MATH529 – Fundamentals of Optimization
Trust Region Algorithms

Marco A. Montes de Oca

Mathematical Sciences, University of Delaware, USA

Line Search vs. Trust Region

- Line Search
  - Select a search (descent) direction $\mathbf{p}_k$.
  - Select step size α_k to ensure sufficient descent along $f(\mathbf{x}_k + \alpha_k \mathbf{p}_k)$.
  - Move to new point $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$.
- Trust Region
  - Build model m_k of f at $\mathbf{x}_k$. (Similar to Newton's method.)
  - Solve $\mathbf{p}_k = \min_{\mathbf{p} \in \mathbb{R}^n} m_k(\mathbf{p}) = f_k + \mathbf{g}_k^T \mathbf{p} + \frac{1}{2} \mathbf{p}^T B_k \mathbf{p} \qquad \text{s.t.} \quad ||\mathbf{p}|| \leq \Delta_k$
  - If predicted decrease is good enough, then $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$. Otherwise, $\mathbf{x}_{k+1} = \mathbf{x}_k$ and improve the model.

Acceptance criterion

To measure how well the predicted decrease matches the actual decrease, we use:

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

Given that $m_k(0) - m_k(\mathbf{p}_k) > 0$, if $\rho_k < 0$ then the predicted reduction is not obtained, the step is rejected and Δ_k is decreased.
If $\rho_k \approx 1$, then accept $\mathbf{p}_k$ and increase Δ_k.
If $\rho_k > 0$ but not ≈ 1, then accept $\mathbf{p}_k$ and do not change Δ_k.
If $\rho_k > 0$ but ≈ 0, the step may be accepted or not, and Δ_k is decreased.

Algorithm

Inttialization: $k = 0$, $\Delta_0 > 0$, and $\mathbf{x}_0$ by educated guess. Set $\eta_g \in (0, 1)$ (typically, $\eta_g = 0.9$), $\eta_a \in (0, \eta_g)$ (typically, $\eta_a = 0.1$), $\gamma_e \geq 1$ (typically, $\gamma_e = 2$), and $\gamma_s \in (0, 1)$ (typically, $\gamma_s = 0.5$).

Until convergence do:

Build model $m_k(\mathbf{p})$.

Solve trust region subproblem (result in $\mathbf{p}_k$)

Test acceptance criterion (result in ρ_k).

If $\rho_k \geq \eta_g$, then $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$ and $\Delta_{k+1} = \gamma_e \Delta_k$

Else If $\rho_k \geq \eta_a$, then $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$

Else If $\rho_k < \eta_a$, then $\Delta_{k+1} = \gamma_s \Delta_k$

Increase k by one

Solving the trust region subproblem approximately

We want to solve the subproblem as efficiently as possible.

We want a solution that at least decreases the model as much as the steepest descent would subject to the size of the trust region.

Solving the trust region subproblem approximately

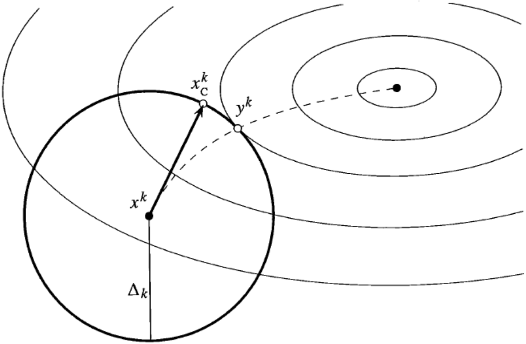


Figure 5.9. The trust region subproblem. The arrow represents the direction of steepest descent and x_C^k is the Cauchy point. The dotted curve represents the solutions of the subproblem for various values of Δ_k.

From Ruszczyński A. “Nonlinear Optimization” pp. 268. Princeton University Press. 2006.

Cauchy Point

The Cauchy point can be found by minimizing the model along a line segment.

Thus, let $\mathbf{p}_k^s = -\Delta_k \frac{\mathbf{g}_k}{||\mathbf{g}_k||}$. (Point at the border of the trust region in the direction of steepest descent.)

