Lectures 9 and 10: Penalty and Augmented Lagrangian (Multiplier) Methods

Prof. Krishna R. Pattipati
Dept. of Electrical and Computer Engineering
University of Connecticut
Contact: krishna@engr.uconn.edu (860) 486-2890

ECE 6437
Computational Methods for Optimization

Fall 2009
October 27 & Nov 3, 2009

Outline of Lectures 9 & 10

- Constrained Optimization Methods
- Penalty Methods
- Multiplier (Augmented Lagrangian) Methods
- Duality and Convergence Issues
- Extensions to Inequality Constraints
- Illustrative Examples

Constrained Optimization Methods

- Three classes of Methods

1. Convert into a sequence of unconstrained minimization problems
   - Penalty functions
   - Barrier functions … we will discuss in the context of LP in Lecture 12
   - Method of multipliers.…. Best available (also called augmented Lagrangian methods)

   Here we satisfy constraints approximately

2. Feasible direction (or) primal methods
   - Work on the original problem by moving in the feasible region
   - Nonlinear function with linearized equality and inequality constrains
   - Manifold sub-optimization Methods
     – Rosen's Gradient projection method
     – Reduced Gradient method
     – Newton type Gradient projection method

3. Successive quadratic programming (SQP) methods… Best available

Basic Idea of Penalty and Multiplier Methods

- Penalty and multiplier methods convert a constrained minimization problem into a series of unconstrained minimization problems. We consider the equality constrained problem first:

$$\min f(\underline{x}) + \sum_{i=1}^{m} \lambda_i h_i(\underline{x}) + \frac{c}{2} \sum_{i=1}^{m} h_i^2(\underline{x})$$

(or)

Penalty parameter and quadratic penalty function

$$\min f(\underline{x}) + \underline{\lambda}^T \underline{h}(\underline{x}) + \frac{c}{2} \underline{h}(\underline{x})^T \underline{h}(\underline{x}) = L_c(\underline{x}, \underline{\lambda})$$

Multiplier vector

- Penalty method: $\underline{\lambda} = 0,\ c$ is changed. Convergence to $\underline{x}^*$ occurs as $c \to \infty$
- Multiplier method: $\underline{\lambda}$ is updated and c is changed. Convergence to $\underline{x}^*$ occurs at a finite value of c

1. D.P. Bertsekas, "Multiplier methods: A survey" Automatica, vol. 12, 1976, pp. 133-145
2. R.T. Rockafeller, "Solving a nonlinear programming problem by way of a dual problem" Symposia Mathematica, vol XXVII, 1976.

Penalty Methods

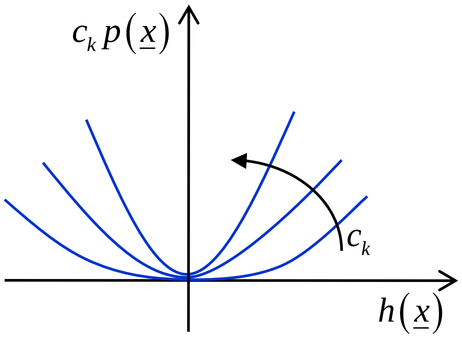
- Penalty methods

$$L_{c_k}(\underline{x}) = f(\underline{x}) + \frac{c_k}{2}\sum_{i=1}^{m} h_i^2(\underline{x}) = f(\underline{x}) + \frac{c_k}{2}\underline{h}^T(\underline{x})\underline{h}(\underline{x}) = f(\underline{x}) + \frac{c_k}{2}p(\underline{x})$$

For nonlinear inequality constraints $\underline{g}(\underline{x}) \le 0$

$$L_{c_k}(\underline{x}) = f(\underline{x}) + \frac{c_k}{2}\underline{h}^T(\underline{x})\underline{h}(\underline{x}) + \frac{c_k}{2}\underline{g}^{+^T}(\underline{x})\underline{g}^{+}(\underline{x})$$

$$\Rightarrow p(\underline{x}) = \underline{h}^T(\underline{x})\underline{h}(\underline{x}) + \underline{g}^{+^T}(\underline{x})\underline{g}^{+}(\underline{x})$$



where

$$\underline{g}_j^+(\underline{x}) = \begin{cases} 0 & \text{if } \underline{g}_j(\underline{x}) \le 0 \\ \underline{g}_j(\underline{x}) & \text{if } \underline{g}_j(\underline{x}) > 0 \end{cases} \Rightarrow \underline{g}_j^+(\underline{x}) = \max(0, \underline{g}_j(\underline{x})); \; j = 1, 2, .., r$$

Suppose we select $\{c_k\} \ni c_{k+1} > c_k$ and $c_k \ge 0\ \forall k$ and $\lim_{k\to\infty} c_k \to \infty$ and perform unconstrained minimization of the form:

$$L_{c_k}(\underline{x}) = f(\underline{x}) + c_k p(\underline{x}) \to \underline{x}_k^*; \; L_{c_{k+1}}(\underline{x}) = f(\underline{x}) + c_{k+1} p(\underline{x}) \to \underline{x}_{k+1}^*$$

Typically, $c_{k+1} = \beta c_k; \beta \in [4, 10]; c_0 = 1$

Does $\underline{x}_k^* \to \underline{x}^*$ and $L_{c_k}(\underline{x}_k^*) \to f(\underline{x}^*)$ as $k \to \infty$

We will consider the equality constrained case, since we will <u>not</u> use this method as is.

Proof of Convergence -1

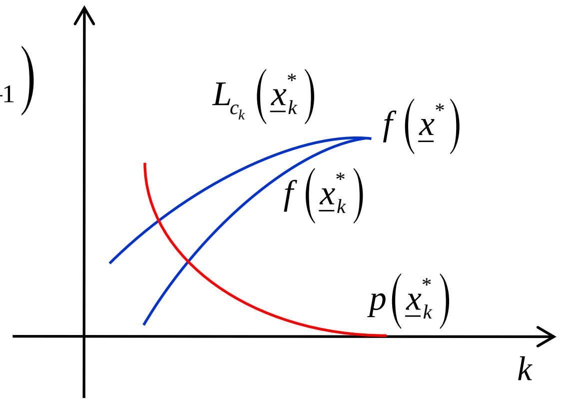
- Proof of convergence:

Results: (1) $L_{c_k}\left(\underline{x}_k^*\right) \le L_{c_{k+1}}\left(\underline{x}_{k+1}^*\right)$

(2) $p\left(\underline{x}_k^*\right) = \frac{1}{2}\underline{h}^T\left(\underline{x}_k^*\right)\underline{h}\left(\underline{x}_k^*\right) \ge \frac{1}{2}\underline{h}^T\left(\underline{x}_{k+1}^*\right)\underline{h}\left(\underline{x}_{k+1}^*\right) = p\left(\underline{x}_{k+1}^*\right)$

(3) $f\left(\underline{x}_k^*\right) \le f\left(\underline{x}_{k+1}^*\right)$

Proof: (1)
$$L_{c_{k+1}}\left(\underline{x}_{k+1}^*\right) \triangleq f\left(\underline{x}_{k+1}^*\right) + \frac{1}{2}c_{k+1}\underline{h}^T\left(\underline{x}_{k+1}^*\right)\underline{h}\left(\underline{x}_{k+1}^*\right)$$
$$\ge f\left(\underline{x}_{k+1}^*\right) + \frac{1}{2}c_k\underline{h}^T\left(\underline{x}_{k+1}^*\right)\underline{h}\left(\underline{x}_{k+1}^*\right)$$
since $c_k < c_{k+1}$
$$\ge f\left(\underline{x}_k^*\right) + \frac{1}{2}c_k\underline{h}^T\left(\underline{x}_k^*\right)\underline{h}\left(\underline{x}_k^*\right)$$
$$= L_{c_k}\left(\underline{x}_k^*\right)$$



$\boxed{L_{c_{k+1}}\left(\underline{x}_{k+1}^*\right) \ge L_{c_k}\left(\underline{x}_k^*\right)} \Rightarrow$ We construct a progressively more constrained version of the original problem

Also, note that $L_{c_k}\left(\underline{x}_k^*\right) \ge f\left(\underline{x}_k^*\right) \forall k$ since $c_k p_k\left(\underline{x}_k^*\right) \ge 0$

Proof of Convergence - 2

- Proof of convergence:

(2) Since $\underline{x}_k^*$ optimizes $L_{c_k}(\underline{x})$ and $\underline{x}_{k+1}^*$ optimizes $L_{c_{k+1}}(\underline{x})$

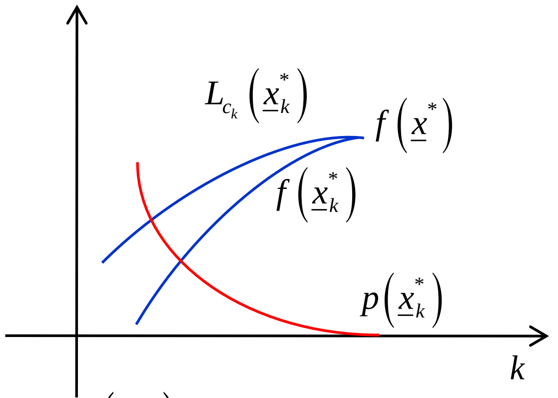
$f(\underline{x}_k^*)+c_k p(\underline{x}_k^*) \le f(\underline{x}_{k+1}^*)+c_k p(\underline{x}_{k+1}^*)$ because $\underline{x}_{k+1}^*$ is suboptimal for $L_{c_k}(\underline{x})$

$f(\underline{x}_{k+1}^*)+c_{k+1} p(\underline{x}_{k+1}^*) \le f(\underline{x}_k^*)+c_{k+1} p(\underline{x}_k^*)$ because $\underline{x}_k^*$ is suboptimal for $L_{c_{k+1}}(\underline{x})$

$$c_k p(\underline{x}_k^*)+c_{k+1} p(\underline{x}_{k+1}^*) \le c_k p(\underline{x}_{k+1}^*)+c_{k+1} p(\underline{x}_k^*)$$

$$(c_{k+1}-c_k)\, p(\underline{x}_{k+1}^*) \le (c_{k+1}-c_k)\, p(\underline{x}_k^*)$$

$$\Rightarrow \boxed{p(\underline{x}_{k+1}^*) \le p(\underline{x}_k^*)}$$



(3) Since $f(\underline{x}_{k+1}^*)+c_k p(\underline{x}_{k+1}^*) \ge f(\underline{x}_k^*)+c_k p(\underline{x}_k^*) = L_{c_k}(\underline{x}^*)$

we have

$$f(\underline{x}_{k+1}^*)-f(\underline{x}_k^*) \ge c_k\left[p(\underline{x}_k^*)-p(\underline{x}_{k+1}^*)\right] \ge 0$$

$$\Rightarrow \boxed{f(\underline{x}_{k+1}^*) \ge f(\underline{x}_k^*)}$$

approach optimal solution from below

Properties of Penalty Methods

- **Property 1:** $f\left(\underline{x}^*\right) \geq L_{c_k}\left(\underline{x}_k^*\right) \geq f\left(\underline{x}_k^*\right)$ since

$$f\left(\underline{x}^*\right) = f\left(\underline{x}^*\right) + c_k p\left(\underline{x}^*\right) \geq f\left(\underline{x}_k^*\right) + c_k p\left(\underline{x}_k^*\right) = L_{c_k}\left(\underline{x}_k^*\right) \geq f\left(\underline{x}_k^*\right)$$

- **Property 2:** $\lim_{k\to\infty} \underline{x}_k \to \underline{x}^*$

Recall "every bounded monotic sequence has a limit"

$$\Rightarrow \lim_{k\to\infty} f\left(\underline{x}_k\right) \to f\left(\bar{x}\right)$$

$\bar{x} =$ limit point and $f\left(\underline{x}\right)$ is a continuous function of $\underline{x}$

