# Computational Aspects Related to the Matrix Sign Function in Lattice QCD

Andreas Frommer
Bergische Universität Wuppertal
Fachbereich Mathematik und Naturwissenschaften
frommer@math.uni-wuppertal.de
http://www.math.uni-wuppertal.de/SciComp

◀◀
▶▶
◀
▶
Back
Close

# joint work with

Henk van der Vorst

Jasper van den Eshof

Katrin Schäfer

Thomas Lippert

Nigel Cundy

Stefan Krieg

Bruno Lang

Tilo Wettig

Jacques Bloch

Valeria Simoncini

◀◀
▶▶
◀
▶
Back
Close

# Outline

1. the setting
   - the Wilson fermion matrix
   - overlap fermions and the sign function
   - partial fraction expansions and multishift CG
2. inner-outer schemes
   - relaxation
   - recursive preconditioning
   - deflation
3. error estimates and bounds
   - Gaussian quadrature
   - estimates from CG

◀◀
▶▶
◀
▶
Back
Close

## 4. non-zero chemical potential

- the sign function revisited
- the Arnoldi process
- deflation
- outlook

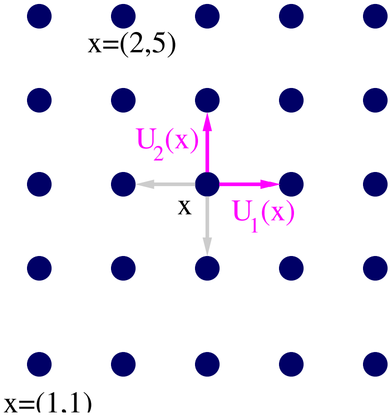
# 1. The Setting

## Wilson fermion matrix: intro

### Lattice Gauge Theory

- QCD = standard theory of strong interaction between quarks
- lattice gauge theory = discretization of QCD
- approximation of gauge fields by configurations $\mathcal{U}$ of gauge links

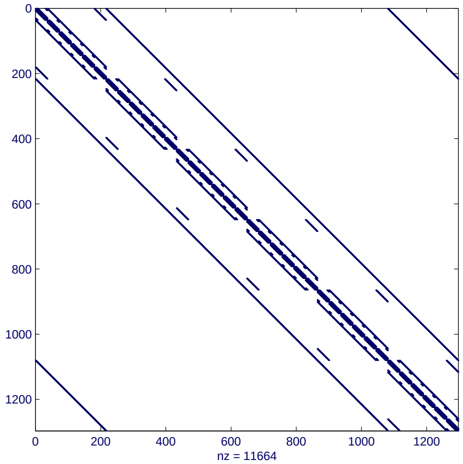
$$\mathcal{U} = \{U_\mu(x) \mid x \in G, \mu = 1, \ldots, 4\}.$$



Back
Close

# Wilson fermion matrix: details 1

- $M = I - \kappa D$
- $M \in \mathbb{C}^{n \times n}$
- nearest neighbor coupling on 4-dimensional torus
- 12 variables per grid point
- $n = 12 \cdot n_1 \cdot n_2 \cdot n_3 \cdot n_4$
- $n_i = 8 \ldots 128$



# Wilson fermion matrix: detail 2

$$(M\psi)_x = \psi_x - \kappa \left( \sum_{\mu=1}^{4} \left( (I - \gamma_\mu) \otimes U_\mu(x) \right) \psi_{x+e_\mu} + \sum_{\mu=1}^{4} \left( (I + \gamma_\mu) \otimes U_\mu^H(x - e_\mu) \right) \psi_{x-e_\mu} \right)$$

◀◀
▶▶
◀
▶
Back
Close

# Wilson fermion matrix: detail 2

$$(M\psi)_x = \psi_x - \kappa \left( \sum_{\mu=1}^{4} \left( (I - \gamma_\mu) \otimes U_\mu(x) \right) \psi_{x+e_\mu} + \sum_{\mu=1}^{4} \left( (I + \gamma_\mu) \otimes U_\mu^H(x - e_\mu) \right) \psi_{x-e_\mu} \right)$$

- $U_\mu(x) \in SU(3)$
- $\gamma_\mu \in \mathbb{C}^{4\times 4}$
- $I \pm \gamma_\mu$ is projector on 2-dimensional subspace
- $= \gamma_1\gamma_2\gamma_3\gamma_4$ satisfies $\gamma_5\gamma_\mu = \gamma_\mu\gamma_5 = 0$.

$$\gamma_5 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Back
Close

# Overlap fermions & sign function: intro

Chiral symmetry is an important physical property which should be reflected in the discretized operator.

- **Wilson fermion matrix:** No chiral symmetry
- **Ginsparg-Wilson relation (GW):** establishes a version of chiral symmetry on the lattice
- **Overlap fermions (Neuberger, 1998):** satisfy GW.

◀◀
▶▶
◀
▶
Back
Close

# Overlap fermions & sign function: overlap operator

## Neuberger's overlap operator:

$$
\begin{aligned}
N &= \rho \cdot I + M \cdot (M^H M)^{-1/2} \\
&= \rho \cdot I + \gamma_5 \cdot \mathsf{sign}(Q)
\end{aligned}
$$

where

- $Q = \gamma_5 \cdot M \quad \Rightarrow Q^H = Q$ hermitian Wilson matrix
- $\mathsf{sign}(Q) = V\mathsf{sign}(\Lambda)V^H$ where $Q = V\Lambda V^H$
- $\rho \geq 1$ ($\rho = 1$: massless operator)
- $\kappa = \frac{4}{3}\kappa_c$

◀◀
▶▶
◀
▶
Back
Close

# Overlap fermions & sign function: inner/outer

**Computational work in simulation**: solve

$$\begin{aligned} & N\psi = \phi \\ \Leftrightarrow \quad & (\rho \cdot I + \gamma_5 \mathsf{sign}(Q))\psi = \phi \end{aligned}$$

