Lecture 9: Penalty Methods and Multiplier Methods

- Quadratic Penalty Methods
- Introduction to Multiplier Methods

Two Convergence Mechanisms

Consider the equality constrained problem

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in X, \ \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}, \end{array}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $\boldsymbol{h} : \mathbb{R}^n \mapsto \mathbb{R}^m$ are continuous, and X is closed.

- The quadratic penalty method:

$$\boldsymbol{x}^r = \arg\min_{\boldsymbol{x} \in X} L_{c^r}(\boldsymbol{x}, \boldsymbol{\lambda}^r) \equiv f(\boldsymbol{x}) + (\boldsymbol{\lambda}^r)'\boldsymbol{h}(\boldsymbol{x}) + \frac{c^r}{2}\|\boldsymbol{h}(\boldsymbol{x})\|^2$$

  where the $\boldsymbol{\lambda}^r$ is a bounded sequence and c^r satisfies $0 < c^r < c^{r+1}$ for all r and $c^r \to \infty$.

- Mechanism 1 for convergence: taking $\boldsymbol{\lambda}^r$ close to a Lagrange multiplier vector
  - ⋆ Assume $X = \mathbb{R}^n$ and $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ is a local min-Lagrange multiplier pair satisfying the 2nd order sufficiency conditions

- ⋆ For c sufficiently large, $\boldsymbol{x}^*$ is a strict local min of $L_c(\cdot, \boldsymbol{\lambda}^*)$

- Mechanism 2 for convergence: Taking c^r very large

  - ⋆ For large c and any $\boldsymbol{\lambda}$, we have

$$L_c(\cdot, \boldsymbol{\lambda}) \approx \begin{cases} f(\boldsymbol{x}) & \text{if } \boldsymbol{x} \in X \text{ and } \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0} \\ \infty & \text{otherwise} \end{cases}$$

- Example:

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) = \frac{1}{2}(x_1^2 + x_2^2) \\ \text{subject to} & x_1 = 1 \end{array}$$

We have $\boldsymbol{x}^* = (1, 0), \ \lambda^* = -1$ and

$$L_c(\boldsymbol{x}, \lambda) = \frac{1}{2}(x_1^2 + x_2^2) + \lambda(x_1 - 1) + \frac{c}{2}(x_1 - 1)^2$$

$$x_1(\lambda, c) = \frac{c - \lambda}{c + 1}, \quad x_2(\lambda, c) = 0$$

Global Convergence

- Suppose $c^r \to \infty$. Then every limit point of $\{\boldsymbol{x}^r\}$ is a global min.
- **Proof:** The optimal value of the problem is

$$f^* = \inf_{\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{0}, \boldsymbol{x}\in X} L_{c^r}(\boldsymbol{x}, \boldsymbol{\lambda}^r).$$

We have $L_{c^r}(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) \le L_{c^r}(\boldsymbol{x}, \boldsymbol{\lambda}^r), \quad \forall\, \boldsymbol{x} \in X$ so taking the inf of the RHS over $x \in X,\ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}$ yields

$$L_{c^r}(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) = f(\boldsymbol{x}^r) + (\boldsymbol{\lambda}^r)'\boldsymbol{h}(\boldsymbol{x}^r) + \frac{c^r}{2}\|\boldsymbol{h}(\boldsymbol{x}^r)\|^2 \le f^*$$

Let $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\lambda}})$ be a limit point of $\{\boldsymbol{x}^r, \boldsymbol{\lambda}^r\}$. Without loss of generality, assume that $\{\boldsymbol{x}^r, \boldsymbol{\lambda}^r\} \to (\bar{\boldsymbol{x}}, \bar{\boldsymbol{\lambda}})$. Taking the limsup above

$$f(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{\lambda}}'\boldsymbol{h}(\bar{\boldsymbol{x}}) + \limsup_{r\to\infty} \frac{c^r}{2}\|\boldsymbol{h}(\boldsymbol{x}^r)\|^2 \le f^*$$

By $\|\boldsymbol{h}(\boldsymbol{x}^r)\|^2 \ge 0$ and $\{c^r\} \to \infty$, we have $\boldsymbol{h}(\boldsymbol{x}^r) \to \boldsymbol{0}$ and $\boldsymbol{h}(\bar{\boldsymbol{x}}) = \boldsymbol{0}$. Hence, $\bar{\boldsymbol{x}}$ is feasible, and since the above inequality implies $f(\bar{\boldsymbol{x}}) \le f^*$, so $\bar{\boldsymbol{x}}$ is optimal.

