- 1. O(N) (vs. conventional $O(N^3)$) eigensolver
- 2. Krylov subspace



B. N. Parlett

The Symmetric Eigenvalue Problem
(Prentice-Hall, '80) Secs. 11-13



Rayleigh Quotient

Theorem

Let A be an $n \times n$ real symmetric matrix, $\lambda_1[A] \leq ... \leq \lambda_n[A]$ its eigenvalues in ascending order, $x \in \mathbb{R}^n$, & the Rayleigh quotient

$$\rho(\mathbf{x}; \mathbf{A}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \quad \text{then} \quad \begin{cases} \lambda_1[\mathbf{A}] = \min_{\mathbf{x} \in \mathbb{R}^n} \rho(\mathbf{x}; \mathbf{A}) \\ \lambda_n[\mathbf{A}] = \max_{\mathbf{x} \in \mathbb{R}^n} \rho(\mathbf{x}; \mathbf{A}) \end{cases}$$

Proof

Let $\mathbf{q}^{(k)}$ be the k-th orthonormalized eigenvector of \mathbf{A} , $\mathbf{A}\mathbf{q}_k = \lambda_k \mathbf{q}_k$, & orthogonal transformation matrix, $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n]$, then

$$\mathbf{Q}^T \mathbf{A} \mathbf{Q} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$$

Let x = Qz (note $Q^TQ = I$), then

$$\rho(\mathbf{x}; \mathbf{A}) = \frac{\mathbf{z}^T \mathbf{Q}^T \mathbf{A} \mathbf{Q} \mathbf{z}}{\mathbf{z}^T \mathbf{Q}^T \mathbf{Q} \mathbf{z}} = \frac{z_1^2 \lambda_1 + \dots + z_n^2 \lambda_n}{z_1^2 + \dots + z_n^2}$$

which is a weighted average of $\lambda_1, ..., \lambda_n$, & the minimum is when $\mathbf{z}^T = (1,0,...,0) = \mathbf{e}_1$ & $\mathbf{x} = \mathbf{Q}\mathbf{e}_1 = \mathbf{q}_1$.

Rayleigh-Ritz Procedure

Theorem

Let $\{q_1,...,q_m\}$ $(q_j \in \mathbb{R}^n; j = 1,...,m; m < n)$ be an orthonormal set, so that any vector $\mathbf{x} \in \mathbb{R}^n$ in the range is expressed as a linear combination of $q_1,...,q_m$:

$$\mathbf{x} = z_1 \mathbf{q}_1 + \dots + z_m \mathbf{q}_m \quad \text{or} \quad n \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = n \begin{bmatrix} \mathbf{q}_1 & \dots & \mathbf{q}_m \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix} m = \mathbf{Q} \mathbf{z}$$

$$\mathbf{x} \in \text{rank-}m \text{ subspace} \subset \mathbf{R}^n$$

then the best approximations for $\lambda_1[A]$ & $\lambda_n[A]$ are obtained by diagonalizing

$$m \times m \quad m \times n \, n \times n \, n \times m$$

$$\mathbf{H} = \mathbf{Q}^T \quad \mathbf{A} \quad \mathbf{Q}$$

as $\lambda_1[H] \& \lambda_m[H]$.

Variation in a vector subspace!

Proof

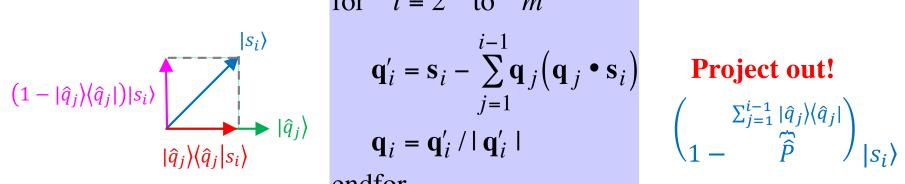
Note
$$(\mathbf{Q}^T\mathbf{Q})_{ij} = \sum_{k=1}^n Q_{ki}Q_{kj} = \mathbf{q}_i \cdot \mathbf{q}_j = \delta_{ij}$$
 orthonormal basis set

then
$$\rho(\mathbf{x}; \mathbf{A}) = \frac{\mathbf{z}^T \mathbf{Q}^T \mathbf{A} \mathbf{Q} \mathbf{z}}{\mathbf{z}^T \mathbf{Q}^T \mathbf{Q} \mathbf{z}} = \frac{\mathbf{z}^T \mathbf{H} \mathbf{z}}{\mathbf{z}^T \mathbf{z}} = \frac{z_1^2 \lambda_1(\mathbf{H}) + \dots + z_m^2 \lambda_m(\mathbf{H})}{z_1^2 + \dots + z_m^2}$$

the minimum of which is $\lambda_1[H]$ (cf. proof in the previous page).

