

Trust Region Method

Lectures for PHD course on
Numerical Optimization

Enrico Bertolazzi

DII – Università di Trento



Notes

- 1 The Trust Region method
- 2 Convergence analysis
- 3 The exact solution of trust region step
- 4 The dogleg trust region step
- 5 The double dogleg trust region step
- 6 Two dimensional subspace minimization



Notes

- Newton and quasi-Newton methods approximate a solution iteratively by choosing at each step a search direction and minimize in this direction.
- An alternative approach is to find a direction and a step-length, then if the step is successful in some sense the step is accepted. Otherwise another direction and step-length is chosen.
- The choice of the step-length and direction is algorithm dependent but a successful approach is the one based on trust region.



Notes

- Newton and quasi-Newton at each step (approximately) solve the minimization problem

$$\arg \min_{\mathbf{s}} m_k(\mathbf{s})$$

$$m_k(\mathbf{s}) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k) \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H}_k \mathbf{s}$$

in the case \mathbf{H}_k is symmetric and positive definite (SPD).

- If \mathbf{H}_k is SPD the minimum is

$$\mathbf{s} = -\mathbf{H}_k^{-1} \mathbf{g}_k, \quad \mathbf{g}_k = \nabla f(\mathbf{x}_k)^T$$

and \mathbf{s} is the quasi-Newton step.

- If $\mathbf{H}_k = \nabla^2 f(\mathbf{x}_k)$ and is SPD, then $\mathbf{s} = -\nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)^T$ is the Newton step.



Notes

- If \mathbf{H}_k is not positive definite, the search direction $-\mathbf{H}_k^{-1} \mathbf{g}_k$ may fail to be a descent direction and the previous minimization problem can have no solution.
- The problem is that the model $m_k(s)$ is an approximation of $f(\mathbf{x})$

$$m_k(s) \approx f(\mathbf{x}_k + s)$$

and this approximation is valid only in a small neighbors of \mathbf{x}_k .

- So that an alternative minimization problem is the following

$$\arg \min_s m_k(s) \quad \text{subject to } \|s\| \leq \Delta_k$$

Δ_k is the radius of the trust region of the model $m_k(s)$, i.e. the region where we trust the model is valid.



Notes

Algorithm (Generic trust region algorithm)

```

x assigned;  $\Delta$  assigned;
while  $\|\nabla f(x)\| > \epsilon$  do
  — setup the model
   $m(s) = f(x) + \nabla f(x)s + \frac{1}{2}s^T H s;$ 
  — compute the step
   $s \leftarrow \arg \min_{\|s\| \leq \Delta} m(s);$ 
   $x_{new} \leftarrow x + s;$ 
  — check the reduction
  if is  $x_{new}$  acceptable? then
     $x \leftarrow x_{new};$ 
    update  $\Delta;$ 
  else
    reduce  $\Delta;$ 
  end if
end while

```



Notes

When accept the step?

- The point \mathbf{x}_{new} in the previous algorithm can be accepted or rejected. The acceptance criterium can be the Armijo criterium of sufficient decrease

$$f(\mathbf{x}_{new}) \leq f(\mathbf{x}) + \beta_0 \nabla f(\mathbf{x})(\mathbf{x}_{new} - \mathbf{x})$$

where $\beta_0 \in (0, 1)$ is a small constant (typically 10^{-4}).

- Alternatively compute the expected and actual reduction with the ratio ρ :

$$p_{red} = m(\mathbf{0}) - m(\mathbf{s}), \quad a_{red} = f(\mathbf{x}) - f(\mathbf{x} + \mathbf{s}),$$

$$\rho = a_{red}/p_{red}$$

If the ratio ρ is near 1 the match of the model with the real function is good. We accept the step if $\rho > \beta_1$ where $\beta_1 \in (0, 1)$ normally $\beta_1 \approx 0.1$.



Notes

If the step is rejected how to reduce the trust radius ?

- We construct the parabola $p(t)$ such that ($s = x_{new} - x$)

$$p(0) = f(x), \quad p'(0) = \nabla f(x)s, \quad p(\Delta) = f(x_{new}),$$

the solution is

$$p(t) = f(x) + (\nabla f(x)s)t + Ct^2$$

$$C = \frac{f(x_{new}) - f(x) - (\nabla f(x)s)\Delta}{\Delta^2}$$

- The new radius is on the minimum of the parabola:

$$\Delta_{new} = -\frac{(\nabla f(x)s)}{2C} = \frac{\Delta^2(\nabla f(x)s)}{2[f(x) + (\nabla f(x)s)\Delta - f(x_{new})]}$$

- A safety interval is normally assumed; if the new radius is outside $[\Delta/10, \Delta/2]$ then it is put again in this interval.



Notes

If the step is accepted how to modify the trust radius ?

- Compute the expected and actual reduction

$$p_{red} = m(\mathbf{0}) - m(\mathbf{s})$$

$$a_{red} = f(\mathbf{x}) - f(\mathbf{x} + \mathbf{s})$$

- Compute the ratio of expected and actual reduction

$$\rho = \frac{a_{red}}{p_{red}}$$

- Compute the new radius

$$\Delta_{new} = \begin{cases} \max\{2\|\mathbf{s}\|, \Delta\} & \text{if } \rho \geq \beta_2 \\ \Delta & \text{if } \rho \in (\beta_1, \beta_2) \\ \|\mathbf{s}\| / \Delta & \text{if } \rho \leq \beta_1 \end{cases}$$



Notes

Algorithm (Check reduction algorithm)

CheckReduction(x, s, Δ);

$$x_{new} \leftarrow x + s$$

$$\alpha \leftarrow \nabla f(x)s$$

$$a_{red} \leftarrow f(x) - f(x_{new})$$

$$p_{red} \leftarrow -\alpha - s^T H s / 2$$

$$\rho \leftarrow a_{red} / p_{red}$$

$$r_{new} \leftarrow \begin{cases} \max\{2 \|s\|, r\} & \text{if } \rho \geq \beta_2 \\ r & \text{if } \rho \in (\beta_1, \beta_2) \\ \|s\| / 2 & \text{if } \rho \leq \beta_1 \end{cases}$$

if $\rho < \beta_1$ **then**
 — *reject the step*

$$x_{new} \leftarrow x$$

end if



Notes

Lemma

Consider the following constrained quadratic problem where $\mathbf{H} \in \mathbb{R}^{n \times n}$ *symmetric and positive definite*.

$$\begin{aligned} \text{Minimize} \quad & f(\mathbf{s}) = f_0 + \mathbf{g}^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H} \mathbf{s}, \\ \text{Subject to} \quad & \|\mathbf{s}\| \leq \Delta \end{aligned}$$

Then the following curve

$$\mathbf{s}(\mu) \doteq -(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g},$$

for any $\mu \geq 0$ defines a descent direction for $f(\mathbf{s})$. Moreover

- there exists a unique μ_* such that $\|\mathbf{s}(\mu_*)\| = \Delta$ and $\mathbf{s}(\mu_*)$ is the solution of the constrained problem;
- or $\|\mathbf{s}(0)\| < \Delta$ and $\mathbf{s}(0)$ is the solution of the constrained problem.



Notes

Proof:(Proof. (1/2)) If $\|s(0)\| \leq \Delta$ then $s(0)$ is the global minimum of $f(s)$ which is inside the trust region. Otherwise consider the Lagrangian

$$\mathcal{L}(s, \mu) = f_0 + \mathbf{g}^T s + \frac{1}{2} s^T \mathbf{H} s + \frac{1}{2} \mu (s^T s - \Delta^2),$$

Then we have

$$\frac{\partial \mathcal{L}}{\partial s}(s, \mu) = \mathbf{H} s + \mu s + \mathbf{g} = 0 \quad \Rightarrow \quad s = -(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g}$$

and $s^T s = \Delta^2$. Remember that if \mathbf{H} is SPD then $\mathbf{H} + \mu \mathbf{I}$ is SPD for all $\mu \geq 0$. Moreover the inverse of an SPD matrix is SPD. From

$$\mathbf{g}^T s = -\mathbf{g}^T (\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} < 0 \quad \text{for all } \mu \geq 0$$

follows that $s(\mu)$ is a descent direction for all $\mu \geq 0$. □



Notes

Proof:(Proof. (2/2)) To prove the uniqueness expand the gradient \mathbf{g} with the eigenvectors of \mathbf{H}

$$\mathbf{g} = \sum_{i=1}^n \alpha_i \mathbf{u}_i$$

\mathbf{H} is SPD so that \mathbf{u}_i can be chosen orthonormal. It follows

$$(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} = (\mathbf{H} + \mu \mathbf{I})^{-1} \sum_{i=1}^n \alpha_i \mathbf{u}_i = \sum_{i=1}^n \frac{\alpha_i}{\lambda_i + \mu} \mathbf{u}_i$$

$$\|(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g}\|^2 = \sum_{i=1}^n \frac{\alpha_i^2}{(\lambda_i + \mu)^2}$$

and $\|(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g}\|$ is a monotonically decreasing function of μ .

