Multilevel optimization using trust-region and linesearch approaches

S. Gratton$^1$  M. Mouffe$^1$  Ph. Toint$^2$  D. Tomanos$^2$
M. Weber-Mendonça$^2$

$^1$CERFACS and CNES, Toulouse, France
$^2$Department of Mathematics, University of Namur, Belgium
(philippe.toint@fundp.ac.be)

December 2008
Multilevel optimization using trust-region and linesearch approaches

S. Gratton\textsuperscript{1}  M. Mouffe\textsuperscript{1}  Ph. Toint\textsuperscript{2}  D. Tomanos\textsuperscript{2}  M. Weber-Mendonça\textsuperscript{2}

\textsuperscript{1}CERFACS and CNES, Toulouse, France

\textsuperscript{2}Department of Mathematics, University of Namur, Belgium

( philippe.toint@fundp.ac.be )

December 2008
1. Introduction

2. Recursive trust-region methods

3. Multigrid limited memory BFGS
Outline

1. Introduction

2. Recursive trust-region methods

3. Multigrid limited memory BFGS
Optimization of continuous problems occurs in many applications: shape optimization, data assimilation, control problems, ... Recent optimization methods have been designed to cope with these problems, including multilevel/multigrid algorithms. These algorithms involve the computation of a hierarchy of problem descriptions, linked by known operators.

Our purpose: review some trust-region and linesearch recent proposals for unconstrained/ bound-constrained optimization:

\[ \min_{x \geq 0} f(x) \]
Can we use a structure of the form:

- Finest problem description
  - Restriction $\downarrow R$
  - $P \uparrow$ Prolongation

- Fine problem description
  - Restriction $\downarrow R$
  - $P \uparrow$ Prolongation

- Coarse problem description
  - Restriction $\downarrow R$
  - $P \uparrow$ Prolongation

- Coarsest problem description
Grid transfer operators

\[ R_i : \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{n_{i-1}} \quad \text{Restriction} \]

\[ P_i : \mathbb{R}^{n_{i-1}} \rightarrow \mathbb{R}^{n_i} \quad \text{Prolongation} \]
Three keys to multigrid algorithms

- **Oscillatory** components of the error are representable on fine grids, but not on coarse grids.
- Iterative methods reduce **oscillatory components** much faster than smooth ones.
- **Smooth** on fine grids → **Oscillatory** on coarse ones.
Annihilate oscillatory error level by level:

Note: $P$ and $R$ are not orthogonal projectors!

A very efficient method for some linear systems
(when $A$(smooth modes) $\in$ smooth modes)
Past developments

- Wen-Goldfarb (2007) (linesearch, explicit smoothing, convergence on convex problems)
- Gratton-T (2008) (linesearch, implicit smoothing, convergence?)
Outline

1. Introduction

2. Recursive trust-region methods

3. Multigrid limited memory BFGS
Recursive multilevel trust region

At each iteration at the **fine** level:

1. **consider a coarser description** model with a trust region

   
   - **compute fine** $g$ (and $H$)  
   - **step and trial point**
   
   
   
   

2. **Restriction** $R$  
   
   - minimize the **coarse** model within the **fine** TR  
   
   
   
   
3. **evaluate** $f$ at the trial point

4. **if achieved decrease $\approx$ predicted decrease:**
   - **accept** the trial point
   - (possibly) **enlarge** the trust region

5. **else:**
   - **keep** current point
   - **shrink** the trust region
Until convergence:

- Choose either a Taylor or recursive model
  - Taylor model: compute a Taylor step
  - Recursive: apply the Algo recursively
- Evaluate change in the objective function
- If achieved reduction $\approx$ predicted reduction,
  - accept trial point as new iterate
  - (possibly) enlarge the trust region
  else
  - reject the trial point
  - shrink the trust region
- Impose: current TR $\subseteq$ upper level TR
Recursive trust-region methods

Norms and trust-region shapes

**RMTR**
- 2-norm TR and criticality measure
- good results, but trust region scaling problem (recursion)

**RMTR-\(\infty\)**
- \(\infty\)-norm (bound constraints)
- new criticality measure
- new possibilities for step length

\[ \Delta_{low,k} \]
\[ x_{low,0} \]
\[ x_{low,k} \]
\[ \Delta_{up} \]

\[ B_{low,k} \]
\[ x_{low,0} \]
\[ x_{low,k} \]
\[ R(B_{up}) \]

