Trust-Region, the Levenberg-Marquardt method, Termination rules

Niclas Börlin

5DA001 Non-linear Optimization

Line search and trust-region

- Line search and trust-region and two examples of global strategies that modify a (usually) locally convergent algorithm, e.g. Newton, to become globally convergent.
- At every iteration k, both global strategies enforce the descent condition

$$f(x_{k+1}) < f(x_k)$$

by controlling the length and direction of the step.

Line search vs. trust-region

- In the line search strategy, the direction is chosen first, followed by the distance.
- In the trust-region strategy, the maximum distance is chosen first, followed by the direction.

The trust-region model

- Trust-region methods use the following local quadratic model of the objective function:

$$m_k(p) = f_k + p^T g_k + \frac{1}{2} p^T B_k p,$$
$$f_k = f(x_k), \quad g_k = \nabla f(x_k).$$

- Newton-type trust-region methods have $B_k = \nabla^2 f(x_k)$.
- The model is "trusted" within a limited region around the current point x_k defined by

$$\|p\| \leq \Delta_k.$$

- This will limit the length of the step from x_k to x_{k+1}.
- The value of Δ_k will be increased if the model is found to be in "good" agreement with the objective function, and decreased if the model is a "poor" approximation.

The trust-region subproblem

- At iteration k of a trust-region method, the following subproblem must be solved:

$$\min_p m_k(p) = f_k + p^T g_k + \frac{1}{2} p^T B_k p,$$
$$\text{s.t. } \|p\| \le \Delta_k.$$

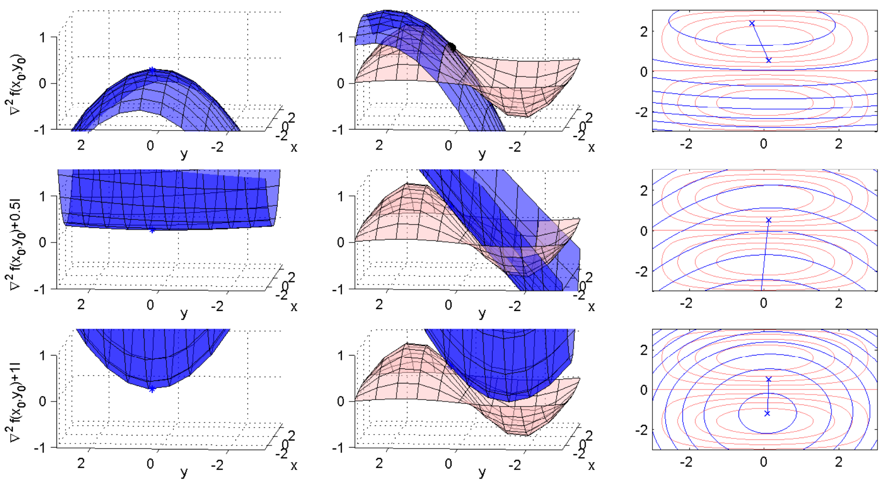
- It can be shown that the solution p^* of this constrained problem is the solution of the linear equation system

$$(B_k + \lambda I)p^* = -g_k$$

for some value of the Lagrange multiplier $\lambda \ge 0$ such that the matrix $(B_k + \lambda I)$ is positive semidefinite.

- Furthermore, the following condition always holds:

$$\lambda(\Delta_k - \|p^*\|) = 0.$$



The trust-region subproblem
Cont'd

- Note that if $B_k = \nabla^2 f(x_k)$ is positive definite and Δ_k big enough, the solution of the trust-region subproblem is the solution of

$$\nabla^2 f(x_k)p = -\nabla f(x_k),$$

i.e. p is a Newton direction.

- Otherwise,

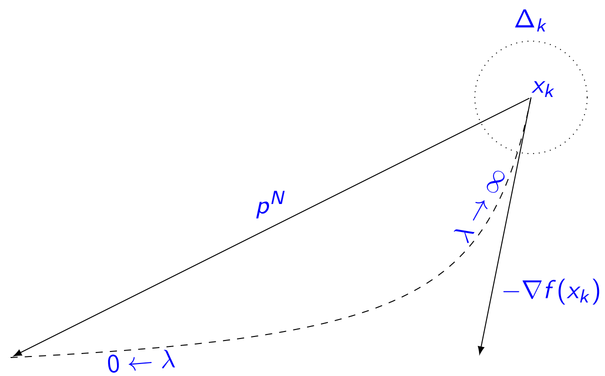
$$\Delta_k \ge \|p_k\| = \|(\nabla^2 f(x_k) + \lambda I)^{-1} \nabla f(x_k)\|,$$

so if $\Delta_k \to 0$, then $\lambda \to \infty$ and

$$p_k \to -\frac{1}{\lambda} \nabla f(x_k).$$

The trust-region search direction

- When λ varies between 0 and ∞, the corresponding search direction $p_k(\lambda)$ will vary between the Newton direction and a multiple of the negative gradient.



