

Part 5: SQP methods for equality constrained optimization

Nick Gould (RAL)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c(x) = 0$$

Misc course on nonlinear optimization

OPTIMALITY AND NEWTON'S METHOD

1st order optimality:

$$g(x, y) \equiv g(x) - A^T(x)y = 0 \quad \text{and} \quad c(x) = 0$$

nonlinear system (linear in y)

\implies

use Newton's method to find a correction (s, w) to (x, y)

\implies

$$\begin{pmatrix} H(x, y) & -A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = - \begin{pmatrix} g(x, y) \\ c(x) \end{pmatrix}$$

EQUALITY CONSTRAINED MINIMIZATION

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad c(x) = 0$$

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and the **constraints** $c : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ ($m \leq n$)

- assume that $f, c \in C^1$ (sometimes C^2) and Lipschitz
- often in practice this assumption violated, but not necessary
- easily generalized to inequality constraints ... but may be better to use interior-point methods for these

ALTERNATIVE FORMULATIONS

unsymmetric:

$$\begin{pmatrix} H(x, y) & -A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = - \begin{pmatrix} g(x, y) \\ c(x) \end{pmatrix}$$

or symmetric:

$$\begin{pmatrix} H(x, y) & A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ -w \end{pmatrix} = - \begin{pmatrix} g(x, y) \\ c(x) \end{pmatrix}$$

or (with $y^+ = y + w$) unsymmetric:

$$\begin{pmatrix} H(x, y) & -A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ y^+ \end{pmatrix} = - \begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$

or symmetric:

$$\begin{pmatrix} H(x, y) & A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ -y^+ \end{pmatrix} = - \begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$

DETAILS

- Often approximate with symmetric $B \approx H(x, y) \implies$ e.g.

$$\begin{pmatrix} B & A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ -y^+ \end{pmatrix} = - \begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$
- solve system using
 - unsymmetric (LU) factorization of $\begin{pmatrix} B & -A^T(x) \\ A(x) & 0 \end{pmatrix}$
 - symmetric (indefinite) factorization of $\begin{pmatrix} B & A^T(x) \\ A(x) & 0 \end{pmatrix}$
 - symmetric factorizations of B and the Schur Complement $A(x)B^{-1}A^T(x)$
 - iterative method (GMRES(k), MINRES, CG within $\mathcal{N}(A), \dots$)

SEQUENTIAL QUADRATIC PROGRAMMING - SQP

or **successive** quadratic programming
or **recursive** quadratic programming (RQP)

Given (x_0, y_0) , set $k = 0$
 Until “convergence” iterate:
 Compute a suitable symmetric B_k using (x_k, y_k)
 Find
 $s_k = \arg \min_{s \in \mathbb{R}^n} g_k^T s + \frac{1}{2} s^T B_k s$ subject to $A_k s = -c_k$
 along with associated Lagrange multiplier estimates y_{k+1}
 Set $x_{k+1} = x_k + s_k$ and increase k by 1

AN ALTERNATIVE INTERPRETATION

- QP** : minimize $g(x)^T s + \frac{1}{2} s^T B s$ subject to $A(x)s = -c(x)$
 $s \in \mathbb{R}^n$
- QP = **quadratic program**
 - first-order model of constraints $c(x + s)$
 - second-order model of objective $f(x + s) \dots$ but B includes curvature of constraints

solution to QP satisfies

$$\begin{pmatrix} B & A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ -y^+ \end{pmatrix} = - \begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$

ADVANTAGES

- simple
- fast
 - quadratically convergent with $B_k = H(x_k, y_k)$
 - superlinearly convergent with good $B_k \approx H(x_k, y_k)$
 - ▷ don't actually need $B_k \longrightarrow H(x_k, y_k)$

PROBLEMS WITH PURE SQP

- how to choose B_k ?
- what if QP_k is unbounded from below? and when?
- how do we globalize this iteration?

