A Line search Multigrid Method for Large-Scale Nonlinear Optimization

Zaiwen Wen[1] Donald Goldfarb[1]

[1] Department of Industrial Engineering and Operations Research
Columbia University

2008 Siam Conference on Optimization

What's New?

Multilevel/Multigrid Methods in “Siam Conference on Optimization 2008”

- One plenary talk: “Multiscale Optimization”
- One dedicated session
- Four sessions on PDE-Based problems which are mainly handled by multigrid methods
- Almost 20 invited and contributed talks

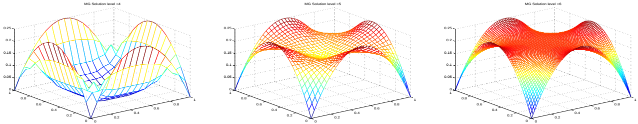
Statement of Problem

Consider solving problem $\boxed{\min_{u \in \mathcal{V}} \mathcal{F}(u)}$

- Infinite-dimensional problem: $\mathcal{F}$ is a functional
- $\mathcal{F}$ has a closely related representations $\{f_h\}$ on a hierarchical discretization levels h.

  $\boxed{\text{Discretize-then-Optimize scheme: } \min f_h}$
- Solutions of $\{\min f_h\}$ might have similar structures

Figure: Solution Structure of $\mathcal{F}(u) = \int_\Omega \sqrt{1 + \|\nabla u(x)\|^2}\,\mathrm{dx}$



(a) Level 4 (b) Level 5 (c) Level 6

Sources of Problems

- Applications in nonlinear PDEs, image processing:

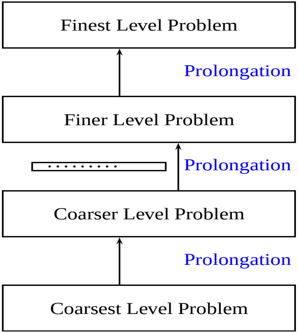
$$\min_{u \in U} \mathcal{F}(u) = \int_{\Omega} \mathcal{L}(\nabla u, u, x)\, dx$$

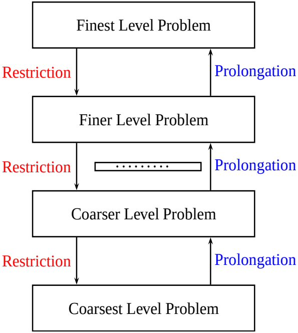
- PDE-constrained optimization: optimal control problems and inverse Problems. Example: finding a local volatility $\sigma(t, x)$ such that the prices $C(T, S)$ from the Black-Scholes PDEs match the observed prices on the market.

$$\begin{cases} \min \quad J(C, \sigma) := \sum_{I} |C(S_i, T_i) - z(S_i, T_i)|^2 + \alpha J_r(\sigma) \\ \text{s.t.} \quad \partial_\tau C - \frac{\sigma^2 K^2}{2} \partial^2_{KK} C + (r - q) K \partial_K C + qC = 0, \\ \qquad C(0, K) = (S - K)_+, \quad K > 0, \quad \tau \in (0, +\infty), \end{cases}$$

where $J_r(\sigma)$ is the regularization term.

Mesh-refinement Method





Linear Multigrid Methods

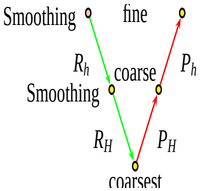
Multigrid method for solving $A_h x_h = b_h$, $A_h \succeq 0$

- Inter-grid operations: Restriction R_h, Prolongation P_h.
- Smoothing: reducing high frequency errors efficiently
- Coarse grid Correction: Let $A_{h-1} = R_h A_h P_h$.

Algorithm

Multigrid-cycle: $x_{h,k+1} = MGCYCLE(h, A_h, b_h, x_{h,k})$

```
-PRE-SMOOTHING: Compute $\bar{x}_{h,k} = x_{h,k} + B_h(b_h - A_h x_{h,k})$.
-COARSE GRID CORRECTION:
    Compute the residual $\bar{r}_{h,k} = b_h - A_h \bar{x}_{h,k}$.
    Restrict the residual $\bar{r}_{h-1,k} = R_h \bar{r}_{h,k}$.
    -Solve the Coarse Grid Residual Equation $A_{h-1} e_{h-1,k} = \bar{r}_{h-1,k}$
    IF $h - 1 = \mathrm{N}_0$, solve $e_{h-1,k} = A_{h-1}^{-1} \bar{r}_{h-1,k}$,
    ELSE call $e_{h-1,k} = MGCYCLE(h-1, A_{h-1}, \bar{r}_{h-1,k}, 0)$.
    Interpolate the correction: $e_{h,k} = P_h e_{h-1,k}$.
    Compute the new approximation solution: $x_{h,k+1} = \bar{x}_{h,k} + e_{h,k}$.
```