The Cauchy point is $\mathbf{p}_k^C = \tau_k \mathbf{p}_k^s = -\tau_k \Delta_k \frac{\mathbf{g}_k}{||\mathbf{g}_k||}$.

To find τ_k, consider
$g(\tau) = m_k(\tau \mathbf{p}_k^s) = f_k + \mathbf{g}_k^T(\tau \mathbf{p}_k^s) + \frac{1}{2}(\tau \mathbf{p}_k^s)^T B_k (\tau \mathbf{p}_k^s)$
$m_k(\tau \mathbf{p}_k^s) = f_k + \tau \mathbf{g}_k^T \mathbf{p}_k^s + \frac{\tau^2}{2}(\mathbf{p}_k^s)^T B_k \mathbf{p}_k^s$
Differentiating wrt τ:

$0 = g'(\tau) = \mathbf{g}_k^T \mathbf{p}_k^s + \tau (\mathbf{p}_k^s)^T B_k \mathbf{p}_k^s$, which means that

Cauchy Point

$$\tau_k = -\frac{\mathbf{g}_k^T \mathbf{p}_k^s}{(\mathbf{p}_k^s)^T B_k \mathbf{p}_k^s}. \qquad (1)$$

Substituting $\mathbf{p}_k^s = -\Delta_k \frac{\mathbf{g}_k}{||\mathbf{g}_k||}$ in (1):

$$\tau_k = -\frac{\mathbf{g_k}^T(-\Delta_k \frac{\mathbf{g}_k}{||\mathbf{g}_k||})}{(-\Delta_k \frac{\mathbf{g}_k}{||\mathbf{g}_k||})^T B_k (-\Delta_k \frac{\mathbf{g}_k}{||\mathbf{g}_k||})} = \frac{1}{\Delta_k} \frac{||\mathbf{g_k}||}{\frac{1}{||\mathbf{g_k}||^2}(\mathbf{g}_k^T B_k \mathbf{g}_k)} = \frac{1}{\Delta_k} \frac{||\mathbf{g_k}||^3}{\mathbf{g}_k^T B_k \mathbf{g}_k}.$$

However, there may be two problems:

a) $\tau_k > \Delta_k$, or
b) $\mathbf{g}_k^T B_k \mathbf{g}_k \leq 0$, that is, B_k is not positive definite.

So, we define the Cauchy point as follows:

Definition (Cauchy Point)

$\mathbf{p}_k^C = \tau_k \mathbf{p}_k^s = -\tau_k \Delta_k \frac{\mathbf{g}_k}{||\mathbf{g}_k||}$, where

$\tau_k = 1$ if $\mathbf{g}_k^T B_k \mathbf{g}_k \leq 0$, or $\tau_k = \min\{1, \frac{1}{\Delta_k} \frac{||\mathbf{g_k}||^3}{\mathbf{g}_k^T B_k \mathbf{g}_k}\}$ otherwise.

Cauchy step is a baseline of performance

- A reduction at least as good as the one obtained with the Cauchy step guarantees that the trust-region method is convergent.
- The Cauchy step is just a steepest descent step with fixed length (Δ_k). (Thus, it is inefficient.)
- The direction of the Cauchy step does not depend directly on B_k, which means that curvature information is not exploited in its calculation.

Improvements over Cauchy step

The main idea is to incorporate information provided by the "full step" (Newton step for the local model m_k): $\mathbf{p}_k^B = -B_k^{-1}\mathbf{g}_k$ whenever $||\mathbf{p}_k^B|| \leq \Delta_k$.