- $N$ is represented by a dense matrix
  $\Rightarrow$ cannot be determined explicitly
- nested iteration for

$$\underbrace{(\rho I + \gamma_5 \mathsf{sign}(Q))}_{=N}\psi = \phi$$

  – outer iteration: MVM with $N$
  – inner iteration: approximate $\mathsf{sign}(Q)b$ in $N \cdot b$

◀◀
▶▶
◀
▶
Back
Close

# Lanczos approach

Krylov subspace $K_m(Q,b) = \langle b, Qb, Q^2b, \ldots, Q^{m-1}b\rangle$

Lanczos method generates basis $v_1, \ldots, v_m$:

Put $V_m = [v_1|v_2|\ldots v_m]$. Then

$$QV_m = V_mT_m + \beta_{m+1}v_{m+1}e_m^T, \quad T_m \text{ tridiagonal } .$$

Note: $T_m = V_m^H Q V_m$

Approximate via the Galerkin approximation

$$\text{sign}(Q)b \approx V_m \text{sign}(T_m)e_1 \cdot \|b\|.$$

## Improvement:

- Lanczos for $Q^2$, start with $Qb$
- use $\text{sign}(t) = t \cdot (t^2)^{-1/2}$
- approximate $\text{sign}(Q)b = V_m(T_m)^{-1/2}e^1 \cdot \beta_0$

## Advantages:

- smooth convergence
- less vectors to store
- easily computable error bound

# Partial fraction expansions & multishift cg: Zolotarev

**Zolotarev:** $l_\infty$ best approx. of sign on $[-b,-a]\cup[a,b]$

Assume spec$(Q)\subset[-b,-a]\cup[a,b]$. Then

$$\begin{aligned} Z_p &= \delta\cdot Q\prod_{i=1}^{p-1}(Q^2+c_{2i}I)\cdot\prod_{i=1}^{p}(Q^2+c_{2i-1}I)^{-1} \\ &= \delta\cdot Q\sum_{i=1}^{p}\omega_i(Q^2+\tau_i I)^{-1}, \end{aligned}$$

where

$$c_i=\frac{\mathrm{sn}^2\left(iK/(2m);\sqrt{1-(b/a)^2}\right)}{1-\mathrm{sn}^2\left(iK/(2m);\sqrt{1-(b/a)^2}\right)},$$

$K$ is the complete elliptic integral.

# Partial fraction expansions & multishift cg: cg

$$\text{sign}(Q)v \approx \sum_{i=1}^{p} \omega_i Q \left(Q^2 - \sigma_i I\right)^{-1} v.$$

$(\sigma_i < 0)$.

Solve all $p$ systems $\left(Q^2 - \sigma_i I\right) x_i = v$ in one stroke (‘multishift CG’), since

$$K_m(Q^2, b) = K_m(Q^2 - \sigma_i I, b),\ i = 1, 2, \ldots, m.$$

◀◀
▶▶
◀
▶
Back
Close

## Summary of methods

1. both (Lanczos and multishift CG) compute $\mathrm{sign}(Q)b \approx p_m(Q)b$, $p_m$ polynomial
2. Zolotarev needs storage prop. to number of poles
3. Lanczos needs storage prop. to $m$
4. Lanczos adapts itself to $b$ (finite termination)
5. Zolotarev: converged systems can be removed for efficiency
6. both benefit from deflation

◀◀
▶▶
◀
▶
Back
Close

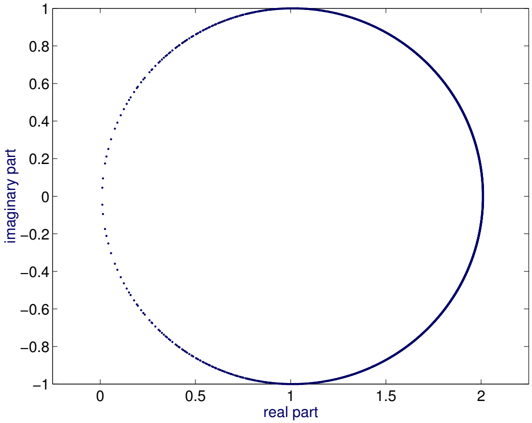
# 2. Inner-outer scheme

## SUMR: intro

Shifted unitary form of $N = \rho I + \gamma_5 \cdot \mathrm{sign}(Q)$

**Method:** 'SUMR' = GMRES for shifted unitary matrices (Reichel and Jagels, 1995)

- isometric Arnoldi
- minimal residual property
- short (coupled) recurrence



## relaxation 1

Each iterative step needs an evaluation of sign$(Q)x$.

Relaxation: Relax accuracy condition for sign$(Q)x$ as iteration proceeds.

◀◀
▶▶
◀
▶
Back
Close

## relaxation 1

Each iterative step needs an evaluation of sign$(Q)x$.

Relaxation: Relax accuracy condition for sign$(Q)x$ as iteration proceeds.

**Theory** [Simoncini & Szyld, v. d. Eshof & Sleijpen, 2003]:

System $Ax = b$.