□



Notes

Remark

As a consequence of the previous Lemma we have:

- as the radius of the trust region becomes smaller as the scalar μ becomes larger. This means that the search direction become more and more oriented toward the gradient direction.
- as the radius of the trust region becomes larger as the scalar μ becomes smaller. This means that the search direction become more and more oriented toward the Newton direction.

Thus a trust region technique not only change the size of the step-length but also its direction. This results in a more robust numerical technique. The price to pay is that the solution of the minimization is more costly than the inexact line search.

but what happen when \mathbf{H} is not positive definite ?



Notes

Lemma

Consider the following constrained quadratic problem where $\mathbf{H} \in \mathbb{R}^{n \times n}$ is **symmetric** with $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ its eigenvalues.

$$\arg \min_{\|\mathbf{s}\| \leq \Delta} f(\mathbf{s}), \quad f(\mathbf{s}) = f_0 + \mathbf{g}^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H} \mathbf{s},$$

Then the following curve

$$\mathbf{s}(\mu) \doteq -(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g},$$

for any $\mu > -\lambda_1$ defines a descent direction for $f(\mathbf{s})$ and $\mathbf{H} + \mu \mathbf{I}$ is positive definite. Moreover

- or $\|\mathbf{s}(0)\| < \Delta$ with $\mathbf{g}^T \mathbf{s}(0) < 0$ and $\mathbf{s}(0)$ is **a local minima** of the problem;
- or there exists a $\mu_* > -\lambda_n$ such that $\|\mathbf{s}(\mu_*)\| = \Delta$ and $\mathbf{s}(\mu_*)$ is **a local minima** of the problem;



Notes

Proof:(Proof.

(1/6)) Consider the Lagrangian

$$\begin{aligned}\mathcal{L}(\mathbf{s}, \mu, \epsilon) = & f_0 + \mathbf{g}^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H} \mathbf{s} \\ & + \frac{1}{2} \mu (\mathbf{s}^T \mathbf{s} + \epsilon^2 - \Delta^2) + \omega (\mathbf{g}^T \mathbf{s} + \delta^2),\end{aligned}$$

where

$$\mathbf{s}^T \mathbf{s} + \epsilon^2 - \Delta^2$$

is the constraint $\|\mathbf{s}\| \leq \Delta$ on the length of the step and

$$\mathbf{g}^T \mathbf{s} + \delta^2$$

is the constraint $\mathbf{g}^T \mathbf{s} \leq 0$ on the step that must be descent □

Notes

Proof:(Proof. (2/6)) Then we must solve the nonlinear system:

$$\partial_s \mathcal{L}(s, \mu, \omega, \epsilon, \delta) = \mathbf{H}s + \mu s + (1 + \omega)\mathbf{g} = 0$$

$$2\partial_\mu \mathcal{L}(s, \mu, \omega, \epsilon, \delta) = s^T s + \epsilon^2 - \Delta^2 = 0$$

$$\partial_\omega \mathcal{L}(s, \mu, \omega, \epsilon, \delta) = \mathbf{g}^T s + \delta^2 = 0$$

$$\partial_\epsilon \mathcal{L}(s, \mu, \omega, \epsilon, \delta) = \mu\epsilon = 0$$

$$\partial_\delta \mathcal{L}(s, \mu, \omega, \epsilon, \delta) = 2\delta\omega = 0$$

from the first equation we have:

$$s = \frac{-1}{1 + \omega}(\mathbf{H} + \mu\mathbf{I})^{-1}\mathbf{g}$$

and if we want a descent direction $\mathbf{g}^T s < 0$ which imply $\omega = 0$.

□



Notes

Proof:(Proof. (3/6)) So that we must solve the reduced non linear system

$$\mathbf{s} = -(\mathbf{H} + \mu\mathbf{I})^{-1}\mathbf{g}$$

$$\mathbf{s}^T \mathbf{s} + \epsilon^2 - \Delta^2 = 0$$

$$\mathbf{g}^T \mathbf{s} = -\delta^2$$

$$\mu\epsilon = 0$$

combining the first and third equation we have

$$\mathbf{g}^T (\mathbf{H} + \mu\mathbf{I})^{-1} \mathbf{g} = \delta^2 \geq 0$$

□



Notes

Proof:(Proof. (4/6)) If $\epsilon \neq 0$ then we must have $\mu = 0$ and

$$\| -\mathbf{H}^{-1}\mathbf{g} \| = \|\mathbf{s}\| \leq \Delta$$

with $\mathbf{g}^T \mathbf{H}^{-1} \mathbf{g} \geq 0$. If $\epsilon = 0$ then we must have

$$\| -(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} \| = \|\mathbf{s}\| = \Delta$$

with $\mathbf{g}^T (\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} \geq 0$. Expand $\mathbf{g} = \sum_{i=1}^n \alpha_i \mathbf{u}_i$ with an orthonormal base of eigenvectors of \mathbf{H} it follows

$$\| (\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} \| = \sum_{i=1}^n \frac{\alpha_i^2}{(\lambda_i + \mu)^2}$$

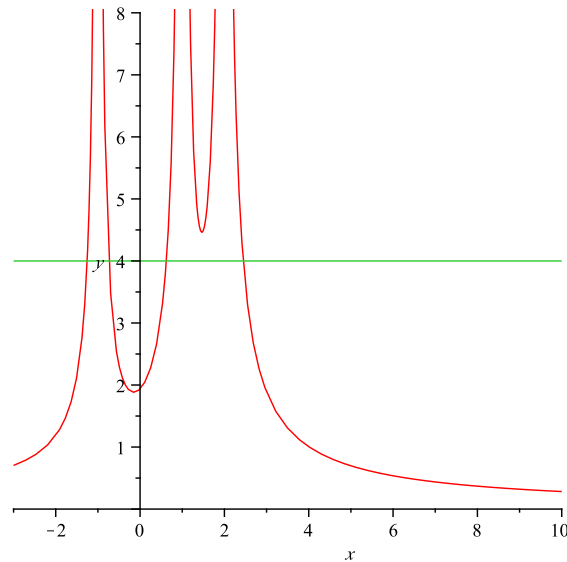
$$\mathbf{g}^T (\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} = \sum_{i=1}^n \frac{\alpha_i^2}{\lambda_i + \mu}$$



Notes

Proof:(Proof. (5/6)) $\|(\mathbf{H} + \mu\mathbf{I})^{-1}\mathbf{g}\|$ is a monotonically decreasing function of μ for $\mu > -\lambda_k$ where k is the first index such that $\alpha_k \neq 0$. For example

$$\|(\mathbf{H} + \mu\mathbf{I})^{-1}\mathbf{g}\| = (\mu + 1)^{-2} + 2(\mu - 1)^{-2} + 3(\mu - 2)^{-2}$$



Notes

Proof:(Proof.

(6/6)) Thus, or

$$\| -\mathbf{H}^{-1}\mathbf{g} \| = \|\mathbf{s}\| \leq \Delta \text{ with } \mathbf{g}^T \mathbf{H}^{-1} \mathbf{g} > 0.$$

or let be k the first index such that $\alpha_k \neq 0$, we can find a $\mu > -\lambda_k$ such that

$$\| -(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} \| = \sum_{i=k}^n \frac{\alpha_i^2}{(\lambda_i + \mu)^2} = \Delta$$

$$\mathbf{g}(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} = \sum_{i=k}^n \frac{\alpha_i^2}{\lambda_i + \mu} > 0$$

□



Notes

- 1 The Trust Region method
- 2 Convergence analysis**
- 3 The exact solution of trust region step
- 4 The dogleg trust region step
- 5 The double dogleg trust region step
- 6 Two dimensional subspace minimization



Notes

Algorithm (Basic trust region algorithm)

```

 $x_0$  assigned;  $\Delta_0$  assigned;  $k \leftarrow 0$ ;
while  $\|\nabla f(x_k)\| \neq 0$  do
     $m_k(s) = f(x_k) + \nabla f(x_k)s + \frac{1}{2}s^T H_k s$ ;           — setup the model
     $s_k \leftarrow \arg \min_{\|s\| \leq \Delta_k} m_k(s)$ ;           — compute the step
     $x_{k+1} \leftarrow x_k + s_k$ ;
     $\rho_k \leftarrow (f(x_k) - f(x_{k+1})) / (m_k(0) - m_k(s_k))$ ;
    — check the reduction
    if  $\rho_k > \beta_2$  then
         $\Delta_{k+1} \leftarrow 2\Delta_k$ ;                         — very successful
    else if  $\rho_k > \beta_1$  then
         $\Delta_{k+1} \leftarrow \Delta_k$ ;                         — successful
    else
         $\Delta_{k+1} \leftarrow \Delta_k/2$ ;  $x_{k+1} \leftarrow x_k$ ; — failure
    end if
     $k \leftarrow k + 1$ ;
end while

```



Notes

Cauchy point

Definition

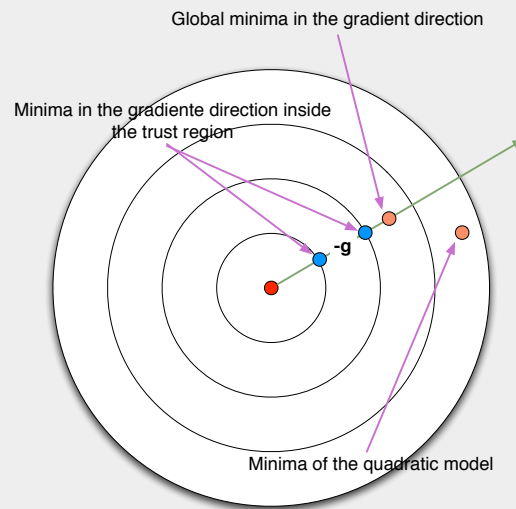
Consider the quadratic

$$m(\mathbf{s}) = f_0 + \mathbf{g}^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H} \mathbf{s}$$

and the minimization problem

$$\mathbf{s}^c(\Delta) = \arg \min_{\mathbf{s} \in \{-t\mathbf{g} \mid t \geq 0, \|\mathbf{s}\| \leq \Delta\}} m(\mathbf{s})$$

The point $\mathbf{s}^c(\Delta)$ is called *Cauchy point* or *step*.



