Convergence of trust-region method for unconstrained smooth optimization

Andersen Ang
ECS, Uni. Southampton, UK
`andersen.ang@soton.ac.uk`

Homepage `angms.science`

Version: May 19, 2023
First draft: Nov 30, 2022

Content

Quadratic model $m(\boldsymbol{s};\boldsymbol{x}) := f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2$

Sufficient descent: $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$ then $\Delta m(\boldsymbol{s}) \geq \frac{\|\nabla f(\boldsymbol{x})\|_2}{2} \min\left\{ \frac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|}, \delta \right\}$

Theory of TR convergence

1. $f - m$ gap: $\left| f(\boldsymbol{x}+\boldsymbol{s}) - m(\boldsymbol{s};\boldsymbol{x}) \right| \leq \frac{\kappa_H + \kappa_B}{2}\delta^2$
2. Progress (small radius $\implies$ success): $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}$, $\delta_k \leq \frac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H + \kappa_B} \min(1, 1 - \eta_{vs}) \implies k \in \mathcal{V}, \delta_{k+1} \geq \delta_k$
3. TR radius will not shrink to 0 at non-sol.
4. Possible finite termination
5. Global convergence of some subsequence

⚠ These notes assume you have seen trust-region method (TRM)
You should be familiar with terms: model, radius, success, model decrease, actual decrease

Table of Contents

Quadratic model $m(\boldsymbol{s};\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \right\rangle + \frac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2$

Sufficient descent: $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$ then $\Delta m(\boldsymbol{s}) \ \geq \ \frac{\|\nabla f(\boldsymbol{x})\|_2}{2} \min\left\{ \frac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|}, \delta \right\}$

Theory of TR convergence

1. $f - m$ gap: $\left| f(\boldsymbol{x}+\boldsymbol{s}) - m(\boldsymbol{s};\boldsymbol{x}) \right| \ \leq \ \frac{\kappa_H + \kappa_B}{2} \delta^2$
2. Progress (small radius $\Longrightarrow$ success): $\nabla f(\boldsymbol{x}_k) \neq \mathbf{0}, \delta_k \leq \frac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H + \kappa_B} \min\left(1, 1 - \eta_{vs}\right) \ \Longrightarrow \ k \in \mathcal{V}, \delta_{k+1} \geq \delta_k$
3. TR radius will not shrink to 0 at non-sol.
4. Possible finite termination
5. Global convergence of some subsequence

Problem setup: smooth unconstrained optimization

$$(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} \ f(\boldsymbol{x}).$$

- $\mathcal{C}^2 \ni f : \mathbb{R}^n \to \mathbb{R}$
  - $\operatorname{dom} f = \mathbb{R}^n$ and the target of f is $\mathbb{R}$. standard Euclidean space and inner product
  - f is twice differentiable. $f \in \mathcal{C}^2$
    - For all point $\boldsymbol{\xi}$, we have gradient $\nabla_{\boldsymbol{x}} f(\boldsymbol{\xi})$ and Hessian $\boldsymbol{H}(\boldsymbol{\xi}) \coloneqq \nabla_{\boldsymbol{x}\boldsymbol{x}} f(\boldsymbol{\xi})$ and they are continuous
  - f is possibly nonconvex
- $\boldsymbol{x} \in \mathbb{R}^n$ is the optimization variable. No constraint: all $\boldsymbol{x} \in \mathbb{R}^n$ feasible.
- Solve $\mathcal{P}$ by iterative method: starting from $\boldsymbol{x}_0$, generate a sequence $\{\boldsymbol{x}_k\}_{k \in \mathbb{N}}$.
  Two ways:
  - Gradient descent / line-search method
  - Trust-region method ← our focus here

Review of gradient descent (GD)

$$\boxed{(\mathcal{P}) \ : \ \min_{\boldsymbol{x}} f(\boldsymbol{x})}$$

- Line search: generate $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}}$ as $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$.
    - $\alpha_k \in \mathbb{R}_{++}$ stepsize
    - $\boldsymbol{d}_k \in \mathbb{R}^n$ update direction
    - GD set $\boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$
- Understanding GD: $\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$ comes from a local quadratic model $m(\boldsymbol{\xi}; \boldsymbol{x}, \alpha)$, and GD is doing is $\boldsymbol{x}_{k+1} = \underset{\boldsymbol{\xi}}{\operatorname{argmin}}\, m(\boldsymbol{\xi}; \boldsymbol{x}_k, \alpha_k)$

$$\boldsymbol{x} - \alpha\nabla f(\boldsymbol{x}) \ = \ \underset{\boldsymbol{\xi}}{\operatorname{argmin}} \ m(\boldsymbol{\xi}; \boldsymbol{x}, \alpha) \coloneqq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{\xi} - \boldsymbol{x} \rangle + \frac{1}{2\alpha}\|\boldsymbol{\xi} - \boldsymbol{x}\|_2^2$$

  How to see it: take $\nabla_{\boldsymbol{\xi}} m(\boldsymbol{\xi}; \boldsymbol{x}, \alpha) = \mathbf{0}$, see here for details.

- See angms.science for more on GD
    - introduction
    - descent lemma
    - convergence on convex smooth function
    - convergence on strongly convex smooth function
    - projected gradient descent
    - proximal gradient descent

Trust-region (TR) method: a “dual” of GD

- **Notation**: $\delta > 0$ denotes the TR radius.

- Given a fixed δ, TR finds an update direction $\boldsymbol{s}$ via solving a *model* m

$$\boldsymbol{s} \;=\; \underset{\|\boldsymbol{s}\|\leq\delta}{\operatorname{argmin}}\; m(\boldsymbol{s};\boldsymbol{x}) \;:=\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \frac{1}{2}\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s}\rangle, \qquad \text{(TR subproblem)}$$

  and then perform the update

$$\boldsymbol{x}^{+} = \begin{cases} \boldsymbol{x}+\boldsymbol{s} & \text{if } f(\boldsymbol{x}+\boldsymbol{s}) \text{ ``} < \text{''}\ f(\boldsymbol{x}) \\ \boldsymbol{x} & \text{otherwise} \end{cases}$$

- We minimize m instead of f to get $\boldsymbol{s}$
    - $m(\boldsymbol{s};\boldsymbol{x})$ is a simple local approximation of f at $\boldsymbol{x}$
        - $m(\boldsymbol{s};\boldsymbol{x})$ may not resemble $f(\boldsymbol{x}+\boldsymbol{s})$ for big $\boldsymbol{s}$ $\implies$ limit $\|\boldsymbol{s}\| \leq \delta$
            - we have to choose a norm $\|\cdot\|$ (Euclidean may not be the best)
        - easier to find $\boldsymbol{s}$ via $\min m$ than $\min f$

          ⚠ TR subproblem can be hard to solve.