Lagrange Multiplier Estimates

- Assume that $X = \mathbb{R}^n$, and f and h are continuously differentiable. Let $\{\boldsymbol{\lambda}^r\}$ be bounded, and $\{c^r\} \to \infty$. Assume $\boldsymbol{x}^r$ satisfies $\nabla_{\boldsymbol{x}} L_{c^r}(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) = 0$ for all r, and that $\boldsymbol{x}^r \to \boldsymbol{x}^*$, where $\boldsymbol{x}^*$ is such that $\text{rank}(\nabla \boldsymbol{h}(\boldsymbol{x}^*)) = m$. Then $\boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0}$ and $\tilde{\boldsymbol{\lambda}}^r \to \boldsymbol{\lambda}^*$, where

$$\tilde{\boldsymbol{\lambda}}^r = \boldsymbol{\lambda}^r + c^r \boldsymbol{h}(\boldsymbol{x}^r), \quad \nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}.$$

- **Proof:** We have

$$\mathbf{0} = \nabla_{\boldsymbol{x}} L_{c^r}(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) = \nabla f(\boldsymbol{x}^r) + \nabla \boldsymbol{h}(\boldsymbol{x}^r) \left(\boldsymbol{\lambda}^r + c^r \boldsymbol{h}(\boldsymbol{x}^r)\right) = \nabla f(\boldsymbol{x}^r) + \nabla \boldsymbol{h}(\boldsymbol{x}^r) \tilde{\boldsymbol{\lambda}}^r.$$

- Multiply with

$$\left(\nabla \boldsymbol{h}(\boldsymbol{x}^r)' \nabla \boldsymbol{h}(\boldsymbol{x}^r)\right)^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}^r)'$$

and take lim to obtain $\tilde{\boldsymbol{\lambda}}^r \to \boldsymbol{\lambda}^*$ with

$$\boldsymbol{\lambda}^* = -\left(\nabla \boldsymbol{h}(\boldsymbol{x}^*)' \nabla \boldsymbol{h}(\boldsymbol{x}^*)\right)^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}^*)' \nabla f(\boldsymbol{x}^*).$$

We also have $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}$ and $\boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0}$ (since $\tilde{\boldsymbol{\lambda}}^r$ converges).

Practical Behaviors

- Three possibilities:
  - ⋆ The method breaks down because an $\boldsymbol{x}^r$ with $\nabla_{\boldsymbol{x}} L_{c^r}(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) \approx \mathbf{0}$ cannot be found.
  - ⋆ A sequence $\{\boldsymbol{x}^r\}$ with $\nabla_{\boldsymbol{x}} L_{c^r}(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) \approx \mathbf{0}$ is obtained, but it either has no limit points, or for each of its limit points $\boldsymbol{x}^*$ the matrix $\nabla \boldsymbol{h}(\boldsymbol{x}^*)$ has rank $< m$.
  - ⋆ A sequence $\{\boldsymbol{x}^r\}$ with with $\nabla_{\boldsymbol{x}} L_{c^r}(\boldsymbol{x}^r, \boldsymbol{\lambda}^r) \approx \mathbf{0}$ is found and it has a limit point $\boldsymbol{x}^*$ such that $\nabla \boldsymbol{h}(\boldsymbol{x}^*)$ has rank m. Then, $\boldsymbol{x}^*$ together with $\boldsymbol{\lambda}^*$ [the corresponding limit point of $\{\boldsymbol{\lambda}^r + c^r \boldsymbol{h}(\boldsymbol{x}^r)\}$] satisfies the first-order necessary conditions.

- Ill-conditioning: The condition number of the Hessian $\nabla^2_{\boldsymbol{x}\boldsymbol{x}} L_{c^r}(\boldsymbol{x}^r, \boldsymbol{\lambda}^r)$ tends to increase with c^r.

