Merging Trust-Region and Limited Memory Technologies for Large-Scale Nonlinear Optimization

James Burke   Andreas Weigmann   Xu Liang

April 7, 2008

Outline

1. Introduction.
    - General format of large-scale nonlinear optimization problem.
    - Main difficulties.
2. Large-scale unconstrained problem.
    - Description of the algorithm.
    - Numerical Algebra.
    - Numerical Results.
3. Active-set Trust-region Algorithm(ASTRAL) for bounded constrained problems.
    - Description of the algorithm.
    - Numerical Algebra.
    - Numerical Results.

Problem Format

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & l \leq x \leq u \end{array}$$

Assumptions:

1. $f : \mathbb{R}^n \to \mathbb{R}$ is smooth, but difficult to evaluate.
   high dimensional integration
   simulation
   systems of PDE/ODE's
   multiple levels of optimization

2. $-\infty \leq l_i \leq u_i \leq \infty$

3. n is large. ($n \geq 1000$, e.g. $n = 2^{19}$ for imaging applications).

Computational Costs

(1) Dominant Costs: function and gradient evaluations.

Jorge Moré and Nick Gould
No test sets now available for *hard* to evaluate functions.

(2) Numerical linear algebra.

A very significant concern due to dimensionality.
But, due to *expensive* functions, we want to exploit all data as fully as possible.

Computational Costs

(1) Dominant Costs: function and gradient evaluations.

line-search vs trust-region

(2) Numerical linear algebra.

matrix multiplies vs equation solves

The Unconstrained Problem

Objective: $\min f(x)$

Algorithm: $x^{k+1} = x^k + s^k$

Notation: $g^k = \nabla f(x^k), \quad B_k \approx \nabla^2 f(x^k)^{-1}, \quad H_k \approx \nabla^2 f(x^k)$

- Line-Search: $s^k = -\lambda_k B_k d^k$

  λ_k a stepsize (weak or strong Wolfe conditions).
  Requires repeated function and gradient evaluations.

- Trust-Region: s^k solves

$$\begin{array}{ll} \min & (g^k)^T s + \frac{1}{2} s^T H_k s \\ \text{s.t.} & \|s\| \le \Delta_k \end{array}$$

Hessian Approximations

Scalar Secant Equation:

$$H_k = \frac{\nabla f(x^k) - \nabla f(x^{k-1})}{x^k - x^{k-1}}$$

Matrix Secant Equation:

$$H_k(x^k - x^{k-1}) = \nabla f(x^k) - \nabla f(x^{k-1})$$

n linear equations in $\frac{n(n+1)}{2}$ unknowns

BFGS Update:

H_0 sym. and positive definite $\Rightarrow H_k$ sym. and positive definite

$$H_{k+1} = H_k - \frac{H_k s^k {s^k}^T H_k}{{s^k}^T H_k s^k} + \frac{y^k {y^k}^T}{{s^k}^T y^k} \qquad \text{whenever } {s^k}^T y^k > 0$$

$$= H_k - [y^k \;\; H_k s^k] \begin{pmatrix} -{s^k}^T y^k & 0 \\ 0 & {s^k}^T H_k s^k \end{pmatrix}^{-1} [y^k \;\; H_k s^k]^T$$

where

$$s^k = x^{k+1} - x^k \quad \text{and} \quad y^k = g^{k+1} - g^k.$$

Compact m-Step Representation

Nocedal, Byrd and Schnabel (1994)

$$H_{k+m} \;=\; H_k - [Y \;\; H_k S] \begin{pmatrix} -D & L^T \\ L & {s^k}^T H_k s^k \end{pmatrix}^{-1} [Y \;\; H_k S]^T$$

where

$$S = [s^k, \cdots, s^{k+m-1}], \qquad Y = [y^k, \cdots, y^{k+m-1}],$$

and

$$S^T Y = L + D + R$$

with L strictly lower triangular,
D diagonal, and
R strictly upper triangular.

Limited Memory BFGS Updating

$$H = \lambda I - \Psi\Gamma^{-1}\Psi^T.$$

where

$$\lambda = \frac{{y^k}^T y^k}{{s^k}^T y}, \qquad \Psi = [Y,\ \lambda S], \qquad \Gamma = \begin{bmatrix} -D & L^T \\ L & \lambda S^T S \end{bmatrix}_{2m\times 2m},$$

$$S^T Y = L + D + R$$

$$S = [s^{k_1}, \cdots, s^{k_m}], \qquad Y = [y^{k_1}, \cdots, y^{k_m}]$$

Typically, $m = 5$.
λ-scaling – an approximate Rayleigh quotient
Oren-Spedicato (1976), Phua-Shanno (1980), Barzilai-Borwein (1988).

