Lecture 15: SQP methods for equality constrained optimization

Coralia Cartis, Mathematical Institute, University of Oxford

C6.2/B2: Continuous Optimization
Nonlinear equality-constrained problems – again!

\[ \min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad c(x) = 0, \]  

(eCP)

where \( f : \mathbb{R}^n \to \mathbb{R}, \quad c = (c_1, \ldots, c_m) : \mathbb{R}^n \to \mathbb{R}^m \) are \( C^1 \) or \( C^2 \) (when needed), with \( m \leq n \).

- easily generalized to inequality constraints ... but may be better to use interior-point methods for these.

- (eCP): attempt to find local solutions or at least KKT points:

\[ \nabla L(x, y) = \nabla f(x) - J(x)^T y = 0 \quad \text{and} \quad c(x) = 0 \]

nonlinear and square system in \( x \) and \( y \) (linear in \( y \)) \( \Rightarrow \) use Newton’s method to find a correction \((s, w)\) to \((x, y)\)

\[
\begin{pmatrix}
\nabla^2 L(x, y) & -J(x)^T \\
J(x) & 0
\end{pmatrix}
\begin{pmatrix}
s \\
w
\end{pmatrix}
= -
\begin{pmatrix}
\nabla L(x, y) \\
c(x)
\end{pmatrix}
\]
Alternative formulations

Recall: the Lagrangian function is $\mathcal{L}(x, y) = f(x) - y^T c(x)$. 

 unsymmetric: 

$$
\left( \begin{array}{cc} 
\nabla^2 \mathcal{L}(x, y) & -J(x)^T \\
J(x) & 0 
\end{array} \right) \left( \begin{array}{c}
s \\
w 
\end{array} \right) = - \left( \begin{array}{c} 
\nabla \mathcal{L}(x, y) \\
c(x) 
\end{array} \right)
$$

 or symmetric: 

$$
\left( \begin{array}{cc} 
\nabla^2 \mathcal{L}(x, y) & J(x)^T \\
J(x) & 0 
\end{array} \right) \left( \begin{array}{c}
s \\
-w 
\end{array} \right) = - \left( \begin{array}{c} 
\nabla \mathcal{L}(x, y) \\
c(x) 
\end{array} \right)
$$

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

[unsymmetric also possible]

$$
\left( \begin{array}{cc} 
\nabla^2 \mathcal{L}(x, y) & J(x)^T \\
J(x) & 0 
\end{array} \right) \left( \begin{array}{c}
s \\
-y^+ 
\end{array} \right) = - \left( \begin{array}{c} 
\nabla f(x) \\
c(x) 
\end{array} \right)
$$

 Lecture 15: SQP methods for equality constrained optimization – p. 3/22
Some more details

- often approximate with symmetric \( B \approx \nabla^2 \mathcal{L}(x, y) \) e.g.

\[
\begin{pmatrix}
  B & J(x)^T \\
  J(x) & 0
\end{pmatrix}
\begin{pmatrix}
  s \\
  -y^+
\end{pmatrix}
= -
\begin{pmatrix}
  \nabla f(x) \\
  c(x)
\end{pmatrix}
\]

- solve system using
  - unsymmetric (LU) factorization of
    \[
    \begin{pmatrix}
      B & -J(x)^T \\
      J(x) & 0
    \end{pmatrix}
    \]
  - symmetric (indefinite) factorization of
    \[
    \begin{pmatrix}
      B & J(x)^T \\
      J(x) & 0
    \end{pmatrix}
    \]
  - symmetric factorizations of \( B \) and the Schur complement \( J(x)B^{-1}J(x)^T \)
  - iterative method (GMRES(k), MINRES, CG within \( \mathcal{N}(J) \), etc.)
An alternative interpretation

\[ \text{QP : minimize } \begin{array}{c}
\minimize_{s \in \mathbb{R}^n} \\
\nabla f(x)^T s + \frac{1}{2} s^T B s \nabla f(x)
\end{array}
\text{ subject to } J(x) s = -c(x) \]

