Lecture 13–14: Nonlinearly constrained optimization

Michael Patriksson

2, 4 March 2004

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 the constraints are normally 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.

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

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

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 a convex set). Need to be numerically "warned" about being 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

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

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

- Choose a 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:

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

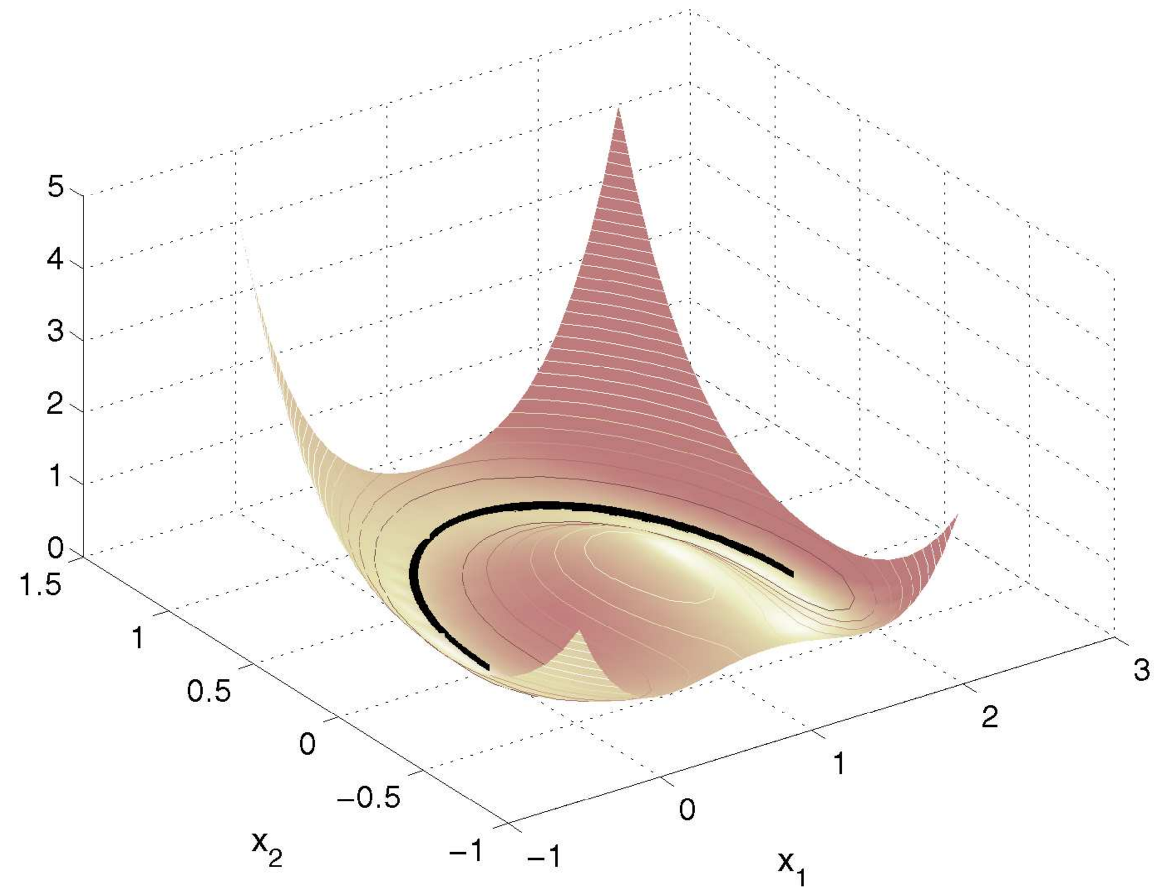
- $\nu > 0$ is a *penalty parameter.*
- Different treatment of inequality/equality constraints since an equality constraint is violated whenever $h_j \neq 0$, while an inequality constraint is violated only when $g_i > 0$; equivalent to $\max\{0, g_i(\boldsymbol{x})\} \neq 0$.

