Part 3: Trust-region methods for unconstrained optimization

Nick Gould (RAL)

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

MSc course on nonlinear optimization

UNCONSTRAINED MINIMIZATION

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

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

- assume that $f \in C^1$ (sometimes C^2) and Lipschitz
- often in practice this assumption violated, but not necessary

LINESEARCH VS TRUST-REGION METHODS

- **Linesearch 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 decrease in model inherited by $f(x_k + s_k)$
  - otherwise set $x_{k+1} = x_k$, "refine" model

TRUST-REGION MODEL PROBLEM

Model $f(x_k + s)$ by:

- linear model

$$m_k^L(s) = f_k + s^T g_k$$

- quadratic model — symmetric B_k

$$m_k^Q(s) = f_k + s^T g_k + \tfrac{1}{2} s^T B_k s$$

Major difficulties:

- models may not resemble $f(x_k + s)$ if s is large
- models may be unbounded from below
  - linear model - always unless $g_k = 0$
  - quadratic model - always if B_k is indefinite, possibly if B_k is only positive semi-definite

THE TRUST REGION

Prevent model $m_k(s)$ from unboundedness by imposing a **trust-region** constraint

$$\|s\| \leq \Delta_k$$

for some "suitable" scalar **radius** $\Delta_k > 0$

$\Longrightarrow$ **trust-region subproblem**

$$\text{approx } \underset{s \in \mathbb{R}^n}{\text{minimize}} \; m_k(s) \;\; \text{subject to} \;\; \|s\| \leq \Delta_k$$

- in theory does not depend on norm $\|\cdot\|$
- in practice it might!

OUR MODEL

For simplicity, concentrate on the second-order (Newton-like) model

$$m_k(s) = m_k^Q(s) = f_k + s^T g_k + \tfrac{1}{2} s^T B_k s$$

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

$$\kappa_s \|\cdot\| \leq \|\cdot\|_2 \leq \kappa_l \|\cdot\|$$

for some $\kappa_l \geq \kappa_s > 0$

Note:

- $B_k = H_k$ is allowed
- important norms in $\mathbb{R}^n$
  - $\|\cdot\|_2 \leq \|\cdot\|_2 \leq \|\cdot\|_2$ (!!!)
  - $n^{-\frac{1}{2}} \|\cdot\|_1 \leq \|\cdot\|_2 \leq \|\cdot\|_1$
  - $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq n \|\cdot\|_\infty$

BASIC TRUST-REGION METHOD

Given $k = 0$, $\Delta_0 > 0$ and x_0, until "convergence" do:

Build the second-order model $m(s)$ of $f(x_k + s)$.

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

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

If $\rho_k \geq \eta_v$ [**very successful**] $\boxed{0 < \eta_v < 1}$

set $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \gamma_i \Delta_k$ $\boxed{\gamma_i \geq 1}$

Otherwise if $\rho_k \geq \eta_s$ then [**successful**] $\boxed{0 < \eta_s \leq \eta_v < 1}$

set $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \Delta_k$

Otherwise [**unsuccessful**]

set $x_{k+1} = x_k$ and $\Delta_{k+1} = \gamma_d \Delta_k$ $\boxed{0 < \gamma_d < 1}$

Increase k by 1

"SOLVE" THE TRUST REGION SUBPROBLEM?

At the very least

- aim to achieve as much reduction in the model as would an iteration of steepest descent
- **Cauchy point**: $s_k^{\text{C}} = -\alpha_k^{\text{C}} g_k$ where

$$\begin{aligned} \alpha_k^{\text{C}} &= \arg\min_{\alpha > 0} m_k(-\alpha g_k) \;\text{ subject to }\; \alpha\|g_k\| \leq \Delta_k \\ &= \arg\min_{0 < \alpha \leq \Delta_k / \|g_k\|} m_k(-\alpha g_k) \end{aligned}$$