QP SUB-PROBLEM

$$\text{minimize}_{s \in \mathbb{R}^n} g^T s + \frac{1}{2} s^T B s \text{ subject to } A s = -c$$

- need constraints to be consistent
 - OK if A is full rank

- need B to be positive (semi-) definite when $A s = 0$

\Leftrightarrow

$N^T B N$ positive (semi-) definite where the columns of N form a basis for $\text{null}(A)$

\Leftrightarrow

$$\begin{pmatrix} B & A^T \\ A & 0 \end{pmatrix}$$

(is non-singular and) has $m - \text{ve}$ eigenvalues

LINESEARCH SQP METHODS

$$s_k = \arg \min_{s \in \mathbb{R}^n} g_k^T s + \frac{1}{2} s^T B_k s \text{ subject to } A_k s = -c_k$$

Basic idea:

- Pick $x_{k+1} = x_k + \alpha_k s_k$, where
- α_k is chosen so that

$$\Phi(x_k + \alpha_k s_k, p_k) \stackrel{<}{\sim} \Phi(x_k, p_k)$$

- $\Phi(x, p)$ is a ‘suitable’ merit function
- p_k are parameters

- vital that s_k is a descent direction for $\Phi(x, p_k)$ at x_k
- normally require that B_k is positive definite

SUITABLE MERIT FUNCTIONS. I

The **quadratic penalty function**:

$$\Phi(x, \mu) = f(x) + \frac{1}{2\mu} \|c(x)\|_2^2$$

Theorem 5.1. Suppose that B_k is positive definite, and that (s_k, y_{k+1}) are the SQP search direction and its associated Lagrange multiplier estimates for the problem

$$\text{minimize}_{x \in \mathbb{R}^n} f(x) \text{ subject to } c(x) = 0$$

at x_k . Then if x_k is not a first-order critical point, s_k is a descent direction for the quadratic penalty function $\Phi(x, \mu_k)$ at x_k whenever

$$\mu_k \leq \frac{\|c(x_k)\|_2}{\|y_{k+1}\|_2}$$

PROOF OF THEOREM 5.1

SQP direction s_k and associated multiplier estimates y_{k+1} satisfy

$$B_k s_k - A_k^T y_{k+1} = -g_k \tag{1}$$

and

$$A_k s_k = -c_k. \tag{2}$$

$$(1) + (2) \implies s_k^T g_k = -s_k^T B_k s_k + s_k^T A_k^T y_{k+1} = -s_k^T B_k s_k - c_k^T y_{k+1} \tag{3}$$

$$(2) \implies \frac{1}{\mu_k} s_k^T A_k^T c_k = -\frac{\|c_k\|_2^2}{\mu_k}. \tag{4}$$

(3) + (4), the positive definiteness of B_k , the Cauchy-Schwarz inequality, the required bound on μ_k , and $s_k \neq 0$ if x_k is not critical \implies

$$\begin{aligned} s_k^T \nabla_x \Phi(x_k) &= s_k^T \left(g_k + \frac{1}{\mu_k} A_k^T c_k \right) = -s_k^T B_k s_k - c_k^T y_{k+1} - \frac{\|c_k\|_2^2}{\mu_k} \\ &< -\|c_k\|_2 \left(\frac{\|c_k\|_2}{\mu_k} - \|y_{k+1}\|_2 \right) \leq 0 \end{aligned}$$

NON-DIFFERENTIABLE EXACT PENALTIES

The **non-differentiable exact penalty function**:

$$\Phi(x, \rho) = f(x) + \rho \|c(x)\|$$

for any norm $\|\cdot\|$ and scalar $\rho > 0$.

Theorem 5.2. Suppose that $f, c \in C^2$, and that x_* is an isolated local minimizer of $f(x)$ subject to $c(x) = 0$, with corresponding Lagrange multipliers y_* . Then x_* is also an isolated local minimizer of $\Phi(x, \rho)$ provided that

$$\rho > \|y_*\|_D,$$

where the **dual norm**

$$\|y\|_D = \sup_{x \neq 0} \frac{y^T x}{\|x\|}.$$

SUITABLE MERIT FUNCTIONS. II

The non-differentiable exact penalty function:

$$\Phi(x, \rho) = f(x) + \rho \|c(x)\|$$

for any norm $\|\cdot\|$ (with dual norm $\|\cdot\|_D$) and scalar $\rho > 0$.