Methods for Solving nonlinear PDEs

- Global linearization method (Newton's Method)
- Local linearization method, such as the full approximation scheme (Brandt, Hackbusch)
- Combination of global and local linearization method (Irad Yavneh, Gregory Dardyk)
- Projection based multilevel method (Stephen Mccormick, Thomas Manteuffel, Oliver Rohrle, John Ruge)
- Multigrid methods for obstacle problems (Ralf Kornhuber, Carsten Graser)
- ...

Multigrid Methods for Optimization

- Traditional optimization methods where the system of linear equations is solved by multigrid methods (A.Borzi, K.Kunisch, Volker Schulz, Thomas Dreyer, Bernd Maar, U.M.Ascher, E.Haber)
- The multigrid optimization framework proposed by Nash and Lewis. (extensions given by A.Borzi)
- The recursive trust region method for unconstrained and box-constrained optimization (S. Gratton, A. Sartenaer, P. Toint, M. Weber, D. Tomanos, M.Mouffe, M. Ulbrich, S. Ulbrich, B. Loesch)
- Multigrid method for image processing (Raymond Chan, Ke Chen, Xue-cheng Tai, Tony Chan, Elad Haber, Jan Moderstitzki)
- ...

General Idea

Consider

$$\min_{x_h \in \mathcal{V}_h} f_h(x_h)$$

on a class of nested spaces $\mathcal{V}_{\mathrm{N}_0} \subset \cdots \subset \mathcal{V}_{\mathrm{N}-1} \subset \mathcal{V}_{\mathrm{N}} \subset \mathcal{V}$.

- General idea of line search:

$$x_{h,k+1} = x_{h,k} + \alpha_{h,k} d_{h,k}, \quad d_{h,k} \in \mathcal{V}_h,$$

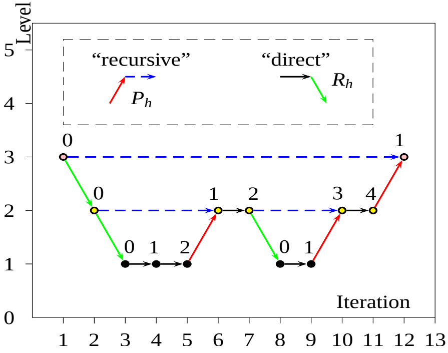
where $d_{h,k}$ is the search direction and $\alpha_{h,k}$ is the step size.

- General idea of using coarse grid information:

$$d_{h,k} = \arg\min_{d_H \in \mathcal{V}_H} f_h(x_{h,k} + d_H).$$

An illustration of the multigrid optimization framework.

- A direct search direction (Taylor iteration), marked by →, is generated on the current level.
- A recursive search direction, marked by - - →, is generated from the coarser levels.



Construction of recursive search direction

If condition (Gratton, Sartenaer, Toint)

$$\|R_h g_{h,k}\| \geq \kappa \|g_{h,k}\|, \quad \|R_h g_{h,k}\| \geq \epsilon_h$$

hold, we compute

recursive search direction

$$d_{h,k} = P_h d_H^* = P_h(x_{H,i^*} - x_{H,0}) = P_h \left(\sum_{i=0}^{i^*-1} \alpha_{H,i} d_{H,i} \right),$$

Two major difficulties:

1. The calculation of $f_h(x_{h,k} + d_H)$ might be expensive.
2. The recursive direction might not be a descent direction.

Construction of the Coarse Level Model

Coarse Level Model (Nash, 2000)

$$\min_{x_H} \{\psi_H(x_H) \equiv f_H(x_H) - (v_H)^\top x_H\},$$

where $v_H = \nabla f_{H,0} - R_h g_{h,k}$ and $g_{h,k} = \nabla \psi_h(x_{h,k})$.

- Define $v_{\mathrm{N}} = 0$, then $f_{\mathrm{N}}(x_{\mathrm{N}}) = \psi_{\mathrm{N}}(x_{\mathrm{N}})$
- The scheme is compatible to the FAS scheme
- Second order coherence about the Hessian can also be enforced

Properties of “Recursive Search Direction” $d_{h,k}$

Assumption

Assume $\sigma_h P_h = R_h^\top$.

