# On classical and modern approximations for neutron transport in a unified framework

**Matthias Schlottbom**

October 26th, 2023

Mathematics for Nuclear Applications Seminar
Port-au-rocs, le Croisic October 23–27, 2023

UNIVERSITY OF TWENTE.

# Outline

## Iterative solution for dG discretization

- Source iteration for NTE in slab geometry
- Accelerating the source iteration
- Accelerated scheme in a variational context

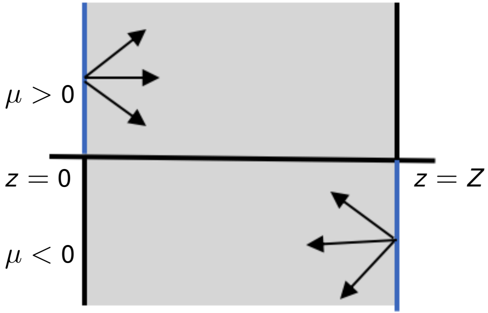
## Low-rank approximations

- Overview of different approaches
- Rank control
- Preconditioning

# Recall: NTE in slab geometry

$$\mu \partial_z \phi + \sigma \phi = \sigma_s \bar{\phi} + q \quad \text{in } (0, Z) \times (-1, 1)$$
$$\phi(0, \mu) = g_0(\mu) \qquad \mu > 0$$
$$\phi(Z, \mu) = g_Z(\mu) \qquad \mu < 0$$

with $\bar{\phi}(z, \mu) = \frac{1}{2} \int_{-1}^{1} \phi(z, \mu')\, \mathrm{d}\mu'$.



# Recall: NTE in slab geometry

$$\mu\partial_z\phi + \sigma\phi = \sigma_s\bar{\phi} + q \quad \text{in } (0,Z)\times(-1,1)$$
$$\phi(0,\mu) = g_0(\mu) \qquad \mu > 0$$
$$\phi(Z,\mu) = g_Z(\mu) \qquad \mu < 0$$

with $\bar{\phi}(z,\mu) = \frac{1}{2}\int_{-1}^{1}\phi(z,\mu')\,\mathrm{d}\mu'$.

**Existence theory:** Fixed-point iteration $T : L^2 \to L^2, \phi^n \mapsto \phi^{n+1}$ with

$$\mu\partial_z\phi^{n+1} + \sigma\phi^{n+1} = \sigma_s\bar{\phi}^n + q \quad \text{in } (0,Z)\times(-1,1)$$
$$\phi^{n+1} = g \qquad \text{on } \Gamma_-$$

# Recall: NTE in slab geometry

$$
\begin{aligned}
\mu\partial_z\phi + \sigma\phi &= \sigma_s\bar{\phi} + q && \text{in } (0,Z)\times(-1,1)\\
\phi(0,\mu) &= g_0(\mu) && \mu > 0\\
\phi(Z,\mu) &= g_Z(\mu) && \mu < 0
\end{aligned}
$$

$$
\text{with } \bar{\phi}(z,\mu) = \tfrac{1}{2}\int_{-1}^{1}\phi(z,\mu')\,\mathrm{d}\mu'.
$$

**Existence theory:** Fixed-point iteration $T : L^2 \to L^2, \phi^n \mapsto \phi^{n+1}$ with

$$
\begin{aligned}
\mu\partial_z\phi^{n+1} + \sigma\phi^{n+1} &= \sigma_s\bar{\phi}^n + q && \text{in } (0,Z)\times(-1,1)\\
\phi^{n+1} &= g && \text{on } \Gamma_-
\end{aligned}
$$

For each $\mu$: family of decoupled advection equations for $\phi^{n+1}$.

# Proof of convergence in $L^2$

Let $\phi, \psi \in L^2$. Then $w = T\phi - T\psi = T(\phi - \psi)$ satisfies

$$\begin{aligned} \mu\partial_z w + \sigma w &= \sigma_s \overline{(\phi - \psi)} && \text{in } (0, Z) \times (-1, 1) \\ w &= 0 && \text{on } \Gamma_- \end{aligned}$$

# Proof of convergence in $L^2$

Let $\phi, \psi \in L^2$. Then $w = T\phi - T\psi = T(\phi - \psi)$ satisfies

$$\begin{aligned} \mu\partial_z w + \sigma w &= \sigma_s(\overline{\phi - \psi}) && \text{in } (0, Z) \times (-1, 1) \\ w &= 0 && \text{on } \Gamma_- \end{aligned}$$

Multiply by $w$ and integrate over $(0, Z) \times (-1, 1)$:

$$(\mu\partial_z w, w) + \|\sqrt{\sigma} w\|_{L^2}^2 = (\sigma_s(\overline{\phi - \psi}), w).$$

# Proof of convergence in $L^2$

Let $\phi, \psi \in L^2$. Then $w = T\phi - T\psi = T(\phi - \psi)$ satisfies

$$\mu\partial_z w + \sigma w = \sigma_s(\overline{\phi - \psi}) \quad \text{in } (0, Z) \times (-1, 1)$$
$$w = 0 \quad \text{on } \Gamma_-$$

Multiply by $w$ and integrate over $(0, Z) \times (-1, 1)$:

$$(\mu\partial_z w, w) + \|\sqrt{\sigma} w\|_{L^2}^2 = (\sigma_s(\overline{\phi - \psi}), w).$$

**Observations.** Integration-by-parts:

$$(\mu\partial_z w, w) = -(w, \mu\partial_z w) + (w, w\mu)_\Gamma = -(w, \mu\partial_z w) + \langle w, w|\mu|\rangle_{\Gamma_+}.$$

# Proof of convergence in $L^2$

Let $\phi, \psi \in L^2$. Then $w = T\phi - T\psi = T(\phi - \psi)$ satisfies

$$\begin{aligned} \mu\partial_z w + \sigma w &= \sigma_s \overline{(\phi - \psi)} && \text{in } (0, Z) \times (-1, 1) \\ w &= 0 && \text{on } \Gamma_- \end{aligned}$$

Multiply by $w$ and integrate over $(0, Z) \times (-1, 1)$:

$$(\mu\partial_z w, w) + \|\sqrt{\sigma} w\|_{L^2}^2 = (\sigma_s \overline{(\phi - \psi)}, w).$$

**Observations.** Integration-by-parts:

$$(\mu\partial_z w, w) = -(w, \mu\partial_z w) + (w, w\mu)_\Gamma = -(w, \mu\partial_z w) + \langle w, w|\mu|\rangle_{\Gamma_+}.$$

**Cauchy-Schwarz** and $\sigma_s \leq \sigma$ and $\sigma > 0$:

$$\begin{aligned} (\sigma_s \overline{(\phi - \psi)}, w) &\leq \|\sqrt{\sigma_s}\overline{(\phi - \psi)}\|_{L^2} \|\sqrt{\sigma_s} w\|_{L^2} \\ &\leq \|\frac{\sigma_s}{\sigma}\|_\infty \|\sqrt{\sigma}\overline{(\phi - \psi)}\|_{L^2} \|\sqrt{\sigma} w\|_{L^2}. \end{aligned}$$

# Proof of convergence in $L^2$

Let $\phi, \psi \in L^2$. Then $w = T\phi - T\psi = T(\phi - \psi)$ satisfies

$$\mu \partial_z w + \sigma w = \sigma_s \overline{(\phi - \psi)} \quad \text{in } (0, Z) \times (-1, 1)$$
$$w = 0 \quad \text{on } \Gamma_-$$

**Conclusion:**

$$\|\sqrt{\sigma} w\|_{L^2} \leq \|\frac{\sigma_s}{\sigma}\|_\infty \|\sqrt{\sigma}(\phi - \psi)\|_{L^2},$$

i.e., $T : L^2 \to L^2$ is a contraction if $\|\frac{\sigma_s}{\sigma}\|_\infty < 1$.

**Remarks:** The iteration $\phi^n \mapsto \phi^{n+1}$

- converges slowly if $\sigma_a \ll \sigma_s$, i.e., $\sigma_s/\sigma \approx 1$.
- is also called source iteration.

# Towards accelerating the iteration: error equation

**Update equation.**

$$\begin{aligned} \mu\partial_z\phi^{n+1} + \sigma\phi^{n+1} &= \sigma_s\bar{\phi}^n + q && \text{in } (0,Z)\times(-1,1) \\ \phi^{n+1} &= g && \text{on } \Gamma_- \end{aligned}$$

The **error** $e^n = \phi - \phi^n$ satisfies

$$\begin{aligned} \mu\partial_z e^{n+1} + \sigma e^{n+1} &= \sigma_s\overline{e^n} && \text{in } (0,Z)\times(-1,1) \\ e^{n+1} &= 0 && \text{on } \Gamma_- \end{aligned}$$

Equivalently (using that $e^n - e^{n+1} = \phi^{n+1} - \phi^n$)

$$\begin{aligned} \mu\partial_z e^{n+1} + \sigma e^{n+1} &= \sigma_s\bar{e}^{n+1} + \sigma_s(\overline{\phi^{n+1} - \phi^n}) && \text{in } (0,Z)\times(-1,1) \\ e^{n+1} &= 0 && \text{on } \Gamma_- \end{aligned}$$

# Towards accelerating the iteration: error equation

Equivalently (using that $e^n - e^{n+1} = \phi^{n+1} - \phi^n$)

$$\begin{aligned} \mu \partial_z e^{n+1} + \sigma e^{n+1} &= \sigma_s \bar{e}^{n+1} + \sigma_s \left(\overline{\phi^{n+1} - \phi^n}\right) \quad &\text{in } (0, Z) \times (-1, 1) \\ e^{n+1} &= 0 \quad &\text{on } \Gamma_- \end{aligned}$$

**Observations:**

- The error satisfies the NTE with source term $\sigma_s \left(\overline{\phi^{n+1} - \phi^n}\right)$.
- Solving the error equation is as difficult as solving the NTE.

**Idea:**

- Approximate the error by $\phi_e^{n+1} \approx e^{n+1}$.
- New iterate $\phi^{n+1} + \phi_e^{n+1}$.

# How to obtain a good and computable correction $\phi_e^{n+1}$?

## Diffusion limit

**Recall:** Convergence is slow if scattering dominates absorption $\sigma_s \gg \sigma_a$.

**Consider:** $\sigma_s = \frac{\bar{\sigma}_s}{\varepsilon}$, $\sigma_a = \varepsilon\bar{\sigma}_a$ with $\bar{\sigma}_s, \bar{\sigma}_a > 0$.

Denote $\phi^\varepsilon$ solution to scaled equations

$$\begin{aligned} \mu\partial_z\phi^\varepsilon + \frac{1}{\varepsilon}\left(\bar{\sigma}_s + \varepsilon^2\bar{\sigma}_a\right)\bar{\sigma}\phi^\varepsilon &= \frac{\sigma_s}{\varepsilon}\bar{\phi}^\varepsilon + \varepsilon\bar{q} \quad \text{in } (0, Z) \times (-1, 1) \\ \phi^\varepsilon &= 0 \quad \text{on } \Gamma_- \end{aligned}$$

[Habetler & Matkowsky '75] [Larsen & Keller '74] [Bardos et al '84] [Blankenship & Papanicolaou '78] [Egger & S 2014].