Orthogonalization by QR Decomposition

• Gram-Schmidt orthonormalization: The orthonormal set Q required for the Rayleigh-Ritz procedure is obtained starting from an arbitrary set of m vectors, $S = [s_1...s_m]$ ($s_i \in \mathbb{R}^n$) as (see supplementary note):



$$\mathbf{q}_{1} = \mathbf{s}_{1} / |\mathbf{s}_{1}|$$
for $i = 2$ to m

$$\mathbf{q}'_{i} = \mathbf{s}_{i} - \sum_{j=1}^{i-1} \mathbf{q}_{j} (\mathbf{q}_{j} \cdot \mathbf{s}_{i})$$

$$\mathbf{q}_{i} = \mathbf{q}'_{i} / |\mathbf{q}'_{i}|$$
endfor

$$\begin{pmatrix} \sum_{j=1}^{i-1} |\hat{q}_j\rangle\langle\hat{q}_j| \\ 1 - \hat{\hat{P}} \end{pmatrix} |s_i\rangle$$

• The Gram-Schmidt procedure amounts to QR decomposition, S = QR, where R is an $m \times m$ right-triangle matrix:

$$n \begin{bmatrix} \mathbf{s}_{1} & \mathbf{s}_{2} & \mathbf{s}_{3} & \mathbf{s}_{4} \end{bmatrix} = n \begin{bmatrix} \mathbf{q}_{1} & \mathbf{q}_{2} & \mathbf{q}_{3} & \mathbf{q}_{4} \end{bmatrix} \begin{bmatrix} \mathbf{q}_{1} & \mathbf{q}_{1} \cdot \mathbf{s}_{2} & \mathbf{q}_{1} \cdot \mathbf{s}_{3} & \mathbf{q}_{1} \cdot \mathbf{s}_{4} \\ 0 & |\mathbf{q}_{2}| & \mathbf{q}_{2} \cdot \mathbf{s}_{3} & \mathbf{q}_{2} \cdot \mathbf{s}_{4} \\ 0 & 0 & |\mathbf{q}_{3}| & \mathbf{q}_{3} \cdot \mathbf{s}_{4} \\ 0 & 0 & 0 & |\mathbf{q}_{4}| \end{bmatrix} m$$

$$\therefore \mathbf{s}_{i} = |\mathbf{q}_{i}'| \mathbf{q}_{i} + \sum_{j=1}^{i-1} \mathbf{q}_{j} (\mathbf{q}_{j} \cdot \mathbf{s}_{i}) \qquad \text{cf. QR decomposition } A = Q[]$$

$$\mathbf{s}_i = |\mathbf{q}_i'|\mathbf{q}_i + \sum_{j=1}^{t-1} \mathbf{q}_j(\mathbf{q}_j \bullet \mathbf{s}_i)$$

$$\mathbf{cf. QR decomposition}$$

$$A = Q$$

Rayleigh-Ritz Algorithm

- 1. Start from $S = [s_1...s_m]$ $(s_j \in \mathbb{R}^n)$ & do Gram-Schmidt orthonormalization, $S = \mathbb{Q}\mathbb{R}$, to obtain an orthonormal set $\mathbb{Q} = [\mathbf{q}_1...\mathbf{q}_m]$
- 2. Form $\mathbf{H} = \mathbf{Q}^{\mathrm{T}} \mathbf{A} \mathbf{Q}$
- 3. Diagonalize **H** to get $\lambda_1[\mathbf{H}], \dots, \lambda_m[\mathbf{H}]$: $\mathbf{Hg}_k = \lambda_k[\mathbf{H}]\mathbf{g}_k \ (k = 1, \dots, m)$
- **4.** Approximations of $\lambda_1[\mathbf{A}]$ & $\lambda_n[\mathbf{A}]$ are given by $\lambda_1[\mathbf{H}]$ & $\lambda_m[\mathbf{H}]$ with the corresponding eigenvectors, $\mathbf{y}_k = \mathbf{Q}\mathbf{g}_k$ (k = 1 & m).