- First-order coherence

$$g_{H,0} = R_h g_{h,k}, \quad (d_{h,k})^\top g_{h,k} = (d_H^*)^\top g_{H,0}.$$

- If $f(x)$ is convex, the direction $d_{h,k}$ is a descent direction $(d_{h,k})^\top g_{h,k} < 0$ and the directional derivative $(d_{h,k})^\top g_{h,k}$ satisfies

$$-(d_{h,k})^\top g_{h,k} \geq \psi_{H,0} - \psi_{H,i^*}.$$

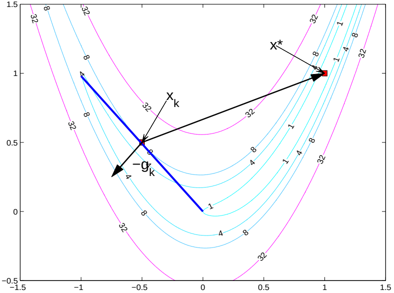
Failure of recursive search direction

Consider “Rosenbrock” function

$$\varphi(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

- Local minimizer $x^* = (1,1)^\top$, initial point $x_0 = (-0.5, 0.5)^\top$.
- Search along the direction $x^* - x_0$?

$$\nabla\varphi(x_0)^\top(x^* - x_0) = (47, 50)(1.5, 0.5)^\top = 95.5 > 0$$



Obtaining a “descent” recursive direction

- Convexify the function $f(x)$ on the coarser level. For example, setting $\widehat{\psi}_H = \psi_H + \lambda_H \|x_H - x_{H,0}\|_2^2$.
- Trust region method:

$$\min m_h(s_{h,k}), \quad \text{subject to } s_{h,k} \in \mathcal{B}_h$$

- Line search method, but checking the decent condition at each step.
- Non-monotone line search
- Watch-dog technique

A New Line Search Scheme

Given the direction $d_{h,k} = P_h \widetilde{d}_{H,i} = P_h(x_{H,i} - x_{H,0})$.

- One can check $g_{h,k}^\top d_{h,k}$ on the coarser levels since the first order coherence: $g_{h,k}^\top d_{h,k} = g_{H,0}^\top \widetilde{d}_{H,i}$
- Since $\psi_H(x_{H,i}) \approx \psi_H(x_{H,0}) + g_{H,0}^\top (x_{H,i} - x_{H,0})$, we check

$$\psi_H(x_{H,i}) > \psi_{H,0} + \rho_2 g_{H,0}^\top \widetilde{d}_{H,i} = \psi_{H,0} + \rho_2 g_{h,k}^\top d_{h,k}$$

  Then

$$g_{h,k}^\top d_{h,k} < \rho_2^{-1}(\psi_H(x_{H,i}) - \psi_{H,0}) \leq 0$$

- Sufficient reduction of the function value (Armijo Condition):

$$\psi_h(x_{h,k} + \alpha_{h,k} d_{h,k}) \leq \psi_{h,k} + \rho_1 \alpha_{h,k} (g_{h,k})^\top d_{h,k}$$

A New Line Search Scheme

Line search conditions on the coarser level

$$
\begin{aligned}
&(A) && \psi_h(x_{h,k}+\alpha_{h,k}d_{h,k}) \le \psi_{h,k} + \rho_1\alpha_{h,k}(g_{h,k})^\top d_{h,k}. \\
&(B) && \psi_h(x_{h,k}+\alpha_{h,k}d_{h,k}) > \psi_{h,0} + \rho_2 g_{h,0}^\top(x_{h,k}+\alpha_{h,k}d_{h,k} - x_{h,0})
\end{aligned}
$$

- Condition (B) is similar to the Goldstein rule if $k = 0$ and $\rho_2 = 1 - \rho_1$.