# How to obtain a good and computable correction $\phi_e^{n+1}$?

## Diffusion limit

**Recall:** Convergence is slow if scattering dominates absorption $\sigma_s \gg \sigma_a$.

**Consider:** $\sigma_s = \frac{\bar\sigma_s}{\varepsilon}$, $\sigma_a = \varepsilon\bar\sigma_a$ with $\bar\sigma_s, \bar\sigma_a > 0$.

Denote $\phi^\varepsilon$ solution to scaled equations

$$\mu\partial_z\phi^\varepsilon + \frac{1}{\varepsilon}\left(\bar\sigma_s + \varepsilon^2\bar\sigma_a\right)\bar\sigma\phi^\varepsilon = \frac{\sigma_s}{\varepsilon}\bar\phi^\varepsilon + \varepsilon\bar q \quad \text{in } (0,Z)\times(-1,1)$$
$$\phi^\varepsilon = 0 \quad \text{on } \Gamma_-$$

**Limit:** $\phi^\varepsilon \to \bar\phi^0$ in $L^2$ as $\varepsilon \to 0$, with $\bar\phi^0 \in H_0^1(0,Z)$ solution to

$$-\mathrm{div}\left(\frac{1}{3\bar\sigma_s}\nabla\bar\phi^0\right) + \bar\sigma_a\bar\phi^0 = \bar q \quad \text{in } (0,Z).$$

**Idea:** Solve the diffusion eq. with RHS $\bar\sigma_s(\bar\phi^{n+1} - \bar\phi^n)$ to obtain $\phi_e^{n+1}$.

---

[Habetler & Matkowsky '75] [Larsen & Keller '74] [Bardos et al '84] [Blankenship & Papanicolaou '78] [Egger & S 2014].

# Summary of DSA scheme

1. Given $\phi^n \in L^2$, compute $\phi^{n+1/2} \in L^2$ solution to

$$\begin{aligned} \mu\partial_z\phi^{n+1/2} + \sigma\phi^{n+1/2} &= \sigma_s\bar{\phi}^n + q \quad \text{in } (0,Z)\times(-1,1),\\ \phi^{n+1/2} &= g \qquad\qquad\quad \text{on } \Gamma_-. \end{aligned}$$

2. Compute correction $\bar{\phi}_c^{n+1/2} \in H_0^1(0,Z)$ solution to

$$-\mathrm{div}\left(\frac{1}{3\sigma}\nabla\bar{\phi}_c^{n+1/2}\right) + \sigma_a\bar{\phi}_c^{n+1/2} = \sigma_s\left(\bar{\phi}^{n+1/2} - \bar{\phi}^n\right) \quad \text{in } (0,Z).$$

3. Define new iterate $\phi^{n+1} = \phi^{n+1/2} + \bar{\phi}_c^{n+1/2}$.

[Adams & Larsen 2002]

# Summary of DSA scheme

1. Given $\phi^n \in L^2$, compute $\phi^{n+1/2} \in L^2$ solution to

$$\begin{aligned} \mu\partial_z\phi^{n+1/2} + \sigma\phi^{n+1/2} &= \sigma_s\bar{\phi}^n + q \quad \text{in } (0,Z)\times(-1,1), \\ \phi^{n+1/2} &= g \qquad\qquad \text{on } \Gamma_-. \end{aligned}$$

2. Compute correction $\bar{\phi}_c^{n+1/2} \in H_0^1(0,Z)$ solution to

$$-\mathrm{div}\left(\frac{1}{3\sigma}\nabla\bar{\phi}_c^{n+1/2}\right) + \sigma_a\bar{\phi}_c^{n+1/2} = \sigma_s(\bar{\phi}^{n+1/2} - \bar{\phi}^n) \quad \text{in } (0,Z).$$

3. Define new iterate $\phi^{n+1} = \phi^{n+1/2} + \bar{\phi}_c^{n+1/2}$.

**Remarks:**

- Step 2 is also called diffusion synthetic acceleration (DSA).
- Amplification factor of the scheme is $\approx 0.2247\|\sigma_s/\sigma\|_\infty$ for unbounded domains/periodic boundary conditions, constant coefficients.
- Incompatible numerical schemes for 1. and 2. may imply divergence.

[Adams & Larsen 2002]

## Iterative solution for dG discretization



## Low-rank approximations

# DSA scheme in a variational context

**Recall variational formulation:** Find $\phi = \phi^+ + \phi^- \in \mathbb{W}^+ \oplus \mathbb{V}^-$ such that for all $\psi = \psi^+ + \psi^- \in \mathbb{W}^+ \oplus \mathbb{V}^-$

$$\langle|\mu|\phi^+, \psi^+\rangle_\Gamma - (\phi^-, \mu\partial_z\psi^+) + (\mu\partial_z\phi^+, \psi^-) + (\sigma\phi, \psi) = (\sigma_s\bar{\phi}, \psi^+) + (q, \psi) + 2\langle|\mu|g, \psi^+\rangle_{\Gamma_-}.$$

# DSA scheme in a variational context

**Recall variational formulation:** Find $\phi = \phi^+ + \phi^- \in \mathbb{W}^+ \oplus \mathbb{V}^-$ such that for all $\psi = \psi^+ + \psi^- \in \mathbb{W}^+ \oplus \mathbb{V}^-$

$$\langle|\mu|\phi^+, \psi^+\rangle_\Gamma - (\phi^-, \mu\partial_z\psi^+) + (\mu\partial_z\phi^+, \psi^-) + (\sigma\phi, \psi) = (\sigma_s\bar{\phi}, \psi^+) \\ +(q, \psi) + 2\langle|\mu|g, \psi^+\rangle_{\Gamma_-}.$$

Testing with $\psi = \psi^-$ yields

$$(\mu\partial_z\phi^+, \psi^-) + (\sigma\phi^-, \psi^-) = +(q^-, \psi^-),$$

i.e., $\phi^- = (q^- - \mu\partial_z\phi^+)/\sigma$.

Inserting $\phi^-$ yields a new variational principle:

# DSA scheme in a variational context

Testing with $\psi = \psi^-$ yields

$$(\mu\partial_z\phi^+, \psi^-) + (\sigma\phi^-, \psi^-) = +(q^-, \psi^-),$$

i.e., $\phi^- = (q^- - \mu\partial_z\phi^+)/\sigma$.

Inserting $\phi^-$ yields a new variational principle:

**Even-parity equation**: Find $u \in \mathbb{W}^+$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathbb{W}^+,$$

where

$$\begin{aligned} a(u, v) &= b(u, v) - k(u, v), \\ b(u, v) &= \langle u, v\rangle_{\Gamma_-} + (\frac{\mu}{\sigma}\partial_z u, \mu\partial_z v) + (\sigma u, v) \\ k(u, v) &= (\sigma_s \bar{u}, v), \\ \ell(v) &= 2\langle g, v\rangle_{\Gamma_-} + (q, v + \frac{\mu}{\sigma}\partial_z v). \end{aligned}$$

# DSA scheme in a variational context

**Even-parity equation**: Find $u \in \mathbb{W}^+$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathbb{W}^+,$$

where

$$a(u, v) = b(u, v) - k(u, v),$$
$$b(u, v) = \langle u, v \rangle_{\Gamma_-} + (\frac{\mu}{\sigma}\partial_z u, \mu \partial_z v) + (\sigma u, v)$$
$$k(u, v) = (\sigma_s \bar{u}, v),$$
$$\ell(v) = 2\langle g, v \rangle_{\Gamma_-} + (q, v + \frac{\mu}{\sigma}\partial_z v).$$

**Observations:**

- $a$ is symmetric positive definite bilinear form on $\mathbb{W}^+$.
- Even-parity equations are well-posed.
- $\|v\|_a := a(v, v)^{1/2}$ is a norm.
- $\phi^+ = u$ and $\phi^- = (q^- - \mu \partial_z u)/\sigma$ can be retrieved from $u$.

# Iterative scheme, without correction

Given $u^n \in \mathbb{W}^+$, find $u^{n+1} \in \mathbb{W}^+$ such that

$$b(u^{n+1}, v) = k(u^n, v) + \ell(v) \quad \forall v \in \mathbb{W}^+.$$

**Error iteration:** $e^n = u - u^n$,

$$b(e^{n+1}, v) = k(e^n, v) \quad \forall v \in \mathbb{W}^+.$$

**Convergence in $L^2$:** Test with $v = e^{n+1}$, and use that

$$b(e^{n+1}, e^{n+1}) = \|e^{n+1}\|_\Gamma^2 + \|\frac{\mu}{\sqrt{\sigma}}\partial_z e^{n+1}\|_{L^2}^2 + \|\sqrt{\sigma} e^{n+1}\|_{L^2}^2,$$

$$k(e^n, e^{n+1}) \leq \|\frac{\sigma_s}{\sigma}\|_\infty \|\sqrt{\sigma} e^n\|_{L^2} \|\sqrt{\sigma} e^{n+1}\|_{L^2}.$$

Hence

$$\|\sqrt{\sigma} e^{n+1}\|_{L^2} \leq \|\frac{\sigma_s}{\sigma}\|_\infty \|\sqrt{\sigma} e^n\|_{L^2}.$$

This result will turn out to be too weak for our purpose.

# Convergence in stronger norm $\|\cdot\|_a$.

**Eigenvalue problem:** Find $(v_j, \lambda_j) \in \mathbb{W}^+ \times \mathbb{R}$ such that

$$a(v_j, v) = \lambda_j b(v_j, v) \quad \forall v \in \mathbb{W}^+, \quad \text{normalization: } b(v_i, v_j) = \delta_{i,j}.$$

**Expand errors in eigenvectors:**

$$e^n = \sum_{j=1}^{\infty} e_j^n v_j \qquad \text{with } e_j^n = b(e^n, v_j).$$

# Convergence in stronger norm $\|\cdot\|_a$.

**Eigenvalue problem:** Find $(v_j, \lambda_j) \in \mathbb{W}^+ \times \mathbb{R}$ such that

$$a(v_j, v) = \lambda_j b(v_j, v) \quad \forall v \in \mathbb{W}^+, \quad \text{normalization: } b(v_i, v_j) = \delta_{i,j}.$$

**Expand errors in eigenvectors:**

$$e^n = \sum_{j=1}^{\infty} e_j^n v_j \qquad \text{with } e_j^n = b(e^n, v_j).$$

**Error estimate:**

$$\|e^{n+1}\|_a^2 = \sum_j \lambda_j |e_j^{n+1}|^2$$

# Convergence in stronger norm $\|\cdot\|_a$.

**Eigenvalue problem:** Find $(v_j, \lambda_j) \in \mathbb{W}^+ \times \mathbb{R}$ such that

$$a(v_j, v) = \lambda_j b(v_j, v) \quad \forall v \in \mathbb{W}^+, \quad \text{normalization: } b(v_i, v_j) = \delta_{i,j}.$$

