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 Rotated Anisotropic Diffusion Problem - Linear
  $p$-Laplacian Problem - Nonlinear
MG/OPT Framework [Nas00]

Consider
\[ x^h_\ast = \arg \min_{x^h \in \mathbb{R}^N} \mathcal{F}^h(x^h). \]

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

Consider \( (N_c < N) \)
\[ x^{H\ast} = \arg \min_{x^H \in \mathbb{R}^{N_c}} \mathcal{F}^H(x^H) - v_k^T x^H, \quad – \text{Coarse Problem} \]

where \( v_k = \nabla \mathcal{F}^H(x^H_k) - R \nabla \mathcal{F}^h(x^h_k). \)
\( R \in \mathbb{R}^{N_c \times N} – \text{Restriction} \quad \& \quad x^H_k = Rx^h_k \)

Define \( P \in \mathbb{R}^{N \times N_c} – \text{Prolongation} \)
MG/OPT - two-level:
\[ x_0 \xrightarrow{\text{Relax.}} x_k \xrightarrow{\text{CGC}} x_k = x_k + \beta P(x^H_\ast - x^H_k) \xrightarrow{\text{Relax.}} \ldots \]

CGC: Coarse-Grid Correction
Multilevel: recursively
Sequential Subspace Optimization (SESOP) [Zib13]

Consider

\[
\min_{x^h \in \mathbb{R}^N} F^h(x^h).
\]

Formulate a subspace

\[
\mathcal{P}_k = [\Phi \nabla F^h(x_k^h), \delta_k, \delta_{k-1}, \ldots, \delta_{k-\Pi+1}], \quad \Pi \geq 0.
\]

\(\Phi\): Preconditioner \& \(\delta_k = x_k^h - x_{k-1}^h\) \& \(\Pi\): Size of histories

SESOP:

\[
x_0 \xrightarrow{\mathcal{P}_k} \alpha_k = \arg\min_{\alpha} F^h(x_k^h + \mathcal{P}_k \alpha) \Rightarrow x_{k+1} = x_k + \mathcal{P}_k \alpha_k \xrightarrow{\mathcal{P}_{k+1}} \ldots
\]

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

\[ \mathcal{P}_k = [\Phi \nabla F^h(x^h_k), \delta_k, \delta_{k-1}, \ldots, \delta_{k-\Pi+1}] \]

Our scheme – add CGC in \( \mathcal{P}_k \):

\[ \tilde{\mathcal{P}}_k = [\Phi \nabla F^h(x^h_k), \underline{P}(x^H_\ast - x^H_k), \delta_k, \delta_{k-1}, \ldots, \delta_{k-\Pi+1}] \]

SESOP-TG-Π: TG means two-grid

\[
\begin{align*}
\alpha_k & = \text{arg min}_{\alpha} F^h(x^h_k + \tilde{\mathcal{P}}_k \alpha) \\
\Rightarrow x_{k+1} & = x_k + \tilde{\mathcal{P}}_k \alpha_k \\
\end{align*}
\]

\[ x_0 \xrightarrow{\text{CGC}& \tilde{\mathcal{P}}_k} \alpha_k = \text{arg min}_{\alpha} F^h(x^h_k + \tilde{\mathcal{P}}_k \alpha) \Rightarrow x_{k+1} = x_k + \tilde{\mathcal{P}}_k \alpha_k \xrightarrow{\text{CGC}& \tilde{\mathcal{P}}_{k+1}} \ldots \]
Convergence Factor Analysis on Linear Problems

Consider

\[ x^* = \arg \min_x \frac{1}{2} x^T Ax - f^T x, \quad A \succ 0 \]

SESOP-TG-1:

\[ x_k = x_{k-1} + c_1 (x_{k-1} - x_{k-2}) + c_2 \Phi (f - Ax_{k-1}) + c_3 PA_H^{-1} R (f - Ax_{k-1}) \]

\( A_H \): coarse-grid matrix approximating \( A \)

Elliptic PDE: \( A_H \) rediscretization or Galerkin formula - \( A_H = RAP \)

Denote by \( e_k = x^* - x_k \). We have

\[ e_k = \Gamma e_{k-1} - c_1 e_{k-2}, \]

where \( \Gamma = (1 + c_1) I - (c_2 \Phi + c_3 PA_H^{-1} R) A \).
Convergence Factor Analysis Continued

Remind

\[ e_k = \Gamma e_{k-1} - c_1 e_{k-2}. \]

Define \( E_k = \begin{bmatrix} e_k \\ e_{k-1} \end{bmatrix} \). We have

\[ E_k = \Upsilon E_{k-1}, \quad \Upsilon \triangleq \begin{bmatrix} \Gamma & -c_1 I \\ I & 0 \end{bmatrix} \]

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

\[ r = \rho(\Upsilon) \]

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?

\[
 r(c_1, c_2, c_3) = \min_{c_1, c_2, c_3} \rho(\mathcal{Y})
\]

linear search?

