Trust Region Method

Lectures for PHD course on
Unconstrained Numerical Optimization

Enrico Bertolazzi

DIMS – Università di Trento

May 2008

Outline

- 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 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.

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

$$\arg\min_{\boldsymbol{s}} \quad m_k(\boldsymbol{s})$$

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

  in the case $\boldsymbol{H}_k$ is symmetric and positive definite (SPD).
- If $\boldsymbol{H}_k$ is SPD the minimum is

$$\boldsymbol{s} = -\boldsymbol{H}_k^{-1}\boldsymbol{g}_k, \qquad \boldsymbol{g}_k = \nabla\mathsf{f}(\boldsymbol{x}_k)^T$$

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

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

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

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

- So that an alternative minimization problem is the following

$$\arg\min_{\boldsymbol{s}} \ m_k(\boldsymbol{s}) \qquad \text{subject to } \|\boldsymbol{s}\| \leq \Delta_k$$

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

Algorithm (Generic trust region algorithm)

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

When accept the step?

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

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

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

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

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

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

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 $\beta_1 \approx 0.1$.

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

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

$$p(0) = f(\boldsymbol{x}), \qquad p'(0) = \nabla f(\boldsymbol{x})\boldsymbol{s}, \qquad p(\Delta) = f(\boldsymbol{x}_{new}),$$

the solution is

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

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

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

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

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

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

- Compute the expected and actual reduction

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

- Compute the ratio of expected and actual reduction

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

- Compute the new radius

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

Algorithm (Check reduction algorithm)

CheckReduction$(\boldsymbol{x}, \boldsymbol{s}, \Delta)$;
$\boldsymbol{x}_{new} \leftarrow \boldsymbol{x}+\boldsymbol{s}$
$\alpha \leftarrow \nabla f(\boldsymbol{x})\boldsymbol{s}$
$a_{red} \leftarrow f(\boldsymbol{x}) - f(\boldsymbol{x}_{new})$
$p_{red} \leftarrow -\alpha - \boldsymbol{s}^T\boldsymbol{H}\boldsymbol{s}/2$
$\rho \leftarrow a_{red}/p_{red}$

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

if $\rho < \beta_1$ **then**
  — reject the step
  $\boldsymbol{x}_{new} \leftarrow \boldsymbol{x}$
end if

Lemma

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

$$\textit{Minimize} \quad f(\boldsymbol{s}) = f_0 + \boldsymbol{g}^T\boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^T\boldsymbol{H}\boldsymbol{s},$$
$$\textit{Subject to} \quad \|\boldsymbol{s}\| \leq \Delta$$

Then the following curve

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

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

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

Proof. (1/2).

If $\|\boldsymbol{s}(0)\| \leq \Delta$ then $\boldsymbol{s}(0)$ is the global minimum of $f(\boldsymbol{s})$ which is inside the trust region. Otherwise consider the Lagrangian

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

Then we have

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

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

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

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

Proof. (2/2).

To prove the uniqueness expand the gradient $\boldsymbol{g}$ with the eigenvectors of $\boldsymbol{H}$

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

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

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

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

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

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 $\boldsymbol{H}$ is not positive definite ?

Lemma

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

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

Then the following curve

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

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

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

Proof. (1/6).

Consider the Lagrangian

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

where

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

is the constraint $\|\boldsymbol{s}\|\leq\Delta^2$ on the length of the step and

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

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

Proof. (2/6).