The reduction ratio

- To enable adaption of the trust-region size Δ_k, we define the reduction ratio

$$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)} = \frac{\text{actual reduction}}{\text{predicted reduction}}.$$

- If the reduction ratio is:
  - **large** e.g. $\rho_k > \frac{3}{4}$, the step p_k is accepted and the trust-region size is increased in the next iteration.
  - **good enough** e.g. $\frac{3}{4} > \rho_k > \frac{1}{4}$, the step p_k is accepted but the trust-region size is unchanged in the next iteration.
  - **small** e.g. $\rho_k < \frac{1}{4}$, the step p_k is rejected and the trust-region size is decreased in the next iteration.

The trust-region algorithm

- Specify starting approximation x_0, maximum step length $\hat{\Delta}$, initial trust-region size $\Delta_0 \in (0, \hat{\Delta})$ and acceptance constant $\eta \in [0, \frac{1}{4})$.
- For $k = 0, 1, \ldots$ until x_k is optimal
  - Solve
    $$\min_p m_k(p) = f_k + p^T g_k + \frac{1}{2} p^T B_k p, \quad \text{s.t. } \|p\| \le \Delta_k$$
    approximately for a trial step p_k.
  - Calculate the reduction ratio
    $$\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}.$$
  - Update the current point
    $$x_{k+1} = \begin{cases} x_k + p_k & \text{if } \rho_k > \eta, \\ x_k & \text{otherwise.} \end{cases}$$
  - Update the trust-region radius
    $$\Delta_{k+1} = \begin{cases} \frac{1}{4}\Delta_k & \text{if } \rho_k < \frac{1}{4}, \\ \min(2\Delta_k, \hat{\Delta}) & \text{if } \rho_k > \frac{3}{4} \text{ and } \|p_k\| = \Delta_k, \\ \Delta_k & \text{otherwise.} \end{cases}$$

The Levenberg-Marquardt algorithm

Original formulation

- The first trust-region algorithm was developed for least squares problems by Levenberg (1944) and Marquardt (1963).
- The original algorithm uses the approximation $B_k = J_k^T J_k$ and solves

$$(B_k + \lambda_k I)p = -g_k$$

  for different values of λ_k.
- The original algorithm adapts by modifying the λ value, i.e. if the reduction produced by p is good enough, $\lambda_{k+1} = \frac{1}{10}\lambda_k$, otherwise $\lambda_{k+1} = 10\lambda_k$ and the step is rejected.

The Levenberg-Marquardt algorithm

Trust-region formulation

- The Levenberg-Marquardt algorithm was put into the trust-region framework (Δ-parameterized) in the early 80-ies (Moré, 1981).
- The Δ version of Levenberg-Marquardts has a number of advantages over the λ version:
  - λ is nontrivially related to the problem. Δ is related to the size of x. E.g. $\Delta_0 = \|x_0\|$ is often a reasonable choice.
  - The transition to $\lambda = 0$ is handled transparently.
  - The λ algorithm need to re-solve the equation system
    $$(B_k + \lambda_k I)p = -g_k$$
    when a step is rejected and λ is reduced. The Δ algorithm has ways to avoid that.
- However, many popular implementation of Levenberg-Marquardt still use the original, λ-parameterized, formulation.

The Trust-region subproblem

- The trust-region subproblem

$$\min_p m_k(p) = f_k + p^T g_k + \frac{1}{2} p^T B_k p,$$
$$\text{s.t. } \|p\| \le \Delta_k$$

is a hard problem.

- If the unconstrained solution

$$p^B = -B_k^{-1} g_k$$

is too long, $\|p^B\| > \Delta_k$, we have to find a λ such that

$$\|p_k(\lambda)\| = \|(B_k + \lambda I)^{-1} g_k\| = \Delta_k.$$

- This is a non-linear equation in λ.

