Chapter 4: Trust-Region Methods

Outline

1. Outline of the Trust-Region Approach
2. Algorithms Based on the Cauchy Point
3. Global Convergence
4. Local Convergence of Trust-Region Newton Methods
5. Other Enhancements
6. Reference

Outline

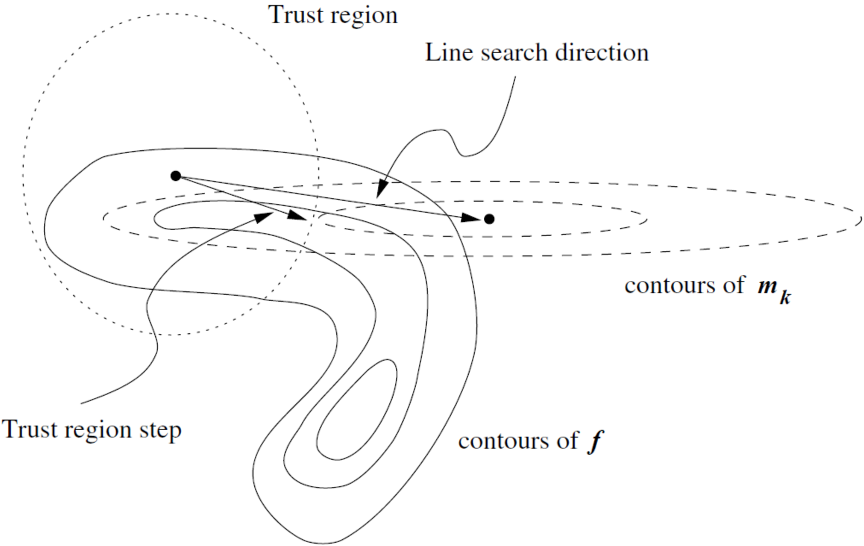
1. Outline of the Trust-Region Approach
2. Algorithms Based on the Cauchy Point
3. Global Convergence
4. Local Convergence of Trust-Region Newton Methods
5. Other Enhancements
6. Reference

Outline of the Trust-Region Approach

Line search methods and trust-region methods both generate steps with the help of a quadratic model of the objective function, but they use this model in different ways.

- Line search methods use it to generate a search direction p, and then focus their efforts on fining a suitable step length α along this direction.
- Trust-region methods choose the direction and length of the step simultaneously.
  - Define a region around the current iterate within which they *trust* the model to be an adequate representation of the objective function.
  - Then choose the step to be the approximate minimizer of the model in this region.
  - If a step is not acceptable, they reduce the size of the region and find a new minimizer.

Trust-region and line search steps



Model Function

- We assume that the model function m_k used at each iteration is quadratic.
- Recall the Taylor-series expansion of f around x_k, which is

$$f(x_k + p) = f_k + g_k^T p + \frac{1}{2} p^T \nabla^2 f(x_k + tp) p, \tag{1.1}$$

  where $f_k = f(x_k)$, $g_k = \nabla f(x_k)$ and $t \in (0, 1)$.
- By using B_k to approximate the Hessian in the second-order term, m_k is defined as follows:

$$m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p, \tag{1.2}$$

  where B_k is some symmetric matrix.
- The difference between $m_k(p)$ and $f(x_k + p)$ is $O(\|p\|^2)$, which is small when p is small.

Model Function

- When $B_k = \nabla^2 f(x_k)$, the approximation error in the model function m_k is $O(\|p\|^3)$, so this model is especially accurate when $\|p\|$ is small.

- In the other part, we emphasis the generality of the trust-region approach by assuming little about B_k except symmetry and uniform boundedness.

Trial Step

To obtain each step, we seek a solution of the trust-region subproblem

$$\min_{p \in \Re^n} m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p, \qquad \text{s.t.} \|p\| \leq \Delta_k, \tag{1.3}$$

where $\Delta_k > 0$ is the trust-region radius. Its solution s_k is called as trial step.