Then we must solve the nonlinear system:

$$\begin{aligned}
\partial_{\boldsymbol{s}}\mathcal{L}(\boldsymbol{s},\mu,\omega,\epsilon,\delta) &= \boldsymbol{H}\boldsymbol{s}+\mu\boldsymbol{s}+(1+\omega)\boldsymbol{g}=0\\
2\partial_{\mu}\mathcal{L}(\boldsymbol{s},\mu,\omega,\epsilon,\delta) &= \boldsymbol{s}^T\boldsymbol{s}+\epsilon^2-\Delta^2=0\\
\partial_{\omega}\mathcal{L}(\boldsymbol{s},\mu,\omega,\epsilon,\delta) &= \boldsymbol{g}^T\boldsymbol{s}+\delta^2=0\\
\partial_{\epsilon}\mathcal{L}(\boldsymbol{s},\mu,\omega,\epsilon,\delta) &= \mu\epsilon=0\\
\partial_{\delta}\mathcal{L}(\boldsymbol{s},\mu,\omega,\epsilon,\delta) &= 2\delta\omega=0
\end{aligned}$$

from the first equation we have:

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

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

Proof. (3/6).

So that we must solve the reduced non linear system

$$\begin{aligned}
\boldsymbol{s} &= -(\boldsymbol{H}+\mu\boldsymbol{I})^{-1}\boldsymbol{g}\\
\boldsymbol{s}^T\boldsymbol{s}+\epsilon^2-\Delta^2 &= 0\\
\boldsymbol{g}^T\boldsymbol{s} &= -\delta^2\\
\mu\epsilon &= 0
\end{aligned}$$

combining the first and third equation we have

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

Proof. (4/6).

If $\epsilon\neq 0$ then we must have $\mu=0$ and

$$\left\|-\boldsymbol{H}^{-1}\boldsymbol{g}\right\|=\|\boldsymbol{s}\|\leq\Delta$$

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

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

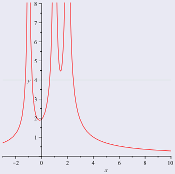
with $\boldsymbol{g}^T(\boldsymbol{H}+\mu\boldsymbol{I})^{-1}\boldsymbol{g}\geq 0$. Expand $\boldsymbol{g}=\sum_{i=1}^n\alpha_i\boldsymbol{u}_i$ with an orthonormal base of eigenvectors of $\boldsymbol{H}$ it follows

$$\begin{aligned}
\left\|(\boldsymbol{H}+\mu\boldsymbol{I})^{-1}\boldsymbol{g}\right\| &= \sum_{i=1}^n\frac{\alpha_i^2}{(\lambda_i+\mu)^2}\\
\boldsymbol{g}(\boldsymbol{H}+\mu\boldsymbol{I})^{-1}\boldsymbol{g} &= \sum_{i=1}^n\frac{\alpha_i^2}{\lambda_i+\mu}
\end{aligned}$$

Proof. (5/6).

$\left\|(\boldsymbol{H}+\mu\boldsymbol{I})^{-1}\boldsymbol{g}\right\|$ 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

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



Proof. (6/6).

Thus, or

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

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

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

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

□

Outline

Algorithm (Basic trust region algorithm)

$\boldsymbol{x}_0$ *assigned;* Δ_0 *assigned;* $k \leftarrow 0$;
while $\|\nabla \mathsf{f}(\boldsymbol{x}_k)\| \neq 0$ **do**
  $m_k(\boldsymbol{s}) = \mathsf{f}(\boldsymbol{x}_k) + \nabla\mathsf{f}(\boldsymbol{x}_k)\boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^T\boldsymbol{H}_k\boldsymbol{s}$; *— setup the model*
  $\boldsymbol{s}_k \leftarrow \arg\min_{\|\boldsymbol{s}\|\le\Delta_k} m_k(\boldsymbol{s})$; *— compute the step*
  $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{s}_k$;
  $\rho_k \leftarrow (f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}))/(m_k(0) - m_k(\boldsymbol{s}_k))$;
  — 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
    $\Delta_{k+1} \leftarrow \Delta_k/2$; $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k$; *— failure*
  end if
  $k \leftarrow k+1$;
end while

Cauchy point

Definition

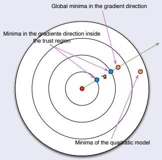
Consider the quadratic

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

and the minimization problem

$$\boldsymbol{s}^c(\Delta) = \arg\min_{\boldsymbol{s}\in\{-t\boldsymbol{g} \mid t\ge 0, \|-t\boldsymbol{g}\|\le\Delta\}} m(\boldsymbol{s})$$

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



Estimate the length of the Cauchy step

Lemma

For the Cauchy step the following characterization is valid:

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

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

Moreover

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

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

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. (cont.)

