Global Convergence Technique

GIAN Short Course on Optimization: Applications, Algorithms, and Computation

Sven Leyffer

Argonne National Laboratory

September 12-24, 2016

Outline

Global Convergence Techniques

Still consider

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable.

Question

How can we ensure convergence from remote starting points?

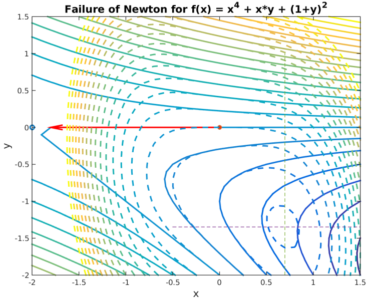
Methods can fail if step is too large ... or too small

Two mechanisms restrict steps:

1. Line-Search Methods ... search along descend direction $s^{(k)}$
2. Trust-Region Methods ... restrict computation of step.

Both converge, because steps revert to steepest descend.

Failures of Newton's Method



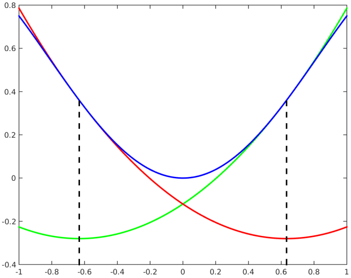
Failure of Newton

$f(x) = x_1^4 + x_1 x_2 + (1 + x_2)^2)$

No descend direction

$$\underset{x}{\text{minimize}} \;\; f(x) = x^2 - \frac{1}{4}x^4.$$

Alternates $-\sqrt{2/5}$ and $\sqrt{2/5}$.



Step too large

Outline

General Line-Search Method

Recall line-search method for $\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$

General Line-Search Method

Let $\sigma > 0$ constant. Given $x^{(0)}$, set $k = 0$.

repeat

Find search direction $s^{(k)}$ such that ${s^{(k)}}^T g(x^{(k)}) < 0$.

Compute steplength α_k such that Wolfe condition holds.

Set $x^{(k+1)} := x^{(k)} + \alpha_k s^{(k)}$ and $k = k + 1$.

until $x^{(k)}$ *is (local) optimum*;

Wolfe Line-Search Conditions

$$f(x^{(k)} + \alpha_k s^{(k)}) - f^{(k)} \leq \delta \alpha_k {g^{(k)}}^T s^{(k)}$$

$$g(x^{(k)} + \alpha_k s^{(k)})^T s^{(k)} \geq \sigma {g^{(k)}}^T s^{(k)}.$$

Illustration of Wolf Conditions

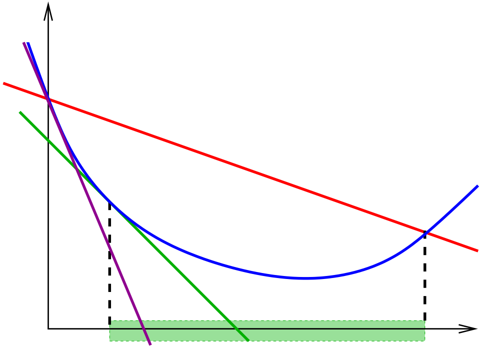
Wolfe Line-Search Conditions

$$f(x^{(k)} + \alpha_k s^{(k)}) \leq f^{(k)} + \delta \alpha_k {g^{(k)}}^T s^{(k)}$$

$$g(x^{(k)} + \alpha_k s^{(k)})^T s^{(k)} \geq \sigma {g^{(k)}}^T s^{(k)}$$

Slope at $x^{(k)}$ in direction $s^{(k)}$ is ${s^{(k)}}^T g^{(k)}$

- 1st condition requires sufficient decrease
- 2nd condition moves $x^{(k+1)}$ away from $x^{(k)}$



General Line-Search Method

Theorem (Convergence of Line-Search Methods)

- *$f(x)$ continuously differentiable and gradient*
- *$g(x) = \nabla f(x)$ Lipschitz continuous on $\mathbb{R}^n$.*