- To overcome ill-conditioning:

- ⋆ Use Newton-like method (and double precision).
- ⋆ Use good starting points.
- ⋆ Increase $\{c^r\}$ at a moderate rate (if $\{c^r\}$ is increased at a fast rate, $\{\boldsymbol{x}^r\}$ converges faster, but the likelihood of ill-conditioning is greater).

Inequality Constraints

- Convert them to equality constraints by using squared slack variables that are eliminated later.

- Convert inequality constraint $g_j(\boldsymbol{x}) \le 0$ to equality constraint $g_j(\boldsymbol{x}) + z_j^2 = 0$.

- The penalty method solves problems of the form

$$\min_{\boldsymbol{x},\boldsymbol{z}} \bar{L}_c(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{\mu}) = L_c(\boldsymbol{x},\boldsymbol{\lambda}) + \sum_{j=1}^{r} \left[\mu_j \left(g_j(\boldsymbol{x}) + z_j^2\right) + \frac{c}{2}|g_j(\boldsymbol{x}) + z_j^2|^2\right],$$

for various values of $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and c.

- First minimize $\bar{L}_c(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{\mu})$ with respect to $\boldsymbol{z}$ to compute $L_c(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})$ by

$$\min_{\boldsymbol{z}} \bar{L}_c(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda},\boldsymbol{\mu}) = L_c(\boldsymbol{x},\boldsymbol{\lambda}) + \sum_{j=1}^{r} \min_{z_j} \left[\mu_j \left(g_j(\boldsymbol{x}) + z_j^2\right) + \frac{c}{2}|g_j(\boldsymbol{x}) + z_j^2|^2\right]$$

and then minimize $L_c(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu})$ with respect to $\boldsymbol{x}$.

Multiplier Methods

- Recall that if $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ is a local min-Lagrange multiplier pair satisfying the 2nd order sufficiency conditions, then for c sufficiently large, $\boldsymbol{x}^*$ is a strict local min of $L_c(\cdot, \boldsymbol{\lambda}^*)$.

- This suggests that for $\boldsymbol{\lambda}^r \approx \boldsymbol{\lambda}^*$, $\boldsymbol{x}^r \approx \boldsymbol{x}^*$.

- Hence it is a good idea to use $\boldsymbol{\lambda}^r \approx \boldsymbol{\lambda}^*$, such as

$$\boldsymbol{\lambda}^{r+1} = \tilde{\boldsymbol{\lambda}}^r = \boldsymbol{\lambda}^r + c^r \boldsymbol{h}(\boldsymbol{x}^r)$$

  This is the (1st order) method of multipliers.

- Key advantages to be shown:

  ⋆ Less ill-conditioning: It is not necessary that $c^r \to \infty$ (only that c^r exceeds some threshold).

- ⋆ Faster convergence when $\boldsymbol{\lambda}^r$ is updated than when $\boldsymbol{\lambda}^r$ is kept constant (whether $c^r \to \infty$ or not).

- Consider the equality constrained problem

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \end{array}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $\boldsymbol{h} : \mathbb{R}^n \mapsto \mathbb{R}^m$ are continuously differentiable.

- The (1st order) multiplier method finds

$$\boldsymbol{x}^r = \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} L_{c^r}(\boldsymbol{x}, \boldsymbol{\lambda}^r) \equiv f(\boldsymbol{x}) + (\boldsymbol{\lambda}^r)'\boldsymbol{h}(\boldsymbol{x}) + \frac{c^r}{2}\|\boldsymbol{h}(\boldsymbol{x})\|^2$$

and updates $\boldsymbol{\lambda}^r$ using

$$\boldsymbol{\lambda}^{r+1} = \boldsymbol{\lambda}^r + c^r \boldsymbol{h}(\boldsymbol{x}^r)$$

Convex Example

- Problem: $\min_{x_1=1} = \frac{1}{2}(x_1^2 + x_2^2)$ with optimal solution $\boldsymbol{x}^* = (1,0)$ and Lagrangian multiplier $\lambda^* = -1$.

