Utrecht, May 1995

Maintaining the convergence properties of BiCGstab methods in finite precision arithmetic

by

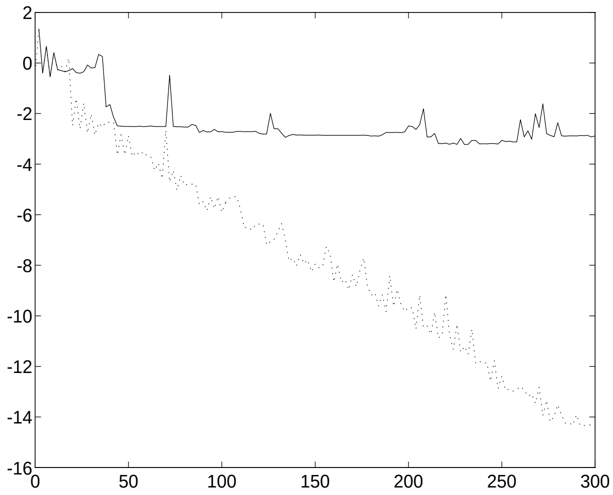
Gerard L.G. Sleijpen

sleijpen@math.ruu.nl

joint work with

Henk A. van der Vorst

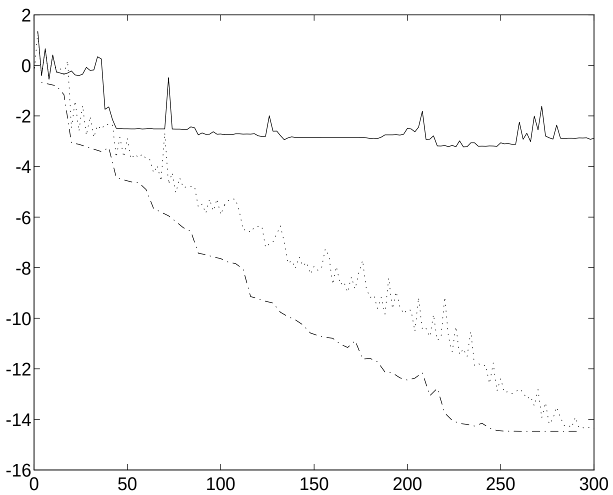
Department of Mathematics
University Utrecht
Utrecht, The Netherlands



$u_{xx} + u_{yy} + u_{zz} + 1000\, u_x = f.$

f is defined by the solution
$u(x,y,z) = \exp(xyz)\sin(\pi x)\sin(\pi y)\sin(\pi z).$

Discretized using $(10 \times 10 \times 10)$ volumes.
No preconditioning was used.



$u_{xx} + u_{yy} + u_{zz} + 1000\, u_x = f.$

f is defined by the solution
$u(x,y,z) = \exp(xyz)\sin(\pi x)\sin(\pi y)\sin(\pi z).$

Discretized using $(10 \times 10 \times 10)$ volumes.
No preconditioning was used.

Bi-CG for $Ax = b$

$$\text{Bi-CG} \left\{ \begin{array}{rcl} u_k &=& r_k - \beta_k u_{k-1} \\ r_{k+1} &=& r_k - \alpha_k A u_k, \end{array} \right.$$
$$x_{k+1} = x_k + \alpha_k u_k.$$

$r_k, u_k \in \mathcal{K}_{k+1}(A; r_0) \equiv \text{span}\,(r_0, Ar_0, \ldots, A^k r_0)$
$Au_k,\ r_k \perp \mathcal{K}_k(A^T; \tilde{r}_0)$

$$\beta_k = \frac{\theta_{k-1}}{\theta_k}\,\frac{\rho_k}{\sigma_{k-1}}, \qquad \alpha_k = \frac{\rho_k}{\sigma_k}$$

with

$$\begin{array}{rcl} \rho_k &\equiv& (r_k, \psi_k(A^T)\,\tilde{r}_0) \\ \sigma_k &\equiv& (Au_k, \psi_k(A^T)\,\tilde{r}_0) \end{array} \quad \& \quad \begin{array}{c} \psi_k \text{ pol. degree } k \\ \psi_k(t) = \theta_k\, t^k + \ldots \end{array}$$