Then, one of three outcomes applies:

1. *finite termination: $g^{(k)} = 0$ for some $k > 0$, or*
2. *unbounded iterates: $\lim_{k \to \infty} f^{(k)} = -\infty$, or*
3. *directional convergence:*

$$\lim_{k \to \infty} \min \left(\left| {s^{(k)}}^T g^{(k)} \right|, \frac{\left| {s^{(k)}}^T g^{(k)} \right|}{\left\| s^{(k)} \right\|} \right) = 0.$$

The third outcome only somewhat successful:
... in the limit there is no descend along $s^{(k)}$.

General Line-Search Method

Corollary (Convergence of Steepest Descend Method)

- *$f(x)$ continuously differentiable and gradient*
- *$g(x) = \nabla f(x)$ Lipschitz continuous on $\mathbb{R}^n$.*

Then steepest descend algorithm results in:

1. *finite termination: $g^{(k)} = 0$ for some $k > 0$, or*
2. *unbounded iterates: $\lim_{k\to\infty} f^{(k)} = -\infty$, or*
3. *convergence to a stationary point: $\lim_{k\to\infty} g^{(k)} = 0$.*

Strengthen descend condition from ${s^{(k)}}^T g(x^{(k)}) < 0$ to

$$ {s^{(k)}}^T g(x^{(k)}) < -\sigma \| g(x^{(k)}) \|^2 $$

... $s^{(k)}$ has σ component of steepest descend direction
$\Rightarrow$ any line-search method with stronger descend converges.

Outline

Trust-Region Methods

More conservative than line-search methods:

- Computation of search direction inside a trust-region
- Revert to steepest descend as trust-region is reduced
- Computationally more expensive per iteration

... enjoy stronger convergence properties

Motivation for Trust-Region Methods

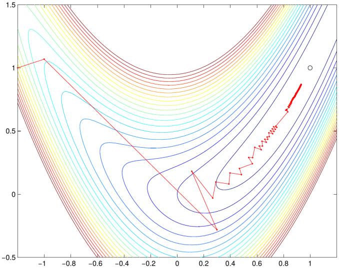
- Taylor model around $x^{(k)}$ accurate in neighborhood of $x^{(k)}$
- Minimize Taylor model inside some neighborhood.

How to define neighborhood?

- Depends on function
- Shape may be very complex

Use simple trust-region:

$$\|x - x^{(k)}\|_2 \leq \Delta_k$$



Trust-Region Methods

Trust-region method for $\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$

Basic Idea of Trust-Region Methods

1. Minimize model of $f(x)$ inside trust-region $\|x - x^{(k)}\|_2 \leq \Delta_k$
2. Move to new point, if we make progress
3. Reduce radius Δ_k, if we do not make progress

Trust-Region Methods

Trust-region models for

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

Trust-Region Models

- Linear model:

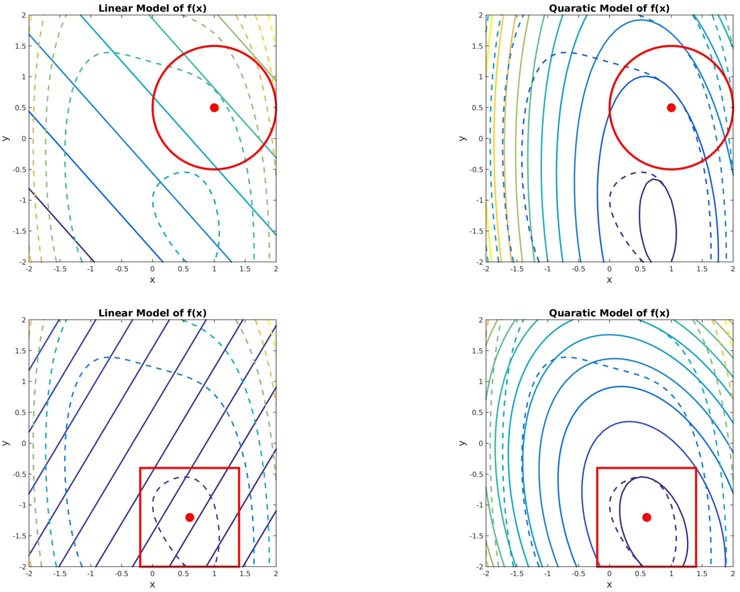
$$l_k(s) = f^{(k)} + s^T g^{(k)} \quad \simeq \quad f(x^{(k)} + s)$$