- We have

$$\boldsymbol{x}^r = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} L_{c^r}(\boldsymbol{x}, \lambda^r) = \left(\frac{c^r - \lambda^r}{c^r + 1}, 0\right)$$
$$\lambda^{r+1} = \lambda^r + c^r \left(\frac{c^r - \lambda^r}{c^r + 1} - 1\right)$$
$$\lambda^{r+1} - \lambda^* = \frac{\lambda^r - \lambda^*}{c^r + 1}$$

- We see that:
  - ⋆ $\lambda^r \to \lambda^* = -1$ and $\boldsymbol{x}^r \to \boldsymbol{x}^* = (1,0)$ for every nondecreasing sequence $\{c^r\}$. It is NOT necessary to increase c^r to ∞.
  - ⋆ The convergence rate becomes faster as c^r becomes larger; in fact $\{|\lambda^r - \lambda^*|\}$ converges superlinearly if $c^r \to \infty$.

Nonconvex Example

- Problem: $\min_{x_1=1} = \frac{1}{2}(-x_1^2 + x_2^2)$ with optimal solution $\boldsymbol{x}^* = (1,0)$ and Lagrangian multiplier $\lambda^* = 1$.

- We have

$$\boldsymbol{x}^r = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} L_{c^r}(\boldsymbol{x}, \lambda^r) = \left(\frac{c^r - \lambda^r}{c^r - 1}, 0\right)$$

provided $c^r > 1$ (otherwise the min does not exist)

$$\lambda^{r+1} = \lambda^r + c^r\left(\frac{c^r-\lambda^r}{c^r-1} - 1\right)$$
$$\lambda^{r+1} - \lambda^* = -\frac{\lambda^r-\lambda^*}{c^r-1}$$

- We see that:
  - ⋆ No need to increase c^r to ∞ for convergence; doing so results in faster convergence rate.
  - ⋆ To obtain convergence, c^r must eventually exceed the threshold 2.

Primal Functional

- Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ be a regular local min-Lagrangian pair satisfying the 2nd order sufficient conditions are satisfied.

- The primal functional

$$p(u) = \min_{\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{u}} f(\boldsymbol{x}),$$

defined for $\boldsymbol{u}$ in an open sphere centered at $\boldsymbol{u} = \mathbf{0}$, and we have

$$p(\mathbf{0}) = f(\boldsymbol{x}^*), \quad \nabla p(\mathbf{0}) = -\boldsymbol{\lambda}^*.$$

- Two examples:

$$p(u) = \min_{x_1 - 1 = u} \frac{1}{2}(x_1^2 + x_2^2) = \frac{1}{2}(1+u)^2,\ p(0) = f(\boldsymbol{x}^*) = \frac{1}{2},\ p'(0) = 1 = -\lambda^*$$

and

$$p(u) = \min_{x_1 - 1 = u} \frac{1}{2}(-x_1^2 + x_2^2) = -\frac{1}{2}(1+u)^2,\ p'(0) = -1 = -\lambda^*$$

Augmented Lagrangian Minimization

- Break down the minimization of $L_c(\boldsymbol{x}, \boldsymbol{\lambda})$:

$$\begin{aligned} \min_{\boldsymbol{x}} L_c(\boldsymbol{x}, \boldsymbol{\lambda}) &= \min_{\boldsymbol{u}} \min_{\boldsymbol{h}(\boldsymbol{x})=\boldsymbol{u}} \left\{ f(\boldsymbol{x}) + \boldsymbol{\lambda}'\boldsymbol{h}(\boldsymbol{x}) + \frac{c}{2}\|\boldsymbol{h}(\boldsymbol{x})\|^2 \right\} \\ &= \min_{\boldsymbol{u}} \left\{ p(\boldsymbol{u}) + \boldsymbol{\lambda}'\boldsymbol{u} + \frac{c}{2}\|\boldsymbol{u}\|^2 \right\}, \end{aligned}$$

  where the minimization above is understood to be local in a neighborhood of $\boldsymbol{u} = \mathbf{0}$.