Example

- Let $S = \{\, \boldsymbol{x} \in \mathbb{R}^2 \mid -x_2 \le 0, (x_1 - 1)^2 + x_2^2 = 1 \,\}$.
- 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
x2
−1
0
1
2
3
x1

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

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

  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 algorithm and its convergence properties

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

- Theorem 13.4: *Let f, g_i $(i = 1, \dots, m)$, and h_j $(j = 1, \dots, \ell)$, be continuously differentiable. Further assume that the penalty function ψ is continuously differentiable and that $\psi'(s) \geq 0$ for all $s \geq 0$. Consider a sequence $\{\boldsymbol{x}_k\}$ of stationary points in (2), corresponding to a positive sequence of penalty parameters $\{\nu_k\}$ converging to $+\infty$. Assume that $\lim_{k \to +\infty} \boldsymbol{x}_k = \hat{\boldsymbol{x}}$, and that the LICQ holds at $\hat{\boldsymbol{x}}$. Then, $\hat{\boldsymbol{x}}$ is a KKT-point for (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:

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

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

- Approximation of χ_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 famous 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.

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 Figure 1 for various values of ν (note how $\nu\hat{\chi}_S$ converges to χ_S as $\nu \downarrow 0$!):

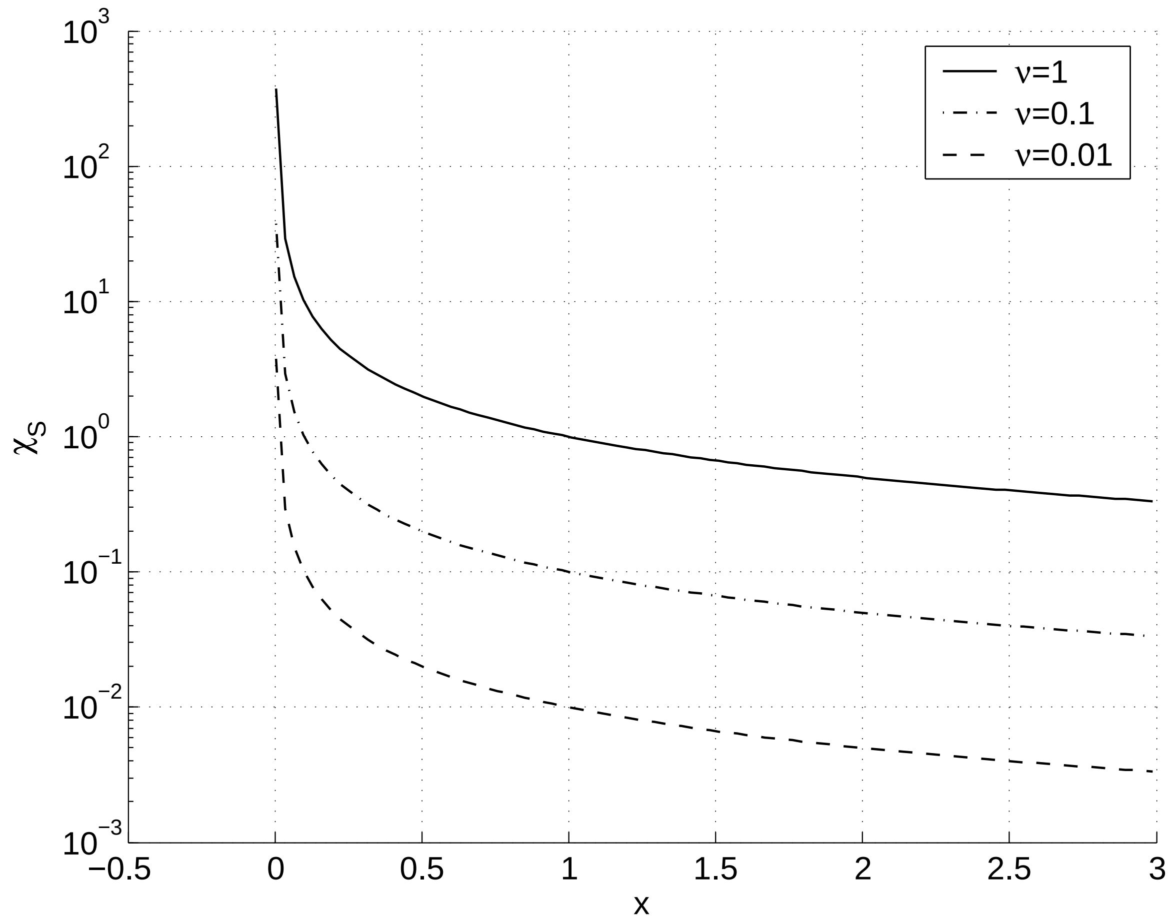


Figure 1: The graph of $\nu\hat{\chi}_S$ for various choices of ν. Note the logarithmic scale.