- Quadratic model

$$q_k(s) = f^{(k)} + s^T g^{(k)} + \frac{1}{2} s^T B^{(k)} s \quad \simeq \quad f(x^{(k)} + s)$$

where $f^{(k)} = f(x^{(k)})$, $g^{(k)} = \nabla f(x^{(k)})$, and $B^{(k)} \approx \nabla^2 f(x^{(k)})$

Illustration of Linear/Quadratic Trust-Region Models



Quadratic Trust-Region Subproblem

Quadratic trust-region subproblem

$$\underset{s}{\text{minimize}} \;\; q_k(s) = f^{(k)} + s^T g^{(k)} + \frac{1}{2} s^T B^{(k)} s \quad \text{subject to } \|s\|_2 \leq \Delta_k$$

... only needs to be solved "approximately" ... more later!

ℓ_2 norm is natural choice for unconstrained optimization.

M-norm for positive definite matrix, M, is a useful alternative:

$$\|x - x^{(k)}\|_M := \sqrt{\left(x - x^{(k)}\right)^T M \left(x - x^{(k)}\right)} \leq \Delta_k \quad M\text{-norm TR}$$

- Mitigates poor scaling of variables
- Trust-region subproblem easy to solve
- Interpret M as a preconditioner for trust-region subproblem

Trust-Region Radius Adjustment

Adjust Δ_k based on agreement of actual and predicted reduction

$$r_k := \frac{\text{actual reduction}}{\text{predicted reduction}} := \frac{f^{(k)} - f(x^{(k)} + s^{(k)})}{f^{(k)} - q_k(s^{(k)})}$$

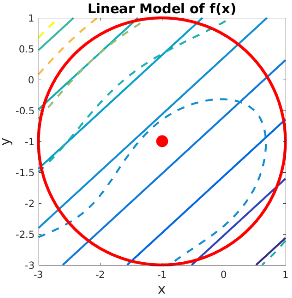
- $r_k \approx 1 \Rightarrow q_k(s)$ close to $f(x)$ .... accept
- $r_k < 0 \Rightarrow f(x)$ increases over step $s^{(k)}$ .... reject

Trust-Region Radius Adjustment

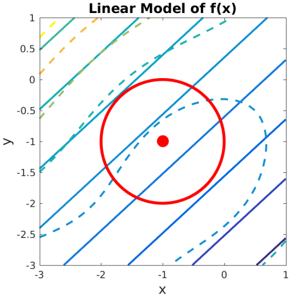
- If $r_k \geq \eta_s > 0$ then accept step & possibly increase Δ_k
- If $r_k < \eta_s$ then reject step & decrease Δ_k
  ... resolve TR subproblem to get better agreement, r_k

Trust-Region Radius Adjustment

Illustration of trust-region adjustment



reject



accept

General Trust-Region Method

Let $0 < \eta_s < \eta_v$ and $0 < \gamma_d < 1 < \gamma_i$.
Given $x^{(0)}$, set $k = 0$, initialize $\Delta_0 > 0$.
repeat
- Approximately solve the trust-region subproblem.
- Compute $r_k = \frac{f^{(k)} - f(x^{(k)} + s^{(k)})}{f^{(k)} - q_k(s^{(k)})}$.
- **if** $r_k \geq \eta_v$ *very successful step* **then**
  - Accept the step $x^{(k+1)} := x^{(k)} + s^{(k)}$.
  - Increase the trust-region radius, $\Delta_{k+1} := \gamma_i \Delta_k$.
- **else if** $r_k \geq \eta_s$ *successful step* **then**
  - Accept the step $x^{(k+1)} := x^{(k)} + s^{(k)}$.
  - Keep the trust-region radius unchanged, $\Delta_{k+1} := \Delta_k$.
- **else if** $r_k < \eta_s$ *unsuccessful step* **then**
  - Reject the step $x^{(k+1)} := x^{(k)}$.
  - Decrease the trust-region radius, $\Delta_{k+1} := \gamma_d \Delta_k$.
- **end**
- Set $k = k + 1$.