**Expand errors in eigenvectors:**

$$e^n = \sum_{j=1}^{\infty} e_j^n v_j \qquad \text{with } e_j^n = b(e^n, v_j).$$

**Error estimate:**

$$\|e^{n+1}\|_a^2 = \sum_j \lambda_j |e_j^{n+1}|^2 = \sum_j |1 - \lambda_j|^2 \lambda_j |e_j^n|^2$$

**Claim 1:** $e_j^{n+1} = (1 - \lambda_j) e_j^n$.

# Convergence in stronger norm $\|\cdot\|_a$.

**Eigenvalue problem:** Find $(v_j, \lambda_j) \in \mathbb{W}^+ \times \mathbb{R}$ such that

$$a(v_j, v) = \lambda_j b(v_j, v) \quad \forall v \in \mathbb{W}^+, \quad \text{normalization: } b(v_i, v_j) = \delta_{i,j}.$$

**Expand errors in eigenvectors:**

$$e^n = \sum_{j=1}^{\infty} e_j^n v_j \qquad \text{with } e_j^n = b(e^n, v_j).$$

**Error estimate:**

$$\|e^{n+1}\|_a^2 = \sum_j \lambda_j |e_j^{n+1}|^2 = \sum_j |1-\lambda_j|^2 \lambda_j |e_j^n|^2 \leq \|\frac{\sigma_s}{\sigma}\|_\infty^2 \|e^n\|_a^2.$$

**Claim 1:** $e_j^{n+1} = (1-\lambda_j) e_j^n$.

**Claim 2:** $1 - \|\frac{\sigma_s}{\sigma}\|_\infty \leq \lambda_j \leq 1$.

# Proof of claim 1: $e_j^{n+1} = (1 - \lambda_j) e_j^n$

By definition $b(e^{n+1}, v_j) = e_j^{n+1}$.

# Proof of claim 1: $e_j^{n+1} = (1-\lambda_j)e_j^n$

By definition $b(e^{n+1}, v_j) = e_j^{n+1}$.

**Recall error equation**

$$b(e^{n+1}, v) = k(e^n, v) \quad \forall v \in \mathbb{W}^+.$$

Since $k = b - a$, we obtain that

$$e_j^{n+1} = k(e^n, v_j) = b(e^n, v_j) - a(e^n, v_j) = (1-\lambda_j)e_j^n.$$

# Proof of claim 2: $1 - \|\frac{\sigma_s}{\sigma}\|_\infty \leq \lambda_j \leq 1$

By definition

$$\lambda_j = \lambda_j b(v_j, v_j) = a(v_j, v_j) = b(v_j, v_j) - k(v_j, v_j) = 1 - k(v_j, v_j).$$

# Proof of claim 2: $1 - \|\frac{\sigma_s}{\sigma}\|_\infty \leq \lambda_j \leq 1$

By definition

$$\lambda_j = \lambda_j b(v_j, v_j) = a(v_j, v_j) = b(v_j, v_j) - k(v_j, v_j) = 1 - k(v_j, v_j).$$

Since for all $v \in \mathbb{W}^+$

$$0 \leq k(v, v) = (\sigma_s \bar{v}, \bar{v}) \leq (\sigma_s v, v) \leq \|\frac{\sigma_s}{\sigma}\|_\infty (\sigma v, v) \leq \|\frac{\sigma_s}{\sigma}\|_\infty b(v, v),$$

we obtain the claim.

# Error equation and subspace correction

**Error equation:**

$$b(e^{n+1}, v) = k(e^n, v) \quad \forall v \in \mathbb{W}^+.$$

**NTE for error:** Using $a = b - k$,

$$a(e^{n+1}, v) = k(u^{n+1} - u^n, v) \quad \forall v \in \mathbb{W}^+.$$

**Subspace:** $\mathbb{W}_0^+ = \{v \in \mathbb{W}^+ : v(z, \mu) = \bar{v}(z)\}$.

**Correction equation:** Find $u_e^{n+1} \in \mathbb{W}_0^+$ such that

$$a(u_e^{n+1}, v) = k(u^{n+1} - u^n, v) \quad \forall v \in \mathbb{W}_0^+.$$

**New iterate:** $u^{n+1} + u_e^{n+1}$.

# Iterative scheme with correction

Given $u^n \in \mathbb{W}^+$, find $u^{n+1/2} \in \mathbb{W}^+$ such that

$$b(u^{n+1/2}, v) = k(u^n, v) + \ell(v) \quad \forall v \in \mathbb{W}^+.$$

**Subspace:** $\mathbb{W}_0^+ = \{v \in \mathbb{W}^+ : v(z,\mu) = \bar{v}(z)\}$.

**Correction equation:** Find $u_e^{n+1} \in \mathbb{W}_0^+$ such that

$$a(u_e^{n+1}, v) = k(u^{n+1/2} - u^n, v) \quad \forall v \in \mathbb{W}_0^+.$$

**New iterate:** $u^{n+1} := u^{n+1/2} + u_e^{n+1}$.

**Theorem:** For any $u^0 \in \mathbb{W}^+$, the iteration $u^n \mapsto u^{n+1}$ converges to the solution $u = \phi^+$ of the even-parity equation, and

$$\|u^{n+1} - u\|_a \leq \|\frac{\sigma_s}{\sigma}\|_\infty \|\|u^n - u\|_a.$$

# Convergence proof

cf. Ceá's lemma

**Galerkin orthogonality:**

$$a(e^{n+1}, v) = a(u_e^{n+1/2}, v) \qquad \forall v \in \mathbb{W}_0^+.$$

# Convergence proof

## cf. Céa's lemma

**Galerkin orthogonality:**

$$a(e^{n+1}, v) = a(u_e^{n+1/2}, v) \qquad \forall v \in \mathbb{W}_0^+.$$

For any $v \in \mathbb{W}_0^+$

$$\begin{aligned}
\|e^{n+1}\|_a^2 &= a(e^{n+1}, e^{n+1}) && \text{(Definition } \|\cdot\|_a\text{)} \\
&= a(e^{n+1}, e^{n+1/2} - u_e^{n+1/2}) && (e^{n+1} = e^{n+1/2} - u_e^{n+1/2}) \\
&= a(e^{n+1}, e^{n+1/2} - v) && \text{(Galerkin orthogonality)} \\
&\le \|e^{n+1}\|_a \|e^{n+1/2} - v\|_a. && \text{(Cauchy-Schwarz)}
\end{aligned}$$

# Convergence proof

cf. Ceá's lemma

**Galerkin orthogonality:**

$$a(e^{n+1}, v) = a(u_e^{n+1/2}, v) \qquad \forall v \in \mathbb{W}_0^+.$$

For any $v \in \mathbb{W}_0^+$

$$\begin{aligned}
\|e^{n+1}\|_a^2 &= a(e^{n+1}, e^{n+1}) && \text{(Definition } \|\cdot\|_a) \\
&= a(e^{n+1}, e^{n+1/2} - u_e^{n+1/2}) && (e^{n+1} = e^{n+1/2} - u_e^{n+1/2}) \\
&= a(e^{n+1}, e^{n+1/2} - v) && \text{(Galerkin orthogonality)} \\
&\leq \|e^{n+1}\|_a \|e^{n+1/2} - v\|_a. && \text{(Cauchy-Schwarz)}
\end{aligned}$$

Therefore

$$\|e^{n+1}\|_a \leq \inf_{v \in \mathbb{W}_0^+} \|e^{n+1/2} - v\|_a \leq \|e^{n+1/2}\|_a \leq \|\frac{\sigma_s}{\sigma}\|_\infty \|e^n\|_a.$$

# Discrete iterative scheme with correction

Choose $\mathbb{W}_h^+ \subset \mathbb{W}^+$.

Given $u_h^n \in \mathbb{W}_h^+$, find $u_h^{n+1/2} \in \mathbb{W}_h^+$ such that

$$b(u_h^{n+1/2}, v_h) = k(u_h^n, v_h) + \ell(v_h) \quad \forall v_h \in \mathbb{W}_h^+.$$

**Subspace:** $\mathbb{W}_{0,h}^+ = \{v_h \in \mathbb{W}_h^+ : v_h(z,\mu) = \bar{v}_h(z)\}$.

**Correction equation:** Find $u_{e,h}^{n+1} \in \mathbb{W}_{0,h}^+$ such that

$$a(u_{e,h}^{n+1}, v_h) = k(u_h^{n+1/2} - u_h^n, v_h) \quad \forall v_h \in \mathbb{W}_{0,h}^+.$$

**New iterate:** $u_h^{n+1} := u_h^{n+1/2} + u_{e,h}^{n+1}$.

**Theorem:** For any $u_h^0 \in \mathbb{W}_h^+$, the iteration $u_h^n \mapsto u_h^{n+1}$ converges to the solution $u_h = \phi^+$ of the discrete even-parity equation, and

$$\|u_h^{n+1} - u_h\|_a \leq \|\frac{\sigma_s}{\sigma}\|_\infty \|\|u_h^n - u_h\|_a.$$

# Relation of correction equations to PDEs

**Correction equation:** Find $u_e^{n+1} \in \mathbb{W}_0$ such that

$$b(u_e^{n+1}, \psi) = k(u_e^{n+1}, v) + k(u^{n+1/2} - u^n, v) \quad \forall v \in \mathbb{W}_0.$$

is the weak formulation of the diffusion equation

$$-\partial_z(\frac{1}{3\sigma}\partial_z u_e) + \sigma_a u_e = \sigma_s(\bar{u}^{n+1/2} - \bar{u}^n) \quad \text{in } (0, Z).$$

---

**Discrete correction equation:** Find $u_{e,h}^{n+1} \in \mathbb{W}_{0,h}$ such that

$$b(u_{e,h}^{n+1}, v) = k(u_h^{n+1}, v) + k(u_h^{n+1/2} - \phi_h^n, v) \quad \forall v \in \mathbb{W}_{0,h}$$

is the weak formulation of the diffusion equation

$$-\partial_z(D_N \partial_z u_{e,h}) + \sigma_a u_{e,h} = \sigma_s(\bar{u}_{e,h}^{n+1/2} - \bar{u}_{e,h}^n) \quad \text{in } (0, Z).$$

with $D_N(z) = \frac{1}{3\sigma}\left(1 + \frac{1}{4}\sum_n \Delta\mu^3\right)$.

# Numerical realization of the scheme: Transport step

Choose $\mathbb{W}_h^+ \subset \mathbb{W}^+$ as in dG method, i.e.,

$$v_h(z,\mu) = \sum_{n=1}^{N} \sum_{j=0}^{J} c_{j,n}^+ \varphi_j(z) Q_n^+(\mu),$$

with hat functions $\varphi_j$ and piecewise constant $Q_n^+$.