Also, $\lim_{k\to\infty} L_{c_k}\left(\underline{x}_k^*\right) \to L^* \leq f\left(\underline{x}^*\right)$ where $L^* = f\left(\underline{x}^*\right) + \lim_{k\to\infty} c_k p\left(\underline{x}_k^*\right)$

$$\Rightarrow \lim_{k\to\infty} c_k p\left(\underline{x}_k\right) = L^* - f\left(\underline{x}^*\right)$$

$$\Rightarrow \lim_{k\to\infty} p\left(\underline{x}_k\right) \to 0 \text{ since } 0 \leq p\left(\underline{x}_{k+1}\right) \leq p\left(\underline{x}_k\right) \forall k$$

$$\Rightarrow \bar{x} \text{ is feasible since } h\left(\bar{x}\right) = \underline{0}$$

To show optimality, recall that

$f\left(\bar{x}\right) = \lim_{k\to\infty} f\left(\underline{x}_k\right) \leq f\left(\underline{x}^*\right)$; since $\bar{x}$ is feasible,

$f\left(\bar{x}\right) \geq f\left(\underline{x}^*\right) \Rightarrow f\left(\bar{x}\right) = f\left(\underline{x}^*\right)$ and $\bar{x} \to \underline{x}^*$

In practice, convergence at large values of $c_k = c_k^*$

Ill-conditioning of Hessian in Penalty Methods -1

Problem of Ill-Conditioning of Hessian

Consider $L_{c_k}(\underline{x}) = f(\underline{x}) + \frac{c_k}{2}\underline{h}^T(\underline{x})\underline{h}(\underline{x})$; at optimum $c_k, \underline{x}_k^*$

$$\nabla \underline{L}_{c_k}(\underline{x}_k^*) = \nabla \underline{f}(\underline{x}_k^*) + \nabla \underline{\underline{h}}(\underline{x}_k^*)\left[c_k \underline{h}(\underline{x}_k^*)\right]$$

$$= \nabla \underline{f}(\underline{x}_k^*) + c_k \sum_{i=1}^{m} h_i(\underline{x}_k^*) \nabla \underline{h}_i(\underline{x}_k^*)$$

As $\underline{x}_k^* \to \underline{x}^*$, i.e., c_k large

$$\nabla \underline{L}_{c_k}(\underline{x}_k^*) \approx \nabla \underline{L}(\underline{x}^*) = \nabla \underline{f}(\underline{x}^*) + \nabla \underline{\underline{h}}(\underline{x}^*)\underline{\lambda}^*$$

$$\Rightarrow \underline{\lambda}^* \approx c_k \underline{h}(\underline{x}_k^*)$$

$$\nabla^2 L_{c_k}(\underline{x}_k^*) = \underbrace{\nabla^2 f(\underline{x}_k^*) + c_k \sum_{i=1}^{m} h_i(\underline{x}_k^*) \nabla^2 h_i(\underline{x}_k^*)}_{\nabla^2 L_o(\underline{x}_k^*)} + c_k \nabla \underline{\underline{h}}(\underline{x}_k^*) \nabla \underline{\underline{h}}^T(\underline{x}_k^*)$$

$$= \nabla^2 L_o(\underline{x}_k^*) + c_k \nabla \underline{h}(\underline{x}_k^*) \nabla \underline{h}^T(\underline{x}_k^*)$$

$\nabla^2 L_o(\underline{x}_k^*) \to \nabla^2 L(\underline{x}^*)$ as $k \to \infty$

$\nabla^2 L(\underline{x}^*)$ must be PD in the subspace $\nabla \underline{h}_i^T(\underline{x}^*) y = 0; \forall i = 1, 2, \ldots m$

The convergence rate of the penalty method depends on

$$\kappa\left[\nabla^2 L_{c_k}(\underline{x}_k^*)\right] = \lambda_{\max}\left(\nabla^2 L_{c_k}(\underline{x}_k^*)\right) \Big/ \lambda_{\min}\left(\nabla^2 L_{c_k}(\underline{x}_k^*)\right)$$

Ill-conditioning of Hessian in Penalty Methods -2

- A basic perturbation of Eigen values result for symmetric matrices

If A and E are symmetric matrices $\ni \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$

Then

$$\lambda_k(A)+\lambda_n(E) \leq \lambda_k(A+E) \leq \lambda_k(A)+\lambda_1(E)$$

or $\lambda_k(A)+\lambda_{\min}(E) \leq \lambda_k(A+E) \leq \lambda_k(A)+\lambda_{\max}(E)$

$$A = \overbrace{c_k \nabla \underline{h}(\underline{x}_k^*) \nabla \underline{h}^T(\underline{x}_k^*)}^{PSD} \ldots m \text{ non-zero Eigen values, } n-m \text{ zero eigenvalues}$$

$$E = \nabla^2 L_o(\underline{x}_k^*) \ldots \text{ assume } n \text{ positive eigen values} \Rightarrow \text{finite}$$

$$\text{so, } \lambda_{\min}\left(\nabla^2 L_{c_k}(\underline{x}_k^*)\right) \leq \lambda_{\max}\left(\nabla^2 L_o(\underline{x}_k^*)\right)$$

$$\kappa\left(\nabla^2 L_{c_k}(\underline{x}_k^*)\right) = \frac{\lambda_{\max}\left(\nabla^2 L_{c_k}(\underline{x}_k^*)\right)}{\lambda_{\min}\left(\nabla^2 L_{c_k}(\underline{x}_k^*)\right)}$$

Use lower bound in numerator and upper bound in denominator

$$\geq \frac{\lambda_{\max}\left(c_k \nabla \underline{h}(\underline{x}_k^*) \nabla \underline{h}^T(\underline{x}_k^*)\right) + \lambda_{\min}\left(\nabla^2 L_o(\underline{x}_k^*)\right)}{\lambda_{\max}\left(\nabla^2 L_o(\underline{x}_k^*)\right)} \to \infty \text{ as } c_k \to \infty$$

Condition number of Hessian $\to \infty$

Ill-conditioning of Hessian in Penalty Methods -3

- For high c_k, the unconstrained minimization problem becomes ill-conditioned $\Rightarrow$ SD out of the question.
- Since rank $\left[\nabla \underline{h}\left(\underline{x}^*\right)\nabla \underline{h}^T\left(\underline{x}^*\right)\right] = m$, m of the eigen values $\to \infty$ as $c_k \to \infty$

  $(n-m)$ are eigen values of $\nabla^2 L_0\left(\underline{x}_k^*\right)$ constrained to subspace $\nabla h^T\left(\underline{x}_k^*\right)\underline{y} = \underline{0}$
- For inequality constraints, $\left[m + \left|A\left(\underline{x}_k^*\right)\right|\right]$ eigenvalues $\to \infty$ as $c_k \to \infty$.

  Here, $\left|A\left(\underline{x}_k^*\right)\right|$ = Cardinality of active (binding) constraints
- Ideal set up for $(m+1)$ (or) $\left[m+1+\left|A\left(\underline{x}_k^*\right)\right|\right]$ PCG algorithm

  a x x x x b x x x x x x m large
- $$L_{C_k}\left(\underline{x}_{k+m}\right) \le \left(\frac{b-a}{b+a}\right)^2 L_{C_k}\left(\underline{x}_k\right) = \left(\frac{\kappa_e - 1}{\kappa_e + 1}\right)^2 L_{C_k}\left(\underline{x}_k\right);\ \kappa_e = \frac{b}{a}$$

  Typically, use $\underline{x}_k^*$ as the starting point for minimization of $L_{c_{k+1}}\left(\underline{x}\right)$

Augmented Lagrangian Methods - 1

- Method of multipliers (augmented Lagrangian methods)

Min $f(\underline{x})$

s.t. $\underline{h}(\underline{x}) = \underline{0}$

Augmented Lagrangian fuction:

$$L_c(\underline{x}, \underline{\lambda}) = f(\underline{x}) + \underbrace{\underline{\lambda}^T}_{\text{Multiplier vector}} \underline{h}(\underline{x}) + \underbrace{\frac{c}{2} \underline{h}^T(\underline{x}) \underline{h}(\underline{x})}_{\text{Penalty term}}$$

At the optimum: with $c = 0$

$$\nabla^2 L_o(\underline{x}^*, \underline{\lambda}^*) = \nabla^2 f(\underline{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(\underline{x}^*)$$

Assume $\nabla^2 L_o(\underline{x}^*, \underline{\lambda}^*)$ PD on the subspace

$$\nabla h^T(\underline{x}^*)\underline{y} = \underline{0} \ \Rightarrow \ \mathbb{N}\left(\nabla h^T(\underline{x}^*)\right)$$

Augmented Lagrangian Methods - 2

- Method: Given Lagrange multiplier vector $\underline{\lambda}_k$ and a penalty parameter c_k
  1. Minimize $L_{c_k}(\underline{x}, \underline{\lambda}_k)$ wrt $\underline{x}$ to obtain $\underline{x}_k$ via CG or NM or QN
  2. Update $\underline{\lambda}_{k+1} = \underline{\lambda}_k + c_k \underline{h}(\underline{x}_k)$ $\exists$ other updates
  3. Select $c_{k+1} \geq c_k$

     Go to step 1
- Geometric Interpretation

$$\min L_c(\underline{x}, \underline{\lambda}_k) = \min_{\underline{x}} \left\{ f(\underline{x}) + \underline{\lambda}_k^T \underline{h}(\underline{x}) + \frac{c}{2} \underline{h}^T(\underline{x}) \underline{h}(\underline{x}) \right\}$$

$$= \min_{\underline{x},\, \underline{u}} \left\{ f(\underline{x}) + \underline{\lambda}_k^T \underline{h}(\underline{x}) + \frac{c}{2} \underline{h}^T(\underline{x}) \underline{h}(\underline{x}) : \ \underline{h}(\underline{x}) = \underline{u} \right\}$$

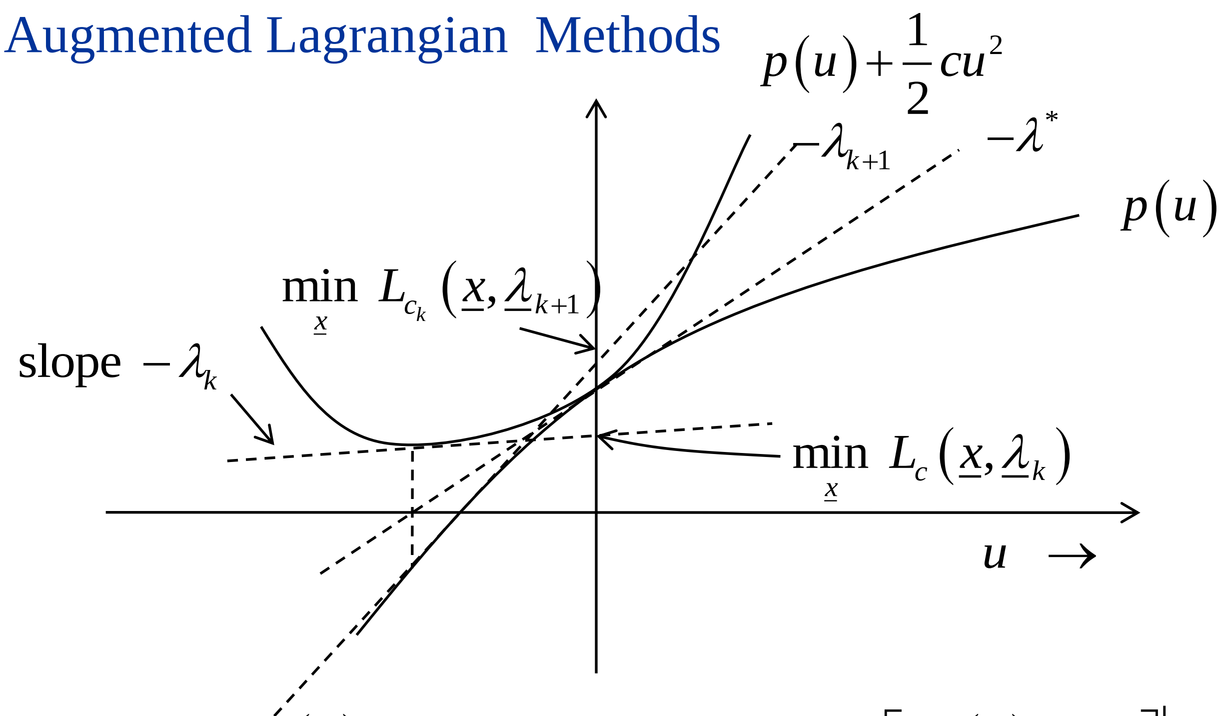
$$= \min_{\underline{u}} \min_{\underline{h}(\underline{x})=\underline{u}} \left\{ f(\underline{x}) + \underline{\lambda}_k^T \underline{h}(\underline{x}) + \frac{c}{2} \underline{h}^T(\underline{x}) \underline{h}(\underline{x}) \right\}$$