Classification of the Trust-Region Methods

The classification of trust-region methods are decided by the choice of B_k and norm for trust region in the model (1.3). For example,

- If $B_k = 0$ in (1.3) and define the trust region using the Euclidean norm, the trust-region method identifies with the steepest descent line search approach;
- If B_k is chosen to be the exact Hessian $\nabla^2 f_k$, the resulting approach is called the *trust-region Newton method*;
- If B_k is defined by means of a quasi-Newton approximation, we obtain a *trust-region quasi-Newton method*.

Trust-Region Size

The size of the trust region is critical to the effectiveness of each step.

- If the region is too small, the algorithm misses an opportunity to take a substantial step that will move it much closer to the minimizer of the objective function.
- If too large, the minimizer of the model may be far from the minimizer of the objective function in the region, so we may have to reduce the size of the region and try again.

Adaptive Trust-Region Adjustment

In practical algorithms, we choose the size of the region according to the performance of the algorithm during previous iterations.

- If the model is consistently reliable, producing good steps and accurately predicting the behavior of the objective function along these steps, the size of the trust region may be increased to allow longer, more ambitious, steps to be taken.
- A failed step is an indication that our model is an inadequate representation of the objective function over the current trust region. After such a step, we reduce the size of the region and try again.

Reduction Ratio

Define the reduction ratio

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(0) - m_k(s_k)} \tag{1.4}$$

the numerator is called the *actual reduction*, and the denominator is the *predicted reduction*. Note that the predicted reduction will always be positive, provided that x_k is not a stationary point. Thus

- if ρ_k is negative, the new objective value $f(x_k + s_k)$ is greater than the current value $f(x_k)$, so the step must be rejected.
- if ρ_k is close to 1, there is good agreement between the model m_k and the function f over this step, so it is safe to expand the trust region for the next iteration.
- If ρ_k is positive but not close to 1, we do not alter the trust region, but if it is close to zero or negative, we shrink the trust region.

Algorithm 1 (Framework of Trust Region Method).

```
Given $x_0 \in \mathbb{R}^n$, $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, and $\eta \in [0, \frac{1}{4}]$:
for $k = 0, 1, 2, \cdots$
    Obtain $s_k$ by (approximately) solving (1.3);
    Evaluate $\rho_k$ from (1.4);
    if $\rho_k < \frac{1}{4}$
        $$\Delta_{k+1} = \frac{1}{4}\Delta_k$$
    elseif $\rho_k > \frac{3}{4}$ and $\|s_k\| = \Delta_k$
        $$\Delta_{k+1} = \min(2\Delta_k, \hat{\Delta}_k)$$
    else
        $$\Delta_{k+1} = \Delta_k;$$
    if $\rho_k > \eta$
        $$x_{k+1} = x_k + s_k$$
    else
        $$x_{k+1} = x_k$$
;
```

Exact Solution of Trust-Region Subproblem

Theorem 1

The vector p^* is a global solution of the trust-region problem

$$\min_{p\in\Re^n} m(p) = f + g^T p + \frac{1}{2} p^T B p, \qquad \texttt{s.t.} \quad \|p\| \le \Delta \tag{1.5}$$

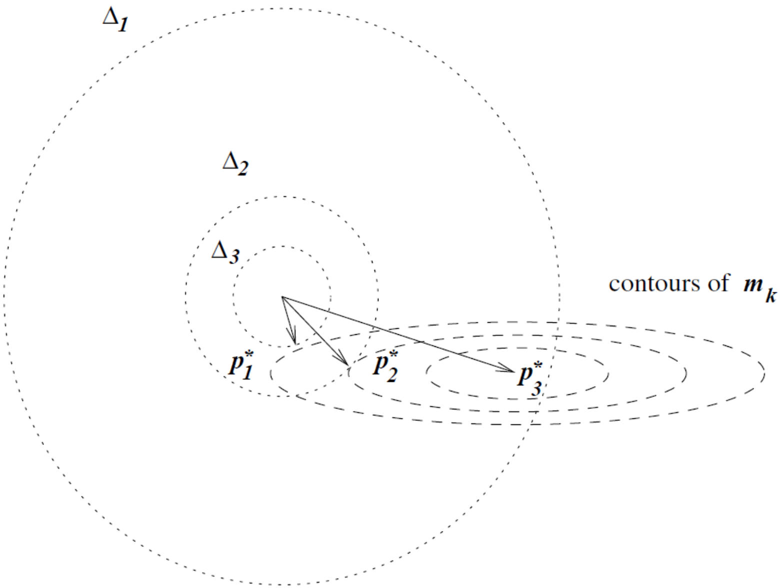
if and only if p^* is feasible and there is a scalar $\lambda \ge 0$ such that the following conditions are satisfied:

$$(B+\lambda I)p^* = -g, \tag{1.6a}$$

$$\lambda(\Delta - \|p^*\|) = 0, \tag{1.6b}$$

$$(B+\lambda I) \quad \textit{is positive semidefinite.} \tag{1.6c}$$

Solution of trust-region subproblem for different radii



Exact Solution of Trust-Region Subproblem

- If B_k is positive definite and $\|B^{-1}g\| \leq \Delta$, then $p^* = -B^{-1}g$.
- If B_k is not positive semidefinite, $\lambda > 0$ and $\|p^*\| = \Delta$
- In this case, find $\lambda > 0$ such that $B + \lambda I$ is positive definite and $\|(B + \lambda I)^{-1}g\| = \Delta$.
- Well-suited for smaller scale problems, but would be costly when the problem dimension is large

Outline

1. Outline of the Trust-Region Approach

2. Algorithms Based on the Cauchy Point

3. Global Convergence

4. Local Convergence of Trust-Region Newton Methods

5. Other Enhancements

6. Reference

The Cauchy Point

- Line search methods do not require optimal step lengths to be globally convergent. In fact, only a crude approximation to the optimal step length that satisfies certain loose criteria is needed.
- A similar situation applies in trust-region methods. Although in principle we are seeking the optimal solution of the subproblem (1.3), it is enough for global convergence purposes to find an approximate solution s_k that lies within the trust region and gives a sufficient reduction in the model.
- The sufficient reduction can be quantified in terms of the Cauchy point, which we denote by s_k^c.

The Cauchy Point

Algorithm 2 (Cauchy Point Calculation).

Find the vector s_k^c that solves

$$\begin{aligned} s_k^c = \text{arg}\min \quad & m_k(p) \\ \text{s.t.} \quad & p = -\tau \nabla f_k \\ & \|p\| \leq \Delta_k, \tau \geq 0. \end{aligned}$$

The Cauchy Point

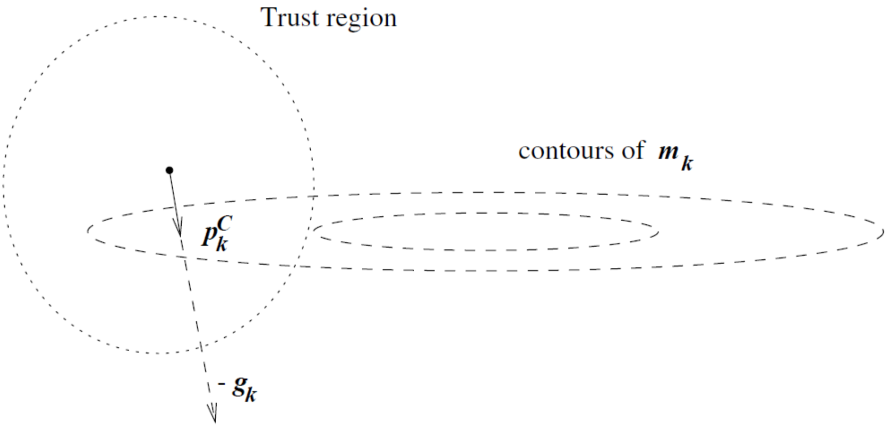
A closed-form definition of the Cauchy point can be written in the following

$$s_k^C = -\tau_k \frac{\Delta_k}{\|\nabla f_k\|} \nabla f_k, \tag{2.1}$$

where

$$\tau_k = \begin{cases} 1 & \text{if } \nabla f_k^T B_k \nabla f_k \leq 0; \\ \min\left(\dfrac{\|\nabla f_k\|^3}{\Delta_k \nabla f_k^T B_k \nabla f_k}, 1 \right) & \text{otherwise.} \end{cases} \tag{2.2a}$$

