UIUC IE510 Applied Nonlinear Programming

Lecture 13: Lagrangian Multiplier Algorithms

Ruoyu Sun

From Opt over convex sets to Duality

"Graph" Summary of Last 5 Lectures

3~5 mins.

Can you summarize major contents in one page, with graphs?

$\min f(x)$ s.t. $x \in X$

no explicit → $\min f(x)$ s.t. $x \in X$, X convex → optimal condition $\langle \nabla f(x^*), x - x^* \rangle \geq 0$, $\forall x \in X$.

Gradient projection → fixed point.

→ feasible direction proof

explicit

equality → $\min f(x)$ s.t. $h(x) = 0$ → 1st order condition: $\nabla f(x) + \lambda \nabla h(x) = 0$.

2nd order.

→ penalty proof

inequality → $\min f(x)$ s.t. $g(x) \leq 0$ → 1st order: $\nabla f(x) + \mu \nabla g(x) = 0$

$\mu \geq 0 \perp g(x) \leq 0$.

2nd order.

transform → equality.

Convex → Assume f, g convex, h linear. → 1st order ⇒ global-min

⇕

strong duality. (different perspective); weak duality

(3 conditions equivalent:
global-min;
KKT condition.
$x^* = \arg\min L(x, \lambda^*)$
(duality).)

"Linear" Summary of Last 5 Lectures

- Optimality condition for optimization over convex sets
  - Inequality condition
  - Gradient projection method

- KKT condition
  - Equality case: Lagrangian multipliers/functions; 1st/2nd order conditions
  - Inequality case: complementarity
  - Proofs: feasible direction; penalty

- Duality
  - Motivation: convex case
  - Dual problem; max-min and min-max
  - Weak duality; strong duality

min max $\geq$ max min

What Problems Can We Solve Till Now?

Simple constraints: Gradient Projection

Apply to: Simplex, ball, bounds

SVM: Dual Coordinate Ascent

Apply to: when dual problems have simple constraints

How to solve more general problems?

This Lecture

- Today: penalty method, multiplier method and barrier method
- After this lecture, you should be able to
  - Desribe two convergence mechanisms
  - Tuesday Apply quadratic penalty method and ALM (augmented Lagrangian method) to solve a constrained problem
  - Tell the pros and cons of the two methods

Thursday

Unconstrained.

Optimality conditions. → Algorithm.

GD

Convergence: Convergence rate.

Momentum

CD/SGD
decompn

BFGS/BB
2nd order

Constrained.

Opt conditions. → Algorithm → Basic algorithm.

Advanced algo

Outline

Overview of Algorithms

One common idea of solving constrained problem is: transfer to unconstrained problems.

- Replace the original problem by a sequence of subproblems, in which constraints are represented by terms added to the objective
- There're different ways to represent the constraints, leading to different algorithms
- **Examples**
  - Quadratic penalty: adds a multiple of square of violation of each constraint to the objective
  - Method of multiplier: explicit Lagrangian multipliers estimate are used together with quadratic penalty

Quadratic Penalty Method

Consider the equality constrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \quad h(x) = 0, \end{aligned} \tag{1}$$

where $f : R^n \mapsto R$ and $h = (h_1, \ldots, h_m) : R^n \mapsto R^m$ are continuous, and X is closed.

Augmented Lagrangian function:

$$L_c(x, \lambda) = \overbrace{f(x) + \lambda^T h(x)}^{\text{Lagrangian function}} + \frac{c}{2} \|h(x)\|^2 .$$

- The quadratic penalty method:

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

where λ^r is any bounded sequence and c^r satisfies $0 < c^r < c^{r+1}$ for all r and $c^r \to \infty$.

Two Extreme Cases

Our purpose: by minimizing L we can solve (1) (*).

Case 1: $\lambda = 0$.

- $L(x, 0) = f(x) + \frac{c}{2} h(x)^2$.
- When is (*) possible?
- Recall Page 17 of Lec 11a, Proof Method 2 by Penalty Method: as $c \to \infty$, local-mins of L converge to a local-min of (1).

Case 2: $c = 0$.

- $L(x, \lambda) = f(x) + \lambda^T h(x)$.
- When is (*) possible?
- Recall 3rd condition of Prop 12.1: when $\lambda = \lambda^*$, x^* is a regular global-min of $L(x, \lambda^*)$, then x^* is a global-min of (1).

We know:

$x^* = \arg\min_x L(x, \lambda^*)$

i.e. $x^* = \arg\min_x L(x, \lambda)$ when $\lambda = \lambda^*$.

Want:

min $L \Rightarrow$ min f s.t. $g = 0$.

