Scientific Computing WS 2020/2021

Slide lecture 2

Jürgen Fuhrmann

juergen.fuhrmann@wias-berlin.de

Elements of iterative methods (Saad Ch.4)

Let $V = \mathbb{R}^n$ be equipped with the inner product $(\cdot, \cdot)$. Let A be an $n \times n$ nonsingular matrix.

Solve $Au = b$ iteratively. For this purpose, two components are needed:

- **Preconditioner**: a matrix $M \approx A$ "approximating" the matrix A but with the property that the system $Mv = f$ is easy to solve
- **Iteration scheme**: algorithmic sequence using M and A which updates the solution step by step

Simple iteration with preconditioning

Assume we know the exact solution $\hat{u}$: $A\hat{u} = b$.

Then it must fulfill the identity

$$\hat{u} = \hat{u} - M^{-1}(A\hat{u} - b)$$

$\Rightarrow$ iterative scheme: put the "old" value on the right hand side and the "new" value on the left hand side:

$$u_{k+1} = u_k - M^{-1}(Au_k - b) \quad (k = 0, 1 \ldots)$$

Obviously, if $u_k = \hat{u}$, the process would be stationary.

Otherwise it leads to a sequence of approximations

$$u_0, u_1, \ldots, u_k, u_{k+1}, \ldots$$

Implementation of the iterative process

Aim: solve $Au = b$ with tolerance ε:

1. Choose initial value u_0, set $k = 0$
2. Calculate *residuum* $r_k = Au_k - b$
3. Test convergence: if $||r_k|| < \varepsilon$ set $u = u_k$, finish
4. Calculate *update*: solve $Mv_k = r_k$
5. Update solution: $u_{k+1} = u_k - v_k$, set $k = k + 1$, repeat with step 2.

The Jacobi method

- Let $A = D - E - F$, where D: main diagonal, E: negative lower triangular part F: negative upper triangular part
- Preconditioner: $M = D$, where D is the main diagonal of $A \Rightarrow$

$$u_{k+1,i} = u_{k,i} - \frac{1}{a_{ii}}\left(\sum_{j=1\dots n} a_{ij}u_{k,j} - b_i\right) \quad (i = 1\dots n)$$

- Equivalent to the succesive (row by row) solution of

$$a_{ii}u_{k+1,i} + \sum_{j=1\dots n, j\neq i} a_{ij}u_{k,j} = b_i \quad (i = 1\dots n)$$

- Already calculated results not taken into account
- Variable ordering does not matter

The Gauss-Seidel method

- Solve for main diagonal element row by row
- Take already calculated results into account
- Run in ascending order: forward GS

$$a_{ii}u_{k+1,i} + \sum_{j<i} a_{ij}u_{k+1,j} + \sum_{j>i} a_{ij}u_{k,j} = b_i \qquad (i = 1 \dots n)$$

$$(D-E)u_{k+1} - Fu_k = b$$

$$M = D - E$$

- Run in descending order: backward GS

$$a_{ii}u_{k+1,i} + \sum_{j>i} a_{ij}u_{k+1,j} + \sum_{j<i} a_{ij}u_{k,j} = b_i \qquad (i = n \dots 1)$$

$$(D-F)u_{k+1} - Eu_k = b$$

$$M = D - F$$

- May be it is faster
- Variable order probably matters

SOR and SSOR

- SOR: Successive overrelaxation: solve $\omega A = \omega B$ and use splitting

$$\omega A = (D - \omega E) - (\omega F + (1 - \omega D))$$
$$M = \frac{1}{\omega}(D - \omega E)$$

leading to

$$(D - \omega E)u_{k+1} = (\omega F + (1 - \omega D))u_k + \omega b$$

- SSOR: Symmetric successive overrelaxation

$$(D - \omega E)u_{k+\frac{1}{2}} = (\omega F + (1 - \omega D))u_k + \omega b$$
$$(D - \omega F)u_{k+1} = (\omega E + (1 - \omega D))u_{k+\frac{1}{2}} + \omega b$$

- Preconditioner:

$$M = \frac{1}{\omega(2 - \omega)}(D - \omega E)D^{-1}(D - \omega F)$$

- Gauss-Seidel and symmetric Gauss-Seidel are special cases for $\omega = 1$.

