Accelerating Multigrid Optimization via SESOP

Tao Hong
Joint work with Irad Yavneh and Michael Zibulevsky

Computer Science Department
Technion - Israel Institute of Technology

Outline

Background

- MG/OPT Framework
- Sequential Subspace Optimization (SESOP)

Merge MG/OPT and SESOP

- SESOP-TG Scheme
- Convergence Factor Analysis with Optimized Stepsizes

Numerical Results

- The Roated Anisotropic Diffusion Problem - Linear
- p-Laplacian Problem - Nonlinear

MG/OPT Framework [Nas00]

Consider

$$\boldsymbol{x}_*^h = \arg \min_{\boldsymbol{x}^h \in \Re^N} \mathcal{F}^h(\boldsymbol{x}^h).$$

$\mathcal{F}^h(\cdot)$ is smooth $\Rightarrow$ "Relaxation" $\rightarrow$ Jacobi, Gauss-Seidel, Gradient Descent, Nesterov's Acceleration, LBFGS etc.

Consider ($N_c < N$)

$$\boldsymbol{x}_*^H = \arg \min_{\boldsymbol{x}^H \in \Re^{N_c}} \mathcal{F}^H(\boldsymbol{x}^H) - \boldsymbol{v}_k^{\mathcal{T}} \boldsymbol{x}^H, \text{ – Coarse Problem}$$

where $\boldsymbol{v}_k = \nabla \mathcal{F}^H(\boldsymbol{x}_k^H) - \boldsymbol{R} \nabla \mathcal{F}^h(\boldsymbol{x}_k^h)$.
$\boldsymbol{R} \in \Re^{N_c \times N}$ – Restriction & $\boldsymbol{x}_k^H = \boldsymbol{R}\boldsymbol{x}_k^h$

Define $\boldsymbol{P} \in \Re^{N \times N_c}$ – Prolongation

MG/OPT - two-level:

$$\boldsymbol{x}_0 \xrightarrow{\text{Relax.}} \boldsymbol{x}_k \xrightarrow{\text{CGC}} \boxed{\boldsymbol{x}_k = \boldsymbol{x}_k + \beta \boldsymbol{P}(\boldsymbol{x}_*^H - \boldsymbol{x}_k^H)} \xrightarrow{\text{Relax.}} \ldots$$

CGC: Coarse-Grid Correction

Multilevel: recursively

Sequential Subspace Optimization (SESOP) [Zib13]

Consider

$$\min_{\boldsymbol{x}^h \in \Re^N} \mathcal{F}^h(\boldsymbol{x}^h).$$

Formulate a subspace

$$\mathfrak{P}_k = \left[\boldsymbol{\Phi}\nabla\mathcal{F}^h(\boldsymbol{x}_k^h), \boldsymbol{\delta}_k, \boldsymbol{\delta}_{k-1}, \cdots, \boldsymbol{\delta}_{k-\Pi+1}\right], \ \ \Pi \geq 0.$$

$\boldsymbol{\Phi}$: Preconditioner & $\boldsymbol{\delta}_k = \boldsymbol{x}_k^h - \boldsymbol{x}_{k-1}^h$ & Π : Size of histories

SESOP:

$$\boldsymbol{x}_0 \xRightarrow{\mathfrak{P}_k} \boldsymbol{\alpha}_k = \arg\min_{\boldsymbol{\alpha}} \mathcal{F}^h(\boldsymbol{x}_k^h + \mathfrak{P}_k \boldsymbol{\alpha}) \Rightarrow \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mathfrak{P}_k \boldsymbol{\alpha}_k \xRightarrow{\mathfrak{P}_{k+1}} \ldots$$

Pros: General framework & Optimal convergence rate $\rightarrow \mathcal{O}(\frac{1}{k^2})$ &
Same as Conjugate-Gradient (CG) – Quadratic
Cons: May need high complexity $\rightarrow$ solving $\min_{\boldsymbol{\alpha}} \mathcal{F}^h(\boldsymbol{x}_k^h + \mathfrak{P}_k \boldsymbol{\alpha})$

