Lecture 16: SQP methods for equality constrained optimization

Coralia Cartis, Mathematical Institute, University of Oxford

C6.2/B2: Continuous Optimization

Nonlinear equality-constrained problems – again!

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

where $f : \mathbb{R}^n \to \mathbb{R}$, $c = (c_1, \ldots, c_m) : \mathbb{R}^n \to \mathbb{R}^m$ are $\mathcal{C}^1$ or $\mathcal{C}^2$ (when needed), with $m \leq n$; $J(x)$ Jacobian of c, with rows $\nabla c_i(x)^T$.

- 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:

$$\nabla_x \mathcal{L}(x, y) = \nabla f(x) - J(x)^T y = 0 \text{ and } c(x) = 0 \qquad (*)$$

where $\mathcal{L}(x, y) = f(x) - y^T c(x) = f(x) - \sum_{i=1}^m y_i c_i(x)$ Lagrangian.
(*) nonlinear and square system in x and y (linear in y) $\Longrightarrow$ use Newton's method for root finding, to find a change (s, w) to (x, y)

Newton iteration for KKT system

Newton step for the KKT system $\nabla_x \mathcal{L}(x,y) = \nabla f(x) - J(x)^T y = 0;\ c(x) = 0$ is:

$$\begin{cases} \left(\nabla^2 f(x) - \sum_{i=1}^m y_i \nabla^2 c_i(x) \quad - J(x)^T\right) \begin{pmatrix} s \\ w \end{pmatrix} = -(\nabla f(x) - J(x)^T y) \\ J(x)s = -c(x) \end{cases}$$

which is equivalent to

$$\begin{pmatrix} \nabla_x^2 \mathcal{L}(x,y) & -J(x)^T \\ J(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = - \begin{pmatrix} \nabla_x \mathcal{L}(x,y) \\ c(x) \end{pmatrix}$$

or (with $y^+ = y + w$) : [symmetric formulations also possible]

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

often approximate $\nabla_x^2 \mathcal{L}(x,y)$ with symmetric B

An alternative interpretation

QP : $\underset{s \in \mathbb{R}^n}{\text{minimize}} \quad m(s) := \nabla f(x)^T s + \frac{1}{2} s^T B s$ 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 of (QP) satisfies KKT conditions with Lagrange multipliers $y^+ \in \mathbb{R}^m$: $\nabla_s m(s) = \nabla f(x) + Bs = J(x)^T y^+$ and $J(x)s = -c(x) \Leftrightarrow$

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

Sequential quadratic programming - SQP

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$_k$)

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

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

- Set $x^{k+1} = x^k + s^k$ and let $k := k + 1$. ◇

"convergence" verifies approximate KKT conditions for (eCP).

Sequential quadratic programming - SQP...

Advantanges of SQP:

- simple
- fast
  - quadratically convergent with $B^k = \nabla^2 \mathcal{L}(x^k, y^k)$
  - superlinearly convergent with good $B^k \approx \nabla^2 \mathcal{L}(x^k, y^k)$

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?

Sequential quadratic programming - SQP...

For the QP$_k$ subproblem:

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

to be well-defined, we need:

- constraints to be consistent
  - OK if $J(x^k)$ is full row rank (and so $m \leq n$).
- B^k to be positive definite when $J(x^k)s = 0 \iff N_k^T B^k N_k$ positive definite where the columns of N_k form a basis for the null space of $J(x^k) \Longrightarrow$

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

  is non-singular if $J(x^k)$ full row rank. [see next slide for explanations]

The QP_k subproblem

(QP_k) : $\min_{s\in\mathbb{R}^n} \underbrace{\nabla f(x^k)^T s + \frac{1}{2} s^T B^k s}_{:= m_k(s)}$ s.t. $J(x^k)s = -c(x^k)$

$\iff J(x^k)s + c(x^k) = 0.$

$(m \leq n)$

(QP_k) is bounded below on the set defined by the constraints if:

(a) B^k is positive definite on the null space of the constraints

Any point s that is feasible satisfies $J(x^k)s = -c(x^k)$. Take a (fixed) feasible point $\bar{s}$: $J(x^k)\bar{s} = -c(x^k)$ (exists if constraints are consistent). Then any feasible s can be expressed as: $s = \bar{s} + z$ where $z \in \underbrace{\ker(J(x^k))}_{\text{null space of constraints}}$.

Thus $\Omega = \{s : J(x^k)s = -c(x^k)\} = \bar{s} + \ker(J(x^k))$.

As $\bar{s}$ is fixed, if B^k is positive definite on $\ker(J(x^k))$, then $m_k(s)$ is ~~convex~~ strictly convex on $\ker(J(x^k))$ and hence on Ω and so (QP_k) is a well defined, convex programming, bounded below problem.

(b) $N_k^T B^k N_k$ positive definite where N_k is a basis of $\ker(J(x^k))$.