Notes

Estimate the length of the Cauchy step

Lemma

For the Cauchy step the following characterization is valid:

$$\mathbf{s}^c(\Delta) = -\tau(\Delta) \frac{\mathbf{g}}{\|\mathbf{g}\|}$$

$$\tau(\Delta) = \begin{cases} \Delta & \text{if } \mathbf{g}^T \mathbf{H} \mathbf{g} \leq 0 \\ \min \left\{ \frac{\|\mathbf{g}\|^3}{\mathbf{g}^T \mathbf{H} \mathbf{g}}, \Delta \right\} & \text{if } \mathbf{g}^T \mathbf{H} \mathbf{g} > 0 \end{cases}$$

Moreover

$$\tau(\Delta) \geq \min \left\{ \frac{\|\mathbf{g}\|}{\varrho(\mathbf{H})}, \Delta \right\}$$

where $\varrho(\mathbf{H})$ is the spectral radius of \mathbf{H}



Notes

Proof: Consider

$$h(t) = m(-tg / \|g\|) = f_0 - t \|g\| + \frac{t^2}{2} \frac{g^T H g}{\|g\|^2}$$

$h(t)$ is a parabola in t and if $g^T H g \leq 0$ then the parabola decrease monotonically for $t \geq 0$. In this case the point is on the boundary of the trust region ($t = \Delta$).

If $g^T H g > 0$ the parabola is decreasing until the global mimima at

$$t = \frac{\|g\|^3}{g^T H g}$$

Otherwise we separate the case if the minimum of the parabola is inside or outside the trust region. □



Notes

Proof: Consider an orthonormal base of eigenvectors for \mathbf{H} and write \mathbf{g} in this coordinate:

$$\mathbf{g} = \sum_{i=1}^n \alpha_i \mathbf{u}_i$$

so that

$$\frac{\mathbf{g}^T \mathbf{H} \mathbf{g}}{\mathbf{g}^T \mathbf{g}} = \frac{\sum_{i=1}^n \lambda_i \alpha_i^2}{\sum_{i=1}^n \alpha_i^2} \leq \frac{\sum_{i=1}^n |\lambda_i| \alpha_i^2}{\sum_{i=1}^n \alpha_i^2} \leq \varrho(\mathbf{H})$$

and finally

$$\frac{\|\mathbf{g}\|^3}{\mathbf{g}^T \mathbf{H} \mathbf{g}} = \|\mathbf{g}\| \frac{\mathbf{g}^T \mathbf{g}}{\mathbf{g}^T \mathbf{H} \mathbf{g}} \geq \frac{\|\mathbf{g}\|}{\varrho(\mathbf{H})}$$



Notes

Estimate the reduction

Estimate the reduction obtained by the Cauchy step

In the convergence analysis is important to obtain estimation of the reduction of the function to be minimized.

A first step in this direction is the estimation of the reduction of the model quadratic function.

Lemma

Consider the quadratic

$$m(\mathbf{s}) = f_0 + \mathbf{g}^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H} \mathbf{s}$$

then for the Cauchy step we have:

$$m(\mathbf{0}) - m(\mathbf{s}^c(\Delta)) \geq \frac{1}{2} \|\mathbf{g}\| \min \left\{ \Delta, \frac{\|\mathbf{g}\|}{\varrho(\mathbf{H})} \right\}$$



Notes

Proof: Compute

$$m(\mathbf{0}) - m(\mathbf{s}^c(\Delta)) = \tau(\Delta) \|\mathbf{g}\| - \frac{\tau(\Delta)^2}{2 \|\mathbf{g}\|^2} \mathbf{g}^T \mathbf{H} \mathbf{g}$$

If $\mathbf{g}^T \mathbf{H} \mathbf{g} \leq 0$ for lemma on slide N.25 we have $\tau(\Delta) = \Delta$

$$\begin{aligned} m(\mathbf{0}) - m(\mathbf{s}^c(\Delta)) &= \Delta \|\mathbf{g}\| - \frac{\Delta^2}{2 \|\mathbf{g}\|^2} \mathbf{g}^T \mathbf{H} \mathbf{g} \\ &= \Delta \left(\|\mathbf{g}\| - \frac{\Delta \mathbf{g}^T \mathbf{H} \mathbf{g}}{2 \|\mathbf{g}\|^2} \right) \\ &\geq \Delta \|\mathbf{g}\| \end{aligned}$$



Notes

Proof: If $g^T H g >$ we have

$$\tau(\Delta) = \min \left\{ \| \mathbf{g} \|^3 / (\mathbf{g}^T \mathbf{H} \mathbf{g}), \quad \Delta \right\}$$

and

$$\begin{aligned} m(\mathbf{0}) - m(\mathbf{s}^c(\Delta)) &= \tau(\Delta) \left(\|\mathbf{g}\| - \frac{1}{2} \min \left\{ \|\mathbf{g}\|, \Delta \frac{\mathbf{g}^T \mathbf{H} \mathbf{g}}{\|\mathbf{g}\|^2} \right\} \right) \\ &\geq \tau(\Delta) \left(\|\mathbf{g}\| - \frac{1}{2} \|\mathbf{g}\| \right) \\ &\geq \tau(\Delta) \frac{1}{2} \|\mathbf{g}\| \end{aligned}$$

so that in general $m(\mathbf{0}) - m(\mathbf{s}^c(\Delta)) \geq \tau(\Delta) \frac{1}{2} \|\mathbf{g}\|$.



Notes

- A successful step in trust region algorithm imply that the ratio

$$\rho_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k)}{m_k(\mathbf{0}) - m_k(\mathbf{s}_k)}$$

is greater than a constant $\beta_1 > 0$.

- Any reasonable step in a trust region algorithm should be no (asymptotically) worse than a Cauchy step. So we require

$$m_k(\mathbf{0}) - m_k(\mathbf{s}_k) \geq \eta [m_k(\mathbf{0}) - m_k(\mathbf{s}^c(\Delta_k))]$$

for a constant $\eta > 0$.

- Using lemma on slide N.28

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k) &= \rho_k (m_k(\mathbf{0}) - m_k(\mathbf{s}_k)) \\ &\geq \rho_k \eta [m_k(\mathbf{0}) - m_k(\mathbf{s}^c(\Delta_k))] \\ &\geq \frac{\eta \beta_1}{2} \|\nabla f(\mathbf{x}_k)\| \min \left\{ \Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|}{\varrho(\mathbf{H}_k)} \right\} \end{aligned}$$



Notes

- Thus any reasonable trust region numerical scheme satisfy

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{\eta\beta_1}{2} \|\nabla f(\mathbf{x}_k)\| \min \left\{ \Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|}{\varrho(\mathbf{H}_k)} \right\}$$

for any successful step (for unsuccessful step $\mathbf{x}_{k+1} = \mathbf{x}_k$).

- Let \mathcal{S} the index set of successful step, then

$$f(\mathbf{x}_0) - \lim_{k \in \mathcal{S}} f(\mathbf{x}_k) \geq \frac{\eta\beta_1}{2} \sum_{k \in \mathcal{S}} \|\nabla f(\mathbf{x}_k)\| \min \left\{ \Delta_k, \frac{\|\nabla f(\mathbf{x}_k)\|}{\varrho(\mathbf{H}_k)} \right\}$$

thus we can use arguments similar to Zoutendijk theorem to prove convergence.

- To complete the argument we must set conditions that guarantees that $\Delta_k \not\rightarrow 0$ as $k \rightarrow \infty$ and that cardinality of \mathcal{S} is not finite.



Notes

Technical assumption

The following assumptions permits to characterize a class of convergent trust region algorithm.

Assumption

For any **successful** step in trust region algorithm, the ratio

$$\rho_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k)}{m_k(\mathbf{0}) - m_k(\mathbf{s}_k)}$$

is greater than a constant $\beta_1 > 0$.