* $\mathbf{Q}\mathbf{Q}^T \neq \mathbf{I}^{n \times n}$ but spans a subspace of the *n*-dimensional space cf. $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}^{m \times m}$ — orthonormal in \mathbf{R}^m subspace but not complete in total \mathbf{R}^n space

Davidson method augments the vector subspace by residual, $\mathbf{r}_k = \mathbf{A}\mathbf{y}_k - \lambda_k \mathbf{y}_k$.

Krylov Subspace

• Krylov subspace S_m is spanned by a Krylov matrix, $K^m(f) = [f Af ... A^{m-1}f]$ $(\mathbf{f} \in \mathbf{R}^n)$

Theorem

Let Q_m be the orthonormal basis obtained by QR factorization, $K_m(f) =$ $Q_m R$, then $T_m = Q_m^T A Q_m$ is a *tridiagonal* matrix

Proof (see supplementary note)

For i > j+1, $q_i^T(Aq_j) = 0$, since $Aq_j \subset S_{j+1}$ by construction & $q_i \perp S_{j+1}$ by Gram-Schmidt orthonormalization for i > j+1. By the symmetry of A, $q_i^T(Aq_i) = q_i^T(A^Tq_i) = q_i^T(Aq_i) = 0 \text{ for } j > i+1 \text{ or } i < j-1.$

$$\mathbf{T}_{m} = \begin{bmatrix} \alpha_{1} & \beta_{1} & & & \\ \beta_{1} & \alpha_{2} & \beta_{2} & & \\ & \ddots & \ddots & & \\ & & \beta_{m-2} & \alpha_{m-1} & \beta_{m-1} \\ & & & \beta_{m-1} & \alpha_{m} \end{bmatrix} \quad \begin{cases} \alpha_{j} = \mathbf{q}_{j}^{T} \mathbf{A} \mathbf{q}_{j} & j = 1, \dots, m \\ \beta_{j} = \mathbf{q}_{j+1}^{T} \mathbf{A} \mathbf{q}_{j} & j = 1, \dots, m-1 \end{cases}$$

$$\begin{cases} \alpha_j = \mathbf{q}_j^T \mathbf{A} \mathbf{q}_j & j = 1, ..., m \\ \beta_j = \mathbf{q}_{j+1}^T \mathbf{A} \mathbf{q}_j & j = 1, ..., m - 1 \end{cases}$$

Tridiagonal matrix can be diagonalized in O(N) time cf. tqli() in Numerical Recipes

Alexei Krylov with daughter Anna, later Anna Kapitsa, wife of Pyotr Kapitsa (1904)

Recursion Formula

• Due to the tridiagonality, Aq_i is a linear combination of q_{i-1} , $q_i & q_{i+1}$:

$$\mathbf{A}\mathbf{q}_i = \beta_{i-1}\mathbf{q}_{i-1} + \alpha_i\mathbf{q}_i + \beta_i\mathbf{q}_{i+1} \quad 2 \le i \le m-1$$

If we define $q_0 = 0$, the above equation is valid for i = 1 as well. Let $r_i = \beta_i q_{i+1}$ (r_i is a component of Aq_i orthogonal to q_i for $j \le i$), then

$$\mathbf{r}_{i} = \mathbf{A}\mathbf{q}_{i} - \beta_{i-1}\mathbf{q}_{i-1} - \alpha_{i}\mathbf{q}_{i} \qquad 1 \le i \le m-1$$

$$\mathbf{A}\mathbf{q}_{i} = \beta_{i-1}\mathbf{q}_{i-1} + \alpha_{i}\mathbf{q}_{i} + \beta_{i}\mathbf{q}_{i+1} \qquad 2 \le i \le m-1$$

• Lanczos algorithm:

Given
$$\mathbf{r}_{0}$$
, $\beta_{0} = \|\mathbf{r}_{0}\|$ ($\mathbf{q}_{0} = 0$)
for $i = 1, ..., m$
 $\mathbf{q}_{i} \leftarrow \mathbf{r}_{i-1}/\beta_{i-1}$
 $\mathbf{r}_{i} \leftarrow \mathbf{A}\mathbf{q}_{i} - \beta_{i-1}\mathbf{q}_{i-1}$
 $\alpha_{i} \leftarrow \mathbf{q}_{i}^{T}\mathbf{r}_{i} : \mathbf{q}_{i}^{T}(\mathbf{A}\mathbf{q}_{i} - \beta_{i-1}\mathbf{q}_{i-1}) = \mathbf{q}_{i}^{T}\mathbf{A}\mathbf{q}_{i} = \alpha_{i} \text{ (orthogonality)}$
 $\mathbf{r}_{i} \leftarrow \mathbf{r}_{i} - \alpha_{i}\mathbf{q}_{i}$
 $\beta_{i} = \|\mathbf{r}_{i}\|$ (only when $i \leq m-1$) $\beta_{i} = \mathbf{q}_{i+1}^{T}\mathbf{A}\mathbf{q}_{i}$
endfor

Keep increasing m until $\lambda_1[T_m]$ converges

Application of Rayleigh-Ritz/Lanczos

- Search for transition states (with a negative eigenvalue of the Hessian matrix, $\partial^2 E/\partial r_i \partial r_i$) by following the eigenvector with the smallest eigenvalue
 - -Rayleigh-Ritz: Kumeda, Wales & Munro, Chem. Phys. Lett. 341, 185 ('01)
 - **Lanczos:** Mousseau et al., J. Mol. Graph. Model. 19, 78 ('01)

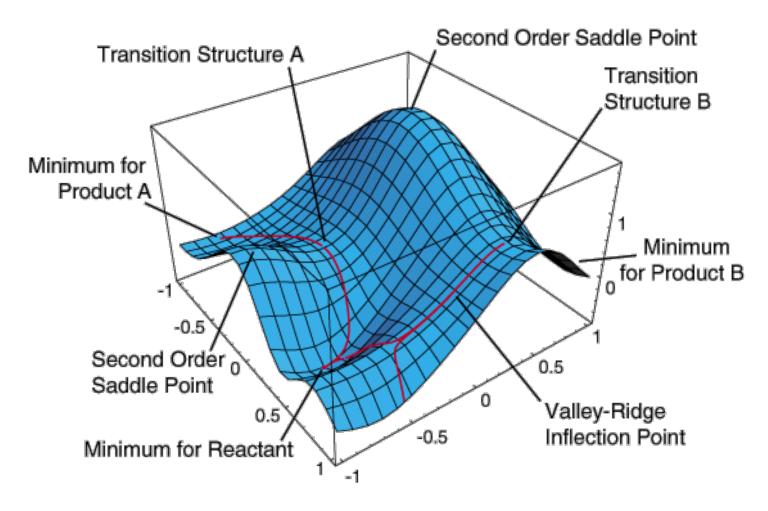


Figure from Prof. H. B. Schlegel; http://chem.wayne.edu/schlegel

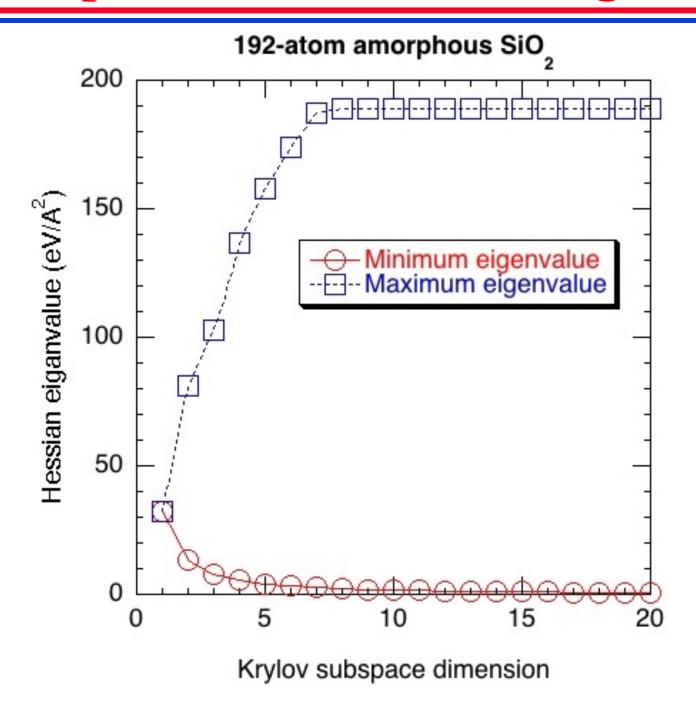
Lanczos Algorithm for Hessian Calculation