- QP = quadratic program
- first-order model of constraints \( c(x + s) \)
- second-order model of objective \( f(x + s) \) but \( B \) includes curvature of constraints

Solution to (QP) satisfies

\[
\begin{pmatrix}
B & J(x)^T \\
J(x) & 0
\end{pmatrix}
\begin{pmatrix}
s \\
-y^+
\end{pmatrix}
= -
\begin{pmatrix}
\nabla f(x) \\
c(x)
\end{pmatrix}
\]
Sequential quadratic programming - SQP

or successive quadratic programming
or recursive quadratic programming (RQP)

A basic SQP method

Given \((x^0, y^0)\), set \(k = 0\)

Until “convergence” iterate:

- Compute a suitable symmetric \(B^k\) using \((x^k, y^k)\)

- Find the solution of (QP\(_k\))

\[
s^k = \arg \min_{s \in \mathbb{R}^n} \nabla f(x^k)^T s + \frac{1}{2} s^T B^k s \quad \text{subject to} \quad J(x^k)s = -c(x^k)
\]

along with associated Lagrange multiplier estimates \(y^{k+1}\).

- Set \(x^{k+1} = x^k + s^k\) and let \(k := k + 1\).
Sequential quadratic programming - SQP...

Advantages of SQP:

- simple
- fast
  - quadratically convergent with \( B^k = \nabla^2 \mathcal{L}(x^k, y^k) \)
  - superlinearly convergent with good \( B^k \approx \nabla^2 \mathcal{L}(x^k, y^k) \)
  - don’t actually need \( B^k \rightarrow \nabla^2 \mathcal{L}(x^k, y^k) \)

Issues with pure SQP [similar to Newton’s method for unconstrained opt and systems]

- how to choose \( B^k \)?
- what if \( QP_k \) is unbounded from below? and when?
- how do we globalize the SQP iteration?
The QP\(_k\) subproblem:

\[
\begin{align*}
\text{minimize} & \quad \nabla f(x^k)^T s + \frac{1}{2} s B^k s \\
\text{subject to} & \quad J(x^k) s = -c(x^k)
\end{align*}
\]

- need constraints to be consistent
  - OK if \(J(x^k)\) is full rank
- need \(B^k\) to be positive (semi-) definite when \(J(x^k) s = 0\)

\[
\iff \quad N_k^T B^k N_k \text{ positive (semi-) definite where the columns of }\ N_k \text{ form a basis for the null space of } J(x^k) \iff
\]

\[
\begin{pmatrix}
B^k & J(x^k)^T \\
J(x^k) & 0
\end{pmatrix}
\]

(is non-singular and) has \(m\) –ve eigenvalues
Linesearch SQP methods

\[ s^k = \arg \min_{s \in \mathbb{R}^n} \nabla f(x^k)^T s + \frac{1}{2} s^T B^k s \quad \text{subject to} \quad J(x^k)^T s = -c(x^k) \]

Linesearch SQP:

- Set \( x^{k+1} = x^k + \alpha^k s^k \), where \( \alpha^k \) is chosen so that

\[
\Phi(x^k + \alpha^k s^k, \sigma^k) \lesssim \Phi(x^k, \sigma^k)
\]

where \( \Phi(x, \sigma) \) is a “suitable” merit function and \( \sigma^k \) are parameters.

Recall unconstrained GLM: crucial that \( s^k \) is a descent direction for \( \Phi(x, \sigma^k) \) at \( x^k \); normally require that \( B^k \) is positive definite.
Suitable merit functions for SQP

Recall the quadratic penalty function:

\[ \Phi(x, \sigma) = f(x) + \frac{1}{2\sigma} \| c(x) \|^2 \]

**Theorem 30:** 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 the (eCP) problem at \( x^k \). Then if \( x^k \) is not a KKT point of (eCP), then \( s^k \) is a descent direction for the quadratic penalty function \( \Phi(x, \sigma^k) \) at \( x^k \) whenever

\[ \sigma^k \leq \frac{\| c(x^k) \|}{\| y^{k+1} \|} . \]
Suitable merit functions for SQP ...