Let $p(\underline{u}) = \min_{\underline{h}(\underline{x})=\underline{u}} f(\underline{x})$

$$\Rightarrow \quad \min L_c(\underline{x}, \underline{\lambda}_k) = \min_{\underline{u}} \left[p(\underline{u}) + \underline{\lambda}_k^T \underline{u} + \frac{c}{2} \underline{u}^T \underline{u} \right]$$

Geometric Interpretation - 1

- Augmented Lagrangian Methods



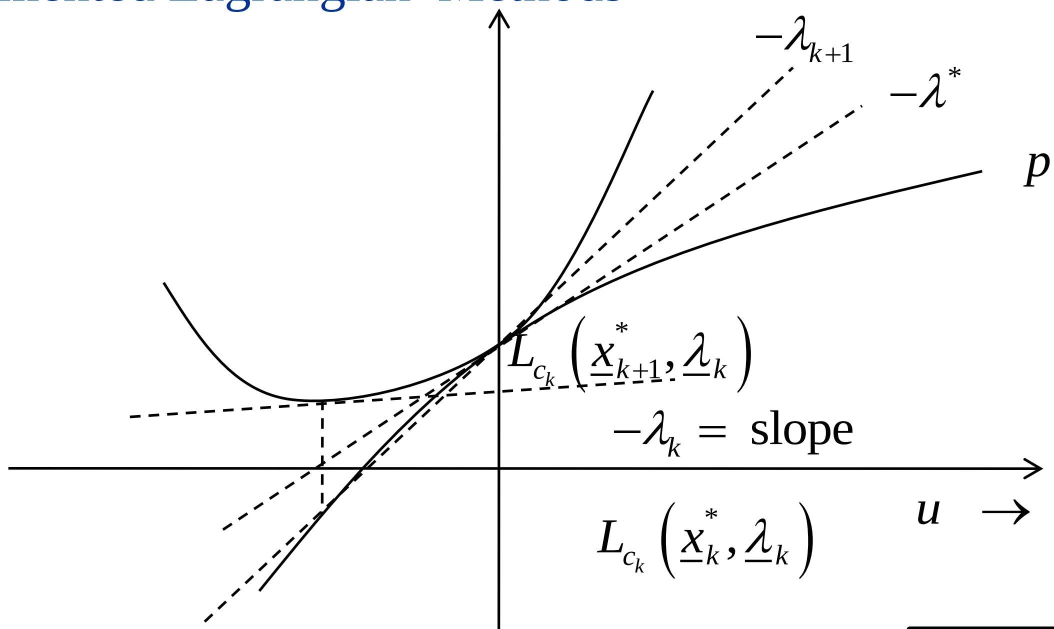
At optimum $\frac{\partial p(u)}{\partial u} + c_k u + \lambda_k = 0 \;\Rightarrow\; \lambda_k = -\left[\frac{\partial p(u)}{\partial u} + c_k u\right]\Bigg|_{u=u_k}$

$\lambda_{k+1} = \lambda_k + c_k u = -\frac{\partial p(u)}{\partial u}\Bigg|_{u=u_k}$

Note: $\lambda^* = \frac{-\partial p(u)}{\partial u}\Bigg|_{u=0}$

Geometric Interpretation - 2

- Augmented Lagrangian Methods



Augmented Lagrangian Methods Convexify Functions

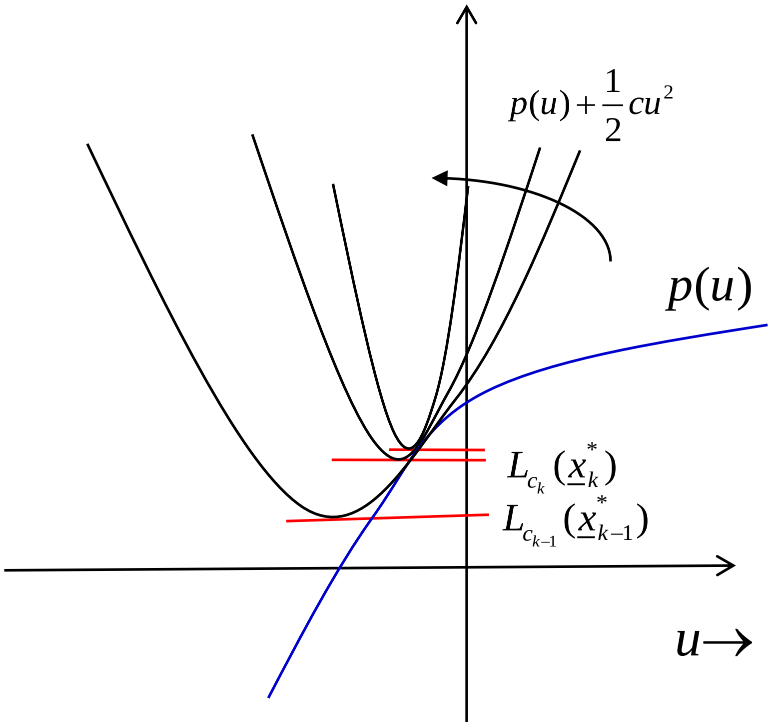
If λ_k is close to λ^*
or c_k is sufficiently large $\Rightarrow \lambda_{k+1} \rightarrow \lambda^*$

$p(u)$ linear $\Rightarrow$ convergence in one iteration

Key: c need not be large for convergence. All we need is $c > \bar{c}$.

Geometric Interpretation - 3

- In contrast, notice the ill-conditioning in Penalty Methods



$$\min_{\underline{x}}\{f(\underline{x})+\frac{1}{2}c\underline{h}^T(\underline{x})\underline{h}(\underline{x})\}$$

$$=\min_{\underline{x},\underline{u}}\{f(\underline{x})+\frac{1}{2}c\underline{h}^T(\underline{x})\underline{h}(\underline{x}):h(\underline{x})=0\}$$

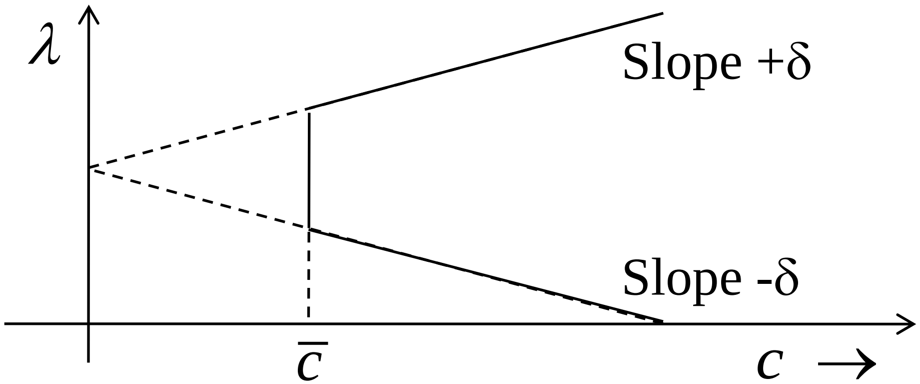
$$=\min_{\underline{u}}\{p(\underline{u})+\frac{1}{2}c\underline{u}^T\underline{u}\}$$

$$p(\underline{u})=\min_{\underline{x}} f(\underline{x})$$

Geometric Interpretation - 4

Multiplier methods, in some sense, balance the condition numbers of dual (λ update) and primal (x update) problems.

- In general $\exists$ a threshold $c=\bar{c}$ and a slope $\delta \ni$ for all $\underline{\lambda}$, c in the set $D \subset R^{m+1}$ defined by the picture below, the method converges



We will come back to this later

$\bar{c}$ is related to the eigen values of

$$\left[\nabla \underline{h}^T\left(\underline{x}^*\right)\left[\nabla_{xx}^2 L_0 + c_k \nabla \underline{h}\left(\underline{x}^*\right)\nabla \underline{h}^T\left(\underline{x}^*\right)\right]^{-1}\nabla \underline{h}\left(\underline{x}^*\right)\right]^{-1}$$

For convergence, need $\bar{c} > \max\left(0, -2\lambda_1, -2\lambda_2, \ \ldots \ , \ -2\lambda_m\right)$

If $\nabla_{xx}^2 \ L_0$ is invertible, use it to compute λ_i

How to Select c_k ?

How to select c_k

- $c_{k+1} = \beta c_k, \qquad \beta \in [4,\ 10]$

- c_0 from eigen value analysis of

  $\nabla \underline{h}^T\left(\underline{x}^*\right)\left[\nabla_{xx}^2 L_0\right]^{-1} \nabla \underline{h}^T\left(\underline{x}^*\right)$. If not, set $c_0 = 1$.

- Increase c_k only if constraint violation is not decreased by a factor $\gamma < 1$ over the previous minimization.

$$c_{k+1} = \begin{cases} \beta c_k & \text{if } \left|h\left(\underline{x}_k^*, c_k, \underline{\lambda}_k\right)\right| > \gamma \left|h\left(\underline{x}_{k+1}^*, c_{k+1}, \underline{\lambda}_{k+1}\right)\right| \\ c_k & \text{otherwise} \end{cases}$$

$\gamma = 0.25$ typically, and $\beta = 4$

Implementation of AL Methods

- Given $\underline{\lambda}_0,\ c_0 = 1,\ \beta \in (4,10)\ \gamma \cong \frac{1}{4},\ k = 0$

Method of multipliers

$$\begin{cases} \text{Step 1:} & \text{Solve for optimal } \underline{x}_k^* \text{ of } L_c(\underline{x}, \underline{\lambda}_k) \ldots \underline{x}_k^* = \textit{function of } (c_k, \underline{\lambda}_k) \\ \text{Step 2:} & \text{Update } \underline{\lambda}_{k+1} = \underline{\lambda}_k + c_k h(\underline{x}_k^*) \\ & \text{If } |h(\underline{x}_k^*)| < \gamma |h(\underline{x}_{k-1}^*)| \\ & \qquad c_{k+1} = c_k \\ & \text{else} \\ & \qquad c_{k+1} = \beta c_k \\ & \text{end if} \\ \text{Step 3:} & \text{Check for convergence of } \underline{x}_k^*. \text{ If not converged, go to step 1} \end{cases}$$

