Lecture 12–13: Constrained optimization

Basic ideas

- A nonlinearly constrained problem must somehow be converted—relaxed—into a problem which we can solve (a linear/quadratic or unconstrained problem)
- We solve a sequence of such problems
- To make sure that we tend towards a solution to the original problem, we must impose properties of the original problem more and more
- How is this done?
- In simpler problem like linearly constrained ones, a line search in f is enough

- For more general problems, where (normally) the constraints are manipulated, this is not enough
- We can include *penalty functions* for constraints that we relax
- We can produce estimates of the Lagrange multipliers and invoke them
- We will look at both types of approaches
- Additional techniques:
  - *Sequential Linear Programming* (SLP)—based on linear approximations of *all* functions
  - *Augmented Lagrangian methods*—combines Lagrangian *and* penalty terms

Penalty functions

- Consider the optimization problem to

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}), \\ &\text{subject to } \boldsymbol{x} \in S, \end{aligned} \tag{1}$$

  where $S \subset \mathbb{R}^n$ is non-empty, closed, and $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable

- Basic idea behind all penalty methods: to replace the problem (1) with the equivalent unconstrained one:

$$\text{minimize } f(\boldsymbol{x}) + \chi_S(\boldsymbol{x}),$$

where

$$\chi_S(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x} \in S, \\ +\infty, & \text{otherwise} \end{cases}$$

is the *indicator function* of the set S

- Feasibility is top priority; only when achieving feasibility can we concentrate on minimizing f
- Computationally bad: non-differentiable, discontinuous, and even not finite (though it is convex provided S is convex). Better: numerical “warning” before becoming infeasible or near-infeasible
- Replace the indicator function with a numerically better behaving function

Exterior penalty methods

- SUMT—Sequential Unconstrained Minimization Techniques—were devised in the late 1960s by Fiacco and McCormick. They are still among the more popular ones for some classes of problems, although there are later modifications that are more often used
- Suppose

$$\begin{aligned} S = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid g_i(\boldsymbol{x}) &\leq 0, \quad i = 1, \ldots, m, \\ h_j(\boldsymbol{x}) &= 0, \quad j = 1, \ldots, \ell \,\}, \end{aligned}$$

$g_i \in C(\mathbb{R}^n)$, $i = 1, \ldots, m$, $h_j \in C(\mathbb{R}^n)$, $j = 1, \ldots, \ell$

- Choose a C^0 function $\psi : \mathbb{R} \to \mathbb{R}_+$ such that $\psi(s) = 0$ if and only if $s = 0$ [typical examples of $\psi(\cdot)$ will be $\psi_1(s) = |s|$, or $\psi_2(s) = s^2$]. Approximation to χ_S:

$$\nu \check{\chi}_S(\boldsymbol{x}) := \nu \left(\sum_{i=1}^{m} \psi\big(\max\{0, g_i(\boldsymbol{x})\}\big) + \sum_{j=1}^{\ell} \psi\big(h_j(\boldsymbol{x})\big) \right)$$

