Newton Trust Region method for numerical unconstrained optimisation

Main theory and examples

Giulio Crognaletti

Advanced Numerical Analysis Exam, July 2021

Overview of the Problem

Problem Definition

It is possible to formally define the problem of unconstrained optimisation as follows:

Unconstrained optimisation (global minimiser)

Given a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, we seek $\mathrm{x}^* \in \mathbb{R}^n$ such that
$f(\mathrm{x}^*) \leq f(\mathrm{x}) \ \forall \mathrm{x} \in \mathbb{R}^n$
The element x^* is said to be a *global minimiser* of f

Usually, the information available on f is only *local* (e.g. ∇f), and hence all we are able to get are local minima:

Local minimiser

The element x^* is said to be a *local minimiser* of f if $\exists B(\mathrm{x}^*, \delta)$ such that
$f(\mathrm{x}^*) \leq f(\mathrm{x}) \ \forall \mathrm{x} \in B(\mathrm{x}^*, \delta)$

Newton's Method

Let's define $g(x) = \nabla f(x)$ and $B(x)$ the hessian of f (or a suitable approximation). Follows that $B(x)$ is also symmetric.

Local minimum characterisation

Let $x^* \in \mathbb{R}^n$ satisfy $g(x^*) = 0$, and $B(x^*) \geq 0$. Then x^* is a local minimiser of f.

So, a possible way of solving the problem is to use Newton's Method to find roots of the gradient g.

At each iteration k, the Newton step p_k can be found by *unconstrained* minimisation of the model function $m_k(p)$:

$$m_k(p) = f(x_k) + g(x_k)^T p + \frac{1}{2} p^T B(x_k) p \approx f(x_k + p)$$

If $B(x_k)$ is SPD, the Newton step $p_k = -B(x_k)^{-1} g(x_k)$ is also a global minimiser of $m_k(p)$.

Some Issues

However, there are some issues in characterising the Newton step this way:

- If B_k is not positive definite, then the unconstrained minimum of $m_k(p)$ does NOT exist.

  In fact, if $p \in Aut(\lambda)$, where λ is a negative eigenvalue of B_k, then $m_k(p) \approx -\lambda ||p||_2^2$, so the newton step cannot be interpreted as the model function minimum

- Even if the minimum exists, $m_k(p)$ is just a *local* approximation of f in a certain point x_k, ad as such is accurate only for small values of $||p||$. If the newton step is too large, the behavior of f and m can disagree importantly.

  In these cases the decrease of function f and the convergence properties to a root of g can be lost.

Local Convergence

The plain Newton Method is extremely fast (quadratic convergence), convergence results apply only locally, that is, if the initial guess x_0 is close enough to the solution x^*.

There are some way to extend this result to Global Convergence:

- **Line Search Approach**

  Here the Newton direction is left unchanged, and the effort is directed to finding the optimal scaling factor α_k in order to achieve a sufficient reduction $f(x_k) - f(x_k + \alpha_k p_k)$ (E.g. the Armijo rule)

- **Trust Region**

  Here instead, the effort is directed to finding both the direction and the length together, by only considering steps that lie in a region where the approximation $f(x_k + p) \approx m_k(p)$ holds

Trust Region Approach

Trust Region approach

The trust region method introduces the parameter Δ_k to define a neighborhood $T(\Delta_k)$ in which $m_k(p)$ is trusted, i.e. sufficiently close to $f(x_k + p)$:

$$T(\Delta_k) = \{p \in \mathbb{R}^n \text{ s.t. } ||p||_2 \leq \Delta_k\}$$

At each iteration, the step is chosen by solving the *constrained* minimisation

$$p_k = \arg\min_{p \in T(\Delta_k)} m_k(p) \tag{1}$$

Finding a numerical solution of (1) will be denoted as "TR subproblem"

To evaluate whether $T(\Delta_k)$ is appropriate, we define the factor ρ_k:

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)} = \frac{actual(p_k)}{pred(p_k)} \tag{2}$$

For accurate models, $\rho_k \approx 1$, while if $\rho_k << 1$ then the predicted reduction was too optimistic.