A. Nakano / Computer Physics Communications 176 (2007) 292-299

```
Algorithm Lanczos
Input:
    \mathbf{R} \in \mathbb{R}^{3N}: a state
    logical initialize: TRUE for the first call in each event generation; FALSE otherwise
    \lambda_1: the minimum eigenvalue of the Hessian matrix, \mathbf{H}(\mathbf{R}) = \frac{\partial^2 V}{\partial \mathbf{R}^2}
    \mathbf{V}^1 \in \mathbb{R}^{3N}: the Hessian eigenvector corresponding to \lambda_1
Steps:
    if initialize
        randomize \Delta \in \mathbb{R}^{3N}, such that it contains no translational motion
    s \leftarrow 0
    \beta^s \leftarrow \|\Delta\|
    \mathbf{Q}^s \ (\in \mathbb{R}^{3N}) \leftarrow 0
                                                                                                              \vec{F}(\vec{R} + \vec{Q}) = \vec{F}(\vec{R}) + \underbrace{\partial \vec{F}/\partial \vec{R}}^{-\vec{H}(\vec{R})} \bullet \vec{Q}
          s \leftarrow s + 1
          \mathbf{Q}^s \leftarrow \Delta/\beta^{s-1}
          c_{\text{fd}} \leftarrow \max_{i\alpha} \{|q_{i\alpha}^s| \mid i = 1, \dots, N; \alpha = x, y, z\}/\delta_{\text{fd}}
                                                                                                             : \vec{H}(\vec{R}) \bullet \vec{Q} = -\vec{F}(\vec{R} + \vec{Q}) + \vec{F}(\vec{R})
          \Delta \leftarrow c_{\text{fd}}[-\mathbf{F}(\mathbf{R} + \mathbf{Q}^s/c_{\text{fd}}) + \mathbf{F}(\mathbf{R})] - \beta^{s-1}\mathbf{Q}^{s-1}
          \alpha^s \leftarrow \mathbf{O}^{sT} \Delta
          \Delta \leftarrow \Delta - \alpha^s O^s
          \beta^s \leftarrow \|\Delta\|
         while |(\tilde{\lambda}_1^s - \tilde{\lambda}_1^{s-1})/\tilde{\lambda}_1^{s-1}| > \Delta_{eigen}
         \mathbf{V}^1 \leftarrow \sum_{k=1}^{s} \mathbf{Q}^k \tilde{q}_k^1
          \mathbf{V}^1 \leftarrow \mathbf{V}^1 / \|\mathbf{V}^1\|
```

^{*} diag $(\tilde{\lambda}_1^s, \dots, \tilde{\lambda}_s^s)$ is an s by s diagonal matrix, with its diagonal elements given by $\tilde{\lambda}_1^s, \dots, \tilde{\lambda}_s^s$. $\tilde{\mathbf{Q}}^s = [\tilde{\mathbf{q}}^1, \dots, \tilde{\mathbf{q}}^s]$ is an s by s orthogonal matrix, with $\tilde{\mathbf{q}}^m \in \mathbb{R}^s$ is the mth eigenvector of \mathbf{T}_s .

