Trust Region Methods for Unconstrained Optimisation

Lecture 9, Numerical Linear Algebra and Optimisation
Oxford University Computing Laboratory, MT 2007
Dr Raphael Hauser (hauser@comlab.ox.ac.uk)

The Trust Region Framework

For the purposes of this lecture we once again consider the unconstrained minimisation problem

$$\text{(UCM)} \quad \min_{x \in \mathbb{R}^n} f(x),$$

where $f \in C^1(\mathbb{R}^n, \mathbb{R})$ with Lipschitz continous gradient $g(x)$.

- In practice, these smoothness assumptions are sometimes violated, but the algorithms we will develop are still observed to work well.

- As in Lecture 8, the algorithms we will construct are iterative descent methods that converge to a point where first and second order optimality conditions hold.

Iterative optimisation algorithms typically solve a much easier optimisation problem than (UCM) in each iteratiion.

- In the case of the line search methods of Lecture 8, the subproblems were easy because they are 1-dimensional.

- In the case of the trust-region methods we discuss today, the subproblems are n-dimensional but based on a simpler objective function – a linear or quadratic model – which is trusted in a simple region – a ball of specified radius in a specified norm.

Conceptually, the trust-region approach replaces a n-dimensional unconstrained optimisation problem by a n-dimensional constrained one. The replacement pays off because

1. The subproblem need not be solved to high accuracy, an approximate solution is enough.

2. The model function belongs to a class for which highly effective specialised algorithms have been developed.

Line Search vs Trust Region Methods:

- Line search methods:
  - pick descent direction p_k
  - pick stepsize α_k to “reduce” $f(x_k + \alpha p_k)$
  - $x_{k+1} = x_k + \alpha_k p_k$

- Trust-region methods:
  - pick step s_k to reduce “model” of $f(x_k + s)$
  - accept $x_{k+1} = x_k + s_k$ if the decrease promised by the model is inherited by $f(x_k + s_k)$,
  - otherwise set $x_{k+1} = x_k$ and improve the model.

The Trust-Region Subproblem:

We model $f(x_k + s)$ by either of the following,

- linear model

$$\mathsf{m}_k^{\mathsf{L}}(s) = f_k + s^\top g_k,$$

- quadratic model – (choose a symmetric matrix B_k)

$$\mathsf{m}_k^{\mathsf{Q}}(s) = f_k + s^\top g_k + \frac{1}{2} s^\top B_k s$$

Challenges:

- Models may not resemble $f(x_k + s)$ if s is large.

- Models may be unbounded from below:

  - m^{L} always unless $g_k = 0$,

  - m^{Q} always if B_k is indefinite, and possibly if B_k is only positive semi-definite.

To prevent both problems, we impose a *trust-region constraint*

$$\|s\| \le \Delta_k$$

for some suitable scalar radius $\Delta_k > 0$ and norm $\|\cdot\|$.

Therefore, the *trust-region subproblem* is the constrained optimisation problem

$$\text{(TRS)} \quad \min_{s \in \mathbb{R}^n} \ \mathrm{m}_k(s) \quad \text{s.t.} \quad \|s\| \le \Delta_k.$$

In theory the success of the method does not depend on the choice of the norm $\|\cdot\|$, but in practice is can!

For simplicity, we concentrate on the quadratic (Newton-like) model

$$\mathsf{m}_k(s) = \mathsf{m}_k^Q(s) = f_k + s^\top g_k + \frac{1}{2} s^\top B_k s$$

and any trust-region norm $\|\cdot\|$ for which

$$\kappa_s \|\cdot\| \le \|\cdot\|_2 \le \kappa_l \|\cdot\|$$

for some $\kappa_l \ge \kappa_s > 0$.

Norms on $\mathbb{R}^n$ we might want to consider:

- $\|\cdot\|_2 \le \|\cdot\|_2 \le \|\cdot\|_2$,
- $n^{-\frac{1}{2}} \|\cdot\|_1 \le \|\cdot\|_2 \le \|\cdot\|_1$,
- $\|\cdot\|_\infty \le \|\cdot\|_2 \le n \|\cdot\|_\infty$.

Choice of B_k:

$B_k = H_k$ is allowed but may be impractical (due to the problem dimension) or undesirable (due to indefiniteness).

As an alternative, any of the Hessian "approximations" discussed in Lecture 7 can be used.