$\Rightarrow$ global-min of (1)

Two Extreme Cases

Our purpose: by minimizing L we can solve (1) (*).

Case 1: $\lambda = 0$.

- $L(x, 0) = f(x) + \frac{c}{2} h(x)^2$.
- When is (*) possible?
- Recall Page 17 of Lec 11a, Proof Method 2 by Penalty Method: as $c \to \infty$, local-mins of L converge to a local-min of (1).

Case 2: $c = 0$.

- $L(x, \lambda) = f(x) + \lambda^T h(x)$.
- When is (*) possible?
- Recall 3rd condition of Prop 12.1: when $\lambda = \lambda^*$, x^* is a regular global-min of $L(x, \lambda^*)$, then x^* is a global-min of (1).

Convergence Mechanism 1

There are two convergence mechanisms, based on two cases above.

Mechanism 1 for convergence: taking $c^r \to \infty$.

- ◇ In early 60's, people pick $\lambda = 0$.
- ◇ Here, we allow λ to be nonzero, or changing but bounded.
- ◇ For $c \to \infty$ and bounded λ, we have $\langle \lambda, h(x) \rangle + f(x) + \underset{\downarrow \atop \infty}{c}\, h(x)^2 \approx \begin{cases} f(x), & \text{if } h(x)=0 \\ \infty, & \text{if } h(x)\neq 0. \end{cases}$

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

⋆ **Prop 13.1** (short version): If c^r is increasing and $\to \infty$ and λ^r is bounded, and let the global minimizer of $L_{c^r}(x, \lambda^r)$ be x^r. Then every limit point of x^r is a global minimizer of (1).

Prop 13.1 (Full Version)

Proposition 13.1

- Problem setup: Consider (1), where X is closed, and f and h are continuous. There exists a feasible point.
- Algorithm Setup: Let $\{\lambda^r\}$ be bounded, and $\{c^r\} \to \infty$.
- Algorithm Assumption: Assume $x^r = \operatorname{argmin}_x L_{c^r}(x, \lambda^r)$, and x^* is a limit point of the sequence $\{x^r\}$.
- Conclusion: Then x^* is a global minimizer of (1).

Proposition 13.1 Textbook Version

Proposition 4.2.1: Assume that f and h are continuous functions, that X is a closed set, and that the constraint set $\{x \in X \mid h(x) = 0\}$ is nonempty. For $k = 0, 1, \ldots$, let x^k be a global minimum of the problem

$$\begin{aligned} &\text{minimize} \quad L_{c^k}(x, \lambda^k) \\ &\text{subject to} \quad x \in X, \end{aligned}$$

where $\{\lambda^k\}$ is bounded, $0 < c^k < c^{k+1}$ for all k, and $c^k \to \infty$. Then every limit point of the sequence $\{x^k\}$ is a global minimum of the original problem (4.21).

Proof of Prop 12.1 (Reading)

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

$$f^* = \inf_{h(x)=0, x\in X} L_{c^r}(x, \lambda^r). = f(x) + \underbrace{\langle \lambda^r, h(x)\rangle + \frac{c^r}{2}\|h(x^r)\|^2}_{=0 \text{ if } h(x)=0.}$$

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

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

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

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

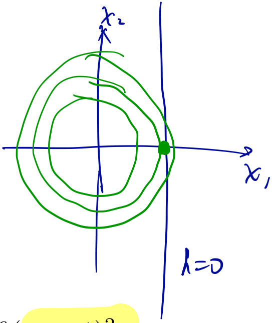
By $\|h(x^r)\|^2 \ge 0$ and $\{c^r\} \to \infty$, we have $h(x^r) \to 0$ and $h(\bar{x}) = 0$. Hence, $\bar{x}$ is feasible, and since the above inequality implies $f(\bar{x}) \le f^*$, so $\bar{x}$ is optimal.

Convergence Mechanism 2

Mechanism 2 for convergence: take $\lambda^r \to \lambda^*$.

- $\diamond$ Here, sufficiently large c is enough (no need to grow to infinity).
- $\diamond$ Is $c = 0$ enough?
- $\diamond$ Prop 12.1 shows a regular global-min of L is desirable
- $\star$ Assume $X = R^n$ and (x^*, λ^*) is a local min-Lagrange multiplier pair satisfying the 2nd order sufficiency conditions
- $\star$ For c sufficiently large, x^* is a strict local min of $L_c(\cdot, \lambda^*)$

Example



Consider the example

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

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

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

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

How to pick λ^r, c^r s.t. $x(\lambda^r, c^r) \to x^*$?

- Mechanism 1 (penalty):

$c^r \to \infty$, $\{\lambda^r\}$ bounded.