SESOP-TG: Merge MG/OPT and SESOP [HYZ18]

Remind SESOP:

$$\mathfrak{P}_k = \left[\boldsymbol{\Phi} \nabla \mathcal{F}^h(\boldsymbol{x}_k^h), \boldsymbol{\delta}_k, \boldsymbol{\delta}_{k-1}, \cdots, \boldsymbol{\delta}_{k-\Pi+1} \right]$$

Our scheme – add CGC in $\mathfrak{P}_k$:

$$\tilde{\mathfrak{P}}_k = \left[\boldsymbol{\Phi} \nabla \mathcal{F}^h(\boldsymbol{x}_k^h), \boxed{\boldsymbol{P}(\boldsymbol{x}_*^H - \boldsymbol{x}_k^H)}, \boldsymbol{\delta}_k, \boldsymbol{\delta}_{k-1}, \cdots, \boldsymbol{\delta}_{k-\Pi+1} \right]$$

SESOP-TG-Π: TG means two-grid

$$\boldsymbol{x}_0 \xrightarrow{\text{CGC}\&\tilde{\mathfrak{P}}_k} \boldsymbol{\alpha}_k = \arg\min_{\boldsymbol{\alpha}} \mathcal{F}^h(\boldsymbol{x}_k^h + \tilde{\mathfrak{P}}_k \boldsymbol{\alpha}) \Rightarrow \boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \tilde{\mathfrak{P}}_k \boldsymbol{\alpha}_k \xrightarrow{\text{CGC}\&\tilde{\mathfrak{P}}_{k+1}} \ldots$$

Convergence Factor Analysis on Linear Problems

Consider

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} \frac{1}{2}\boldsymbol{x}^{\mathcal{T}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{f}^{\mathcal{T}}\boldsymbol{x}, \quad \boldsymbol{A} \succ 0$$

SESOP-TG-1:

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + c_1 \underbrace{\left(\boldsymbol{x}_{k-1} - \boldsymbol{x}_{k-2}\right)}_{\text{History}} + c_2 \underbrace{\boldsymbol{\Phi}\left(\boldsymbol{f} - \boldsymbol{A}\boldsymbol{x}_{k-1}\right)}_{\text{Pre. Gradient}} + c_3 \underbrace{\boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\left(\boldsymbol{f} - \boldsymbol{A}\boldsymbol{x}_{k-1}\right)}_{\text{CGC}}$$

$\boldsymbol{A}_H$: coarse-grid matrix approximating $\boldsymbol{A}$

Elliptic PDE: $\boldsymbol{A}_H$ rediscretization or Galerkin formula - $\boldsymbol{A}_H = \boldsymbol{RAP}$

Denote by $\boldsymbol{e}_k = \boldsymbol{x}^* - \boldsymbol{x}_k$. We have

$$\boldsymbol{e}_k = \boldsymbol{\Gamma}\boldsymbol{e}_{k-1} - c_1\boldsymbol{e}_{k-2},$$

where $\boldsymbol{\Gamma} = (1 + c_1)\boldsymbol{I} - \left(c_2\boldsymbol{\Phi} + c_3\boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\right)\boldsymbol{A}$.

Convergence Factor Analysis Continued

Remind

$$\boldsymbol{e}_k = \boldsymbol{\Gamma}\boldsymbol{e}_{k-1} - c_1\boldsymbol{e}_{k-2}.$$

Define $\boldsymbol{E}_k = \begin{bmatrix} \boldsymbol{e}_k \\ \boldsymbol{e}_{k-1} \end{bmatrix}$. We have

$$\boldsymbol{E}_k = \boldsymbol{\Upsilon}\boldsymbol{E}_{k-1},\ \boldsymbol{\Upsilon} \triangleq \begin{bmatrix} \boldsymbol{\Gamma} & -c_1\boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}$$

By a giving c_1, c_2, c_3, the asymptotic convergence factor r is

$$r = \rho(\boldsymbol{\Upsilon})$$

where $\rho(\cdot)$ the spectral radius operator.