The Trust-Region Subproblem

$$\begin{array}{ll} \min & (g^k)^T s + \frac{1}{2} s^T H_k s \\ \text{s.t.} & \|s\| \leq \Delta_k \end{array}$$

Apply Newton's method to find $\mu > 0$ so that

$$\phi(\mu) = 0$$

where

$$\phi(\mu) = \frac{1}{\Delta_k} - \frac{1}{\|s(\mu)\|} \qquad \text{and} \qquad s(\mu) = -(\mu I + H_k)^{-1} g^k \ .$$

$$\mu_+ = \mu - \frac{\phi(\mu)}{\phi'(\mu)}, \quad \text{where} \quad \phi'(\mu) = -\frac{g^T(\mu I + H)^{-3} g}{\|s(\mu)\|^3}.$$

(Our implementation avoids the *hard case*.)

Powers of $(\mu I + H)^{-1}$

$$(\mu I + H)^{-1} = \frac{1}{\tau}[I + \Psi(\tau\Gamma - \Psi^T\Psi)^{-1}\Psi^T]$$

$$(\mu I + H)^{-2} =$$

$$\frac{1}{\tau^2}[I + \Psi(\tau\Gamma - \Psi^T\Psi)^{-1}\Psi^T + \tau\Psi(\tau\Gamma - \Psi^T\Psi)^{-1}\Gamma(\tau\Gamma - \Psi^T\Psi)^{-1}\Psi^T],$$

where $\tau = \mu + \lambda$.

Triangular Factorization of $\tau\Gamma - \Psi^T\Psi$

Set

$$\hat{D} = (\mu + \lambda)D + Y^T Y, \quad \hat{W} = \lambda\mu S^T S, \quad \text{and} \quad \hat{L} = \mu L - \lambda(R + D).$$

Compute Cholesky factorizations

$$\begin{array}{rcl} \hat{D} & = & M M^T \\ \hat{W} + \hat{L}\hat{D}^{-1}\hat{L}^T & = & J J^T \end{array}$$

Then

$$\tau\Gamma - \Psi^T\Psi = \begin{bmatrix} M & 0 \\ -\hat{L}M^{-T} & J \end{bmatrix} \begin{bmatrix} -M^T & M^{-1}\hat{L}^T \\ 0 & J^T \end{bmatrix}.$$

The trust-region Newton iteration occurs entirely in dimension $2m$.
Cost $\approx O(mn)$ assuming $m^3 << n$.

Numerical Results

TEST SET: 24 problems from MINPACK-2 set, $2500 \leq n \leq 160,000$.

Termination Criteria:

$$f_{\text{best}} \sim \text{ best known function value.}$$

1. $|f^k - f_{\text{best}}| / \max(1, |f_{\text{best}}|) \leq \epsilon$.

2. $\text{nf} \geq 1000$.

Performance Profile: (Dolen and Moré 2001)

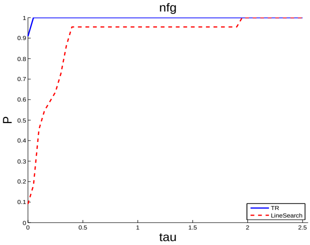
Given a problem set $\mathcal{P}$,

$$
\begin{aligned}
t_{i,s_j} &= \text{nfg(CPU time) of solving prob i by alg } s_j \\
r_{i,s_j} &= \frac{t_{i,s_j}}{\min_j t_{i,s_j}}.
\end{aligned}
$$

Set

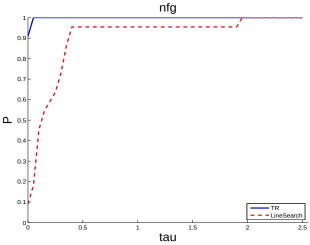
$$\rho_{s_j}(\tau) = \frac{1}{n_p}|\{i \in \mathcal{P} : r_{i,s_j} \leq \tau\}|.$$

Comparison of number of the function and gradient evaluations

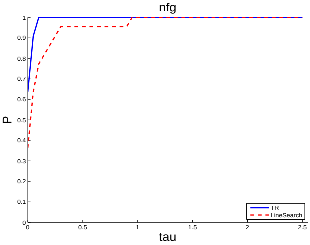


(a) relative accuracy 10^{-5}

Comparison of number of the function and gradient evaluations

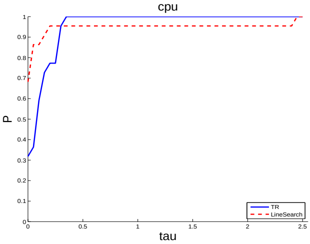


(c) relative accuracy 10^{-5}

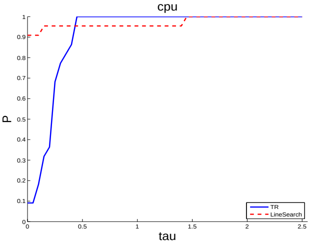


(d) relative accuracy 10^{-3}

Comparison of CPU time



(e) relative accuracy 10^{-5}



(f) relative accuracy 10^{-3}

More Implementation Details

Initialization: $x^0,\ g^0 = \nabla f(x^0),\ H_0 \succeq 0,\ 0 < \kappa < 1,\ 0 < \sigma < 1.$