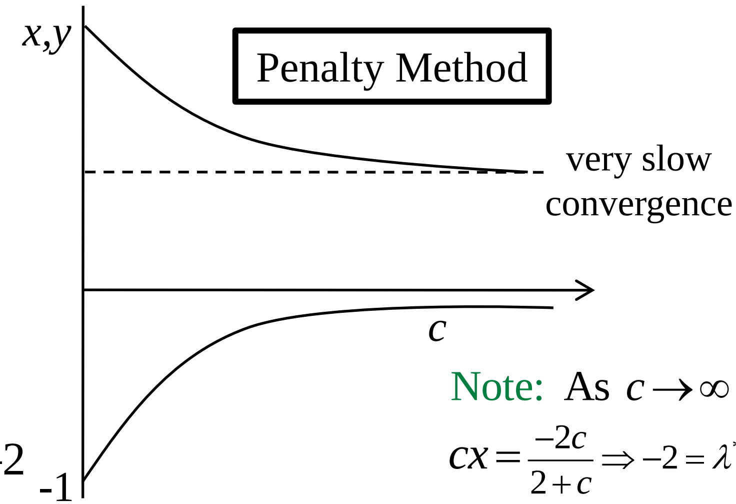
Illustrative Examples - 1

Example 1: Convex function

$$\min\ 2x^2 + 2xy + y^2 - 2y$$
$$\text{s.t.}\quad x = 0$$

$$L(x, y, \lambda) = 2x^2 + 2xy + y^2 - 2y + \lambda x$$

$$\left.\begin{aligned} \nabla_x L = 0 &\Rightarrow \quad 4x + 2y + \lambda = 0 \\ \nabla_y L = 0 &\Rightarrow \quad 2x + 2y - 2 = 0 \\ \nabla_\lambda L = 0 &\Rightarrow \quad x = 0 \end{aligned}\right\} \Rightarrow \quad \begin{aligned} x^* &= 0 \\ y^* &= 1 \\ \lambda^* &= -2 \end{aligned}$$



Note: As $c \to \infty$

$$cx = \frac{-2c}{2+c} \Rightarrow -2 = \lambda^*$$

- Penalty :

$$L_c(x, y) = 2x^2 + 2xy + y^2 - 2y + \frac{1}{2}cx^2$$

$$\nabla_x L_c \qquad 4x + 2y + cx = 0 \ \Rightarrow 4x + 2 - 2x + cx = 0 \Rightarrow x = \frac{-2}{2+c}$$

$$\nabla_y L_c \qquad 2x + 2y - 2 = 0 \quad \Rightarrow y = 1 - x = 1 + \frac{2}{2+c} = \frac{4+c}{2+c}$$

Illustrative Examples - 2

- Primal-dual method :

– for a given λ, $\min_{\underline{x}} L(\underline{x},\lambda) = q(\lambda)$ primal

– $\max_{\lambda} \; q(\lambda)$ dual

$$L(\underline{x},\lambda) = 2x^2 + 2xy + y^2 - 2y + \lambda x$$

$$\nabla_x L = 0 \Rightarrow \quad 4x + 2y + \lambda = 0 \Rightarrow x = -\left(1 + \frac{\lambda}{2}\right) \text{ from } y = 1 - x$$

$$\nabla_y L = 0 \Rightarrow \quad 2x + 2y - 2 = 0 \Rightarrow y = 1 - x \Rightarrow y = 2 + \frac{\lambda}{2} \text{ from } x = -\left(1 + \frac{\lambda}{2}\right)$$

$$q(\lambda) = (x + y)^2 + x^2 - 2y + \lambda x$$

$$= 1 + \left(\frac{\lambda}{2} + 1\right)^2 - 2\left(2 + \frac{\lambda}{2}\right) - \lambda\left(1 + \frac{\lambda}{2}\right)$$

$$= \frac{-\lambda^2}{4} - \lambda - 2$$

$$q'(\lambda) = \frac{-2\lambda}{4} - 1 = 0 \Rightarrow \lambda = -2, x = 0, y = 1. \text{ But, we will use an iterative update for } \lambda$$

Illustrative Examples - 3

- Steepest Ascent Iteration :

$$\lambda_{k+1} = \lambda_k + \alpha \overbrace{\left(\frac{-2\lambda_k}{4} - 1 \right)}^{\nabla q(\lambda)}$$

$$= \lambda_k + \alpha\left(\frac{-\lambda_k}{2} - 1\right)$$

$$= \lambda_k + \alpha x_1^*(\lambda) = \lambda_k + \alpha h(\underline{x}^*(\lambda_k))$$

- Newton's Method:

$$\lambda_{k+1} = \lambda_k + 2\left(\frac{-\lambda_k}{2} - 1 \right) = -2$$

$$\lambda_{k+1} = \lambda_k - [\nabla^2 q(\underline{\lambda})]^{-1} \nabla \underline{q}(\lambda); \nabla^2 q(\underline{\lambda}) = -\frac{1}{2}; \nabla \underline{q}(\lambda) = -\frac{\lambda}{2} - 1$$

Use damped Newton's method, etc.

Illustrative Examples - 4

- Augmented Lagrangian Method:

$$L_C(\underline{x},\lambda) = 2x^2 + 2xy + y^2 - 2y + \lambda_k x + \frac{1}{2}cx^2$$

$$\text{min at}: 4x + 2y + \lambda_k + cx = 0 \quad (1)$$

$$2x + 2y - 2 = 0 \quad (2)$$

$$\text{Using (2)}: \; 2 + 2x + \lambda_k + cx = 0$$

$$\Rightarrow x = \frac{-(\lambda_k + 2)}{(c+2)}$$

$$y = \frac{4 + \lambda_k + c}{c+2}$$

$$\lambda_{k+1} = \lambda_k - c\frac{\lambda_k + 2}{c+2} = \left(\frac{2}{2+c}\right)\lambda_k - \left(\frac{2c}{2+c}\right)$$

Note that for small enough c, $\lambda_k \to -2$ as $k \to \infty$

$$c = 2: \lambda_0 = 0 \;\Rightarrow x_0 = -\frac{1}{2} \;\Rightarrow \lambda_1 = -1 \Rightarrow x = -\frac{1}{3} \Rightarrow \lambda_2 = -1.5 \;\Rightarrow x = -\frac{1}{8} \quad \text{etc.}$$

$$y = 1.5 \qquad\qquad y = \frac{5}{4} \qquad\qquad y = 1.125$$

Illustrative Examples - 5

- Can also use Newton's method

$$\nabla^2 q = -\nabla h^T(\underline{x}_k^*)\,[\nabla_{xx}^2 L_c(\underline{x}_k^*, \underline{\lambda}_k)]^{-1}\,\nabla h(\underline{x}_k^*)\;;\quad \nabla q = \underline{h}(\underline{x}_k^*)$$

$$\nabla h = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \nabla_{xx}^2 L_c = \nabla^2 f + c_k \nabla h \nabla h^T$$

$$= \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} + c_k \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 4 + c_k & 0 \\ 0 & 2 \end{bmatrix}$$

$$\nabla^2 q = -\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1/(4+c_k) & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = -\frac{1}{4+c_k}$$

$$\lambda_{k+1} = \lambda_k - \left(\frac{1}{4+c_k}\right)^{-1} \cdot \frac{(\lambda_k + 2)}{(c_k + 2)}$$

$$= \frac{-2}{2+c_k}\lambda_k - 2\left[1 + \frac{2}{2+c_k}\right]$$

$$= -2 - \frac{2}{2+c_k}(\lambda_k + 2)$$

$$c_k = 2 \Rightarrow \lambda_{k+1} = -2 - \frac{1}{2}.2 = -3, x = \frac{1}{4}, y = \frac{3}{4} \Rightarrow \lambda = -\frac{3}{2}, x = -\frac{1}{8}, y = 1.125, \text{etc.}$$

Illustrative Examples - 6

- Example 2: Non-convex function

$$\min f(x) = \frac{1}{2}(x_1^2 - x_2^2) - x_2$$
$$\text{s.t. } x_2 = 0$$

$$\frac{1}{2}(x_1^2 - x_2^2) - x_2 + \lambda x_2$$
$$\nabla_{\underline{x}} L = \underline{0} \Rightarrow \quad x_1 = 0$$
$$-x_2 - 1 + \lambda = 0 \ \Rightarrow x_2 = \lambda - 1$$
$$\nabla_{\lambda} L = 0 \Rightarrow x_2 = 0$$
$$x_2 \ = 0 \Rightarrow \lambda = 1$$

- Penalty

$$L_c(\underline{x}) = \frac{1}{2}(x_1^2 - x_2^2) - x_2 + \frac{1}{2} c x_2^2$$
$$x_1 = 0$$
$$-x_2 - 1 + c x_2 = 0 \quad \Rightarrow \quad x_2 = \frac{1}{c-1} \qquad c > 1$$
$$\text{as } \ c \to \infty \quad x_2 \to 0$$
$$\text{Also as } \ c \to \infty \quad c x_2 = \frac{c}{c-1} \ \to 1 = \lambda^*$$

Illustrative Examples - 7

- Primal-Dual:

$$L(\underline{x},\lambda) = \frac{1}{2}(x_1^2 - x_2^2) - x_2 + \lambda x_2$$

$$\nabla_{\underline{x}} L = \underline{0} \Rightarrow \quad x_1 = 0$$

$$-x_2 - 1 + \lambda = 0 \quad \Rightarrow x_2 = \lambda - 1 \quad \text{It is not minimum}$$

$$q(\lambda) = -\frac{1}{2}(\lambda - 1)^2 - 1 + \lambda + \lambda(\lambda - 1)$$

$$= \frac{-\lambda^2}{2} + \lambda - \frac{1}{2} - 1 + \lambda + \lambda^2 - \lambda$$

$$= \frac{\lambda^2}{2} + \lambda - \frac{3}{2}$$

No maximum $\quad \lambda \to \infty$

Primal-dual methods do not work for non convex functions!!

Illustrative Examples - 8

- Augmented Lagrangian Methods:
  - Convexifies of the function

$$L_c(\underline{x},\lambda)=\frac{1}{2}(x_1^2-x_2^2)-x_2+\lambda_k x_2+\frac{1}{2}cx_2^2$$

$$\nabla_{\underline{x}}L=\underline{0}\Rightarrow \quad x_1=0$$

$$-x_2-1+\lambda_k+cx_2=0 \quad \Rightarrow \quad x_2=\frac{1-\lambda_k}{c-1}$$

$$\lambda_{k+1}=\lambda_k+\frac{c}{c-1}(1-\lambda_k)$$

$$=-\frac{1}{c-1}\lambda_k+\frac{c}{c-1}$$

All we need is $c>1$. In fact, need $c>2$ to make $\frac{1}{c-1}<1$

- The threshold is due to non-convexity of f

$$\bar{c}>\max(-2\lambda_i)$$

$$\lambda_i=\text{eigen values of } \nabla h^T[\nabla_{xx}^2L]^{-1}\nabla h$$

$$\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}^{-1}\begin{bmatrix}0\\ 1\end{bmatrix}=-1$$

Primal-Dual Viewpoint

- Primal-Dual viewpoint of multiplier (augmented Lagrangian) methods:

Consider the problem

$$\min\ f(\underline{x})+\frac{1}{2}c\underline{h}^T(\underline{x})\underline{h}(\underline{x})$$
$$\text{s.t. } \underline{h}(\underline{x})=\underline{0}$$

Dual: $q(\underline{\lambda})=\min_{\underline{x}}\left[f(\underline{x})+\frac{1}{2}c\underline{h}^T(\underline{x})\underline{h}(\underline{x})+\underline{\lambda}^T\underline{h}(\underline{x})\right]=\min_{\underline{x}}\ L_c(\underline{x},\underline{\lambda})$

Shown earlier that

$$\min\ f(\underline{x})+\frac{c}{2}\underline{h}^T(\underline{x})\underline{h}(\underline{x}) \quad \Leftrightarrow \quad \max_{\underline{\lambda}}\ q(\underline{\lambda})$$
$$\text{s.t. } \underline{h}(\underline{x})=\underline{0}$$