Optimizing Fixed Stepsizes

c_1, c_2, c_3: subspace minimization & each iteration - SESOP.

Existing optimal fixed one?

The answer is *Positive*

But How?

$$r(c_1, c_2, c_3) = \min_{c_1, c_2, c_3} \rho(\Upsilon)$$

linear search ?

Let us see :-)

Optimizing Fixed Stepsizes Continued

Remind

$$\boldsymbol{e}_k = \boldsymbol{\Gamma}\boldsymbol{e}_{k-1} - c_1\boldsymbol{e}_{k-2},$$

where $\boldsymbol{\Gamma} = (1+c_1)\boldsymbol{I} - \left(c_2\boldsymbol{\Phi} + c_3\boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\right)\boldsymbol{A}$.
Define $\boldsymbol{W}_\alpha = \alpha\boldsymbol{\Phi}\boldsymbol{A} + (1-\alpha)\boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\boldsymbol{A}$ with $\alpha \in [0,1]$. Then

$$\boldsymbol{\Gamma} = (1+c_1)\boldsymbol{I} - c_{23}\boldsymbol{W}_\alpha$$

with $c_{23} = c_2 + c_3$.

Denote $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ the condition number of $\boldsymbol{W}_\alpha$.

Optimal Convergence Factor of SESOP-TG-1 [HYZ18]:

$$r_{opt} = \boxed{\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}}$$

by choosing $c_1 = \boxed{\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2}$ and $c_{23} = \boxed{\frac{4}{\lambda_{min}(\sqrt{\kappa}+1)^2}}$ with a given α.

Optimizing Fixed Stepsizes Continued

Remind

$$r_{opt} = \boxed{\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}} \text{ and } \kappa = cond(\boldsymbol{W}_\alpha)$$

where $\boldsymbol{W}_\alpha = \alpha \boldsymbol{\Phi} \boldsymbol{A} + (1-\alpha)\boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\boldsymbol{A}$ with $\alpha \in [0,1]$.

Remarks:

- $c_1 = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$ - ill-conditioned & using history is significant.
- $\alpha = 1$, retain Conjugate-Gradient (CG) rate
- $\alpha = 1$ & $c_1 = 0$, $r_{opt} = \frac{\kappa-1}{\kappa+1}$, retain gradient descent rate
- only need to find a bounded α for minimizing κ rather three $\rightarrow$ $\min_{c_1,c_2,c_3} \rho(\boldsymbol{\Upsilon})$ – relatively simple

Left:

Find α to minimize κ

Optimizing κ - Theoretic Insights

Assume $\boldsymbol{A}_H = \boldsymbol{RAP}$ (Galerkin form), $\boldsymbol{\Phi} = \boldsymbol{I}$, and the columns of $\boldsymbol{P}$ are a subset of the eigenectors of $\boldsymbol{A}$.
Denote by $\mathcal{R}(\boldsymbol{I}_H^h)$ the range of the prolongation and

$$\eta_{fmax} = \max_{i:\boldsymbol{w}_i \notin \mathcal{R}(\boldsymbol{P})} \eta_i, \qquad \eta_{fmin} = \min_{i:\boldsymbol{w}_i \notin \mathcal{R}(\boldsymbol{P})} \eta_i,$$
$$\eta_{cmax} = \max_{i:\boldsymbol{w}_i \in \mathcal{R}(\boldsymbol{P})} \eta_i, \qquad \eta_{cmin} = \min_{i:\boldsymbol{w}_i \in \mathcal{R}(\boldsymbol{P})} \eta_i.$$

where η_i and $\boldsymbol{w}_i$ are the eigenvalues and corresponding eigenvectors of $\boldsymbol{A}$.
We have

$$\alpha_{opt} = \boxed{\frac{1}{1+\eta_{fmin}-\eta_{cmin}} \leq 1,}$$

$$\kappa_{opt} = \boxed{\begin{cases} \frac{\eta_{fmax}}{\eta_{fmin}} & \text{if } \eta_{fmax} - \eta_{fmin} \geq \eta_{cmax} - \eta_{cmin}, \\ 1 + \frac{\eta_{cmax}-\eta_{cmin}}{\eta_{fmin}} & \text{otherwise.} \end{cases}}$$