until $x^{(k)}$ *is (local) optimum*;

General Trust-Region Method

Reasonable values for Trust-Region parameters:

- Very successful step agreement: $\eta_v = 0.9$ or 0.99
- Successful step agreement: $\eta_s = 0.1$ or 0.01,
- Trust-region increase/decrease factors $\gamma_i = 2, \gamma_d = {}^1\!/_2$

Do not increase trust-region radius, unless step is on boundary

Trust-region algorithm much simpler than previous methods

- Computational difficulty hidden in subproblem solve
- Must be careful to solve TR subproblem efficiently.

The Cauchy Point & Steepest Descend Directions

Use steepest descend for minimalist conditions on TR subproblem

Definition (Cauchy Point)

Cauchy point: minimizer of model in steepest descend direction

$$\begin{aligned} \alpha_c := & \underset{\alpha}{\text{argmin}} \; q_k(-\alpha g^{(k)}) \text{ subject to } 0 \le \alpha \|g^{(k)}\| \le \Delta_k \\ = & \underset{\alpha}{\text{argmin}} \; q_k(-\alpha g^{(k)}) \text{ subject to } 0 \le \alpha \le \frac{\Delta_k}{\|g^{(k)}\|}. \end{aligned}$$

then Cauchy point is $s_c^{(k)} = -\alpha_c g^{(k)}$

- Cauchy point is cheap to compute
- Cauchy point is minimalistic assumption for convergence:

$$q_k(s^{(k)}) \le q_k(s_c^{(k)}) \quad \text{and} \quad \|s^{(k)}\| \le \Delta_k$$

Outline of Convergence of Trust-Region Methods

Outline of convergence proof ... ideas apply in other areas

1. Lower bound on predicted reduction from Cauchy point:

$$\text{pred. reduct.} \quad f^{(k)} - q_k(s_c^{(k)}) \geq \frac{1}{2} \|g^{(k)}\|_2 \min\left(\frac{\|g^{(k)}\|_2}{1 + \|B^{(k)}\|}, \kappa \Delta_k\right).$$

2. Corollary TR subproblem solution $s^{(k)}$, satisfies lower bound.
   - TR step makes at least as much progress as $s_c^{(k)}$
3. Bound agreement between objective and quadratic model:

$$\left| f(x^{(k)} + s^{(k)}) - q_k(s^{(k)}) \right| \leq \kappa \Delta_k^2,$$

$\kappa > 0$ depends Hessian bounds ... from Taylor's theorem.

Outline of Convergence of Trust-Region Methods

Cont. outline of convergence proof ...

1. Crucial Result

   Can always make progress from non-critical point $g^{(k)} \neq 0$:

   $$\text{If} \quad \Delta_k \leq \|g^{(k)}\|_2 \kappa(1-\eta_v), \quad \text{then very successful step}$$

   ... and $\Delta_{k+1} \geq \Delta_k$

   - Here $\kappa(1-\eta_v)$ constant
   - η_s threshold for very successful step

   Intuitive: reducing Δ gives better agreement

   ... make progress with $r_k \simeq 1$

2. If gradient norm bounded away from zero, i.e. $\|g^{(k)}\| \geq \epsilon > 0$,

   ... then trust-region radius also bounded away from zero:

   $$\|g^{(k)}\| \geq \epsilon > 0 \;\Rightarrow\; \Delta_k \geq \epsilon\kappa(1-\eta_v).$$

3. If number of iteration finite, then final iterate is stationary.

Outline of Convergence of Trust-Region Methods

Summarize results in theorem ...