Given $u_h^n \in \mathbb{W}_h^+$, find $u_h^{n+1/2} \in \mathbb{W}_h^+$ such that

$$b(u_h^{n+1/2}, v_h) = k(u_h^n, v_h) + \ell(v_h) \quad \forall v_h \in \mathbb{W}_h^+,$$

translates to: Given $\mathbf{u}^n$, solve for $\mathbf{u}^{n+1/2}$

$$\left(\mathbf{R} + \mathbf{N} \otimes \mathbf{M}(\sigma)^+ + (\mathbf{P}^T\mathbf{N}^{-1}\mathbf{P} \otimes \mathbf{D}^T\mathbf{C}\mathbf{D})\right) \mathbf{u}^{n+1/2} = (\mathbf{K} \otimes \mathbf{M}(\sigma_s)^+)\mathbf{u}^n + \mathbf{f}.$$

- matrices on LHS are sparse
- $\mathbf{P}^T\mathbf{N}^{-1}\mathbf{P}$, and $\mathbf{N}$ are diagonal: matrix on LHS is block-diagonal.
  - can be solved in parallel.
  - each system corresponds to an elliptic equation.
- application of dense matrix $\mathbf{K}$ is cheap.

# Numerical realization of the scheme: Subspace correction

**Correction equation:** Find $u_{e,h}^{n+1} \in \mathbb{W}_{0,h}^{+}$ such that

$$a(u_{e,h}^{n+1}, v_h) = k(u_h^{n+1/2} - u_h^n, v_h) \quad \forall v_h \in \mathbb{W}_{0,h}^{+}.$$

**New iterate:** $u_h^{n+1} := u_h^{n+1/2} + u_{e,h}^{n+1}$.

Translates to: Given $\mathbf{u}^n$, $\mathbf{u}^{n+1/2}$, solve for $\mathbf{u}_e^{n+1}$

$$(\mathbf{B} + \mathbf{M}(\sigma_a)^{+} + \mathbf{D}^T\mathbf{C}\mathbf{D})\mathbf{u}_e^{n+1} = (\frac{1}{2}\mathbf{e}^T\mathbf{K} \otimes \mathbf{M}(\sigma_s)^{+})(\mathbf{u}^{n+1/2} - \mathbf{u}^n).$$

$\mathbf{u}^{n+1} = \mathbf{u}^{n+1/2} + \mathbf{Q}\mathbf{u}_e^{n+1}$.

- $\mathbf{Q} = \mathbf{e} \otimes \mathbf{I}$ prolongates coefficients of functions in $\mathbb{W}_{h,0}^{+}$ to $\mathbb{W}_h^{+}$.
- Correction equation is a small elliptic equation.

# Numerical tests

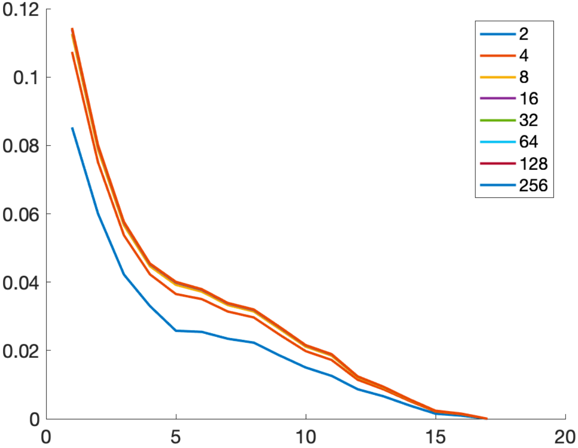
$$\sigma_s(z) = \begin{cases} 2 + \sin(2\pi z), & z \leq \frac{1}{2} \\ 102 + \sin(2\pi z), & z > \frac{1}{2}, \end{cases} \qquad \sigma_a(z) = \begin{cases} 10.01, & z \leq \frac{1}{2} \\ 0.01, & z > \frac{1}{2}. \end{cases}$$

Proven convergence rate for iteration without subspace correction

$$\|\sigma_s/\sigma_t\|_\infty \approx 0.9999$$

# Spectrum of error propagator $\bar{e}_h^n \mapsto \bar{e}_h^{n+1}$ for different $N$

J=16

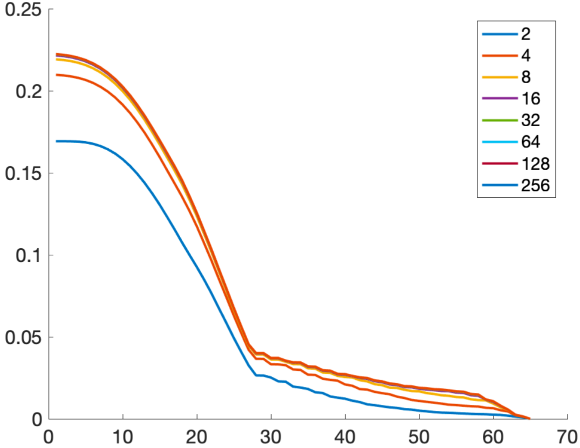



Largest eigenvalues $\rho \leq 0.12$, i.e.,

$$\|u_h^n - u_h\|_a \leq \rho^n \|\|u_h^0 - u_h\|_a.$$

# Spectrum of error propagator $\bar{e}_h^n \mapsto \bar{e}_h^{n+1}$ for different $N$

J=64

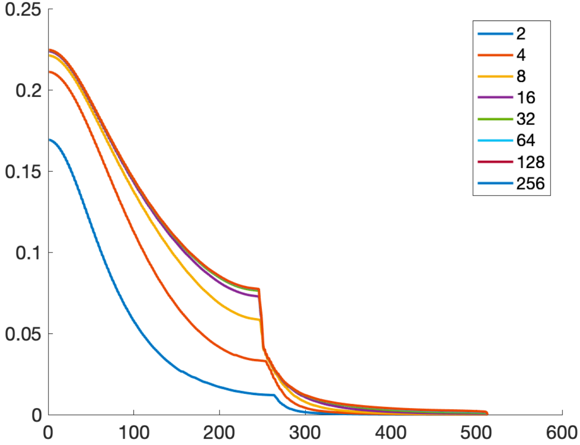



Largest eigenvalues $\rho \leq 0.25$, i.e.,

$$\|u_h^n - u_h\|_a \leq \rho^n \|\|u_h^0 - u_h\|_a.$$

# Spectrum of error propagator $\bar{e}_h^n \mapsto \bar{e}_h^{n+1}$ for different $N$

J=512




Largest eigenvalues $\rho \leq 0.25$, i.e.,

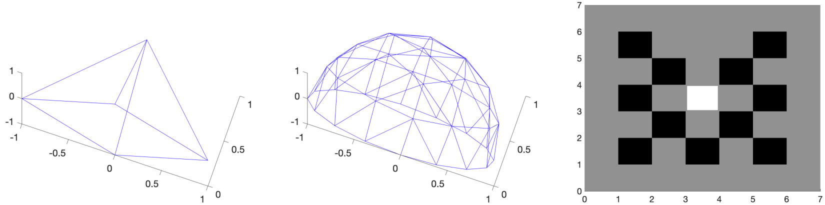
$$\|u_h^n - u_h\|_a \leq \rho^n \| \|u_h^0 - u_h\|_a.$$

# Multi-D: Lattice problem

$$\mathbf{s} \cdot \nabla_{\mathbf{r}} \phi(\mathbf{r}, \mathbf{s}) + \sigma(\mathbf{r})\phi(\mathbf{r}, \mathbf{s}) = \sigma_s(\mathbf{r})\bar{\phi} + q(\mathbf{r}, \mathbf{s}), \qquad \text{(NTE)}$$

for $\mathbf{r} \in (0,7) \times (0,7)$, $\mathbf{s} \in \mathbb{S}^2$.

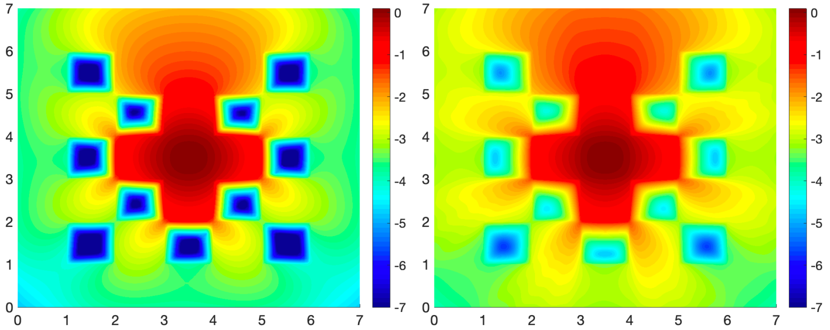
All results translate verbatim!




Left and middle: Approximation of the half-sphere with $N = 4$ and $N = 64$ triangles. Right: Geometry of the lattice problem. Here, $\sigma_a = 10$ and $\sigma = \sigma_a$ in black regions, $\sigma_a = 0$ and $\sigma = 1$ else; $q = 1$ in white region, $q = 0$ else.

[Palii & S 2020]

# Lattice problem: results

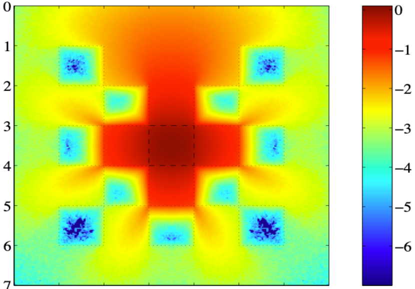



Neutron density in a $\log_{10}$-scale for the lattice problem for $J = 9\,801$ spatial vertices and $N = 4$ triangles on a half-sphere (left) and $J = 78\,961$ spatial vertices and $N = 64$ triangles on a half-sphere (right).

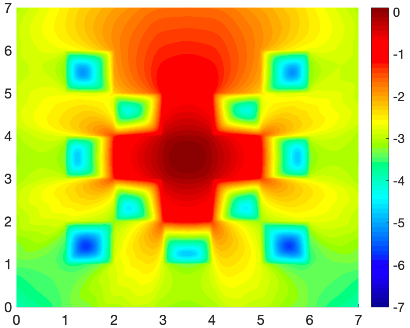
Stopping criterion: $\|u_h^{n+1} - u_h^n\|_a \leq 10^{-10}$.
Observed rate: $\|u_h^{n+1} - u_h^n\|_a \leq 0.2\|u_h^n - u_h^{n-1}\|_a$, i.e., 17 iterations.

[Palii & S 2020]

# Lattice problem: results




(f) Implicit Monte Carlo

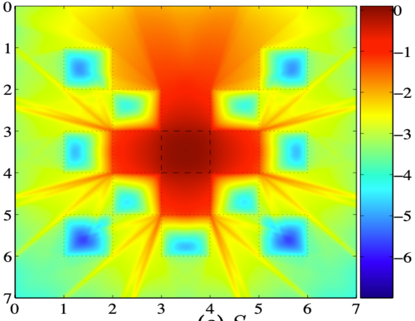



Comparison to Monte Carlo (left) [Brunner] and our approximation with $N = 64$ (right).