Transpose-free Bi-CG

$\rho_k = (r_k, \psi_k(A^T)\,\tilde{r}_0) = (\psi_k(A) r_k, \tilde{r}_0),$
$\sigma_k = (A u_k, \psi_k(A^T)\,\tilde{r}_0) = (A\psi_k(A) u_k, \tilde{r}_0)$

$\Psi_k \equiv \psi_k(A)$

$$\text{(Bi-CG)} \quad \begin{cases} \rho_k, & \Psi_k u_k = \Psi_k r_k - \beta_k \Psi_k u_{k-1}, \\ \sigma_k, & \Psi_k r_{k+1} = \Psi_k r_k - \alpha_k A \Psi_k u_k, \end{cases}$$
$$\hat{\mathbf{x}} = \mathbf{x}_k + \alpha_k \Psi_k u_k$$

(Pol) Compute ψ_{k+1} of degree $k+1$
such that $\psi_{k+1}(0) = 1$.

Compute

$$\mathbf{u}_k \equiv \Psi_{k+1} u_k, \quad \mathbf{r}_{k+1} \equiv \Psi_{k+1} r_{k+1}, \quad \mathbf{x}_{k+1}$$

Hybrid Bi-CG

$$\boxed{\mathbf{r}_{k+1} = \Psi_{k+1} r_{k+1} = \psi_{k+1}(A)\phi_{k+1}(A) r_0}$$

$r_{k+1} = \phi_{k+1}(A) r_0$:
ϕ_{k+1} is the “Bi-CG polynomial”
of degree $k+1$.

CGS (Conjugate Gradients squared):

$$\psi_k = \phi_k, \qquad \mathbf{r}_k = \phi_k(A)\,\phi_k(A)\,r_0$$

Exploit recurrence relations for Bi-CG leading to a very efficient algorithm.

Bi-CGSTAB (Bi–CG stabilized):

$$\psi_{k+1}(t) = (1 - \omega_k t)\,\psi_k(t)$$

$$\begin{aligned} \Psi_{k+1} r_{k+1} &= (1 - \omega_k A)\Psi_k r_{k+1}, \\ \mathbf{x}_{k+1} &= \mathbf{x} + \omega_k \Psi_k r_{k+1} \\ \Psi_{k+1} u_k &= (1 - \omega_k A)\Psi_k u_k \end{aligned}$$

where

$$\boxed{\omega_k \quad \text{minimizes} \quad \|(1 - \omega A)\ \underbrace{\Psi_k r_{k+1}}_{\hat{\mathbf{r}}}\|}$$

$$\omega_k = \frac{(\hat{\mathbf{r}}, A\hat{\mathbf{r}})}{(A\hat{\mathbf{r}}, A\hat{\mathbf{r}})} \quad \text{and} \quad \frac{\theta_{k-1}}{\theta_k} = \frac{1}{-\omega_{k-1}}$$

Breakdown of the minimization

Exact arithmetic, $\omega_k = 0$:

— No reduction of the residual by
$\Psi_{k+1} r_{k+1} = (I - \omega_k A)\Psi_k r_{k+1}$.

— ψ_{k+1} is of degree k;
Bi-CG scalars cannot be computed:
breakdown of the incorporated Bi-CG.

Finite precision arithmetic, $\omega_k \approx 0$:

— Poor reduction of the residual by
$\Psi_{k+1} r_{k+1} = (I - \omega_k A)\Psi_k r_{k+1}$.

— Bi-CG scalars affected by rounding errors:
drop of speed of convergence.

$$\omega_k = \frac{(\,\widehat{\mathbf{r}}\,, A\,\widehat{\mathbf{r}}\,)}{(A\,\widehat{\mathbf{r}}\,, A\,\widehat{\mathbf{r}}\,)}$$