  - minimize quadratic on line segment $\Longrightarrow$ very easy!
- require that

$$m_k(s_k) \leq m_k(s_k^{\text{C}}) \;\text{ and }\; \|s_k\| \leq \Delta_k$$

- in practice, hope to do far better than this

ACHIEVABLE MODEL DECREASE

Theorem 3.1. If $m_k(s)$ is the second-order model and s_k^{C} is its Cauchy point within the trust-region $\|s\| \leq \Delta_k$,

$$f_k - m_k(s_k^{\mathrm{C}}) \geq \tfrac{1}{2}\|g_k\|_2 \min\left[\frac{\|g_k\|_2}{1+\|B_k\|_2}, \kappa_s\Delta_k\right].$$

Corollary 3.2. If $m_k(s)$ is the second-order model, and s_k is an improvement on the Cauchy point within the trust-region $\|s\| \leq \Delta_k$,

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

DIFFERENCE BETWEEN MODEL AND FUNCTION

Lemma 3.3. Suppose that $f \in C^2$, and that the true and model Hessians satisfy the bounds $\|H_k\|_2 \leq \kappa_h$ and $\|B_k\|_2 \leq \kappa_b$ for all k and some $\kappa_h \geq 1$ and $\kappa_b \geq 0$. Then

$$|f(x_k + s_k) - m_k(s_k)| \leq \kappa_d \Delta_k^2,$$

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

ULTIMATE PROGRESS AT NON-OPTIMAL POINTS

Lemma 3.4. Suppose that $f \in C^2$, that the true and model Hessians satisfy the bounds $\|H_k\|_2 \leq \kappa_h$ and $\|B_k\|_2 \leq \kappa_b$ for all k and some $\kappa_h \geq 1$ and $\kappa_b \geq 0$, and that $\kappa_d = \frac{1}{2}\kappa_l^2(\kappa_h + \kappa_b)$. Suppose furthermore that $g_k \neq 0$ and that

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

Then iteration k is very successful and

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

RADIUS WON'T SHRINK TO ZERO AT NON-OPTIMAL POINTS

Lemma 3.5. Suppose that $f \in C^2$, that the true and model Hessians satisfy the bounds $\|H_k\|_2 \leq \kappa_h$ and $\|B_k\|_2 \leq \kappa_b$ for all k and some $\kappa_h \geq 1$ and $\kappa_b \geq 0$, and that $\kappa_d = \frac{1}{2}\kappa_l^2(\kappa_h + \kappa_b)$. Suppose furthermore that there exists a constant $\epsilon > 0$ such that $\|g_k\|_2 \geq \epsilon$ for all k. Then

$$\Delta_k \geq \kappa_\epsilon \stackrel{\text{def}}{=} \epsilon\gamma_d \min\left(\frac{1}{\kappa_s(\kappa_h + \kappa_b)}, \frac{\kappa_s(1 - \eta_v)}{2\kappa_d}\right)$$

for all k.

POSSIBLE FINITE TERMINATION

Lemma 3.6. Suppose that $f \in C^2$, and that both the true and model Hessians remain bounded for all k. Suppose furthermore that there are only finitely many successful iterations. Then $x_k = x_*$ for all sufficiently large k and $g(x_*) = 0$.

GLOBAL CONVERGENCE OF ONE SEQUENCE

Theorem 3.7. Suppose that $f \in C^2$, and that both the true and model Hessians remain bounded for all k. Then either

$$g_l = 0 \text{ for some } l \geq 0$$

or

$$\lim_{k\to\infty} f_k = -\infty$$

or

$$\liminf_{k\to\infty} \|g_k\| = 0.$$

Theorem 3.8. Suppose that $f \in C^2$, and that both the true and model Hessians remain bounded for all k. Then either

$$g_l = 0 \text{ for some } l \geq 0$$

or

$$\lim_{k \to \infty} f_k = -\infty$$

or

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

II: SOLVING THE TRUST-REGION SUBPROBLEM

$$\text{(approximately) } \underset{s \in \mathbb{R}^n}{\text{minimize}} \; q(s) \equiv s^T g + \tfrac{1}{2} s^T B s \text{ subject to } \|s\| \leq \Delta$$