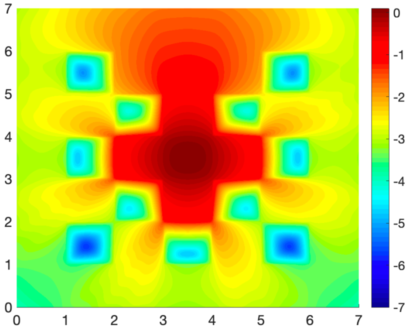
Stopping criterion: $\|u_h^{n+1} - u_h^n\|_a \leq 10^{-10}$.
Observed rate: $\|u_h^{n+1} - u_h^n\|_a \leq 0.2\|u_h^n - u_h^{n-1}\|_a$, i.e., 17 iterations.

[Palii & S 2020]

# Lattice problem: results




(e) $S_6$




Comparison to standard $S_6$ discrete ordinates method (left) [Brunner] and our approximation with $N = 64$ (right).

Stopping criterion: $\|u_h^{n+1} - u_h^n\|_a \leq 10^{-10}$.
Observed rate: $\|u_h^{n+1} - u_h^n\|_a \leq 0.2\|u_h^n - u_h^{n-1}\|_a$, i.e., 17 iterations.
Side note: No "ray effect" in DG solution (left)

[Palii & S 2020]

# Convergence behavior in a diffusion scaling

$$\sigma_s^\varepsilon(\mathbf{r}) = \frac{\sigma_s(\mathbf{r}) + 1/10}{\varepsilon}, \qquad \sigma_a^\varepsilon = \varepsilon(\sigma_a(\mathbf{r}) + 1/10), \qquad q^\varepsilon(x,s) = \varepsilon q(\mathbf{r}).$$

$\|\sigma_s/\sigma\|_\infty = O(1-\varepsilon^2)$ for $\varepsilon \to 0$

| | $J = 9\,801$ | | | | $J = 78\,961$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $N=4$ | | $N=64$ | | $N=4$ | | $N=64$ | |
| $1/\varepsilon$ | $n$ | rate | $n$ | rate | $n$ | rate | $n$ | rate |
| 1 | 9 | 0.04 | 15 | 0.16 | 9 | 0.04 | 15 | 0.17 |
| 10 | 9 | 0.06 | 15 | 0.22 | 9 | 0.06 | 16 | 0.25 |
| 100 | 8 | 0.06 | 13 | 0.22 | 9 | 0.07 | 15 | 0.27 |
| 1000 | 5 | 0.01 | 7 | 0.06 | 6 | 0.05 | 10 | 0.17 |

Iteration counts $n$ and minimal reduction rates for $\|\phi_h^n - \phi_h^{n-1}\|_a$ for the lattice problem with scaled parameters $\sigma_s^\varepsilon$, $\sigma_a^\varepsilon$ and $q^\varepsilon$ for different $\varepsilon$ and discretizations with $N$ triangles on a half-sphere and $J$ vertices in the spatial mesh.

[Palii & S 2020]

# Convergence of DSA scheme: classical vs variational

**Classical discrete ordinates method** [Adams & Larsen 2002]

- Diffusion synthetic acceleration motivated by asymptotic analysis.
- For semidiscrete problem with periodic b.c. and constant coefficients

$$\|\bar{e}^{n+1}\|_2 \leq \|\frac{\sigma_s}{\sigma}\|_\infty \|\bar{e}^n\|_2.$$

- Inconsistend discretization can lead to divergence.

**Variational approach** [Palii & S 2020]

- The iteration always converges:
  - varying and (possibly) discontinuous coefficients
  - non-periodic b.c.
  - independent of the spatial discretization
- convergence is also fast (mathematical proof misses)

---

[Habetler & Matkowsky '75] [Larsen & Keller '74] [Dautray, Lions '93] [Bardos et al '87] [Egger & S 2014] [Adams & Larsen 2001]

# Convergence of DSA scheme: classical vs variational

**Classical discrete ordinates method** [Adams & Larsen 2002]

- Diffusion synthetic acceleration motivated by asymptotic analysis.
- For semidiscrete problem with periodic b.c. and constant coefficients

$$\|\bar{e}^{n+1}\|_2 \leq \|\frac{\sigma_s}{\sigma}\|_\infty \|\bar{e}^n\|_2.$$

- Inconsistend discretization can lead to divergence.

**Variational approach** [Palii & S 2020]

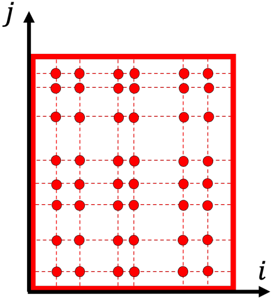
- The iteration always converges:
  - varying and (possibly) discontinuous coefficients
  - non-periodic b.c.
  - independent of the spatial discretization
- convergence is also fast (mathematical proof misses)
- can be extended to anisotropic scattering [Dölz et al, 2022]:
  - matrix compression techniques for applying scattering integral
  - larger subspaces for correction

[Habetler & Matkowsky '75] [Larsen & Keller '74] [Dautray, Lions '93] [Bardos et al '87] [Egger & S 2014] [Adams & Larsen 2001]

# Classical approximations

use a tensor product approximation

$$\phi(\mathbf{r}, \mathbf{s}) \approx \sum_{i=1}^{I} \sum_{j=1}^{J} u_{i,j} \phi_i(\mathbf{r}) H_j(\mathbf{s})$$




If $\mathcal{R}$ and $\mathbb{S}^2$ are partitioned by quasi-uniform triangulations with mesh-size $h$:

$$I \sim h^{-d}, \qquad J \sim h^{-d+1}.$$

**Storage** is proportional to $IJ \approx h^{-2d+1}$

[Chandrasekhar ('50)] [Case+Zweifel ('67)] [Duderstadt+Martin ('79)] [Lewis+Miller ('84)] [Manteuffel et al (2000)] [Egger+S (2010)], and many more

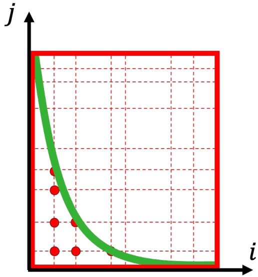
# More efficient approaches

**Sparse tensor products**
[Widmer et al (2008)],

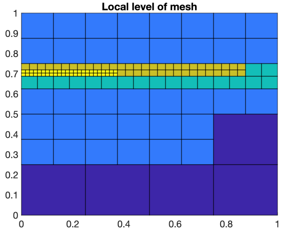
[Grella+Schwab (2011a,b)]

$$\phi(\mathbf{r}, \mathbf{s}) \approx \sum_{1 \le f(i,j) \le I} u_{i,j} \phi_i(\mathbf{r}) H_j(\mathbf{s})$$

**Storage** is proportional to $I \log I$




**Phase-space adaptive methods**
[Kophazy+Lathouwers (2015)]
[Dahmen et al (2020)]
[Palii+S (2022)]



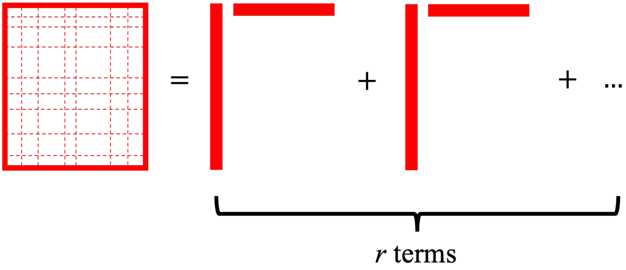
# Low-rank tensor product approximations

**Coefficient matrix of solution U** $= (u_{i,j})_{i,j} \in \mathbb{R}^{I \times J}$

**Approximate U** by short sums of rank one matrices:

$$\mathbf{U} \approx \sum_{k=1}^{r} \mathbf{v}_k \otimes \mathbf{w}_k, \qquad \mathbf{v}_k \in \mathbb{R}^I, \mathbf{w}_k \in \mathbb{R}^J$$

**Storage** is proportional to $r(I + J)$



# Recall: Even-parity formulation

**Even-parity equation** (in variational form): Find $u = \phi^+ \in \mathbb{W}^+$ such that

$$a(u, v) = \ell(v) \quad \forall v \in \mathbb{W}^+,$$

where

$$a(u, v) = \langle|\mu|u, v\rangle_\Gamma + (\frac{\mu}{\sigma}\partial_z u, \mu\partial_z v) + (\sigma u, v) - (\sigma_s \bar{u}, v),$$
$$\ell(v) = (q, v + \frac{\mu}{\sigma}\partial_z v) + 2\langle|\mu|g, v\rangle_{\Gamma_-}.$$

**Observations:**

- $a$ is symmetric positive definite bilinear form on $\mathbb{W}^+$.
- Even-parity equations are well-posed (Lax-Milgram lemma).
- $\|v\|_a := a(v, v)^{1/2}$ is a norm.
- Odd part $\phi^- = \frac{1}{\sigma}(q^- - \mu\partial_z u)$ can be retrieved from $u$.

# Structure of even-parity system

**Recall:** After discretization $\mathbf{A}\mathbf{u} = \mathbf{f}$ with

$$\mathbf{A} := \mathbf{R} + \mathbf{A}^{+} + \left(\mathbf{P}^{T}\mathbf{N}^{-1}\mathbf{P} \otimes \mathbf{D}^{T}\mathbf{C}\mathbf{D}\right)$$

with

$$\mathbf{R} = \mathbf{B} \otimes \mathrm{diag}(1, 0, \ldots, 0, 1), \qquad \text{'boundary' matrix}$$
$$\mathbf{A}^{+} = \mathbf{N} \otimes \mathbf{M}(\sigma)^{+} - \mathbf{K} \otimes \mathbf{M}(\sigma_s)^{+}, \quad \text{'attenuation' matrix}$$

is a *short sum of Kronecker products*:

$$\mathbf{A} = \sum_{k=1}^{4} \mathbf{A}_k \otimes \mathbf{B}_k$$

with sparse or low-rank matrices $\mathbf{A}_k \in \mathbb{R}^{J \times J}$ and $\mathbf{B}_k \in \mathbb{R}^{I \times I}$.

# Computational complexity of matrix-vector products

For $\mathbf{A} = \sum_{k=1}^{4} \mathbf{A}_k \otimes \mathbf{B}_k$ and

$$\mathbf{U} = \mathrm{mat}(\mathbf{u})$$

we have

$$\mathrm{mat}(\mathbf{Au}) = \sum_{i=1}^{4} \underbrace{\mathbf{B}_i \mathbf{U} \mathbf{A}_i^T}_{O(IJ)}.$$

**Storage and Flops** for MatVec **Au** are $O(IJ)$.

Iterative schemes are suitable.