Algorithm 1. [Basic Trust-Region Method]

1. Initialisation: Set $k = 0$, $\Delta_0 > 0$ and choose starting point x_0 by educated guess. Fix $\eta_v \in (0, 1)$ (typically, $\eta_v = 0.9$), $\eta_s \in (0, \eta_v)$ (typically, $\eta_s = 0.1$), $\gamma_i \geq 1$ (typically, $\gamma_i = 2$), and $\gamma_d \in (0, 1)$ (typically, $\gamma_d = 0.5$).

2. Until "convergence" repeat

   i) Build a quadratic model $\mathrm{m}(s)$ of $s \mapsto f(x_k + s)$.

   ii) Solve the trust-region subproblem approximately to find s_k for which $m(s_k)$ "$<$" f_k and $\|s_k\| \leq \Delta_k$, and define

   $$\rho_k = \frac{f_k - f(x_k + s_k)}{f_k - \mathrm{m}_k(s_k)}.$$

   iii) If $\rho_k \geq \eta_v$ ("very successful" TR step), set $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \gamma_i \Delta_k$.

   iv) Else, if $\rho_k \geq \eta_s$ ("successful" TR step), set $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \Delta_k$.

   v) Else ($\rho_k < \eta_s$, "unsuccessful" TR step), set $x_{k+1} = x_k$ and $\Delta_{k+1} = \gamma_d \Delta_k$.

   vi) Increase k by 1.

The Effect of Approximately Solving the TRS

Each trust-region subproblem has to be solved approximately, and this approximate solution should be obtained cheaply.

In order to be able to guarantee convergence of the overall method, we want to aim at the very least for an approximate solution that achieves as much reduction in the model as would a steepest descent step constrained by the trust-region:

- The *Cauchy point* is defined by $s_k^c := -\alpha_k^c g_k$, where

$$\begin{aligned} \alpha_k^c &:= \arg\min\{\mathrm{m}_k(-\alpha g_k) :\ \alpha > 0,\ \alpha\|g_k\| \le \Delta_k\} \\ &= \arg\min\{\mathrm{m}_k(-\alpha g_k) :\ 0 < \alpha \le \Delta_k/\|g_k\|\}. \end{aligned}$$

  Computing the C.p. is very easy (minimise a quadratic over a line segment).

- For the approximate solution of the trust region subproblem we then require that

$$\mathrm{m}_k(s_k) \le \mathrm{m}_k(s_k^c) \text{ and } \|s_k\| \le \Delta_k.$$

  In practice, hope to do far better than this.

Convergence Theory for TRM with Approximate Solves:

Theorem 2. *[Achievable Model Decrease]*
Let $\mathsf{m}_k(s)$ *be a quadratic model of* f *and* s_k^c *its Cauchy point within the trust-region* $\{s : \|s\| \leq \Delta_k\}$*. Then the achievable model decrease is at least*

$$f_k - \mathsf{m}_k(s_k^c) \geq \frac{1}{2}\|g_k\|_2 \cdot \min\left[\frac{\|g_k\|_2}{1 + \|B_k\|_2}, \kappa_s \Delta_k\right].$$

Corollary 3. *Let* $\mathsf{m}_k(s)$ *be a quadratic model of* f *and* s_k *an improvement on its Cauchy point within the trust-region* $\{s : \|s\| \leq \Delta_k\}$*. Then*

$$f_k - \mathsf{m}_k(s_k) \geq \frac{1}{2}\|g_k\|_2 \cdot \min\left[\frac{\|g_k\|_2}{1 + \|B_k\|_2}, \kappa_s \Delta_k\right].$$

Further, if the trust region step s_k *is "very successful", then*

$$f_k - f_{k+1} \geq \eta_v \frac{1}{2}\|g_k\|_2 \cdot \min\left[\frac{\|g_k\|_2}{1 + \|B_k\|_2}, \kappa_s \Delta_k\right].$$

Lemma 4. *[Difference between Model and Function]*
Let $f \in C^2$, and let there exist some constants $\kappa_h \geq 1$ and $\kappa_b \geq 0$ such that $\|H_k\|_2 \leq \kappa_h$ and $\|B_k\|_2 \leq \kappa_b$ for all k. Then

$$|f(x_k + s_k) - \mathsf{m}_k(s_k)| \leq \kappa_d \cdot \Delta_k^2, \quad (k \in \mathbb{N}),$$

where $\kappa_d = \frac{1}{2}\kappa_l^2(\kappa_h + \kappa_b)$.