About $m(\boldsymbol{s};\boldsymbol{x}) := f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2}\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s} \rangle$

$$\boldsymbol{s} \;=\; \underset{\|\boldsymbol{s}\|\le\delta}{\operatorname{argmin}}\; m(\boldsymbol{s};\boldsymbol{x}) \;:=\; \underbrace{f(\boldsymbol{x})}_{\text{constant}} + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2}\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s} \rangle. \qquad \text{(TR subproblem)}$$

- argmin ignores constant:

$$\boldsymbol{s} \;=\; \underset{\|\boldsymbol{s}\|\le\delta}{\operatorname{argmin}}\; m(\boldsymbol{s};\boldsymbol{x}) \;:=\; \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2}\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s} \rangle.$$

- The constant term $f(\boldsymbol{x})$ is actually m with $\boldsymbol{s} = \boldsymbol{0}$

$$m(\boldsymbol{0};\boldsymbol{x}) = f(\boldsymbol{x}) \,. \qquad \text{(0}^{\text{th}}\text{-order equivalence)}$$

Equivalence between f and $m(\boldsymbol{s};\boldsymbol{x}) := f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \frac{1}{2}\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s}\rangle$

- "Tangential properties" / "Taylor-equivalence" / coincident property
  - 0th-order equivalence: $m(\mathbf{0};\boldsymbol{x}) = f(\boldsymbol{x})$. f and m coincide at current iterate
  - 1st-order equivalence: $\nabla_{\boldsymbol{s}} m(\boldsymbol{s};\boldsymbol{x})\Big|_{\boldsymbol{s}=\mathbf{0}} = \nabla f(\boldsymbol{x})$. $\mathrm{grad} f$ and $\mathrm{grad} m$ coincide at current iterate
  - 2nd-order equivalence: **If** $\boldsymbol{B} = \underbrace{\text{Hessian } \boldsymbol{H}(\boldsymbol{\xi}) \text{ of } f \text{ at } \boldsymbol{\xi} \in [\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{s}]}_{\text{this is mean value theorem}}$ (this assumes $f \in \mathcal{C}^2$)

    Then $\nabla^2_{\boldsymbol{s}} m(\boldsymbol{s};\boldsymbol{x})\Big|_{\boldsymbol{s}=\mathbf{0}} = \boldsymbol{H}(\boldsymbol{\xi})$.

- **Predicted decrease / model decrease**

$$\begin{aligned}\Delta m(\boldsymbol{s}) &:= m(\mathbf{0};\boldsymbol{x}) - m(\boldsymbol{s};\boldsymbol{x}) \\ &= f(\boldsymbol{x}) - \Big(f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \frac{1}{2}\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s}\rangle \Big) \\ &= -\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle - \frac{1}{2}\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s}\rangle.\end{aligned}$$

About $\boldsymbol{B}$

$$\boxed{\boldsymbol{s} \;=\; \underset{\|\boldsymbol{s}\|\le\delta}{\operatorname{argmin}}\; m(\boldsymbol{s};\boldsymbol{x}) \;:=\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \frac{1}{2}\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s}\rangle}$$

- $\boldsymbol{B} \in \mathbb{S}$: $\boldsymbol{B}$ is symmetric
    - Indefinite $\boldsymbol{B}$: TR subproblem is unbounded below.
    - Positive semi-definite $\boldsymbol{B}$: TR subproblem is possibly unbounded below.
        - This includes the case $\boldsymbol{B} = \mathbf{0}_{n\times n}$
        - recall $\mathbf{0}_{n\times n}$ is both positive semi-definite and negative semi-definite
        - $\boldsymbol{B} = \mathbf{0}_{n\times n}$: we have linear model m
    - Positive definite $\boldsymbol{B}$: TR subproblem is bounded below.
- If $\boldsymbol{B} = \boldsymbol{H}(\boldsymbol{x})$ (Hessian of f at $\boldsymbol{x}$) then we have a Newton-type quadratic model $\boldsymbol{m}$.
- Quasi-Newton approach use $\boldsymbol{B}$ to approximate $\boldsymbol{H}$.
- Importance of 🟥: in this case $\boldsymbol{s}^*$ is simply the extreme value in the constraint set $\|\boldsymbol{s}\| \le \delta$.

If $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$ (GD direction)

$$\boxed{\boldsymbol{s} \;=\; \underset{\|\boldsymbol{s}\| \leq \delta}{\operatorname{argmin}} \; m(\boldsymbol{s}; \boldsymbol{x}) \;:=\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2} \langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s} \rangle}$$

$$\begin{aligned} \alpha^* &= \underset{0 \leq \alpha \|\nabla f(\boldsymbol{x})\| \leq \delta}{\operatorname{argmin}} \; f(\boldsymbol{x}) + \big\langle \nabla f(\boldsymbol{x}), -\alpha \nabla f(\boldsymbol{x}) \big\rangle + \frac{1}{2} \big\langle \boldsymbol{B} \alpha \nabla f(\boldsymbol{x}), \alpha \nabla f(\boldsymbol{x}) \big\rangle \\ &= \underset{0 \leq \alpha \leq \frac{\delta}{\|\nabla f(\boldsymbol{x})\|}}{\operatorname{argmin}} \; \frac{\big\langle \boldsymbol{B} \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \big\rangle}{2} \alpha^2 - \|\nabla f(\boldsymbol{x})\|_2^2 \, \alpha \end{aligned}$$

- A simple quadratic scalar optimization problem

$$x = \underset{0 \leq x \leq u}{\operatorname{argmin}} \; a x^2 - b x \qquad b, u \geq 0$$

⚠ a can be negative if $\boldsymbol{B}$ is indefinite / semi-positive definite.

- $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$ is called *Cauchy point* in some books.

On $x = \underset{0 \le x \le u}{\operatorname{argmin}}\ ax^2 - bx$ with $b \geq 0, u \geq 0$

- **Case** $a \leq 0$
  Problem is unbounded below: $\underbrace{\underbrace{a}_{\leq 0} \underbrace{x^2}_{\geq 0}}_{\leq 0} - \underbrace{\underbrace{b}_{\geq 0} \underbrace{x}_{\geq 0}}_{\leq 0}$. Optimal x is at the boundary $x^* = u$.

- **Case** $a > 0$
  Completing the squares $ax^2 - bx \,=\, a\big(x^2 - \frac{b}{a}\big)$ gives

$$ax^2 - bx \,=\, a\Big(x^2 - \frac{b}{a} + \big(\frac{b}{2a}\big)^2 - \big(\frac{b}{2a}\big)^2\Big) \,=\, a\Big(\big(x - \frac{b}{2a}\big)^2 - \frac{b^2}{4a^2}\Big) \,=\, a\big(x - \frac{b}{2a}\big)^2 - \frac{b^2}{4a}.$$

  The minimum of the quadratic occurs at $x = \frac{b}{2a}$. Depends on where is $\frac{b}{2a}$, we have

$$x^* \;=\; \operatorname{median}\Big(0,\, \frac{b}{2a}\,, u\Big) \;=\; \begin{cases} 0 & \dfrac{b}{2a} \leq 0 \\ \dfrac{b}{2a} & 0 < \dfrac{b}{2a} \leq u \\ u & \dfrac{b}{2a} > u \end{cases}$$

Summary of TR-subproblem

$$\boldsymbol{s} \ = \ \operatorname*{argmin}_{\|\boldsymbol{s}\| \leq \delta} m(\boldsymbol{s}; \boldsymbol{x}) \ \coloneqq \ \underbrace{f(\boldsymbol{x})}_{=: m(\boldsymbol{0}; \boldsymbol{x})} + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2} \langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s} \rangle.$$

- *Predicted decrease* $\Delta m(\boldsymbol{s}) \coloneqq m(\boldsymbol{0}; \boldsymbol{x}) - m(\boldsymbol{s}; \boldsymbol{x}) = -\langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle - \frac{1}{2} \langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s} \rangle$
- If $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$,

$$\Delta m\big(-\alpha \nabla f(\boldsymbol{x})\big) = -\frac{\big\langle \boldsymbol{B} \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \big\rangle}{2} \alpha^2 + \|\nabla f(\boldsymbol{x})\|_2^2 \, \alpha.$$

Two cases

1. $\alpha^* = \frac{\delta}{\|\nabla f(\boldsymbol{x})\|}$ $\qquad \langle \boldsymbol{B} \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \rangle \leq 0$
2. $\alpha^* = \begin{cases} 0 & \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\langle \boldsymbol{B} \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \rangle} \leq 0 \\ \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\langle \boldsymbol{B} \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \rangle} & 0 < \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\langle \boldsymbol{B} \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \rangle} \leq \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} \\ \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} & \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\langle \boldsymbol{B} \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \rangle} > \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} \end{cases}$ $\qquad \langle \boldsymbol{B} \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \rangle > 0$

Usually we use positive definite $\boldsymbol{B}$ so $\alpha^* = 0$ is impossible.