Philippe Toint (University of Namur)    Multilevel optimization (TR and LS)    December 2008
Taylor iterations in the 2-norm version satisfy the sufficient decrease condition

\[ m_i(x) - m_i(x + s) \geq \kappa_{\text{red}} g(x) \min \left[ \frac{g(x)}{\beta}, \Delta \right]. \]

Taylor iterations in the \( \infty \)-norm are constrained; they satisfy

\[ h_i(x) - h_i(x + s) \geq \kappa_{\text{red}} \chi_i(x) \min \left[ 1, \frac{\chi_i(x)}{\beta}, \Delta \right]. \]

where

\[ \chi(x) = \min_{d \in R^n, ||d|| \leq 1} \langle g, d \rangle. \]
Until convergence:

- Choose either a Taylor or recursive model
  - Taylor model: compute a Taylor step ($\infty$-norm)
  - Recursive: apply the Algo recursively
- Evaluate change in the objective function
- If achieved reduction $\approx$ predicted reduction,
  - accept trial point as new iterate
  - (possibly) enlarge the trust region
else
  - reject the trial point
  - shrink the trust region
- Impose: current TR $\subseteq$ Restricted upper level TR
Computing **good starting points:**

- Solve the problem on the coarsest level
  ⇒ Good starting point for the next fine level

- Do the same on each level
  ⇒ Good starting point for the finest level

- Finally solve the problem on the finest level
FMG : Combination of mesh refinement and V-cycles
Consider the minimum surface problem

$$\min_{v \in K} \int_0^1 \int_0^1 (1 + (\partial_x v)^2 + (\partial_y v)^2)^{\frac{1}{2}} \, dx \, dy,$$

where $K = \{ v \in H^1(S_2) \mid v(x, y) = v_0(x, y) \text{ on } \partial S_2 \}$ with

$$v_0(x, y) = \begin{cases} f(x), & y = 0, \quad 0 \leq x \leq 1, \\ 0, & x = 0, \quad 0 \leq y \leq 1, \\ f(x), & y = 1, \quad 0 \leq x \leq 1, \\ 0, & x = 1, \quad 0 \leq y \leq 1, \end{cases}$$

where $f(x) = x(1 - x)$.

Finite element basis (P1 on triangles) $\rightarrow$ convex problem.
### Recursive trust-region methods

#### Some typical results on MS \((n = 127^2, 6 \text{ levels})\)

<table>
<thead>
<tr>
<th></th>
<th>Mesh ref.</th>
<th>RMTR(_2)</th>
<th>RMTR(_\infty)</th>
<th>Mesh ref.</th>
<th>RMTR(_\infty)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nit</td>
<td>1057</td>
<td>23</td>
<td>10</td>
<td>2768</td>
<td>214</td>
</tr>
<tr>
<td>nf</td>
<td>23</td>
<td>38</td>
<td>15</td>
<td>649</td>
<td>240</td>
</tr>
<tr>
<td>ng</td>
<td>16</td>
<td>28</td>
<td>14</td>
<td>640</td>
<td>236</td>
</tr>
<tr>
<td>nH</td>
<td>17</td>
<td>20</td>
<td>6</td>
<td>32</td>
<td>101</td>
</tr>
</tbody>
</table>

---

Philippe Toint (University of Namur)  Multilevel optimization (TR and LS)  December 2008  20 / 38
Excellent numerical experience!
Adaptable to bound-constrained problems
Fully supported by (simpler?) theory
Fortran code in the polishing stages (→ GALAHAD)
Outline

1 Introduction

2 Recursive trust-region methods

3 Multigrid limited memory BFGS
Until convergence:

- Compute a search direction \( d = -Hg \)
- Perform a linesearch along \( d \), yieiding
  \[
  f(x^+) \leq f(x) + \alpha \langle g, d \rangle \quad \text{and} \quad \langle g^+, d \rangle \geq \beta \langle g, d \rangle
  \]
- Update the Hessian approximation to satisfy
  \[
  H^+(g^+ - g) = x^+ - x \quad \text{(secant equation)}
  \]

**BFGS update:**

\[
H^+ = \left( I - \frac{ys^T}{y^Ts} \right) H \left( I - \frac{ys^T}{y^Ts} \right) + \frac{ss^T}{y^Ts}
\]

with

\[
y = g^+ - g \quad \text{and} \quad s = x^+ - x
\]
Generating new secant equations

The fundamental secant equation: \( H^+ y = s \)

Motivation:

\[
G^{-1} y = s \quad \text{where} \quad G = \int_0^1 \nabla_{xx} f(x + ts) \, dt
\]

Assume:
- known invariants subspaces \( \{S_i\}_{i=1}^p \) of \( G \).
- known orthogonal projectors onto \( S_i \)

\[
G^{-1} S_i y = S_i G^{-1} y = S_i s
\]

\( \Rightarrow \) new secant equation: \( H^+ y_i = s_i \) with \( s_i = S_i s \) and \( y_i = S_i y \)
How accurate are these equations?