- Interpretation of this minimization: Penalized Primal Function $p(\boldsymbol{u}) + \frac{c}{2}\|\boldsymbol{u}\|^2$

- If c is sufficiently large, $p(\boldsymbol{u}) + \boldsymbol{\lambda}'\boldsymbol{u} + \frac{c}{2}\|\boldsymbol{u}\|^2$ is convex in a neighborhood of $\mathbf{0}$. Also, for $\boldsymbol{\lambda} \approx \boldsymbol{\lambda}^*$ and large c, the value $\min_{\boldsymbol{x}} L_c(\boldsymbol{x}, \boldsymbol{\lambda}) \approx p(\mathbf{0}) = f(\boldsymbol{x}^*)$.

Interpretation of This Method

- Geometric interpretation of the iteration

$$\boldsymbol{\lambda}^{r+1} = \boldsymbol{\lambda}^r + c^r \boldsymbol{h}(\boldsymbol{x}^r).$$

- If $\boldsymbol{\lambda}^r$ is sufficiently close to $\boldsymbol{\lambda}^*$ and/or c^r is sufficiently large, $\boldsymbol{\lambda}^{r+1}$ will be closer to $\boldsymbol{\lambda}^*$ than $\boldsymbol{\lambda}^r$.

- c^r need not be increased to ∞ in order to obtain convergence; it is sufficient that c^r eventually exceeds some threshold level.

- If $p(\boldsymbol{u})$ is linear, convergence to $\boldsymbol{\lambda}^*$ will be achieved in one iteration.

Computational Aspects

- Key issue is how to select $\{c^r\}$.
- c^r should eventually become larger than the "threshold" of the given problem.
- c^0 should not be so large as to cause ill-conditioning at the 1st minimization.
- c^r should not be increased so fast that too much ill-conditioning is forced upon the unconstrained minimization too early.
- c^r should not be increased so slowly that the multiplier iteration has poor convergence rate.
- A good practical scheme is to choose a moderate value c^0, and use $c^{r+1} = \beta c^r$, where $\beta > 1$ is a scalar (typically $\beta \in [5, 10]$ if a Newton like method is used).
- In practice the minimization of $L_{c^r}(\boldsymbol{x}, \boldsymbol{\lambda}^r)$ is typically inexact (usually exact asymptotically). In some variants of the method, only one Newton step per minimization is used (with safeguards).

Duality Framework

- Consider the problem

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) + \frac{c}{2}\|\boldsymbol{h}(\boldsymbol{x})\|^2 \\ \text{subject to} & \|\boldsymbol{x} - \boldsymbol{x}^*\| < \epsilon, \ \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \end{array}$$

  where ϵ is small enough for a local analysis to hold based on the implicit function theorem, and c is large enough for the minimum to exist.

- Consider the dual function and its gradient

$$\begin{aligned} q_c(\boldsymbol{\lambda}) &= \min_{\|\boldsymbol{x}-\boldsymbol{x}^*\|<\epsilon} L_c(\boldsymbol{x}, \boldsymbol{\lambda}) = L_c\left(\boldsymbol{x}(\boldsymbol{\lambda}, c), \boldsymbol{\lambda}\right), \\ \nabla q_c(\boldsymbol{\lambda}) &= \nabla_{\boldsymbol{\lambda}} \boldsymbol{x}(\boldsymbol{\lambda}, c) \nabla_{\boldsymbol{x}} L_c\left(\boldsymbol{x}(\boldsymbol{\lambda}, c), \boldsymbol{\lambda}\right) + \boldsymbol{h}\left(\boldsymbol{x}(\boldsymbol{\lambda}, c)\right) = \boldsymbol{h}\left(\boldsymbol{x}(\boldsymbol{\lambda}, c)\right) \end{aligned}$$

  We have $\nabla q_c(\boldsymbol{\lambda}^*) = \boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\nabla^2 q_c(\boldsymbol{\lambda}^*) \succ 0$.

- The multiplier method is a steepest ascent iteration for maximizing q_{c^r}

$$\boldsymbol{\lambda}^{r+1} = \boldsymbol{\lambda}^r + c^r \nabla q_{c^r}(\boldsymbol{\lambda}^r).$$