Block methods

- Jacobi, Gauss-Seidel, (S)SOR methods can as well be used block-wise, based on a partition of the system matrix into larger blocks,
- The blocks on the diagonal should be square matrices, and invertible
- Interesting variant for systems of partial differential equations, where multiple species interact with each other

Convergence

- Let $\hat{u}$ be the solution of $Au = b$.
- Let $e_k = u_k - \hat{u}$ be the error of the k-th iteration step

$$
\begin{aligned}
u_{k+1} &= u_k - M^{-1}(Au_k - b) \\
&= (I - M^{-1}A)u_k + M^{-1}b \\
u_{k+1} - \hat{u} &= u_k - \hat{u} - M^{-1}(Au_k - A\hat{u}) \\
&= (I - M^{-1}A)(u_k - \hat{u}) \\
&= (I - M^{-1}A)^k(u_0 - \hat{u})
\end{aligned}
$$

resulting in

$$e_{k+1} = (I - M^{-1}A)^k e_0$$

- So when does $(I - M^{-1}A)^k$ converge to zero for $k \to \infty$?
- Let $B = I - M^{-1}A$

Jordan canonical form of a matrix B

- λ_i $(i = 1 \dots p)$: eigenvalues of B
- $\sigma(B) = \{\lambda_1 \dots \lambda_p\}$: spectrum of B
- μ_i: algebraic multiplicity of λ_i:
  multiplicity as zero of the characteristic polynomial $\det(B - \lambda I)$
- γ_i geometric multiplicity of λ_i: dimension of $\mathrm{Ker}(B - \lambda I)$
- l_i: index of the eigenvalue: the smallest integer for which
  $\mathrm{Ker}(B - \lambda I)^{l_i+1} = \mathrm{Ker}(B - \lambda I)^{l_i}$
- $l_i \leq \mu_i$

Theorem (Saad, Th. 1.8) B can be transformed to a block diagonal matrix consisting of p diagonal blocks $D_1 \dots D_p$, each associated with a distinct eigenvalue λ_i.

- Each of the diagonal blocks D_i has itself a block diagonal structure consisting of γ_i *Jordan blocks* $J_{i,1} \dots J_{i,\gamma_i}$.
- Each of the Jordan blocks is an upper bidiagonal matrix of size not exceeding l_i with λ_i on the diagonal and 1 on the first upper diagonal.

Jordan canonical form of a matrix II

$$X^{-1}BX = J = \begin{pmatrix} D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_p \end{pmatrix}$$

$$D_i = \begin{pmatrix} J_{i,1} & & & \\ & J_{i,2} & & \\ & & \ddots & \\ & & & J_{i,\gamma_i} \end{pmatrix}$$

$$J_{i,k} = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & 1 & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}$$

Each $J_{i,k}$ is of size $\leq l_i$ and corresponds to a different eigenvector of B.

Definition The spectral radius $\rho(B)$ is the largest absolute value of any eigenvalue of B: $\rho(B) = \max_{\lambda \in \sigma(B)} |\lambda|$.

Theorem (Saad, Th. 1.10) $\lim_{k \to \infty} B^k = 0 \Leftrightarrow \rho(B) < 1$.