Cauchy point for a subproblem in which B_k is positive definite



The Cauchy Point

- The Cauchy step s_k^c is inexpensive to calculate, with no matrix factorizations being required.
- The model reduction obtained by the Cauchy point is

$$m_k(0) - m_k(s_k^c) \geq \frac{1}{2}\|\nabla f_k\| \min(\Delta_k, \frac{\|\nabla f_k\|}{\|B_k\|}).$$

- It plays a crucial importance in deciding if an approximate solution of the trust-region subproblem is acceptable.
- Specifically, a trust-region method will be globally convergent if its steps s_k attain a sufficient reduction in m_k; that is, they give a reduction in the model m_k that is at least some fixed multiple of the decrease attained by the Cauchy step at each iteration, namely, for some $c_2 \in (0, 1]$

$$m_k(0) - m_k(s_k) \geq c_2(m_k(0) - m_k(s_k^c)). \tag{2.3}$$

Improving on the Cauchy Point

Since the Cauchy point s_k^c provides sufficient reduction in the model function m_k to yield global convergence, and since the cost of calculating it is so small, why should we look any further for a better approximate solution of (1.3)?

- By always taking the Cauchy point as our step, we are simply implementing the steepest descent method with a particular choice of step length. Since steepest descent performs poorly even if an optimal step length is used at each iteration, to make the Trust Region algorithm efficient in practice, we have to improve on the Cauchy point.
- Note that B_k only affects the length of s_k. So it is reasonable to explore the potential benefit of B_k in determining the direction of s_k.

Improving on the Cauchy Point

Since we will be focusing on the internal workings of a single iteration, so we drop the subscript "k" from the quantities to simplify the notations. With this simplification, we restate the trust-region subproblem as follows:

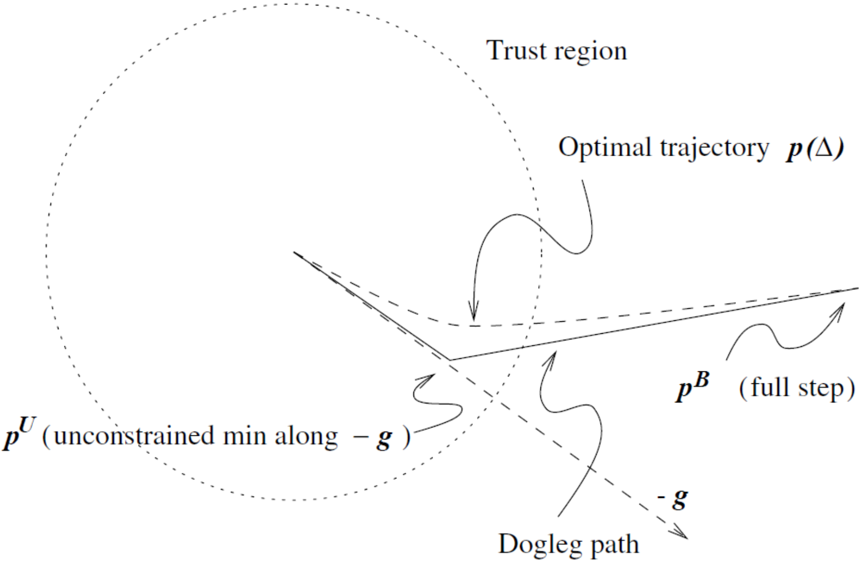
$$\min_{p \in \Re^n} m(p) = f + g^T p + \frac{1}{2} p^T B p, \qquad \text{s.t.} \|p\| \leq \Delta, \tag{2.4}$$

We denote the solution of above problem by $p^*(\Delta)$, to emphasize the dependence on Δ.

Improving on the Cauchy Point

- A number of algorithms for generating approximate solutions s_k to the trust-region problem (2.4) start by computing the Cauchy point and then try to improve on it.
- The improvement strategy is often designed so that the full step $p^B = -B^{-1}g$ is chosen whenever B is positive definite and $\|p^B\| \le \Delta$. When B_k is the exact Hessian or a quasi-Newton approximation, this strategy can be expected to yield superlinear convergence.