Trust Region approach - Algorithm

Data: $\Delta_{max} > 0$, $\Delta_0 \in (0, \Delta_{max})$, ϵ, $\eta \in (0, \epsilon)$, x_0, tol, $\alpha < 1$, $\omega > 1$

while $||g(x_k)|| > tol\ ||g(x_0)||$ **do**

$p_k \leftarrow$ Solve (1)

$\rho_k \leftarrow$ Evaluate (2)

if $\rho_k < \epsilon$ **then**

$\Delta_{k+1} \leftarrow \alpha \Delta_k$ `// Poor agreement, shrink Trust Region`

else if $\rho_k > 1 - \epsilon$ **then**

$\Delta_{k+1} \leftarrow \omega \Delta_k$ `// Good agreement, widen Trust Region`

else

$\Delta_{k+1} \leftarrow \Delta_k$ `// Ok agreement, keep Trust Region`

if $\rho_k > \eta$ **then**

$x_{k+1} = x_k + p_k$

else

$x_{k+1} = x_k$ // Decrease in f is not enough

end

TR subproblem

Cauchy step

The Cauchy step obtained by solving a very rough approximation of (1) based on the gradient direction, very similar to a Line Search approach: $p_k^C = -\tau_k g_k$, where

$$\tau_k = \arg\min_{t \geq 0} m(-t g_k) \;\; \text{s.t.} \;\; p_k^C \in T(\Delta_k)$$

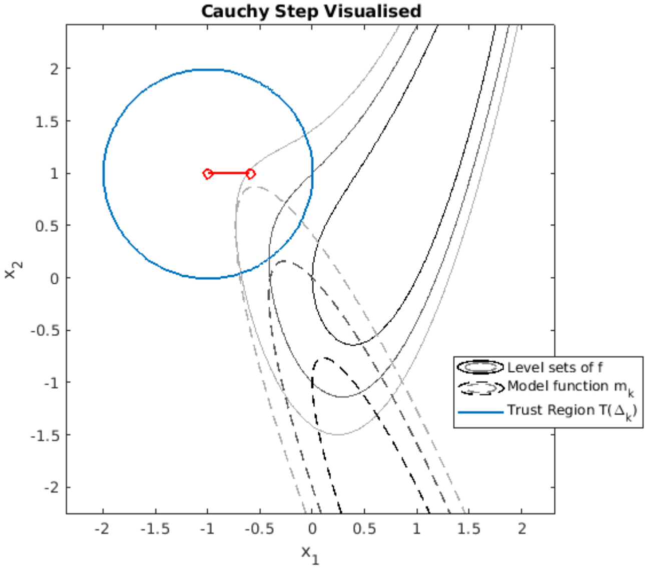
By direct substitution, it is possible to find an easy formula for τ_k:

$$\tau_k = \begin{cases} \frac{\Delta_k}{||g_k||} & \text{if } g_k^T B_k g_k \leq 0 \\ \min\left(\frac{\Delta_k}{||g_k||}, \frac{||g_k||^2}{g_k^T B_k g_k}\right) & \text{otherwise} \end{cases}$$

This method has strong similarities with the steepest descent

- **PROS**: it is very cheap and is enough to get global convergence
- **CONS**: it is a very rough estimate, and can slow down the overall procedure

Cauchy step - Example (Rosenbrock function)



Dogleg Method

When $B(x)$ is always SPD, we can find an improvement of the Cauchy point with the following observations:

- We already noted that the unconstrained minimiser of $m_k(p)$ is $p_k^B = -B_k^{-1}g$, and for $\Delta_k \geq ||p_k^B||$ this is also the solution of (1). This will always hold in the limit of large Δ_k
- On the other hand, in the limit of small Δ_k, the linear part of the quadratic model is dominant, so the solution can be effectively approximated by the "unconstrained Cauchy step" $p_k^U = -\frac{||g_k||^2}{g_k^T B_k g_k} g_k$

The approximate solution while Δ_k varies can then be approximated by combining the two limits in a piecewise linear path, whose length is governed by the parameter τ:

$$p_k^D(\tau) = \begin{cases} \tau p_k^U & \text{if } \tau \in (0,1] \\ p_k^U + (\tau - 1)(p_k^B - p_k^U) & \text{if } \tau \in (1,2] \end{cases} \tag{3}$$

Dogleg Method

For $\Delta_k < ||p_k^B||$ choice of τ in (3) must be done to enforce $p_k^D(\tau) \in T(\Delta_k)$, and this is made easy by the following lemma

Dogleg monotonicity lemma

If B_k is SPD, then

1. $||p_k^D(\tau)||$ is a monotonically increasing function of τ
2. $m_k(p_k^D(\tau))$ is a monotonically decreasing function of τ

This implies that the equation $||p_k^D(\tau)|| = \Delta_k$ has unique solution, and that the minimum value of m_k along this path is obtained exactly there. Finding such intersection reveals to be trivial, as it is a quadratic equation.