Remark: $1 + \frac{\eta_{cmax}-\eta_{cmin}}{\eta_{fmin}} < \frac{\eta_{fmax}}{\eta_{fmin}} + 1 < 2$ & $\kappa_{opt} = \frac{\eta_{fmax}}{\eta_{fmin}}$ - ill-conditioned

Optimizing κ - In Practice

It is challenge for a general $\boldsymbol{A}$.

But if $\boldsymbol{A}$ is formulated from an elliptic partial differential equation (PDE) with constant coefficients, we can optimize κ in practice.

Strategy I: Local Fourier Analysis

Example: two dimensional & two-grid analysis

Denote:
$T^{\text{low}} : \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)^2$ & L_h the elliptic operator & $\tilde{L}_h(\theta_1, \theta_2)$ the symbol of L_h
Remind: $\boldsymbol{W}_\alpha = \alpha\boldsymbol{A} + (1-\alpha)\boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\boldsymbol{A}$ (extend to $\boldsymbol{\Phi} \neq \boldsymbol{I}$ obviously)
The eigenvalues of $\boldsymbol{W}_\alpha \Leftrightarrow 4 \times 4$, $\tilde{\boldsymbol{W}}_\alpha^{\theta_1,\theta_2}$ over the whole $(\theta_1, \theta_2) \in T^{low}$

$$\tilde{\boldsymbol{W}}_\alpha^{\theta_1,\theta_2} = \alpha\tilde{\boldsymbol{A}}^{\theta_1,\theta_2} + (1-\alpha)\tilde{\boldsymbol{P}}^{\theta_1,\theta_2}\left(\tilde{\boldsymbol{A}}_H^{\theta_1,\theta_2}\right)^{-1}\tilde{\boldsymbol{R}}^{\theta_1,\theta_2}\tilde{\boldsymbol{A}}^{\theta_1,\theta_2}$$

Optimizing κ - In Practice Continued

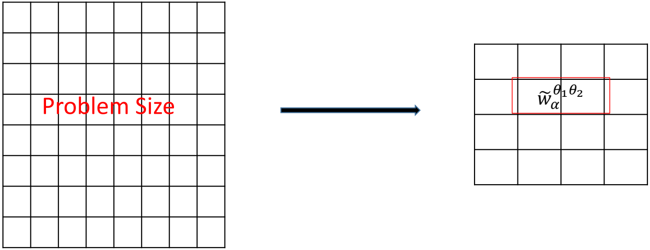
Strategy I and Example continued:

$$
\begin{aligned}
\tilde{\boldsymbol{W}}_{\alpha}^{\theta_1,\theta_2} &= \alpha \tilde{\boldsymbol{A}}^{\theta_1,\theta_2} + (1-\alpha)\tilde{\boldsymbol{P}}^{\theta_1,\theta_2}\left(\tilde{\boldsymbol{A}}_H^{\theta_1,\theta_2}\right)^{-1}\tilde{\boldsymbol{R}}^{\theta_1,\theta_2}\tilde{\boldsymbol{A}}^{\theta_1,\theta_2} \\
\tilde{\boldsymbol{A}}^{\theta_1,\theta_2} &= \begin{bmatrix} \tilde{L}_h(\theta_1,\theta_2) & & & \\ & \tilde{L}_h(\bar{\theta}_1,\theta_2) & & \\ & & \tilde{L}_h(\theta_1,\bar{\theta}_2) & \\ & & & \tilde{L}_h(\bar{\theta}_1,\bar{\theta}_2) \end{bmatrix} \\
\tilde{\boldsymbol{A}}_H^{\theta_1,\theta_2} &= \tfrac{1}{4}\tilde{L}_h(2\theta_1,2\theta_2) - \text{rediscretization} \\
&\text{or} \\
\tilde{\boldsymbol{A}}_H^{\theta_1,\theta_2} &= \tilde{\boldsymbol{R}}^{\theta_1,\theta_2}\tilde{L}_h(2\theta_1,2\theta_2)\tilde{\boldsymbol{P}}^{\theta_1,\theta_2} - \text{Galerkin form} \\
\bar{\theta}_i &= \begin{cases} \theta_i + \pi, & \text{if } \theta_i < 0 \\ \theta_i - \pi, & \text{if } \theta_i > 0 \end{cases}, \quad i = 1,2
\end{aligned}
$$

where $\tilde{\boldsymbol{R}}^{\theta_1,\theta_2} \in \Re^{4\times 1}$ and $\tilde{\boldsymbol{P}}^{\theta_1,\theta_2} \in \Re^{1\times 4}$ denote the symbols of $\boldsymbol{R}$ and $\boldsymbol{P}$, respectively.

Optimizing κ - In Practice Continued

Strategy II: Evaluate on a small size of grids - deterioration



Result: Evaluating the eigenvalues of $\boldsymbol{W}_{\alpha}$ becomes easily

Linear search $\Rightarrow \min_{\alpha \in [0,1]} cond(\boldsymbol{W}_{\alpha}) \Rightarrow$ e.g., MATLAB “fminbnd”

What Is Left?

- Two-level $\Rightarrow$ Multilevel : recursively
- The connection with h-ellipticity measure

$$E_h(L_h) := \frac{\min\{|\tilde{L}_h(\boldsymbol{\theta})| : \boldsymbol{\theta} \in T^{\text{high}}\}}{\max\{|\tilde{L}_h(\boldsymbol{\theta})| : \boldsymbol{\theta} \in T^{\text{high}}\}}$$

  where $T^{high} : [-\pi, \pi)^2 \setminus \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)^2$.
  ill-conditioned: $\kappa_{opt} = \frac{1}{E_h} \Rightarrow r_{opt} = \frac{1-\sqrt{E_h}}{1+\sqrt{E_h}}$ - Ideal One

  Remind: Theoretic Insights
- Find the details $\Rightarrow$ our paper [HYZ18]

The Roated Anisotropic Diffusion Problem - Linear

Problem description:

$$\mathcal{L}u = f$$

where

$$\mathcal{L}u = (C^2 + \varepsilon S^2)u_{xx} + 2(1-\varepsilon)CSu_{xy} + (\varepsilon C^2 + S^2)u_{yy}$$

with $C = \cos\phi$ and $S = \sin\phi$.

Discretization:

$$\mathcal{L}^h = \frac{1}{h^2}\begin{bmatrix} -\frac{1}{2}(1-\varepsilon)CS & \varepsilon C^2 + S^2 & \frac{1}{2}(1-\varepsilon)CS \\ C^2 + \varepsilon S^2 & -2(1+\varepsilon) & C^2 + \varepsilon S^2 \\ \frac{1}{2}(1-\varepsilon)CS & \varepsilon C^2 + S^2 & -\frac{1}{2}(1-\varepsilon)CS \end{bmatrix}$$

Coarse problem: rediscretization

Bilinear and Full-weighting

Linear Continued - Stepsizes & Subspace Minimization

Fine 64×64 grids & Dirichlet Boundary Condition

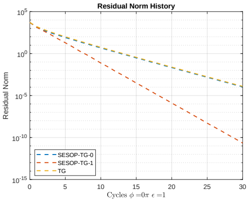
TG: Jacobi with optimally damped factor

Residual Norm:

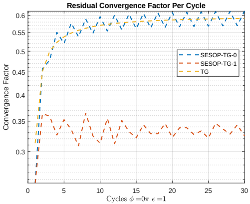
$$\|\mathcal{L}^h \boldsymbol{u}_k^h - f^h\|_F$$

Convergence Factor:

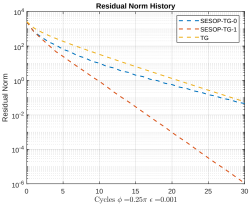
$$\frac{\|\mathcal{L}^h \boldsymbol{u}_k^h - f^h\|_F}{\|\mathcal{L}^h \boldsymbol{u}_{k-1}^h - f^h\|_F}$$



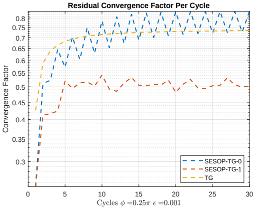
(a) $\varepsilon = 1, \phi = 0$



(b) $\varepsilon = 1, \phi = 0$



(c) $\varepsilon = 10^{-3}, \phi = \frac{\pi}{4}$



(d) $\varepsilon = 10^{-3}, \phi = \frac{\pi}{4}$

Linear Problem Continued - SESOP Vs Fixed Stepsizes

Fine 64×64 & Periodic Boundary Condition

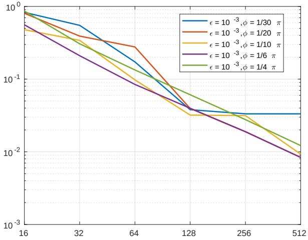
The comparison of convergence factor versus diff. methods
SESOP - Geometric average of the last 10 iterations

ϕ	ε	Bilinear		Bicubic		Ideal One
		SESOP	Fixed	SESOP	Fixed	
0	1	0.332	0.332	0.333	0.331	0.333
$\frac{\pi}{6}$	10^{-3}	0.570	0.563	0.537	0.532	0.587
$\frac{\pi}{6}$	10^{-4}	0.572	0.565	0.538	0.533	0.588
$\frac{\pi}{4}$	10^{-3}	0.509	0.500	0.457	0.443	0.446
$\frac{\pi}{4}$	10^{-4}	0.511	0.502	0.458	0.445	0.446

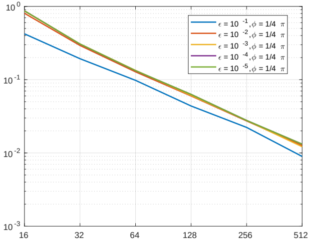
Linear Problem Continued - Deterioration of Strategy II

Denote

$$r_{ratio}(Num) \triangleq \frac{\log r_{1024}^{opt}}{\log r_{Num}^{opt}} - 1$$



(e) Various ϕ



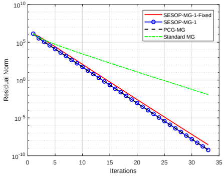
(f) Various ε

Result: working on 128×128 but solving 1024×1024 & less 10% additional computation - benefit if work on a huge problem

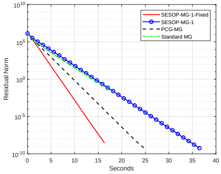
Linear Problem Continued - Multilevel Results

fine 1024×1024 & determine $64 \times 64 - 1.5$ seconds & Dirichlet
W-cycle, 2 pre- and 1 postrelaxation only coarse levels

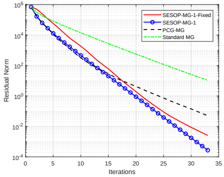
- SESOP-MG-1-Fixed: fixed stepsizes
- SESOP-MG-1: subspace minimization
- Standard MG: Jacobi relaxation with optimally damped factor + Coarse-Grid Correction
- PCG-MG: Preconditioned CG with standard MG as the preconditioner