The Dogleg Method


Trust region
Optimal trajectory $\boldsymbol{p}(\Delta)$
$\boldsymbol{p}^B$ (full step)
$\boldsymbol{p}^U$ (unconstrained min along $-\boldsymbol{g}$)
$-\boldsymbol{g}$
Dogleg path

The Dogleg Method

$$\tilde{p}(\tau) = \begin{cases} \tau p^U, & 0 \le \tau \le 1, \\ p^U + (\tau - 1)(p^B - p^U), & 1 \le \tau \le 2. \end{cases} \tag{2.5a}$$

where

$$p^U = -\frac{g^T g}{g^T B g} g$$

is the unconstrained minimizer of $m(\cdot)$ along the steepest descent direction and

$$p^B = -B^{-1} g.$$

The Dogleg Method

The dogleg method chooses p to minimize the model m along this path, subject to the trust-region bound. In fact, it is not even necessary to carry out a search, because the dogleg path intersects the trust-region boundary at most once and the intersection point can be computed analytically.The following theorem proves these claims.

Theorem 2

Let B be positive definite. Then
(i) $\|\tilde{p}(\tau)\|$ is an increasing function of τ, and
(ii) $m(\tilde{p}(\tau))$ is a decreasing function of τ.

The Dogleg Method

- It follows from above theorem that the path $\tilde{p}(\tau)$ intersects the trust-region boundary $\|p\| = \Delta$ at exactly one point if $\|p^B\| \geq \Delta$, and nowhere otherwise.
- Since m is decreasing along the path, the chosen value of p will be at p^B if $\|p^B\| \leq \Delta$, otherwise at the point of intersection of the dogleg and the trust-region boundary.
- In the latter case, we compute the appropriate value of τ by solving the following scalar quadratic equation:

$$\|p^U + (\tau - 1)(p^B - p^U)\|^2 = \Delta^2.$$

Two-Dimensional Subspace Minimization

When B is positive definite, the dogleg method strategy can be made slightly more sophisticated by widening the search for p to the entire two-dimensional subspace spanned by p^U and p^B (equivalently, g and $-B^{-1}g$). The subproblem (2.4) is replaced by

$$\begin{aligned} \min_{p \in \Re^n} \quad & m(p) = f + g^T p + \frac{1}{2} p^T B p, \\ \text{s.t.} \quad & \|p\| \leq \Delta, \\ & p \in \operatorname{span}[g, B^{-1} g]. \end{aligned}$$

When B has negative eigenvalues, change the subspace to

$$\operatorname{span}[g, (B + \alpha I)^{-1} g]$$

for some $\alpha \in (-\lambda_1, -2\lambda_1]$. Here λ_1 denotes the most negative eigenvalues of B.

Outline

1. Outline of the Trust-Region Approach
2. Algorithms Based on the Cauchy Point
3. Global Convergence
4. Local Convergence of Trust-Region Newton Methods
5. Other Enhancements
6. Reference

Sufficient Reduction

Our first main result is that the dogleg and two dimensional subspace minimization algorithms produce approximate solution s_k of the subproblem (1.3) that satisfy the following estimate of decrease in the model function:

$$m_k(0) - m_k(s_k) \geq c_1 \|g_k\| \min(\Delta_k, \frac{\|g_k\|}{\|B_k\|}), \tag{3.1}$$

for some constant $c_1 \in (0, 1]$.

Theorem 3

Let s_k be any vector such that

$$\|s_k\| \leq \Delta_k \text{ and } m_k(0) - m_k(s_k) \geq c_2(m_k(0) - m_k(s_k^c)).$$

Then s_k satisfies (3.1) with $c_1 = c_2/2$.