Let us show that (a) $\iff$ (b).

We have that $\ker(J(x^k)) = \{s : J(x^k)s = 0\} = \{N_k u : u \in \mathbb{R}^{|N_k|}\}$.

(a) $\iff s^T B^k s > 0 \ \forall s \in \ker(J(x^k)) \underset{s = N_k u}{\iff} u^T N_k^T B^k N_k u > 0 \ \forall u \in \mathbb{R}^{|N_k|}$.

$J(x^k)$ full row rank and (a) [or (b)]

$$\Rightarrow \begin{pmatrix} B^k & J(x^k)^T \\ J(x^k) & 0 \end{pmatrix} \text{ nonsingular.}$$

let $\begin{pmatrix} s \\ w \end{pmatrix}$ s.t. $\begin{pmatrix} B^k & J(x^k)^T \\ J(x^k) & 0 \end{pmatrix}\begin{pmatrix} s \\ w \end{pmatrix} = 0 \Leftrightarrow \begin{cases} B^k s + J(x^k)^T w = 0 \\ J(x^k) s = 0. \quad (+) \end{cases}$

$\Rightarrow s \in \ker(J(x^k))$

$$\Rightarrow (s^T \; w^T)\begin{pmatrix} B^k & J(x^k)^T \\ J(x^k) & 0 \end{pmatrix}\begin{pmatrix} s \\ w \end{pmatrix} = 0 \Leftrightarrow (s^T \; w^T)\begin{pmatrix} B^k s + J(x^k)^T w \\ J(x^k) s \end{pmatrix} = 0$$

$$\Leftrightarrow s^T B^k s + s^T J(x^k)^T w + w^T J(x^k) s = 0$$

$$\Leftrightarrow s^T B^k s + 2\, w^T \underbrace{J(x^k) s}_{=0 \text{ due to } (+)} = 0 \quad \Leftrightarrow s^T B^k s = 0 \underset{\substack{B^k \text{ pos. def} \\ \text{on } \ker(J(x^k))}}{\Rightarrow} \boxed{s = 0}$$

$$\underset{(+)}{\Rightarrow} J(x^k)^T w = 0 \underset{\substack{J(x^k) \\ \text{full row} \\ \text{rank}}}{\Rightarrow} \boxed{w = 0}$$

As the only vector in the null space of $\begin{pmatrix} B^k & J(x^k)^T \\ J(x^k) & 0 \end{pmatrix}$ is $\begin{pmatrix} s \\ w \end{pmatrix} = 0$,

this matrix must be nonsingular.

Linesearch SQP methods

$$s^k = \underset{s \in \mathbb{R}^n}{\text{arg min}} \ \nabla f(x^k)^T s + \tfrac{1}{2} s^T B^k s \ \text{ subject to } \ J(x^k) s = -c(x^k)$$

Linesearch SQP:

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

$$\Phi(x^k + \alpha^k s^k, \sigma^k) \text{``} <'' \Phi(x^k, \sigma^k)$$

  where $\Phi(x, \sigma)$ is a "suitable" merit function and σ^k are parameters.

Recall unconstrained GLM: crucial that s^k is a descent direction for $\Phi(x, \sigma^k)$ at x^k.

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 on the feasible set of QP_k, and that (s^k, y^{k+1}) are the SQP search direction and its associated Lagrange multiplier estimates 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

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

$$J(x^k) s^k = -c(x^k) \quad (2)$$

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

$$= -(s^k)^T B^k s^k - c(x^k)^T y^{k+1} \quad (3)$$

(2) $\implies$ $c(x^k)^T J(x^k) s^k = -\|c(x^k)\|^2.$ (4)

Suitable merit functions for SQP ...

Proof of Theorem 30: (continued) $\Phi(x) = f(x) + \frac{1}{2\sigma^k}\|c(x)\|^2$
Since x^k is not KKT, then either $c(x^k) \neq 0$ and so from (4), $s^k \neq 0$. Or there is no y such that $\nabla f(x^k) = J(x^k)^T y$; in particular, $\nabla f(x^k) \neq J(x^k)^T y^{k+1}$ and so from (1), $B^k s^k \neq 0$ and so $s^k \neq 0$ since B^k is positive definite in the feasible set $\{s : J(x^k)s = -c(x^k)\}$. It remains to show that $(s^k)^T \nabla_x \Phi(x^k, \sigma^k) < 0$. To see this, calculate

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

$$= -(s^k)^T B^k s^k - c(x^k)^T y^{k+1} - \frac{\|c(x^k)\|^2}{\sigma^k}, \quad \text{by (3), (4)}$$

$$< -c(x^k)^T y^{k+1} - \frac{\|c(x^k)\|^2}{\sigma^k}, \quad \text{since } (s^k)^T B^k s^k > 0$$

$$\leq \|c(x^k)\| \left(\|y^{k+1}\| - \frac{\|c(x^k)\|}{\sigma^k} \right), \quad \text{by Cauchy-Schwarz}$$

≤ 0 and so s^k is descent for $\Phi(x^k, \sigma^k)$. □

Suitable merit functions for SQP ...