Lemma 5. *[Ultimate Progress at Nonoptimal Points]*
Let $f \in C^2$*, and let there exist some constants* $\kappa_h \geq 1$ *and* $\kappa_b \geq 0$ *such that* $\|H_k\|_2 \leq \kappa_h$ *and* $\|B_k\|_2 \leq \kappa_b$ *for all* k*. Let* $\kappa_d = \frac{1}{2}\kappa_l^2(\kappa_h + \kappa_b)$*. If at iteration* k *we have* $g_k \neq 0$ *and*

$$\Delta_k \leq \|g_k\|_2 \cdot \min\left[\frac{1}{\kappa_s(\kappa_h + \kappa_b)}, \frac{\kappa_s(1-\eta_v)}{2\kappa_d}\right],$$

then iteration k *is very successful and* $\Delta_{k+1} \geq \Delta_k$*.*

Corollary 6. *TR Radius Won't Shrink to Zero at Nonoptimal Points]* *Let $f \in C^2$, and let there exist some constants $\kappa_h \geq 1$ and $\kappa_b \geq 0$ such that $\|H_k\|_2 \leq \kappa_h$ and $\|B_k\|_2 \leq \kappa_b$ for all k. Let $\kappa_d = \frac{1}{2}\kappa_l^2(\kappa_h + \kappa_b)$. If there exists a constant $\varepsilon > 0$ such that $\|g_k\|_2 \geq \varepsilon$ for all k, then*

$$\Delta_k \geq \kappa_\varepsilon := \varepsilon\gamma_d \cdot \min\left[\frac{1}{\kappa_s(\kappa_h + \kappa_b)}, \frac{\kappa_s(1\eta_v)}{2\kappa_d}\right], \quad \forall\, k.$$

Corollary 7. *[Possible Finite Termination] Let $f \in C^2$, and let both the true and model Hessians be bounded away from zero for all k. If the basic trust region method has only finitely many successful iterations, then $x_k = x^*$ and $g(x^*) = 0$ for all k large enough.*

Theorem 8. *[Global Convergence]*
Let $f \in C^2$, and let both the true and model Hessians be bounded away from zero for all k. Then one of the following cases occurs,

i) $g_k = 0$ *for some* $k \in \mathbb{N}$,

ii) $\lim_{k\to\infty} f_k = -\infty$,

iii) $\lim_{k\to\infty} g_k = 0$.

Methods for Solving the TR Subproblem

Let us now discuss how to solve the trust region subproblem

$$\min_{s \in \mathbb{R}^n} q(s) = s^\top g + \frac{1}{2} s^\top B s$$
$$\text{s.t.} \quad \|s\| \leq \Delta$$

such that the convergence theory above applies, that is, we aim to find $s^* \in \mathbb{R}^n$ such that

$$q(s^*) \leq q(s^c) \text{ and } \|s^*\| \leq \Delta.$$

Might solve

- exactly $\Rightarrow$ Newton-like method
- approximately $\Rightarrow$ steepest descent/conjugate gradients

From now on we choose the ℓ_2-norm to determine trust regions, so that we have to approximately solve

$$\text{(TRS)} \quad \min_{s \in \mathbb{R}^n} \{q(s) : \; s \in \mathbb{R}^n, \, \|s\|_2 \leq \Delta\},$$

where $q(s) = s^\top g + \frac{1}{2} s^\top B s$. The exact optimal solution can be characterised using the optimality conditions of Lecture 7:

Theorem 9. *Any* global *minimiser s^* of (TR) must satisfy*

i) $(B + \lambda^* \mathrm{I}) s^* = -g$,

ii) $B + \lambda^* \mathrm{I} \succeq 0$ *(positive semi-definite),*

iii) $\lambda^* \geq 0$,

iv) $\lambda^* \cdot (\|s^*\|_2 - \Delta) = 0$.

Furthermore, if $B + \lambda^* \mathrm{I} \succ 0$ (positive definite) then s^* is unique.

Exact solutions of (TRS):

1. If $B \succ 0$ and the solution of $Bs = -g$ satisfies $\|s\|_2 \leq \Delta$, then $s^* = s$, i.e., solve the symmetric positive definite linear system $Bs = -g$.

2. If B is indefinite or the solution to $Bs = -g$ satisfies $\|s\|_2 > \Delta$. Then solve the nonlinear system

$$(B + \lambda \mathrm{I})s = -g,$$
$$s^\top s = \Delta^2,$$

   for s and λ using Newton's method. Complications occur

   - possibly when multiple local solutions occur,
   - or when g is close to orthogonal to the eigenvector(s) corresponding to the most negative eigenvalue of B.