# Computational complexity of matrix-vector products

For $\mathbf{A} = \sum_{k=1}^{4} \mathbf{A}_k \otimes \mathbf{B}_k$ and

$$\mathbf{U} = \mathrm{mat}(\mathbf{u}) = \sum_{k=1}^{r} \mathbf{v}_k \otimes \mathbf{w}_k$$

we have

$$\mathrm{mat}(\mathbf{Au}) = \sum_{i=1}^{4} \underbrace{\mathbf{B}_i \mathbf{U} \mathbf{A}_i^T}_{O(IJ)} = \sum_{i=1}^{4} \underbrace{\sum_{k=1}^{r} (\mathbf{B}_i \mathbf{v}_k) \otimes (\mathbf{A}_i \mathbf{w}_k)}_{O(r(I+J))}.$$

**Storage and Flops** for MatVec **Au** are $O(r(I+J))$.
However: Rank $r$ grows by a factor of 4.
**Conclusion** for iterative schemes that should exploit low rank of **U**:

- control growth of ranks
- aim for few iterations ($\rightsquigarrow$ preconditioning)

# Composition of contractions

**Lemma.** Let $S : X \to Y$, $T : Y \to Z$ be Lipschitz continuous with Lipschitz constants $L_S$ and $L_T$. Then $T \circ S$ is Lipschitz with

$$\|T(S(x_1)) - T(S(x_2))\|_Z \leq L_S L_T \|x_1 - x_2\|_X \quad \forall x_1, x_2 \in X.$$

**Implications:** If $S$ describes a preconditioned Richardson iteration from above, then

- $L_S < 1$ (contraction).
- If $T$ describes rank truncation, we require $L_T \leq 1$ (non-expansive). Then $T \circ S$ is a convergent scheme.
- The norms are important!

# Truncated singular value decomposition

Let $\mathbf{U} \in \mathbb{R}^{I\times J}$ be of rank $n$.

**Singular value decomposition**

$$\mathbf{U} = \mathbf{W}\Sigma\mathbf{V}^T, \quad \mathbf{W} \in \mathbb{R}^{I\times n}, \Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_n), \mathbf{V} \in \mathbb{R}^{J\times n},$$

with $\sigma_j > 0$, $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

**Eckart-Young-Mirsky Theorem:**

$$\min_{\mathbf{Z}\in\mathbb{R}^{I\times J}, \mathrm{rank}(\mathbf{Z})=k} \|\mathbf{U} - \mathbf{Z}\|_F = \sigma_{k+1} = \|\mathbf{U} - \mathbf{W}_k\Sigma_k\mathbf{V}_k^T\|_F.$$

with $\mathbf{W}_k$ the first $k$-columns of $\mathbf{W}$, $\mathbf{V}_k$ the first $k$-columns of $\mathbf{V}$, $\Sigma_k = \mathrm{diag}(\sigma_1, \dots, \sigma_k)$.

**Issues:**

- How to interpret the Frobenius norm $\|\mathbf{U}\|_F^2 = \sum_{i,j} |U_{i,j}|^2$ in function space context?
- Denoting truncated SVD of $\mathbf{U}$ by $T_k(\mathbf{U}) = \mathbf{W}_k\Sigma_k\mathbf{V}_k^T$, we do not have

$$\|T_k(\mathbf{U}_1) - T_k(\mathbf{U}_2)\|_F \leq \|\mathbf{U}_1 - \mathbf{U}_2\|_F.$$

# Non-expansive rank truncation

Let $\mathbf{U} \in \mathbb{R}^{I\times J}$ be of rank $n$. Singular value decomposition

$$\mathbf{U} = \mathbf{W}\Sigma\mathbf{V}^T, \quad \mathbf{W} \in \mathbb{R}^{I\times n}, \Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_n), \mathbf{V} \in \mathbb{R}^{J\times n},$$

with $\sigma_j > 0$, $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

**Soft-thresholding:** $s_\delta(t) = \mathrm{sgn}(t)\max\{|t| - \delta, 0\}$.

Define $\mathbf{S}_\delta(\mathbf{U}) = \mathbf{W}\,\mathrm{diag}(s_\delta(\sigma_1), \dots, s_\delta(\sigma_n))\,\mathbf{V}^T$.
Note: all singular values $\sigma_j < \delta$ are set to zero.

**$\mathbf{S}_\delta$ is non-expansive:** $\|\mathbf{S}_\delta(\mathbf{U}_1) - \mathbf{S}_\delta(\mathbf{U}_2)\|_F \leq \|\mathbf{U}_1 - \mathbf{U}_2\|_F$.

---

[Bachmayr & Schneider, 2017]

# Non-expansive rank truncation

Let $\mathbf{U} \in \mathbb{R}^{I \times J}$ be of rank $n$. Singular value decomposition

$$\mathbf{U} = \mathbf{W}\Sigma\mathbf{V}^T, \quad \mathbf{W} \in \mathbb{R}^{I \times n}, \Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_n), \mathbf{V} \in \mathbb{R}^{J \times n},$$

with $\sigma_j > 0$, $\mathbf{W}^T\mathbf{W} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$.

**Soft-thresholding:** $s_\delta(t) = \mathrm{sgn}(t) \max\{|t| - \delta, 0\}$.

Define $\mathbf{S}_\delta(\mathbf{U}) = \mathbf{W}\, \mathrm{diag}(s_\delta(\sigma_1), \dots, s_\delta(\sigma_n))\, \mathbf{V}^T$.
Note: all singular values $\sigma_j < \delta$ are set to zero.

**$\mathbf{S}_\delta$ is non-expansive:** $\|\mathbf{S}_\delta(\mathbf{U}_1) - \mathbf{S}_\delta(\mathbf{U}_2)\|_F \leq \|\mathbf{U}_1 - \mathbf{U}_2\|_F$.

**Next step:** Find transformation $\mathbf{W} = f(\mathbf{U})$ s.t. $\|\mathbf{W}\|_F \sim \|\mathbf{U}\|_{\mathbf{A}} = \|u_h\|_a$, and apply rank truncation to $\mathbf{W}$.

---

[Bachmayr & Schneider, 2017]

# Equivalent inner products

**Recall:** Even-parity bilinear form

$$a(u,v) = \langle |\mu| u, v\rangle_\Gamma + (\frac{\mu}{\sigma}\partial_z u, \mu\partial_z v) + (\sigma u, v) - (\sigma_s \bar{u}, v)$$

**Lemma.** There exists constants $\gamma_1, \gamma_2 > 0$ such that

$$\gamma_1 p_*(v,v) \leq a(v,v) \leq \gamma_2 p_*(v,v) \quad \forall v \in \mathbb{W}^+,$$

with

$$p_*(u,v) := (\mu^2 \partial_z u, \partial_z v) + ((1+\mu^2)u, v).$$

# Spectrally equivalent matrices

$$a(u,v) = \langle u,v\rangle_{\Gamma_-} + (\frac{\mu}{\sigma}\partial_z u, \mu\partial_z v) + (\sigma u, v) - (\sigma_s \bar{u}, v),$$
$$p_*(u,v) = (\mu^2\partial_z u, \partial_z v) + ((1+\mu^2)u, v).$$

**Corollary.** For all $\mathbf{x} \in \mathbb{R}^{IJ}$ it holds

$$\gamma_1\langle \mathbf{P}_*\mathbf{x}, \mathbf{x}\rangle \leq \langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle \leq \gamma_2\langle \mathbf{P}_*\mathbf{x}, \mathbf{x}\rangle$$

with matrix

$$\mathbf{P}_* = \mathbf{T} \otimes (\mathbf{K} + \mathbf{M}) + \mathbf{N} \otimes \mathbf{M},$$
$$\mathbf{M} = \mathbf{M}(1)^+, \quad \mathbf{K} = \mathbf{D}^T(\mathbf{M}(1)^-)^{-1}\mathbf{D}.$$

# Variable transformation

**Recall:** $\mathbf{P}_* = \mathbf{T} \otimes (\mathbf{K} + \mathbf{M}) + \mathbf{N} \otimes \mathbf{M}$

**Cholesky factorization:** $\mathbf{U}_z^T \mathbf{U}_z = \mathbf{K} + \mathbf{M}$ with bidiagonal $\mathbf{U}_z$, yields

$$\mathbf{P}_* = \left(\mathbf{N}^{1/2} \otimes \mathbf{U}_z^T\right)\left(\tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}\right)\left(\mathbf{N}^{1/2} \otimes \mathbf{U}_z\right).$$

**Lemma.** For all $u_h \in \mathbb{W}_h^+$ it holds

$$\gamma_1 \|\mathbf{w}\|_2^2 \leq \|u_h\|_a^2 \leq \gamma_2 \|\mathbf{w}\|_2^2,$$

with $\mathbf{w} = \tilde{\mathbf{P}}_*^{1/2}(\mathbf{N}^{1/2} \otimes \mathbf{U}_z)\mathbf{u}$, and $\tilde{\mathbf{P}}_* = \tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}$.

**Take away:** Control over $\mathbf{w}$ in Euclidean norm implies control over $u_h$ in energy norm.

# Variable transformation

**Recall:** $\mathbf{P}_* = \mathbf{T} \otimes (\mathbf{K} + \mathbf{M}) + \mathbf{N} \otimes \mathbf{M}$

**Cholesky factorization:** $\mathbf{U}_z^T \mathbf{U}_z = \mathbf{K} + \mathbf{M}$ with bidiagonal $\mathbf{U}_z$, yields

$$\mathbf{P}_* = \left(\mathbf{N}^{1/2} \otimes \mathbf{U}_z^T\right)\left(\tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}\right)\left(\mathbf{N}^{1/2} \otimes \mathbf{U}_z\right).$$

**Lemma.** For all $u_h \in \mathbb{W}_h^+$ it holds

$$\gamma_1 \|\mathbf{w}\|_2^2 \leq \|u_h\|_a^2 \leq \gamma_2 \|\mathbf{w}\|_2^2,$$

with $\mathbf{w} = \tilde{\mathbf{P}}_*^{1/2}(\mathbf{N}^{1/2} \otimes \mathbf{U}_z)\mathbf{u}$, and $\tilde{\mathbf{P}}_* = \tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}$.