Proof.

Consider an onthonormal base of eigenvectors for H and write g if this coordinate:

$$g = \sum_{i=1}^{n} \alpha_i u_i$$

so that

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

and finally

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

□

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

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

then for the Cauchy step we have:

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

Proof.

Compute

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

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

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

(cont.)

Proof.

If $\boldsymbol{g}^T\boldsymbol{H}\boldsymbol{g} >$ we have

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

and

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

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

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

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

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

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

for a constant $\eta > 0$.

- Using lemma on slide N.28

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

- Thus any reasonable trust region numerical scheme satisfy

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

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

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

$$f(\boldsymbol{x}_0) - \lim_{k\in\mathcal{S}} f(\boldsymbol{x}_k) \geq$$

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

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

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

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

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

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:*

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

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

Lemma

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

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

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

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

for any accepted step.

Proof.

By using Taylor's theorem

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

so that

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

and

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

(cont.)

Proof.

using these inequalities we can estimate the ratio

$$\left| \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{s}_k)}{m_k(\boldsymbol{0}) - m_k(\boldsymbol{s}_k)} - 1 \right| = \frac{|m_k(\boldsymbol{s}_k) - f(\boldsymbol{x}_k + \boldsymbol{s}_k)|}{|m_k(\boldsymbol{0}) - m_k(\boldsymbol{s}_k)|}$$

$$\leq \frac{1}{2\eta} \frac{(\varrho(\boldsymbol{H}_k) + \gamma)\|\boldsymbol{s}_k\|^2}{|m_k(\boldsymbol{0}) - m_k(\boldsymbol{s}^c(\Delta))|}$$

$$\leq \frac{(\varrho(\boldsymbol{H}_k) + \gamma)\Delta^2}{\eta \|\nabla f(\boldsymbol{x}_k)\| \min\left\{\Delta, \frac{\|\nabla f(\boldsymbol{x}_k)\|}{\varrho(\boldsymbol{H}_k)}\right\}}$$

(cont.)

Proof.

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

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

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

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

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

Corollary

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

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

then we have

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

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

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

Convergence theorem

Theorem (Convergence to stationary points)

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

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

if the set

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

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

$$\lim_{k\to\infty} \nabla f(\boldsymbol{x}_k) = \boldsymbol{0}$$

Proof.

A trivial application of previous corollary. □

Convergence theorem

Theorem (Convergence to minima)

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

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

is compact then:

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

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.

Outline

The Newton approach (1/7)

- Consider the Lagrangian

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

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

- Then we can try to solve the nonlinear system

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

- Using Newton method we have

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

The Newton approach (2/7)

Lemma

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

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

Proof.

It enough to differentiate the relation

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

two times:

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

□

The Newton approach (3/7)

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

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

- To build Newton method we need to evaluate

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

- Putting all in a Newton step we obtain

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

The Newton approach (4/7)

- Newton step can be reorganized as follows

$$\boldsymbol{a} = (\boldsymbol{H} + \mu_k\boldsymbol{I})^{-1}\boldsymbol{g}$$
$$\boldsymbol{b} = (\boldsymbol{H} + \mu_k\boldsymbol{I})^{-1}\boldsymbol{a}$$
$$\beta = \|\boldsymbol{a}\|$$
$$\mu_{k+1} = \mu_k + \beta\frac{\beta - \Delta}{\boldsymbol{a}^T\boldsymbol{b}}$$

- 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.

The Newton approach (5/7)

Lemma

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

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

Proof.

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

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

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

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

(cont.)