Theorem 5.3. Suppose that B_k is positive definite, and that (s_k, y_{k+1}) are the SQP search direction and its associated Lagrange multiplier estimates for the problem

$$\text{minimize } f(x) \text{ subject to } c(x) = 0 \\ x \in \mathbb{R}^n$$

at x_k . Then if x_k is not a first-order critical point, s_k is a descent direction for the non-differentiable penalty function $\Phi(x, \rho_k)$ at x_k whenever $\rho_k \geq \|y_{k+1}\|_D$

PROOF OF THEOREM 5.3

Taylor's theorem applied to f and c + (2) \implies (for small α)

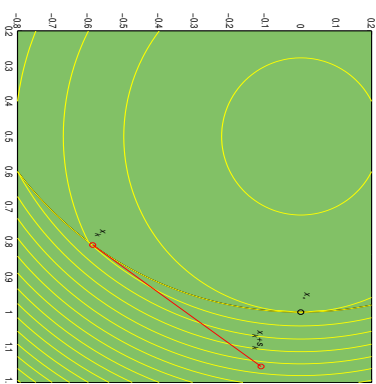
$$\begin{aligned} \Phi(x_k + \alpha s_k, \rho_k) - \Phi(x_k, \rho_k) &= \alpha s_k^T g_k + \rho_k (\|c_k + \alpha A_k s_k\| - \|c_k\|) + O(\alpha^2) \\ &= \alpha s_k^T g_k + \rho_k (\|(1 - \alpha)c_k\| - \|c_k\|) + O(\alpha^2) \\ &= \alpha (s_k^T g_k - \rho_k \|c_k\|) + O(\alpha^2) \end{aligned}$$

+ (3), the positive definiteness of B_k , the Hölder inequality, and $s_k \neq 0$ if x_k is not critical \implies

$$\begin{aligned} \Phi(x_k + \alpha s_k, \rho_k) - \Phi(x_k, \rho_k) &= -\alpha (s_k^T B_k s_k + c_k^T y_{k+1} + \rho_k \|c_k\|) + O(\alpha^2) \\ &< -\alpha (-\|c_k\| \|y_{k+1}\|_D + \rho_k \|c_k\|) + O(\alpha^2) \\ &= -\alpha \|c_k\| (\rho_k - \|y_{k+1}\|_D) + O(\alpha^2) < 0 \end{aligned}$$

because of the required bound on ρ_k , for sufficiently small α . Hence sufficiently small steps along s_k from non-critical x_k reduce $\Phi(x, \rho_k)$.

THE MARATOS EFFECT



l_1 non-differentiable exact penalty function ($\rho = 1$):
 $f(x) = 2(x_1^2 + x_2^2 - 1) - x_1$
 and $c(x) = x_1^2 + x_2^2 - 1$
 solution: $x_* = (1, 0)$, $y_* = \frac{3}{2}$

Maratos effect: merit function may prevent acceptance of the SQP step arbitrarily close to x_* \implies slow convergence

AVOIDING THE MARATOS EFFECT

The Maratos effect occurs because the curvature of the constraints is not adequately represented by linearization in the SQP model:

$$c(x_k + s_k) = O(\|s_k\|^2)$$

⇒ need to correct for this curvature

⇒ use a **second-order correction** from $x_k + s_k$:

$$c(x_k + s_k + s_k^c) = o(\|s_k\|^2)$$

also do not want to destroy potential for fast convergence ⇒

$$s_k^c = o(s_k)$$

POPULAR 2ND-ORDER CORRECTIONS

○ minimum norm solution to $c(x_k + s_k) + A(x_k + s_k)s_k^c = 0$

$$\begin{pmatrix} I & A^T(x_k + s_k) \\ A(x_k + s_k) & 0 \end{pmatrix} \begin{pmatrix} s_k^c \\ -y_{k+1}^c \end{pmatrix} = - \begin{pmatrix} 0 \\ c(x_k + s_k) \end{pmatrix}$$