Assumption

For any step in trust region algorithm, the model reduction for a constant $\eta > 0$ satisfy the inequality:

$$m_k(\mathbf{0}) - m_k(\mathbf{s}_k) \geq \eta [m_k(\mathbf{0}) - m_k(\mathbf{s}^c(\Delta_k))]$$



Notes

The following lemma permits to estimate the reduction ratio ρ_k and conclude that there exists a positive trust ray Δ_k for which the step is accepted!.

Lemma

Let be $f \in C^1(\mathbb{R}^n)$ with Lipschitz continuous gradient

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|$$

and apply basic trust region algorithm of slide N.23 with assumption of slide N.33 then we have

$$\Delta_k \geq \frac{(1 - \beta_2)\eta \|\nabla f(\mathbf{x}_k)\|}{2(\varrho(\mathbf{H}_k) + \gamma)}$$

for any accepted step.



Notes

Proof: By using Taylor's theorem

$$f(\mathbf{x}_k + \mathbf{s}_k) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k) \mathbf{s}_k + \int_0^1 [\nabla f(\mathbf{x}_k + t\mathbf{s}_k) - \nabla f(\mathbf{x}_k)] \mathbf{s}_k dt$$

so that

$$m_k(\mathbf{s}_k) - f(\mathbf{x}_k + \mathbf{s}_k) = (\mathbf{s}_k^T \mathbf{H}_k \mathbf{s}_k)/2 - \int_0^1 [\nabla f(\mathbf{x}_k + t\mathbf{s}_k) - \nabla f(\mathbf{x}_k)] \mathbf{s}_k dt$$

and

$$|m_k(\mathbf{s}_k) - f(\mathbf{x}_k + \mathbf{s}_k)| \leq \frac{\mathbf{s}_k^T \mathbf{H}_k \mathbf{s}_k}{2} + \frac{\gamma}{2} \|\mathbf{s}_k\|^2 \leq \frac{\varrho(\mathbf{H}_k) + \gamma}{2} \|\mathbf{s}_k\|^2$$

□



Notes

Proof: using these inequalities we can estimate the ratio

$$\begin{aligned}
 \left| \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k)}{m_k(\mathbf{0}) - m_k(\mathbf{s}_k)} - 1 \right| &= \frac{|m_k(\mathbf{s}_k) - f(\mathbf{x}_k + \mathbf{s}_k)|}{|m_k(\mathbf{0}) - m_k(\mathbf{s}_k)|} \\
 &\leq \frac{1}{2\eta} \frac{(\varrho(\mathbf{H}_k) + \gamma) \|\mathbf{s}_k\|^2}{|m_k(\mathbf{0}) - m_k(\mathbf{s}^c(\Delta))|} \\
 &\leq \frac{(\varrho(\mathbf{H}_k) + \gamma) \Delta^2}{\eta \|\nabla f(\mathbf{x}_k)\| \min \left\{ \Delta, \frac{\|\nabla f(\mathbf{x}_k)\|}{\varrho(\mathbf{H}_k)} \right\}}
 \end{aligned}$$

□



Notes

Proof: If $\Delta \leq \|\nabla f(\mathbf{x}_k)\| / \varrho(\mathbf{H}_k)$ we obtain

$$|\rho_k - 1| \leq \frac{(\varrho(\mathbf{H}_k) + \gamma)\Delta}{\eta \|\nabla f(\mathbf{x}_k)\|}$$

so that when $\Delta_k \leq \Delta$:

$$\Delta = \frac{(1 - \beta_2)\eta \|\nabla f(\mathbf{x}_k)\|}{(\varrho(\mathbf{H}_k) + \gamma)}$$

than $\rho_k \geq 1 - \beta_2$ and the step is accepted □



Notes

Corollary

Apply basic trust region algorithm of slide N.23 with assumption of slide N.33 to $f \in \mathcal{C}^1(\mathbb{R}^n)$ with Lipschitz continuous gradient

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|$$

then we have

$$f(\mathbf{x}_0) - \lim_{k \in \mathcal{S}} f(\mathbf{x}_k) \geq \frac{\eta^2 \beta_1 (1 - \beta_2)}{4} \sum_{k \in \mathcal{S}} \frac{\|\nabla f(\mathbf{x}_k)\|^2}{\varrho(\mathbf{H}_k) + \gamma}$$

moreover if $\varrho(\mathbf{H}_k) \leq C$ for all k we have

$$f(\mathbf{x}_0) - \lim_{k \in \mathcal{S}} f(\mathbf{x}_k) \geq \frac{\eta^2 \beta_1 (1 - \beta_2)}{4(C + \gamma)} \sum_{k \in \mathcal{S}} \|\nabla f(\mathbf{x}_k)\|^2$$



Notes

Convergence theorem

Theorem (Convergence to stationary points)

Apply basic trust region algorithm of slide N.23 with assumption of slide N.33 to $f \in \mathcal{C}^1(\mathbb{R}^n)$ with Lipschitz continuous gradient

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \gamma \|\mathbf{x} - \mathbf{y}\|$$

if the set

$$\mathcal{K} = \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$$

is compact and $\varrho(\mathbf{H}_k) \leq C$ for all k we have

$$\lim_{k \rightarrow \infty} \nabla f(\mathbf{x}_k) = \mathbf{0}$$

Proof: A trivial application of previous corollary. □



Notes

Convergence theorem

Theorem (Convergence to minima)

Apply basic trust region algorithm of slide N.23 with assumption of slide N.33 to $f \in \mathcal{C}^2(\mathbb{R}^n)$. If $\mathbf{H}_k = \nabla^2 f(\mathbf{x}_k)$ and the set

$$\mathcal{K} = \{\mathbf{x} \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$$

is compact then:

- 1 Or the iteration terminate at \mathbf{x}_k which satisfy second order necessary condition.
- 2 Or the limit point $\mathbf{x}_* = \lim_{k \rightarrow \infty} \mathbf{x}_k$ satisfy second order necessary condition.



J. J. Moré, D.C.Sorensen

Computing a Trust Region Step

SIAM J. Sci. Stat. Comput. **4**, No. 3, 1983



Notes

Constrained minimization problem

Solving the constrained minimization problem

As for the line-search problem we have many alternative for solving the constrained minimization problem:

- We can solve **accurately** the constrained minimization problem. For example by an iterative method.
- We can **approximate** the solution of the constrained minimization problem.

as for the line search the accurate solution of the constrained minimization problem is not paying while a good cheap approximations is normally better performing.



Notes

Outline

- 1 The Trust Region method
- 2 Convergence analysis
- 3 The exact solution of trust region step
- 4 The dogleg trust region step
- 5 The double dogleg trust region step
- 6 Two dimensional subspace minimization



Notes

- Consider the Lagrangian

$$\mathcal{L}(\mathbf{s}, \mu) = a + \mathbf{g}^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H} \mathbf{s} + \frac{1}{2} \mu (\mathbf{s}^T \mathbf{s} - \Delta^2),$$

where $a = f(\mathbf{x})$ and $\mathbf{g} = \nabla f(\mathbf{x})^T$.

- Then we can try to solve the nonlinear system

$$\frac{\partial \mathcal{L}}{\partial (\mathbf{s}, \mu)}(\mathbf{s}, \mu) = \begin{pmatrix} \mathbf{H} \mathbf{s} + \mu \mathbf{s} + \mathbf{g} \\ (\mathbf{s}^T \mathbf{s} - \Delta^2)/2 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}$$

- Using Newton method we have

$$\begin{pmatrix} \mathbf{s}_{k+1} \\ \mu_{k+1} \end{pmatrix} = \begin{pmatrix} \mathbf{s}_k \\ \mu_k \end{pmatrix} - \begin{pmatrix} \mathbf{H} + \mu \mathbf{I} & \mathbf{s} \\ \mathbf{s}^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{H} \mathbf{s}_k + \mu_k \mathbf{s}_k + \mathbf{g} \\ (\mathbf{s}_k^T \mathbf{s}_k - \Delta^2)/2 \end{pmatrix}$$



Notes

Lemma

let be $s(\mu)$ the solution of $(\mathbf{H} + \mu\mathbf{I})s(\mu) = -\mathbf{g}$ than we have

$$s'(\mu) = -(\mathbf{H} + \mu\mathbf{I})^{-1}\mathbf{s}(\mu) \quad \text{and} \quad s''(\mu) = 2(\mathbf{H} + \mu\mathbf{I})^{-2}\mathbf{s}(\mu)$$

Proof: It enough to differentiate the relation

$$\mathbf{H}s(\mu) + \mu\mathbf{s}(\mu) = \mathbf{g}$$

two times:

$$\mathbf{H}s'(\mu) + \mu s'(\mu) + s(\mu) = \mathbf{0}$$

$$\mathbf{H}s''(\mu) + \mu s''(\mu) + 2s'(\mu) = \mathbf{0}$$



Notes

- A better approach to compute μ is given by solving $\Phi(\mu) = 0$ where

$$\Phi(\mu) = \|\mathbf{s}(\mu)\| - \Delta, \quad \text{and} \quad \mathbf{s}(\mu) = -(\mathbf{H} + \mu\mathbf{I})^{-1}\mathbf{g}$$

- To build Newton method we need to evaluate

$$\Phi'(\mu) = \frac{\mathbf{s}(\mu)^T \mathbf{s}'(\mu)}{\|\mathbf{s}(\mu)\|}, \quad \mathbf{s}'(\mu) = -(\mathbf{H} + \mu\mathbf{I})^{-1}\mathbf{s}(\mu)$$

- Putting all in a Newton step we obtain

$$\mu_{k+1} = \mu_k + \frac{\Delta - \|\mathbf{s}(\mu_k)\|}{\mathbf{s}(\mu_k)^T \mathbf{s}'(\mu_k)} \|\mathbf{s}(\mu_k)\|$$



Notes

- Newton step can be reorganized as follows

$$\mathbf{a} = (\mathbf{H} + \mu_k \mathbf{I})^{-1} \mathbf{g}$$

$$\mathbf{b} = (\mathbf{H} + \mu_k \mathbf{I})^{-1} \mathbf{a}$$

$$\beta = \|\mathbf{a}\|$$

$$\mu_{k+1} = \mu_k + \beta \frac{\beta - \Delta}{\mathbf{a}^T \mathbf{b}}$$

- Thus Newton step require **two** linear system solution per step. However the coefficient matrix is the same so that only **one** LU factorization, thus the cost per step is essentially due to the LU factorization.