**Take away:** Control over $\mathbf{w}$ in Euclidean norm implies control over $u_h$ in energy norm.

**Proof:**

$$\begin{aligned}
\|u_h\|_a^2 &= a(u_h, u_h) \sim p_*(u_h, u_h) && \text{(Equivalence } a \sim p_*\text{)} \\
&= \mathbf{u}^T \mathbf{P}_* \mathbf{u} && \text{(using coordinates)} \\
&= ((\mathbf{N}^{1/2} \otimes \mathbf{U}_z)\mathbf{u})^T \tilde{\mathbf{P}}_* ((\mathbf{N}^{1/2} \otimes \mathbf{U}_z)\mathbf{u}) && \text{(Factorization of } \mathbf{P}_*\text{)} \\
&= \|\tilde{\mathbf{P}}_*^{1/2}(\mathbf{N}^{1/2} \otimes \mathbf{U}_z)\mathbf{u}\|_2^2.
\end{aligned}$$

# Transformed linear system

**Linear system $\mathbf{A}\mathbf{u} = \mathbf{f}$** is equivalent to

**Preconditioned linear system**

$$\tilde{\mathbf{P}}_*^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{P}}_*^{-1/2}\mathbf{w} = \tilde{\mathbf{f}}$$

with

$$\begin{aligned}
\tilde{\mathbf{A}} &:= (\mathbf{N}^{1/2} \otimes \mathbf{U}_z^T)^{-1}\mathbf{A}(\mathbf{N}^{1/2} \otimes \mathbf{U}_z)^{-1} \\
\tilde{\mathbf{P}}_* &= \tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}} \\
\tilde{\mathbf{f}} &:= \tilde{\mathbf{P}}_*^{-1/2}(\mathbf{N}^{1/2} \otimes \mathbf{U}_z^T)^{-1}\mathbf{f}.
\end{aligned}$$

# Transformed linear system

**Linear system $\mathbf{A}\mathbf{u} = \mathbf{f}$** is equivalent to

**Preconditioned linear system**

$$\tilde{\mathbf{P}}_*^{-1/2} \tilde{\mathbf{A}} \tilde{\mathbf{P}}_*^{-1/2} \mathbf{w} = \tilde{\mathbf{f}}$$

with

$$\tilde{\mathbf{A}} := (\mathbf{N}^{1/2} \otimes \mathbf{U}_z^T)^{-1} \mathbf{A} (\mathbf{N}^{1/2} \otimes \mathbf{U}_z)^{-1}$$
$$\tilde{\mathbf{P}}_* = \tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}$$
$$\tilde{\mathbf{f}} := \tilde{\mathbf{P}}_*^{-1/2} (\mathbf{N}^{1/2} \otimes \mathbf{U}_z^T)^{-1} \mathbf{f}.$$

**Question:** How to apply $\tilde{\mathbf{P}}_*^{-1/2}$ in a way that is compatible with the low-rank approach?

# Transformed linear system

**Linear system $\mathbf{A}\mathbf{u} = \mathbf{f}$** is equivalent to

**Preconditioned linear system**

$$\tilde{\mathbf{P}}_*^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{P}}_*^{-1/2}\mathbf{w} = \tilde{\mathbf{f}}$$

with

$$\begin{aligned}
\tilde{\mathbf{A}} &:= (\mathbf{N}^{1/2} \otimes \mathbf{U}_z^T)^{-1}\mathbf{A}(\mathbf{N}^{1/2} \otimes \mathbf{U}_z)^{-1} \\
\tilde{\mathbf{P}}_* &= \tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}} \\
\tilde{\mathbf{f}} &:= \tilde{\mathbf{P}}_*^{-1/2}(\mathbf{N}^{1/2} \otimes \mathbf{U}_z^T)^{-1}\mathbf{f}.
\end{aligned}$$

**Question:** How to apply $\tilde{\mathbf{P}}_*^{-1/2}$ in a way that is compatible with the low-rank approach?

**Calculus** $\exp(\tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}) = \exp(\tilde{\mathbf{T}}) \otimes \exp(\tilde{\mathbf{M}})$.

**Approach:** 'Interpolate' $\tilde{\mathbf{P}}_*^{-1/2}$ using sums of exponentials.

# Complex function theory

**Gamma function**

$$\Gamma(z) = \int_0^{\infty} e^{-t} t^{z-1}\, dt, \quad \mathrm{Re}(z) > 0.$$

**[Scholz & Yserentant (2017)]:** It holds for $r > 0$ and $\mathrm{Re}(z) > 0$ :

$$\Gamma(z) = r^z \int_{-\infty}^{\infty} \exp(-re^t + zt)\, dt.$$

Therefore, for arbitrary $\beta > 0$:

$$\frac{1}{r^\beta} = \frac{1}{\Gamma(\beta)} \int_{-\infty}^{\infty} \exp(-re^t + \beta t)\, dt$$

**Observation:** The integrand decays rapidly for $t \to \pm\infty$.

**Idea:** Approximate the integral with the trapezoidal rule, and truncate:

$$\frac{1}{r^\beta} \approx \frac{h}{\Gamma(\beta)} \sum_{k=k_1}^{k_2} \exp(-re^{kh}) e^{kh\beta}$$

# Exponential sum approximation of the ideal preconditioner

Recall $\exp(\tilde{\mathbf{P}}_*) = \exp(\tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}) = \exp(\tilde{\mathbf{T}}) \otimes \exp(\tilde{\mathbf{M}})$.

We obtain

$$\begin{aligned}
\tilde{\mathbf{P}}_*^{-1/2} &= \frac{1}{\Gamma(1/2)} \int_{-\infty}^{\infty} \exp(-\tilde{\mathbf{P}}_* e^t) e^{t/2}\, dt && \text{(Functional calculus)} \\
&\approx \frac{h}{\Gamma(\beta)} \sum_{k=k_1}^{k_2} \exp(-\tilde{\mathbf{P}}_* e^{kh}) e^{kh/2} && \text{(Trapezoidal rule)} \\
&= \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \rho_k \exp(-\alpha_k \tilde{\mathbf{T}}) \otimes \exp(-\alpha_k \tilde{\mathbf{M}}) \\
&=: \tilde{\mathbf{P}}^{-1/2}
\end{aligned}$$

Use $\tilde{\mathbf{P}}^{-1/2}$ instead of $\mathbf{P}_*^{-1/2}$ in the numerical scheme.

---

[Braess+Hackbusch (2005)][Beylkin+Monzon (2010)], [Scholz+Yserentant (2017)],

[Yserentant (2020)]

# Accuracy of exponential sum approximation

$$\frac{1}{r^{1/2}} \approx \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} e^{kh/2} \exp(-e^{kh} r)$$

For $\epsilon > 0$, choose $h, k_1, k_2$ such that for all eigenvalues $r$ of $\tilde{\mathbf{P}}_*$:

$$1 - \epsilon \leq \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+\ln(r)} + (kh + \ln(r))/2) \leq 1 + \epsilon.$$

Then, by functional calculus,

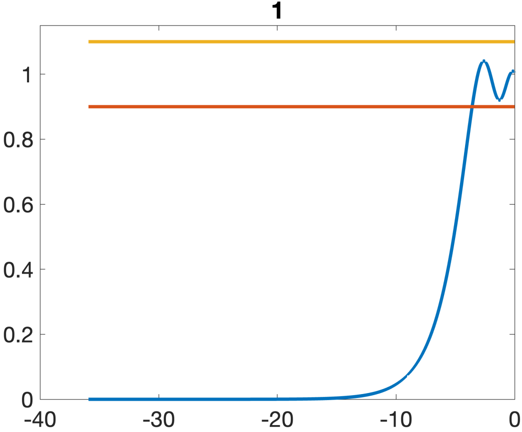
$$\|(\tilde{\mathbf{P}}^{-1/2} - \tilde{\mathbf{P}}_*^{-1/2})\mathbf{x}\|_2 \leq \epsilon \|\tilde{\mathbf{P}}_*^{-1/2}\mathbf{x}\|_2.$$

# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 1$.

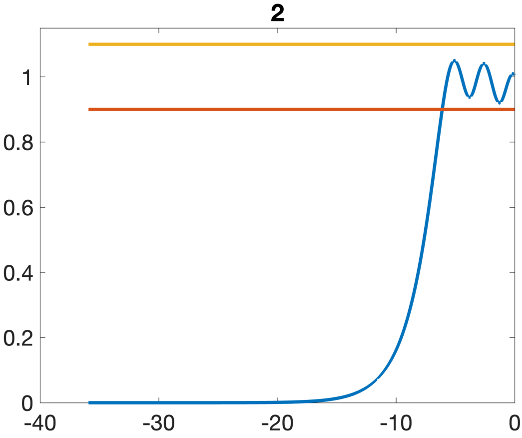


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 2$.

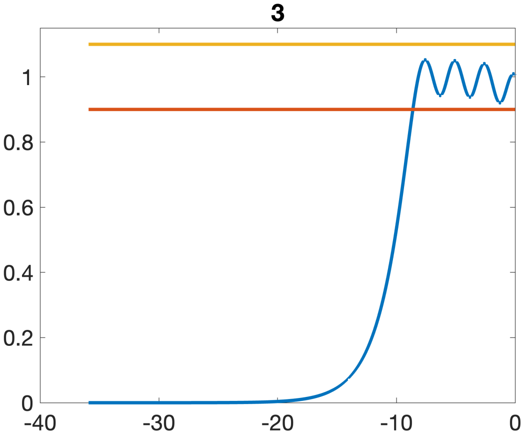


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 3$.

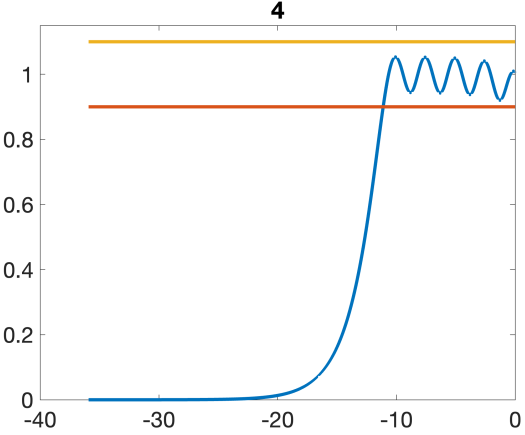


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2, \quad k_2 = 4$.

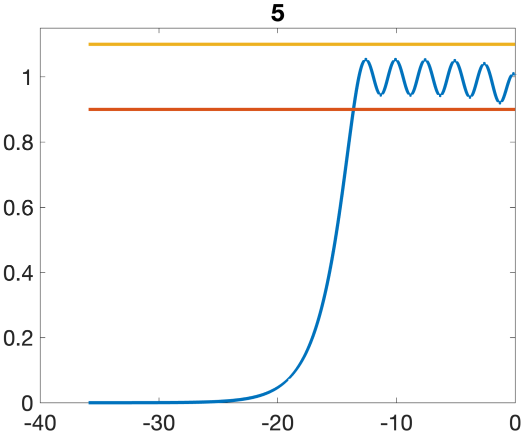


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $\quad k_2 = 5$.

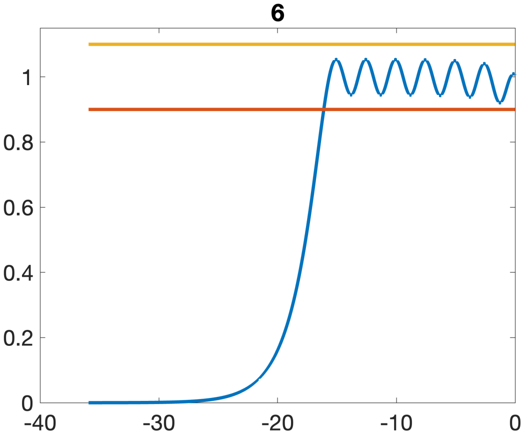


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 6$.

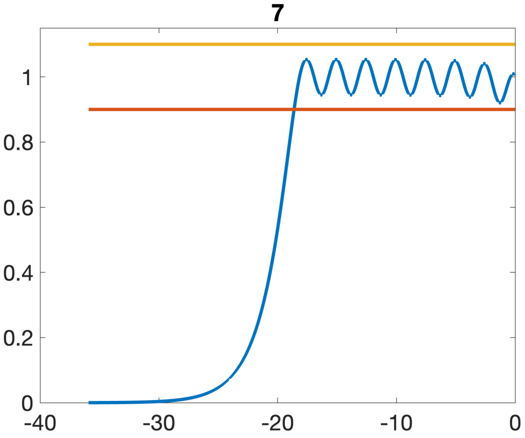


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 7$.