○ minimum norm solution to $c(x_k + s_k) + A(x_k)s_k^c = 0$

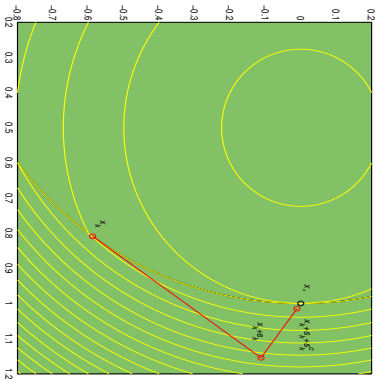
$$\begin{pmatrix} I & A^T(x_k) \\ A(x_k) & 0 \end{pmatrix} \begin{pmatrix} s_k^c \\ -y_{k+1}^c \end{pmatrix} = - \begin{pmatrix} 0 \\ c(x_k + s_k) \end{pmatrix}$$

○ another SQP step from $x_k + s_k$

$$\begin{pmatrix} H(x_k + s_k, y_k^+) & A^T(x_k + s_k) \\ A(x_k + s_k) & 0 \end{pmatrix} \begin{pmatrix} s_k^c \\ -y_{k+1}^c \end{pmatrix} = - \begin{pmatrix} q(x_k + s_k) \\ c(x_k + s_k) \end{pmatrix}$$

○ etc., etc.

2ND-ORDER CORRECTIONS IN ACTION



l_1 non-differentiable exact penalty function ($\rho = 1$):
 $f(x) = 2(x_1^2 + x_2^2 - 1) - x_1$
 and $c(x) = x_1^2 + x_2^2 - 1$
 solution: $x_* = (1, 0)$, $y_* = \frac{3}{2}$

○ (very) fast convergence

○ $x_k + s_k + s_k^c$ reduces $\Phi \Rightarrow$ global convergence

TRUST-REGION SQP METHODS

Obvious trust-region approach:

$s_k = \arg \min_{s \in \mathbb{R}^n} q_k^T s + \frac{1}{2} s^T B_k s$ subject to $A_k s = -c_k$ and $\|s\| \leq \Delta_k$

○ do not require that B_k be positive definite

⇒ can use $B_k = H(x_k, y_k)$

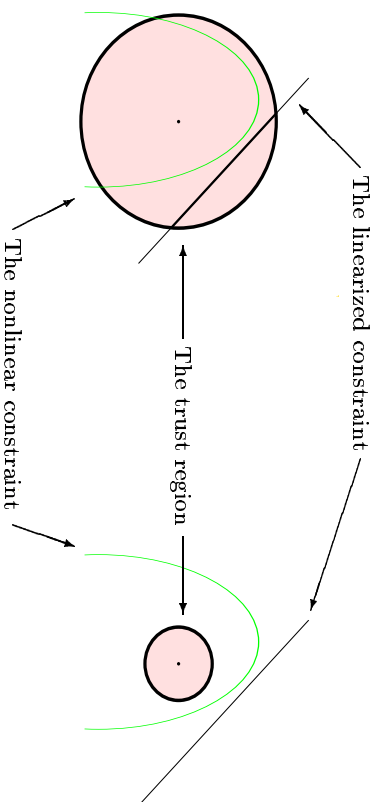
○ if $\Delta_k < \Delta_{\text{CRTR}}$ where

$$\Delta_{\text{CRTR}} \stackrel{\text{def}}{=} \min \|s\| \text{ subject to } A_k s = -c_k$$

⇒ **no solution to trust-region subproblem**

⇒ simple trust-region approach to SQP is flawed if $c_k \neq 0 \Rightarrow$ need to consider alternatives

INFEASIBILITY OF THE SQP STEP



ALTERNATIVES

- the $S\ell_p$ QP method of Fletcher
- composite step SQP methods
 - ◊ constraint relaxation (Vardi)
 - ◊ constraint reduction (Byrd–Omojokun)
 - ◊ constraint lumping (Celis–Dennis–Tapia)
- the filter-SQP approach of Fletcher and Leyffer

THE $S\ell_p$ QP METHOD

Try to minimize the ℓ_p -(exact) penalty function

$$\Phi(x, \rho) = f(x) + \rho \|c(x)\|_p$$

for sufficiently large $\rho > 0$ and some ℓ_p norm ($1 \leq p \leq \infty$), using a trust-region approach