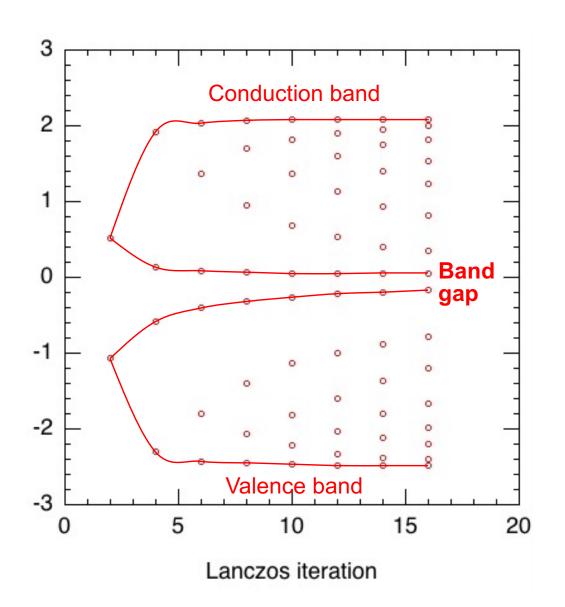
Sample Run of Lanczos Program



Electronic Energy Bands of GaAs

• 8-band k•p model

$$\begin{array}{c} A & 0 & V^* & 0 & \sqrt{3}V & -\sqrt{2}U & -U & \sqrt{2}V^* \\ 0 & A & -\sqrt{2}U & -\sqrt{3}V^* & 0 & -V & \sqrt{2}V & U \\ V & -\sqrt{2}U & -P+Q & -S^* & R & 0 & \sqrt{\frac{1}{2}} \ S & -\sqrt{2}Q \\ 0 & -\sqrt{3}V & -S & -P-Q & 0 & R & -\sqrt{2}R & \frac{1}{\sqrt{2}}S \\ -\sqrt{2}U & -V^* & 0 & R^* & 0 & -P-Q & S^* & \frac{1}{\sqrt{2}}S^* & \sqrt{2}R^* \\ -V^2U & -V^* & 0 & R^* & S & -P+Q & \sqrt{2}Q & \sqrt{\frac{3}{2}}S^* \\ -U & \sqrt{2}V^* & \sqrt{\frac{1}{2}}S^* & -\sqrt{2}R^* & \frac{1}{\sqrt{2}}S & \sqrt{2}Q & -P-\Delta & 0 \\ \sqrt{2}V & U & -\sqrt{2}Q & \frac{1}{\sqrt{2}}S^* & \sqrt{2}R & \sqrt{\frac{3}{2}}S & 0 & -P-\Delta \\ \end{array} \right) \\ A = E_c - \frac{\hbar^2}{2m_0}(\partial_x^2 + \partial_y^2 + \partial_z^2), \\ P = -E_v - \gamma_1 \frac{\hbar^2}{2m_0}(\partial_x^2 + \partial_y^2 + \partial_z^2), \\ Q = -\gamma_2 \frac{\hbar^2}{2m_0}(\partial_x^2 + \partial_y^2 - 2\partial_z^2), \\ R = \sqrt{3} \frac{\hbar^2}{2m_0} [\gamma_2(\partial_x^2 - \partial_y^2) - 2i\gamma_3\partial_x\partial_y], \\ S = -\sqrt{3} \gamma_3 \frac{\hbar^2}{m_0} \partial_z(\partial_x - i\partial_y), \\ U = \frac{-i}{\sqrt{3}} P_0 \partial_z, \\ V = \frac{-i}{\sqrt{6}} P_0(\partial_x - i\partial_y). \end{array}$$