Notes

Lemma

If \mathbf{H} is SPD for all $\mu > 0$ we have:

$$\Phi'(\mu) < 0 \quad \text{and} \quad \Phi''(\mu) > 0$$

Proof: If $\mu > 0$ then $\mathbf{s}(\mu) \neq \mathbf{0}$. Evaluating $\Phi'(\mu)$ and using lemma of slide N.44 we have

$$\|\mathbf{s}(\mu)\| \Phi'(\mu) = \mathbf{s}(\mu)^T \mathbf{s}'(\mu) = -\mathbf{s}(\mu)^T (\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{s}(\mu) < 0$$

Evaluating $\Phi''(\mu)$ and using lemma of slide N.44 we have

$$\Phi''(\mu) = \frac{\mathbf{s}'(\mu)^T \mathbf{s}'(\mu) + \mathbf{s}(\mu)^T \mathbf{s}''(\mu)}{\|\mathbf{s}(\mu)\|} - \frac{(\mathbf{s}(\mu)^T \mathbf{s}'(\mu))^2}{\|\mathbf{s}(\mu)\|^3}$$



Notes

Proof: Using Cauchy–Schwartz inequality

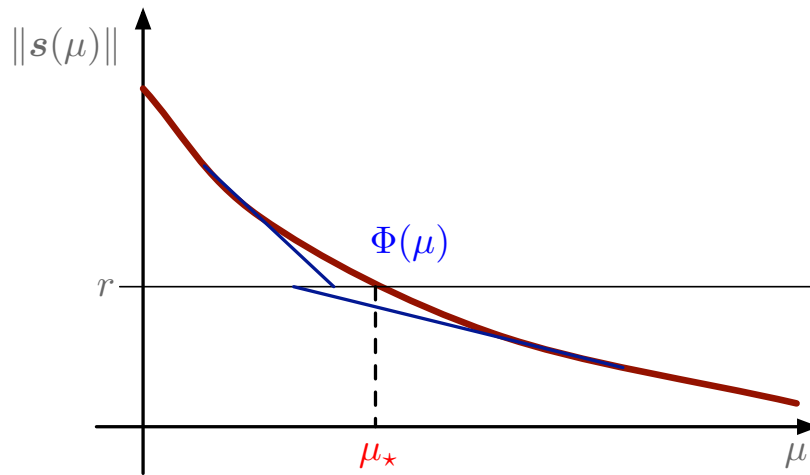
$$\begin{aligned}
 \Phi''(\mu) &\geq \frac{\mathbf{s}'(\mu)^T \mathbf{s}'(\mu) + \mathbf{s}(\mu)^T \mathbf{s}''(\mu)}{\|\mathbf{s}(\mu)\|} - \frac{\|\mathbf{s}(\mu)\|^2 \|\mathbf{s}'(\mu)\|^2}{\|\mathbf{s}(\mu)\|^3} \\
 &= \frac{\mathbf{s}(\mu)^T \mathbf{s}''(\mu)}{\|\mathbf{s}(\mu)\|} \\
 &= 2 \frac{\mathbf{s}(\mu)^T (\mathbf{H} + \mu \mathbf{I})^{-2} \mathbf{s}(\mu)}{\|\mathbf{s}(\mu)\|} > 0
 \end{aligned}$$

□



Notes

- From $\Phi''(\mu) > 0$ we have that Newton is monotonically convergent and steps underestimates μ .



Notes

- If we develop the vector \mathbf{g} with the orthonormal bases given by the eigenvectors of \mathbf{H} we have

$$\mathbf{g} = \sum_{i=1}^n \alpha_i \mathbf{u}_i$$

- Using this expression to evaluate $\mathbf{s}(\mu)$ we have

$$\mathbf{s}(\mu) = -(\mathbf{H} + \mu \mathbf{I})^{-1} \mathbf{g} = \sum_{i=1}^n \frac{\alpha_i}{\mu + \lambda_i} \mathbf{u}_i$$

$$\|\mathbf{s}(\mu)\| = \left(\sum_{i=1}^n \frac{\alpha_i^2}{(\mu + \lambda_i)^2} \right)^{1/2}$$

- This expression suggest to use as a model for $\Phi(\mu)$ the following expression

$$m_k(\mu) = \frac{\alpha_k}{\beta_k + \mu} - \Delta$$



Notes

- The model consists of **two** parameter α_k and β_k . To set this parameter we can impose

$$m_k(\mu_k) = \frac{\alpha_k}{\beta_k + \mu_k} - \Delta = \Phi(\mu_k)$$

$$m'_k(\mu_k) = -\frac{\alpha_k}{(\beta_k + \mu_k)^2} = \Phi'(\mu_k)$$

- solving for α_k and β_k we have

$$\alpha_k = -\frac{(\Phi(\mu_k) + \Delta)^2}{\Phi'(\mu_k)} \quad \beta_k = -\frac{\Phi(\mu_k) + \Delta}{\Phi'(\mu_k)} - \mu_k$$

where

$$\Phi(\mu_k) = \|\mathbf{s}(\mu_k)\| - \Delta \quad \Phi'(\mu_k) = -\frac{\mathbf{s}(\mu_k)^T (\mathbf{H} + \mu_k \mathbf{I})^{-1} \mathbf{s}(\mu_k)}{\|\mathbf{s}(\mu_k)\|^2}$$

- Having α_k and β_k it is possible to solve $m_k(\mu) = 0$ obtaining

$$\mu_{k+1} = \frac{\alpha_k}{\Delta} - \beta_k$$



Notes

- Substituting α_k and β_k the step become

$$\mu_{k+1} = \mu_k - \frac{\Phi(\mu_k)}{\Phi'(\mu_k)} - \frac{\Phi(\mu_k)^2}{\Phi'(\mu_k)\Delta} = \mu_k - \frac{\Phi(\mu_k)}{\Phi'(\mu_k)} \left(1 + \frac{\Phi(\mu_k)}{\Delta} \right)$$

- Comparing with the Newton step

$$\mu_{k+1} = \mu_k - \frac{\Phi(\mu_k)}{\Phi'(\mu_k)}$$

we see that this method perform larger step by a factor $1 + \Phi(\mu_k)\Delta^{-1}$.

- Notice that $1 + \Phi(\mu_k)\Delta^{-1}$ converge to 1 as $\mu_k \rightarrow \mu_*$. So that this iteration become the Newton iteration as μ_k becomes near the solution.



Notes

Algorithm (Exact trust region algorithm)

```

exact_trust_region( $\Delta, g, H$ )
 $\mu \leftarrow 0;$ 
 $s \leftarrow H^{-1}g;$ 
while  $\|s\| - \Delta > \epsilon$  and  $\mu \geq 0$  do
    — compute the model
     $s' \leftarrow -(H + \mu I)^{-1}s;$ 
     $\Phi \leftarrow \|s\| - \Delta;$ 
     $\Phi' \leftarrow (s^T s') / \|s\|$ 
    — update  $\mu$  and  $s$ 
     $\mu \leftarrow \mu - \frac{\Phi}{\Phi'} \frac{\|s\|}{\Delta};$ 
     $s \leftarrow -(H + \mu I)^{-1}g;$ 
end while
if  $\mu < 0$  then
     $s \leftarrow -H^{-1}g;$ 
end if

```



Notes

Outline

- 1 The Trust Region method
- 2 Convergence analysis
- 3 The exact solution of trust region step
- 4 The dogleg trust region step**
- 5 The double dogleg trust region step
- 6 Two dimensional subspace minimization



Notes

- The computation of the μ such that $\|s(\mu)\| = \Delta$ of the **exact** trust region computation can be very expensive.
- An alternative was proposed by Powell:



M.J.D. Powell

A hybrid method for nonlinear equations
in: Numerical Methods for Nonlinear Algebraic
Equations

ed. Ph. Rabinowitz, Gordon and Breach, pages 87-114,
1970.

where instead of computing exactly the curve $s(\mu)$ a
piecewise linear approximation $s_{dl}(\mu)$ is used in
computation.