Algorithm 3 (Trust Region Methods).

```
Given $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, and $\eta \geq 0$:
for $k = 0, 1, 2, \cdots$
    Calculate $s_k$;
    Evaluate $\rho_k$ from (1.4);
    if $\rho_k < \frac{1}{4}$
        $\Delta_{k+1} = \frac{1}{4}\Delta_k$
    elseif $\rho_k > \frac{3}{4}$ and $\|s_k\| = \Delta_k$
        $\Delta_{k+1} = \min(2\Delta_k, \hat{\Delta}_k)$
    else
        $\Delta_{k+1} = \Delta_k;$
    if $\rho_k > \eta$
        $x_{k+1} = x_k + s_k$
    else
        $x_{k+1} = x_k;$
end(for)
```

Convergence to Stationary Points

Global convergence results for trust-region methods come in two varieties, depending on whether we set the parameter η in the algorithm to zero or to some small positive value.

- when $\eta = 0$ (that is, the step is taken whenever it products a lower value of f), we can show that the sequence of gradients $\{\nabla f_k\}$ has an accumulation point zero, namely $\liminf \|\nabla f_k\| = 0$.
- For the more stringent acceptance test with $\eta > 0$, which requires the actual decrease in f to be at least some small fraction of the predicted decrease, we have the stronger result that any accumulation point of $\{\nabla f_k\}$ is zero, namely $\lim \|\nabla f_k\| = 0$.

We provide the global convergence results for both cases.

Convergence to Stationary Points

- We assume that the approximate Hessians B_k are bounded in norm, and that f is bounded below on the level set

$$S \equiv \{x | f(x) \leq f(x_0)\}. \tag{3.2}$$

For later reference, we define an open neighborhood of this set by

$$S(R_0) \equiv \{x | \|x - y\| < R_0 \text{ for some } y \in S\}.$$

where R_0 is a positive constant.

- To allow our results to be applied more generally, we also allow the length of the approximate solution s_k of (1.3) to exceed the trust-region bound, provided that it stays within some fixed multiple of the bound; that is,

$$\|s_k\| \leq \gamma \Delta_k, \qquad \text{for some constant } \gamma \geq 1. \tag{3.3}$$

Convergence to Stationary Points

The first result deals with the case $\eta = 0$.

Theorem 4

Let $\eta = 0$ in the trust region algorithm. Suppose that $\|B_k\| \le \beta$ for some constants β, that f is bounded below on the level set S defined by (3.2) and Lipschitz continuously differentiable in the neighborhood $S(R_0)$ for some $R_0 > 0$, and that all approximate solution of (1.3) satisfy the inequalities (3.1) and (3.3), for some positive constants c_1 and γ. We then have

$$\liminf_{k\to\infty} \|\nabla f_k\| = 0. \tag{3.4}$$

Proof Sketch

Proof Sketch: By contradiction. Assume that $\exists\, \epsilon > 0$ such that

$$\|\nabla f_k\| \geq \epsilon$$

for any k. Then sufficient decrease condition (3.1) indicates that

$$m_k(0) - m_k(s_k) \geq c_1 \|g_k\| \min(\Delta_k, \frac{\|g_k\|}{\|B_k\|}) \geq c_1 \epsilon \min(\Delta_k, \frac{\epsilon}{\beta}).$$

Then $\Delta_k \to 0$. On the other hand, since

$$|m_k(s_k) - f(x_k + s_k)| = \left| \frac{1}{2} s_k^T B_k s_k - \int_0^1 [g(x_k + ts_k) - g(x_k)]^T s_k \right| = O(\Delta_k^2)$$

we have

$$|\rho_k - 1| = \left| \frac{m(s_k) - f(x_k + s_k)}{m_k(0) - m_k(s_k)} \right| = O(\Delta_k) \to 0.$$

Therefore, $\Delta_{k+1} \geq \Delta_k$ which contradicts $\Delta_k \to 0$.

