Lecture 14

Constrained optimization

Kin Cheong Sou
Department of Mathematical Sciences
Chalmers University of Technology and Göteborg University
December 16, 2014

CHALMERS | GÖTEBORGS UNIVERSITET

- Consider the optimization problem to

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

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:

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

where

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

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
- Approximate the indicator function with a numerically better behaving function

- 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

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

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

- 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:

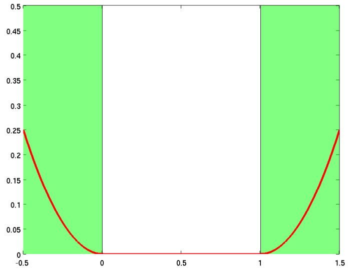
$$\nu \check{\chi}_S(x) := \nu \left(\sum_{i=1}^{m} \psi\big(\max\{0, g_i(x)\}\big) + \sum_{j=1}^{\ell} \psi\big(h_j(x)\big) \right)$$

- $S = \{x \mid -x \leq 0, x \leq 1\}$

- Indicator function

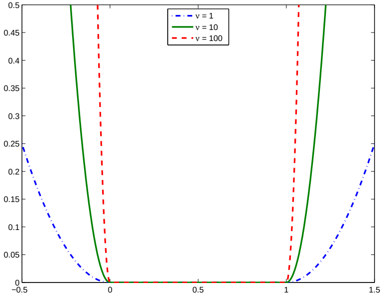
$$\chi_S(x) = \begin{cases} 0 & \text{if } 0 \leq x \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

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



  - Penalty function $\psi(s) = s^2$
  - Approximate function (i.e. substitute for indicator function)

$$\nu\check{\chi}_S = \nu\Big((\max\{0, x-1\})^2 + (\max\{0, -x\})^2\Big)$$



- $\nu > 0$ is *penalty parameter*
- $\nu \check{\chi}_S(x) \to \chi_S(x)$ as $\nu \to \infty$.

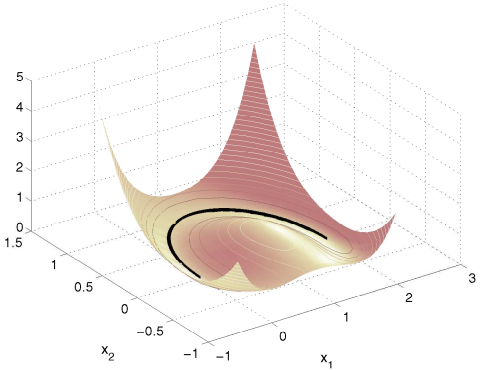
- Approximate function (i.e. substitute for indicator function)

$$\nu \check{\chi}_S = \nu \Big((\max\{0, x-1\})^2 + (\max\{0, -x\})^2 \Big)$$

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

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

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



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

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

  has at least one optimal solution x^*_ν
- $\check{\chi}_S \geq 0$; $\check{\chi}_S(x) = 0$ if and only if $x \in S$
- The Relaxation Theorem states that the inequality

$$f(x^*_\nu) + \nu \check{\chi}_S(x^*_\nu) \leq f(x^*) + \nu \check{\chi}_S(x^*) = f(x^*)$$

  holds for every positive ν (lower bound on the optimal value)
- The problem (2) is convex if (1) and $\psi(s)$ are, and $\psi(s)$ increasing for $s \geq 0$.

Assume that the problem (1) possesses optimal solutions. Then, as $\nu \to +\infty$ every limit point of the sequence $\{x_\nu^*\}$ of globally optimal solutions to (2) is globally optimal in the problem (1)

- Of interest for convex problems, since global minimum can be found relatively easily.
- Statement not very useful for general nonconvex 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 $s \geq 0$. Consider a sequence $\nu_k \to \infty$.
>
> $$\left.\begin{array}{r} x_k \text{ stationary in (2) with } \nu_k \\ x_k \to \hat{x} \text{ as } k \to +\infty \\ \text{LICQ holds at } \hat{x} \\ \hat{x} \text{ feasible in (1)} \end{array}\right\} \Longrightarrow \hat{x} \text{ stationary (KKT) in (1)}$$

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

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

- When the penalty parameter ν is very large, the unconstrained minimization subproblem becomes very badly conditioned, and hard to solve.
- In subproblem k we must start at a point x such that $x^*_{\nu_k} \approx x$.
- If we increase the penalty slowly a good guess is that $x^*_{\nu_k} \approx x^*_{\nu_{k-1}}$.
- This guess can be improved.