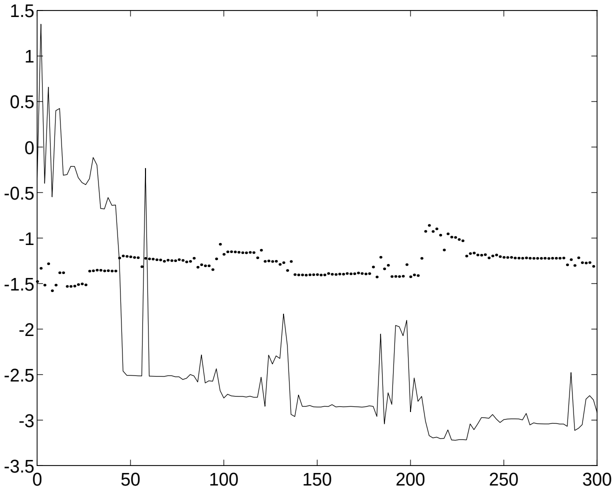
$$\omega_k = \frac{(\,\widehat{\mathbf{r}}\,, A\,\widehat{\mathbf{r}}\,)}{\|\,\widehat{\mathbf{r}}\,\|\,\|A\,\widehat{\mathbf{r}}\,\|}\frac{\|\,\widehat{\mathbf{r}}\,\|}{\|A\,\widehat{\mathbf{r}}\,\|} = \widehat{\omega}_k\frac{\|\,\widehat{\mathbf{r}}\,\|}{\|A\,\widehat{\mathbf{r}}\,\|}$$

$\omega_k \approx 0$ to be expected if

A has eigenvalues

with relatively large imaginary part

ω_k is real!

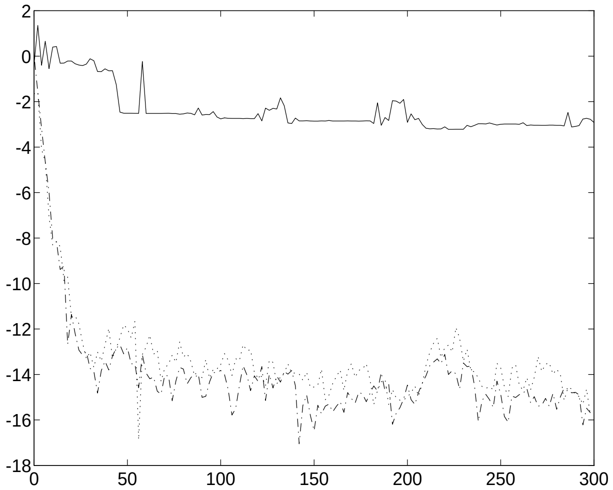


$$u_{xx} + u_{yy} + u_{zz} + 1000\,u_x = f.$$

$$\rho_k = (\mathbf{r}_k, \widetilde{r}_0), \qquad \rho_k^\star = \rho_k(1+\epsilon)$$

$$|\epsilon| \le n\,\overline{\xi}\,\frac{\|\mathbf{r}_k\|\,\|\widetilde{r}_0\|}{|(\mathbf{r}_k, \widetilde{r}_0)|} = \frac{n\,\overline{\xi}}{\widehat{\rho}_k}$$

where $\quad \widehat{\rho}_k \equiv \dfrac{|(\mathbf{r}_k, \widetilde{r}_0)|}{\|\mathbf{r}_k\|\,\|\widetilde{r}_0\|},$



$$u_{xx} + u_{yy} + u_{zz} + 1000\,u_x = f.$$

<u>Near breakdowns</u>

<u>and inaccurate Bi-CG coefficients</u>

- Is there a polynomial ψ_k for which

$$\rho_k: \quad \frac{|(\psi_k(A)\,r_k, \widetilde{r}_0)|}{\|\,\psi_k(A)\,r_k\|\,\|\widetilde{r}_0\|} \quad \text{is not very small,}$$

$$\sigma_k: \quad \frac{|(A\psi_k(A)\,u_k, \widetilde{r}_0)|}{\|\,\psi_k(A)\,u_k\|\,\|\widetilde{r}_0\|} \quad \text{is not very small?}$$

- If so, do we have these properties for the specific polynomial ψ_k used in the iterative method?

- If not, can the method efficiently be modified such that the resulting polynomial ψ_k has these properties?