Iteration:

1. Set $\bar{s} = -H_k^{-1} g^k$ and $r(\bar{s}) = \frac{f(x^k+\bar{s})-f(x^k)}{q(\bar{s})}$. (98% acceptance)

2. WHILE $r(\bar{s}) < \kappa$

   a. Let $\delta = \sigma \|\bar{s}\|$.

   b. Solve

$$\begin{array}{ll} \min & q(s) = s^T g^k + \frac{1}{2} s^T H_k s \\ \text{s.t.} & \qquad ||s|| \le \delta. \end{array}$$

   c. Compute $r(\bar{s}) = \frac{f(x^k+\bar{s})-f(x^k)}{q(\bar{s})}$.

   END WHILE

3. Set $x^{k+1} = x^k + \bar{s}$. Update H_{k+1}.

Minimization with bounded constraints

$$(\mathcal{P}) \qquad \min \qquad f(x)$$
$$l \le x \le u.$$

Active-set Trust-region Algorithm (ASTRAL).

We use an ℓ^∞ trust-region to conform with the constraint geometry.

$$\Omega = \{x \mid l \le x \le u\}, \quad \mathbb{B}_\infty = \{x \mid \|x\|_\infty \le 1\}$$

and

$$\Omega_k = \Omega \cap (x_k + \Delta_k \mathbb{B}_\infty) = \left\{x \,\middle|\, l^k \le x \le u^k \right\}.$$

Binding Constraints

Active Constraints

$$A(x) = \{i \mid x_i = l_i \text{ or } x_i = u_i\}$$

Binding Constraints

$$\mathcal{B}(x) = \left\{ i \;\middle|\; \begin{array}{c} x_i = l_i \text{ and } (\nabla f(x))_i > 0, \\ \text{or } x_i = u_i \text{ and } (\nabla f(x))_i < 0 \end{array} \right\}$$

Non-Binding Constraints

$$\mathcal{B}^c(x) = \{1, 2, \ldots, n\} \setminus \mathcal{B}(x), \qquad \nu(x) = |\mathcal{B}^c(x)|.$$

$\Phi(x)$ is an $n \times \nu(x)$ matrix whose columns are those of the identity matrix corresponding to the non-binding constraints $\mathcal{B}^c(x)$.

Active-set Trust-region Algorithm

1. Identify $\mathcal{B}(x^k)$ and set $\Phi_k = \Phi(x^k)$.

2. Set $\quad \tilde{H}_k = \Phi_k^T H_k \Phi_k, \quad \tilde{g}^k = \Phi_k^T g^k, \quad \tilde{l}^k = \Phi_k^T l^k,$ and $\tilde{u}^k = \Phi_k^T u^k$.

3. Solve trust-region subproblem

$$\begin{array}{ll} \min & \frac{1}{2} s^T \tilde{H}_k s + \tilde{g}^k s \\ \text{subject to} & \tilde{l}^k \leq s \leq \tilde{u}^k \end{array}$$

4. Form the ratio $r = \frac{f(x^k) - f(x^k + \bar{s})}{q_k(\bar{s})}$ and update.

Trust Region Sub-Probroblem Reduction

Let $\quad w = s - \tilde{l}^k, h = \tilde{u}^k - \tilde{l}^k.$

$$\begin{array}{ll} \min & \frac{1}{2}s^T\tilde{H}_k s + \tilde{g}^k s \\ \text{s.t.} & \tilde{l}^k \le s \le \tilde{u}^k \end{array} \equiv \begin{array}{ll} \min & \frac{1}{2}w^T\tilde{H}w + k^T w \\ \text{s.t.} & 0 \le w \le h. \end{array}.$$

Since $\tilde{H} = \Phi^T H \Phi$ is positive definite, this QP is convex.

We solve using an interior point algorithm.

$$W = \operatorname{diag}(w)$$

Interior Point Newton Equations

$$\begin{aligned}
p &= z - v - k - t\,U^{-1}e + U^{-1}V\,h + t\,W^{-1}e \\
r &= (\tilde{H} + U^{-1}V + W^{-1}Z)^{-1}p, \\
\Delta w &= -w + r \\
\Delta u &= -w + h - u - \Delta w, \\
\Delta v &= t\,U^{-1}e - v - U^{-1}V\,\Delta s, \\
\Delta z &= -W^{-1}Z\,\Delta w + t\,W^{-1}e - z.
\end{aligned}$$

$t =$ homotopy parameter.