Weighted norm

- We will see the term $\left\langle \boldsymbol{B}\nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \right\rangle$ many times.
- Shorthand notation: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{A}} := \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y} \rangle$ is called *weighted inner product* under the weight $\boldsymbol{A}$
- Weighted norm: $\|\boldsymbol{x}\|_{\boldsymbol{A}} := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{A}}} = \sqrt{\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle}$
- Weighted norm-squared: $\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{A}} = \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle$
- Easy careless-mistake: $\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 \neq \|\boldsymbol{A}\boldsymbol{x}\|_2^2$

$$\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{A}} = \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \rangle \quad \neq \quad \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle = \|\boldsymbol{A}\boldsymbol{x}\|_2^2$$

- Using weighted norm, $\left\langle \boldsymbol{B}\nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \right\rangle = \left\langle \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x}) \right\rangle_{\boldsymbol{B}} = \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2$

Summary of TR-subproblem, in weighted norm

$$\boldsymbol{s} \;=\; \operatorname*{argmin}_{\|\boldsymbol{s}\|\le\delta} m(\boldsymbol{s};\boldsymbol{x}) \;:=\; \underbrace{f(\boldsymbol{x})}_{=:m(\boldsymbol{0};\boldsymbol{x})} + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \frac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2.$$

- *Predicted decrease* $\Delta m(\boldsymbol{s}) \coloneqq m(\boldsymbol{0};\boldsymbol{x}) - m(\boldsymbol{s};\boldsymbol{x}) = -\langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle - \dfrac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2$
- If $\boldsymbol{s} = -\alpha\nabla f(\boldsymbol{x})$,

$$\Delta m\big(-\alpha\nabla f(\boldsymbol{x})\big) = -\frac{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}{2}\alpha^2 + \|\nabla f(\boldsymbol{x})\|_2^2\,\alpha.$$

  Two cases

  1. $\alpha^* = \frac{\delta}{\|\nabla f(\boldsymbol{x})\|}$ $\qquad \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \le 0$
  2. $\alpha^* = \begin{cases} 0 & \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \le 0 \\ \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} & 0 < \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \le \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} \\ \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} & \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} > \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} \end{cases}$ $\qquad \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 > 0$

  Usually we use positive definite $\boldsymbol{B}$ so $\alpha^* = 0$ is impossible.

Summary of TR-subproblem, in compact form

$$\boldsymbol{s} \ = \ \underset{\|\boldsymbol{s}\| \le \delta}{\operatorname{argmin}} \ m(\boldsymbol{s}; \boldsymbol{x}) \ \coloneqq \ \underbrace{f(\boldsymbol{x})}_{=: m(\boldsymbol{0}; \boldsymbol{x})} + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2} \|\boldsymbol{s}\|_{\boldsymbol{B}}^2 .$$

- *Predicted decrease* $\Delta m(\boldsymbol{s}) \coloneqq m(\boldsymbol{0}; \boldsymbol{x}) - m(\boldsymbol{s}; \boldsymbol{x}) = -\langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle - \dfrac{1}{2} \|\boldsymbol{s}\|_{\boldsymbol{B}}^2$

- If $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$,

$$\Delta m\big(- \alpha \nabla f(\boldsymbol{x})\big) = -\frac{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}{2} \alpha^2 + \|\nabla f(\boldsymbol{x})\|_2^2 \, \alpha .$$

  Two cases

  1. $\alpha^* = \frac{\delta}{\|\nabla f(\boldsymbol{x})\|}$ $\qquad \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \le 0$
  2. $\alpha^* = \operatorname{median}\Big(0, \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}, \frac{\delta}{\|\nabla f(\boldsymbol{x})\|}\Big)$ $\qquad \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 > 0$

  Usually we use positive definite $\boldsymbol{B}$ so $\alpha^* = 0$ is impossible.

Table of Contents

Quadratic model $m(\boldsymbol{s};\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2}\|\boldsymbol{s}\|_B^2$

Sufficient descent: $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$ then $\Delta m(\boldsymbol{s}) \;\geq\; \frac{\|\nabla f(\boldsymbol{x})\|_2}{2} \min\left\{ \frac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|}, \delta \right\}$

Theory of TR convergence

1. $f - m$ gap: $\left| f(\boldsymbol{x}+\boldsymbol{s}) - m(\boldsymbol{s};\boldsymbol{x}) \right| \;\leq\; \frac{\kappa_H + \kappa_B}{2} \delta^2$
2. Progress (small radius $\Longrightarrow$ success): $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}, \delta_k \leq \frac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H + \kappa_B} \min\left(1, 1 - \eta_{vs}\right) \implies k \in \mathcal{V}, \delta_{k+1} \geq \delta_k$
3. TR radius will not shrink to 0 at non-sol.
4. Possible finite termination
5. Global convergence of some subsequence

If $\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \leq 0$

$$\boxed{m(\boldsymbol{s};\boldsymbol{x}) \;:=\; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \frac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2}$$

- Form previous slide, $\alpha^* = \frac{\delta}{\|\nabla f(\boldsymbol{x})\|_2}$ if 🟥.
- Put $\boldsymbol{s} = -\alpha\nabla f(\boldsymbol{x})$ in $m(\boldsymbol{s};\boldsymbol{x})$

$$\begin{aligned}
m\big(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x}\big) &= f(\boldsymbol{x}) - \alpha\|\nabla f(\boldsymbol{x})\|_2^2 + \tfrac{\alpha^2}{2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 && (i)\\
\tfrac{\alpha^2}{2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 &\overset{\text{🟥}}{\leq} 0 && (ii)\\
m\big(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x}\big) &\leq f(\boldsymbol{x}) - \alpha\|\nabla f(\boldsymbol{x})\|_2^2 && (i)+(ii)\\
&= f(\boldsymbol{x}) - \delta\|\nabla f(\boldsymbol{x})\|_2 && \alpha^* = \tfrac{\delta}{\|\nabla f(\boldsymbol{x})\|_2} \text{ if } \text{🟥}\\
&\overset{m(\mathbf{0};\boldsymbol{x}) = f(\boldsymbol{x})}{=} m(\mathbf{0};\boldsymbol{x}) - \delta\|\nabla f(\boldsymbol{x})\|_2
\end{aligned}$$

Hence

$$\Delta m\big(-\alpha\nabla f(\boldsymbol{x})\big) := m(\mathbf{0};\boldsymbol{x}) - m\big(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x}\big) \;\geq\; \delta\|\nabla f(\boldsymbol{x})\|_2.$$