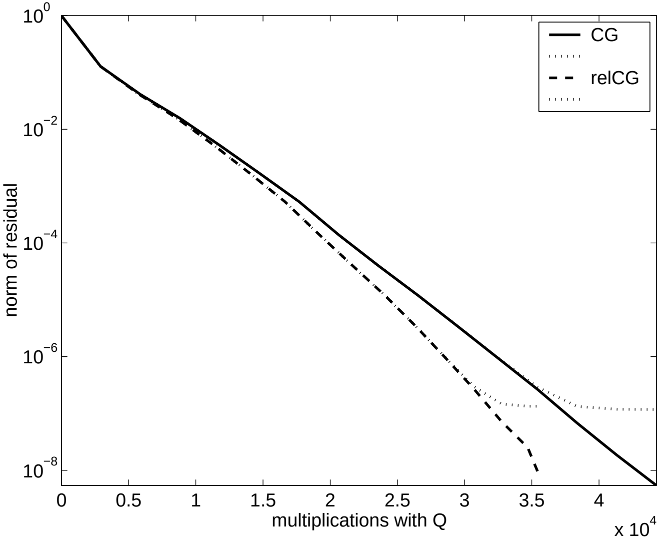
Investigate

$$\| \underbrace{b - Ax^k}_{\text{true residual}} \| \le \| \underbrace{r^k - (b - Ax^k)}_{\text{residual gap}} \| + \| \underbrace{r^k}_{\text{computed residual}} \|.$$

Develop strategy to bound residual gap below required accuracy $\epsilon$.

◀◀
▶▶
◀
▶
Back
Close

| matrix properties | method | rel. tolerance $\eta_j$ |
|---|---|---|
| herm. pos. def. $(N^H N$ | CG | $\eta_j = \epsilon\sqrt{\sum_{i=0}^{j} \|r^i\|^{-2}}$ |
| herm. indefinite $\gamma_5 N$ | MINRES | $\eta_j = \epsilon/\|r^j\|$ |
| shifted unitary $(N)$ | SUMR | $\eta_j = \epsilon/\|r^j\|$ |



Back
Close

# SUMR: recursive preconditioning

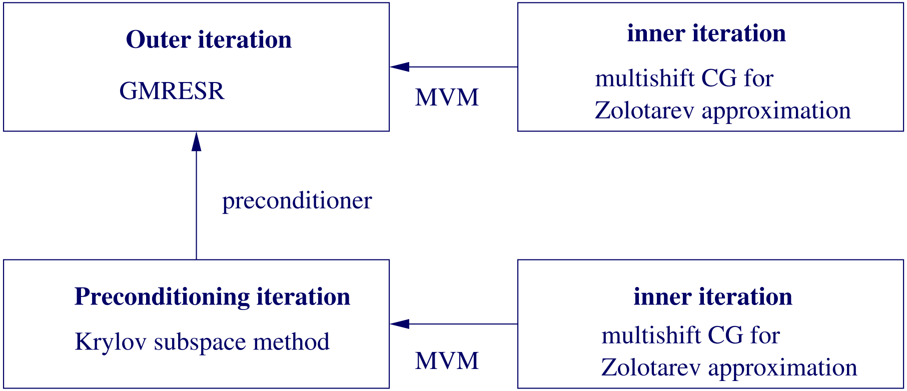
**Idea:** Relaxation pays more if convergence is fast.

- Use low accuracy SUMR as preconditioner
- outer: take adequate iterative method like GMRESR

# SUMR: recursive preconditioning

**Idea:** Relaxation pays more if convergence is fast.

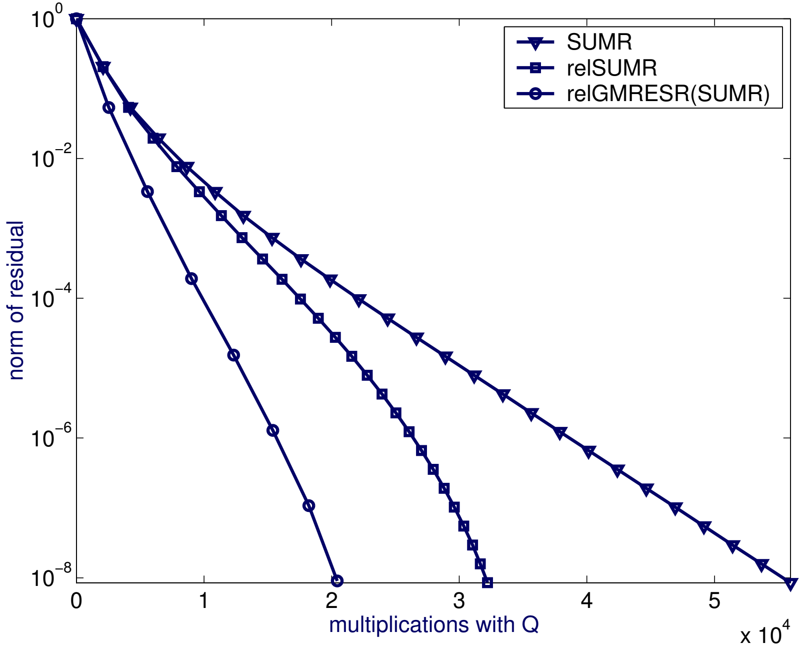
- Use low accuracy SUMR as preconditioner
- outer: take adequate iterative method like GMRESR



```
relGMRESR($A, b, \epsilon$)
    {computes $x$ with $\|Ax-b\| \leq \epsilon\cdot\|b\|$ via relaxed GMRESR}
    $x = 0,\ r = b$            {initial values}
    $C = [],\ U = []$;            {empty matrix}
    while $\|r\| > \epsilon \cdot \|b\|$ do
        solve $Au = r$ to relative accuracy $\xi$   {preconditioner}
                (for example $u = \mathrm{relSUMR}(A, r, \xi)$)
        compute $c$ with $\|Au - c\| \leq \epsilon \cdot \|b\| \cdot \|u\|/\|r\|$
        for i=1:size(C,2) do
            $\beta = C[:, i]^H \cdot c$
            $c = c - \beta \cdot C[:, i]$
            $u = u - \beta \cdot U[:, i]$
        end for
        $c = c/\|c\|,\ u = u/\|c\|$
        $C = [C, c],\ U = [U, u]$
        $\alpha = c^H \cdot r$
        $x = x + \alpha \cdot u$
        $r = r - \alpha \cdot c$
    end while
```

◀◀
▶▶
◀
▶
Back
Close

# SUMR: numerical results



$8^4$ lattice, $\rho = (1+\mu)/(1-\mu)$, $\mu = 0.1$
(precond.: accuracy $10^{-1}$)

Back
Close

| Method | $\mu = 0.03$ | $\mu = 0.1$ | $\mu = 0.3$ |
|---|---|---|---|
| SUMR | 31550 | 8312 | 3200 |
| relSUMR | 18840(1.87) | 6038(1.38) | 2656(1.20) |
| relGMRESR(SUMR) | 5974(5.82) | 2252(3.69) | 1382(2.32) |

Times (in seconds) on (quenched) $16^4$ configuration, run on 16 processors of ALiCE.