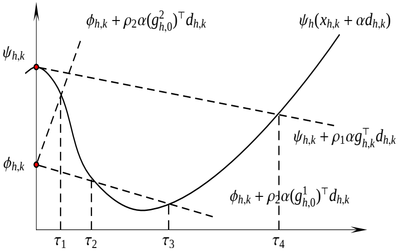
Algorithm

Backtracking Line Search

```
Step 1.  Given $\alpha_\rho > 0$, $0 < \rho_1 < \frac{1}{2}$ and $1 - \rho_1 \le \rho_2 \le 1$. Let $\alpha^{(0)} = \alpha_\rho$. Set $l = 0$.
Step 2.  If ($h = \mathrm{N}$ and condition (A) is satisfied) or if ($h < \mathrm{N}$ and both conditions (A) and (B) are satisfied),
         RETURN $\alpha_{h,k} = \alpha^{(l)}$.
Step 3.  Set $\alpha^{(l+1)} = \tau\alpha^{(l)}$, where $\tau \in (0, 1)$. Set $l = l + 1$ and go to Step 2.
```

Existence of the step size

- The first step size $\alpha_{h,0}$ always exists!
- Suppose $\alpha_{h,i}$ for $i = 0, \cdots, k$ exist, then $\alpha_{h,k+1}$ also exists.
  - Define $\phi_{h,k} := \psi_{h,0} + \rho_2 g_{h,0}^\top (x_{h,k} - x_{h,0})$.
  - $\psi_{h,k} > \psi_{h,0} + \rho_2 g_{h,0}^\top (x_{h,k-1} + \alpha_{h,k-1} d_{h,k-1} - x_{h,0}) = \phi_{h,k}$
  - Condition (B): $\psi_h(x_{h,k} + \alpha_{h,k} d_{h,k}) > \phi_{h,k} + \rho_2 \alpha_{h,k} g_{h,0}^\top d_{h,k}$


$\phi_{h,k} + \rho_2 \alpha (g_{h,0}^2)^\top d_{h,k}$
$\psi_h(x_{h,k} + \alpha d_{h,k})$
$\psi_{h,k}$
$\psi_{h,k} + \rho_1 \alpha g_{h,k}^\top d_{h,k}$
$\phi_{h,k}$
$\phi_{h,k} + \rho_2 \alpha (g_{h,0}^1)^\top d_{h,k}$
τ_1
τ_2
τ_3
τ_4

- Exit MG if the step size is too small

Choice of direct search direction

We require that a direct search satisfy

Direct Search Condition

$$\|d_{h,k}\| \le \beta_{\mathcal{T}} \|g_{h,k}\| \text{ and } -(d_{h,k})^{\top} g_{h,k} \ge \eta_{\mathcal{T}} \|g_{h,k}\|^2.$$

The following directions satisfy the condition (convex)

- Steepest descent direction $d_{h,k} = -g_{h,k}$.
- Exact Newton's direction $d_{h,k} = -G_{h,k}^{-1} g_{h,k}$.
- Inexact Newton's direction generated by the conjugate gradient method.

Other possible choices:

- Inexact Newton's direction generated by the linear multigrid method
- Limited-memory BFGS direction

The Algorithm

Algorithm

$x_h = MGLS(h, x_{h,0}, \tilde{g}_{h,0})$

```
Step 1.  Given $\kappa > 0$, $\epsilon_h > 0$ and $\xi > 0$ and an integer $K$.
Step 2.  IF $h < \mathrm{N}$, compute $v_h = \nabla f_{h,0} - \tilde{g}_{h,0}$, set $g_{h,0} = \tilde{g}_{h,0}$;
         ELSE set $v_h = 0$ and compute $g_{h,0} = \nabla f_{h,0}$.
Step 3.  FOR $k = 0, 1, 2, \cdots$
    3.1.  IF $\|g_{h,k}\| \leq \epsilon_h$ or IF $h < \mathrm{N}$ and $k \geq \mathrm{K}$,
              RETURN solution $x_{h,k}$;
    3.2.  IF $h = \mathrm{N}_0$ or $\|R_h g_{h,k}\| < \kappa \|g_{h,k}\|$ or $\|R_h g_{h,k}\| < \epsilon_h$

              -Direct Search Direction Computation.
              Compute a descent search direction $d_{h,k}$ on the current level.

          ELSE

              -Recursive Search Direction Computation.
              Call $x_{h-1,i*} = MGLS(h-1, R_h x_{h,k}, R_h g_{h,k})$ to return a solution (or approximate solution)
              $x_{h-1,i*}$ of "$\min_{x_{h-1}} \psi_{h-1}(x_{h-1})$".
              Compute $d_{h,k} = P_h \widetilde{d}_{h-1,i*} = P_h(x_{h-1,i*} - R_h x_{h,k})$.
    3.3.  Call the backtracking line search Algorithm to obtain a step size $\alpha_{h,k}$.
    3.4.  Set $x_{h,k+1} = x_{h,k} + \alpha_{h,k} d_{h,k}$. IF $\alpha_{h,k} \leq \xi$ and $h < \mathrm{N}$, RETURN solution $x_{h,k+1}$.
```

Convergence for Convex functions

Let $\rho_2 = 1$. Assume $f_h(x)$ is twice continuously differentiable and uniformly convex. Suppose the “direct search” Condition is satisfied by all direct search steps.