AIM: find s_* so that

$$q(s_*) \leq q(s^{\mathrm{C}}) \text{ and } \|s_*\| \leq \Delta$$

Might solve

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

THE ℓ_2-NORM TRUST-REGION SUBPROBLEM

$$\underset{s \in \mathbb{R}^n}{\text{minimize}} \;\; q(s) \equiv s^T g + \tfrac{1}{2} s^T B s \;\; \text{subject to} \;\; \|s\|_2 \le \Delta$$

Solution characterisation result:

Theorem 3.9. Any *global* minimizer s_* of $q(s)$ subject to $\|s\|_2 \le \Delta$ satisfies the equation

$$(B + \lambda_* I) s_* = -g,$$

where $B + \lambda_* I$ is positive semi-definite, $\lambda_* \ge 0$ and $\lambda_*(\|s_*\|_2 - \Delta) = 0$. If $B + \lambda_* I$ is positive definite, s_* is unique.

ALGORITHMS FOR THE ℓ_2-NORM SUBPROBLEM

Two cases:

- B positive-semi definite and $Bs = -g$ satisfies $\|s\|_2 \le \Delta \Longrightarrow s_* = s$
- B indefinite or $Bs = -g$ satisfies $\|s\|_2 > \Delta$
  In this case
  - $(B + \lambda_* I) s_* = -g$ and $s_*^T s_* = \Delta^2$
  - nonlinear (quadratic) system in s and λ
  - concentrate on this

EQUALITY CONSTRAINED ℓ_2-NORM SUBPROBLEM

Suppose B has spectral decomposition

$$B = U^T \Lambda U$$

- U eigenvectors
- Λ diagonal eigenvalues: $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$

Require $B + \lambda I$ positive semi-definite $\Longrightarrow \lambda \ge -\lambda_1$

Define

$$s(\lambda) = -(B + \lambda I)^{-1} g$$

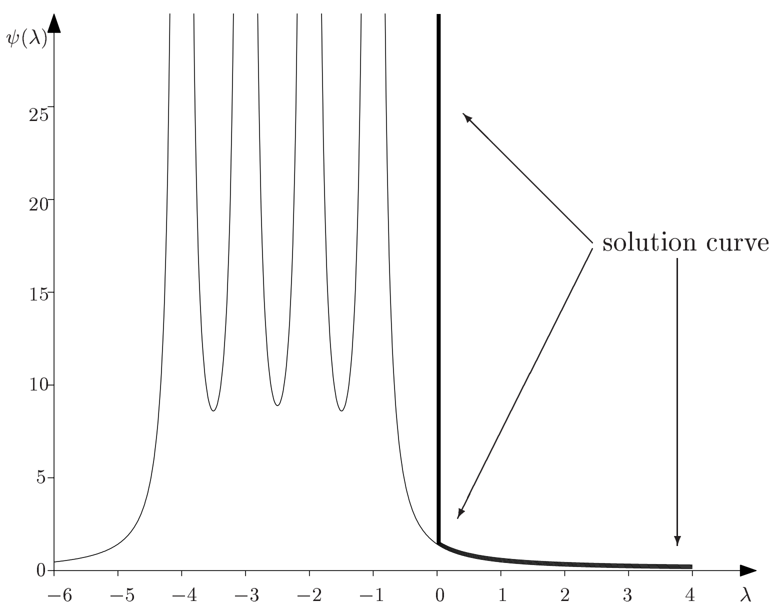
Require

$$\psi(\lambda) \stackrel{\text{def}}{=} \|s(\lambda)\|_2^2 = \Delta^2$$

Note $\qquad (\gamma_i = e_i^T U g)$

$$\psi(\lambda) = \|U^T(\Lambda + \lambda I)^{-1} U g\|_2^2 = \sum_{i=1}^{n} \frac{\gamma_i^2}{(\lambda_i + \lambda)^2}$$