- We have:

$$\text{IF} \quad \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \leq 0 \quad \text{THEN} \quad \underbrace{\Delta m\big(-\alpha\nabla f(\boldsymbol{x})\big)}_{=:m(\mathbf{0};\boldsymbol{x}) - m(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x})} \;\geq\; \delta\|\nabla f(\boldsymbol{x})\|_2.$$

If $\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 > 0$, case 1

$$\boxed{\begin{aligned} m(\boldsymbol{s};\boldsymbol{x}) &:= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \tfrac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}} \\ m(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x}) &= f(\boldsymbol{x}) - \alpha\|\nabla f(\boldsymbol{x})\|_2^2 + \tfrac{\alpha^2}{2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \end{aligned}}$$

$$\alpha^* = \operatorname{median}\Big(0, \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}, \frac{\delta}{\|\nabla f(\boldsymbol{x})\|}\Big) = \begin{cases} 0 & \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \le 0 \\ \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} & 0 < \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \le \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} \\ \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} & \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} > \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} \end{cases}$$

- What we want: to derive bound for

$$\Delta m\big(-\alpha^*\nabla f(\boldsymbol{x})\big) := m(\mathbf{0};\boldsymbol{x}) - m(-\alpha^*\nabla f(\boldsymbol{x});\boldsymbol{x}) = \alpha^*\|\nabla f(\boldsymbol{x})\|_2^2 - \frac{{\alpha^*}^2}{2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2.$$

- Consider case 1 $\alpha^* = 0$: we have no update: $\Delta m\big(-\alpha^*\nabla f(\boldsymbol{x})\big) = 0$.

- Note that this case is impossible if we use positive definite $\boldsymbol{B}$

If $\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 > 0$, case 2

$$\boxed{\begin{array}{rcl} m(\boldsymbol{s};\boldsymbol{x}) & := & f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \frac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}} \\ m(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x}) & = & f(\boldsymbol{x}) - \alpha\|\nabla f(\boldsymbol{x})\|_2^2 + \frac{\alpha^2}{2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \end{array}}$$

$$\alpha^* = \operatorname{median}\left(0, \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}, \frac{\delta}{\|\nabla f(\boldsymbol{x})\|}\right) = \begin{cases} 0 & \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \leq 0 \\ \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} & 0 < \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \leq \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|} \\ \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|} & \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} > \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|} \end{cases}$$

- Consider case 2 $\alpha^* = \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}$

$$\begin{aligned} \Delta m\big(-\alpha^*\nabla f(\boldsymbol{x})\big) &= \alpha^*\|\nabla f(\boldsymbol{x})\|_2^2 - \frac{\alpha^{*2}}{2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \\ &\overset{\alpha^*}{=} \frac{\|\nabla f(\boldsymbol{x})\|_2^4}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} - \frac{\|\nabla f(\boldsymbol{x})\|_2^4}{2\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \\ &= \frac{\|\nabla f(\boldsymbol{x})\|_2^4}{2\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \\ &\overset{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \leq \|\nabla f(\boldsymbol{x})\|_2^2\|\boldsymbol{B}\|_2}{\geq} \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{2\|\boldsymbol{B}\|_2}. \end{aligned}$$

Where $\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \leq \|\boldsymbol{B}\nabla f(\boldsymbol{x})\|_2\|\nabla f(\boldsymbol{x})\|_2 \leq \|\boldsymbol{B}\|_2\|\nabla f(\boldsymbol{x})\|_2\|\nabla f(\boldsymbol{x})\|_2 = \|\boldsymbol{B}\|_2\|\nabla f(\boldsymbol{x})\|_2^2$

If $\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 > 0$, case 3

$$\boxed{\begin{aligned} m(\boldsymbol{s};\boldsymbol{x}) &:= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s}\rangle + \tfrac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}} \\ m(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x}) &= f(\boldsymbol{x}) - \alpha\|\nabla f(\boldsymbol{x})\|_2^2 + \tfrac{\alpha^2}{2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \end{aligned}}$$

$$\alpha^* = \operatorname{median}\Big(0, \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}, \frac{\delta}{\|\nabla f(\boldsymbol{x})\|}\Big) = \begin{cases} 0 & \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \le 0 \\ \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} & 0 < \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \le \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|} \\ \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|} & \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} > \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|} \end{cases}$$

- For case 3 $\alpha^* = \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|}$:

$$\Delta m\big(-\alpha^*\nabla f(\boldsymbol{x})\big) = \alpha^*\|\nabla f(\boldsymbol{x})\|_2^2 - \frac{\alpha^{*2}}{2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \overset{\alpha^*}{=} \delta\|\nabla f(\boldsymbol{x})\| - \frac{\delta^2}{2\|\nabla f(\boldsymbol{x})\|_2^2}\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \qquad (*)$$

- Because we are in case 3,

$$\frac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} > \frac{\delta}{\|\nabla f(\boldsymbol{x})\|} \iff \frac{\|\nabla f(\boldsymbol{x})\|}{\delta} > \frac{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}{\|\nabla f(\boldsymbol{x})\|_2^2} \implies -\frac{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2}{\|\nabla f(\boldsymbol{x})\|_2^2} > -\frac{\|\nabla f(\boldsymbol{x})\|}{\delta} \qquad (**)$$

- Put $(**)$ into $(*)$ gives

$$\Delta m\big(-\alpha^*\nabla f(\boldsymbol{x})\big) \ge \frac{\delta}{2}\|\nabla f(\boldsymbol{x})\|_2.$$

Summary: sufficient descent condition of m if $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$

- From the last 4 slides: after solving the TR-subproblem with $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$, if $\alpha^* \neq 0$,

$$\Delta m\big(-\alpha\nabla f(\boldsymbol{x})\big) \;\;\coloneqq\;\; m(\mathbf{0};\boldsymbol{x}) - m(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x}) \;\;\geq\;\; \begin{cases} \delta\|\nabla f(\boldsymbol{x})\|_2 & \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \leq 0 \\ \dfrac{\delta}{2}\|\nabla f(\boldsymbol{x})\|_2 & \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 > 0,\ \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} > \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|_2} \\ \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{2\|\boldsymbol{B}\|} & \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 > 0,\ \dfrac{\|\nabla f(\boldsymbol{x})\|_2^2}{\|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2} \leq \dfrac{\delta}{\|\nabla f(\boldsymbol{x})\|_2} \end{cases}$$

$$= \begin{cases} \delta\|\nabla f(\boldsymbol{x})\|_2 & \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 \leq 0 \\ \dfrac{\|\nabla f(\boldsymbol{x})\|_2}{2}\min\left\{\dfrac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|},\delta\right\} & \|\nabla f(\boldsymbol{x})\|_{\boldsymbol{B}}^2 > 0 \end{cases}$$

- If we use positive definite $\boldsymbol{B}$, the first case is impossible

$$\Delta m\Big(-\alpha\nabla f(\boldsymbol{x})\Big) \;\geq\; \frac{\|\nabla f(\boldsymbol{x})\|_2}{2}\min\left\{\frac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|},\delta\right\}. \qquad (\dagger)$$