1. The step size $\alpha_{h,k}$ is bounded below by a constant.
2. $-d_{h,k}^{\top} g_{h,k} \geq \eta_h \|g_{h,k}\|^2$ and $\cos(\theta_{h,k}) \geq \delta_h$, where $\theta_{h,k}$ is the angle between $d_{h,k}$ and $-g_{h,k}$.
3. $\{x_{\mathrm{N},k}\}$ converges to the unique minimizer $\{x_{\mathrm{N}}^*\}$ of $f_{\mathrm{N}}(x_{\mathrm{N}})$.
4. The rate of convergence is at least R-linear.
5. For any $\epsilon > 0$, after at most $\tau = \frac{\log((f_{\mathrm{N}}(x_{\mathrm{N},0}) - f_{\mathrm{N}}(x_{\mathrm{N}}^*))/\epsilon)}{\log(1/c)}$ iterations, where $0 < c = 1 - \frac{\chi\alpha^*\eta_{\mathrm{N}}}{2} < 1$, we have $f_{\mathrm{N}}(x_{\mathrm{N},k}) - f_{\mathrm{N}}(x_{\mathrm{N}}^*) \leq \epsilon$.

Convergence for general nonconvex functions

Let $1 - \rho_1 \leq \rho_2 < 1$. Assume all direct search direction $d_{h,k}$ satisfy the “direct search” condition; the level set

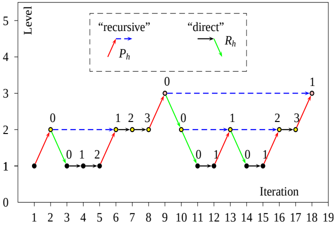
$$\mathcal{D}_h = \{x_h : \psi_h(x_h) \leq \psi_h(x_{h,0})\}$$

is bounded; the objective function ψ_h is continuously differentiable and the gradient $\nabla\psi_h$ is Lipschitz continuous.

1. The step size and the directional derivative of the first iteration of each minimization sequence on the coarser levels can be bounded from below by the norm of gradient raised to some finite power
2. At the uppermost level $\lim_{k\to\infty}\|\nabla f_{\mathrm{N}}(x_{\mathrm{N},k})\| = 0$.

Implementation Issues

- Discretization of the problems
- Choices of the direct search direction
- Prolongation and Restriction operator
- Full multigrid method



- Sharing information on different minimization sequences

$$\mathcal{P}^{(l)}: \quad \min \psi_h^{(l)}(x_h) = f_h(x_h) - (v_h^{(l)})^\top x_h,$$

$$\mathcal{P}^{(l+1)}: \quad \min \psi_h^{(l+1)}(x_h) = f_h(x_h) - (v_h^{(l+1)})^\top x_h,$$

Test Examples

We compared the following three algorithms:

1. The standard L-BFGS method, denoted by “L-BFGS”.
2. The mesh refinement technique, denoted by “MRLS”.
3. The full multigrid Algorithm with one “smoothing” step, denoted by “FMLS”.

Implementation detail:

- The problems are discretized by finite difference.
- The initial point in the “FMLS” Algorithm is taken to be the zero vector. For the “MGLS” Algorithm , we set

$$\kappa = 10^{-4}, \epsilon_h = 10^{-5}/5^{\mathrm{N}-h}, \mathrm{K} = 100, \rho_1 = 10^{-3}, \rho_2 = 1 - \rho_1.$$

- The number of gradient and step difference pairs stored by the L-BFGS method was set to 5.