# Deflation: intro

## Features:

- precompute some ($\approx 30$) smallest eigenvalues and -vectors of $Q^2$
- 'project those out' (effect on sign function is known)
  $b = b^+ + b^- + b^\perp \Rightarrow \mathrm{sign}(Q)b = b^+ - b^- + \mathrm{sign}(Q)b^\perp$
  $\mathrm{sign}(Q)b^\perp = \mathrm{sign}(\Pi^H Q \Pi)b^\perp$
  - improves cond. no. of $Q$
  - significant decrease in no. of poles in Zolotarev PFE (for example $28$ for $10^{-10}$)
  - decreases no. of iterations in multishift CG
- relaxed GMRES(SUMR)

◀◀
▶▶
◀
▶
Back
Close

# Deflation: results

| $n_p$ | Inversion | Calls to Wilson op. | Eigenval. calc. | Total time |
|---:|---:|---:|---:|---:|
| 1 | 9144 | 1032172 | 0 | 9144 |
| 10 | 1269 | 189514 | 111 | 1380 |
| 20 | 796 | 112862 | 118 | 914 |
| 30 | 568 | 78548 | 172 | 740 |
| 40 | 459 | 63566 | 274 | 733 |
| 50 | 387 | 52758 | 361 | 748 |
| 60 | 340 | 45732 | 410 | 750 |

total time for one relGMRESR(CG) + projection of $n_p$ eigenmodes,
$8^4$ lattice, $\mu = 0.1$

◀◀
▶▶
◀
▶
Back
Close

# 3. Error Estimates and Bounds

(Aggressive) relaxation requires good estimates or upper bounds for approximation error

$$\|\mathsf{sign}(Q)b - p_m(Q)b\|$$

## Lanczos

Lanczos for $Q^2$:

$$\|\mathsf{sign}(Q)b - p_m(Q)b\| \leq \rho_m,$$

where $\rho_m$ is norm of $m$-th CG residual for $Q^2x = b$ (initial guess $0$)
[van den Eshof et al 2002]

◀◀
▶▶
◀
▶
Back
Close

## Zolotarev I: basics

**Notation:** Zolotarev $= t \cdot g(t^2)$ with

$$g(t) = \sum_{i=1}^{s} \omega_i \frac{1}{t - \sigma_i}.$$

**Remember:** For all poles $\sigma_i$, the $m$-th CG residuals for $(Q^2 - \sigma_i I)x = b$ are collinear to the Lanczos vector $v_m$,

$$r_i^m = b - (Q^2 - \sigma_i I)x_i^m = \rho_i^m v_m.$$

**Approximation** and **error**:

$$x^m = \beta_0 V_m g(T_m) e_1 = \sum_{i=1}^{s} \omega_i x_i^m, \qquad e^m = x^m - g(A)b.$$

**Classical estimate**: If convergence is monotone or even superlinear

$$\|e^m\| \approx \|x^m - x^{m+d}\|, \quad d \geq 1 \text{ moderately large}$$

◀◀
▶▶
◀
▶
Back
Close

# Zolotarev: Gaussian quadrature

Expand error in terms of residuals:

$$e^m = \sum_{i=1}^{s} \rho_i^m \omega_i (Q^2 - \sigma_i I)^{-1} v^m, \quad \|e^m\|^2 = (v^m)^H h(Q^2) v^m,$$

where

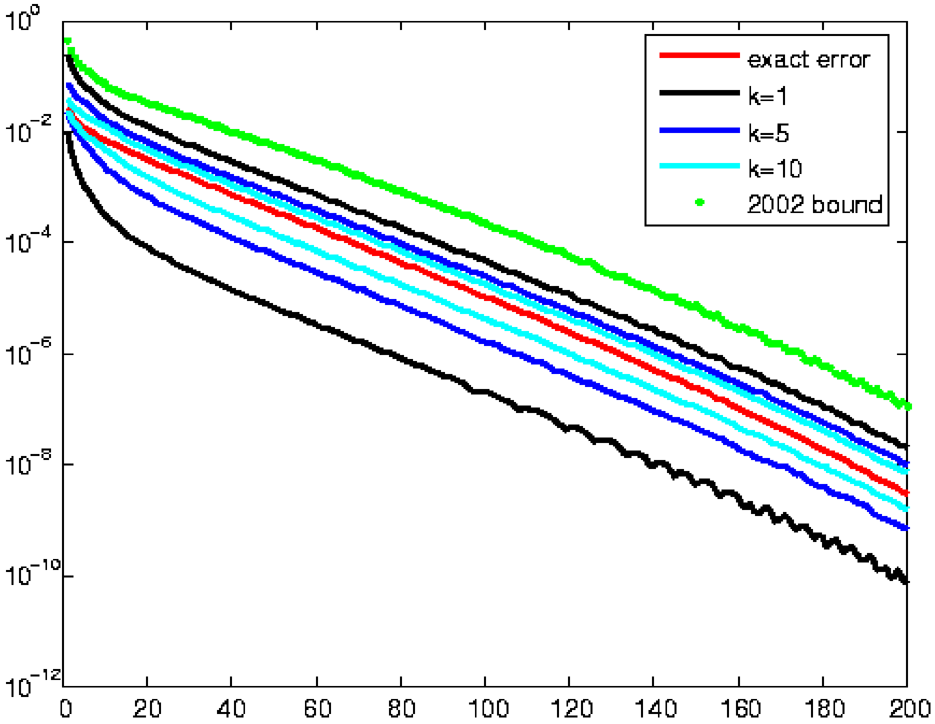
$$h(Q^2) = \sum_{i,j=1}^{s} \rho_i^m \rho_j^m \omega_i \omega_j (Q^2 - \sigma_i I)^{-1} (Q^2 - \sigma_j I)^{-1}$$