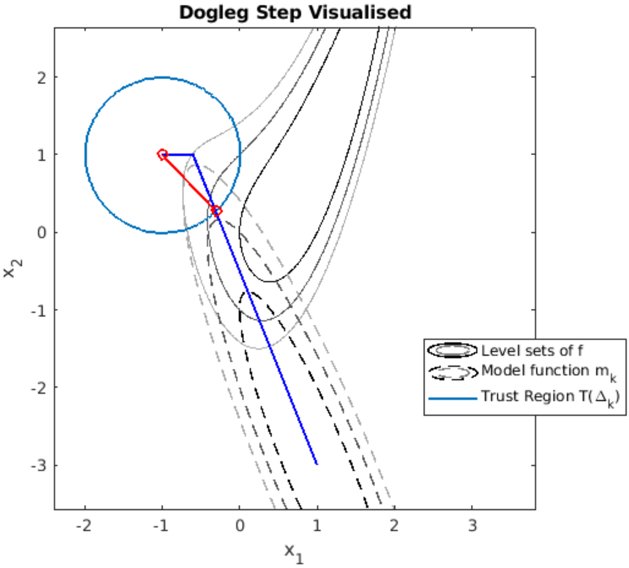
PROS

- it is cheap, requiring only one factorisation
- when $\Delta_k \geq ||p_k^B||$ it gives the exact solution

CONS

- It only works when B(x) is SPD

Dogleg step - Example (Rosenbrock function)



Exact solution - characterisation

The exact solution of the constrained minimisation (1) can be characterised by the following theorem:

Theroem

The vector $p^* \in T(\Delta)$ is a solution to

$$\arg\min_{p \in T(\Delta)} f + g^T p + \frac{1}{2} p^T B p$$

if and only if $\exists \lambda > 0$ such that

$$(B + \lambda I)p^* = -g$$
$$\lambda(||p^*|| - \Delta) = 0$$
$$(B + \lambda I) \geq 0$$

To approximate p^*, it's possible to numerically solve the above system of equations (exact solution approximation)

Exact solution - approximation

Some observations on this result:

- If B is SPD and the Newton step $p^N = -B^{-1}g \in T(\Delta)$, then $\lambda = 0$ and p^N is the solution to (1).
- When B is not SPD, this theorem implies that $\lambda \neq 0$, and therefore $||p^*|| = \Delta$, i.e. the minimum lies on the border of $T(\Delta)$.

When either $B \not\geq 0$ or p^N is not feasible, follows from $-(B + \lambda I)p^* = g$ that

$$\exists \lambda > 0 \quad \text{s.t} \quad p(\lambda) = -(B + \lambda I)^{-1} g \quad \text{and} \quad ||p(\lambda)|| = \Delta \tag{4}$$

is the solution of (1).

Considering the factorisation $B = Q\Lambda Q^T$, where $\Lambda = diag(\lambda_1, ...\lambda_n)$, with $\lambda_1 \leq \lambda_2 ... \leq \lambda_n$, the last condition $(B + \lambda I) \geq 0$ implies $\lambda \geq -\lambda_1$.

Exact solution - approximation

Under the simplifying hypothesis that $q_i^T g \neq 0 \ \forall i$, the scalar function

$$||p(\lambda)||^2 = ||Q(\Lambda + \lambda I)^{-1} Q^T g||^2 = \sum_{i=1}^{n} \frac{(q_i^T g)^2}{(\lambda_i + \lambda)^2}$$

is monotonically decreasing function in $(-\lambda_1, +\infty)$ and since

$$\lim_{\lambda \to -\lambda_1} ||p(\lambda)||^2 = +\infty \ \text{ and } \lim_{\lambda \to +\infty} ||p(\lambda)||^2 = 0$$

there is a unique value λ^* consistent with (4), which can be computed using Newton's Method as the root of $\phi(\lambda) = \frac{1}{||p(\lambda)||} - \frac{1}{\Delta}$.

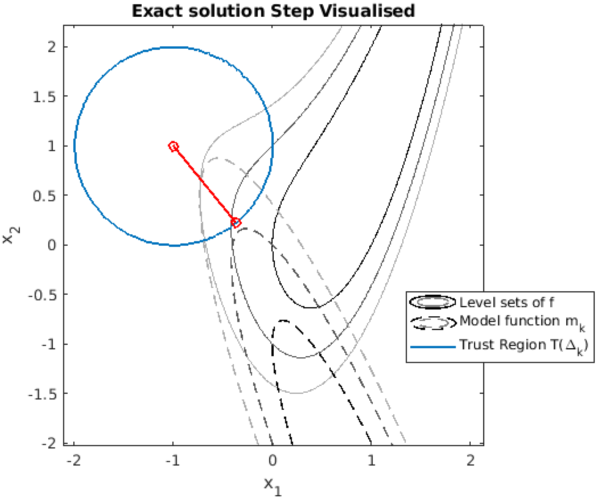
PROS

- Very accurate approximation of p^*

CONS

- Evaluation of ϕ is very costly (requires the factorisation of B)
- An estimate of λ_1 is required