Proof of Theorem 30:
SQP direction $s^k$ and associated multiplier estimates $y^{k+1}$ satisfy

$$B^k s^k - J(x^k)^T y^{k+1} = -\nabla f(x^k) \quad (*)$$

$$J(x^k) s^k = -c(x^k) \quad (***)$$

$$(*) + (***) \implies (s^k)^T \nabla f(x^k) = -(s^k)^T B^k s^k + (s^k)^T J(x^k)^T y^{k+1}$$

$$= -(s^k)^T B^k s^k - c(x^k)^T y^{k+1}$$

$$(***) \implies c(x^k)^T J(x^k) s^k = -\|c(x^k)\|^2.$$ 

These, the positive definiteness of $B^k$, the Cauchy-Schwarz inequality, the required bound on $\sigma^k$, and $s^k \neq 0$ if $x^k$ is not critical
Proof of Theorem 30: (continued)

\[(s^k)^T \nabla_x \Phi(x^k, \sigma^k) = (s^k)^T \left( \nabla f(x^k) + \frac{1}{\sigma^k} J(x^k)^T c(x^k) \right)\]

\[= -(s^k)^T B^k s^k - c(x^k)^T y^{k+1} - \frac{\|c(x^k)\|^2}{\sigma^k}\]

\[< -\|c(x^k)\| \left( \frac{\|c(x^k)\|}{\sigma^k} - \|y^{k+1}\| \right) \leq 0,\]

and so \(s^k\) is descent for \(\Phi(x^k, \sigma^k)\). \(\square\)

Other suitable merit functions:

The non-differentiable exact penalty function: [widely used]

\[\Psi(x, \rho) = f(x) + \rho \|c(x)\|, \text{ where } \| \cdot \| \text{ can be any norm and } \rho > 0.\]

- recall that minimizers of (eCP) correspond to those of \(\Psi(x, \rho)\) for \(\rho\) sufficiently large (and finite! \(\rho > \|y^*\|_D\));
- equivalent of Th 30 holds \((\rho^k \geq \|y^{k+1}\|_D)\).
The Maratos effect

- merit function may prevent acceptance of the full SQP step (so $\alpha^k \neq 1$) arbitrarily close to $x^*$ $\implies$ slow convergence

$$f(x) = 2(x_1^2 + x_2^2 - 1) - x_1$$ and $$c(x) = x_1^2 + x_2^2 - 1;$$ solution: $x^* = (1, 0), y^* = \frac{3}{2}.$ Here: $\ell_1$ non-differentiable exact penalty function ($\rho = 1$) but other merit fcts. have similar behaviour.
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 \( x^k + s^k \):

\[ c(x^k + s^k + s_C^k) = o(\|s^k\|^2). \]

Also, do not want to destroy potential for fast convergence

⇒ \( s_C^k = o(s^k) \).
Popular second-order corrections

- minimum norm solution to \( c(x^k + s^k) + J(x^k + s^k)s_C^k = 0 \)

\[
\begin{pmatrix}
I & J(x^k + s^k)^T \\
J(x^k + s^k) & 0 \\
\end{pmatrix}
\begin{pmatrix}
s_C^k \\
-y_C^{k+1} \\
\end{pmatrix}
= -
\begin{pmatrix}
0 \\
c(x^k + s^k) \\
\end{pmatrix}
\]

- minimum norm solution to \( c(x^k + s^k) + J(x^k)s_C^k = 0 \)

\[
\begin{pmatrix}
I & J(x^k)^T \\
J(x^k) & 0 \\
\end{pmatrix}
\begin{pmatrix}
s_C^k \\
-y_C^{k+1} \\
\end{pmatrix}
= -
\begin{pmatrix}
0 \\
c(x^k + s^k) \\
\end{pmatrix}
\]