Golub/Meurant (1994, 1997): Use Gaussian quadrature w.r.t. discrete measure to get upper and lower bounds for the moment $(v^m)^H h(Q^2) v^m$.

- Elegant theory, lower and upper bounds
- One more node in quadrature rule amounts to one further step of Lanczos for $Q^2$ and $v^m$
- MVMs cannot be recycled to improve the solutions to the systems

# Zolotarev: example

Zolotarev for $(Q^2)^{-1/2}Qb$ with $\mathrm{spec}(Q) \subset [-32,-1] \cup [1,32]$.



## New estimates based on CG coefficients

Recall CG algorithm:

```
Choose $x^0 = 0$, set $r^{(0)} = b$, $p^0 = r^0$
for $k = 1, 2, \ldots$ do
    $\gamma^{k-1} = \langle r^{k-1}, r^{k-1} \rangle / \langle p^{k-1}, Ap^{k-1} \rangle$
    $x^k = x^{k-1} + \gamma^{k-1} p^{k-1}$
    $r^k = r^{k-1} - \gamma^{k-1} Ap^{k-1}$
    $\delta^k = \langle r^k, r^k \rangle / \langle r^{k-1}, r^{k-1} \rangle$
    $p^k = r^k + \delta^k p^{k-1}$
end for
```

◀◀
▶▶
◀
▶
Back
Close

# New estimates based on CG coefficients

Recall CG algorithm:

```
Choose $x^0 = 0$, set $r^{(0)} = b$, $p^0 = r^0$
for $k = 1, 2, \ldots$ do
    $\gamma^{k-1} = \langle r^{k-1}, r^{k-1} \rangle / \langle p^{k-1}, Ap^{k-1} \rangle$
    $x^k = x^{k-1} + \gamma^{k-1} p^{k-1}$
    $r^k = r^{k-1} - \gamma^{k-1} Ap^{k-1}$
    $\delta^k = \langle r^k, r^k \rangle / \langle r^{k-1}, r^{k-1} \rangle$
    $p^k = r^k + \delta^k p^{k-1}$
end for
```

For $d \in \mathbb{N}$, denote

$$\eta^{k,d} := \sum_{i=0}^{d-1} \gamma^{k+i} \langle r^{k+i}, r^{k+i} \rangle$$

$$\varphi^{k,d} := \sum_{i=0}^{d} \frac{\langle p^{k+i}, p^{k+i} \rangle}{\langle p^{k+i}, Ap^{k+i} \rangle} \cdot \left( \langle r^{k+i}, e^{k+i} \rangle + \langle r^{k+i+1}, e^{k+i+1} \rangle \right).$$

◀◀
▶▶
◀
▶
Back
Close

**Lemma**:

$$\begin{aligned}\langle r^k, e^k\rangle &= \langle r^{k+d}, e^{k+d}\rangle + \eta^{k,d} \geq \eta^{k,d},\\ \langle e^k, e^k\rangle &= \langle e^{k+d}, e^{k+d}\rangle + \varphi^{k,d} \geq \varphi^{k,d}.\end{aligned}$$

**Note:** $\langle r^{k+i}, e^{k+i}\rangle$ in $\varphi^{k,d}$ is not available.
Replacing by $\eta^{k+i,d}$ gives the estimate

$$\begin{aligned}\tau^{k,d} &= \sum_{i=0}^{d} \frac{\langle p^{k+i}, p^{k+i}\rangle}{\langle p^{k+i}, Ap^{k+i}\rangle}(\eta^{k+i,d} + \eta^{k+i+1,d})\\ &\leq \langle e^k, e^k\rangle.\end{aligned}$$

[Hestenes-Stiefel 1952, Strakos-Tichy 2002, Meurant 2005]

◀◀
▶▶
◀
▶
Back
Close

For Galerkin approximation to $g(A)b$ we have

$$\|e^k\|^2 = \sum_{i,j=1}^{s} \omega_i \omega_j \langle e_i^k, e_j^k \rangle.$$

For $\sigma_i \neq \sigma_j$ one has

$$\frac{1}{(t-\sigma_i)(t-\sigma_j)} = \frac{1}{\sigma_i - \sigma_j} \cdot \left( \frac{1}{t-\sigma_i} - \frac{1}{t-\sigma_j} \right),$$

thus

$$\langle e_i^k, e_j^k \rangle = \frac{1}{\sigma_i - \sigma_j} \cdot \left( \langle r_i^k, e_j^k \rangle - \langle r_j^k, e_i^k \rangle \right)$$