Steepest Ascent Iteration - 1

- Steepest ascent iteration for $\underline{\lambda}$:

$$\underline{\lambda}_{k+1} = \underline{\lambda}_k + \alpha \nabla \underline{q}(\underline{\lambda})$$

We can show that $\nabla \underline{q}(\underline{\lambda}) = \underline{h}(\underline{x}) = \underline{h}\left(\underline{x}^*(\underline{\lambda})\right)$ (1)

Opt. $\underline{x}$ for a given $\underline{\lambda}$

$$\nabla^2 q(\underline{\lambda}) = -\nabla \underline{\underline{h}}^T\left(\underline{x}^*(\underline{\lambda})\right)\left[\nabla_{xx}^2 L_c\left(\underline{x}^*(\underline{\lambda}), \underline{\lambda}\right)\right]^{-1} \nabla \underline{\underline{h}}\left(\underline{x}^*(\underline{\lambda})\right) \quad (2)$$

full rank PD

$$\Rightarrow \quad \nabla^2 q(\underline{\lambda}) < 0$$

- Proof of (1)

$$q(\underline{\lambda}) = f\left(\underline{x}^*(\underline{\lambda})\right) + \frac{1}{2} c \underline{h}^T\left(\underline{x}^*(\underline{\lambda})\right)\underline{h}\left(\underline{x}^*(\underline{\lambda})\right) + \underline{\lambda}^T \underline{h}\left(\underline{x}^*(\underline{\lambda})\right)$$

$$\nabla \underline{q}(\underline{\lambda}) = \nabla_{\underline{\lambda}} \underline{x}^*(\underline{\lambda})\left[\nabla \underline{f}\left(\underline{x}^*(\underline{\lambda})\right) + \nabla \underline{\underline{h}}\left(\underline{x}^*(\underline{\lambda})\right)\underline{\lambda} + c \nabla \underline{\underline{h}}\left(\underline{x}^*(\underline{\lambda})\right)\underline{h}\left(\underline{x}^*(\underline{\lambda})\right)\right]$$

$$+ \underline{h}\left(\underline{x}^*(\underline{\lambda})\right)$$

Steepest Ascent Iteration - 2

since by definition of $q(\underline{\lambda})$, the first term is zero

$$\nabla \underline{q}(\underline{\lambda}) = \underline{h}\left(x^*(\underline{\lambda})\right)$$

$\Rightarrow \quad \underline{\lambda}_{k+1} = \underline{\lambda}_k + \alpha\ \underline{h}\left(x^*(\underline{\lambda})\right)$ is the steepest ascent method

When $\alpha = c \quad \Rightarrow$ same as penalty viewpoint.

So, $\underline{\lambda}$ iteration can be viewed as steepest ascent iteration for maximizing the dual functional

- Convergence rate depends on eigen values of $\nabla^2 q(\lambda)$: Proof of (2)

Know $\nabla \underline{f}\left(\underline{x}^*(\underline{\lambda})\right) + \nabla \underline{\underline{h}}\left(\underline{x}^*(\underline{\lambda})\right)\underline{\lambda} + c\nabla\underline{\underline{h}}\left(\underline{x}^*(\underline{\lambda})\right)\underline{h}\left(\underline{x}^*(\underline{\lambda})\right) = 0$

$$\nabla_{\underline{\lambda}}\left(\underline{x}^*(\underline{\lambda})\right)\Bigg[\underbrace{\nabla^2 f\left(\underline{x}^*(\underline{\lambda})\right) + \sum_{i=1}^{m}\lambda_i \nabla^2 h_i\left(\underline{x}^*(\underline{\lambda})\right) + c\sum_{i=1}^{m} h_i \nabla^2 h_i\left(\underline{x}^*(\underline{\lambda})\right)}_{\nabla^2_{xx} L_0} +$$

$$c\nabla\underline{\underline{h}}\left(\underline{x}^*(\underline{\lambda})\right)\nabla\underline{\underline{h}}^T\left(\underline{x}^*(\underline{\lambda})\right)\Bigg] + \nabla\underline{\underline{h}}^T\left(\underline{x}^*(\underline{\lambda})\right) = 0$$

Steepest Ascent Iteration - 3

$$\nabla_{\underline{\lambda}}\left(\underline{x}^{*}(\underline{\lambda})\right)=-\nabla\underline{\underline{h}}^{T}(\underline{x}^{*}(\underline{\lambda}))\left[\nabla_{xx}^{2}L_{0}\left(\underline{x}^{*}(\underline{\lambda}),\underline{\lambda}\right)+c\nabla\underline{\underline{h}}\left(\underline{x}^{*}(\underline{\lambda})\right)\nabla\underline{\underline{h}}^{T}\left(\underline{x}^{*}(\underline{\lambda})\right)\right]^{-1}$$

$$=-\nabla\underline{\underline{h}}^{T}(\underline{x}^{*}(\underline{\lambda}))\left[\nabla_{xx}^{2}L_{c}\left(\underline{x}^{*}(\underline{\lambda}),\underline{\lambda}\right)\right]^{-1}$$

Know $\nabla\underline{q}(\underline{\lambda})=\underline{h}\left(\underline{x}^{*}(\underline{\lambda})\right)$

$$\nabla^{2}q(\underline{\lambda})=\nabla_{\underline{\lambda}}\left(\underline{x}^{*}(\underline{\lambda})\right)\ \nabla\underline{\underline{h}}\left(\underline{x}^{*}(\underline{\lambda})\right)$$

$$=-\nabla\underline{\underline{h}}^{T}\left(\underline{x}^{*}(\underline{\lambda})\right)\left[\nabla_{xx}^{2}L_{c}\left(\underline{x}^{*}(\underline{\lambda}),\underline{\lambda}\right)\right]^{-1}\nabla\underline{\underline{h}}\left(\underline{x}^{*}(\underline{\lambda})\right)$$

For large values of c

$$\nabla^{2}q(\underline{\lambda})\cong\frac{1}{c}I\ \Rightarrow\ \underline{\lambda}_{k+1}=\underline{\lambda}_{k}+c\underline{h}\left(\underline{x}^{*}(\underline{\lambda}_{k})\right)\ \Rightarrow$$ Newton's method for large c

- Can use Newton's method to update $\underline{\lambda}_{k}^{s}$

$$\underline{\lambda}_{k+1}=\underline{\lambda}_{k}-\left[\nabla^{2}q(\underline{\lambda})\right]^{-1}\underline{h}\left(\underline{x}^{*}(\underline{\lambda})\right)$$

Plus All the tricks of unconstrained minimization!!

AL = Primal-dual on a Penalty function

- Augmented Lagrangian = Primal-dual on a penalty function

Primal

$$\min\{f(\underline{x})+\frac{1}{2}c\underline{h}^T(\underline{x})\underline{h}(\underline{x})\}$$

$$\text{s.t.}\quad \underline{h}(\underline{x})=0$$

Dual

$$\max_{\underline{\lambda}} q(\underline{\lambda}), \text{where } q(\underline{\lambda})=\min_{\underline{x}}\{f(\underline{x}) +(1/2)c\underline{h}^T(\underline{x})\underline{h}(\underline{x})+\underline{\lambda}^T\underline{h}(\underline{x})\}$$

$\therefore$ Steepest ascent iteration for the dual:

$$\underline{\lambda}_{k+1}=\underline{\lambda}_k+\alpha\nabla \underline{q}(\underline{\lambda}_k)=\underline{\lambda}_k+\alpha_k\underline{h}(\underline{x}_k^*)$$

If $\alpha_k=c_k$, the previous iteration is the steepest ascent iteration for the dual.

So, multiplier (augemented Lagrangian) method is a primal-dual method:

Do until convergence

$\boxed{\text{Solve for } \underline{x}_k^* \text{ by } \min L_{c_k}(\underline{x},\underline{\lambda}_k)}$ Primal

$\boxed{\text{Update } \underline{\lambda}_k : \underline{\lambda}_{k+1}=\underline{\lambda}_k+c_k h(\underline{x}_k); \text{Update } c_k}$ Dual

Why Steepest Ascent Iteration ?

- The convergence rate of the primal-dual method depends on the primal and dual Hessians

For primal method, Hessian is

$$\nabla^2_{xx} L_{c_k}(\underline{x}_k^*, \underline{\lambda}_k^*) = \nabla^2_{xx} L_0(\underline{x}_k^*, \underline{\lambda}_k^*) + c_k \nabla \underline{\underline{h}}(\underline{x}_k^*) \nabla \underline{\underline{h}}^T(\underline{x}_k^*)$$

$$\nabla^2_{xx} L_0(\underline{x}_k^*, \underline{\lambda}_k^*) = \nabla^2 f(\underline{x}_k^*) + \sum_{i=1}^{m} \underbrace{\left[\lambda_k + c_k \underline{h}(\underline{x}_k^*)\right]_i}_{\lambda_{k+1,\,i}} \nabla^2 \underline{h}_i(\underline{x}_k^*) \to \nabla^2_{xx} L_0(\underline{x}^*, \underline{\lambda}^*)$$

i^{th} component of next $\underline{\lambda}$

As $c_k \to \infty$, primal is ill-conditioned

On the other hand, the convergence of dual depends on

$$\nabla^2 q(\underline{\lambda}_k) = -\nabla \underline{\underline{h}}^T(\underline{x}_k^*) \left[\nabla^2_{xx} L_c(\underline{x}_k^*, \underline{\lambda}_k)\right]^{-1} \nabla \underline{\underline{h}}(\underline{x}_k^*) \quad m \times m \text{ matrix}$$

As $c_k \to \infty$, $\nabla^2 q(\underline{\lambda}_k) \to -\frac{1}{c} I \Rightarrow \kappa\left[\nabla^2 q(\underline{\lambda}_k)\right] = 1 \Rightarrow$ well-conditioned

Multiplier methods ensure convergence at reasonable values of c_k

AL balances primal and dual condition numbers

Choices for c_k -1

- What are the reasonable values of c_k?

  Recall that $\nabla^2_{xx} L_{c_k}(\underline{x}^*, \underline{\lambda}^*) = \nabla^2_{xx} L_0(\underline{x}^*, \underline{\lambda}^*) + c_k \nabla \underline{\underline{h}}(\underline{x}^*) \nabla \underline{\underline{h}}^T(\underline{x}^*)$

  Assume second order Kuhn-Tucker Condition is valid:

$$\underline{y}^T \nabla^2_{xx} L_0(\underline{x}^*, \underline{\lambda}^*) \underline{y} > 0 \quad \forall \underline{y} \ni \nabla \underline{\underline{h}}^T(\underline{x}^*) \underline{y} = \underline{0}$$

$$\Rightarrow \underline{y}^T \nabla^2_{xx} L_0(\underline{x}^*, \underline{\lambda}^*) \underline{y} > 0 \quad \forall \underline{y} \ni \underline{y}^T \nabla \underline{\underline{h}}(\underline{x}^*) \nabla \underline{\underline{h}}^T(\underline{x}^*) \underline{y} = 0$$

  Any vector $\underline{w} \ni \|\underline{w}\| = 1$ can be written as

$$\underline{w} = \underline{y} + \underline{z} \qquad \underline{y} = N(\nabla \underline{\underline{h}}^T); \ \underline{z} \perp \underline{y}$$

$$\underline{w}^T \nabla^2_{xx} L_{c_k}(\underline{x}^*, \underline{\lambda}^*) \underline{w} = \underline{w}^T \nabla^2_{xx} L_0 \underline{w} + c \underline{z}^T \nabla \underline{\underline{h}}(\underline{x}^*) \nabla \underline{\underline{h}}^T(\underline{x}^*) \underline{z}$$

  since $\|\underline{w}\| = 1, \ \exists$ a scalar $\bar{c} \ni$ the sum $\nabla^2_{xx} L_0 + c \nabla \underline{\underline{h}}(\underline{x}^*) \nabla \underline{\underline{h}}^T(\underline{x}^*) > 0$