- another SQP step from \( x^k + s^k \) ...and so on...

\[
\begin{pmatrix}
\nabla^2 \mathcal{L}(x^k + s^k, y^k_+) & J(x^k + s^k)^T \\
J(x^k + s^k) & 0 \\
\end{pmatrix}
\begin{pmatrix}
s_C^k \\
-y_C^{k+1} \\
\end{pmatrix}
= -
\begin{pmatrix}
\nabla \mathcal{L}(x^k + s^k) \\
c(x^k + s^k) \\
\end{pmatrix}
\]
Second-order corrections in action

\[ f(x) = 2(x_1^2 + x_2^2 - 1) - x_1 \text{ and } c(x) = x_1^2 + x_2^2 - 1; \text{ solution:} \]
\[ x^* = (1, 0), \quad y^* = \frac{3}{2}. \]

Here: \( \ell_1 \) non-differentiable exact penalty function (\( \rho = 1 \)) but other merit fcts. have similar behaviour.

- fast convergence
- \( x^k + s^k + s^k_C \) reduces \( \Phi \implies \text{global convergence} \)
Trust-region SQP methods

Obvious trust-region approach:

\[
s^k = \arg \min_{s \in \mathbb{R}^n} \nabla f(x^k)^T s + \frac{1}{2} s^T B_k s
\]

subject to \( J(x^k)s = -c(x^k) \) and \( \|s\| \leq \Delta_k \)

- do not require that \( B_k \) be positive definite
  \[\implies\] can use \( B_k = \nabla^2 \mathcal{L}(x^k, y^k) \)

- if \( \Delta_k < \Delta_{\text{CRIT}} \) where

\[
\Delta_{\text{CRIT}} \overset{\text{def}}{=} \min \|s\| \text{ subject to } J(x^k)s = -c(x^k)
\]

\[\implies\] no solution to trust-region subproblem

\[\implies\] simple trust-region approach to SQP is flawed if \( c(x^k) \neq 0 \)

\[\implies\] need to consider alternatives
Infeasibility of the SQP step

The linearized constraint

The trust region

The nonlinear constraint
An alternative: 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)

\[ \| J(x^k)n^k + c(x^k) \| < \| c(x^k) \| \]

(model objective may get worse)

the tangential step \( t^k \) reduces the model objective function (within the trust-region) without sacrificing feasibility obtained from \( n^k \)

\[ J(x^k)(n^k + t^k) = J(x^k)n^k \implies J(x^k)t^k = 0 \]
Normal and tangential steps

Points on dotted line are all potential tangential steps

The linearized constraint

Nearest point on linearized constraint

Close to nearest point

The trust region
Constraint reduction [Byrd-Omojokun]

**normal step:** replace

\[ J(x^k)s = -c(x^k) \text{ and } \|s\| \leq \Delta_k \]

by

approximately minimize \[ \| J(x^k)n + c(x^k) \| \text{ subject to } \|n\| \leq \Delta_k \]

**tangential step:**

(approximate) \[ \arg \min_{t \in \mathbb{R}^n} (\nabla f(x^k) + B^k n^k)^T t + \frac{1}{2} t^T B^k t \]

subject to \[ J(x^k)t = 0 \text{ and } \|n^k + t\| \leq \Delta_k \]

- use conjugate gradients to solve both subproblems
  \[ \Rightarrow \text{ Cauchy conditions satisfied in both cases} \]

- globally convergent

- basis of successful **KNITRO** package
Other alternatives

- composite step SQP methods
  - constraint relaxation (Vardi) - *not addressed here*
  - constraint reduction (Byrd–Omojokun) - *covered*
  - constraint lumping (Celis–Dennis–Tapia) - *not addressed*
- the $S\ell_p$QP method of Fletcher - *not addressed here*
- the filter-SQP approach of Fletcher and Leyffer - *not addressed*

An important class of methods not addressed: active-set methods for linearly-constrained nonlinear problems (ie, generalization of simplex methods from the LP to the nonlinear case).