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:

- Theorem 13.6: *Let f and g_i $(i = 1, \ldots, m)$, be continuously differentiable. Further assume that the barrier function ϕ is continuously differentiable and that $\phi'(s) \geq 0$ for all $s < 0$.*

  Consider a sequence $\{\boldsymbol{x}_k\}$ of stationary points in (3) corresponding to a positive sequence of penalty parameters $\{\nu_k\}$ converging to 0. Assume that $\lim_{k \to +\infty} \boldsymbol{x}_k = \hat{\boldsymbol{x}}$, and that the LICQ holds at $\hat{\boldsymbol{x}}$. Then, $\hat{\boldsymbol{x}}$ is a KKT-point for (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

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

and the corresponding system of optimality conditions:

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

- 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:

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

- 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 $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $h_j : \mathbb{R}^n \mapsto \mathbb{R}$ 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,

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

- Interpretation: the KKT system for the QP problem to

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

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

- 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.

A penalty function based SQP algorithm

- New problem:

$$\text{minimize } f(\boldsymbol{x}), \tag{9a}$$

$$\text{subject to } g_i(\boldsymbol{x}) \le 0, \qquad i = 1, \ldots, m, \tag{9b}$$

- Penalty function:

$$P(\boldsymbol{x}) = \text{maximum}\{0, g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x})\}.$$

- Interesting connection between penalty function and solutions to (9): Proposition 13.10: *Suppose that* $\boldsymbol{x}^*$ *is a local minimum of* f *over the feasible set of the problem* (9), *which satisfies the linear independence CQ (LICQ) and together with Lagrange multipliers* $\boldsymbol{\mu}^*$ *satisfies the KKT conditions as well as a second-order sufficiency condition. Then, if the value of* c *is large enough such that*

$$c > \sum_{i=1}^{m} \mu_i^*,$$

then the vector $\boldsymbol{x}^*$ *is a strict local minimum of the function* $f + cP$. ■

- SQP methods are based on a combination of a method for minimizing $f + cP$ for some parameter $c > 0$ and a method for updating c in order to try to achieve the (unknown) threshold value stated in the Proposition. In the convex case, the result will be a globally optimal solution; in other cases, a KKT point.

- Let c be given. Solve the problem to

$$\underset{(\boldsymbol{p},\xi)}{\operatorname{minimize}}\ \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\boldsymbol{H}_k\boldsymbol{p} + c\xi, \tag{10a}$$

$$\text{subject to } g_i(\boldsymbol{x}) + \nabla g_i(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p} \leq \xi, \qquad i = 1, \dots, m, \tag{10b}$$

  where $\boldsymbol{H}_k \in \mathbb{R}^n \times n$ is symmetric, positive definite.

- The resulting search direction $\boldsymbol{p}_k$ is a direction of descent for $f + cP$ at $\boldsymbol{x}_k$ (Proposition 13.11). We then let $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k$, where α_k is determined by an exact line search or the Armijo rule.

- Occasionally, we fix ξ to zero in (10); this is a more exact representation of the second-order approximation of the original problem. If it has no solution—go back to the real problem(10). If it has a solution, then we update the value of c as follows:

$$c := \text{maximum}\left\{ c, \sum_{i=1}^{m} \mu_i + \beta \right\},$$

  where μ_i is the Lagrange multiplier value for the constraint i in the problem (10), and β is some positive scalar.

- The solver `fmincon` is an SQP method.

Numerical considerations

- Ill-conditioning: Penalty methods in general suffer from ill-conditioning. For some problems, like LP, the ill-conditioning is avoided thanks to the special structure of LP.
- 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.