The Newton approach (6/7)

Proof.

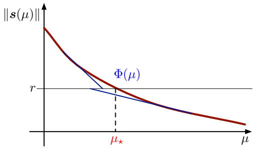
Using Cauchy–Schwartz inequality

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

□

The Newton approach (7/7)

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



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

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

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

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

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

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

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

- 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

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

where

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

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

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

- 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 $\mu_k \to \mu_\star$. So that this iteration become the Newton iteration as μ_k becomes near the solution.

Algorithm (Exact trust region algorithm)

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

Outline

The DogLeg approach (1/3)

- The computation of the μ such that $\|\boldsymbol{s}(\mu)\| = \Delta$ 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 $\boldsymbol{s}(\mu)$ a piecewise linear approximation $\boldsymbol{s}_{dl}(\mu)$ is used in computation.
- This approximation also permits to solve $\|\boldsymbol{s}_{dl}(\mu)\| = \Delta$ explicitly.

The DogLeg approach (2/3)

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

$$\boldsymbol{s}(0) = -\boldsymbol{H}^{-1}\boldsymbol{g}, \qquad \lim_{\mu\to\infty} \mu^2\boldsymbol{s}'(\mu) = -\boldsymbol{g}$$

  i.e. the curve start from the Newton step and reduce to zero in the direction opposite to the gradient step.
- The direction $-\boldsymbol{g}$ is a descent direction, so that a first piece of the piecewise approximation should be a straight line from $\boldsymbol{x}$ to the minimum of $m_k(-\lambda\boldsymbol{g})$. The minimum $\lambda_\star$ is found at

$$\lambda_\star = \frac{\|\boldsymbol{g}\|^2}{\boldsymbol{g}^T\boldsymbol{H}\boldsymbol{g}}$$

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

The DogLeg approach (3/3)

- We denote by

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

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 curve[1] $x_{dl}(\mu) = x + s_{dl}(\mu)$ where

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

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

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.

By using Kantorovich we can prove:

Lemma

We denote by

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

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

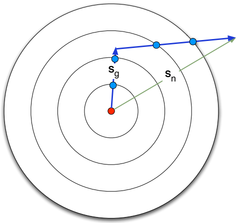
Proof.

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 in proved. □

the Dogleg piecewise curve



Lemma

Consider the dogleg curve connecting $\boldsymbol{x}+\boldsymbol{s}_n$, $\boldsymbol{x}+\boldsymbol{s}_g$ *and* $\boldsymbol{x}$. *The curve can be expressed as* $\boldsymbol{x}_{dl}(\mu) = \boldsymbol{x} + \boldsymbol{s}_{dl}(\mu)$ *where*

$$\boldsymbol{s}_{dl}(\mu) = \begin{cases} \mu \boldsymbol{s}_g + (1-\mu)\boldsymbol{s}_n & \text{for } \mu \in [0,1] \\ (2-\mu)\boldsymbol{s}_g & \text{for } \mu \in [1,2] \end{cases}$$

for this curve if $\boldsymbol{s}_g$ *is not parallel to* $\boldsymbol{s}_n$ *we have that the function*

$$d(\mu) = \|\boldsymbol{x}_{dl}(\mu) - \boldsymbol{x}\| = \|\boldsymbol{s}_{dl}(\mu)\|$$

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

Proof. (1/4).

In order to have a unique solution to the problem $\|\boldsymbol{s}_{dl}(\mu)\| = \Delta$ we must have that $\|\boldsymbol{s}_{dl}(\mu)\|$ is a monotone decreasing function:

$$\|\boldsymbol{s}_{dl}(\mu)\|^2 = \begin{cases} \mu^2 \boldsymbol{s}_g^2 + (1-\mu)^2 \boldsymbol{s}_n^2 + 2\mu(1-\mu)\boldsymbol{s}_g^T \boldsymbol{s}_n & \mu \in [0,1] \\ (2-\mu)^2 \boldsymbol{s}_g^2 & \mu \in [1,2] \end{cases}$$