Minimal Surface Problem

Consider the minimal surface problem

$$\begin{aligned} \min \quad & f(u) = \int_\Omega \sqrt{1 + \|\nabla u(x)\|^2}\, \mathrm{dx} \\ \text{s.t.} \quad & u(x) \in K = \{u \in H^1(\Omega) : u(x) = u_\Omega^1(x) \text{ for } x \in \partial\Omega\}, \end{aligned} \tag{1}$$

where $\Omega = [0, 1] \times [0, 1]$ and

$$u_\Omega(x) = \begin{cases} y(1-y), & x = 0, 1, \\ x(1-x), & y = 0, 1. \end{cases}$$

Minimal Surface Problem

Table: Summary of computational costs

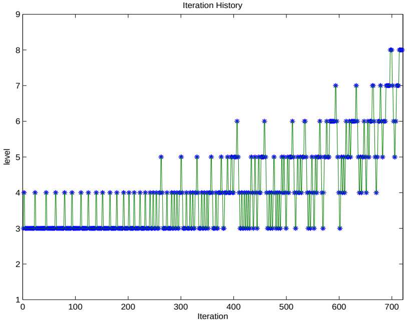
L-BFGS					
h	nls	nfe	nge	$\|\|g_{\mathrm{N}}^{*}\|\|_2$	CPU
8	820	837	821	7.614301e-06	169.0924

	FMLS						MRLS					
h	3	4	5	6	7	8	3	4	5	6	7	8
nls	268	150	96	61	22	10	17	26	45	66	88	25
nfe	355	236	140	80	28	12	20	27	47	71	91	26
nge	324	181	112	67	24	11	18	27	46	67	89	26
$\|\|g_{\mathrm{N}}^{*}\|\|_2$	7.198984e-06						9.031382e-06					
CPU	3.913212						7.363142					

h indicates the level; “nls", “nfe” and “nge" denote the total number of line searches, the total number of function evaluations and the total number of gradient evaluations at that level, respectively. We also report the total CPU time measured in seconds and the accuracy attained, which is measured by the Euclidean-norm $||g_{\mathrm{N}}^{*}||_2$ of the gradient at the final iteration.

Minimal Surface Problem

Figure: Iteration history of the “FMLS” method



Nonlinear PDE

Consider the nonlinear PDE:

$$\begin{aligned} -\Delta u + \lambda u e^{u} = f \qquad & \text{in} \quad \Omega, \\ u = 0 \qquad & \text{on} \quad \partial\Omega, \end{aligned} \tag{2}$$

where $\lambda = 10$, $\Omega = [0,1] \times [0,1]$ and

$$f = \left(9\pi^2 + \lambda e^{((x^2-x^3)\sin(3\pi y))}(x^2 - x^3) + 6x - 2\right)\sin(3\pi y),$$

and the exact solution is $u = (x^2 - x^3)\sin(3\pi y)$. The corresponding variational problem is

$$\min \mathcal{F}(u) = \int_{\Omega} \frac{1}{2}|\nabla u|^2 - \lambda(ue^{u} - e^{u}) - fu\, dx.$$

Nonlinear PDE

Table: Summary of computational costs

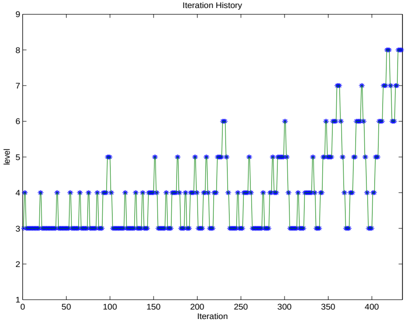
L-BFGS					
h	nls	nfe	nge	$\|\|g_{\mathrm{N}}^*\|\|_2$	CPU
8	431	463	432	8.858409e-06	56.4617

	FMLS						MRLS					
h	3	4	5	6	7	8	3	4	5	6	7	8
nls	182	101	50	28	13	8	16	27	49	58	69	59
nfe	234	144	73	41	18	11	20	31	53	61	75	60
nge	208	114	56	32	15	9	17	28	50	59	70	60
$\|\|g_{\mathrm{N}}^*\|\|_2$	5.190493e-06						9.401884e-06					
CPU	3.052931						12.225726					

h indicates the level; “nls", “nfe” and “nge" denote the total number of line searches, the total number of function evaluations and the total number of gradient evaluations at that level, respectively. We also report the total CPU time measured in seconds and the accuracy attained, which is measured by the Euclidean-norm $\|\|g_{\mathrm{N}}^*\|\|_2$ of the gradient at the final iteration.

Nonlinear PDE

Figure: Iteration history of the “FMLS” method



Summary

- Interpretation of Multigrid in optimization
- A new globally convergent line search algorithm by imposing a new condition on a modified backtracking line search procedure
- Outlook
  - More efficient “direct search” directions?
  - Multigrid for constrained optimization?
- Thanks
  - S.Nash and M.Lewis for the multigrid framework
  - P.Toint and his group for the Recursive TR method
  - Y. Yuan for the subspace minimization