**Theorem:** We have

$$\|g(A)b - \sum_{i=1}^{s} \omega_i x_i^k\|_2^2 \geq \boldsymbol{\eta}^{k,d} + \boldsymbol{\tau}^{k,d},$$

where

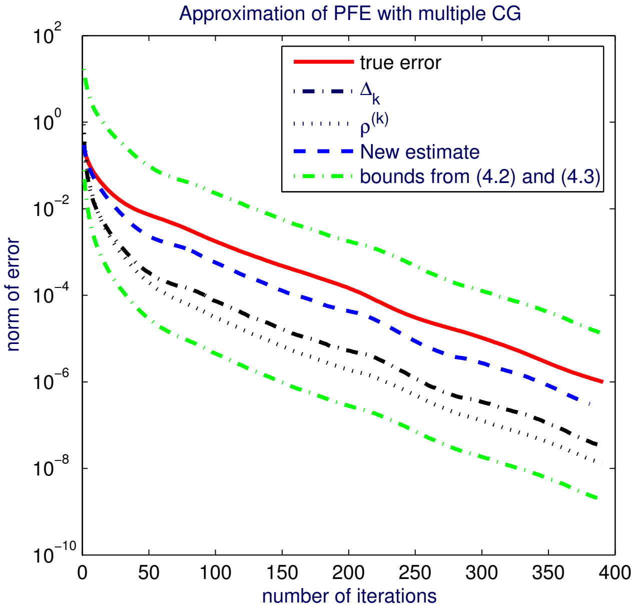
$$\begin{aligned} \boldsymbol{\eta}^{k,d} &= \sum_{i,j=1,\sigma_i\neq\sigma_j}^{s} \frac{\omega_i\omega_j}{\sigma_i-\sigma_j}\left(\frac{\rho_j^k}{\rho_i^k}\eta_i^{(k,d)} - \frac{\rho_i^k}{\rho_j^k}\eta_j^{k,d}\right), \\ \boldsymbol{\tau}^{(k,d)} &= \sum_{i,j=1,\sigma_i=\sigma_j}^{s} \omega_i\omega_j\tau_j^{k,d}. \end{aligned}$$

**Corollary:** If estimates are positive, error norm decreases.

◀◀
▶▶
◀
▶
Back
Close

# Error estimates: numerical results

**Example 1:** sign$(A)b$, including deflation of small eigenvalues. Configuration `conf5.4-00l8x8-2000.mtx` from MatrixMarket, $d = 5$



# 4. Nonzero chemical potential

Wilson matrix is modified:

$$\begin{aligned}(M(\mu)\psi)_x = \psi_x - \kappa \Bigg( & \sum_{\nu=1}^{3} \left((I-\gamma_\nu)\otimes U_\nu(x)\right)\psi_{x+e_\nu} \\ & + \sum_{\nu=1}^{3} \left((I+\gamma_\nu)\otimes U_\nu^H(x-e_\nu)\right)\psi_{x-e_\nu} \Bigg) \\ & -\kappa\left(e^{-\mu}\,(I-\gamma_4)\otimes U_4(x)\right)\psi_{x+e_4} \\ & -\kappa\left(e^{\mu}\,(I+\gamma_4)\otimes U_4^H(x-e_4)\right)\psi_{x-e_4})\end{aligned}$$

**Consequence:** $Q=\gamma_5 M$ is not hermitian any more.

◀◀
▶▶
◀
▶
Back
Close

# sign function revisited

We need **alternatives** to the spectral definition.

**Function theory**: $f$ analytic in neighborhood of spec$(A)$, Γ contour:

$$f(A) = \frac{1}{2\pi i} \oint_\Gamma f(z)(zI - A)^{-1} dz.$$

$A$ is **diagonalizable**, $A = U \Lambda U^{-1}$, then

$$f(A) = U f(\Lambda) U^{-1} \text{ with } f(\Lambda) = \text{diag}(f(\lambda_i)).$$

◀◀
▶▶
◀
▶
Back
Close

# sign function revisited II

$A$ **not diagonalizable**, Jordan decomposition

$$A = U\left(\bigoplus_i J_i\right)U^{-1}, \qquad J_i = \begin{pmatrix} \lambda_i & 1 & \cdots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i \end{pmatrix}.$$

Then

$$f(A) = U\left(\bigoplus_i f(J_i)\right)U^{-1},$$

where

$$f(J_i) = \begin{pmatrix} f(\lambda_i) & f^{(1)}(\lambda_i) & \cdots & \frac{f^{(m_i-1)}(\lambda_i)}{(m_i-1)!} \\ 0 & f(\lambda_i) & \ddots & \vdots \\ \vdots & \ddots & \ddots & f^{(1)}(\lambda_i) \\ 0 & \cdots & 0 & f(\lambda_i) \end{pmatrix}.$$

◀◀
▶▶
◀
▶
Back
Close

# sign function revisited: Galerkin

**Consequence:**

$$\mathsf{sign}(Q) = p(Q),$$

$p$ Hermite interpolating polynomial on $\mathsf{spec}(Q)$.

**Problem:** Fix $m \in \mathbb{N}$. Find "best" approximating polynomial $p_{m-1}$ s.t.

$$p_{m-1}(Q)b - \mathsf{sign}(Q)b \;\rightarrow\; \mathsf{min!} \quad \text{for all } p_{m-1} \in \mathcal{P}_{m-1}.$$

**Solution:** Use Galerkin condition

$$p_{m-1}(Q)b - \mathsf{sign}(Q)b \perp K_m(Q, b)$$

Back
Close

# sign function revisited: computation

Use Arnoldi process to construct orthogonal basis $v_1, \ldots, v_m$ of $K_m(Q, b)$:

$$QV_m = V_m H_m + \beta_m v_{m+1} e_1^T, \quad H_m \text{ upper Hessenberg }.$$

Results in long recurrence!.