Suitable model problem: ℓ_p QP

minimize $(f_k +) g_k^T s + \frac{1}{2} s^T B_k s + \rho \|c_k + A_k s\|_p$ subject to $\|s\| \leq \Delta_k$
 $s \in \mathbb{R}^n$

- model problem always consistent
- when ρ and Δ_k are large enough, model minimizer = SQP direction
- when the norms are polyhedral (e.g., ℓ_1 or ℓ_∞ norms), ℓ_p QP is equivalent to a quadratic program ...

THE ℓ_1 QP SUBPROBLEM

ℓ_1 QP model problem with an ℓ_∞ trust region

minimize $g_k^T s + \frac{1}{2} s^T B_k s + \rho \|c_k + A_k s\|_1$ subject to $\|s\|_\infty \leq \Delta_k$
 $s \in \mathbb{R}^n$

But

$$c_k + A_k s = u - v, \quad \text{where } (u, v) \geq 0$$

$\implies \ell_1$ QP equivalent to quadratic program (QP):

minimize $g_k^T s + \frac{1}{2} s^T B_k s + \rho(e^T u + e^T v)$
 $s \in \mathbb{R}^n, u, v \in \mathbb{R}^m$
 subject to $A_k s - u + v = -c_k$
 and $u \geq 0, v \geq 0$
 and $-\Delta_k e \leq s \leq \Delta_k e$

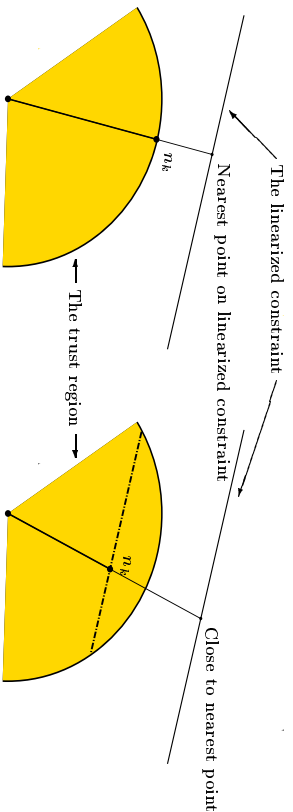
- good methods for solving QP
- can exploit structure of u and v variables

PRACTICAL S_{l_1} QP METHODS

- Cauchy point requires solution to l_1 LP model:

$$\underset{s \in \mathbb{R}^n}{\text{minimize}} \quad g_k^T s + \rho \|c_k + A_k s\| \quad \text{subject to} \quad \|s\|_\infty \leq \Delta_k$$
- approximate solutions to both l_1 LP and l_1 QP subproblems suffice
- need to adjust ρ as method progresses
- easy to generalize to inequality constraints
- globally convergent, but needs second-order correction for fast asymptotic convergence
- if $c(x) = 0$ are inconsistent, converges to (locally) least value of infeasibility $\|c(x)\|$

NORMAL AND TANGENTIAL STEPS



Points on dotted line are all potential tangential steps

COMPOSITE-STEP METHODS

Aim: find **composite step**

$$s_k = n_k + t_k$$

where

the **normal step** n_k moves towards feasibility of the linearized constraints (within the trust region)

$$\|A_k n_k + c_k\| < \|c_k\|$$

(model objective may get worse)

the **tangential step** t_k reduces the model objective function (within the trust-region) without sacrificing feasibility obtained from n_k

$$A_k(n_k + t_k) = A_k n_k \implies A_k t_k = 0$$

CONSTRAINT RELAXATION — YARDI

normal step: relax

$$A_k s = -c_k \quad \text{and} \quad \|s\| \leq \Delta_k$$

to

$$A_k n = -\sigma_k c_k \quad \text{and} \quad \|n\| \leq \Delta_k$$

where $\sigma_k \in [0, 1]$ is small enough so that there is a feasible n_k

tangential step:

$$\begin{aligned} & \text{(approximate)} \quad \arg \min_{t \in \mathbb{R}^n} (g_k + B_k n_k)^T t + \frac{1}{2} t^T B_k t \\ & \text{subject to} \quad A_k t = 0 \quad \text{and} \quad \|n_k + t\| \leq \Delta_k \end{aligned}$$