When n is large, factorisation to solve $Bs = -g$ may be impossible. However, we only need an approximate solution of (TRS), so use an iterative method.

Approximate solutions of (TRS):

1. Steepest descent leads to the Cauchy point s^c.

2. Use conjugate gradients to improve from s^c.

Issues to address:

- Staying in the trust region.

- Dealing with negative curvature.

Algorithm 10. [Conjugate Gradients to Minimise $q(s)$]

1. Initialisation: Set $s^{(0)} = 0$, $g^{(0)} = g$, $d^{(0)} = -g$ and $i = 0$.
2. Until $\|g^{(i)}\|_2$ is sufficiently small or breakdown occurs, repeat
   i) $\alpha^{(i)} = \|g^{(i)}\|_2^2 / [d^{(i)}]^\top B d^{(i)}$,
   ii) $s^{(i+1)} = s^{(i)} + \alpha^{(i)} d^{(i)}$,
   iii) $g^{(i+1)} = g^{(i)} + \alpha^{(i)} B d^{(i)}$,
   iv) $\beta^{(i)} = \|g^{(i+1)}\|_2^2 / \|g^{(i)}\|_2^2$,
   v) $d^{(i+1)} = -g^{(i+1)} + \beta^{(i)} d^{(i)}$,
   vi) increment i by 1.

Important features of conjugate gradients:

- $g^{(j)} = B s^{(j)} + g$ for $j = 0, \ldots, i$,
- $[d^{(j)}]^\top g^{(i+1)} = 0$ for $j = 0, \ldots, i$,
- $[g^{(j)}]^\top g^{(i+1)} = 0$ for $j = 0, \ldots, i$.
- $\alpha^{(i)} = \arg\min_{\alpha > 0} \; q(s^{(i)} + \alpha d^{(i)})$.

The following lemma motivates the truncated CG method we are about to introduce:

Lemma 11. *[Crucial Property of CG]*
Let Algorithm 10 be applied to minimize $q(s)$*. If* $[d^{(i)}]^\top B d^{(i)} > 0$ *for* $0 \le i \le k$*, then the iterates* $s^{(j)}$ *grow in 2-norm,*

$$\|s^{(j)}\|_2 < \|s^{(j+1)}\|_2, \quad (0 \le j \le k-1).$$

Algorithm 12. [Truncated CG to Minimise $q(s)$]
Apply CG steps as in Algorithm 10, but terminate at iteration i if either of the following occurs,

- $[d^{(i)}]^\mathsf{T} B d^{(i)} \leq 0$ (in this case the line search

$$\min_{\alpha>0} \; q(s^{(i)} + \alpha d^{(i)})$$

  is unbounded below).

- $\|s^{(i)} + \alpha^{(i)} d^{(i)}\|_2 > \Delta$ (in this case Lemma 11 implies that the solution lies on the TR boundary).

In both cases, stop with $s^* = s^{(i)} + \alpha^\mathsf{B} d^{(i)}$, where α^B is chosen as the positive root of

$$\|s^{(i)} + \alpha^\mathsf{B} d^{(i)}\|_2 = \Delta.$$

Since the first step of Algorithm 12 takes us to the Cauchy point $s^{(1)} = s^c$, and all further steps are descent steps, we have

$$q(s^*) \leq q(s^c) \text{ and } \|s^*\|_2 \leq \Delta.$$

Therefore, our convergence theory applies and the TR algorithm with truncated CG solves converges to a first-order stationary point.

When q is convex, Algorithm 12 is very good:

Theorem 13. *Let B be positive definite and let Algorithm 12 be applied to the minimisation of $q(s)$. Let s^* be the computed solution, and let s^{M} be the exact solution of the (TRS). Then*

$$q(s^*) \leq \frac{1}{2} q(s^{\mathsf{M}}).$$

Note that $q(0) = 0$, so that $q(s^{\mathsf{M}}) \leq 0$ and $|q(s^{\mathsf{M}})|$ is the achievable model decrease. Theorem 13 says that at least half the achievable model decrease is realised.

In the non-convex case Algorithm 12 may yield a poor solution with respect to the achiebable model decrease: For example, if $g = 0$ and B is indefinite, then $q(s^*) = 0$. In this case use Lanczos' method to move around trust-region boundary – effective in practice.