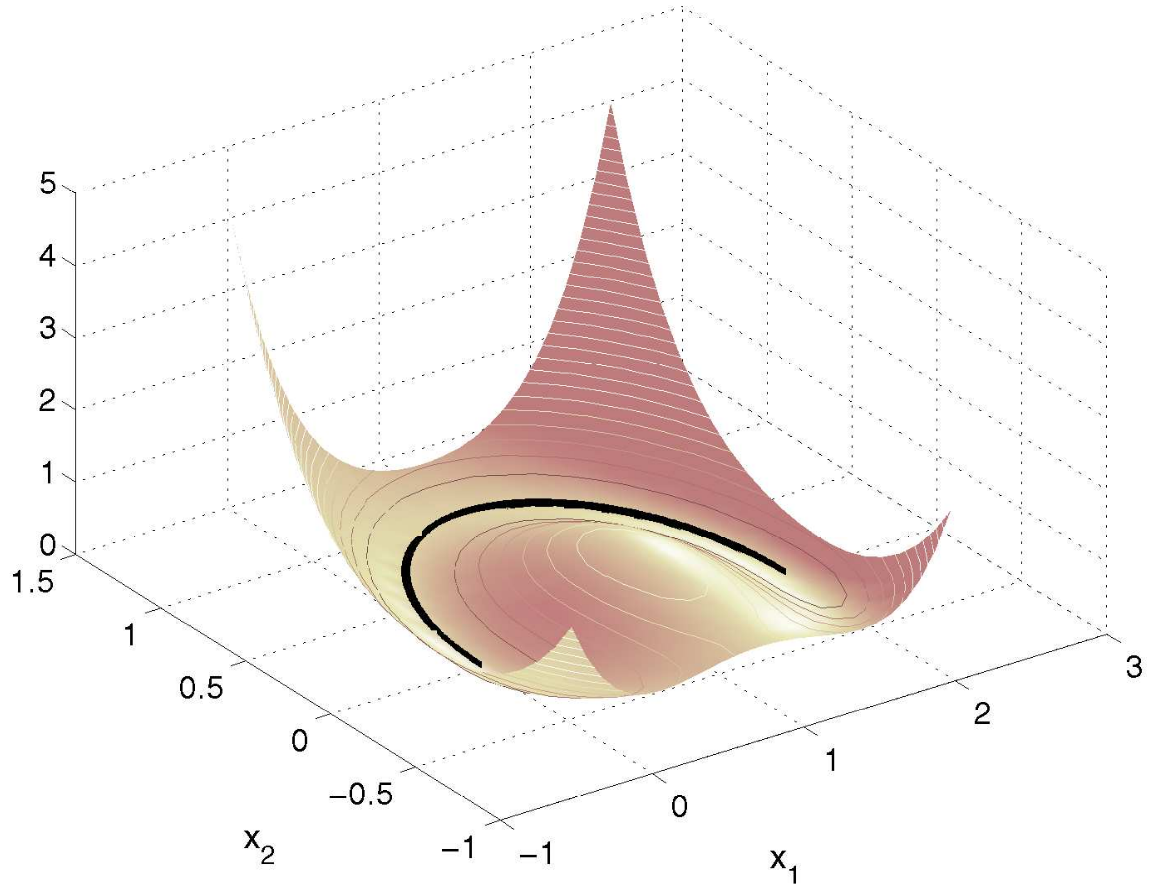
- $\nu > 0$ is a *penalty parameter*
- Different treatment of inequality/equality constraints since an equality constraint is violated whenever $h_j(\boldsymbol{x}) \neq 0$, while an inequality constraint is violated only when $g_i(\boldsymbol{x}) > 0$; equivalent to $\max\{0, g_i(\boldsymbol{x})\} \neq 0$
- $\check{\chi}_S$ approximates χ_S from *below* ($\check{\chi}_S \leq \chi_S$)

Example

- Let $S = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid -x_2 \leq 0, (x_1 - 1)^2 + x_2^2 = 1 \,\}$
- Let $\psi(s) = s^2$. Then,

$$\check{\chi}_S(\boldsymbol{x}) = [\max\{0, -x_2\}]^2 + [(x_1 - 1)^2 + x_2^2 - 1]^2$$

- Graph of $\check{\chi}_S$ and S:


5
4
3
2
1
0
1.5
1
0.5
0
−0.5
−1
x_2
−1
0
1
2
3
x_1

Properties of the penalty problem

- We assume the problem (1) has an optimal solution $\boldsymbol{x}^*$
- We assume that for every $\nu > 0$ the problem to

$$\operatorname*{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x}) + \nu\check{\chi}_S(\boldsymbol{x}) \tag{2}$$

  has at least one optimal solution $\boldsymbol{x}^*_\nu$

- $\check{\chi}_S \geq 0$; $\check{\chi}_S(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x} \in S$
- The Relaxation Theorem 7.1 states that the inequality $f(\boldsymbol{x}^*_\nu) + \nu\check{\chi}(\boldsymbol{x}^*_\nu) \leq f(\boldsymbol{x}^*) + \chi_S(\boldsymbol{x}^*) = f(\boldsymbol{x}^*)$ holds for every positive ν. (Lower bound on the optimal value.)
- The problem (2) is convex if (1) is

The algorithm and its convergence properties

- *Assume that the problem* (1) *possesses optimal solutions. Then, as* $\nu \to +\infty$ *every limit point of the sequence* $\{\boldsymbol{x}_\nu^*\}$ *of globally optimal solutions to* (2) *is globally optimal in the problem* (1)
- Of interest for convex problems. What about general problems?