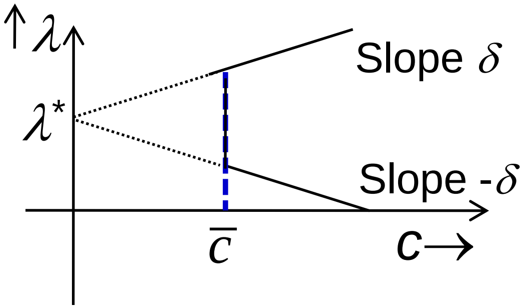
  $\Rightarrow$ For $c > \bar{c}$, the primal has a local minimum

$$c > \max(-2\lambda_i)$$

$$\lambda_i \text{ of } \{-\nabla \underline{\underline{h}}^T [\nabla^2_{xx} L]^{-1} \nabla \underline{\underline{h}}\} = \nabla^2 q$$

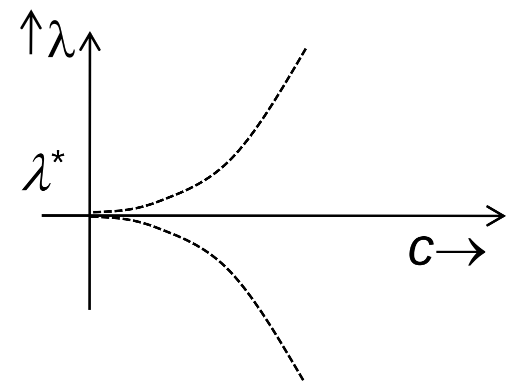
Choices for c_k -2

- In fact, (λ, c) pair should be in a region for convergence to occur



Proof is technical
See Bertsekas

❑ Example: $\min -x^4;$ s.t. $x=0$



$$\nabla_x L_c(x,\lambda) = -4x^3 + \lambda + cx = 0$$

$$\nabla^2_{xx} L_c(x,\lambda) = -12x^2 + c \Rightarrow \nabla^2_{xx} L_c(x,\lambda) > 0 \ \forall |x| < \sqrt{c/12}$$

$$\nabla_x L_c(x,\lambda) = 0 \Rightarrow \lambda = x[4x^2 - c]$$

$$so,\ |\lambda| < \frac{2}{3}\frac{1}{\sqrt{12}}c^{3/2} = \frac{1}{3\sqrt{3}}c^{3/2}$$

In general for $\min -x^p;$ s.t. $x=0$

$$\{(\lambda,c) \,|\, |\lambda| < \frac{p-2}{p-1}\left[\frac{1}{p(p-1)}\right]^{\frac{1}{p-2}} c^{(p-1)/(p-2)},\ c > 0\} \Rightarrow |\lambda| < \delta c \text{ as } p \to \infty$$

Why α = c Works? - 1

- Step-size Analysis for Gradient Methods (or) Why $\alpha = c$ works

Consider a quadratic function $\frac{1}{2}\underline{x}^T Q\underline{x}$; Q need not be PD

Want to minimize this function subject to $A\underline{x} = \underline{b}$

$$\min_{\underline{x}} \quad \frac{1}{2}\underline{x}^T Q\underline{x} + \frac{1}{2}c(A\underline{x}-\underline{b})^T(A\underline{x}-\underline{b})$$

s.t. $A\underline{x} = \underline{b}$

Assume Q is PD in $N(A)$

$$\Rightarrow \underline{y}^T Q\underline{y} > 0 \ \forall \underline{y} \ni A\underline{y} = \underline{0}$$

$$\text{Dual Problem: } q(\underline{\lambda}) = \min_{\underline{x}}\left[\frac{1}{2}\underline{x}^T Q\underline{x} + \underline{\lambda}^T(A\underline{x}-\underline{b}) + \frac{c}{2}(A\underline{x}-\underline{b})^T(A\underline{x}-\underline{b})\right]$$

$$= \min_{\underline{x}}\left[L_c(\underline{x},\underline{\lambda})\right]$$

$$\nabla_x L_c(\underline{x},\underline{\lambda}) = 0 \Rightarrow Q\underline{x}_\lambda^* + A^T\underline{\lambda} + cA^T(A\underline{x}_\lambda^* - \underline{b})$$

$$\underline{x}_\lambda^* = -(Q + cA^TA)^{-1}A^T[\underline{\lambda} - c\underline{b}]$$

Why α = *c* Works? - 2

The dual function is

$$q(\underline{\lambda}) = \frac{1}{2}\, x_\lambda^{*T} Q\, x_\lambda^* + \underline{\lambda}^T A x_\lambda^* + \frac{c}{2}(A x_\lambda^* - \underline{b})^T (A x_\lambda^* - \underline{b})$$

$$= \frac{1}{2}\underbrace{\left[\, x_\lambda^{*T} Q + \underline{\lambda}^T A + c(A x_\lambda^* - \underline{b})^T A \right]}_{=0} x_\lambda^* + \frac{\underline{\lambda}^T A x_\lambda^*}{2} - \frac{c}{2}(A x_\lambda^* - \underline{b})^T \underline{b}$$

$$= \frac{\underline{\lambda}^T A x_\lambda^*}{2} - \frac{c}{2}(A x_\lambda^* - \underline{b})^T \underline{b} = \frac{c}{2}\underline{b}^T \underline{b} + \frac{1}{2}(\underline{\lambda} - c\underline{b})^T A x_\lambda^*$$

$$= \frac{c}{2}\underline{b}^T \underline{b} - \frac{1}{2}(\underline{\lambda} - c\underline{b})^T A(Q + cA^T A)^{-1} A^T [\underline{\lambda} - c\underline{b}]$$

Note that $c\underline{b} = cA\underline{x}^* = \underline{\lambda}^*$ and optimal solution to $\min \frac{1}{2}\, \underline{x}^T Q \underline{x}$ s.t. $A\underline{x} = \underline{b}$

is $\frac{1}{2}\, \underline{\lambda}^{*T} \underline{b} = \frac{c}{2}\, \underline{b}^T \underline{b} \Rightarrow q(\underline{\lambda}) = \frac{1}{2}(\underline{\lambda} - \underline{\lambda}^*)^T \overbrace{A(Q + cA^T A)^{-1} A^T}^{-\nabla^2 q} (\underline{\lambda} - \underline{\lambda}^*) \leq f(\underline{x}^*)$

Also, note that $q(\underline{\lambda}^*) = f(\underline{x}^*)$ as it must.

Why α = *c* Works? - 3

Now consider steepest ascent iteration:

$$\underline{\lambda}_{k+1} = \underline{\lambda}_k + \alpha \nabla \underline{q}(\underline{\lambda}_k) = \underline{\lambda}_k + \alpha \underline{h}(\underline{x}_k^*); \quad \underline{x}_k^* \text{ opt. for } \underline{\lambda}_k$$

For steepest ascent iteration

$$(\underline{\lambda}_{k+1} - \underline{\lambda}^*)^T (\underline{\lambda}_{k+1} - \underline{\lambda}^*) = \left[\underline{\lambda}_k - \underline{\lambda}^* + \alpha \nabla \underline{q}(\underline{\lambda}_k)\right]^T \left[\underline{\lambda}_k - \underline{\lambda}^* + \alpha \nabla \underline{q}(\underline{\lambda}_k)\right]$$

$$= (\underline{\lambda}_k - \underline{\lambda}^*)^T \left[I + \alpha \nabla^2 q\right]^2 (\underline{\lambda}_k - \underline{\lambda}^*)$$

Recall $\nabla \underline{q}(\underline{\lambda}_k) = \nabla^2 q(\underline{\lambda}_k - \underline{\lambda}^*)$ to get the above equation

$$\left\|\underline{\lambda}_{k+1} - \underline{\lambda}^*\right\|^2 \le \lambda_{\max}\left[(I + \alpha \nabla^2 q)^2\right] \left\|\underline{\lambda}_k - \underline{\lambda}^*\right\|^2$$

Let $\{\omega_i\}$ be the eigen values of $-\nabla^2 q = A(Q + cA^T A)^{-1}$. The eigen values of $(I + \alpha \nabla^2 q)$ are $\{(1 - \alpha\omega_i)\} \Rightarrow \lambda_{\max}(I + \alpha \nabla^2 q) = \max\left[\left|1 - \alpha\omega_{\min}\right|, \left|1 - \alpha\omega_{\max}\right|\right]$

$$\Rightarrow \frac{\left\|\underline{\lambda}_{k+1} - \underline{\lambda}^*\right\|}{\left\|\underline{\lambda}_k - \underline{\lambda}^*\right\|} \le \max\left[\left|1 - \alpha\omega_{\min}\right|, \left|1 - \alpha\omega_{\max}\right|\right] = r(\alpha)$$

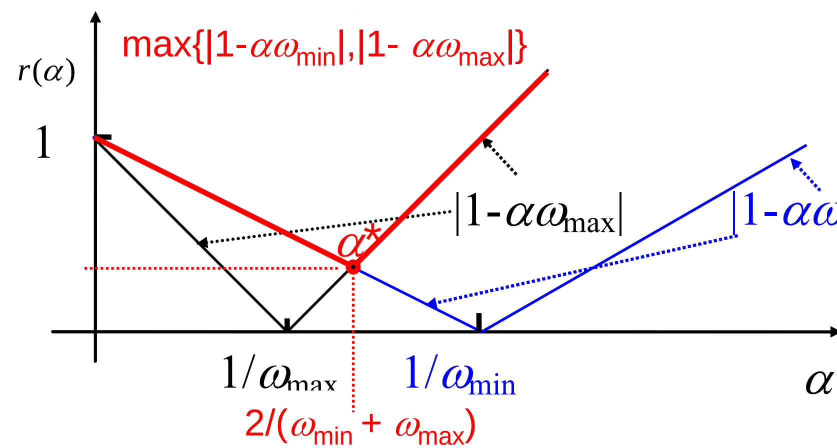
Why α = c Works? - 4

Clearly, need $0 < \alpha < 1/2\omega_{\max}$

optimal $\alpha^* \Rightarrow \left|1-\alpha^*\omega_{\min}\right| = \left|1-\alpha^*\omega_{\max}\right|$

$$\Rightarrow 1-\alpha^*\omega_{\min} = \alpha^*\omega_{\max} - 1$$

$$\Rightarrow \alpha^* = \frac{2}{(\omega_{\min}+\omega_{\max})}$$



The eigen values $\omega_{\min}$ and $\omega_{\max}$ are related to eigen values $\lambda_{\min}$ and $\lambda_{\max}$ of $(AQ^{-1}A^T)^{-1}$

since $(I + cAQ^{-1}A^T)^{-1} = I - cA(Q + cA^T A)^{-1}A^T$

$$\Rightarrow \frac{1}{1+c/\lambda_i} = 1 - c\omega_i \Rightarrow \omega_i = \frac{1}{c+\lambda_i} \Rightarrow \text{Note that } AQ^{-1}A^T \leftrightarrow \nabla \underline{\underline{h}}^T \left[\nabla_{xx}^2 L_0\right]^{-1} \nabla \underline{\underline{h}}$$

So, convergence occurs for $0 < \alpha < 2\left[c + \lambda_{\min}\right]$

$$\text{optimal } \alpha^* = \frac{2}{(\omega_{\min}+\omega_{\max})} = \frac{2(c+\lambda_{\min})(c+\lambda_{\max})}{(2c+\lambda_{\min}+\lambda_{\max})}$$

$$= 2c\left[1 - \frac{c}{(2c+\lambda_{\min}+\lambda_{\max})} + \frac{\lambda_{\min}\lambda_{\max}}{c(c+\lambda_{\min}+\lambda_{\max})}\right] \geq 2c\left[1 - \frac{c}{(2c+\lambda_{\min}+\lambda_{\max})}\right]$$

As $c \to \infty$, $\alpha^* \to c$

When is α ≠ c Better?