- 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:

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

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

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

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

  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

- Approximate function convex if g_i and ϕ are convex functions, and $\phi(s)$ increasing for $s < 0$.

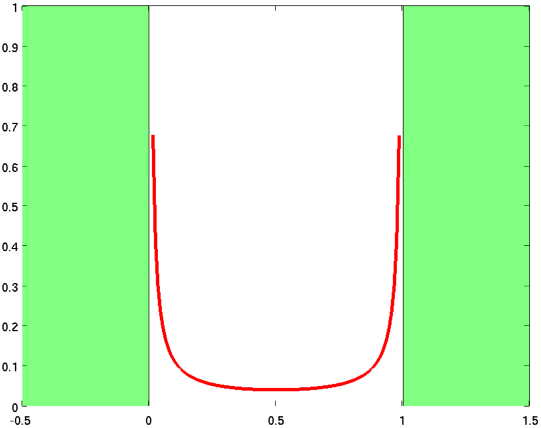
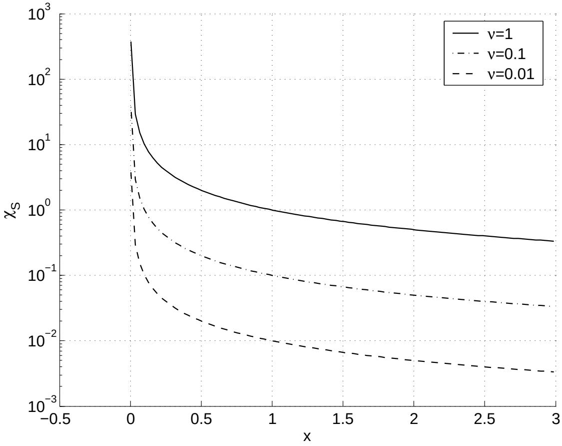


Figure : Feasible set is $S = \{x \mid -x \leq 0, x \leq 1\}$. Barrier function $\phi(s) = -1/s$, barrier parameter $\nu = 0.01$.

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 $\nu \downarrow 0$!):



- Penalty problem:

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

- 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 $s < 0$. Consider sequence $\nu_k \to 0$. Then:

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

- If we use $\phi(s) = \phi_1(s) = -1/s$, then $\phi'(s) = 1/s^2$, and the sequence $\{\nu_k/g_i^2(x_k)\} \to \hat{\mu}_i$.

- Consider the LP

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

and the corresponding system of optimality conditions:

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

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

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

- The KKT conditions for this problem is:

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

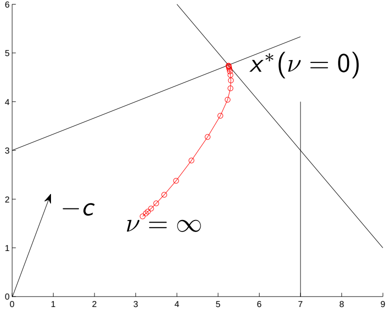
- Perturbation in the complementary conditions!

Optimal solutions to subproblems

$$\text{minimize } -b^T y - \nu \sum_{j=1}^{n} \log(s_j)$$

$$\text{subject to } A^T y + s = c$$

for different ν's form the central path.



- 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 size of the problem

- Provide guarantee that LP can be solved in polynomial time (the simplex method computation effort can grow exponentially, but this is rare).

Consider problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq \mathbf{0} \\ & h(x) = \mathbf{0} \end{aligned}$$

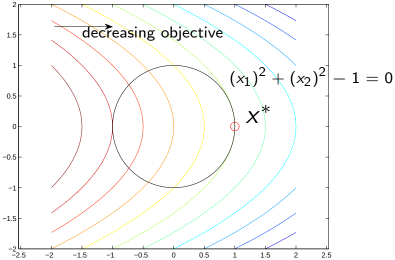
- We have good solution methods for quadratic programs (QP) (e.g., simplicial decomposition and gradient projection method)
- At iterate x_k, approximate original problem with QP subproblem. Find search direction p by solving QP subproblem

$$\begin{aligned} \underset{p}{\text{minimize}} \quad & \tfrac{1}{2} p^T \nabla^2 f(x_k) p + \nabla f(x_k)^T p \\ \text{subject to} \quad & g_i(x_k) + \nabla g_i(x_k)^T p \leq 0, \quad i = 1, \ldots, m \\ & h_j(x_k) + \nabla h_j(x_k)^T p = 0, \quad j = 1, \ldots, l \end{aligned}$$

- Suggested method does not always work!