To check monotonicity we take first derivative

$$\frac{\mathrm{d}}{\mathrm{d}\mu} \|\boldsymbol{s}_{dl}(\mu)\|^2$$

$$= \begin{cases} 2\mu \boldsymbol{s}_g^2 - 2(1-\mu)\boldsymbol{s}_n^2 + (2-4\mu)\boldsymbol{s}_g^T \boldsymbol{s}_n & \mu \in [0,1] \\ (2\mu - 4)\boldsymbol{s}_g^2 & \mu \in [1,2] \end{cases}$$

$$= \begin{cases} 2\mu(\boldsymbol{s}_g^2 + \boldsymbol{s}_n^2 - 2\boldsymbol{s}_g^T \boldsymbol{s}_n) - 2\boldsymbol{s}_n^2 + 2\boldsymbol{s}_g^T \boldsymbol{s}_n & \mu \in [0,1] \\ (2\mu - 4)\boldsymbol{s}_g^2 & \mu \in [1,2] \end{cases}$$

Proof. (2/4).

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

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

moreover

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

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

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

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

Proof. (3/4).

By using

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

of the previous lemma

$$\boldsymbol{s}_g^2 - \boldsymbol{s}_g^T \boldsymbol{s}_n = \|\boldsymbol{s}_g\|^2 \left(1 - \frac{\boldsymbol{s}_n^T \boldsymbol{s}_g}{\|\boldsymbol{s}_g\|^2}\right)$$

$$= \|\boldsymbol{s}_g\|^2 \left(1 - \frac{1}{\gamma_*}\right) < 0$$

Proof. (4/4).

To prove that $s_{dl}(\mu)$ is a descent direction it is enough top 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

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

For s_n we have

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

□

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. □

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 $s_{dl} = s_n$;

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:

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

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

to avoid cancellation the computation formula is the following

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

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

Algorithm (Computing DogLeg step)

DoglegStep(s_g, s_n, Δ);
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;

Outline

The Double DogLeg approach

- We denote by

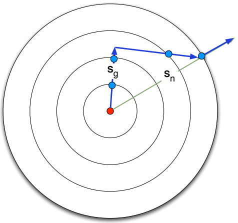
$$\boldsymbol{s}_g = -\boldsymbol{g}\frac{\|\boldsymbol{g}\|^2}{\boldsymbol{g}^T\boldsymbol{H}\boldsymbol{g}}, \qquad \boldsymbol{s}_n = -\boldsymbol{H}^{-1}\boldsymbol{g}, \qquad \gamma_* = \frac{\|\boldsymbol{s}_g\|^2}{\boldsymbol{s}_g^T\boldsymbol{s}_n}$$

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

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

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

The Double Dogleg piecewise curve



Lemma

Consider the double dogleg curve connecting $\boldsymbol{x}+\boldsymbol{s}_n$, $\boldsymbol{x}+\gamma_*\boldsymbol{s}_n$, $\boldsymbol{x}+\boldsymbol{s}_g$ and $\boldsymbol{x}$. The curve can be expressed as $\boldsymbol{x}_{ddl}(\mu) = \boldsymbol{x}+\boldsymbol{s}_{ddl}(\mu)$ where

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

for this curve if $\boldsymbol{s}_g$ is not parallel to $\boldsymbol{s}_n$ we have that the function

$$d(\mu) = \|\boldsymbol{s}_{ddl}(\mu)\|$$

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

Proof. (1/2).

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

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

To check monotonicity we take first derivative

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

Proof. (2/2).