Theorem (Convergence of TR Method with Cauchy Condition)

$f(x)$ twice continuously differentiable and Hessian matrices $B^{(k)}, H^{(k)}$ bounded. Then, TR algorithm has on of three outcomes:

1. *finite termination: $g^{(k)} = 0$ for some $k > 0$, or*
2. *unbounded iterates: $\lim_{k\to\infty} f^{(k)} = -\infty$, or*
3. *convergence to a stationary point: $\lim_{k\to\infty} g^{(k)} = 0$.*

Solving the Trust-Region Subproblem

Remarkable Result about TR Subproblem

With ℓ_2-norm TR, can solve TR subproblem to global optimality.

Theorem

Global minimizer, s^*, of trust-region subproblem,

$$\underset{s}{\text{minimize}} \; q(s) := f + g^T s + \frac{1}{2} s^T B s \quad \text{subject to } \|s\|_2 \le \Delta$$

satisfies $(B + \lambda^* I)s^* = -g$ *, where*

- $B + \lambda^* I$ *positive definite,*
- $\lambda^* \ge 0$*, and*
- $\lambda^*(\|s^*\|_2 - \Delta) = 0$.

Moreover, if $B + \lambda^* I$ is positive definite, then s^* is unique.

Solving the Trust-Region Subproblem

Theorem

Global minimizer, s^*, of trust-region subproblem,

$$\underset{s}{\text{minimize}} \;\; q(s) := f + g^T s + \tfrac{1}{2} s^T B s \quad \text{subject to } \|s\|_2 \le \Delta$$

satisfies

$$(B + \lambda^* I) s^* = -g,$$

where $B + \lambda^* I$ positive definite, $\lambda^* \ge 0$, and $\lambda^*(\|s^*\|_2 - \Delta) = 0$. Moreover, if $B + \lambda^* I$ is positive definite, then s^* is unique.

- Necessary and sufficient conditions for global minimizer
- Optimality conditions are KKT conditions of TR subproblem.
- Suggest way to solve TR subproblem to global optimality

Solving the Trust-Region Subproblem

Divide solution of TR subproblem,

$$\underset{s}{\text{minimize}} \;\; q(s) := f + g^T s + \frac{1}{2} s^T B s \quad \text{subject to } \|s\|_2 \le \Delta$$

into two cases:

1. B pos. def. and solution of $Bs = -g$, satisfies $\|s\| \le \Delta$
2. B not pos. def. or solution of $Bs = -g$, satisfies $\|s\| > \Delta$

Case 1: B positive def., and $Bs = -g$, satisfies $\|s\| \le \Delta$

- Solution s is global solution of TR subproblem

... modern factorization routines detect positive definiteness

Solving the Trust-Region Subproblem

Trust-region subproblem

$$\underset{s}{\text{minimize}}\ \ q(s) := f + g^T s + \tfrac{1}{2} s^T B s \quad \text{subject to } \|s\|_2 \le \Delta$$

Case 2: B not pos. def. or solution of $Bs = -g$, satisfies $\|s\| > \Delta$

Optimality conditions of TR subproblem: (s^*, λ^*) satisfies

$$(B + \lambda I)s = -g \quad \text{and} \quad s^T s = \Delta^2,$$

set of $(n+1)$ linear/quadratic equations in $(n+1)$ unknowns.

Methods for solving linear/quadratic equation:

- Compute Cholesky factors of $B + \lambda I$
- Eliminate s from quadratic equation
- Solve nonlinear equation for λ ... repeat

... need to be careful in certain difficult cases.

Solving Large-Scale Trust-Region Subproblems

Trust-region subproblem

$$\underset{s}{\text{minimize}} \;\; q(s) := f + g^T s + \tfrac{1}{2} s^T B s \quad \text{subject to } \|s\|_2 \leq \Delta$$

Cholesky factors are computationally impractical for large n

$\Rightarrow$ consider iterative methods for solving TR subproblem

- Conjugate gradients good choice
  ... first step is steepest descend consistent with Cauchy step!
- Get convergence to stationary points for "free"

Adapting Conjugate Gradient to TR constraint

- What is the interaction between iterates and the trust region?
- What do we do, if B is indefinite?