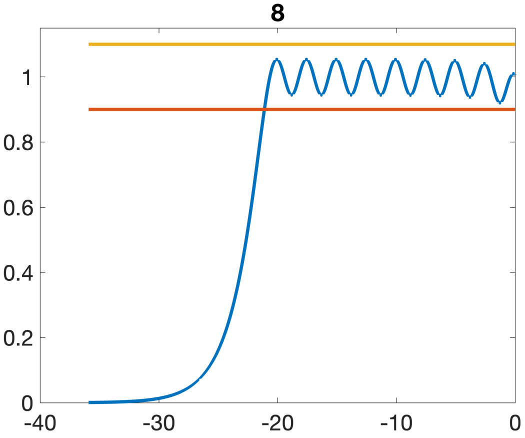


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 8$.

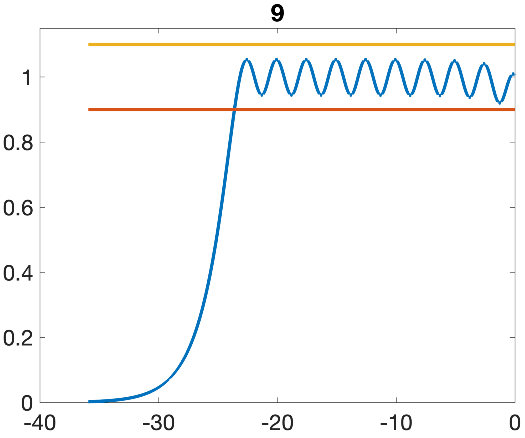


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 9$.

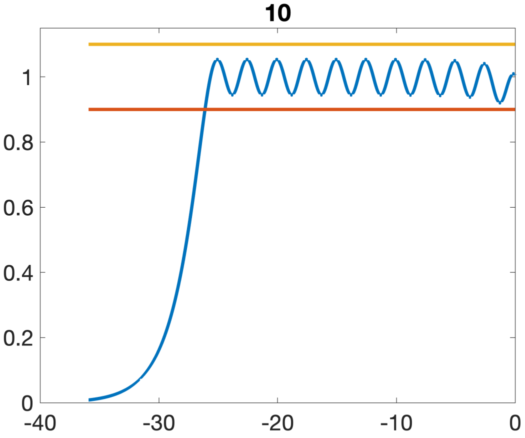


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 10$.

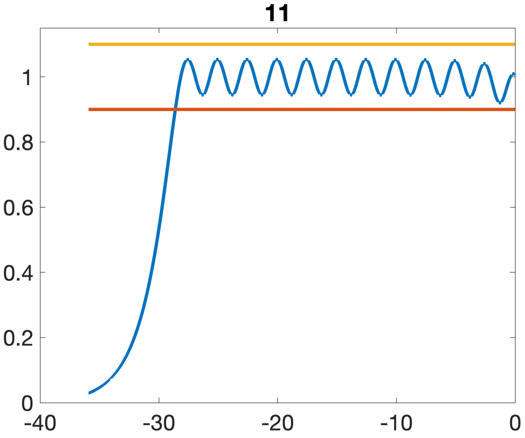


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 11$.

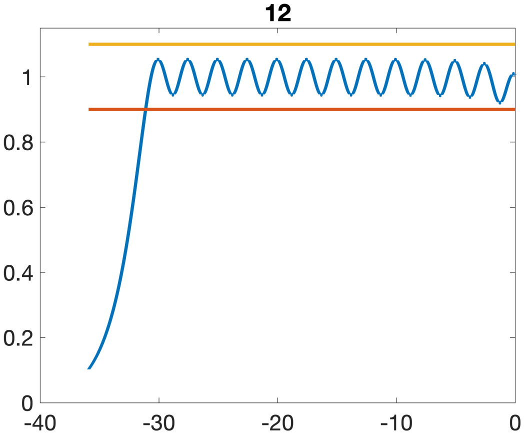


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 12$.

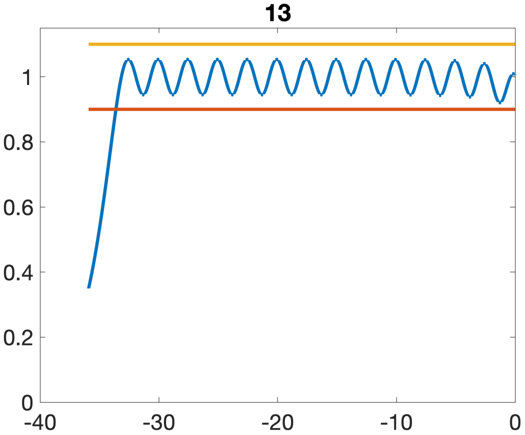


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 13$.

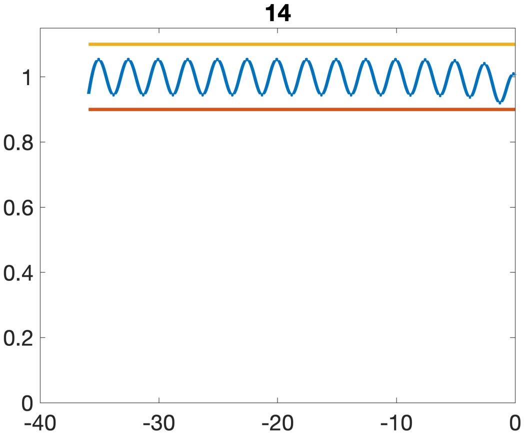


# Impression of the required parameters

Plots of

$$x \mapsto \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+x} + (kh+x)/2)$$

for $x = \ln(r)$ for $r = 10^{-16}$ to $r = 1$, $h = 2.5$, $k_1 = -2$, $k_2 = 14$.



# Eigenvalues for $\tilde{\mathbf{P}}_*$ for example discretizations

For $\epsilon = 1/10$, choose $h, k_1, k_2$ such that for all eigenvalues $\lambda = r$ of $\tilde{\mathbf{P}}_*$:

$$1 - \epsilon \leq \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+\ln(r)} + (kh + \ln(r))/2) \leq 1 + \epsilon.$$

**For dG method in angle, FEM in space** with uniform grid-sizes $\Delta z$, $\Delta \mu$

$$c\Delta z^2 \Delta \mu^2 / 3 \leq \lambda \leq 1.$$

If $\Delta z = \Delta \mu = 10^{-5}$, choose $h = 2.5$, $k_1 = -2$, $k_2 = 18$.

**Note:** Only $\ln \lambda$ enters.

# Eigenvalues for $\tilde{\mathbf{P}}_*$ for example discretizations

For $\epsilon = 1/10$, choose $h, k_1, k_2$ such that for all eigenvalues $\lambda = r$ of $\tilde{\mathbf{P}}_*$:

$$1 - \epsilon \leq \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \exp(-e^{kh+\ln(r)} + (kh + \ln(r))/2) \leq 1 + \epsilon.$$

**For dG method in angle, FEM in space** with uniform grid-sizes $\Delta z$, $\Delta\mu$

$$c\Delta z^2 \Delta\mu^2/3 \leq \lambda \leq 1.$$

If $\Delta z = \Delta\mu = 10^{-5}$, choose $h = 2.5$, $k_1 = -2$, $k_2 = 18$.

**For $P_N$ method in angle, FEM in space** with uniform grid-size $\Delta z$

$$c\Delta z^2/N^4 \leq \lambda \leq 1.$$

If $\Delta z = 10^{-5}$, $N = 100$, choose $h = 2.5$, $k_1 = -2$, $k_2 = 17$.

**Note:** Only $\ln \lambda$ enters.

# Summary: Preconditioner via exponential sums

$\tilde{\mathbf{P}}_* = \tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}$

**Exponential sum approximation:**

$$\tilde{\mathbf{P}}^{-1/2} := \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \rho_k \exp(-\alpha_k \tilde{\mathbf{T}}) \otimes \exp(-\alpha_k \tilde{\mathbf{M}})$$

with error bound (appropriately choosing $h$, $k_1$, $k_2$)

$$\|(\tilde{\mathbf{P}}^{-1/2} - \tilde{\mathbf{P}}_*^{-1/2})\mathbf{x}\|_2 \leq \epsilon \|\tilde{\mathbf{P}}_*^{-1/2}\mathbf{x}\|_2.$$

**Lemma.** For all $\mathbf{x} \in \mathbb{R}^{IJ}$ it holds that

$$\frac{\gamma_1}{1+\epsilon}\langle \tilde{\mathbf{P}}\mathbf{x}, \mathbf{x}\rangle \leq \langle \tilde{\mathbf{A}}\mathbf{x}, \mathbf{x}\rangle \leq \frac{\gamma_2}{1-\epsilon}\langle \tilde{\mathbf{P}}\mathbf{x}, \mathbf{x}\rangle.$$

# Summary: Preconditioner via exponential sums

$\tilde{\mathbf{P}}_* = \tilde{\mathbf{T}} \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{\mathbf{M}}$

**Exponential sum approximation:**

$$\tilde{\mathbf{P}}^{-1/2} := \frac{h}{\sqrt{\pi}} \sum_{k=k_1}^{k_2} \rho_k \exp(-\alpha_k \tilde{\mathbf{T}}) \otimes \exp(-\alpha_k \tilde{\mathbf{M}})$$

with error bound (appropriately choosing $h$, $k_1$, $k_2$)

$$\|(\tilde{\mathbf{P}}^{-1/2} - \tilde{\mathbf{P}}_*^{-1/2})\mathbf{x}\|_2 \leq \epsilon \|\tilde{\mathbf{P}}_*^{-1/2}\mathbf{x}\|_2.$$

**Theorem.** For any $\mathbf{w}^0$, the mapping $\mathbf{w}^n \mapsto \mathbf{w}^{n+1}$ defined by

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \tau(\tilde{\mathbf{P}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{P}}^{-1/2}\mathbf{w}^n - \tilde{\mathbf{f}})$$

converges for any $\tau \in (0, 2(1-\epsilon)/\gamma_2)$ to the solution $\mathbf{w}$ in the 2-norm.

# Conclusion: Rank-controlled iteration

**Rank controlled iteration**

$$\begin{aligned} \tilde{\mathbf{w}}^{n+1} &= \mathbf{w}^n - \tau(\tilde{\mathbf{P}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{P}}^{-1/2})\mathbf{w}^n - \tilde{\mathbf{f}}) \\ \mathbf{w}^{n+1} &= \mathbf{S}_{\delta_n}(\tilde{\mathbf{w}}^{n+1}), \end{aligned}$$

with soft-thresholding $\mathbf{S}_\delta$.

**Properties:**

- Converges linearly (independent of the grid)
- The iterates $\{\mathbf{w}^n\}$ have *quasi-optimal ranks*, assuming the singular values of the limit $\mathbf{w}$ decay algebraically or exponentially.
- Storage and operation count scales like $O\big(r(I+J)\big)$ and not $O(IJ)$ as for usual implementations.

---

[Bachmayr & Schneider (2017)], [Bachmayr & Bardin & S (2023)]