Notice that

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

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

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

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

Using the previous Lemma we can prove

Lemma

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

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

1. if $\Delta \leq \|\boldsymbol{s}_g\|$ we set $\boldsymbol{s}_{ddl} = \Delta\boldsymbol{s}_g / \|\boldsymbol{s}_g\|$;
2. if $\Delta \leq \gamma_*\|\boldsymbol{s}_n\|$ we set $\boldsymbol{s}_{ddl} = \alpha\boldsymbol{s}_g + (1-\alpha)\gamma_*\boldsymbol{s}_n$; where α is the root in the interval $[0,1]$ of:
$$\alpha^2\|\boldsymbol{s}_g\|^2 + \gamma_*^2(1-\alpha)^2\|\boldsymbol{s}_n\|^2 + 2\gamma_*\alpha(1-\alpha)\boldsymbol{s}_g^T\boldsymbol{s}_n = \Delta^2$$
3. if $\Delta \leq \|\boldsymbol{s}_n\|$ we set $\boldsymbol{s}_{ddl} = \Delta\boldsymbol{s}_n / \|\boldsymbol{s}_n\|$;
4. if $\Delta > \|\boldsymbol{s}_n\|$ we set $\boldsymbol{s}_{ddl} = \boldsymbol{s}_n$;

Solving

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

we have that if $\|\boldsymbol{s}_g\| \leq \Delta \leq \gamma_*\|\boldsymbol{s}_n\|$ the root in $[0,1]$ is given by:

$$\begin{aligned}A &= \gamma_*^2\|\boldsymbol{s}_n\|^2 - \|\boldsymbol{s}_g\|^2\\ B &= \Delta^2 - \|\boldsymbol{s}_g\|^2\\ \alpha &= \frac{A-B}{A+\sqrt{AB}}\end{aligned}$$

Algorithm (Computing Double DogLeg step)

DoubleDoglegStep$(\boldsymbol{s}_g, \boldsymbol{s}_n, \Delta)$;
$\gamma_* \leftarrow \|\boldsymbol{s}_g\|^2 / (\boldsymbol{s}_g^T \boldsymbol{s}_n)$;
if $\Delta \leq \|\boldsymbol{s}_g\|$ **then**
  $\boldsymbol{s} \leftarrow \Delta \boldsymbol{s}_g / \|\boldsymbol{s}_g\|$;
else if $\Delta \leq \gamma_* \|\boldsymbol{s}_n\|$ **then**
  $A \leftarrow \gamma_*^2 \|\boldsymbol{s}_n\|^2 - \|\boldsymbol{s}_g\|^2$;
  $B \leftarrow \Delta^2 - \|\boldsymbol{s}_g\|^2$;
  $\alpha \leftarrow (A - B)/(A + \sqrt{AB})$;
  $\boldsymbol{s} \leftarrow \alpha \boldsymbol{s}_g + (1 - \alpha)\boldsymbol{s}_n$;
else if $\Delta \leq \|\boldsymbol{s}_n\|$ **then**
  $\boldsymbol{s} \leftarrow \Delta \boldsymbol{s}_n / \|\boldsymbol{s}_n\|$;
else
  $\boldsymbol{s} \leftarrow \boldsymbol{s}_n$;
end if
return $\boldsymbol{s}$;

Outline

Two dimensional subspace minimization

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

$$\boldsymbol{s} = \mathop{\arg\min}_{\|\alpha \boldsymbol{s}_g + \beta \boldsymbol{s}_n\| \leq \Delta} f(\alpha \boldsymbol{s}_g + \beta \boldsymbol{s}_n)$$

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

- The 2D problem results:

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

Two dimensional subspace minimization

The 2D problem written in matrix form:

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

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

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

and the constraint

$$\|\alpha \boldsymbol{s}_g + \beta \boldsymbol{s}_n\|^2 = \begin{pmatrix} \alpha & \beta \end{pmatrix} \boldsymbol{D} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

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

Lemma

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

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

$$\text{Subject to} \qquad \boldsymbol{s}^T\boldsymbol{D}\boldsymbol{s} \le r^2$$

Then the following curve

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

for any $\mu \ge 0$ *defines a descent direction for* $f(\boldsymbol{s})$. *Moreover*

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

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.