Convergence to Stationary Points

Theorem 5 (Schultz, Schnabel, and Byrd)

Let $\eta \in (0, \frac{1}{4})$ in the trust region algorithm. Suppose that $\|B_k\| \leq \beta$ for some constant β, that f is bounded below on the level set S (3.2) and Lipschitz continuously differentiable in $S(R_0)$ for some $R_0 > 0$, and that all approximate solutions s_k of (1.3) satisfy the inequalities (3.1) and (3.3) for some positive constants c_1 and γ. We then have

$$\lim_{k \to \infty} \nabla f_k = 0. \tag{3.5}$$

Outline

1 Outline of the Trust-Region Approach

2 Algorithms Based on the Cauchy Point

3 Global Convergence

4 Local Convergence of Trust-Region Newton Methods

5 Other Enhancements

6 Reference

Local Convergence of Trust-Region Newton Methods

Theorem 6

Let f be twice Lipschitz continuously differentiable in a neighborhood of a point x^* at which second-order sufficient conditions are satisfied. Suppose the sequence $\{x_k\}$ converges to x^* and that for all k sufficiently large, the trust-region algorithm based on (1.3) with $B_k = \nabla^2 f(x_k)$ chooses steps s_k that satisfy the Cauchy-point-based model reduction criterion (3.1) and are asymptotically similar to Newton steps s_k^N whenever $\|s_k^N\| \leq \frac{1}{2}\Delta_k$, that is,

$$\|s_k - s_k^N\| = o(\|s_k^N\|). \tag{4.1}$$

Then the trust-region bound Δ becomes inactive for all k sufficiently large and the sequence $\{x_k\}$ converges superlinearly to x^*.

It is immediate from the above theorem that if $s_k = s_k^N$ for all k sufficiently large, we have quadratic convergence of $\{x_k\}$ to x^*.

Outline

1. Outline of the Trust-Region Approach

2. Algorithms Based on the Cauchy Point

3. Global Convergence

4. Local Convergence of Trust-Region Newton Methods

5. Other Enhancements

6. Reference

Scaling

Optimization problems are often posed with poor scaling: the objective function f is highly sensitive to small changes in certain components of the vector x and relatively insensitive to changes in other components. Topologically, a symptom of poor scaling is that the minimizer x^* lies in a narrow valley, so that the contours of the objective $f(\cdot)$ near x^* tend towards highly eccentric ellipses. Algorithms can perform poorly unless they compensate for poor scaling.

Scaling

Recalling our definition of a trust region - a region around the current the current iterate within which the model $m_k(\cdot)$ is an adequate representation of the true objective $f(\cdot)$ - it is easy to see that a *spherical* trust region may not be approximate when f is poorly scaled.

- Even if the model Hessian B_k is exact, the rapid changes in f along certain directions probably will cause m_k to be a poor approximation to f along these directions.
- On the other hand, m_k may be a more reliable approximation to f along these directions in which f is changing more slowly.

Since the shape of our trust region should be such that our confidence in the model is more or less the same at all points on the boundary of the region, we are led naturally to consider *elliptical* trust regions in which the axes are short in the sensitive directions and longer in the less sensitive directions.

Scaling

- Elliptical trust regions can be defined by

$$\|Dp\| \leq \Delta, \tag{5.1}$$

where D is a diagonal matrix with positive diagonal elements, yielding the following scaled trust-region subproblem:

$$\min_{p \in \Re^n} m_k(p) \equiv f_k + g_k^T p + \frac{1}{2} p^T B_k p, \text{ s.t. } \|Dp\| \leq \Delta_k. \tag{5.2}$$

- When $f(x)$ is highly sensitive to the value of the ith component x_i, we set the corresponding diagonal element d_{ii} of D to be large, while d_{ii} is smaller for less-sensitive components.
- All algorithms discussed in this chapter can be modified for the case of elliptical trust regions, and the convergence theory continues to hold, with numerous superficial modifications

Trust Region in Other Norms

Trust regions may also be defined in terms of norms other than the Euclidean norm. For instance, we may have

$$\|p\|_1 \leq \Delta_k \text{ or } \|p\|_\infty \leq \Delta_k,$$

or their scaled counterparts

$$\|Dp\|_1 \leq \Delta_k \text{ or } \|Dp\|_\infty \leq \Delta_k.$$

Outline

1. Outline of the Trust-Region Approach
2. Algorithms Based on the Cauchy Point
3. Global Convergence
4. Local Convergence of Trust-Region Newton Methods
5. Other Enhancements
6. Reference

Reference

A.R. Conn, N.I.M. Gould and P.L. Toint, Trust-Region Methods, MPS-SIAM Series on Optimization, SIAM, 2000.

Thanks for your attention!