When is $\alpha \neq c$ better?

- Case 1: $\lambda_{\min}\left[(AQ^{-1}A^T)^{-1}\right] < 0 < \lambda_{\max}\left[(AQ^{-1}A^T)^{-1}\right]$

$$\alpha = c \Rightarrow r(c) = \max\left[\left|1-\frac{c}{\lambda_{\min}+c}\right|, \left|1-\frac{c}{\lambda_{\max}+c}\right|\right] = \max\left[\left|\frac{\lambda_{\min}}{\lambda_{\min}+c}\right|, \left|\frac{\lambda_{\max}}{\lambda_{\max}+c}\right|\right]$$

$$\alpha = \alpha^* \Rightarrow r(\alpha^*) = \left[\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\min}+\lambda_{\max}+2c}\right]$$

$$\lim_{c\to\infty}\frac{r(c)}{r(\alpha^*)} = \max\left[\frac{2|\lambda_{\min}|}{\lambda_{\max}-\lambda_{\min}}, \frac{2|\lambda_{\max}|}{\lambda_{\max}-\lambda_{\min}}\right] < 2$$

The value of $\alpha = c$ is probably OK.

- Case 2: $\lambda_{\min} \le \lambda_{\max} \le 0$ or $0 \le \lambda_{\min} \le \lambda_{\max}$

$$\Rightarrow \lim_{c\to\infty}\frac{r(c)}{r(\alpha^*)} = \max\left[\frac{2|\lambda_{\min}|}{\lambda_{\max}-\lambda_{\min}}, \frac{2|\lambda_{\max}|}{\lambda_{\max}-\lambda_{\min}}\right] \ge 2$$

$\Rightarrow$ It is worth using optimal step size. Greater improvement as $\lambda_{\max} \to \lambda_{\min}$.

- For convex problems with $0 \le \lambda_{\min} \le \lambda_{\max}$, we have

$$\alpha^* \ge 2c\left[1-\frac{c}{\lambda_{\min}+\lambda_{\max}+2c}\right] \Rightarrow c \le \alpha \le 2c$$

Newton Iteration for Dual Updates - 1

- Second Order Iteration

$$\underline{\lambda}_{k+1} = \underline{\lambda}_k + \left[\nabla^2 q(\underline{\lambda}_k)\right]^{-1} \underline{h}(\underline{x}_k^*)$$

$$\nabla^2 q(\underline{\lambda}_k) = -\nabla \underline{\underline{h}}^T(\underline{x}_k^*)\left[\nabla_{xx}^2 L_c(\underline{x}_k^*, \underline{\lambda}_k)\right]^{-1} \nabla \underline{\underline{h}}^T(\underline{x}_k^*)$$

so, $\underline{\lambda}_{k+1} = \underline{\lambda}_k + B_k^{-1} \underline{h}(\underline{x}_k^*)$,

where $B_k = -\nabla^2 q(\underline{\lambda}_k)$

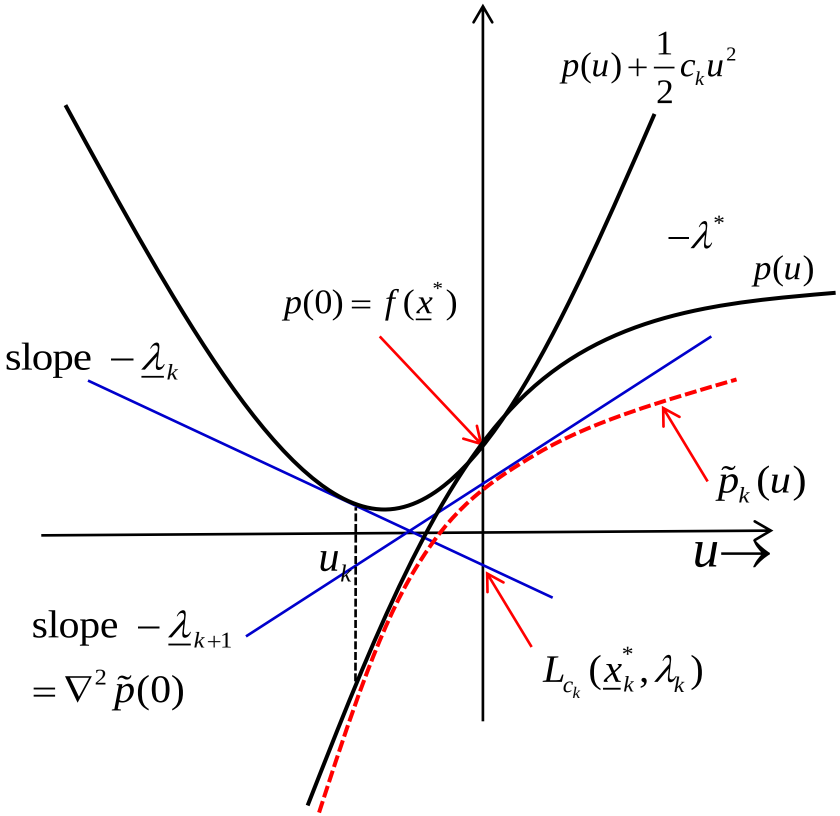
Recall that $\nabla_{xx}^2 L_c(\underline{x}_k^*, \underline{\lambda}_k) = \nabla_{xx}^2 L_0(\underline{x}_k^*, \underline{\lambda}_k + c_k \underline{h}(\underline{x}_k^*)) + c_k \nabla \underline{\underline{h}}(\underline{x}_k^*) \nabla \underline{\underline{h}}^T(\underline{x}_k^*)$

It is easy to see that

$$B_k^{-1} = \underbrace{\left[\nabla \underline{h}^T(\underline{x}_k^*)\left[\nabla_{xx}^2 L_0(\underline{x}_k^*, \underline{\lambda}_k)\right]^{-1} \nabla \underline{h}(\underline{x}_k^*)\right]^{-1}}_{\nabla^2 p(\underline{u}_k) \text{ of the primal}} + c_k I$$

Newton Iteration for Dual Updates - 2

Geometric Interpretation



$$
\begin{aligned}
\underline{\lambda}_{k+1} &= \underline{\lambda}_k + \left[\nabla^2 p(\underline{u}_k) + c_k I\right]\underline{u}_k \\
&= \underline{\lambda}_k + c_k u + \nabla^2 p(\underline{u}_k)\underline{u}_k \\
&= \underline{\tilde{\lambda}}_k + \nabla^2 p(\underline{u}_k)\underline{u}_k
\end{aligned}
$$

$$
\begin{aligned}
\tilde{p}_k(\underline{u}) = p(\underline{u}_k) &+ \nabla \underline{p}^T(\underline{u}_k)(\underline{u}-\underline{u}_k) \\
&+ \frac{1}{2}(\underline{u}-\underline{u}_k)^T \nabla^2 p(\underline{u}_k)(\underline{u}-\underline{u}_k)
\end{aligned}
$$

$$\Rightarrow \nabla \underline{\tilde{p}}_k(\underline{0}) = \nabla \underline{p}(\underline{u}_k) - \nabla^2 p(\underline{u}_k)\underline{u}_k$$

since $\underline{\tilde{\lambda}}_k = -\nabla \underline{p}(\underline{u}_k)$

$$\Rightarrow \underline{\lambda}_{k+1} = -\nabla \underline{\tilde{p}}_k(\underline{0})$$

- For Quasi-Newton version, see Bertsekas’ book

Extension to Inequality Constraints - 1

$$\min \quad f(\underline{x})$$
$$\text{s.t.} \quad \underline{h}(\underline{x}) = \underline{0}, \quad \underline{h} \in R^m \qquad \text{I}$$
$$\underline{g}(\underline{x}) \leq \underline{0}, \quad \underline{g} \in R^r$$

Convert the problem into an equality constrained problem:

$$\min \quad f(\underline{x})$$
$$\text{s.t.} \quad h_1(\underline{x}) = h_2(\underline{x}) = \ldots h_m(\underline{x}) = 0 \qquad \text{II}$$
$$g_1(\underline{x}) + z_1^2 = g_2(\underline{x}) + z_2^2 = \ldots g_r(\underline{x}) + z_r^2 = 0$$

$\underline{x}$* is a minimum of "I" if and only if

$$\underline{x}^* \text{ and } z_j^* = \left[-g_j(\underline{x}^*)\right]^{1/2}, \; j = 1, 2, \ldots, r \text{ is a minimum of "II".}$$

Primal

$$\min \quad f(\underline{x}) + \tfrac{1}{2} c \underline{h}^T(\underline{x}) \underline{h}(\underline{x})$$
$$+ \tfrac{1}{2} c \sum_{j=1}^{r} [g_i(\underline{x}) + z_j^2]^2$$
$$\text{s.t.} \quad \underline{h}(\underline{x}) = \underline{0}, \quad \underline{h} \in R^m$$
$$\underline{g}(\underline{x}) \leq \underline{0}, \quad \underline{g} \in R^r$$

Dual

$$\max_{\underline{\lambda}, \underline{\mu} \geq 0} \quad q(\underline{\mu}, \underline{\lambda})$$

$$\text{where } q(\underline{\mu}, \underline{\lambda}) = \min_{\underline{z}, \underline{x}} \{ f(\underline{x}) + \sum_{j=1}^{r} \mu_j \left[g_j(\underline{x}) + z_j^2 \right]$$
$$+ \frac{c}{2} (g_j(\underline{x}) + z_j^2)^2 + \underline{\lambda}^T \underline{h}(\underline{x}) + \frac{c}{2} \underline{h}^T(\underline{x}) \underline{h}(\underline{x}) \}$$
$$= \min_{\underline{z}, \underline{x}} L_c(\underline{x}, \underline{z}, \underline{\mu}, \underline{\lambda})$$

Extension to Inequality Constraints - 2

Define $v_j = z_j^2$, then

$$q(\underline{\mu},\underline{\lambda}) = \min_{\underline{v}\ge 0,\underline{x}} \{ f(\underline{x}) + \underline{\mu}^T(\underline{g}(\underline{x})+\underline{v}) + \frac{c}{2}(\underline{g}(\underline{x})+\underline{v})^T(\underline{g}(\underline{x})+\underline{v})$$

$$+\underline{\lambda}^T \underline{h}(\underline{x}) + \frac{c}{2}\underline{h}^T(\underline{x})\underline{h}(\underline{x})\}$$

Key: For each fixed $\underline{x}$, can find $\underline{v}^*$ explicity.

since $\underline{v}$ enters through only $\underline{\mu}^T(\underline{g}(\underline{x})+\underline{v}) + \frac{c}{2}(\underline{g}(\underline{x})+\underline{v})^T(\underline{g}(\underline{x})+\underline{v})$

in $L_c(\underline{x},\underline{z},\underline{\mu},\underline{\lambda})$, we have

$$\min_{\underline{v}\ge 0} \underline{\mu}^T(\underline{g}(\underline{x})+\underline{v}) + \frac{c}{2}(\underline{g}(\underline{x})+\underline{v})^T(\underline{g}(\underline{x})+\underline{v})$$

$$= \sum_{j=1}^{r} \mu_j\left[g_j(\underline{x})+v_j\right] + \frac{c}{2}(g_j(\underline{x})+v_j)^2 = \sum_{j=1}^{r} P_j$$

$\Rightarrow$ separable

$$\Rightarrow v_j^* > 0 \Rightarrow \frac{dP_j}{dv_j} = 0$$

$$v_j^* = 0 \Rightarrow \frac{dP_j}{dv_j} \ge 0$$

Extension to Inequality Constraints - 3

$$\text{The unconstrained minimum} = \hat{v}_j = -\left[\frac{\mu_j}{c} + g_j(\underline{x})\right]$$

$$\Rightarrow v_j^* = \max(0, -\frac{\mu_j}{c} - g_j(\underline{x}))$$

$$g_j(\underline{x}) + \underline{v}_j^* = \max(g_j(\underline{x}), -\frac{\mu_j}{c})$$