Proof, $\Rightarrow$: Let u_i be a unit eigenvector associated with an eigenvalue λ_i. Then

$$\begin{aligned}
Bu_i &= \lambda_i u_i \\
B^2 u_i &= \lambda_i B_i u_i = \lambda^2 u_i \\
&\vdots \\
B^k u_i &= \lambda^k u_i \\
\text{therefore} \quad ||B^k u_i||_2 &= |\lambda^k| \\
\text{and} \quad \lim_{k \to \infty} |\lambda^k| &= 0
\end{aligned}$$

so we must have $\rho(B) < 1$

Proof, $\Leftarrow$: Jordan form $X^{-1}BX = J$. Then $X^{-1}B^kX = J^k$.
Sufficient to regard Jordan block $J_i = \lambda I + E$ where $|\lambda| < 1$ and $E^{l_i} = 0$.
Let $k \geq l_i$. Then

$$J_i^k = \sum_{j=0}^{l_{i-1}} \binom{k}{j} \lambda^{k-j} E^j$$

$$||J_i||^k \leq \sum_{j=0}^{l_{i-1}} \binom{k}{j} |\lambda|^{k-j} ||E||^j$$

But $\binom{k}{j} = \frac{k!}{j!(k-j)!} = \sum_{i=0}^{j} \begin{bmatrix} j \\ i \end{bmatrix} \frac{k^i}{j!} = p_j(k)$ is a polynomial of degree j in k
where the Stirling numbers of the first kind are given by
$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = 1, \quad \begin{bmatrix} j \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ j \end{bmatrix} = 0, \quad \begin{bmatrix} j+1 \\ i \end{bmatrix} = j \begin{bmatrix} j \\ i \end{bmatrix} + \begin{bmatrix} j \\ i-1 \end{bmatrix}$.
Thus, $p_j(k)|\lambda|^{k-j} \to 0$ $(k \to \infty)$ as exponential decay beats polynomial growth
$\square$.

Corollary from proof

Theorem (Saad, Th. 1.12)

$$\lim_{k\to\infty} ||B^k||^{\frac{1}{k}} = \rho(B)$$

$\square$

Back to iterative methods

Sufficient condition for convergence: $\rho(I - M^{-1}A) < 1$.

Convergence rate

Assume λ with $|\lambda| = \rho(I - M^{-1}A) < 1$ is the largest eigenvalue and has a single Jordan block of size l. Then the convergence rate is dominated by this Jordan block, and therein by the term with the lowest possible power in λ which due to $E^l = 0$ is

$$\lambda^{k-l+1} \binom{k}{l-1} E^{l-1}$$

$$||(I - M^{-1}A)^k(u_0 - \hat{u})|| = O\left(|\lambda^{k-l+1}| \binom{k}{l-1}\right)$$

and the "worst case" convergence factor ρ equals the spectral radius:

$$\begin{aligned}
\rho &= \lim_{k\to\infty} \left(\max_{u_0} \frac{||(I - M^{-1}A)^k(u_0 - \hat{u})||}{||u_0 - \hat{u}||} \right)^{\frac{1}{k}} \\
&= \lim_{k\to\infty} ||(I - M^{-1}A)^k||^{\frac{1}{k}} \\
&= \rho(I - M^{-1}A)
\end{aligned}$$

Depending on u_0, the rate may be faster, though

Richardson iteration, sufficient criterion for convergence

Assume A has positive real eigenvalues $0 < \lambda_{min} \leq \lambda_i \leq \lambda_{max}$.
E.g. A is symmetric, positive definite (spd).

- Let $\alpha > 0$, $M = \frac{1}{\alpha} I \Rightarrow I - M^{-1}A = I - \alpha A$
- Then for the eigenvalues μ_i of $I - \alpha A$ one has:

$$
\begin{aligned}
1 - \alpha\lambda_{max} \leq \mu_i \leq 1 - \alpha\lambda_{min} \\
\mu_i < 1
\end{aligned}
$$