For minimal rounding errors:

minimize

$$\frac{\|\,\psi_k(A)\,r_k\|\,\|\tilde r_0\|}{|(\psi_k(A)\,r_k,\tilde r_0)|} = \frac{\|\,\psi_k(A)\,r_k\|}{|\theta_k|}\cdot\frac{\|\tilde r_0\|}{|(r_k,(A^T)^k\,\tilde r_0)|}$$

$$\begin{aligned}(\psi_k(A)\,\mathbf{r}_k,\tilde r_0) &= (r_k,\psi_k(A^T)\,\tilde r_0)\\ &= (r_k,\theta_k\,(A^T)^k\,\tilde r_0)\end{aligned}$$

$$r_k\perp\mathcal{K}_k(A^T;\tilde r_0)$$

Solution

$$\left.\begin{array}{l}\psi_k(A)\,r_k\perp\mathcal{K}_k(A;r_k)\\ \psi_k(0)=1\end{array}\right\}\quad\begin{array}{l}\psi_k \text{ is the}\\ \text{FOM polynomial}\end{array}$$

For minimal rounding errors
using degree one factors:

minimize

$$\frac{\|\,\mathbf{r}_{k+1}\|\,\|\,\tilde r_0\,\|}{|(\mathbf{r}_{k+1},\tilde r_0)|} = \frac{\|\,\mathbf{r}_{k+1}\|}{|\,\omega\,|}\,\frac{\|\,\tilde r_0\,\|}{|(A\,\widehat{\mathbf{r}}\,,\tilde r_0)|}$$

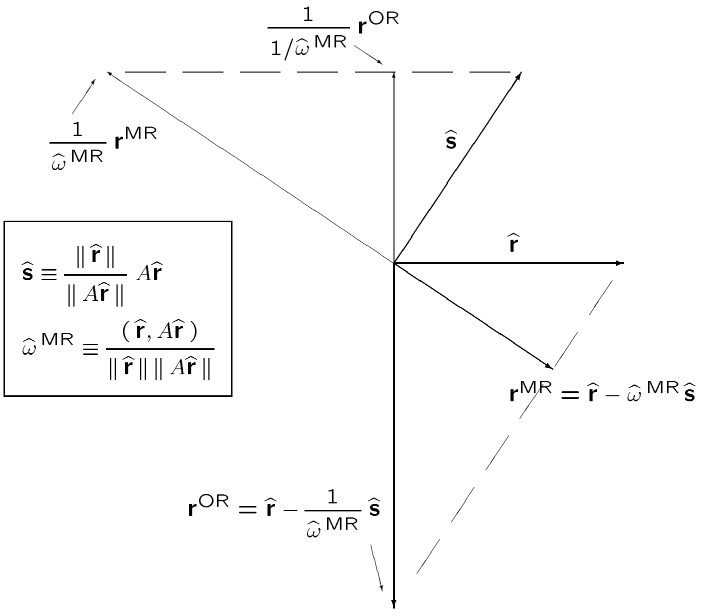
where $\quad\mathbf{r}_{k+1}=(I-\omega\,A)\,\widehat{\mathbf{r}}\,,\quad \widehat{\mathbf{r}}\equiv\Psi_k r_{k+1}$

$$\boxed{\text{minimize}\ \ \frac{\|\,\mathbf{r}_{k+1}\,\|}{|\,\omega\,|}=\frac{\|(I-\omega\,A)\,\widehat{\mathbf{r}}\,\|}{|\,\omega\,|}}$$

Here we used the fact that

$$\begin{aligned}(\mathbf{r}_{k+1},\tilde r_0) &= (r_{k+1},(I-\omega\,A^T)\,\psi_k(A^T)\,\tilde r_0)\\ &= -\omega\,(A\,\widehat{\mathbf{r}}\,,\tilde r_0)\end{aligned}$$

$$\mathbf{r}_{k+1}=\widehat{\mathbf{r}}-\omega\,A\,\widehat{\mathbf{r}}\,,\qquad \frac{\mathbf{r}_{k+1}}{\omega}=\frac{1}{\omega}\,\widehat{\mathbf{r}}-A\,\widehat{\mathbf{r}}$$



Updating by pols. of dg. $\ell\geq 2$

Consider $\quad k=m\ell.$

$$\psi_{m\ell+\ell}=p_m\psi_{m\ell}=p_m\cdot p_{m-1}\cdot\ldots\cdot p_0,$$

where p_m of exact degree ℓ, $\;p_m(0)=1\;$ and,
where, for minimal residuals,
$p_m\;$ minimizes $\;\|p_m(A)\,\widehat{\mathbf{r}}\,\|;\qquad \widehat{\mathbf{r}}\equiv\Psi_{m\ell}r_{m\ell+\ell}.$