- The meaning

$$\underbrace{m(\mathbf{0};\boldsymbol{x})}_{\boldsymbol{x}\text{ not moving}} - \underbrace{m(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x})}_{\substack{\boldsymbol{x}\text{ move along }-\alpha\nabla f(\boldsymbol{x}) \\ -\alpha\nabla f(\boldsymbol{x})\text{ is the steepest descent direction} \\ \text{moving along this direction makes } m \text{ smaller}}} \;\;\geq\;\; \underbrace{\frac{\|\nabla f(\boldsymbol{x})\|_2}{2}\min\left\{\frac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|},\delta\right\}}_{\text{how much is the gap}}.$$

Algorithm 1: Trust-region algorithm

```
Initialize $\boldsymbol{x}_0$                                    % initial starting point
Initialize $\delta_0$                                            % initial trust-region radius
Pick a norm $\|\cdot\|$                                          % trust-region geometry
Pick $0 < \gamma_d < 1 < \gamma_i$, $0 < \eta_s \leq \eta_{vs} < 1$.   % TR parameters
Compute $f(\boldsymbol{x}_0)$

for $k = 1, 2, \dots$ do
    Build $m(\boldsymbol{s}; \boldsymbol{x}_k) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2$

    Find $\boldsymbol{s}$ that satisfies $\|\boldsymbol{s}\| \leq \delta_k$ and $m(\boldsymbol{s}; \boldsymbol{x}_k) \leq m(-\alpha^* \nabla f(\boldsymbol{x}_k); \boldsymbol{x}_k)$

    Let $\rho_k = \dfrac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{s})}{m(\mathbf{0}; \boldsymbol{x}_k) - m(\boldsymbol{s}; \boldsymbol{x}_k)}$

    $\boldsymbol{x}_{k+1} = \begin{cases} \boldsymbol{x}_k + \boldsymbol{s} & \rho_k \geq \eta_{vs} \quad \text{(very successful)} \\ \boldsymbol{x}_k + \boldsymbol{s} & \rho_k \in [\eta_s, \eta_{vs}[ \quad \text{(successful)} \\ \boldsymbol{x}_k & \rho_k < \eta_s \quad \text{(failed)} \end{cases}$    $\delta_{k+1} = \begin{cases} \gamma_i \delta_k & \rho_k \geq \eta_{vs} \quad \text{(very successful)} \\ \delta_k & \rho_k \in [\eta_s, \eta_{vs}[ \quad \text{(successful)} \\ \gamma_d \delta_k & \rho_k < \eta_s \quad \text{(failed)} \end{cases}$
```

Typical value: $\gamma_i = 2$, $\gamma_i = 0.5$.

Compared with gradient descent, TR has a higher cost per-iteration.

Set of iteration counter k

$$
\begin{aligned}
&\text{set of very successful iteration} && \mathcal{V} &:=& \Big\{k \mid \rho_k \geq \eta_{vs}\Big\}\\
&\text{set of successful iteration} && \mathcal{S} &:=& \Big\{k \mid \rho_k \geq \eta_{s}\Big\}\\
&\text{set of failed iteration} && \mathcal{F} &:=& \Big\{k \mid \rho_k < \eta_{v}\Big\}\\
&\text{set of iteration} && \mathcal{K} &:=& \mathbb{N} = \{1, 2, 3, \dots\}
\end{aligned}
$$

- $\mathcal{K}$ is an infinite set
- $\mathcal{V} \subseteq \mathcal{S}$
- $\mathcal{S} \cap \mathcal{F} = \varnothing$, $\mathcal{K} = \mathcal{S} \cup \mathcal{F}$, $\mathcal{F} = \mathcal{K} \setminus \mathcal{S}$ and $|\mathcal{F}| = |\mathcal{K}| - |\mathcal{S}|$
- Fact: if there are finitely many successful & very successful iteration, then there exists a sufficiently large k_0 such that all iterations k after k_0 are failed:
  - finitely many successful and very successful iteration $\implies |\mathcal{S}| \leq \infty$
  - $\mathcal{F} = \mathcal{K} \setminus \mathcal{S}$
  - so there exists k_0 s.t. $k > k_0$ are all in $\mathcal{F}$
  - $|\mathcal{F}| = |\mathcal{K}| - |\mathcal{S}| = |\mathbb{N}| - |\mathcal{S}| = \aleph_0 - |\mathcal{S}| = \infty - |\mathcal{S}| = \infty$ details of $\aleph_0$

  This fact is useful later for proving convergence.

Table of Contents

Quadratic model $m(\boldsymbol{s};\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \frac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2$

Sufficient descent: $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$ then $\Delta m(\boldsymbol{s}) \ \geq \ \frac{\|\nabla f(\boldsymbol{x})\|_2}{2} \min\left\{ \frac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|}, \delta \right\}$

Theory of TR convergence

1. $f - m$ gap: $\left| f(\boldsymbol{x}+\boldsymbol{s}) - m(\boldsymbol{s};\boldsymbol{x}) \right| \ \leq \ \frac{\kappa_H + \kappa_B}{2}\delta^2$
2. Progress (small radius $\Longrightarrow$ success): $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}, \delta_k \leq \frac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H + \kappa_B} \min\left(1, 1 - \eta_{vs}\right) \implies k \in \mathcal{V}, \delta_{k+1} \geq \delta_k$
3. TR radius will not shrink to 0 at non-sol.
4. Possible finite termination
5. Global convergence of some subsequence

Assumptions for TR convergence

- To derive some theories of TR, we assume

  1. $f \in \mathcal{C}^2$.
  2. $\|\boldsymbol{H}(\boldsymbol{x})\|_2 \leq \kappa_H, \ \forall \boldsymbol{x}$.
  3. $\|\boldsymbol{B}(\boldsymbol{x})\|_2 \leq \kappa_B, \ \forall \boldsymbol{x}$.
  4. $\kappa_H \geq 1$ and $\kappa_B \geq 0$.

- Meaning

  1. f is twice differentiable (so Hessian exsits and we can have assumption 2).
  2. For the Hessian of f, its matrix 2-norm is globally bounded above.
     a strong assumption, can be relaxed by the sequence $\{f(\boldsymbol{x}_k)\}_{k\in\mathbb{N}}$ is monotonically decreasing
  3. For $\boldsymbol{B}$ in the model m, its matrix 2-norm of is globally bounded above.
  4. Condition on κ_H (larger than 1) and κ_B (larger than 0).

  2 & 4 also mean $\|\boldsymbol{H}(\boldsymbol{x})\|_2$ is bounded above by at-least-1

Summary of TR convergence results under assumptions

1. $f \in \mathcal{C}^2$.
2. $\|\boldsymbol{H}(\boldsymbol{x})\|_2 \leq \kappa_H, \ \forall \boldsymbol{x}$.
3. $\|\boldsymbol{B}(\boldsymbol{x})\|_2 \leq \kappa_B, \ \forall \boldsymbol{x}$.
4. $\kappa_H \geq 1$ and $\kappa_B \geq 0$.