The Galerkin approximation can be computed as

$$p_{m-1}(Q)b = V_m \text{sign}(H_m) e_1 \cdot \|b\| = V_m \text{sign}(V_m^H Q V_m) V_m^H b.$$

**Note:** $H_m$ is "small", $\text{sign}(H_m)$ can be computed using Roberts' iteration

$$S_{k+1} = S_k + S_k^{-1}, \quad S_0 = H_m.$$

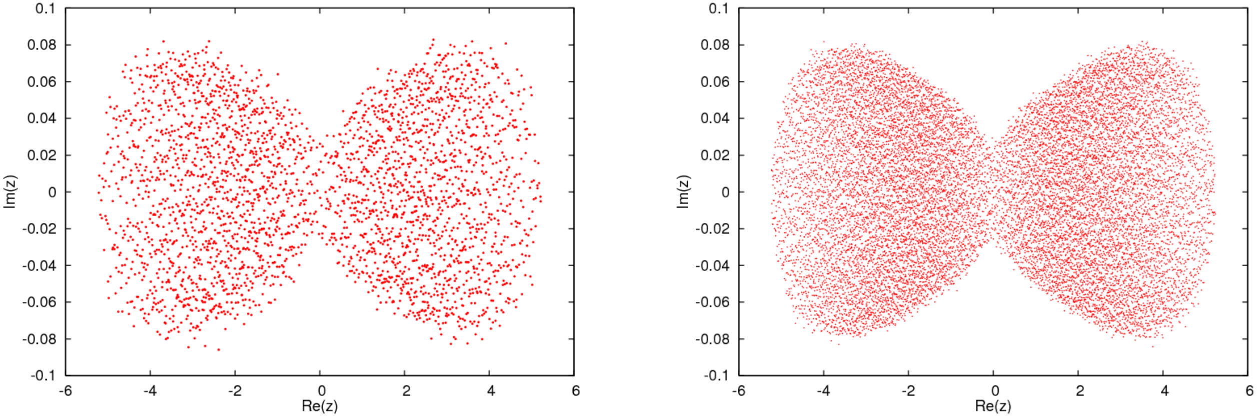
**Option**: Compute $QR$-factorization first.

◀◀
▶▶
◀
▶
Back
Close

# Deflation

**Experience:** Deflation of small eigenvalues is mandatory.

Our test cases: $\mu = 0.3$, lattice $4^4$ (left) and $6^4$ (right)



spectrum of $Q = \gamma_5 M$

Back
Close

# Deflation: augmented subspaces

**Problem:** $w \perp S$ and $QS = S \not\Rightarrow Qw \perp S$!!

**Consequence**: Let $S$ be spanned by "small" eigenvectors. Decompose

$$b = b_{\|} + b_{\perp}.$$

Then $K_m(Q, b_{\perp}) \cap S \neq \{0\}$.

**Solution**: Augmented Krylov subspace approach.

Let $AS = ST$, $T \in \mathbb{R}^{k \times k}$ ($T$ is usually triangular).

Compute orthogonal basis for $K_m(Q, b_{\perp}) + S$ similarly to Arnoldi:

$$Q \begin{pmatrix} S & V_m \end{pmatrix} = \begin{pmatrix} S & V_m \end{pmatrix} \begin{pmatrix} T & S^H A V_m \\ 0 & H_m \end{pmatrix} + \beta_m v_{m+1} e_{k+m+1}^T \,.$$

# Deflation: Galerkin

Define

$$W_m = \left( S \;\; V_m \right), \quad G_m = \begin{pmatrix} T & S^H A V_m \\ 0 & H_m \end{pmatrix}.$$

Imposing the Galerkin condition gives

$$\mathsf{sign}(Q) b_\perp \approx W_m \mathsf{sign}(G_m) e_{m+1} \cdot \|b_\perp\|.$$

**Note**:

$$\mathsf{sign}(G_m) = \begin{pmatrix} \mathsf{sign}(T) & Y \\ 0 & \mathsf{sign}(H_m) \end{pmatrix}$$

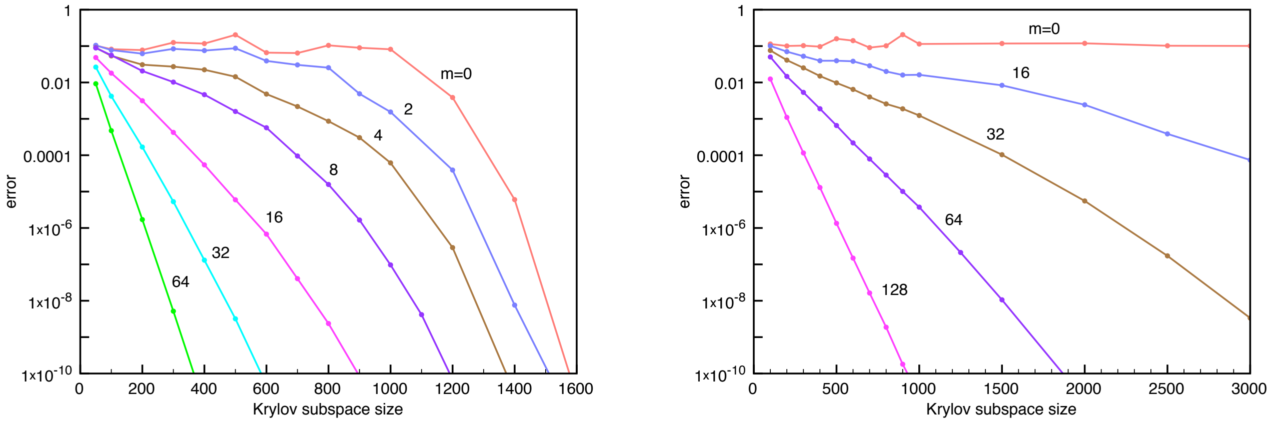
where $Y$ solves the Sylvester equation

$$TY - YH_m = \mathsf{sign}(T)X - X\mathsf{sign}(H_m), \qquad X = S^H Q V_m.$$

◀◀
▶▶
◀
▶
Back
Close

# Numerical results

Convergence histories (left: $4^4$ lattice, right: $6^4$ lattice):