Consider problem

$$\min_x \; -x_1 - \frac{1}{2}(x_2)^2$$
$$\text{s.t.} \; (x_1)^2 + (x_2)^2 - 1 = 0$$



Optimal solution $(1, 0)^T$, consider QP subproblem at $x_1 = 1.1$, $x_2 = 0$:

$$\underset{p}{\text{minimize}} \; -p_1 - \frac{1}{2}(p_2)^2$$
$$\text{subject to} \; p_1 + 0.0955 = 0$$

QP subproblem unbounded – bad linear approx. of nonlinear constraint!

- Linearized constraints might be too inaccurate!
- Account for nonlinear constraints in objective – Lagrangian idea.

$$L(x_k, \mu_k, \lambda_k) = f(x_k) + \mu_k^T g(x_k) + \lambda_k^T h(x_k).$$

- Solve (improved) QP subproblem to find search direction p:

$$\begin{aligned} \underset{p}{\text{minimize}} \quad & \tfrac{1}{2} p^T \nabla_{xx}^2 L(x_k, \mu_k, \lambda_k) p + \nabla f(x_k)^T p \\ \text{subject to} \quad & g_i(x_k) + \nabla g_i(x_k)^T p \leq 0, \quad i = 1, \dots, m \\ & h_j(x_k) + \nabla h_j(x_k)^T p = 0, \quad j = 1, \dots, l \end{aligned}$$

- Direction p, with multipliers μ_{k+1}, λ_{k+1}, define Newton step for solving (nonlinear) KKT conditions (see text for more).
- Lagrangian Hessian $\nabla_{xx}^2 L(x_k, \mu_k, \lambda_k)$ may not be positive definite.

- Given $x_k \in \mathbb{R}^n$ and a vector $(\mu_k, \lambda_k) \in \mathbb{R}^m_+ \times \mathbb{R}^\ell$, choose a positive definite matrix $B_k \in \mathbb{R}^{n \times n}$. $B_k \approx \nabla^2_{xx} L(x_k, \mu_k, \lambda_k)$
- Solve

$$\begin{aligned}
\underset{p}{\text{minimize}} \quad & \frac{1}{2} p^T B_k p + \nabla f(x_k)^T p, && \text{(6a)} \\
\text{subject to} \quad & g_i(x_k) + \nabla g_i(x_k)^T p \leq 0, \; i = 1, \dots, m, && \text{(6b)} \\
& h_j(x_k) + \nabla h_j(x_k)^T p = 0, \; j = 1, \dots, \ell && \text{(6c)}
\end{aligned}$$

- Working version of SQP search direction subproblem
- Quadratic convergence near KKT points. What about global convergence? Perform line search with some merit function.

1. Initialize iterate with (x_0, μ_0, λ_0), B_0 and merit function M.
2. At iteration k with (x_k, μ_k, λ_k) and B_k, solve QP subproblem for search direction p_k:

$$\begin{aligned} \underset{p}{\text{minimize}} \quad & \tfrac{1}{2} p^T B_k p + \nabla f(x_k)^T p \\ \text{subject to} \quad & g_i(x_k) + \nabla g_i(x_k)^T p \le 0, \quad i = 1, \ldots, m \\ & h_j(x_k) + \nabla h_j(x_k)^T p = 0, \quad j = 1, \ldots, l \end{aligned}$$

   Let μ_k^* and λ_k^* be optimal multipliers of QP subproblem. Define $\Delta x = p_k$, $\Delta\mu = \mu_k^* - \mu_k$, $\Delta\lambda = \lambda_k^* - \lambda_k$.

3. Perform line search to find $\alpha_k > 0$ s.t. $M(x_k + \alpha_k \Delta x) < M(x_k)$.
4. Update iterates:
   $x_{k+1} = x_k + \alpha_k \Delta x$, $\mu_{k+1} = \mu_k + \alpha_k \Delta\mu$, $\lambda_{k+1} = \lambda_k + \alpha_k \Delta\lambda$.
5. Stop if converge, otherwise update B_k to B_{k+1}; go to step 2.

Merit function as *non-differentiable* exact penalty function P_e:

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

- For large enough ν, solution to QP subproblem (6) defines a descent direction for P_e at (x_k, μ_k, λ_k).
- For large enough ν, reduction in P_e implies progress towards KKT point in the original constrained optimization problem.
  - Compare convergence results for exterior penalty methods.
  - See text for more (Proposition 13.10).

- Combining the descent direction property and exact penalty function property, one can prove convergence of the merit SQP method.

- Convergence of the SQP method towards KKT points can be established under additional conditions on the choices of matrices $\{B_k\}$
  1. Matrices B_k bounded

- Selecting the value of ν is difficult
- No guarantees that the subproblems (6) 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