Other suitable merit functions:

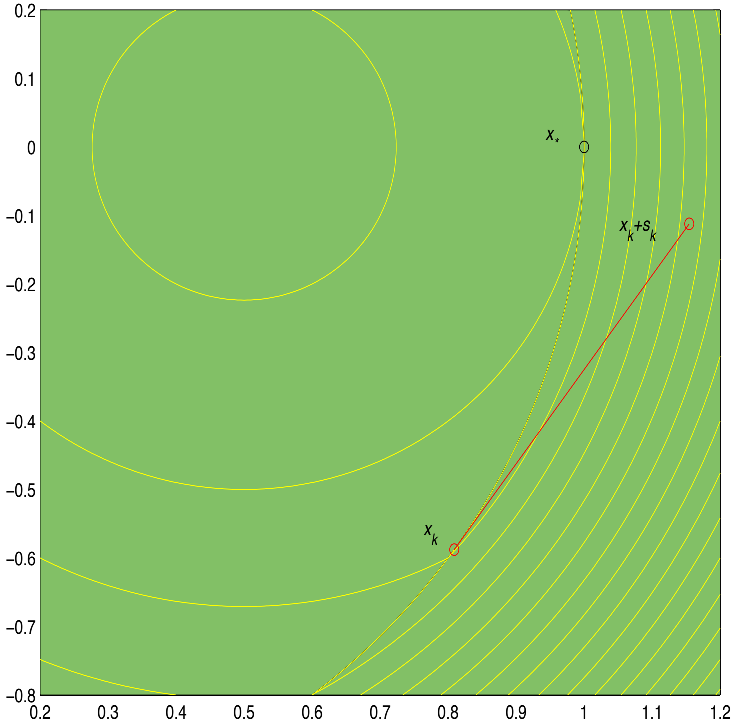
The non-differentiable exact penalty function: [widely used]

$\Psi(x, \rho) = f(x) + \rho \| c(x) \|$, where $\| \cdot \|$ can be any norm and $\rho > 0$.

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

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: ℓ_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 linearisation in the SQP model:

$$c(x^k + s^k) = O(\|s^k\|^2)$$

$\Longrightarrow$ need to correct for this curvature

Use a **second-order correction** from $x^k + s^k$:

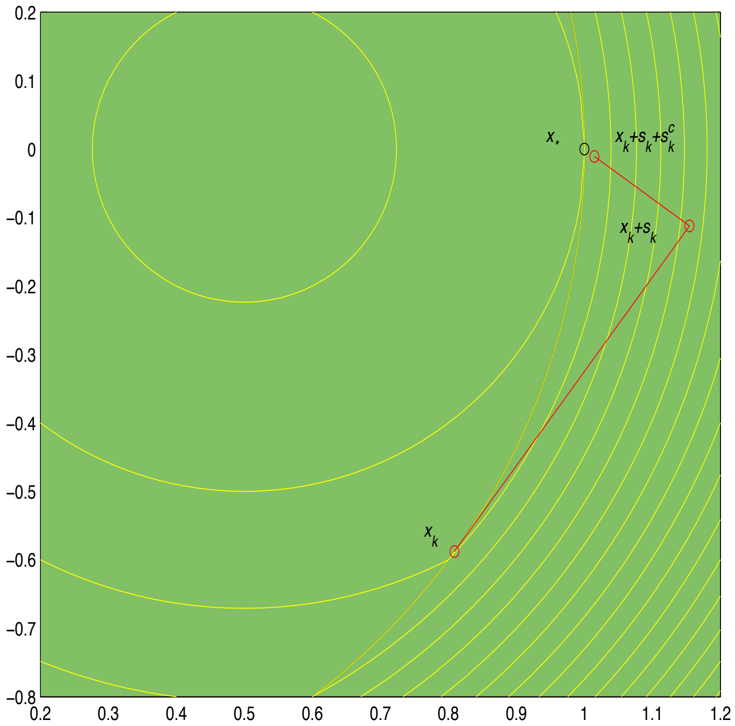
$$c(x^k + s^k + s_C^k) = o(\|s^k\|^2).$$

Also, do not want to destroy potential for fast convergence $\Longrightarrow$ $s_C^k = o(s^k)$.

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

or to $c(x^k + s^k) + J(x^k)s_C^k = 0$

Second-order corrections in action



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

- fast convergence
- $x^k + s^k + s_C^k$ reduces Φ $\implies$ global convergence

Topics not covered - for constrained optimization

- trust-region SQP methods
- the S$\ell_\mathbf{p}$QP method
- the filter-SQP approach
- active-set methods for linearly-constrained nonlinear problems (ie, generalization of simplex methods from the LP to the nonlinear case)