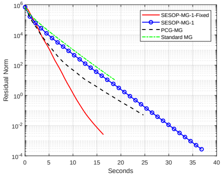
(g) $\varepsilon = 1, \phi = 0$



(h) $\varepsilon = 1, \phi = 0$



(i) $\varepsilon = 10^{-3}, \phi = \frac{\pi}{4}$



(j) $\varepsilon = 10^{-3}, \phi = \frac{\pi}{4}$

p-Laplacian Problem - Nonlinear

Problem description:

$$\begin{cases} \min_u \mathcal{F}\left(u(x,y)\right) = \int_\Omega \|\nabla u(x,y) + \xi\|^p - f(x,y)u(x,y)dxdy \\ \quad \text{such that} \quad u = 0 \quad \text{on} \quad \partial\Omega, \end{cases}$$

where $p \in (1,2)$.

PDE form:

$$\begin{cases} -\nabla \cdot \left(\|\nabla u + \xi\|^{p-2} \nabla u \right) = f \;\; \text{in} \;\; \Omega \\ \qquad u = 0 \qquad\qquad\qquad\qquad \text{on} \;\; \partial\Omega. \end{cases}$$

$\xi > 0$ regularization & avoid a trivial value in the denominator part.

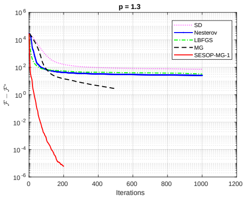
Coarse problem: rediscretization

Bilinear and Full-weighting

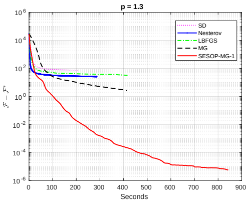
Nonlinear Problem Continued

Fine 1024 $\times$ 1024 & gradient descent as relaxation – SESOP-MG-1 and MG

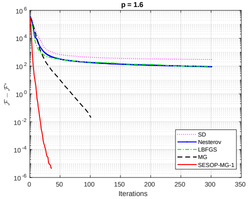
Newton as subspace minimization and BFGS for the coarsest level 9 $\times$ 9



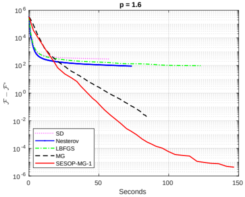
(k) $p = 1.3$



(l) $p = 1.3$



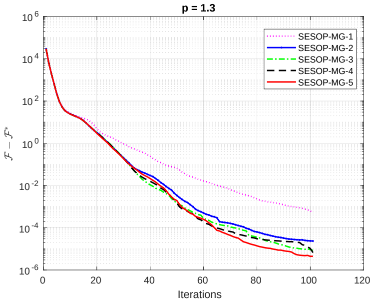
(m) $p = 1.6$



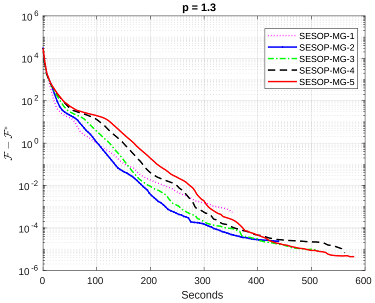
(n) $p = 1.6$

Nonlinear Problem Continued - History

Fine 1024×1024



(o) $p = 1.3$



(p) $p = 1.3$

More experiments and the detail of our analyses $\Rightarrow$ our paper [HYZ18]

Thanks & Questions?

Tao Hong, Irad Yavneh, and Michael Zibulevsky.
Accelerating multigrid optimization via sesop.
arXiv preprint arXiv:1812.06896, 2018.

Stephen G Nash.
A multigrid approach to discretized optimization problems.
Optimization Methods and Software, 14(1-2):99–116, 2000.

Michael Zibulevsky.
Speeding-up convergence via sequential subspace optimization: Current state and future directions.
arXiv preprint arXiv:1401.0159, 2013.