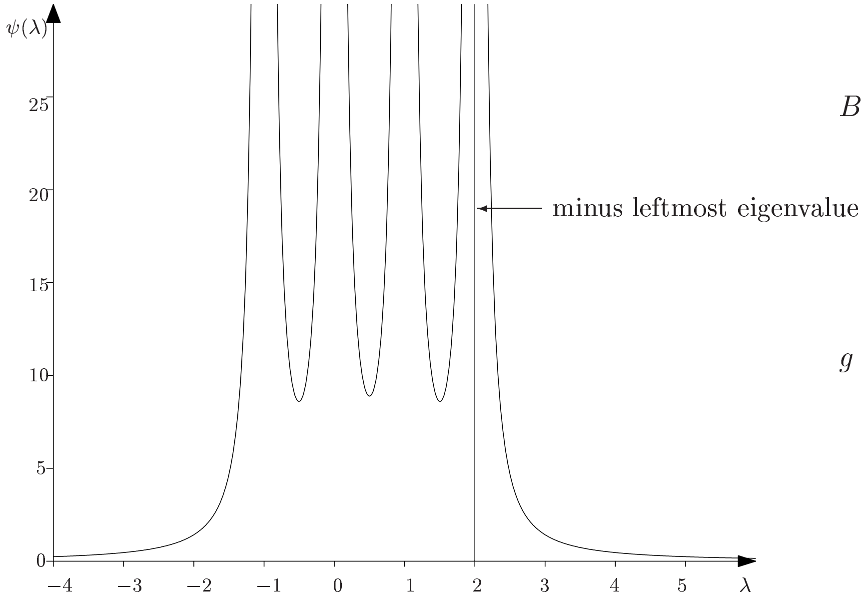
CONVEX EXAMPLE



$$B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

$$g = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

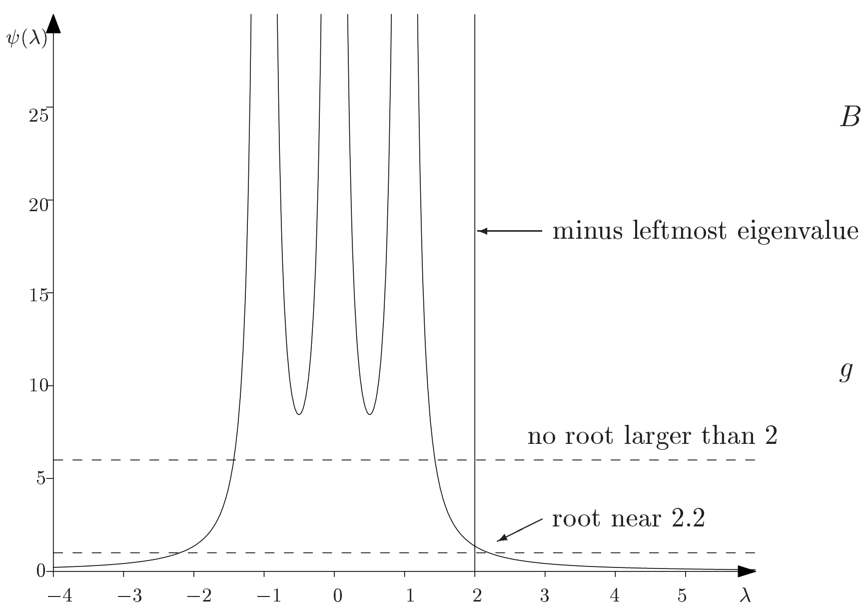
NONCONVEX EXAMPLE



$$B = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$g = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

THE “HARD” CASE



$$B = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$g = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

SUMMARY

For indefinite B,
Hard case occurs when g orthogonal to eigenvector u_1
for most negative eigenvalue λ_1