- *Let* f, g_i $(i = 1, \ldots, m)$, *and* h_j $(j = 1, \ldots, \ell)$, *be in* C^1
- *Assume that the penalty function* ψ *is in* C^1 *and that* $\psi'(s) \geq 0$ *for all* $s \geq 0$
- *Then:*

$$\left.\begin{array}{r} \boldsymbol{x}_k \text{ stationary in } (2) \\ \boldsymbol{x}_k \to \hat{\boldsymbol{x}} \text{ as } k \to +\infty \\ \text{LICQ holds at } \hat{\boldsymbol{x}} \\ \hat{\boldsymbol{x}} \text{ feasible in } (1) \end{array}\right\} \Longrightarrow \hat{\boldsymbol{x}} \text{ stationary (KKT) in } (1)$$

- From the proof we can obtain estimates of Lagrange multipliers: the optimality conditions of (2) gives that

$$\mu_i^* \approx \nu_k \psi'[\max\{0, g_i(\boldsymbol{x}_k)\}] \quad \text{and} \quad \lambda_j^* \approx \nu_k \psi'[h_j(\boldsymbol{x}_k)]$$

Interior penalty methods

- In contrast to exterior methods, interior penalty, or *barrier*, function methods construct approximations *inside* the set S and set a barrier against leaving it
- If a globally optimal solution to (1) is on the boundary of the feasible region, the method generates a sequence of interior points that converge to it
- We assume that the feasible set has the following form:

$$S = \{\, \boldsymbol{x} \in \mathbb{R}^n \mid g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m \,\}$$

- We need to assume that there exists a *strictly feasible* point $\hat{\boldsymbol{x}} \in \mathbb{R}^n$, i.e., such that $g_i(\hat{\boldsymbol{x}}) < 0$, $i = 1, \ldots, m$

- Approximation of χ_S (from *above*, that is, $\hat{\chi}_S \geq \chi_S$):

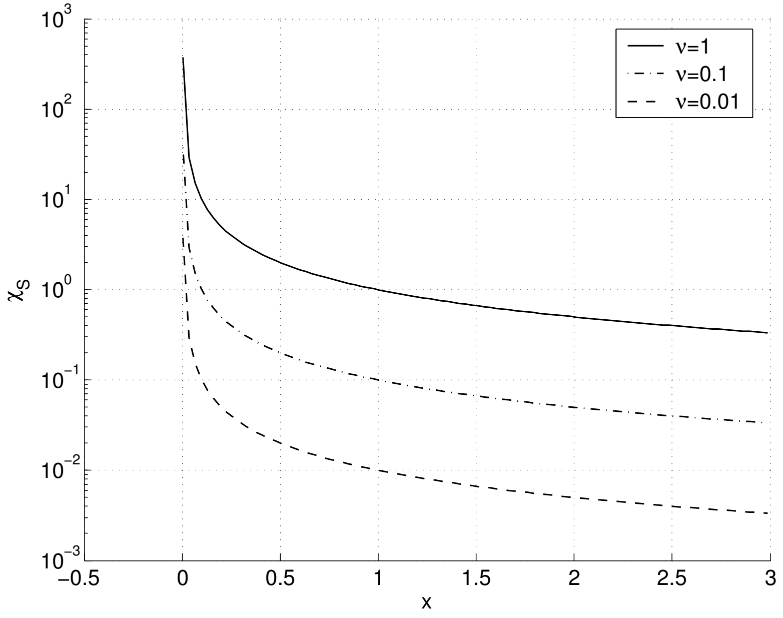
$$\nu\hat{\chi}_S(\boldsymbol{x}) := \begin{cases} \nu \sum_{i=1}^{m} \phi[g_i(\boldsymbol{x})], & \text{if } g_i(\boldsymbol{x}) < 0,\ i = 1, \ldots, m, \\ +\infty, & \text{otherwise,} \end{cases}$$

where $\phi : \mathbb{R}_- \to \mathbb{R}_+$ is a continuous, non-negative function such that $\phi(s_k) \to \infty$ for all *negative* sequences $\{s_k\}$ converging to zero

- Examples: $\phi_1(s) = -s^{-1}$; $\phi_2(s) = -\log[\min\{1, -s\}]$
- The differentiable *logarithmic barrier function* $\widetilde{\phi}_2(s) = -\log(-s)$ gives rise to the same convergence theory, if we drop the non-negativity requirement on ϕ
- Barrier function convex if (1) is

