Lecture 13: Constrained optimization

Michael Patriksson

2010-12-03

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

Penalty functions, I

- Consider the optimization problem to

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

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:

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

where

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

is the *indicator function* of the set S

Penalty functions, II

- 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, I

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

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

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

- 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:

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

Exterior penalty methods, II

- $\nu > 0$ is a *penalty parameter*

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

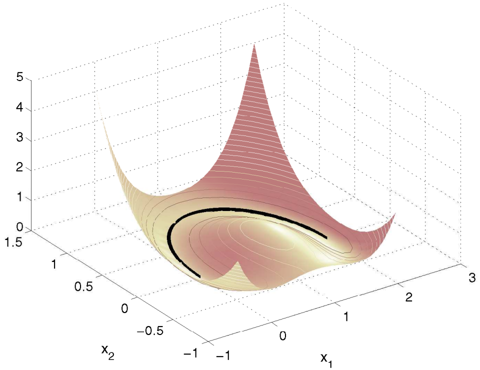
- $\check{\chi}_S$ approximates χ_S from *below* ($\check{\chi}_S \leq \chi_S$)

Example

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

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

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



Properties of the penalty problem

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

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

  has at least one optimal solution $\mathbf{x}_\nu^*$
- $\check{\chi}_S \geq 0$; $\check{\chi}_S(\mathbf{x}) = 0$ if and only if $\mathbf{x} \in S$
- The Relaxation Theorem states that the inequality $f(\mathbf{x}_\nu^*) + \nu \check{\chi}(\mathbf{x}_\nu^*) \leq f(\mathbf{x}^*) + \chi_S(\mathbf{x}^*) = f(\mathbf{x}^*)$ 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, I

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

The algorithm and its convergence properties, II

- *Let f, g_i $(i = 1, \dots, m)$, and h_j $(j = 1, \dots, \ell)$, 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:*

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

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

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

Interior penalty methods, I

- 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:

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

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

Interior penalty methods, II

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

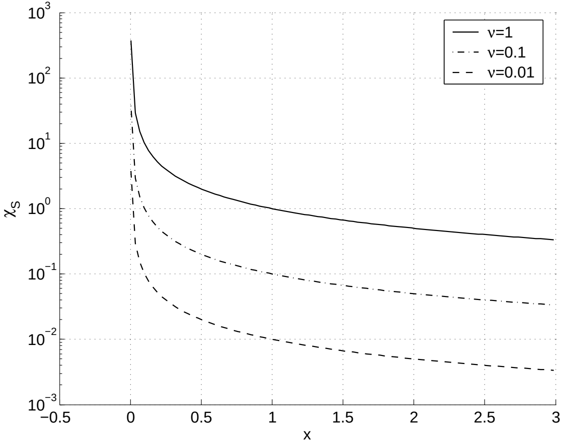
$$\nu\hat{\chi}_S(\mathbf{x}) := \begin{cases} \nu \sum_{i=1}^{m} \phi[g_i(\mathbf{x})], & \text{if } g_i(\mathbf{x}) < 0, \ i = 1, \dots, m, \\ +\infty, & \text{otherwise,} \end{cases}$$

  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:

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

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

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

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

Interior point (polynomial) method for LP, I

- Consider the dual LP to

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

and the corresponding system of optimality conditions:

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

Interior point (polynomial) method for LP, II

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

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

- The KKT conditions for this problem is:

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

- Perturbation in the complementary conditions!

Interior point (polynomial) method for LP, II

- 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, I

- We study the equality constrained problem to

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

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

  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 $\mathbf{x}^*$ of f over the feasible set, where $\mathbf{x}^*$ satisfies some CQ, there exists a vector $\boldsymbol{\lambda}^* \in \mathbb{R}^\ell$ with

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

$$\nabla_{\boldsymbol{\lambda}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) := \mathbf{h}(\mathbf{x}^*) = \mathbf{0}^\ell$$

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

Sequential quadratic programming (SQP) methods, II

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

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

that is,

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

- Interpretation: the KKT system for the QP problem to

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

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

Sequential quadratic programming (SQP) methods, III

- Objective: second-order approximation of the Lagrange function with respect to $\mathbf{x}$. Constraints: first-order approximations at $\mathbf{x}_k$. The vector $\mathbf{v}_k$ appearing in (7) is the vector of Lagrange multipliers for the constraints (8b)

- Unsatisfactory:
  1. Convergence is only *local*
  2. The algorithm requires strong assumptions about the problem

Introducing an exact penalty function

- Consider (1), and

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

- *Suppose* $\mathbf{x}^*$ *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*

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

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

Basic SQP method

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

- Solve

$$\underset{\mathbf{p}}{\text{minimize}} \ \frac{1}{2}\mathbf{p}^{\mathrm{T}}\mathbf{B}_k\mathbf{p} + \nabla f(\mathbf{x}_k)^{\mathrm{T}}\mathbf{p}, \tag{9a}$$

$$\text{subject to} \ g_i(\mathbf{x}_k) + \nabla g_i(\mathbf{x}_k)^{\mathrm{T}}\mathbf{p} \leq 0, \ i = 1, \ldots, m, \tag{9b}$$

$$h_j(\mathbf{x}_k) + \nabla h_j(\mathbf{x}_k)^{\mathrm{T}}\mathbf{p} = 0, \ j = 1, \ldots, \ell \tag{9c}$$

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

*Convergence

- One can show that if the penalty value is large enough, then the vector $\mathbf{p}_k$ from (9) is a direction of descent with respect to the exact penalty function P_e at $(\mathbf{x}_k, \boldsymbol{\mu}_k, \boldsymbol{\lambda}_k)$

- Convergence of the SQP method towards KKT points can then be established under additional conditions on the choices of matrices $\{\mathbf{B}_k\}$

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*: $\mathbf{x}^1$ *dominates* $\mathbf{x}^2$ if $\check{\chi}(\mathbf{x}^1) \leq \check{\chi}(\mathbf{x}^2)$ and $f(\mathbf{x}^1) \leq f(\mathbf{x}^2)$ (if $\mathbf{x}^1$ 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