Back
Close

# Alternatives

**Simplify your life:** Two-sided deflation!

**Specifically**: colspan$(V)$ right evs, colspan$(W)$ left evs

$$Qv_i = \lambda_i v_i, \quad w_i^H Q = \lambda_i w_i^H, \quad w_i^H v_i = \delta_{ij}, i = 1, \ldots, m.$$

$VW^H$ projects on colspan$(V)$ along colspan$(W)$

**Decompose:** $b = VW^H b + (I - VW^H)b$.

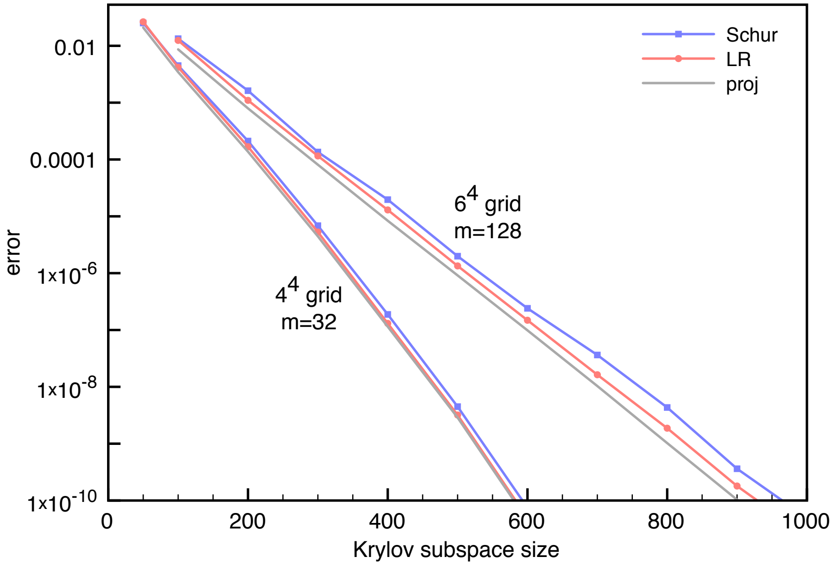
**Then:** $\text{sign}(Q) \cdot VW^H b = \text{diag}(\text{sign}(\lambda_i)) \cdot V \cdot (W^H b)$,
$VW^H \cdot Q \cdot (I - VW^H)b = 0$.

**Consequence:**

$$K_k(Q, (I - VW^H)b) = K_k((I - VW^H)Q(I - VW^H), b).$$

◀◀
▶▶
◀
▶
Back
Close

# Performance one-sided vs two-sided deflation



# Computational cost

$4^4$ lattice, $m = 32$
Schur deflation

initialization time: 14.1 s

| $k$ | Arnoldi | sign($H_k$) | total |
|---|---|---|---|
| 100 | 0.18 | 0.03 | 0.23 |
| 200 | 0.59 | 0.21 | 0.81 |
| 300 | 1.22 | 0.52 | 1.75 |
| 400 | 2.05 | 1.08 | 3.16 |
| 500 | 3.12 | 1.79 | 4.93 |
| 600 | 4.37 | 2.90 | 7.31 |
| 700 | 5.88 | 4.57 | 10.49 |
| 800 | 7.56 | 6.69 | 14.28 |
| 900 | 9.50 | 9.38 | 18.92 |
| 1000 | 11.63 | 12.68 | 24.36 |

◀◀
▶▶
◀
▶
Back
Close

$4^4$ lattice, $m = 32$
LR-deflation

initialization time: 27.5 s

| $k$ | Arnoldi | sign$(H)$ | total |
|---|---|---|---|
| 100 | 0.12 | 0.03 | 0.15 |
| 200 | 0.45 | 0.20 | 0.66 |
| 300 | 1.01 | 0.49 | 1.51 |
| 400 | 1.77 | 1.02 | 2.82 |
| 500 | 2.76 | 1.69 | 4.47 |
| 600 | 3.94 | 2.77 | 6.74 |
| 700 | 5.36 | 4.40 | 9.79 |
| 800 | 6.96 | 6.44 | 13.44 |
| 900 | 8.84 | 9.10 | 17.98 |
| 1000 | 10.84 | 12.33 | 23.21 |

Back
Close

$6^4$ lattice, $m = 128$
Schur deflation

initialization time: 884 s

| $k$ | Arnoldi | sign($H_k$) | total |
|---|---|---|---|
| 100 | 2.03 | 0.05 | 2.13 |
| 200 | 5.16 | 0.22 | 5.45 |
| 300 | 9.27 | 0.56 | 9.91 |
| 400 | 14.59 | 1.15 | 15.85 |
| 500 | 20.95 | 2.09 | 23.17 |
| 600 | 28.12 | 3.35 | 31.61 |
| 700 | 36.81 | 5.17 | 42.15 |
| 800 | 46.32 | 7.39 | 53.88 |
| 900 | 56.83 | 10.37 | 67.39 |
| 1000 | 68.29 | 13.88 | 82.39 |

◀◀
▶▶
◀
▶
Back
Close

$6^4$ lattice, $m = 128$
LR-deflation

initialization time: 1713 s

| $k$ | Arnoldi | sign$(H)$ | total |
|---|---|---|---|
| 100 | 0.66 | 0.03 | 0.75 |
| 200 | 2.39 | 0.15 | 2.62 |
| 300 | 5.16 | 0.42 | 5.69 |
| 400 | 9.01 | 0.94 | 10.06 |
| 500 | 13.96 | 1.73 | 15.84 |
| 600 | 20.03 | 2.80 | 22.98 |
| 700 | 27.09 | 4.44 | 31.70 |
| 800 | 35.09 | 6.49 | 41.78 |
| 900 | 44.38 | 9.10 | 53.70 |
| 1000 | 54.74 | 12.36 | 67.34 |

Back
Close