1. $\left| f(\boldsymbol{x}+\boldsymbol{s}) - m(\boldsymbol{s};\boldsymbol{x}) \right| \leq \frac{\kappa_H + \kappa_B}{2}\delta^2.$ (gap between f and m)
2. $\begin{array}{l} \nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0} \\ \delta_k \leq \frac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H+\kappa_B} \min\left(1, 1-\eta_{vs}\right) \end{array} \implies$ update is $\mathcal{V}$ & $\delta_{k+1} \geq \delta_k$. (progress at non-sol / small δ guarantee successful)
3. If there exist ϵ and $k_0 \in \mathbb{N}$ s.t. $\|\nabla f(\boldsymbol{x}_k)\| \geq \epsilon \geq 0 \ \forall k \geq k_0$,
   then $\delta_k \geq \delta_{\min} := \frac{\|\epsilon \gamma_d\|_2}{\kappa_H+\kappa_B} \min\left(1, 1-\eta_{vs}\right) \forall k \geq k_1$ for some $k_1 \in \mathbb{N}$. (TR radius will not shrink to 0)
4. If there are finitely many very successful & successful iterations,
   then $\boldsymbol{x}_k = \boldsymbol{x}^*$ for sufficiently large k where $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$. (possible finite termination)
5. Either $\begin{cases} \exists k < \infty \,\text{s.t.}\, \nabla f(\boldsymbol{x}_k) = \boldsymbol{0} \\ \lim\limits_{k\to\infty} f(\boldsymbol{x}_k) = -\infty \\ \liminf\limits_{k\to\infty} \|\nabla f(\boldsymbol{x}_k)\| = 0 \end{cases}$ (Global convergence)

Gap between objective function f and model $m = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \rangle + \dfrac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2$

- IF
  1. $f \in \mathcal{C}^2$.
  2. $\|\boldsymbol{H}(\boldsymbol{x})\|_2 \leq \kappa_H, \ \forall \boldsymbol{x}$.
  3. $\|\boldsymbol{B}(\boldsymbol{x})\|_2 \leq \kappa_B, \ \forall \boldsymbol{x}$.
  4. $\kappa_H \geq 1$ and $\kappa_B \geq 0$.

  THEN

  $$\Big| f(\boldsymbol{x} + \boldsymbol{s}) - m(\boldsymbol{s}; \boldsymbol{x}) \Big| \ \leq \ \frac{\kappa_H + \kappa_B}{2} \delta^2. \quad \text{(Gap)}$$

- **Proof**. $f \in \mathcal{C}^2$, apply mean value theorem on f at $\boldsymbol{s}$ for some $\boldsymbol{\xi} \in [\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{s}]$ gives

$$\begin{aligned}
f(\boldsymbol{x}+\boldsymbol{s}) &= f(\boldsymbol{x}) + \big\langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \big\rangle + \frac{1}{2} \big\langle \boldsymbol{H}(\xi)\boldsymbol{s}, \boldsymbol{s} \big\rangle && \boxed{\text{assumption 1 (f twice differentiable)}} \text{ \& mean value theorem} \\
\big| f(\boldsymbol{x}+\boldsymbol{s}) - m(\boldsymbol{s};\boldsymbol{x}) \big| &= \tfrac{1}{2} \Big| \big\langle \boldsymbol{H}(\boldsymbol{\xi})\boldsymbol{s}, \boldsymbol{s} \big\rangle - \big\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s} \big\rangle \Big| \\
&\leq \tfrac{1}{2} \Big| \big\langle \boldsymbol{H}(\boldsymbol{\xi})\boldsymbol{s}, \boldsymbol{s} \big\rangle \Big| + \tfrac{1}{2} \Big| \big\langle \boldsymbol{B}\boldsymbol{s}, \boldsymbol{s} \big\rangle \Big| && \text{triangle inequality} \\
&= \tfrac{1}{2} \|\boldsymbol{H}(\boldsymbol{\xi})\|_2 \|\boldsymbol{s}\|_2^2 + \tfrac{1}{2} \|\boldsymbol{B}\|_2 \|\boldsymbol{s}\|_2^2 && \text{Cauchy-Schwartz inequality} \\
&= \tfrac{1}{2} \Big(\|\boldsymbol{H}(\boldsymbol{\xi})\|_2 + \|\boldsymbol{B}\|_2 \Big) \|\boldsymbol{s}\|_2^2 \\
&\leq \tfrac{1}{2} (\kappa_H + \kappa_B) \delta^2 && \boxed{\text{assumption 2 3}} \text{ \& } \|\boldsymbol{s}\| \leq \delta
\end{aligned}$$

* You don't need assumption 4 here.

Progress at non-sol / small TR radius guarantee successful ... 1/2

- If
  1. $f \in \mathcal{C}^2$.
  2. $\|\boldsymbol{H}(\boldsymbol{x})\|_2 \le \kappa_H, \ \forall \boldsymbol{x}$
  3. $\|\boldsymbol{B}(\boldsymbol{x})\|_2 \le \kappa_B, \ \forall \boldsymbol{x}$
  4. $\kappa_H \ge 1$, $\kappa_B \ge 0$

  &
  - $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$
  - $\delta_k \le \dfrac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H + \kappa_B} \min\left(1, 1 - \eta_{vs}\right)$

  Then
  - the update is very successful
  - $\delta_{k+1} \ge \delta_k$

- **Proof**. 🟥 implies $\delta_k \le \dfrac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H + \kappa_B}$ and $\delta_k \le \dfrac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H + \kappa_B}\left(1 - \eta_{vs}\right)$

$$\|\boldsymbol{B}(\boldsymbol{x}_k)\| \overset{\text{🟩,🟧}}{\le} \kappa_B + \kappa_H \implies \frac{1}{\kappa_B + \kappa_H} \le \frac{1}{\|\boldsymbol{B}(\boldsymbol{x}_k)\|} \implies \frac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_B + \kappa_H} \le \frac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\|\boldsymbol{B}(\boldsymbol{x}_k)\|}$$

Recall

$$\Delta m\Big(- \alpha \nabla f(\boldsymbol{x})\Big) \overset{(\dagger)}{\ge} \frac{\|\nabla f(\boldsymbol{x})\|_2}{2} \min\left\{ \frac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|}, \delta \right\} \overset{\text{🟪,🟦}}{=} \frac{\|\nabla f(\boldsymbol{x})\|_2}{2} \delta \ge 0.$$

(Because by 🟪, 🟦, we have $\delta \le \dfrac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|}$ so the $\min$ gives δ)

- Now we have $\Delta m\big(- \alpha \nabla f(\boldsymbol{x})\big) \ge \dfrac{\|\nabla f(\boldsymbol{x})\|_2}{2} \delta \ge 0$.

Progress at non-sol / small TR radius guarantee successful ... 2/2

$$\Delta m(-\alpha\nabla f(\boldsymbol{x})) \geq \frac{\|\nabla f(\boldsymbol{x})\|_2}{2}\delta \geq 0$$

- Now consider $|\rho - 1|$ with $\rho_k = \dfrac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{s})}{m(\boldsymbol{0};\boldsymbol{x}_k) - m(\boldsymbol{s};\boldsymbol{x}_k)}$, where $\boldsymbol{s} = -\alpha\nabla f(\boldsymbol{x})$, then