- This approximation also permits to solve $\|s_{dl}(\mu)\| = \Delta$
explicitly.



Notes

- Form the definition of $s(\mu) = -(\mathbf{H} + \mu\mathbf{I})^{-1}\mathbf{g}$ and the relation $s'(\mu) = (\mathbf{H} + \mu\mathbf{I})^{-2}\mathbf{g}$ it follows

$$s(0) = -\mathbf{H}^{-1}\mathbf{g}, \quad \lim_{\mu \rightarrow \infty} \mu^2 s'(\mu) = -\mathbf{g}$$

i.e. the curve start from the Newton step and reduce to zero in the direction opposite to the gradient step.

- The direction $-\mathbf{g}$ is a descent direction, so that a first piece of the piecewise approximation should be a straight line from \mathbf{x} to the minimum of $m_k(-\lambda\mathbf{g})$. The minimum λ_* is found at

$$\lambda_* = \frac{\|\mathbf{g}\|^2}{\mathbf{g}^T \mathbf{H} \mathbf{g}}$$

- Having reached the minimum if the $-\mathbf{g}$ direction we can now go to the point $\mathbf{x} + s(0) = \mathbf{x} - \mathbf{H}^{-1}\mathbf{g}$ with another straight line.



Notes

- We denote by

$$s_g = -g \frac{\|g\|^2}{g^T H g}, \quad s_n = -H^{-1}g$$

respectively the step due to the unconstrained minimization in the gradient direction and in the Newton direction.

- The piecewise linear curve connecting $x + s_n$, $x + s_g$ and x is the **DogLeg** curve¹ $x_{dl}(\mu) = x + s_{dl}(\mu)$ where

$$s_{dl}(\mu) = \begin{cases} \mu s_g + (1 - \mu) s_n & \text{for } \mu \in [0, 1] \\ (2 - \mu) s_g & \text{for } \mu \in [1, 2] \end{cases}$$

¹notice that $s(\mu)$ is parametrized in the interval $[0, \infty]$ while $s_{dl}(\mu)$ is parametrized in the interval $[0, 2]$



Notes

Lemma (Kantorovich)

Let $A \in \mathbb{R}^{n \times n}$ an SPD matrix then the following inequality is valid

$$1 \leq \frac{(x^T A x)(x^T A^{-1} x)}{(x^T x)^2} \leq \frac{(M + m)^2}{4 M m}$$

for all $x \neq 0$. Where $m = \lambda_1$ is the smallest eigenvalue of A and $M = \lambda_n$ is the biggest eigenvalue of A .

this lemma can be improved a little bit for the first inequality

Lemma (Kantorovich (bis))

Let $A \in \mathbb{R}^{n \times n}$ an SPD matrix then the following inequality is valid

$$1 < \frac{(x^T A x)(x^T A^{-1} x)}{(x^T x)^2}$$

for all $x \neq 0$ and x not an eigenvector of A .



Notes

By using Kantorovich we can prove:

Lemma

We denote by

$$s_g = -g \frac{\|g\|^2}{g^T H g}, \quad s_n = -H^{-1}g, \quad \gamma_* = \frac{\|s_g\|^2}{s_n^T s_g}$$

then $\gamma_* \leq 1$, moreover if s_n is not parallel to s_g then $\gamma_* < 1$.

Proof: Using

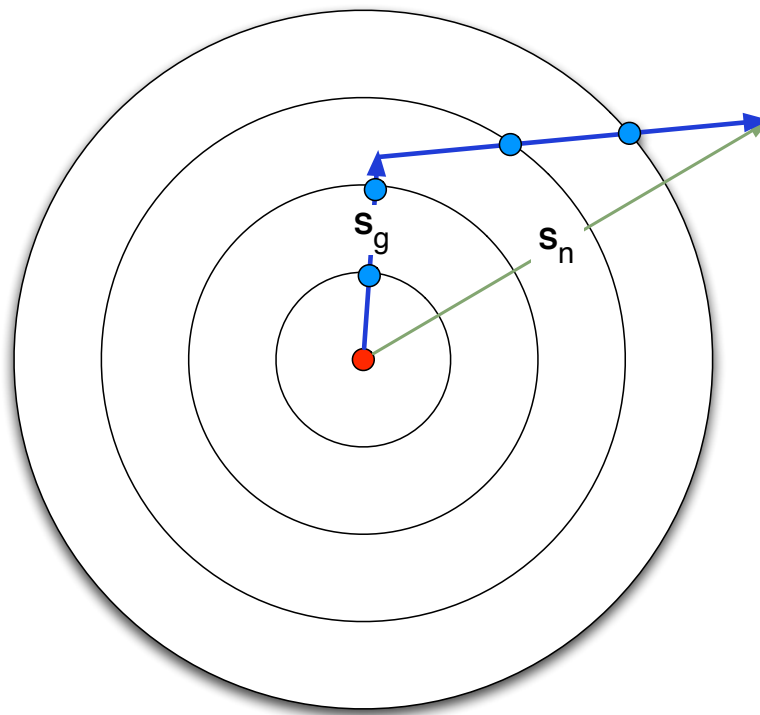
$$s_n^T s_g = \|g\|^2 \frac{g^T H^{-1} g}{g^T H g} \quad \text{and} \quad s_g^2 = \frac{\|g\|^6}{(g^T H g)^2}$$

we have $\gamma_* = \|g\|^4 / [(g^T H g)(g^T H^{-1} g)]$ and using Kantorovich inequality the lemma is proved. □



Notes

the Dogleg piecewise curve



Notes

Lemma

Consider the **dogleg** curve connecting $x + s_n$, $x + s_g$ and x . The curve can be expressed as $x_{dl}(\mu) = x + s_{dl}(\mu)$ where

$$s_{dl}(\mu) = \begin{cases} \mu s_g + (1 - \mu) s_n & \text{for } \mu \in [0, 1] \\ (2 - \mu) s_g & \text{for } \mu \in [1, 2] \end{cases}$$

for this curve if s_g is not parallel to s_n we have that the function

$$d(\mu) = \|x_{dl}(\mu) - x\| = \|s_{dl}(\mu)\|$$

is strictly monotone decreasing, moreover the direction $s_{dl}(\mu)$ is a descent direction for all $\mu \in [0, 2]$.



Notes

Proof: (Proof. (1/4)) In order to have a unique solution to the problem $\|s_{dl}(\mu)\| = \Delta$ we must have that $\|s_{dl}(\mu)\|$ is a monotone decreasing function:

$$\|s_{dl}(\mu)\|^2 = \begin{cases} \mu^2 s_g^2 + (1 - \mu)^2 s_n^2 + 2\mu(1 - \mu) s_g^T s_n & \mu \in [0, 1] \\ (2 - \mu)^2 s_g^2 & \mu \in [1, 2] \end{cases}$$

To check monotonicity we take first derivative

$$\begin{aligned} \frac{d}{d\mu} \|s_{dl}(\mu)\|^2 &= \begin{cases} 2\mu s_g^2 - 2(1 - \mu) s_n^2 + (2 - 4\mu) s_g^T s_n & \mu \in [0, 1] \\ (2\mu - 4) s_g^2 & \mu \in [1, 2] \end{cases} \\ &= \begin{cases} 2\mu(s_g^2 + s_n^2 - 2s_g^T s_n) - 2s_n^2 + 2s_g^T s_n & \mu \in [0, 1] \\ (2\mu - 4) s_g^2 & \mu \in [1, 2] \end{cases} \end{aligned}$$



Notes

Proof:(Proof. (2/4)) Notice that $(2\mu - 4) < 0$ for $\mu \in [1, 2]$ so that we need only to check that

$$2\mu(\mathbf{s}_g^2 + \mathbf{s}_n^2 - 2\mathbf{s}_g^T \mathbf{s}_n) - 2\mathbf{s}_n^2 + 2\mathbf{s}_g^T \mathbf{s}_n < 0 \quad \text{for } \mu \in [0, 1]$$

moreover

$$\mathbf{s}_g^2 + \mathbf{s}_n^2 - 2\mathbf{s}_g^T \mathbf{s}_n = \|\mathbf{s}_g - \mathbf{s}_n\|^2 \geq 0$$

Then it is enough to check the inequality for $\mu = 1$

$$2(\mathbf{s}_g^2 + \mathbf{s}_n^2 - 2\mathbf{s}_g^T \mathbf{s}_n) - 2\mathbf{s}_n^2 + 2\mathbf{s}_g^T \mathbf{s}_n = 2\mathbf{s}_g^2 - 2\mathbf{s}_g^T \mathbf{s}_n$$

i.e. we must check $\mathbf{s}_g^2 - \mathbf{s}_g^T \mathbf{s}_n < 0$. □



Notes

Proof:(Proof.