Dogleg Method

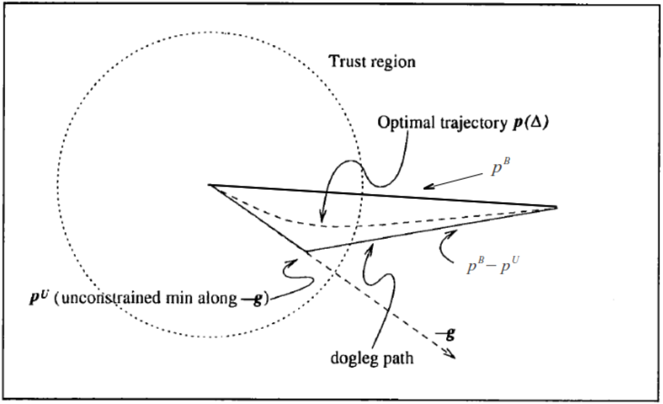
Let $\mathbf{p}_k^\star$ be the solution to the subproblem. If $\Delta_k \geq ||\mathbf{p}_k^B||$, then $\mathbf{p}_k^\star = \mathbf{p}_k^B$. If, however, $\Delta_k << ||\mathbf{p}_k^B||$, then $\mathbf{p}_k^\star \approx \mathbf{p}_k^s = -\Delta_k \frac{\mathbf{g}_k}{||\mathbf{g}_k||}$.

The idea of the dogleg method is to combine these two directions and search the minimum of the model along the resulting path $\widetilde{p}(\tau)$:

$$\widetilde{p}(\tau) = \begin{cases} \tau \mathbf{p}_k^U & 0 \leq \tau \leq 1, \\ \mathbf{p}_k^U + (\tau - 1)(\mathbf{p}_k^B - \mathbf{p}_k^U) & 1 < \tau \leq 2, \end{cases}$$

where $0 \leq \tau \leq 2$, and $\mathbf{p}_k^U = -\frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_k^T B_k \mathbf{g}_k}\mathbf{g}_k$, i.e., the steepest descent step with exact length (see that if $||\mathbf{p}_k^C|| < \Delta_k$, $\mathbf{p}_k^U = \mathbf{p}_k^C$).

Dogleg Method



Adapted from Nocedal J. and Wright S. “Numerical Optimization” 2nd. Ed. pp. 74. Springer. 2006.

Dogleg Method

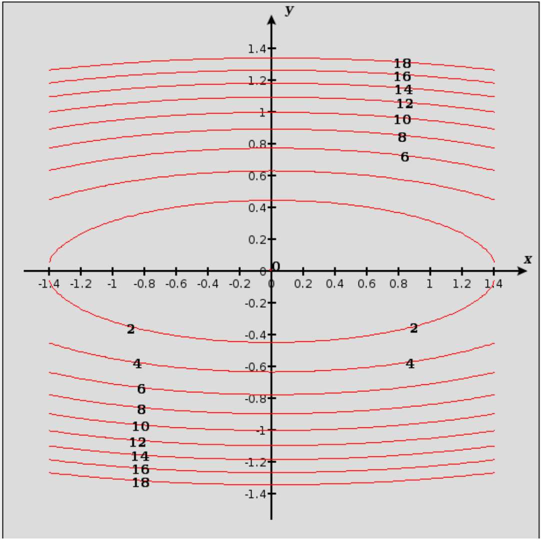
If B_k is positive definite, $m(\widetilde{p}(\tau))$ is a decreasing function of τ (Lemma 4.2, page 75). Therefore:

The minimum along $\widetilde{p}(\tau)$ is attained at $\tau = 2$ if $||\mathbf{p}_k^B|| \leq \Delta_k$.

If $||\mathbf{p}_k^B|| > \Delta_k$, we need to find τ such that $||\widetilde{p}(\tau)|| = \Delta_k$.