Example

- Consider $S = \{\, x \in \mathbb{R} \mid -x \leq 0 \,\}$. Choose $\phi = \phi_1 = -s^{-1}$. Graph of the barrier function $\nu\hat{\chi}_S$ in below figure for various values of ν (note how $\nu\hat{\chi}_S$ converges to χ_S as $\nu \downarrow 0$!):



Algorithm and its convergence

- Penalty problem:

$$\text{minimize } f(\boldsymbol{x}) + \nu \hat{\chi}_S(\boldsymbol{x}) \tag{3}$$

- Convergence of global solutions to (3) to globally optimal solutions to (1) straightforward. Result for stationary (KKT) points more practical:

- *Let f and g_i $(i = 1, \ldots, m)$, an ϕ be in C^1, and that $\phi'(s) \geq 0$ for all $s < 0$*
- *Then:*

$$\left.\begin{array}{r} \boldsymbol{x}_k \textit{ stationary in } (3) \\ \boldsymbol{x}_k \to \hat{\boldsymbol{x}} \textit{ as } k \to +\infty \\ \textit{LICQ holds at } \hat{\boldsymbol{x}} \end{array}\right\} \Longrightarrow \hat{\boldsymbol{x}} \textit{ stationary (KKT) in } (1)$$

- If we use $\phi(s) = \phi_1(s) = -1/s$, then $\phi'(s) = 1/s^2$, and the sequence $\{\nu_k/g_i^2(\boldsymbol{x}_k)\}$ converges towards the Lagrange multiplier $\hat{\mu}_i$ corresponding to the constraint i $(i = 1, \ldots, m)$

Interior point (polynomial) method for LP

- Consider the dual LP to

$$\begin{aligned} &\text{maximize } \boldsymbol{b}^{\mathrm{T}}\boldsymbol{y}, \\ &\text{subject to } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{s} = \boldsymbol{c}, \\ &\qquad\qquad\qquad \boldsymbol{s} \geq \mathbf{0}^n, \end{aligned} \tag{4}$$

  and the corresponding system of optimality conditions:

$$\begin{aligned} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{s} &= \boldsymbol{c}, \\ \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{b}, \\ \boldsymbol{x} \geq \mathbf{0}^n, \ \boldsymbol{s} \geq \mathbf{0}^n, \ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{s} &= 0 \end{aligned}$$

- Apply a barrier method for (4). Subproblem:

$$\begin{aligned} &\text{minimize } -\boldsymbol{b}^{\mathrm{T}}\boldsymbol{y} - \nu \sum_{j=1}^{n} \log(s_j) \\ &\text{subject to } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{s} = \boldsymbol{c} \end{aligned}$$

- The KKT conditions for this problem is:

$$\begin{aligned} \boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{s} &= \boldsymbol{c}, \\ \boldsymbol{A}\boldsymbol{x} &= \boldsymbol{b}, \\ x_j s_j &= \nu, \quad j = 1, \ldots, n \end{aligned} \tag{5}$$

- Perturbation in the complementary conditions!

- Using a Newton method for the system (5) yields a very effective LP method. If the system is solved exactly we trace the *central path* to an optimal solution, but *polynomial* algorithms are generally implemented such that only one Newton step is taken for each value of ν_k before it is reduced
- A polynomial algorithm finds, in theory at least (disregarding the finite precision of computer arithmetic), an optimal solution within a number of floating-point operations that are polynomial in the data of the problem

Sequential quadratic programming (SQP) methods: A first image

- We study the equality constrained problem to

$$\text{minimize } f(\boldsymbol{x}), \tag{6a}$$

$$\text{subject to } h_j(\boldsymbol{x}) = 0, \qquad j = 1, \ldots, \ell, \tag{6b}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h_j : \mathbb{R}^n \to \mathbb{R}$ are in C^1 on $\mathbb{R}^n$

- The KKT conditions state that at a local minimum $\boldsymbol{x}^*$ of f over the feasible set, where $\boldsymbol{x}^*$ satisfies some CQ, there exists a vector $\boldsymbol{\lambda}^* \in \mathbb{R}^\ell$ with

$$\begin{aligned}\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) &:= \nabla f(\boldsymbol{x}^*) + \sum_{j=1}^{\ell} \lambda_j^* \nabla h_j(\boldsymbol{x}^*) = \mathbf{0}^n, \\ \nabla_{\boldsymbol{\lambda}} L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) &:= \boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0}^\ell\end{aligned}$$