C. Pryor, *Phys. Rev. B* 57, 7190 ('98)

Lanczos Program in Fortran

```
do s = 1,NWF
  q(:,:,:,s) = v/bet(s-1)
  call hamiltonian_op(q(:,:,:,s),hv) ! Operates Hamiltonian H on Q(S)
  v = hv-bet(s-1)*q(:,:,:,s-1)
  alp(s) = inner_product(q(:,:,:,s),v)
  v = v-alp(s)*q(:,:,:,s)
  bet(s) = sqrt(inner_product(v,v))
  call tridiag(eval,s) ! Diagonalize the S by S tridiagonal matrix
end do ! Lanczos iteration over s
```

```
Given \mathbf{r}_{0}, \beta_{0} = \|\mathbf{r}_{0}\| (\mathbf{q}_{0} = 0)

for i = 1, ..., m

\mathbf{q}_{i} \leftarrow \mathbf{r}_{i-1}/\beta_{i-1}

\mathbf{r}_{i} \leftarrow \mathbf{A}\mathbf{q}_{i} - \beta_{i-1}\mathbf{q}_{i-1}

\alpha_{i} \leftarrow \mathbf{q}_{i}^{T}\mathbf{r}_{i}

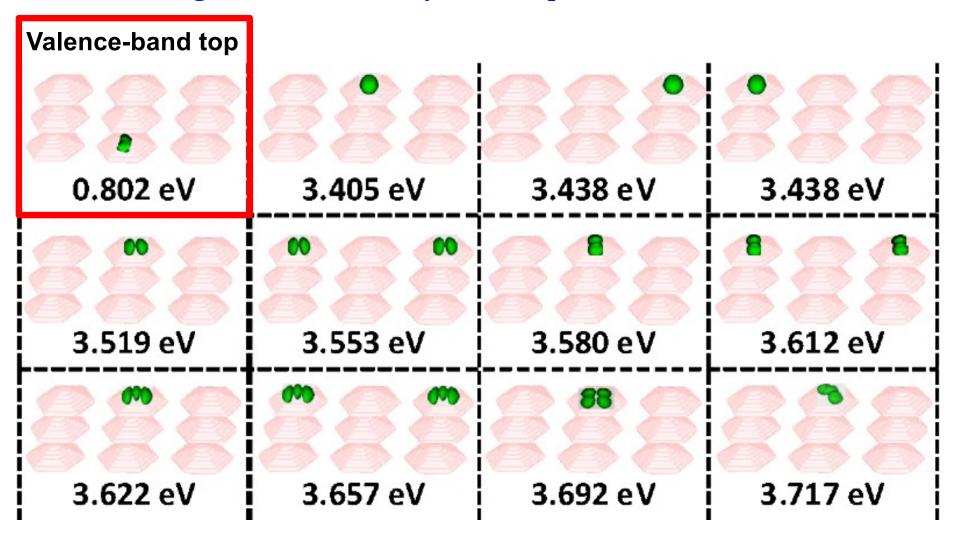
\mathbf{r}_{i} \leftarrow \mathbf{r}_{i} - \alpha_{i}\mathbf{q}_{i}

\beta_{i} = \|\mathbf{r}_{i}\| (only when i \leq m-1)

endfor
```

Band-edge Wave Functions

• Band-edge states in an array of GaN quantum dots in AlN matrix



Conduction-band states

S. Sburlan, Ph.D. dissertation, USC ('13)

Krylov-Subspace Class Project

Article

https://doi.org/10.1038/s41467-024-47685-8

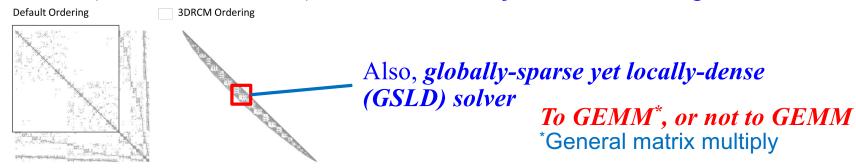
Scalable computation of anisotropic vibrations for large macromolecular

assemblies Jordy Homing Lam ® 1,2,3, Aiichiro Nakano ® 1,4,5 ⋈ & Vsevolod Katritch ® 1,2,3,6 ⋈

Nature Communications | (2024)15:3479

https://github.com/jhmlam/Inching

We map, optimize, and compare several low-complexity *Krylov-subspace eigensolvers*, supplemented by techniques such as *Chebyshev filtering*, sum decomposition, external explicit deflation and shift-and-inverse, to allow *fast GPU-resident calculations*. The method allows accurate calculation of the first 1000 vibrational modes of some largest structures in PDB (> 2.4 million atoms) at least *250 times faster* than existing methods.



Dimensionality reduction via Krylov subspace everywhere!

Example: Krylov subspace in quantum computing

Kirby et al., Quant. 7, 1018 ('23) Kim & Krylov, J. Phys. Chem. A 127, 6552 ('23) Tkachenko et al., Quant. Sci. Tech. 9, 035012 ('24)

Top 10 Algorithms in History

In putting together this issue of *Computing in Science & Engineering*, we knew three things: it would be difficult to list just 10 algorithms; it would be fun to assemble the authors and read their papers; and, whatever we came up with in the end, it would be controversial. We tried to assemble the 10 algorithms with the greatest influence on the development and practice of science and engineering in the 20th century. Following is our list (here, the list is in chronological order; however, the articles appear in no particular order):

- Metropolis Algorithm for Monte Carlo
- Simplex Method for Linear Programming
- Krylov Subspace Iteration Methods
- The Decompositional Approach to Matrix Computations
- The Fortran Optimizing Compiler
- QR Algorithm for Computing Eigenvalues
- Quicksort Algorithm for Sorting
- Fast Fourier Transform
- Integer Relation Detection
- Fast Multipole Method

IEEE CiSE, Jan/Feb (2000)

PHYS 516 CSCI 596 CSCI 653