BiCGstab(ℓ) [S,F,V] $\quad\ell\geq 2.$

With respect to the γ_i, minimize

$$\|p_m(A)\,\widehat{\mathbf{r}}\,\| \;=\; \|\,\widehat{\mathbf{r}}-\sum_{j=1}^{\ell}\gamma_j A^j\,\widehat{\mathbf{r}}\,\|,$$

where $\qquad p_m(t) \;=\; 1-\sum_{j=1}^{\ell}\gamma_j t^j$

For minimal rounding errors
using degree ℓ factors:

$$\text{maximize} \quad \widehat{\rho}_{k+\ell} \equiv \frac{|(\mathbf{r}_{k+\ell}, \widetilde{r}_0)|}{\|\mathbf{r}_{k+\ell}\|\,\|\widetilde{r}_0\|},$$

$$\text{where} \quad \mathbf{r}_{m\ell+\ell} = p_m(A)\,\widehat{\mathbf{r}}, \quad \widehat{\mathbf{r}} \equiv \Psi_k r_{m\ell+\ell}$$

or, equivalently,

$$\boxed{\text{minimize} \quad \bracevert \mathbf{r}_{k+\ell} \bracevert \equiv \frac{\|\,\mathbf{r}_{k+\ell}\,\|}{|\,\gamma_\ell\,|} = \frac{\|p_m(A)\,\widehat{\mathbf{r}}\,\|}{|\gamma_\ell|}}$$

Theorem

$\widehat{\mathbf{r}}_1, \ldots, \widehat{\mathbf{r}}_{\ell-1}$ orthonormal basis $\mathcal{K}_{\ell-1}(A; A\,\widehat{\mathbf{r}})$.

Define

$$\widetilde{\mathbf{r}}_0 \equiv \widehat{\mathbf{r}} - \sum_{j=1}^{\ell-1} (\,\widehat{\mathbf{r}}, \widehat{\mathbf{r}}_j)\,\widehat{\mathbf{r}}_j,$$

$$\widetilde{\mathbf{r}}_\ell \equiv A^\ell\,\widehat{\mathbf{r}} - \sum_{j=1}^{\ell-1} (A^\ell\,\widehat{\mathbf{r}}, \widehat{\mathbf{r}}_j)\,\widehat{\mathbf{r}}_j,$$

$$\widehat{\omega} \equiv \widehat{\omega}^{\mathrm{MR}} \equiv \frac{(\,\widetilde{\mathbf{r}}_\ell, \widetilde{\mathbf{r}}_0)}{\|\,\widetilde{\mathbf{r}}_\ell\|\,\|\,\widetilde{\mathbf{r}}_0\|}.$$

Then

$$\mathbf{r}^{\mathrm{MR}} = \widetilde{\mathbf{r}}_0 - \widehat{\omega}\frac{\|\,\widetilde{\mathbf{r}}_0\|}{\|\,\widetilde{\mathbf{r}}_\ell\|}\widetilde{\mathbf{r}}_\ell \quad \& \quad \mathbf{r}^{\mathrm{OR}} = \widetilde{\mathbf{r}}_0 - \frac{1}{\widehat{\omega}}\frac{\|\,\widetilde{\mathbf{r}}_0\|}{\|\,\widetilde{\mathbf{r}}_\ell\|}\widetilde{\mathbf{r}}_\ell,$$

$$\|\,\mathbf{r}^{\mathrm{MR}}\| = \sqrt{1-\widehat{\omega}^2}\,\|\,\widetilde{\mathbf{r}}_0\| \quad \& \quad \|\,\mathbf{r}^{\mathrm{OR}}\| = \frac{\sqrt{1-\widehat{\omega}^2}}{|\,\widehat{\omega}\,|}\|\,\widetilde{\mathbf{r}}_0\|,$$

$$\bracevert\,\mathbf{r}^{\mathrm{MR}}\bracevert = \frac{\sqrt{1-\widehat{\omega}^2}}{|\,\widehat{\omega}\,|}\|\,\widetilde{\mathbf{r}}_\ell\| \quad \& \quad \bracevert\,\mathbf{r}^{\mathrm{OR}}\bracevert = \sqrt{1-\widehat{\omega}^2}\,\|\,\widetilde{\mathbf{r}}_\ell\|.$$