Let us see :-(
Remind

\[ e_k = \Gamma e_{k-1} - c_1 e_{k-2}, \]

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

Define \( W_\alpha = \alpha \Phi A + (1 - \alpha) PA_H^{-1} RA \) with \( \alpha \in [0, 1] \). Then

\[ \Gamma = (1 + c_1)I - c_{23} W_\alpha \]

with \( c_{23} = c_2 + c_3 \).

Denote \( \kappa = \frac{\lambda_{max}}{\lambda_{min}} \) the condition number of \( W_\alpha \).

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

\[ r_{opt} = \frac{\sqrt{\kappa - 1}}{\sqrt{\kappa + 1}} \]

by choosing \( c_1 = \left( \frac{\sqrt{\kappa - 1}}{\sqrt{\kappa + 1}} \right)^2 \) and \( c_{23} = \frac{4}{\lambda_{min}(\sqrt{\kappa + 1})^2} \) with a given \( \alpha \).
Optimizing Fixed Stepsizes Continued

Remind

\[ r_{opt} = \frac{\sqrt{\kappa - 1}}{\sqrt{\kappa + 1}} \text{ and } \kappa = \text{cond}(W_\alpha) \]

where \( W_\alpha = \alpha \Phi A + (1 - \alpha)PA_H^{-1}RA \) 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 \( \alpha \) for minimizing \( \kappa \) rather three \( \rightarrow \) \( \min_{c_1,c_2,c_3} \rho(\Upsilon) \) – relatively simple

Left:

Find \( \alpha \) to minimize \( \kappa \)
Optimizing $\kappa$ - Theoretic Insights

Assume $A_H = RAP$ (Galerkin form), $\Phi = I$, and the columns of $P$ are a subset of the eigenvectors of $A$.

Denote by $\mathcal{R}(I^h_H)$ the range of the prolongation and

$$
\eta_{fmax} = \max_{i: w_i \notin \mathcal{R}(P)} \eta_i, \quad \eta_{fmin} = \min_{i: w_i \notin \mathcal{R}(P)} \eta_i,
$$

$$
\eta_{cmax} = \max_{i: w_i \in \mathcal{R}(P)} \eta_i, \quad \eta_{cmin} = \min_{i: w_i \in \mathcal{R}(P)} \eta_i.
$$

where $\eta_i$ and $w_i$ are the eigenvalues and corresponding eigenvectors of $A$.

We have

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

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

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 $\kappa$ - In Practice

It is challenge for a general $A$.

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

*Strategy I: Local Fourier Analysis*

Example: two dimensional & two-grid analysis

Denote:

$T_{low} : \left[ -\frac{\pi}{2}, \frac{\pi}{2} \right]^2$ & $L_h$ the elliptic operator & $\tilde{L}_h(\theta_1, \theta_2)$ the symbol of $L_h$

Remind: $W_\alpha = \alpha A + (1 - \alpha) P A H^{-1} R A$ (extend to $\Phi \neq I$ obviously)

The eigenvalues of $W_\alpha \Leftrightarrow 4 \times 4$, $\tilde{W}_{\alpha, \theta_1, \theta_2}$ over the whole $(\theta_1, \theta_2) \in T_{low}^\text{low}$

$$\tilde{W}_{\alpha, \theta_1, \theta_2} = \alpha \tilde{A}^{\theta_1, \theta_2} + (1 - \alpha) \tilde{P}^{\theta_1, \theta_2} \left( \tilde{A}_H^{\theta_1, \theta_2} \right)^{-1} \tilde{R}^{\theta_1, \theta_2} \tilde{A}^{\theta_1, \theta_2}$$
Strategy I and Example continued:

\[
\tilde{W}_{\alpha}^{\theta_1, \theta_2} = \alpha \tilde{A}^{\theta_1, \theta_2} + (1 - \alpha) \tilde{P}^{\theta_1, \theta_2} \left( \tilde{A}^{-1}_{H}^{\theta_1, \theta_2} \right) \tilde{R}^{\theta_1, \theta_2} \tilde{A}^{\theta_1, \theta_2}
\]

\[
\tilde{A}^{\theta_1, \theta_2} = \begin{bmatrix}
\tilde{L}_h(\theta_1, \theta_2) \\
\tilde{L}_h(\bar{\theta}_1, \theta_2) \\
\tilde{L}_h(\theta_1, \bar{\theta}_2) \\
\tilde{L}_h(\bar{\theta}_1, \bar{\theta}_2)
\end{bmatrix}
\]