If we define

$$g_j^+(\underline{x}, \underline{\mu}_j, c) = g_j(\underline{x}) + v_j^* = \max(g_j(\underline{x}), -\frac{\mu_j}{c})$$

$$\text{and } \underline{g}^+(\underline{x}, \underline{\mu}, c) = \begin{bmatrix} g_1^+(\underline{x}, \mu_1, c) \\ g_2^+(\underline{x}, \mu_2, c) \\ \vdots \\ g_r^+(\underline{x}, \mu_r, c) \end{bmatrix}$$

$$L_c(\underline{x}, \underline{\mu}, \underline{\lambda}) = \min_z L_c(\underline{x}, z, \underline{\mu}, \underline{\lambda}) = f(\underline{x}) + \underline{\lambda}^T \underline{h}(\underline{x}) + \frac{c}{2} \underline{h}^T(\underline{x}) \underline{h}(\underline{x})$$

$$+ \underline{\mu}^T g^+(\underline{x}, \underline{\mu}, c) + \frac{c}{2} \underline{g}^{+T}(\underline{x}, \underline{\mu}, c) \underline{g}^+(\underline{x}, \underline{\mu}, c)$$

Extension to Inequality Constraints - 4

- Alternative form for $L_c(\underline{x}, \underline{\mu}, \underline{\lambda})$

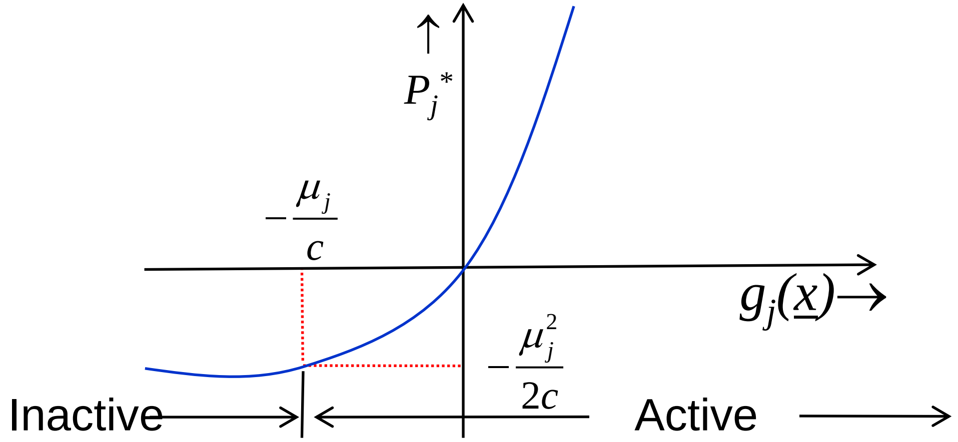
$$\left.\begin{aligned} \text{Suppose } v_j^* = 0 \Rightarrow P_j^* &= \mu_j g_j(\underline{x}) + \frac{c}{2} g_j^2(\underline{x}) \\ &= \frac{1}{2c}\left[2\mu_j c g_j(\underline{x}) + c^2 g_j^2(\underline{x})\right] \\ &= \frac{1}{2c}\left[(\mu_j + c g_j(\underline{x}))^2 - \mu_j^2\right] \end{aligned}\right\} \Rightarrow \text{Active constraints}$$

$$\text{Suppose } v_j^* > 0 \Rightarrow g_j(\underline{x}) + v_j^* = -\frac{\mu_j}{c}$$

$$\left.\begin{aligned} &\Rightarrow P_j^* = \mu_j\left[g_j(\underline{x}) + v_j^*\right]^2 + \frac{c}{2}\left[g_j(\underline{x}) + v_j^*\right]^2 \\ &= -\frac{\mu_j^2}{c} + \frac{\mu_j^2}{2c} = -\frac{\mu_j^2}{2c} \end{aligned}\right\} \Rightarrow \text{Inactive constraints}$$

Extension to Inequality Constraints - 5

Combining the terms $P_j^* = \frac{1}{2c}\left[\{\max(0, \mu_j + c g_j(\underline{x}))\}^2 - \mu_j^2\right]$



Result: $L_c(\underline{x}, \underline{\mu}, \underline{\lambda}) = f(\underline{x}) + \underline{\lambda}^T \underline{h}(\underline{x}) + \frac{c}{2}\underline{h}^T(\underline{x})\underline{h}(\underline{x})$

$$+\frac{1}{2c}\sum_{j=1}^{r}\left(\left[\max(0, \mu_j + c g_j(\underline{x}))\right]^2 - \mu_j^2\right)$$

$$q(\underline{\mu}, \underline{\lambda}) = \min_{\underline{x}} L_c(\underline{x}, \underline{\mu}, \underline{\lambda}) = L_c(\underline{x}^*, \underline{\mu}, \underline{\lambda})$$

- Similar to equality constrained problem
- $\mu_j \geq 0$ and <u>penalty function depends on $\underline{\mu}$</u>

Extension to Inequality Constraints - 6

- Steepest Ascent Iteration

$$\underline{\lambda}_{k+1} = \underline{\lambda}_k + c_k \underline{h}(\underline{x}_k^*)$$

$$\underline{\mu}_{k+1} = \underline{\mu}_k + c_k \underline{g}^+(\underline{x}_k^*, \underline{\mu}_k, c)$$

$$\text{or } (\underline{\mu}_{k+1})_j = (\underline{\mu}_k)_j + c_k \max\left[g_j(\underline{x}_k^*, \underline{\mu}_k, c), -\frac{(\underline{\mu}_k)_j}{c} \right]$$

$$= \max(0, (\underline{\mu}_k)_j + c_k g_j(\underline{x}_k^*, \mu_j, c))$$

$\Rightarrow$ Note that any inequality constraint such that $g_j(\underline{x}_k^*, \mu_j, c) < -\frac{\mu_j}{c}$

is automatically inactive; else, it is active.

Extension to Inequality Constraints - 7

- Newton Iteration

$$\nabla^2 q(\underline{\lambda}, \underline{\mu}) = \begin{bmatrix} -B & 0 \\ 0 & -\frac{1}{c} I \end{bmatrix} \begin{matrix} \text{Active + equality} \\ \text{Inactive} \end{matrix}$$

$$B = N^T \left[\nabla_{xx} L_c(\underline{x}_k^*, \underline{\lambda}, \underline{\mu}) \right]^{-1} N$$

$$N = \left[\underbrace{\nabla \underline{h}_1 \nabla \underline{h}_2 \ldots \nabla \underline{h}_m}_{\text{Equality}} \qquad \underbrace{\nabla \underline{g}_1 \nabla \underline{g}_2 \ldots \nabla \underline{g}_p}_{\text{Active Inequality}} \right]$$

$$\underline{g}^+ = \begin{bmatrix} \underline{g}_{active} \\ \underline{g}_{inactive} \end{bmatrix}$$

$$\underline{g}_{inactive} = \begin{bmatrix} -\frac{\mu_{p+1}}{c_k} \\ \cdot \\ \cdot \\ -\frac{\mu_r}{c_k} \end{bmatrix};$$

$$\underline{h}_{active} = \begin{bmatrix} \underline{h} \\ \underline{g}_{active} \end{bmatrix}$$

$$\underline{\mu} = \begin{bmatrix} \underline{\mu}_{active} \\ \underline{\mu}_{inactive} \end{bmatrix}; \underline{\lambda}_{active} = \begin{bmatrix} \underline{\lambda} \\ \underline{\mu}_{active} \end{bmatrix}$$

Newton Iteration

$$\begin{bmatrix} \underline{\lambda}_{active} \\ \underline{\mu}_{inactive} \end{bmatrix} = \begin{bmatrix} \underline{\lambda}_{active} \\ \underline{\mu}_{inactive} \end{bmatrix} + \begin{bmatrix} B^{-1} & 0 \\ 0 & c \end{bmatrix} \begin{bmatrix} \underline{h}_{active} \\ \underline{g}_{inactive} \end{bmatrix} = \begin{bmatrix} \underline{\lambda}_{active} + B^{-1} \underline{h}_{active} \\ \underline{0} \end{bmatrix}$$

$\Rightarrow$ set multipliers of inactive constraints to 0

$\Rightarrow$ Treat active constraints as if they are equality constraints.

Saddle Point Theorem - 1

- Saddle Point Theorem for Inequality Constrained NLP Problem

$$\left.\begin{array}{l}\min f(\underline{x}) \\ \text{s.t.}\quad g_j(\underline{x}) \le 0 \qquad j = 1, 2, \ldots\ldots, r \qquad \underline{x} \in \Omega\end{array}\right\}$$

Same for equality constraints except λ^s have unrestricted sign

- Lagrangian: $L(\underline{x}, \underline{\mu}) = f(\underline{x}) + \sum_{j=1}^{r} \mu_j g_j(\underline{x})$

  Also, $\min_{\underline{x}\in\Omega} L(\underline{x}, \underline{\mu}^*) = f(\underline{x}^*)$ since $\mu_j^* g_j(\underline{x}^*) = 0$

- Saddle Point Theorem:

$$\boxed{L(\underline{x}^*, \underline{\mu}) \le L(\underline{x}^*, \underline{\mu}^*) \le L(\underline{x}, \underline{\mu}^*) \qquad \forall \underline{x} \in \Omega \text{ and } \underline{\mu} \ge 0}$$

$$L(\underline{x}^*, \underline{\mu}^*) = f(\underline{x}^*) = \min_{\underline{x}\in\Omega} L(\underline{x}, \underline{\mu}^*) \le L(\underline{x}, \underline{\mu}^*)$$

$$= f(\underline{x}) + \sum_{j=1}^{r} \mu_j^* g_j(\underline{x}) \le f(\underline{x}) \text{ since } g_j(\underline{x}) \le 0$$

Also,

$$L(\underline{x}^*, \underline{\mu}) = f(\underline{x}^*) + \sum_{j=1}^{r} \mu_j g_j(\underline{x}^*) \le f(\underline{x}^*) = L(\underline{x}^*, \underline{\mu}^*) \text{ since } \mu_j \ge 0$$

$$f(\underline{x}^*) = \min_{\underline{x}\in\Omega} \max_{\underline{\mu}\ge 0} L(\underline{x}, \underline{\mu})$$

Minimax theorem

Saddle Point Theorem - 2

$$\text{since } \max_{\underline{\mu}\geq 0} L(\underline{x},\underline{\mu}) = \max_{\underline{\mu}\geq 0}\left\{ f(\underline{x}) + \sum_{j=1}^{r}\mu_j g_j(\underline{x})\right\} = \begin{cases} f(x) \text{ if } g_j(\underline{x}) \leq 0 \\ \qquad\qquad j = 1, 2, \ldots, r \\ \infty \quad \text{Otherwise} \end{cases}$$

- If we let

$$q^* = \max_{\underline{\mu}\geq 0} \min_{\underline{x}\in\Omega} L(\underline{x},\underline{\mu}) \leq \max_{\underline{\mu}} L(\underline{z},\underline{\mu}) = L(\underline{z},\underline{\mu}^*) \;\; ; \;\; \underline{z} \in \Omega$$

Taking min over $\underline{z}$

$$\left.\begin{aligned} q^* &\leq f^* \\ \text{Also, } \quad f^* &= \min_{\underline{x}\in\Omega} L(\underline{x},\underline{\mu}^*) \leq q^* \end{aligned}\right\} \Rightarrow f^* = q^*$$

Summary

- Constrained Optimization Methods
- Penalty Methods
- Multiplier (Augmented Lagrangian) Methods
- Duality and Convergence Issues
- Extensions to Inequality Constraints
- Illustrative Examples