Computational details

$\widehat{\mathbf{r}}_0, \widehat{\mathbf{r}}_1, \ldots, \widehat{\mathbf{r}}_\ell$ basis $\mathcal{K}_{\ell+1}(A; \widehat{\mathbf{r}})$

$$\mathbf{R} \equiv [\,\widehat{\mathbf{r}}_0\,|\,\ldots\,|\,\widehat{\mathbf{r}}_\ell],$$

$$\widetilde{\mathbf{r}}_0 = \mathbf{R}\,\vec{\eta}_0, \quad \widetilde{\mathbf{r}}_\ell = \mathbf{R}\,\vec{\eta}_\ell \quad \text{for some} \quad \vec{\eta}_0, \vec{\eta}_\ell \in \mathbf{R}^{\ell+1}.$$

With $\mathbf{V} \equiv \mathbf{R}^T\mathbf{R}$, define

$$<\vec{\eta}, \vec{\mu}> \equiv \vec{\mu}^T\mathbf{V}\vec{\eta} \quad \text{and} \quad |\,\vec{\eta}\,| \equiv \sqrt{<\vec{\eta}, \vec{\eta}>}.$$

Then

$$\boxed{\begin{aligned} &\mathbf{r}_{k+\ell} = \mathbf{R}\left(\vec{\eta}_0 - \widehat{\gamma}_k\frac{|\,\vec{\eta}_0|}{|\,\vec{\eta}_\ell|}\vec{\eta}_\ell\right) \\ &\text{with} \quad \widehat{\gamma}_k \equiv \frac{\widehat{\omega}}{|\widehat{\omega}|}\max(|\,\widehat{\omega}|, \Omega), \quad \widehat{\omega} = \frac{<\vec{\eta}_\ell\,,\,\vec{\eta}_0>}{|\,\vec{\eta}_\ell|\,|\,\vec{\eta}_0|} \end{aligned}}$$

Global effects

$$\psi_{k+\ell} = \psi_{m\ell+\ell} = p_m \cdot \psi_{m\ell} = p_m \cdot p_{m-1} \cdot \ldots \cdot p_0$$

$$\frac{\bracevert\,\psi_{k+\ell}(A) r_{k+\ell}\bracevert}{\bracevert\,\phi^{\mathrm{OR}}_{k+\ell}(A) r_{k+\ell}\bracevert} = \frac{1}{\varrho_m}\frac{\bracevert\,p_m^{\mathrm{OR}}(A)\,\widehat{\mathbf{r}}\,\bracevert}{\bracevert\,\phi^{\mathrm{OR}}_{k+\ell}(A) r_{k+\ell}\bracevert} \geq \frac{1}{\varrho_m}$$

$$\text{where} \quad \varrho_m \equiv \frac{\bracevert p_m^{\mathrm{OR}}(A)\,\widehat{\mathbf{r}}\,\bracevert}{\bracevert p_m(A)\,\widehat{\mathbf{r}}\,\bracevert}.$$

$$\boxed{\frac{\bracevert\,\psi_k(A)\,r_k\bracevert}{\bracevert\,\phi_k^{\mathrm{OR}}(A)\,r_k\bracevert} \gtrapprox \prod_{j<m}\frac{1}{\varrho_j}}$$

Why using pol. fact. of dg.≥ 2?

1 sweep BiCGstab(ℓ) versus ℓ steps Bi-CGSTAB:

- GMRES(ℓ) will reduce residuals better than $\ell\times$ GMRES(1)

- GMRES(ℓ) contributes only once to an increase of $\widehat{\rho}_k$

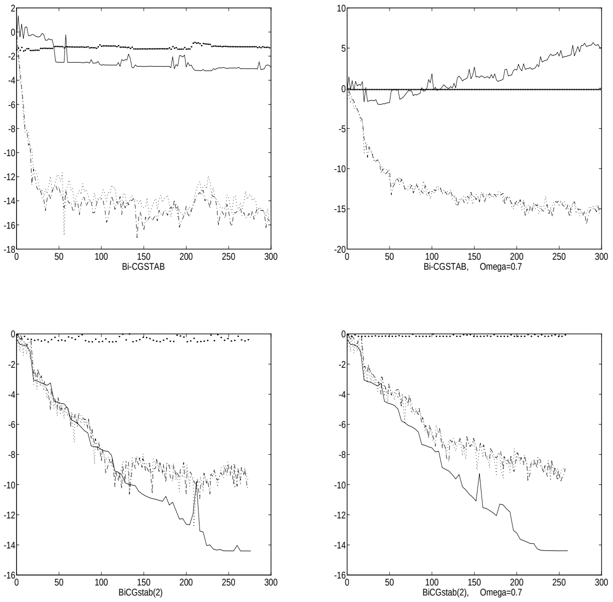
Why not?