(3/4)) By using

$$\gamma_* = \frac{\|\mathbf{s}_g\|^2}{\mathbf{s}_n^T \mathbf{s}_g} < 1$$

of the previous lemma

$$\begin{aligned} \mathbf{s}_g^2 - \mathbf{s}_g^T \mathbf{s}_n &= \|\mathbf{s}_g\|^2 \left(1 - \frac{\mathbf{s}_n^T \mathbf{s}_g}{\|\mathbf{s}_g\|^2} \right) \\ &= \|\mathbf{s}_g\|^2 \left(1 - \frac{1}{\gamma_*} \right) < 0 \end{aligned}$$

□



Notes

Proof:(Proof. (4/4)) To prove that $s_{dl}(\mu)$ is a descent direction it is enough to notice that

- for $\mu \in [0, 1]$ the direction $s_{dl}(\mu)$ is a convex combination of s_g and s_n .
- for $\mu \in [1, 2)$ the direction $s_{dl}(\mu)$ is parallel to s_g .

so that it is enough to verify that s_g and s_n are descent direction. For s_g we have

$$s_g^T g = -\lambda_* g^T g < 0$$

For s_n we have

$$s_n^T g = -g^T H^{-1} g < 0$$

□



Notes

Using the previous Lemma we can prove

Lemma

If $\|s_{dl}(0)\| \geq \Delta$ then there is unique point $\mu \in [0, 2]$ such that $\|s_{dl}(\mu)\| = \Delta$.

Proof: It is enough to notice that $s_{dl}(2) = \mathbf{0}$ and that $\|s_{dl}(\mu)\|$ is strictly monotonically descendent. \square The approximate solution of the constrained minimization can be obtained by this simple algorithm

- 1 if $\Delta \leq \|s_g\|$ we set $s_{dl} = \Delta s_g / \|s_g\|$;
- 2 if $\Delta \leq \|s_n\|$ we set $s_{dl} = \alpha s_g + (1 - \alpha) s_n$; where α is the root in the interval $[0, 1]$ of:

$$\alpha^2 \|s_g\|^2 + (1 - \alpha)^2 \|s_n\|^2 + 2\alpha(1 - \alpha) s_g^T s_n = \Delta^2$$

- 3 if $\Delta > \|s_n\|$ we set $s_{dl} = s_n$;



Notes

Solving

$$\alpha^2 \|s_g\|^2 + (1 - \alpha)^2 \|s_n\|^2 + 2\alpha(1 - \alpha)s_g^T s_n = \Delta^2$$

we have that if $\|s_g\| \leq \Delta \leq \|s_n\|$ the root in $[0, 1]$ is given by:

$$\Delta = \|s_g\|^2 + \|s_n\|^2 - 2s_g^T s_n = \|s_g - s_n\|^2$$

$$\alpha = \frac{\|s_n\|^2 - s_g^T s_n - \sqrt{(s_g^T s_n)^2 - \|s_g\|^2 \|s_n\|^2 + \Delta^2 \Delta}}{\Delta}$$

to avoid cancellation the computation formula is the following

$$\begin{aligned} \alpha &= \frac{1}{\Delta} \frac{\|s_n\|^4 - 2s_g^T s_n \|s_n\|^2 + \|s_g\|^2 \|s_n\|^2 - \Delta^2 \Delta}{\|s_n\|^2 - s_g^T s_n + \sqrt{(s_g^T s_n)^2 - \|s_g\|^2 \|s_n\|^2 + \Delta^2 \Delta}} \\ &= \frac{\|s_n\|^2 - \Delta^2}{\|s_n\|^2 - s_g^T s_n + \sqrt{(s_g^T s_n)^2 - \|s_g\|^2 \|s_n\|^2 + \Delta^2 \|s_g - s_n\|^2}} \end{aligned}$$



Notes

Algorithm (Computing DogLeg step)

```

DoglegStep( $s_g, s_n, \Delta$ );
if  $\Delta \leq \|s_g\|$  then
     $s \leftarrow \Delta \frac{s_g}{\|s_g\|}$ ;
else if  $\Delta \geq \|s_n\|$  then
     $s \leftarrow s_n$ ;
else
     $a \leftarrow \|s_g\|^2$ ;
     $b \leftarrow \|s_n\|^2$ ;
     $c \leftarrow \|s_g - s_n\|^2$ ;
     $d \leftarrow (a + b - c)/2$ ;
     $\alpha \leftarrow \frac{b - \Delta^2}{b - d + \sqrt{d^2 - ab + \Delta^2 c}}$ ;
     $s \leftarrow \alpha s_g + (1 - \alpha) s_n$ ;
end if
return  $s$ ;

```



Notes

Outline

- 1 The Trust Region method
- 2 Convergence analysis
- 3 The exact solution of trust region step
- 4 The dogleg trust region step
- 5 The double dogleg trust region step
- 6 Two dimensional subspace minimization



Notes

The Double DogLeg approach

- We denote by

$$\mathbf{s}_g = -\mathbf{g} \frac{\|\mathbf{g}\|^2}{\mathbf{g}^T \mathbf{H} \mathbf{g}}, \quad \mathbf{s}_n = -\mathbf{H}^{-1} \mathbf{g}, \quad \gamma_* = \frac{\|\mathbf{s}_g\|^2}{\mathbf{s}_g^T \mathbf{s}_n}$$

respectively the step due to the unconstrained minimization in the gradient direction and in the Newton direction.

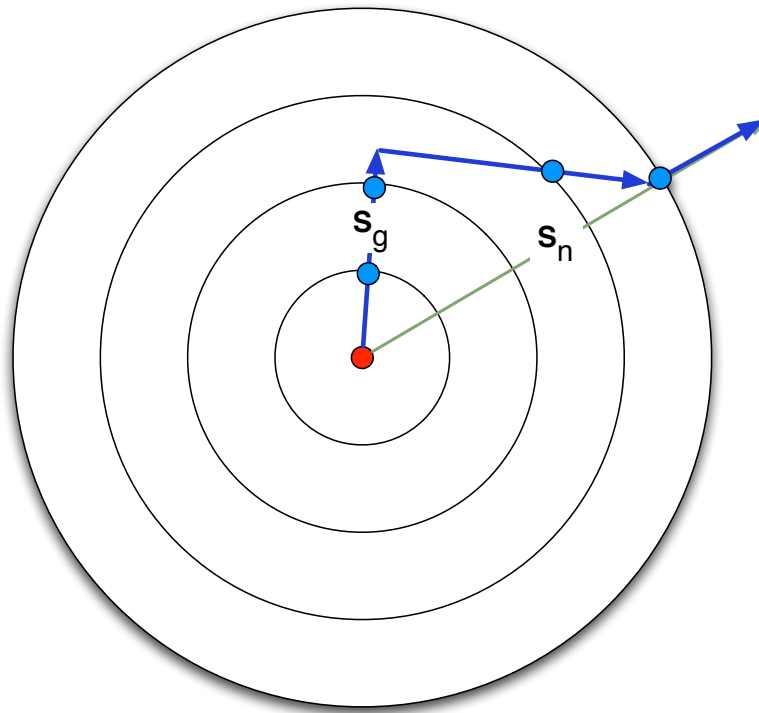
- The piecewise linear curve connecting $\mathbf{x} + \mathbf{s}_n$, $\mathbf{x} + \gamma_* \mathbf{s}_n$, $\mathbf{x} + \gamma_* \mathbf{s}_g$ and \mathbf{x} is the **Double Dogleg** curve $\mathbf{x}_{ddl}(\mu) = \mathbf{x} + \mathbf{s}_{ddl}(\mu)$ where

$$\mathbf{s}_{ddl}(\mu) = \begin{cases} (1 - \mu) \gamma_* \mathbf{s}_n & \text{for } \mu \in [0, 1] \\ (\mu - 1) \mathbf{s}_g + (2 - \mu) \gamma_* \mathbf{s}_n & \text{for } \mu \in [1, 2] \\ (3 - \mu) \mathbf{s}_g & \text{for } \mu \in [2, 3] \end{cases}$$



Notes

The Double Dogleg piecewise curve



Notes

Lemma

Consider the **double dogleg** curve connecting $x + s_n$, $x + \gamma_* s_n$, $x + s_g$ and x . The curve can be expressed as $x_{ddl}(\mu) = x + s_{ddl}(\mu)$ where

$$s_{ddl}(\mu) = \begin{cases} (1 - \mu)\gamma_* s_n & \text{for } \mu \in [0, 1] \\ (\mu - 1)s_g + (2 - \mu)\gamma_* s_n & \text{for } \mu \in [1, 2] \\ (3 - \mu)s_g & \text{for } \mu \in [2, 3] \end{cases}$$

for this curve if s_g is not parallel to s_n we have that the function

$$d(\mu) = \|s_{ddl}(\mu)\|$$

is strictly monotone decreasing, moreover the direction $s_{ddl}(\mu)$ is a descent direction for all $\mu \in [0, 3]$.