We prove

\[
\frac{\|E\|}{\|G\|} \leq \frac{\|Gs_i - y_i\|}{\|s_i\| \|G\|}
\]

Now let \( S_i = Q_i D_i Q_i^T \) and

\[
Q_i^T G Q_i = G_i \quad \text{and} \quad (Q_i^C)^T G Q_i = F_i.
\]

Then

\[
\frac{\|E_i\|}{\|G\|} \leq \frac{\|G_i D_i - D_i G_i\|}{\sigma_{\text{min}}(D_i) \|G\|} + \kappa(D_i) \frac{\|F_i\| \|s\|}{\|G\| \|s_i\|} \leq \kappa(D_i) \left[ 2 \frac{\|G_i\|}{\|G\|} + \frac{\|F_i\| \|s\|}{\|G\| \|s_i\|} \right]
\]
Until convergence:

- Compute a search direction $d = -Hg$
- Perform a linesearch along $d$, yielding
  
  \[ f(x^+) \leq f(x) + \alpha \langle g, d \rangle \text{ and } \langle g^+, d \rangle \geq \beta \langle g, d \rangle \]

- Update the Hessian approximation to satisfy
  
  \[ H^+ y = s \text{ and } H^+ y_i = s_i \quad (i = 1, \ldots, p) \]

Natural setting: limited-memory (BFGS) algorithm

⇒ apply L-BFGS with secant pairs $(s_1, y_1), \ldots, (s_p, y_p), (s, y)$
Are they reasonable settings where the $S_i$ are known?

**Idea:** Grid levels may provide invariant subspace information!

- **Fine grid:** all modes
- **Less fine grid:** all but the most oscillatory modes
- **Coarser grid:** relatively smooth modes
- **Coarsest grid:** smoothest modes

$P^i R^i$ provides a (cheap) approximate $S_i$ operator!
How to order the secant pairs?

- Update for lower grid levels (smooth modes) first or last?

Should we control collinearity?

- Remember nested structure of the $S_i$ subspaces...
- Test cosines of angles between $s$ and $s_i$?

What information should we remember?

- A memory-less BFGS method is possible!

Many possible choices!
It consists [Lewis,Nash,04] in finding the function $a(x)$ defined on $[0, \pi]$, that minimizes

$$\int_0^\pi (\partial_y u(x,0) - \phi(x))^2 \, dx,$$

where $\partial_y u$ is the partial derivative of $u$ with respect to $y$,

and where $u$ is the solution of the boundary value problem

$$\begin{align*}
\Delta u &= 0 \quad \text{in } S, \\
u(x, y) &= a(x) \quad \text{on } \Gamma, \\
u(x, y) &= 0 \quad \text{on } \partial S \setminus \Gamma.
\end{align*}$$
Consider here the two-dimensional model problem for multigrid solvers in the unit square domain $S_2$

$$-\Delta u(x, y) = f \text{ in } S_2$$
$$u(x, y) = 0 \text{ on } \partial S_2,$$

$f$ such that the analytical solution is $u(x, y) = 2y(1 - y) + 2x(1 - x)$.

5-point finite-difference discretization

Consider the variational formulation

$$\min_{x \in \mathbb{R}^{nr}} \frac{1}{2} x^T A_r x - x^T b_r,$$
Data assimilation: the 4D-Var functional

- Consider a dynamical system \( \dot{x} = f(t, x) \) with solution operator \( x(t) = \mathcal{M}(t, x_0) \).

- Observations \( b_i \) at time \( t_i \) modeled by \( b_i = \mathcal{H}x(t_i) + \varepsilon \), where \( \varepsilon \) is a Gaussian noise with covariance matrix \( R_i \).

- The a priori error error covariance matrix on \( x_0 \) is \( B \).

- We wish to find \( x_0 \) which minimizes

\[
\frac{1}{2} \|x_0 - x_b\|_B^{-1} + \frac{1}{2} \sum_{i=0}^{N} \|\mathcal{H}\mathcal{M}(t_i, x_0) - b_i\|_{R_i}^{-1},
\]

- The first term in the cost function is the background term, the second term is the observation term.
The shallow system is often considered as a good approximation of the dynamical systems used in ocean modeling.