- ∘ Efficiency:
  $2\ell+10$ AXPY/4 MV, $\ell+7$ DOT/4 MV
  Memory: $2\ell+5$ vector arrays

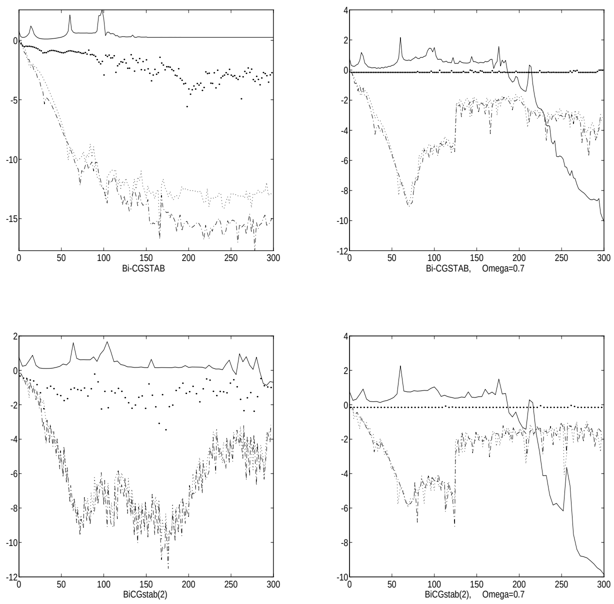
- ∘ Loss of accuracy:

$$\Big|\|\mathbf{r}_k\| - \|b - A\mathbf{x}_k\|\Big| \leq \overline{\xi}\, k\, n_A\ \big\|\,|A|\,\big\|\,\big\|A^{-1}\big\|\ \max\|\mathbf{r}\|$$

max. taken over all intermediate residuals $\mathbf{r}$



$u_{xx}+u_{yy}+u_{zz}+1000\,u_x = f$, (10×10×10) vol.



$$-u_{xx} - u_{yy} + \gamma(xu_x + yu_y) + \beta u = f$$

$\gamma = 100$ & $\beta = -200$, (63×63) vol.

$$-u_{xx} - u_{yy} + \gamma(xu_x + yu_y) + \beta u = f$$

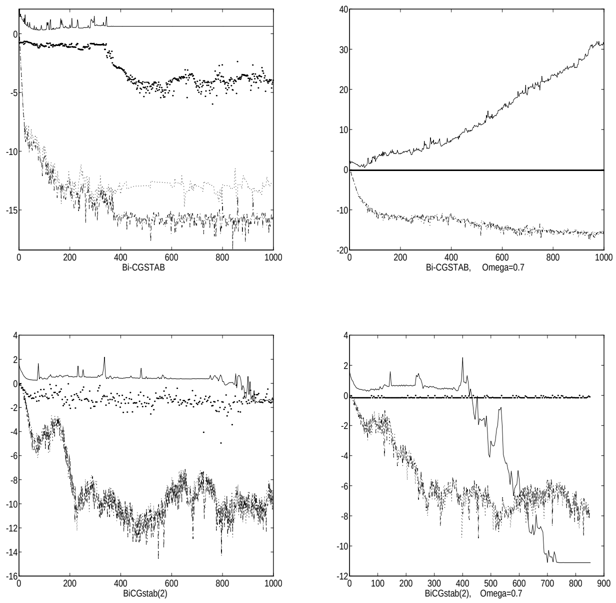
on the unit-square, Dirichlet boudary conditions,

with discretized solution $x = (1, 1, \ldots, 1)^T$,

$\gamma = 1000$ and $\beta = 10$.

Discretized using (66×66) volumes.

No preconditioning was used.



$$-u_{xx} - u_{yy} + \gamma(xu_x + yu_y) + \beta u = f$$

$\gamma = 1000$ & $\beta = 10$, (66×66) vol.