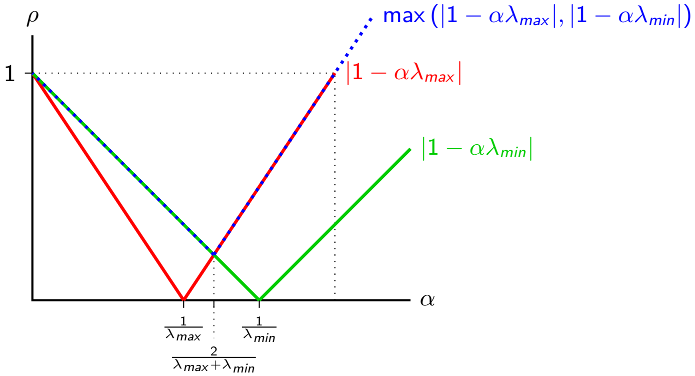
- We also need $1 - \alpha\lambda_{max} > -1$, so we must have $0 < \alpha < \frac{2}{\lambda_{max}}$.

Theorem. The Richardson iteration converges for any α with $0 < \alpha < \frac{2}{\lambda_{max}}$.

The convergence rate is $\rho = \max\left(|1 - \alpha\lambda_{max}|, |1 - \alpha\lambda_{min}|\right)$.

$\square$

Richardson iteration, choice of optimal parameter



- Due to $-(1-\alpha\lambda_{max}) > -(1-\alpha\lambda_{min})$ and $+(1-\alpha\lambda_{min}) > +(1-\alpha\lambda_{max})$,

$$\begin{aligned}\rho &= \max(|1-\alpha\lambda_{max}|, |1-\alpha\lambda_{min}|)\\ &= \max((1-\alpha\lambda_{max}), -(1-\alpha\lambda_{min}))\end{aligned}$$

- $1-\alpha\lambda_{max}$ is monotonically decreasing, the $-(1-\alpha\lambda_{min})$ increases, so the minimum must be at the intersection

$$1-\alpha\lambda_{max} = -1+\alpha\lambda_{min} \quad \Rightarrow \quad 2 = \alpha(\lambda_{max}+\lambda_{min})$$

Theorem. The optimal parameter is $\alpha_{opt} = \frac{2}{\lambda_{min}+\lambda_{max}}$.
For this parameter, the convergence factor is

$$\rho_{opt} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{\kappa - 1}{\kappa + 1}$$

where $\kappa = \kappa(A) = \frac{\lambda_{max}}{\lambda_{min}}$ is the spectral condition number of A. $\square$

Spectral equivalence

Theorem. M, A spd. Assume the *spectral equivalence estimate*

$$0 < \gamma_{min}(Mu, u) \leq (Au, u) \leq \gamma_{max}(Mu, u)$$

Then for the eigenvalues μ_i of $M^{-1}A$ we have

$$\gamma_{min} \leq \mu_{min} \leq \mu_i \leq \mu_{max} \leq \gamma_{max}$$

and $\kappa(M^{-1}A) \leq \frac{\gamma_{max}}{\gamma_{\text{min}}}$

Proof. Let the inner product $(\cdot, \cdot)_M$ be defined via $(u, v)_M = (Mu, v)$. In this inner product, $C = M^{-1}A$ is self-adjoint:

$$\begin{aligned}(Cu, v)_M &= (MM^{-1}Au, v) = (Au, v) = (M^{-1}Mu, Av) = (Mu, M^{-1}Av)\\ &= (u, M^{-1}A)_M = (u, Cv)_M\end{aligned}$$

Minimum and maximum eigenvalues can be obtained as Ritz values in the $(\cdot, \cdot)_M$ scalar product

$$\mu_{min} = \min_{u\neq 0} \frac{(Cu, u)_M}{(u, u)_M} = \min_{u\neq 0} \frac{(Au, u)}{(Mu, u)} \geq \gamma_{min}$$

$$\mu_{max} = \max_{u\neq 0} \frac{(Cu, u)_M}{(u, u)_M} = \max_{u\neq 0} \frac{(Au, u)}{(Mu, u)} \leq \gamma_{max}$$

Matrix preconditioned Richardson iteration

M, A spd.