Exact solution step - Example (Rosenbrock function)



Convergence Properties

Global Convergence

Assume that $\forall k$ the step p_k complies to

$$m_k(0) - m_k(p_k) \geq c_1 ||g_k|| \min\left(\Delta_k, \frac{||g_k||}{||B_k||}\right) \text{ and } ||p_k|| < \gamma\Delta_k \qquad (5)$$

for some $c_1 \in (0, 1]$, and $\gamma \geq 1$. Then

Convergence to stationary points

Assume (5) holds. If B_k is bounded below $\forall k$, f is bounded below on S and continuously Lipschitz differentiable on $S(R_0)$ for some constant R_0, then

$$\lim_k g(x_k) = 0$$

Where $S = \{x \in \mathbb{R}^n \ s.t. \ f(x) \leq f(x_0)\}$ be the level set of f in x_0 and let $S(R) = \{x \in \mathbb{R}^n \ s.t. \ ||x - y|| < R \text{ for some } y \in S\}$

Global Convergence

If an exact solution approximation of (1) is used to detemine p_k, then

Convergence to local minimiser

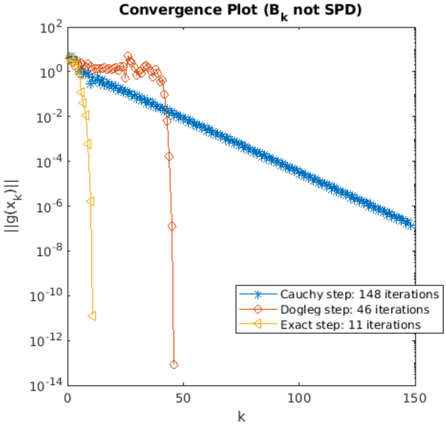
Let the assumptions of the previous theorem hold, $B(x)$ be the Hessian of $f(x)$ and $m(0) - m(p_k) \geq c_1(m(0) - m(p^*)) \; \forall k$. Then $\lim_k g(x_k) = 0$.

Moreover if S is compact, then either x_k converge to a local minimiser x^* or $\{x_k\}$ has a limit point to x^*

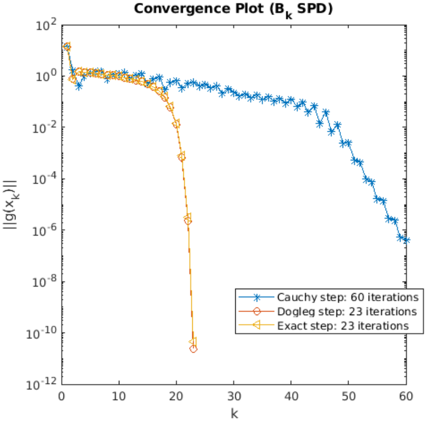
In the case of $x_k \to x^*$, some comments on local convergence:

- While being near the local minimum x^*, $p_k \approx p_k^N$ which suggests fast local convergence is kept.
- It can be proved that for large k, if $||p_k - p_k^N|| \in O(||p_k^N||)$ the Trust Region constraint becomes inactive and the convergence of$\{x_k\}$ superlinear.

Examle - Convergence profiles

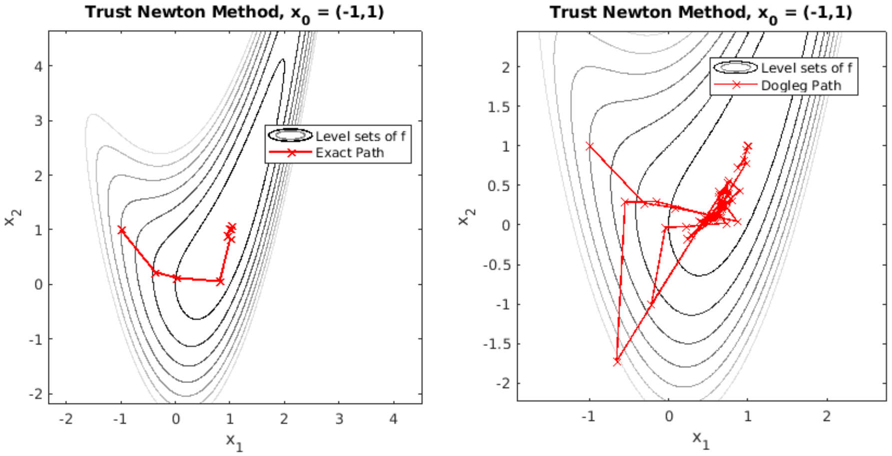


(a) If B_k is not SPD, Dogleg may fail



(b) If B_k is SPD, all methods work

Example (Rosenbrock function)



Thank You for your attention