The Dogleg algorithm

- The dogleg algorithm solves the subproblem by approximating $p_k(\lambda)$ by a *dogleg path* — a piecewise linear polygon $\tilde{p}(\tau)$ and solving $\|\tilde{p}(\tau)\| = \Delta_k$.
- The polygon $\tilde{p}(\tau)$ is defined as

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

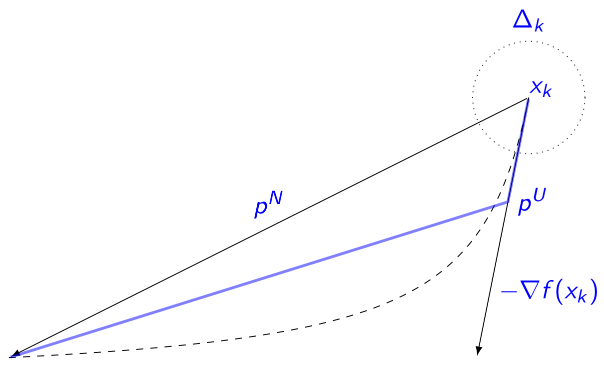
- The point p^U is the *Cauchy point*, i.e. the minimizer of m in the direction of the negative gradient

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

- The dogleg algorithm will produce sufficient descent if B_k is positive definite.

The Dogleg algorithm

The dogleg path



Termination rules

- Mathematically, a minimization algorithm should terminate with a solution x_k satisfying
  - the first order necessary conditions

$$\nabla f(x_k) = 0$$

  - the second order necessary conditions

$$\nabla^2 f(x_k) \text{ positive semi-definite.}$$

- Due to e.g. finite arithmetic, no algorithm is guaranteed to satisfy this condition in finite time.
- Instead, termination rules based on thresholds on e.g. $\|\nabla f(x_k)\|$ are used.

Termination rules

Absolute criteria

- Consider the absolute termination criteria
$$\|\nabla f(x_k)\| \leq \epsilon,$$
for some $\epsilon > 0$.
- This condition is scale dependent, since a change of units in f would rescale ∇f and affect the strength of the condition.
- A change of units from e.g. mm to m corresponds to a scaling of 10^3.

Termination rules

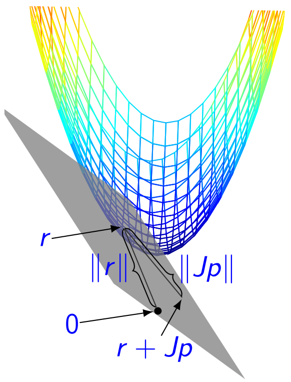
Relative criteria

- A small modification of the absolute termination criteria leads the relative termination criteria
$$\|\nabla f(x_k)\| \leq \epsilon |f(x_k)|,$$
which is not scale dependent.
- However, if $f(x^*) \approx 0$ the relative test will be difficult or impossible to satisfy due to round-off errors.
- A possible combination is
$$\|\nabla f(x_k)\| \leq \epsilon \left(1 + |f(x_k)|\right).$$
- This test will behave like an absolute test if $f(x_k) \approx 0$ and otherwise like a relative test.

Termination rules

Least squares problems

- For least squares problems
$$\min_x f(x) = \frac{1}{2} r(x)^T r(x),$$
the test
$$\|Jp\| = \|J(J^T J)^{-1} J^T r\| \leq \epsilon(1 + \|r\|)$$
may be used instead of the gradient test.
- Since Jp and r belong to the same vector space, the test may be interpreted geometrically.
- The ratio
$$\frac{\|Jp\|}{\|r\|} = \cos\alpha,$$
is related to the angle α between the residual r and the tangent plane at r.



Termination rules

Least squares problems

- For least squares problems
$$\min_x f(x) = \frac{1}{2} r(x)^T r(x),$$
the test
$$\|Jp\| = \|J(J^T J)^{-1} J^T r\| \leq \epsilon(1 + \|r\|)$$
may be used instead of the gradient test.
- Since Jp and r belong to the same vector space, the test may be interpreted geometrically.
- Close to the solution the residual will approach orthogonality with the tangent plane, i.e.
$$\alpha \to \pi/2 \text{ and } \frac{\|Jp\|}{\|r\|} \to 0.$$