\[
\tilde{A}^{-1}_{H}^{\theta_1, \theta_2} = \frac{1}{4} \tilde{L}_h(2\theta_1, 2\theta_2) - \text{rediscretization}
\]
or

\[
\tilde{A}^{-1}_{H}^{\theta_1, \theta_2} = \tilde{R}^{\theta_1, \theta_2} \tilde{L}_h(2\theta_1, 2\theta_2) \tilde{P}^{\theta_1, \theta_2} - \text{Galerkin form}
\]

\[
\bar{\theta}_i = \begin{cases}
\theta_i + \pi, & \text{if } \theta_i < 0 \\
\theta_i - \pi, & \text{if } \theta_i > 0
\end{cases}, \quad i = 1, 2
\]

where \(\tilde{R}^{\theta_1, \theta_2} \in \mathbb{R}^{4 \times 1}\) and \(\tilde{P}^{\theta_1, \theta_2} \in \mathbb{R}^{1 \times 4}\) denote the symbols of \(R\) and \(P\), respectively.
**Strategy II**: Evaluate on a small size of grids - deterioration

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

Linear search $\Rightarrow \min_{\alpha \in [0,1]} \text{cond}(\mathbf{W}_\alpha) \Rightarrow$ e.g., MATLAB “fminbnd”
What Is Left?

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

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

where \( T^{\text{high}} : [-\pi, \pi)^2 \setminus \left[-\frac{\pi}{2}, \frac{\pi}{2}\right)^2 \).

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:

\[ \| L^h u^h_k - f^h \|_F \]

Convergence Factor:

\[ \frac{\| L^h u^h_k - f^h \|_F}{\| L^h u^h_{k-1} - f^h \|_F} \]

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

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

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

(d) \( \varepsilon = 10^{-3}, \phi = \frac{\pi}{4} \)
Fine $64 \times 64$ & Periodic Boundary Condition

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

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>$\varepsilon$</th>
<th>Bilinear</th>
<th>Bicubic</th>
<th>Ideal One</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SESOP</td>
<td>Fixed</td>
<td>SESOP</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.332</td>
<td>0.332</td>
<td>0.333</td>
</tr>
<tr>
<td>$\pi/6$</td>
<td>$10^{-3}$</td>
<td>0.570</td>
<td>0.563</td>
<td>0.537</td>
</tr>
<tr>
<td>$\pi/6$</td>
<td>$10^{-4}$</td>
<td>0.572</td>
<td>0.565</td>
<td>0.538</td>
</tr>
<tr>
<td>$\pi/4$</td>
<td>$10^{-3}$</td>
<td>0.509</td>
<td>0.500</td>
<td>0.457</td>
</tr>
<tr>
<td>$\pi/4$</td>
<td>$10^{-4}$</td>
<td>0.511</td>
<td>0.502</td>
<td>0.458</td>
</tr>
</tbody>
</table>
Linear Problem Continued - Deterioration of Strategy II

Denote

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

Result: working on \( 128 \times 128 \) but solving \( 1024 \times 1024 \) & less

10% additional computation - benefit if work on a huge problem
Linear Problem Continued - Multilevel Results

fine $1024 \times 1024$ & 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) $\epsilon = 1, \phi = 0$
(h) $\epsilon = 1, \phi = 0$

(i) $\epsilon = 10^{-3}, \phi = \frac{\pi}{4}$
(j) $\epsilon = 10^{-3}, \phi = \frac{\pi}{4}$
Problem description:

\[
\begin{align*}
\min_u F (u(x,y)) &= \int_{\Omega} \| \nabla u(x,y) + \xi \|^p - f(x,y)u(x,y) \, dx \, dy \\
\text{such that} \quad u &= 0 \quad \text{on} \quad \partial \Omega,
\end{align*}
\]

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

PDE form:

\[
\begin{align*}
-\nabla \cdot \left( \| \nabla u + \xi \|^p - 2 \nabla u \right) &= f \quad \text{in} \quad \Omega \\
u &= 0 \quad \text{on} \quad \partial \Omega.
\end{align*}
\]

\( \xi > 0 \) 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

More experiments and the detail of our analyses ⇒ our paper [HYZ18]
Thanks & Questions?
Tao Hong, Irad Yavneh, and Michael Zibulevsky.  
Accelerating multigrid optimization via sesop.  

Stephen G Nash.  
A multigrid approach to discretized optimization problems.  

Michael Zibulevsky.  
Speeding-up convergence via sequential subspace optimization: Current state and future directions.  