- Appealing to find a KKT point by directly attacking this system of nonlinear equations, which has $n + \ell$ unknowns as well as equations

- Newton's method! So suppose that f and h_j $(j = 1, \ldots, \ell)$ are in C^2 on $\mathbb{R}^n$. Suppose we have an iteration point $(\boldsymbol{x}_k, \boldsymbol{\lambda}_k) \in \mathbb{R}^n \times \mathbb{R}^\ell$
- Next iterate $(\boldsymbol{x}_{k+1}, \boldsymbol{\lambda}_{k+1})$: $(\boldsymbol{x}_{k+1}, \boldsymbol{\lambda}_{k+1}) = (\boldsymbol{x}_k, \boldsymbol{\lambda}_k) + (\boldsymbol{p}_k, \boldsymbol{v}_k)$, where $(\boldsymbol{p}_k, \boldsymbol{v}_k) \in \mathbb{R}^n \times \mathbb{R}^\ell$ solves the second-order approximation of the stationary point condition for the Lagrange function:

$$\nabla^2 L(\boldsymbol{x}_k, \boldsymbol{\lambda}_k) \begin{pmatrix} \boldsymbol{p}_k \\ \boldsymbol{v}_k \end{pmatrix} = -\nabla L(\boldsymbol{x}_k, \boldsymbol{\lambda}_k),$$

that is,

$$\begin{bmatrix} \nabla^2_{\boldsymbol{x}\boldsymbol{x}} L(\boldsymbol{x}_k, \boldsymbol{\lambda}_k) & \boldsymbol{h}(\boldsymbol{x}_k) \\ \boldsymbol{h}(\boldsymbol{x}_k)^{\mathrm{T}} & \boldsymbol{0}^{m \times m} \end{bmatrix} \begin{pmatrix} \boldsymbol{p}_k \\ \boldsymbol{v}_k \end{pmatrix} = \begin{pmatrix} -\nabla_{\boldsymbol{x}} L(\boldsymbol{x}_k, \boldsymbol{\lambda}_k) \\ -\boldsymbol{h}(\boldsymbol{x}_k) \end{pmatrix} \qquad (7)$$

- Interpretation: the KKT system for the QP problem to

$$\underset{\boldsymbol{p}}{\text{minimize}}\ \frac{1}{2}\boldsymbol{p}^{\mathrm{T}} \nabla^2_{\boldsymbol{x}\boldsymbol{x}} L(\boldsymbol{x}_k, \boldsymbol{\lambda}_k)\boldsymbol{p} + \nabla_{\boldsymbol{x}} L(\boldsymbol{x}_k, \boldsymbol{\lambda}_k)\boldsymbol{p}, \qquad (8a)$$

$$\text{subject to } h_j(\boldsymbol{x}_k) + \nabla h_j(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p} = 0, \qquad j = 1, \ldots, \ell \qquad (8b)$$

- Objective: second-order approximation of the Lagrange function with respect to $\boldsymbol{x}$. Constraints: first-order approximations at $\boldsymbol{x}_k$. The vector $\boldsymbol{v}_k$ appearing in (7) is the vector of Lagrange multipliers for the constraints (8b)
- Unsatisfactory: (a) Convergence is only *local*. (b) The algorithm requires strong assumptions about the problem

An exact penalty function

- Consider (1), and

$$\begin{aligned} \check{\chi}_S(\boldsymbol{x}) &:= \sum_{i=1}^{m} \text{maximum}\,\{0, g_i(\boldsymbol{x})\} + \sum_{j=1}^{\ell} |h_j(\boldsymbol{x})|, \\ P_e(\boldsymbol{x}) &:= f(\boldsymbol{x}) + \nu \check{\chi}_S(\boldsymbol{x}) \end{aligned}$$

- *Suppose* $\boldsymbol{x}^*$ *is a KKT point for* (1)*, with Lagrange multipliers* $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$*, and* (1) *is a convex problem*
- *Then: if the value of* ν *is large enough such that*

$$\nu \geq \text{maximum}\{\mu_i^*,\ i \in \mathcal{I}(x^*); \quad |\lambda_j^*|,\ j = 1, \dots, \ell\,\}$$

then the vector $\boldsymbol{x}^*$ *is also a global minimum of the function* P_e