- ⊙ OK if radius is radius small enough
- ⊙ No "obvious" solution to equations ... but solution is actually of the form

$$s_{\text{lim}} + \sigma u_1$$

where

- ⋄ $s_{\text{lim}} = \lim_{\lambda \xrightarrow{+} -\lambda_1} s(\lambda)$
- ⋄ $\|s_{\text{lim}} + \sigma u_1\|_2 = \Delta$

HOW TO SOLVE $\|s(\lambda)\|_2 = \Delta$

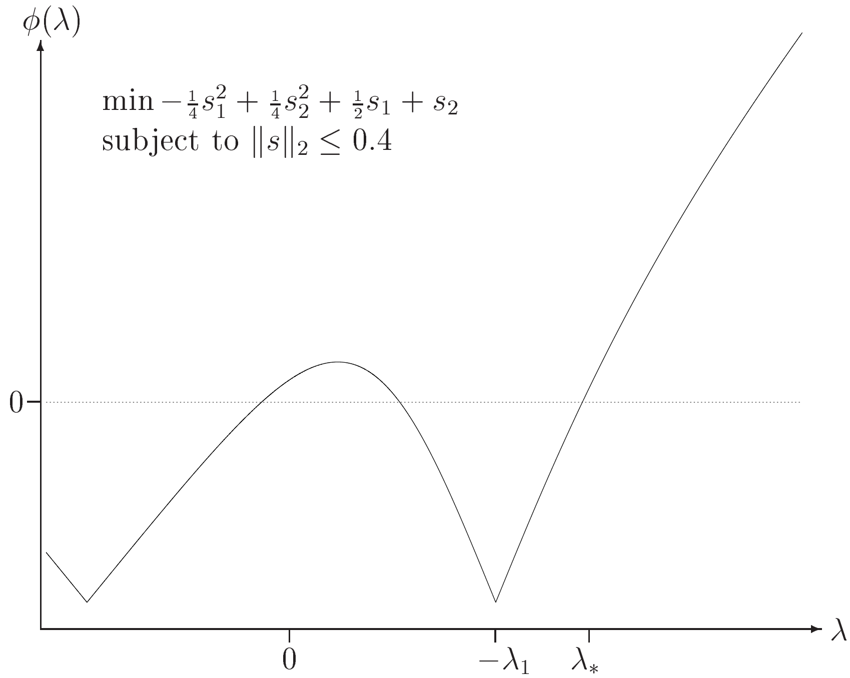
DON'T!!

Solve instead the **secular equation**

$$\phi(\lambda) \stackrel{\text{def}}{=} \frac{1}{\|s(\lambda)\|_2} - \frac{1}{\Delta} = 0$$

- ⊙ no poles
- ⊙ smallest at eigenvalues (except in hard case!)
- ⊙ analytic function $\Longrightarrow$ ideal for Newton
- ⊙ global convergent (ultimately quadratic rate except in hard case)
- ⊙ need to safeguard to protect Newton from the hard & interior solution cases

THE SECULAR EQUATION



NEWTON'S METHOD & THE SECULAR EQUATION

Let $\lambda > -\lambda_1$ and $\Delta > 0$ be given

Until "convergence" do:

Factorize $B + \lambda I = LL^T$

Solve $LL^T s = -g$

Solve $Lw = s$

Replace λ by

$$\lambda + \left(\frac{\|s\|_2 - \Delta}{\Delta}\right)\left(\frac{\|s\|_2^2}{\|w\|_2^2}\right)$$

SOLVING THE LARGE-SCALE PROBLEM

- when n is large, factorization may be impossible
- may instead try to use an iterative method to approximate
  - Steepest descent leads to the Cauchy point
  - obvious generalization: conjugate gradients ... but
    - what about the trust region?
    - what about negative curvature?