Solving Large-Scale Trust-Region Subproblems

Trust-Region Subproblem Conjugate-Gradient Method

Set $s^{(0)} = 0$, $g^{(0)} = g$, $d^{(0)} = -g$, and $i = 0$.

repeat

- Exact line search: $\alpha_i = \|g^{(i)}\|^2 / (d^{(i)^T} B d^{(i)})$
- New iterate: $s^{(i+1)} = s^{(i)} + \alpha_i d^{(i)}$
- Gradient update: $g^{(i+1)} = g^{(i)} + \alpha_i B d^{(i)}$
- Fletcher-Reeves: $\beta_i = \|g^{(i+1)}\|^2 / \|g^{(i)}\|^2$
- New search direction: $d^{(i+1)} = -g^{(i+1)} + \beta_i d^{(i)}$
- Set $i = i + 1$.

until *Breakdown or small* $\|g^{(i)}\|$ *found*;

Breakdown: needs to be defined (reach TR or indefinite)

Solving Large-Scale Trust-Region Subproblems

$$\underset{s}{\text{minimize}}\ q(s) := f + g^T s + \frac{1}{2} s^T B s \quad \text{subject to } \|s\|_2 \le \Delta$$

What is the interaction between iterates and the trust region?

Theorem

Apply conjugate-gradient to trust-region subproblem, assume ${d^{(i)}}^T B d^{(i)} > 0$ *for all* $0 \le i \le k$. *Then*

$$\|s^{(i)}\|_2 \le \|s^{(i+1)}\|_2 \quad \forall\, 0 \le i \le k.$$

- If $\|s^{(i)}\| > \Delta$ at iteration i,
  ... then subsequent iterates lie outside TR too.
- Once we pass TR boundary, then we know that $\|s^*\| = \Delta$

Solving Large-Scale Trust-Region Subproblems

$$\underset{s}{\text{minimize}} \;\; q(s) := f + g^T s + \tfrac{1}{2} s^T B s \quad \text{subject to } \|s\|_2 \le \Delta$$

Termination Conditions for TR Conjugate Gradient

Terminate CG solution of TR subproblem, if

1. Find non-positive curvature: $d^{(i)^T} B d^{(i)} \le 0$:
   $\Rightarrow$ $q(s)$ is unbounded along $d^{(i)}$.
2. Generate iterate outside TR
   $\Rightarrow$ all subsequent iterates lie outside the TR

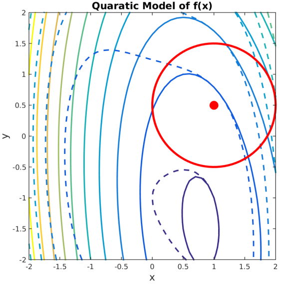
If $\|s^{(i+1)}\| > \Delta$, then compute step to boundary solving for α^B:

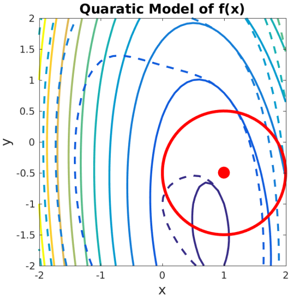
$$\|s^{(i)} + \alpha^B d^{(i)}\|_2^2 = \Delta^2.$$

Approach OK convex case, poor for nonconvex $f(x)$.
Prefer more elaborate Lanczos method for nonconvex $f(x)$.

Conclusions

Introduction to Trust-Region Methods





- Minimize model of $f(x)$ inside trust-region $\|x - x^{(k)}\| \leq \Delta_k$
- Measure progress ratio

$$r = \frac{\text{actual reduct.}}{\text{predicted reduct.}}$$

- Accept step if good progress
- Reject step if poor progress ... and reduce Δ_k
- Solve TR subproblem to global optimality