Basic SQP method

- Given $\boldsymbol{x}_k \in \mathbb{R}^n$ and a vector $(\boldsymbol{\mu}_k, \boldsymbol{\lambda}_k) \in \mathbb{R}^m_+ \times \mathbb{R}^\ell$, choose a positive definite, symmetric matrix $\boldsymbol{B}_k \in \mathbb{R}^n \times n$
- Solve

$$\begin{aligned} \underset{\boldsymbol{p}}{\text{minimize}} \quad & \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\boldsymbol{B}_k\boldsymbol{p} + \nabla f(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p}, & \text{(9a)} \\ \text{subject to} \quad & g_i(\boldsymbol{x}_k) + \nabla g_i(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p} \leq 0, \ i = 1, \dots, m, & \text{(9b)} \\ & h_j(\boldsymbol{x}_k) + \nabla h_j(\boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{p} = 0, \ j = 1, \dots, \ell & \text{(9c)} \end{aligned}$$

- If $\boldsymbol{B}_k \approx \nabla^2_{\boldsymbol{xx}} L(\boldsymbol{x}_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)$ then (9) is a 2^{nd} order approximation of KKT (cf. quasi-Newton!)

Convergence

- *Given* $\boldsymbol{x}_k \in \mathbb{R}^n$ *consider* (9), *where* $\boldsymbol{B}_k \in \mathbb{R}^{n\times n}$ *is positive definite. Suppose* $\boldsymbol{p}_k$ *solves* (9) *together with multipliers* $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$*. Assume that* $\boldsymbol{p}_k \neq \boldsymbol{0}^n$*. Then, if* $\nu \geq \text{maximum}\,\{\mu_1, \ldots, \mu_m, |\lambda_1|, \ldots, |\lambda_\ell|\}$ *the vector* $\boldsymbol{p}_k$ *is a direction of descent with respect to the exact penalty function* P_e *at* $(\boldsymbol{x}_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)$

- *The SQP method either terminates finitely at a KKT point or it produces an infinite sequence* $\{\boldsymbol{x}_k\}$*. In the latter case, assume that* $\{\boldsymbol{x}_k\}$ *lies in a compact set* $X \subset \mathbb{R}^n$ *and that for every* $\boldsymbol{x} \in X$ *and positive definite matrix* $\boldsymbol{B}_k$ *the QP* (9) *has a unique solution, and also unique multiplier vectors* $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$ *satisfying* $\nu \geq \text{maximum}\{\mu_1, \ldots, \mu_m, |\lambda_1|, \ldots, |\lambda_\ell|\}$*, where* $\nu > 0$ *is the penalty parameter. Assume that* $\{\boldsymbol{B}_k\}$ *is bounded and that every accumulation point of this sequence is positive definite (or, the sequence* $\{\boldsymbol{B}_k^{-1}\}$ *of matrices is bounded). Then, every accumulation point of* $\{\boldsymbol{x}_k\}$ *is a KKT point*

Remarks

- Selecting the value of ν is difficult
- No guarantees that the subproblems (9) are feasible; we *assumed* above that the problem is well-defined
- P_e is only continuous; some step length rules infeasible
- Fast convergence not guaranteed (the *Maratos effect*)
- Penalty methods in general suffer from ill-conditioning. For some problems the ill-conditioning is avoided
- Exact penalty SQP methods suffer less from ill-conditioning, and the number of QP:s needed can be small. They can, however, cost a lot computationally
- `fmincon` in MATLAB is an SQP-based solver

Filter-SQP

- Popular development: algorithms where the penalty parameter is avoided altogether—*filter-SQP methods*
- *Multi-objective optimization*: $\boldsymbol{x}^1$ *dominates* $\boldsymbol{x}^2$ if $\check{\chi}(\boldsymbol{x}^1) \leq \check{\chi}(\boldsymbol{x}^2)$ and $f(\boldsymbol{x}^1) \leq f(\boldsymbol{x}^2)$
- (if $\boldsymbol{x}^1$ is better in terms of feasibility *and* optimality)
- *Filter:* a list of pairs $(\check{\chi}_i, f_i)$ such that $\check{\chi}_i < \check{\chi}_j$ or $f_i < f_j$ for all $j \neq i$ in the list
- Its elements build up an *efficient frontier* in the bi-criterion problem
- Filter used in place of the penalty function, when the standard Newton-like step cannot be computed