$$\begin{aligned}
\big|\rho_k - 1\big| &= \left|\frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{s})}{m(\boldsymbol{0};\boldsymbol{x}_k) - m(\boldsymbol{s};\boldsymbol{x}_k)} - \frac{m(\boldsymbol{0};\boldsymbol{x}_k) - m(\boldsymbol{s};\boldsymbol{x}_k)}{m(\boldsymbol{0};\boldsymbol{x}_k) - m(\boldsymbol{s};\boldsymbol{x}_k)}\right| && \\
&= \left|\frac{m(\boldsymbol{s};\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{s})}{m(\boldsymbol{0};\boldsymbol{x}_k) - m(\boldsymbol{s};\boldsymbol{x}_k)}\right| && m(\boldsymbol{0};\boldsymbol{x}_k) = f(\boldsymbol{x}_k) \\
&= \frac{1}{|\Delta m(\boldsymbol{s})|}\Big|f(\boldsymbol{x}_k + \boldsymbol{s}) - m(\boldsymbol{s};\boldsymbol{x}_k)\Big| && m(\boldsymbol{0};\boldsymbol{x}_k) - m(\boldsymbol{s};\boldsymbol{x}_k) = \Delta m(\boldsymbol{s}) \\
&\leq \frac{2}{\|\nabla f(\boldsymbol{x}_k)\|_2\delta}\Big|f(\boldsymbol{x}_k - \alpha\nabla f(\boldsymbol{x})) - m(-\alpha\nabla f(\boldsymbol{x});\boldsymbol{x}_k)\Big| && \textbf{by } 🟦 \text{ and } \boldsymbol{s} = -\alpha\nabla f(\boldsymbol{x}) \\
&\leq \frac{2}{\|\nabla f(\boldsymbol{x}_k)\|_2\delta}\frac{\kappa_H + \kappa_B}{2}\delta^2 && \text{By (Gap), see 2 slides before} \\
&= \frac{\kappa_H + \kappa_B}{\|\nabla f(\boldsymbol{x}_k)\|_2}\delta && \\
&\overset{🟨}{\leq} 1 - \eta_{vs}. &&
\end{aligned}$$

- Now we have $|\rho_k - 1| \leq 1 - \eta_{vs}$, which gives

$$\underbrace{-(1-\eta_{vs}) \;\leq\; \rho - 1}_{\eta_{vs} \leq \rho} \;\leq\; 1 - \eta_{vs} \qquad \Longrightarrow \qquad \rho \geq \eta_{vs} \text{ meaning the iteration is very successful, i.e., } k \in \mathcal{V} \subset \mathcal{S}$$

For very successful iteration, $\delta_{k+1} = \gamma_i\delta_k$. Since $\gamma_i > 1$, thus $\delta_{k+1} > \delta_k$.

TR radius will not shrink to 0 at non-sol.

- IF
  1. $f \in \mathcal{C}^2$.
  2. $\|\boldsymbol{H}(\boldsymbol{x})\|_2 \leq \kappa_H, \ \forall \boldsymbol{x}$
  3. $\|\boldsymbol{B}(\boldsymbol{x})\|_2 \leq \kappa_B, \ \forall \boldsymbol{x}$
  4. $\kappa_H \geq 1$ and $\kappa_B \geq 0$

  & there exists constant ϵ and $k_0 \in \mathbb{N}$ such that $\|\nabla f(\boldsymbol{x}_k)\| \geq \epsilon \geq 0$ for all $k \geq k_0$.

  THEN $\delta_k \geq \delta_{\min} := \dfrac{\epsilon\gamma_d}{\kappa_H + \kappa_B} \min\left(1, 1-\eta_{vs}\right) > 0$ for all $k \geq k_1$ for some $k_1 \in \mathbb{N}$.

- **Proof**. If there is some $k' \geq k_0$ such that $\delta_{k'} \geq \frac{\epsilon \min\left(1, 1-\eta_{vs}\right)}{\kappa_H+\kappa_B}$, then by definition of TR algorithm, in the worse case we have $\delta_k \geq \delta_{\min} := \frac{\epsilon\gamma_d}{\kappa_H+\kappa_B} \min\left(1, 1-\eta_{vs}\right)$ (in other cases we have larger δ_k).

  Now for contradiction, suppose otherwise that $k \geq k'$ is the first iteration such that

  $$\delta_k \geq \delta_{\min} > \delta_{k+1} = \gamma_d \delta_k. \tag{$*$}$$

  Thus
  $$\delta_k = \frac{\delta_{k+1}}{\gamma_d} \leq \frac{\delta_{\min}}{\gamma_d} = \frac{\epsilon}{\kappa_H + \kappa_B} \min\left(1, 1-\eta_{vs}\right) \leq \frac{\|\nabla f(\boldsymbol{x}_k)\|}{\kappa_H + \kappa_B} \min\left(1, 1-\eta_{vs}\right).$$

  Then by the lemma of progress at non-sol., $\delta_{k+1} \geq \delta_k$, which contradicts with $(*)$.

  Now we have to show that $\exists k' \geq k_0$ such that $\delta_{k'} \geq \frac{\epsilon}{\kappa_H+\kappa_B} \min\left(1, 1-\eta_{vs}\right)$.
  By the lemma of progress at non-sol., whenever $\delta_{k'} < \frac{\epsilon}{\kappa_H+\kappa_B} \min\left(1, 1-\eta_{vs}\right)$, we have a very successful iteration, and therefore we strictly increase the radius by the factor $\gamma_i > 1$, i.e., $\delta_{k+1} = \gamma_i \delta_k$.

Possible finite termination

- IF

  1. $f \in \mathcal{C}^2$.
  2. $\|\boldsymbol{H}(\boldsymbol{x})\|_2 \leq \kappa_H, \ \forall \boldsymbol{x}$
  3. $\|\boldsymbol{B}(\boldsymbol{x})\|_2 \leq \kappa_B, \ \forall \boldsymbol{x}$
  4. $\kappa_H \geq 1$ and $\kappa_B \geq 0$

  & there are finitely many very successful & successful iterations.

  THEN $\boldsymbol{x}_k = \boldsymbol{x}^*$ for all sufficiently large k and $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

- **Proof** By assumption, it follows that there exists some $\boldsymbol{x}^*$ such that $\boldsymbol{x}_{k_0+j} = \boldsymbol{x}_{k_0+1} = \boldsymbol{x}^*$ for all $j \geq 1$, where k_0 is the index of the last successful iterate (see page 23).

  Hence, all the remaining infinitely many unsuccessful iterations will eventually shrink the TR radius to zero, i.e.,

$$\lim_{k\to\infty} \delta_k = 0. \tag{$*$}$$

  For the purpose of contradiction, assume $\nabla f(\boldsymbol{x}_{k_0+1}) \neq \mathbf{0}$, let $\epsilon = \|\nabla f(\boldsymbol{x}_{k_0+1})\| > 0$. By the lemma in the previous page, we have

$$\delta_k \geq \delta_{\min} := \frac{\epsilon \gamma_d}{\kappa_H + \kappa_B} \min\left(1, 1 - \eta_{vs}\right) > 0,$$

  contradicting $(*)$. Therefore the assumption is false and we have $\nabla f(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}_{k_0+1}) = \mathbf{0}$.

Global convergence[1] of some subsequence ... 1/3

- If
  1. $f \in \mathcal{C}^2$.
  2. $\|\boldsymbol{H}(\boldsymbol{x})\|_2 \leq \kappa_H, \ \forall \boldsymbol{x}$
  3. $\|\boldsymbol{B}(\boldsymbol{x})\|_2 \leq \kappa_B, \ \forall \boldsymbol{x}$
  4. $\kappa_H \geq 1$ and $\kappa_B \geq 0$

  then either
  1. finite termination: $\exists k < \infty \,\text{s.t.}\, \nabla f(\boldsymbol{x}_k) = \mathbf{0}$.
  2. unbounded objective function: $\min\limits_{k \to \infty} f(\boldsymbol{x}_k) = -\infty$.
  3. convergence of a subsequence of the gradients: $\liminf\limits_{k \to \infty} \|\nabla f(\boldsymbol{x}_k)\| = 0$.