Snags:

- choice of σ_k
- incompatible constraints

CONSTRAINT REDUCTION — BYRD-OMOJOKUN

normal step: replace

$$A_k s = -c_k \text{ and } \|s\| \leq \Delta_k$$

by

approximately minimize $\|A_k n + c_k\|$ subject to $\|n\| \leq \Delta_k$

tangential step: as in Vardi

- use conjugate gradients to solve both subproblems
 \implies Cauchy points in both cases
- globally convergent using ℓ_2 merit function
- basis of successful **KNITRRO** package

FILTER METHODS — FLETCHER AND LEYFFER

Rationale:

- trust-region and linearized constraints compatible if c_k is small enough so long as $c(x) = 0$ is compatible
 \implies if trust-region subproblem incompatible, simply move closer to constraints
- merit functions depend on arbitrary parameters
 \implies use a different mechanism to measure progress

Let $\theta = \|c(x)\|$

A **filter** is a set of pairs $\{(\theta_k, f_k)\}$ such that no member dominates another, i.e., it does not happen that

$$\theta_i < \theta_j \text{ and } f_i < f_j$$

for any pair of filter points $i \neq j$

CONSTRAINT LUMPING — CELIS-DENNIS-TAPIA

normal step: replace

$$A_k s = -c_k \text{ and } \|s\| \leq \Delta_k$$

by

$$\|A_k n + c_k\| \leq \sigma_k \text{ and } \|n\| \leq \Delta_k$$

where $\sigma_k \in [0, \|c_k\|]$ is large enough so that there is a feasible n_k

tangential step:

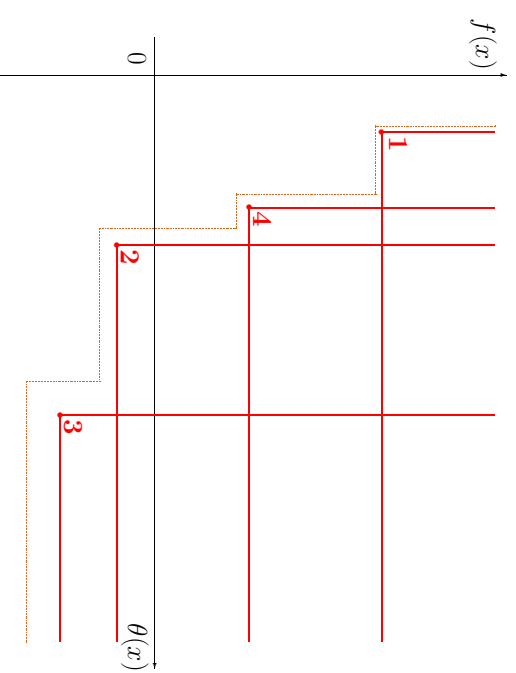
(approximate) $\arg \min_{t \in \mathbb{R}^n} (g_k + B_k n_k)^T t + \frac{1}{2} t^T B_k t$

subject to $\|A_k t + A_k n_k + c_k\| \leq \sigma_k$ and $\|t + n_k\| \leq \Delta_k$

Snags:

- choice of σ_k
- tangential subproblem is (NP?) hard

A FILTER WITH FOUR ENTRIES



BASIC FILTER METHOD

- if possible find

$$s_k = \arg \min_{s \in \mathbb{R}^n} q_k^T s + \frac{1}{2} s^T B_k s \text{ subject to } A_k s = -c_k \text{ and } \|s\| \leq \Delta_k$$

- otherwise, find s_k :

$$\theta(x_k + s_k) \stackrel{!}{\leq} \theta_i \text{ for all } i \leq k$$

- if $x_k + s_k$ is “acceptable” for the filter, set $x_{k+1} = x_k + s_k$ and possibly increase Δ_k and “prune” filter
- otherwise reduce Δ_k and try again

In practice, far more complicated than this!