- Scaled Richardson iteration with preconditoner M

$$u_{k+1} = u_k - \alpha M^{-1}(Au_k - b)$$

- Spectral equivalence estimate

$$0 < \gamma_{min}(Mu, u) \leq (Au, u) \leq \gamma_{max}(Mu, u)$$

- $\Rightarrow \gamma_{min} \leq \lambda_i \leq \gamma_{max}$
- $\Rightarrow$ optimal parameter $\alpha = \frac{2}{\gamma_{max}+\gamma_{min}}$
- Relative condition number estimate: $\kappa(M^{-1}A) \leq \frac{\gamma_{max}}{\gamma_{min}}$
- Convergence rate with optimal parameter: $\rho \leq \frac{\kappa(M^{-1}A)-1}{\kappa(M^{-1}A)+1}$

- Regard the $n \times n$ 1D heat conduction matrix with $h = \frac{1}{n-1}$ and $\alpha = \frac{1}{h}$ (easier to analyze).

$$A = \begin{pmatrix} \frac{2}{h} & -\frac{1}{h} & & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & & & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} \\ & & & & -\frac{1}{h} & \frac{2}{h} \end{pmatrix}$$

- Eigenvalues (tri-diagonal Toeplitz matrix):

$$\lambda_i = \frac{2}{h}\left(1 + \cos\left(\frac{i\pi}{n+1}\right)\right) \quad (i = 1 \dots n)$$

Source: A. Böttcher, S. Grudsky: Spectral Properties of Banded Toeplitz Matrices. SIAM,2005

- Express them in h: $n + 1 = \frac{1}{h} + 2 = \frac{1+2h}{h} \Rightarrow$

$$\lambda_i = \frac{2}{h}\left(1 + \cos\left(\frac{ih\pi}{1+2h}\right)\right) \quad (i = 1 \dots n)$$

1D heat conduction: spectral bounds estimate

- For $i = 1 \dots n$, the argument of cos is in $(0, \pi)$
- cos is monotonically decreasing in $(0, \pi)$, so we get λ_{max} for $i = 1$ and λ_{min} for $i = n = \frac{1+h}{h}$
- Therefore:

$$\lambda_{max} = \frac{2}{h}\left(1 + \cos\left(\pi \frac{h}{1+2h}\right)\right) \approx \frac{2}{h}\left(2 - \frac{\pi^2 h^2}{2(1+2h)^2}\right)$$

$$\lambda_{min} = \frac{2}{h}\left(1 + \cos\left(\pi \frac{1+h}{1+2h}\right)\right) \approx \frac{2}{h}\left(\frac{\pi^2 h^2}{2(1+2h)^2}\right)$$

Here, we used the Taylor expansion

$$cos(\delta) = 1 - \frac{\delta^2}{2} + O(\delta^4) \quad (\delta \to 0)$$

$$cos(\pi - \delta) = -1 + \frac{\delta^2}{2} + O(\delta^4) \quad (\delta \to 0)$$

and $\frac{1+h}{1+2h} = \frac{1+2h}{1+2h} - \frac{h}{1+2h} = 1 - \frac{h}{1+2h}$

Jacobi preconditioned Richardson for 1D heat conduction

- The Jacobi preconditioner just multiplies by $\frac{h}{2}$, therefore for $M^{-1}A$:

$$\mu_{max} \approx 2 - \frac{\pi^2 h^2}{2(1+2h)^2}$$
$$\mu_{min} \approx \frac{\pi^2 h^2}{2(1+2h)^2}$$

- Optimal parameter: $\alpha = \frac{2}{\lambda_{max}+\lambda_{min}} \approx 1 \ (h \to 0)$
- Good news: this is independent of h resp. n
- No need for spectral estimate in order to work with optimal parameter.
- Is this true beyond this special case ?