It is based on the Shallow Water equations

\[
\begin{align*}
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial v}{\partial y} - fv + g \frac{\partial z}{\partial x} &= \lambda \Delta u \\
\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + fu + g \frac{\partial z}{\partial y} &= \lambda \Delta v \\
\frac{\partial z}{\partial t} + u \frac{\partial z}{\partial x} + v \frac{\partial z}{\partial y} + z \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) &= \lambda \Delta z
\end{align*}
\]

Observations: every 5 points in the physical domain at every 5 time steps.

The a priori term is modeled using a diffusion operator [Weaver, Courtier, 2001].

The system is time integrated using a leapfrog scheme.

The damping in $\lambda \Delta$ improves spatial solution smoothness.
Relative accuracy of the multigrid secant equations

Plot $\|E\|/\|G\|$ against $k$

$\Rightarrow$ size of perturbation marginal
Testing a few variants

In our tests:

- old approximate secant pairs are discarded
- the LM updates are started with $\langle y, s \rangle \|y\|^2$ times the identity
- L-BFGS + 8 algorithmic variants:

<table>
<thead>
<tr>
<th></th>
<th>collinearity control (0.999)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
</tr>
<tr>
<td>Update order</td>
<td>mem</td>
</tr>
<tr>
<td>Coarse first</td>
<td>CNM</td>
</tr>
<tr>
<td>Fine first</td>
<td>FNM</td>
</tr>
</tbody>
</table>

Memory management:

*M*: past “exact” secant pairs are used (mem)

*N*: past “exact” secant pairs are not used (nomem)
Multigrid limited memory BFGS

The results

<table>
<thead>
<tr>
<th>Algo</th>
<th>DN ((n = 255))</th>
<th>MG ((n = 127^2))</th>
<th>SW ((n = 63^2))</th>
<th>MS ((n = 127^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>levels/mem</td>
<td>7/10</td>
<td>6/9</td>
<td>3/5</td>
<td>4/5</td>
</tr>
<tr>
<td>L-BFGS</td>
<td>330/319</td>
<td>308/299</td>
<td>64/61</td>
<td>387/378</td>
</tr>
<tr>
<td>CNM</td>
<td>94/84</td>
<td>137/122</td>
<td>66/61</td>
<td>196/170</td>
</tr>
<tr>
<td>CNN</td>
<td>125/100</td>
<td>174/134</td>
<td>57/55</td>
<td>408/338</td>
</tr>
<tr>
<td>CYM</td>
<td>110/92</td>
<td>123/104</td>
<td>64/61</td>
<td>224/192</td>
</tr>
<tr>
<td>CYN</td>
<td>113/89</td>
<td>138/107</td>
<td>57/55</td>
<td>338/267</td>
</tr>
<tr>
<td>FNM</td>
<td>120/100</td>
<td>172/144</td>
<td>63/57</td>
<td>241/208</td>
</tr>
<tr>
<td>FNN</td>
<td>137/89</td>
<td>151/120</td>
<td>65/62</td>
<td>280/221</td>
</tr>
<tr>
<td>FYM</td>
<td>90/76</td>
<td>149/128</td>
<td>63/57</td>
<td>211/176</td>
</tr>
<tr>
<td>FYN</td>
<td>140/107</td>
<td>153/120</td>
<td>65/62</td>
<td>283/216</td>
</tr>
</tbody>
</table>

(NF/NIT)
Observations:

- L-BFGS acts as a smoother
- the step is asymptotically very smooth
- the eigenvalues associated with the smooth subspace are (relatively) close to each other
- the step is asymptotically an approximate eigenvector
- an equation of the form

\[
Hs_i = \frac{\langle y_i, s_i \rangle}{\|y_i\|^2} s_i
\]

can also be included…

⇒ more (efficient) algorithmic variants!
Conclusions

Multilevel/multigrid optimization useful and interesting

Much remains to be explored

Recursive trust-region methods often very effective

Invariant subspace information useful for some problems

Multilevel quasi-Newton information exploitable
Perspectives

- More complicated constraints
- Better understanding of approximate secant/eigen information
- Invariant subspaces without grids?
- Multilevel L-BFGS in RMTR?
- Combination with ACO methods?
- More test problems?

Thank you for your attention!

Papers: http://perso.fundp.ac.be/~phtoint/publications.html