Notes

Proof:(Proof. (1/2)) In order to have a unique solution to the problem $\|\mathbf{s}_{ddl}(\mu)\| = \Delta$ we must have that $\|\mathbf{s}_{ddl}(\mu)\|$ is a monotone decreasing function. It is enough to prove for $\mu \in [1, 2]$:

$$\|\mathbf{s}_{ddl}(1 + \alpha)\|^2 = \alpha^2 \mathbf{s}_g^2 + (1 - \alpha)^2 \gamma_*^2 \mathbf{s}_n^2 + 2\alpha(1 - \alpha) \gamma_* \mathbf{s}_g^T \mathbf{s}_n$$

To check monotonicity we take first derivative

$$\begin{aligned} \frac{d}{d\alpha} \|\mathbf{s}_{ddl}(1 + \alpha)\|^2 &= 2\alpha \mathbf{s}_g^2 - 2(1 - \alpha) \gamma_*^2 \mathbf{s}_n^2 + (2 - 4\alpha) \gamma_* \mathbf{s}_g^T \mathbf{s}_n \\ &= 2\alpha (\mathbf{s}_g^2 + \gamma_*^2 \mathbf{s}_n^2 - 2\gamma_* \mathbf{s}_g^T \mathbf{s}_n) - 2\gamma_*^2 \mathbf{s}_n^2 + 2\gamma_* \mathbf{s}_g^T \mathbf{s}_n \end{aligned}$$

□



Notes

Proof:(Proof. (2/2)) Notice that

$$\mathbf{s}_g^2 + \gamma_*^2 \mathbf{s}_n^2 - 2\gamma_* \mathbf{s}_g^T \mathbf{s}_n = \|\mathbf{s}_g - \gamma_* \mathbf{s}_n\|^2 > 0$$

because \mathbf{s}_g and \mathbf{s}_n are not parallel. Then it is enough to check the inequality for $\alpha = 1$

$$\begin{aligned} 2(\mathbf{s}_g^2 + \gamma_*^2 \mathbf{s}_n^2 - 2\gamma_* \mathbf{s}_g^T \mathbf{s}_n) - 2\gamma_*^2 \mathbf{s}_n^2 + 2\gamma_* \mathbf{s}_g^T \mathbf{s}_n &= 2\mathbf{s}_g^2 - 2\gamma_* \mathbf{s}_g^T \mathbf{s}_n \\ &= 0 \end{aligned}$$

The rest of the proof is similar as for the single dogleg step. \square



Notes

Using the previous Lemma we can prove

Lemma

If $\|s_{ddl}(0)\| \geq \Delta$ then there is unique point $\mu \in [0, 3]$ such that $\|s_{ddl}(\mu)\| = \Delta$.

The approximate solution of the constrained minimization can be obtained by this simple algorithm

- 1 if $\Delta \leq \|s_g\|$ we set $s_{ddl} = \Delta s_g / \|s_g\|$;
- 2 if $\Delta \leq \gamma_* \|s_n\|$ we set $s_{ddl} = \alpha s_g + (1 - \alpha) \gamma_* s_n$; where α is the root in the interval $[0, 1]$ of:

$$\alpha^2 \|s_g\|^2 + \gamma_*^2 (1 - \alpha)^2 \|s_n\|^2 + 2\gamma_* \alpha (1 - \alpha) s_g^T s_n = \Delta^2$$

- 3 if $\Delta \leq \|s_n\|$ we set $s_{ddl} = \Delta s_n / \|s_n\|$;
- 4 if $\Delta > \|s_n\|$ we set $s_{ddl} = s_n$;



Notes

Solving

$$\alpha^2 \|s_g\|^2 + \gamma_*^2 (1 - \alpha)^2 \|s_n\|^2 + 2\gamma_* \alpha (1 - \alpha) s_g^T s_n = \Delta^2$$

we have that if $\|s_g\| \leq \Delta \leq \gamma_* \|s_n\|$ the root in $[0, 1]$ is given by:

$$A = \gamma_*^2 \|s_n\|^2 - \|s_g\|^2$$

$$B = \Delta^2 - \|s_g\|^2$$

$$\alpha = \frac{A - B}{A + \sqrt{AB}}$$



Notes

Algorithm (Computing Double DogLeg step)

DoubleDoglegStep(s_g, s_n, Δ);

$\gamma_* \leftarrow \|s_g\|^2 / (s_g^T s_n);$

if $\Delta \leq \|s_g\|$ **then**

$s \leftarrow \Delta s_g / \|s_g\|;$

else if $\Delta \leq \gamma_* \|s_n\|$ **then**

$A \leftarrow \gamma_*^2 \|s_n\|^2 - \|s_g\|^2;$

$B \leftarrow \Delta^2 - \|s_g\|^2;$

$\alpha \leftarrow (A - B) / (A + \sqrt{AB});$

$s \leftarrow \alpha s_g + (1 - \alpha) s_n;$

else if $\Delta \leq \|s_n\|$ **then**

$s \leftarrow \Delta s_n / \|s_n\|;$

else

$s \leftarrow s_n;$

end if

return $s;$



Notes

Outline

- 1 The Trust Region method
- 2 Convergence analysis
- 3 The exact solution of trust region step
- 4 The dogleg trust region step
- 5 The double dogleg trust region step
- 6 Two dimensional subspace minimization



Notes

Two dimensional subspace minimization

- When \mathbf{H} is positive definite the dogleg step can be improved by widening the search subspace

$$\mathbf{s} = \arg \min_{\|\alpha \mathbf{s}_g + \beta \mathbf{s}_n\| \leq \Delta} f(\alpha \mathbf{s}_g + \beta \mathbf{s}_n)$$

i.e. we must solve a two dimensional constrained problem.

- The 2D problem results:

$$\begin{aligned} f(\alpha \mathbf{s}_g + \beta \mathbf{s}_n) &= f_0 + \mathbf{g}^T (\alpha \mathbf{s}_g + \beta \mathbf{s}_n) \\ &+ \frac{1}{2} (\alpha \mathbf{s}_g + \beta \mathbf{s}_n)^T \mathbf{H} (\alpha \mathbf{s}_g + \beta \mathbf{s}_n) \\ &= f_0 + \alpha \mathbf{g}^T \mathbf{s}_g + \beta \mathbf{g}^T \mathbf{s}_n \\ &+ \frac{1}{2} \alpha^2 \mathbf{s}_g^T \mathbf{H} \mathbf{s}_g + \frac{1}{2} \beta^2 \mathbf{s}_n^T \mathbf{H} \mathbf{s}_n + \alpha \beta \mathbf{s}_g^T \mathbf{H} \mathbf{s}_n \end{aligned}$$



Notes

Two dimensional subspace minimization

The 2D problem written in matrix form:

$$f(\alpha, \beta) = f_0 + \mathbf{b}^T \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \frac{1}{2} (\alpha \quad \beta) \mathbf{A} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} \mathbf{g}^T \mathbf{s}_g \\ \mathbf{g}^T \mathbf{s}_n \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{s}_g^T \mathbf{H} \mathbf{s}_g & \mathbf{s}_g^T \mathbf{H} \mathbf{s}_n \\ \mathbf{s}_n^T \mathbf{H} \mathbf{s}_g & \mathbf{s}_n^T \mathbf{H} \mathbf{s}_n \end{pmatrix}$$

and the constraint

$$\|\alpha \mathbf{s}_g + \beta \mathbf{s}_n\|^2 = (\alpha \quad \beta) \mathbf{D} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

$$\mathbf{D} = \begin{pmatrix} \mathbf{s}_g^T \mathbf{s}_g & \mathbf{s}_g^T \mathbf{s}_n \\ \mathbf{s}_n^T \mathbf{s}_g & \mathbf{s}_n^T \mathbf{s}_n \end{pmatrix}$$



Notes

Lemma

Consider the following constrained quadratic problem where $\mathbf{H} \in \mathbb{R}^{n \times n}$, $\mathbf{D} \in \mathbb{R}^{n \times n}$ are *symmetric and positive definite*.

$$\text{Minimize} \quad f(\mathbf{s}) = f_0 + \mathbf{g}^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \mathbf{H} \mathbf{s},$$

$$\text{Subject to} \quad \mathbf{s}^T \mathbf{D} \mathbf{s} \leq r^2$$

Then the following curve

$$\mathbf{s}(\mu) \doteq -(\mathbf{H} + \mu \mathbf{D})^{-1} \mathbf{g},$$

for any $\mu \geq 0$ defines a descent direction for $f(\mathbf{s})$. Moreover

- there exists a unique μ_* such that $\|\mathbf{s}(\mu_*)\| = \Delta$ and $\mathbf{s}(\mu_*)$ is the solution of the constrained problem;
- or $\|\mathbf{s}(0)\| < \Delta$ and $\mathbf{s}(0)$ is the solution of the constrained problem.



Notes

References



Jorge Nocedal, and Stephen J. Wright

Numerical optimization

Springer, 2006



J. Stoer and R. Bulirsch

Introduction to numerical analysis

Springer-Verlag, Texts in Applied Mathematics, **12**, 2002.



J. E. Dennis, Jr. and Robert B. Schnabel

Numerical Methods for Unconstrained Optimization and
Nonlinear Equations

SIAM, Classics in Applied Mathematics, **16**, 1996.



Notes