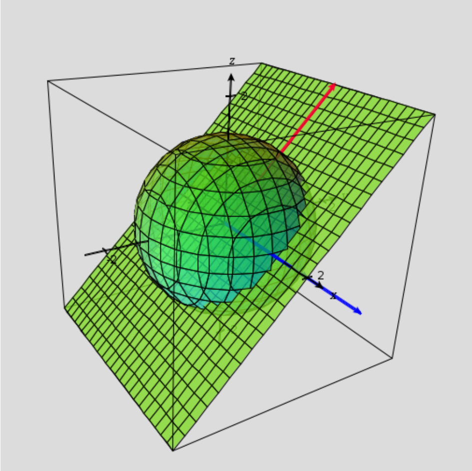
Dogleg Method

Example: $f(x, y) = x^2 + 10y^2$



2D Subspace Minimization

The dogleg is completely contained in the plane spanned by $\mathbf{p}_k^U$ and $\mathbf{p}_k^B$. Therefore, one may extend the search to the whole subspace spanned by $\mathbf{p}_k^U$ and $\mathbf{p}_k^B$, $\mathrm{span}[\mathbf{p}_k^U, \mathbf{p}_k^B]$.



2D Subspace Minimization

Given $\mathrm{span}[\mathbf{p}_k^U, \mathbf{p}_k^B] = \{\mathbf{v} | a\mathbf{p}_k^U + b\mathbf{p}_k^B\}$, $a, b \in \mathbb{R}$. The subproblem is thus:

$$\min_{a,b\in\mathbb{R}} \left[f_k + (a\mathbf{p}_k^U + b\mathbf{p}_k^B)^T \nabla f_k + \frac{1}{2}(a\mathbf{p}_k^U + b\mathbf{p}_k^B)^T B_k (a\mathbf{p}_k^U + b\mathbf{p}_k^B) \right]$$

s.t. $||a\mathbf{p}_k^U + b\mathbf{p}_k^B|| \leq \Delta_k$,

which can be solved using tools from constrained optimization. (To be discussed after break.)

Issues

Indefinite Hessians

Problem: Newton's step may not be decreasing.

Example: Newton's step solves the system $Hf_k\mathbf{p} = -\nabla f_k$. Now,
$\begin{pmatrix} 10 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix} \mathbf{p} = -(1,-3,2)^T = (-1,3,-2)^T$. Thus,
$\mathbf{p} = (-1/10, 1, 2)$. However, $\mathbf{p}^T \nabla f_k > 0$, thus $\mathbf{p}$ is not a descent direction.

Solution approaches:

- Replace negative eigenvalues by some small positive number.
- Replace negative eigenvalues by their negative.

Replace negative eigenvalues by some small positive number

Now $Hf_k = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 10^{-6} \end{pmatrix}$, so $\mathbf{p}^T \nabla f_k < 0$, but $\mathbf{p} = \,?$

Replace negative eigenvalues by some small positive number

Now $Hf_k = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 10^{-6} \end{pmatrix}$, so $\mathbf{p}^T \nabla f_k < 0$, but $\mathbf{p} = ?$

Replace negative eigenvalues by their negative

$$\text{Now } Hf_k = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \text{ so } \mathbf{p}^T \nabla f_k < 0, \text{ but } \mathbf{p} = ?$$

In practice

Perturb B_k with βI such that:

- $(B_k + \beta I)\mathbf{p} = -g$,
- $\beta(\Delta_k - ||\mathbf{p}||) = 0$, and
- $B_k + \beta I$ is positive semidefinite.

with $\beta \in (-\lambda_1, -2\lambda_1]$, where λ_1 is the most negative eigenvalue of B.

Further improvements

- Iterative solution of the subproblem: To avoid direct Hessian manipulation.
- Scaling: $||D\mathbf{p}|| \leq \Delta_k$. This created elliptical trust regions, which reduce the problem of different scaling of some variables.

Other methods

- Conjugate Gradient Methods: A set of nonzero vectors $\{\mathbf{p}_0, \mathbf{p}_1, \ldots, ...\mathbf{p}_n\}$ are conjugate wrt to a symmetric positive definite matrix A if $\mathbf{p}_i^T A \mathbf{p}_j = 0$, for all $i \neq j$.
- Quasi-Newton Methods: Use changes in gradient information to estimate a model of the function in order to achive superlinear convergence. Example: $B_{k+1}\alpha_k \mathbf{p}_k = \nabla f_{k+1} - \nabla f_k$ (BFGS Method).
- Derivative-free methods.
- Heuristic methods.