Jacobi for 1D heat conduction: convergence factor

- Condition number + spectral radius

$$\kappa(M^{-1}A) = \kappa(A) \approx \frac{4(1+2h)^2}{\pi^2 h^2} - 1$$
$$\rho(I - M^{-1}A) = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{\pi^2 h^2}{2(1+2h)^2}$$

- Bad news: $\rho \to 1 \quad (h \to 0)$
- Typical situation with second order PDEs:

$$\kappa(A) = O(h^{-2}) \quad (h \to 0)$$
$$\rho(I - D^{-1}A) = 1 - O(h^2) \quad (h \to 0)$$

- Mean square error of approximation $||u - u_h||_2 < h^\gamma$, in the simplest case $\gamma = 2$.

Estimating Iterative solver complexity I

- Solve linear system iteratively until $||e_k|| = ||(I - M^{-1}A)^k e_0|| \leq \epsilon$

$$\begin{aligned} \rho^k e_0 &\leq \epsilon \\ k \ln \rho &< \ln \epsilon - \ln e_0 \\ k \geq k_\rho &= \left\lceil \frac{\ln e_0 - \ln \epsilon}{\ln \rho} \right\rceil \end{aligned}$$

- $\Rightarrow$ we need at least k_ρ iteration steps to reach accuracy ϵ
- The ideal iterative solver:
  - $\rho(I - M^{-1}A) < \rho_0 < 1$ independent of h resp. N
    $\Rightarrow k_\rho$ independent of N.
  - A sparse $\Rightarrow$ matrix-vector multiplication Au has complexity $O(N)$
  - Solution of $Mv = r$ has complexity $O(N)$.

$\Rightarrow$ Number of iteration steps k_ρ independent of N
Each iteration step has complexity $O(N)$
$\Rightarrow$ Overall complexity $O(N)$

Estimating Iterative solver complexity II

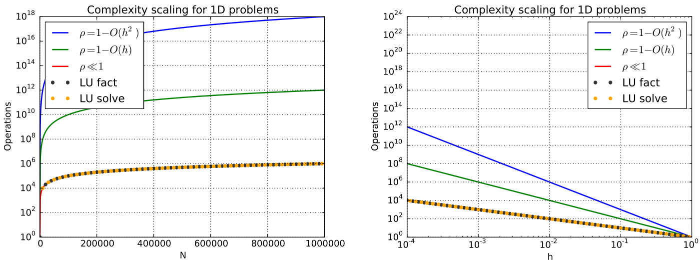
- Assume
  - $\rho = 1 - h^\delta \Rightarrow \ln \rho \approx -h^\delta \to k_\rho = O(h^{-\delta})$
  - d: space dimension: $N \approx n^d$, $h \approx \frac{1}{n} \approx N^{-\frac{1}{d}}$
    $\Rightarrow k_\rho = O(N^{\frac{\delta}{d}})$
  - $O(N)$ complexity of one iteration step (e.g. Jacobi, Gauss-Seidel)

  $\Rightarrow$ Overall complexity $O(N^{1+\frac{\delta}{d}})=O(N^{\frac{d+\delta}{d}})$
- Jacobi: $\delta = 2$ (Gauss-Seidel scales in a similar way)
- Hypothetical "Improved iterative solver" with $\delta = 1$?
- Overview on complexity estimates (SpLU: sparse LU)

Space dim.	$\delta = 2$ $\rho = 1 - O(h^2)$	$\delta = 1$ $\rho = 1 - O(h)$	SpLU fact.	SpLU solve
1	$O(N^3)$	$O(N^2)$	$O(N)$	$O(N)$
2	$O(N^2)$	$O(N^{\frac{3}{2}})$	$O(N^{\frac{3}{2}})$	$O(N \log N)$
3	$O(N^{\frac{5}{3}})$	$O(N^{\frac{4}{3}})$	$O(N^2)$	$O(N^{\frac{4}{3}})$
Tendency	$\downarrow$	$\downarrow$	$\uparrow\uparrow$	$\uparrow$