CONJUGATE GRADIENTS TO "MINIMIZE" q(s)

Given $s^0 = 0$, set $g^0 = g$, $d^0 = -g$ and $i = 0$
Until g^i "small" or breakdown, iterate

$$\alpha^i = \|g^i\|_2^2 / d^{i\,T} B d^i$$
$$s^{i+1} = s^i + \alpha^i d^i$$
$$g^{i+1} = g^i + \alpha^i B d^i$$
$$\beta^i = \|g^{i+1}\|_2^2 / \|g^i\|_2^2$$
$$d^{i+1} = -g^{i+1} + \beta^i d^i$$

and increase i by 1

Important features

- $g^j = Bs^j + g$ for all $j = 0, \ldots, i$
- $d^{j\,T} g^{i+1} = 0$ for all $j = 0, \ldots, i$
- $g^{j\,T} g^{i+1} = 0$ for all $j = 0, \ldots, i$

CRUCIAL PROPERTY OF CONJUGATE GRADIENTS

Theorem 3.10. Suppose that the conjugate gradient method is applied to minimize $q(s)$ starting from $s^0 = 0$, and that $d^{i\,T} B d^i > 0$ for $0 \leq i \leq k$. Then the iterates s^j satisfy the inequalities

$$\|s^j\|_2 < \|s^{j+1}\|_2$$

for $0 \leq j \leq k - 1$.

TRUNCATED CONJUGATE GRADIENTS

Apply the conjugate gradient method, but terminate at iteration i if

1. $d^{i\,T} B d^i \leq 0 \Longrightarrow$ problem unbounded along d^i
2. $\|s^i + \alpha^i d^i\|_2 > \Delta \Longrightarrow$ solution on trust-region boundary

In both cases, stop with $s_* = s^i + \alpha^{\mathrm{B}} d^i$, where α^{B} chosen as positive root of

$$\|s^i + \alpha^{\mathrm{B}} d^i\|_2 = \Delta$$

Crucially

$$q(s_*) \leq q(s^{\mathrm{C}}) \;\text{ and }\; \|s_*\|_2 \leq \Delta$$

$\Longrightarrow$ TR algorithm converges to a first-order critical point

HOW GOOD IS TRUNCATED C.G.?

In the convex case ... very good

Theorem 3.11. Suppose that the truncated conjugate gradient method is applied to minimize $q(s)$ and that B is positive definite. Then the computed and actual solutions to the problem, s_* and s_*^{M}, satisfy the bound

$$q(s_*) \leq \tfrac{1}{2} q(s_*^{\mathrm{M}})$$

In the non-convex case ... maybe poor

- e.g., if $g = 0$ and B is indefinite $\Longrightarrow$ $q(s_*) = 0$

WHAT CAN WE DO IN THE NON-CONVEX CASE?

Solve the problem over a subspace

- instead of the B-conjugate subspace for CG, use the equivalent Lanczos orthogonal basis
  - Gram-Schmidt applied to CG (Krylov) basis $\mathcal{D}^i$
  - Subspace $\mathcal{Q}^i = \{s \mid s = Q^i s_q \text{ for some } s_q \in \mathbb{R}^i\}$
  - Q^i is such that

    $$Q^{i\,T} Q^i = I \quad \text{and} \quad Q^{i\,T} B Q^i = T^i$$

    where T^i is tridiagonal and $Q^{i\,T} g = \|g\|_2\, e_1$
  - Q^i trivial to generate from CG $\mathcal{D}^i$

GENERALIZED LANCZOS TRUST-REGION METHOD

$$s^i = \arg\min_{s \in \mathcal{Q}^i} q(s) \;\text{ subject to }\; \|s\|_2 \leq \Delta$$

$\Longrightarrow \; s^i = Q^i s_q^i$, where

$$s_q^i = \arg\min_{s_q \in \mathbb{R}^i} \|g\|_2 \, e_1^T s_q + \tfrac{1}{2} s_q^T T^i s_q \;\text{ subject to }\; \|s_q\|_2 \leq \Delta$$

- ⊙ advantage T^i has very sparse factors $\Longrightarrow$ can solve the problem using the earlier secular equation approach
- ⊙ can exploit all the structure here $\Longrightarrow$ use solution for one problem to initialize next
- ⊙ until the trust-region boundary is reached, it **is** conjugate gradients $\Longrightarrow$ switch when we get there