$(\tilde{H} + U^{-1}V + W^{-1}Z)^{-1}$

$$\tilde{H} = \Phi^T H \Phi = \Phi^T(\lambda I - \Psi\Gamma^{-1}\Psi^T)\Phi$$

$$(\tilde{H} + U^{-1}V + W^{-1}Z)^{-1}$$
$$=$$
$$G^{-1} + G^{-1}(\Phi^T\Psi)\left[\Gamma - (\Phi^T\Psi)^T G^{-1}(\Phi^T\Psi)\right]^{-1}(\Phi^T\Psi)^T G^{-1}$$

where

$$G = \lambda I + U^{-1}V + W^{-1}Z.$$

Triangular Factorization

Write

$$\Gamma - (\Phi_k^T \Psi)^T G^{-1} (\Phi_k^T \Psi) = \begin{bmatrix} -\hat{D} & \hat{L}^T \\ \hat{L} & \hat{W} \end{bmatrix}.$$

Compute Cholesky factors

$$\hat{D} = MM^T \quad \text{and} \quad \hat{W} + \hat{L}\hat{D}^{-1}\hat{L}^T = JJ^T.$$

Then

$$\Gamma - (\Phi_k^T \Psi)^T G^{-1} (\Phi_k^T \Psi) = \begin{bmatrix} M & 0 \\ -\hat{L}M^{-T} & J \end{bmatrix} \begin{bmatrix} -M^T & M^{-1}\hat{L}^T \\ 0 & J^T \end{bmatrix}.$$

Numerical Results

TEST SET: 23 problems from CUTEr set. $n \geq 1000$.

21 problems have dimension ≥ 10000.

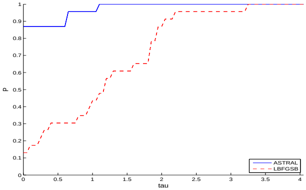
Termination Criteria: $f_{\text{best}} =$ best known function value.

1. $|f^k - f_{\text{best}}| / \max(1, |f_{\text{best}}|) \leq \epsilon$.

2. nf ≥ 1000.

Comparison of nfg between L-BFGS-B and ASTRAL

L-BFGS-B: Nocedal-Zhu-Byrd-Liu (1997)



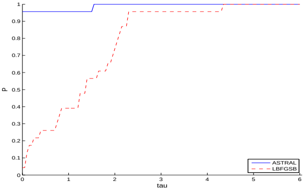


Figure 1a   Figure 1b

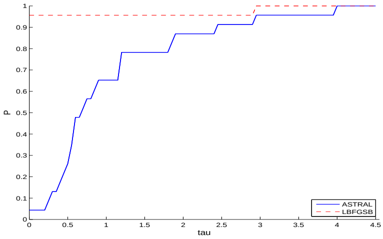
Figure 1a. Performance profiles, sum of the function and gradient evaluations, relative accuracy 10^{-5}. Figure 1b. Performance profiles, sum of the function and gradient evaluations, relative accuracy 10^{-3}.

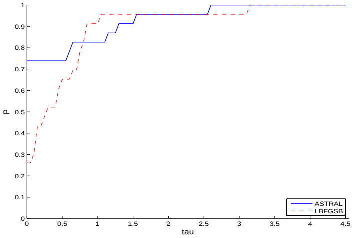
Comparison of CPU time

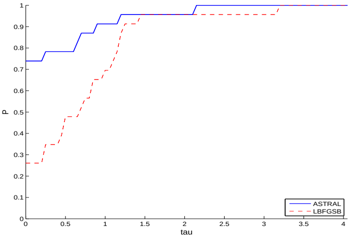
Note that the average cost of each function evaluation in our test set is 0.002s, and the average cost of each gradient evaluation is 0.008s.

Comparison of CPU time

Note that the average cost of each function evaluation in our test set is 0.002s, and the average cost of each gradient evaluation is 0.008s.







cpu(f) = 0.02s, cpu(g) = 0.08s

cpu(f) = 0.04s, cpu(g) = 0.16s

Thank You

Reference.

1. Limited Memory BFGS Updating in a Trust–Region Framework for Unconstrained Optimization, with James V. Burke, and Andreas Wiegmann.

2. ASTRAL: An Active Set l_∞-Trust-Region Algorithm for Box Constrained Optimization, with James V. Burke.

Literature Review:

Hager and Zhang(2006)

Zhang and Hager(2004)

Gilbert and Nocedal(1992)

Liu and Nocedal(1989)

Nash(1984)

Previous Work

- Hager and Zhang(2006)
- Birgin and Martínez(SPG, 2001, 2002)
- Lin and Moré(TRON, 1999)
- Coleman and Li(1994, 1996)
- Byrd, Lu, Norcedal, and Zhu(l-bfgsb, 1995, 1998)
- Conn, Gould, and Toint(1988)