- **Idea of the proof**. We show that under the assumption we will get exactly one of the result.
  - To do so we introduce an object: let $\mathcal{S}$ be the index set of successful and very successful iterations.
  - By definition of the TR (Algorithm 1 in page 22), if at an iteration $k \in \mathcal{S}$, we have

$$\rho_k \geq \eta_s. \tag{$*$}$$

  - Recall the definition of TR (Algorithm 1) on ρ_k, we have

$$\rho_k \overset{\text{definition}}{=} \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - \boldsymbol{s}_k)}{m_k(\mathbf{0}) - m_k(\boldsymbol{s}_k)} \iff f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - \boldsymbol{s}_k) = \rho_k \underbrace{\left(m_k(\mathbf{0}) - m_k(\boldsymbol{s}_k)\right)}_{=:\Delta m_k(\boldsymbol{s}_k)} \overset{(*)}{\geq} \eta_s \Delta m_k(\boldsymbol{s}_k). \tag{$**$}$$

  $(**)$ is the starting point of the proof.

- **Proof**. Let $\mathcal{S}$ be the index set of successful and very successful iterations.
  - Lemma (possible finite termination, previous slide) implies result 1 is true if $|\mathcal{S}| < \infty$. case 1 done
  - Now consider the remaining case $|\mathcal{S}| = \infty$. If f is unbounded below then we have result 2. case 2 done
  - So now we show that if $|\mathcal{S}| = \infty$ and f is bounded below then we have case 3.

---

[1] Here "global convergence" means convergence to a stationary point regardless of starting point

Global convergence of some subsequence[2] ... 2/3

- Goal: show that if $|\mathcal{S}| = \infty$ and f is bounded below then we have case 3.

- For the purpose of contradiction, assume there exists $\epsilon > 0$ and $k_0 \in \mathbb{N}$ such that

$$\|\nabla f(\boldsymbol{x}_k)\| \geq \epsilon > 0 \quad \forall k \geq k_0. \qquad (\pounds)$$

- From $(**)$, we have the following for all $k \in \mathcal{S}$ such that $k \geq k_0$

$$\begin{aligned}
f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{s}_k) &\geq \eta_s \Delta m_k(\boldsymbol{s}_k) && \text{by } (**) \\
&\geq \eta_s \frac{1}{2} \|\nabla f(\boldsymbol{x}_k)\| \min\left\{ \frac{\|\nabla f(\boldsymbol{x}_k)\|}{\|\boldsymbol{B}_k\|}, \delta_k \right\} && \text{by } \boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x}_k) \text{ and sufficient descent condition of } m \\
&\geq \frac{\eta_s}{2} \epsilon \min\left\{ \frac{\epsilon}{\|\boldsymbol{B}_k\|}, \delta_k \right\} && \text{by } (\pounds) \\
&\geq \frac{\eta_s \epsilon}{2} \min\left\{ \frac{\epsilon}{\kappa_{\boldsymbol{B}}}, \delta_k \right\} && \|\boldsymbol{B}_k\| \leq \kappa_{\boldsymbol{B}} \\
&\geq \underbrace{\frac{\eta_s \epsilon}{2} \min\left\{ \frac{\epsilon}{\kappa_{\boldsymbol{B}}}, \delta_{\min} \right\}}_{=:\delta_\epsilon} && \delta_k \geq \delta_{\min} \text{(TR radius will not shrink to 0)} \\
&> 0 && \epsilon > 0, \kappa_{\boldsymbol{B}} \geq 1, \eta_s \geq 1, \delta_{\min} > 0
\end{aligned}$$

Now we have for all $k \in \mathcal{S}$ such that $k \geq k_0$

$$f_k - f_{k+1} := f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{s}_k) \geq \delta_\epsilon > 0. \qquad (\diamond)$$

---

[2] Here subsequence is used because we consider sequence $\{\boldsymbol{x}_k\}_{k \geq k_0}$

Global convergence of some subsequence ... 3/3

$$f_k - f_{k+1} := f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{s}_k) \geq \delta_\epsilon > 0. \qquad (\diamond)$$

- Now we perform telescoping sum: pick $j \geq 1$ and then summing over all $k \leq j$

$$\sum_{k=0}^{j} (f_k - f_{k+1}) \overset{\text{telescope sum}}{=} f_0 - f_{j+1}.$$

- Focus on $k \in \mathcal{S}$ such that $k \leq j$ gives

$$f_0 - f_{j+1} \overset{\text{telescope sum}}{=} \sum_{k=0}^{j} (f_k - f_{k+1}) \overset{(!)}{\geq} \sum_{k=0,k\in\mathcal{S}}^{j} (f_k - f_{k+1}) \overset{(\diamond)}{\geq} \sum_{k=0,k\in\mathcal{S}}^{j} \delta_\epsilon > 0. \qquad (\diamond\diamond)$$

where $\overset{(!)}{\geq}$ is by definition: if $k \notin \mathcal{S}$ then that iteration is unsuccessful, by definition of TR algorithm $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ so $f_k = f_{k+1}$. Since the set of $[0, 1, \ldots, k, \ldots, j]$ is larger than $[0, 1, \ldots, k, \ldots, j] \cap \{k \in \mathcal{S}\}$ so we have $\geq$ sign.

- Now take limit $j \to \infty$ on $(\diamond\diamond)$

$$\lim_{j\to\infty} (f_0 - f_{j+1}) \overset{(\diamond\diamond)}{\geq} \lim_{j\to\infty} \sum_{k=0,k\in\mathcal{S}}^{j} \delta_\epsilon = \sum_{k=0,k\in\mathcal{S}}^{\infty} \delta_\epsilon \overset{\delta_\epsilon \geq 0}{=} +\infty \quad \Longrightarrow \quad f_0 - f_\infty \geq +\infty$$

$\Longrightarrow$ f is unbounded below. This contradicts to the assumption therefore the assumption $(\pounds)$ is false, which means there exists a subsequence of the gradients that converges to zero, i.e., $\liminf_{k\to\infty} \|\nabla f(\boldsymbol{x}_k)\| = 0$.

Last page - summary

Quadratic model $m(\boldsymbol{s};\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{x}), \boldsymbol{s} \right\rangle + \frac{1}{2}\|\boldsymbol{s}\|_{\boldsymbol{B}}^2$

Sufficient descent: $\boldsymbol{s} = -\alpha \nabla f(\boldsymbol{x})$ then $\Delta m(\boldsymbol{s}) \;\geq\; \frac{\|\nabla f(\boldsymbol{x})\|_2}{2} \min \left\{ \frac{\|\nabla f(\boldsymbol{x})\|_2}{\|\boldsymbol{B}\|}, \delta \right\}$

Theory of TR convergence

1. $f - m$ gap: $\left| f(\boldsymbol{x}+\boldsymbol{s}) - m(\boldsymbol{s};\boldsymbol{x}) \right| \;\leq\; \frac{\kappa_H + \kappa_B}{2} \delta^2$
2. Progress (small radius $\Longrightarrow$ success): $\nabla f(\boldsymbol{x}_k) \neq \boldsymbol{0}, \delta_k \leq \frac{\|\nabla f(\boldsymbol{x}_k)\|_2}{\kappa_H + \kappa_B} \min\left(1, 1 - \eta_{vs}\right) \implies k \in \mathcal{V}, \delta_{k+1} \geq \delta_k$
3. TR radius will not shrink to 0 at non-sol.
4. Possible finite termination
5. Global convergence of some subsequence

End of document