Solver complexity scaling for 1D problems

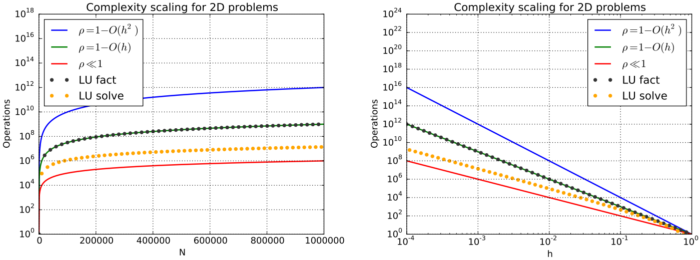
dim	$\rho = 1 - O(h^2)$	$\rho = 1 - O(h)$	LU fact.	LU solve
1	$O(N^3)$	$O(N^2)$	$O(N)$	$O(N)$



- Sparse direct solvers are asymptotically optimal
- Non-ideal iterative solvers significantly worse than optimal

Solver complexity scaling for 2D problems

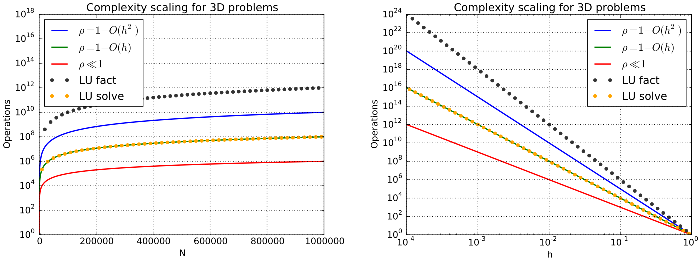
dim	$\rho = 1 - O(h^2)$	$\rho = 1 - O(h)$	LU fact.	LU solve
2	$O(N^2)$	$O(N^{\frac{3}{2}})$	$O(N^{\frac{3}{2}})$	$O(N \log N)$



- Sparse direct solvers better than simple nonideal iterative solvers ($\delta = 2$ Jacobi etc.)
- Sparse direct solvers on par with improved iterative solvers ($\delta = 1$)

Solver complexity scaling for 3D problems

dim	$\rho = 1 - O(h^2)$	$\rho = 1 - O(h)$	LU fact.	LU solve
3	$O(N^{\frac{5}{3}})$	$O(N^{\frac{4}{3}})$	$O(N^2)$	$O(N^{\frac{4}{3}})$



- Sparse LU factorization is expensive: going from h to $h/2$ increases N by a factor of 8 and operation count by a factor of 64!
- Sparse LU solve on par with improved iterative solvers

What could be done ?

- Holy grail: find ideal preconditioner with $\rho \leq \rho_0 < 1$ independent of h, N
- Find "improved preconditioner" with $\kappa(M^{-1}A) = O(h^{-1}) \Rightarrow \delta = 1$
- Find "improved iterative scheme" with $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$:

For Jacobi, we had $\kappa = X^2 - 1$ where $X = \frac{2(1+2h)}{\pi h} = O(h^{-1})$.

$$
\begin{aligned}
\rho &= 1 + \frac{\sqrt{X^2-1}-1}{\sqrt{X^2-1}+1} - 1 \\
&= 1 + \frac{\sqrt{X^2-1}-1-\sqrt{X^2-1}-1}{\sqrt{X^2-1}+1} \\
&= 1 - \frac{1}{\sqrt{X^2-1}+1} = 1 - \frac{1}{X\left(\sqrt{1-\frac{1}{X^2}}+\frac{1}{X}\right)} \\
&= 1 - O(h) \quad (h \to 0)
\end{aligned}
$$

$\Rightarrow$ Similar effect as wihth $\delta = 1$