$x_1 = \frac{c^r - \lambda^r}{c^r + 1} \to 1 = x_1^*$.

- Mechanism 2 (Lag-multiplier):

c^r arbitrary, $\lambda^r \to -1 = \lambda^*$,

$x_1 = \frac{c - \lambda^r}{c+1} \to \frac{c+1}{c+1} = 1 = x_1^*$.

Outline

Subproblem Solver

- Prop 13.1 requires $\{x^r\}$ to be the global-min of L exactly.
  - ◇ Impossible in practice
  - ◇ A common issue in double-loop algorithm
- Do we really need to solve the subproblem to global optimality?
  - ◇ For nonconvex L, solving to stationary point is enough (to get stationary of original problem)
  - ◇ Getting an inexact stationary point (with increasing precision) is enough

Result

Proposition 13.2 (inexact subproblem;)

- Problem setup: Consider (1), where $X = R^n$, and f and h are continuously differentiable.
- Algorithm Setup: Let $\{\lambda^r\}$ be bounded, and $\{c^r\} \to \infty$.
- Algorithm Assumption: Assume x^r satisfies inexact stationary point

$$\|\nabla_x L_{c^r}(x^r, \lambda^r)\| \le \epsilon_r \to 0, \ \forall \ r, \tag{2}$$

  and x^* is a regular limit point of $\{x^r\}$ (i.e. rank$(\nabla h(x^*)) = m$).
- Conclusion: Then the algorithm converges to first-order stationary solutions (KKT points)

$$\lambda^r + c^r h(x^r) \to \lambda^*, \quad \nabla_x L(x^*, \lambda^*) = 0, \quad h(x^*) = 0$$

additional finding: converge to λ^* !

Proposition 13.2 Textbook Version

Proposition 4.2.2: Assume that $X = \Re^n$, and f and h are continuously differentiable. For $k = 0, 1, \ldots$, let x^k satisfy

$$\|\nabla_x L_{c^k}(x^k, \lambda^k)\| \leq \epsilon^k,$$

where $\{\lambda^k\}$ is bounded, and $\{\epsilon^k\}$ and $\{c^k\}$ satisfy

$$0 < c^k < c^{k+1}, \quad \forall\, k, \qquad c^k \to \infty,$$

$$0 \leq \epsilon^k, \quad \forall\, k, \qquad \epsilon^k \to 0.$$

Assume that a subsequence $\{x^k\}_K$ converges to a vector x^* such that $\nabla h(x^*)$ has rank m. Then

$$\{\lambda^k + c^k h(x^k)\}_K \to \lambda^*,$$

where λ^* is a vector satisfying, together with x^*, the first order necessary conditions

$$\nabla f(x^*) + \nabla h(x^*)\lambda^* = 0, \qquad h(x^*) = 0.$$

Compare Prop 13.2 and Prop 13.1

Difference and relation between these two results?

- Differences
  - ◇ Prop 13.2 is for stationary points, Prop 13.1 is for global-min
  - ◇ Prop 13.2 is for inexact subproblem solution, Prop 13.1 is for exact subproblem solution

- **Relation**: For convex problem (i.e. f is convex, h is linear), stationary point is global-min; so Prop 13.2 implies Prop 13.1.

Proof of Prop 13.2 for $\epsilon_k = 0$ case (Reading)

- **Proof:** We have

$$0 = \nabla_x L_{c^r}(x^r, \lambda^r) = \nabla f(x^r) + \nabla h(x^r)(\lambda^r + c^r h(x^r)) =$$
$$\nabla f(x^r) + \nabla h(x^r)\bar{\lambda}^r,$$

where $\bar{\lambda}^r = \lambda^r + c^r h(x^r)$.
Multiply with

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

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

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

We also have $\nabla_x L(x^*, \lambda^*) = 0$ and $h(x^*) = 0$ (since $\bar{\lambda}^r$ converges).
Q.E.D.

For general ϵ_k case, see the textbook for the proof.

END of TUESDAY LECTURE.

THURSDAY Lecture

Quadratic Penalty Method

Let's review the theory/algorithm so far.

Problem: (P) $\min_x f(x), \text{subject to } x \in X,\ h(x) = 0.$
Quadratic Penalty Method:

- Define $L_c(x, \lambda) = f(x) + \lambda^T h(x) + \frac{c}{2} h(x)^2$.
- Pick initial λ^0, c^0.
- For $r = 0, 1, 2, \ldots$
  - **Inner loop**: Solve $\min_x L_{c^r}(x, \lambda^r)$, to obtain x^r s.t.

$$\|\nabla L_{c^r}(x^r, \lambda^r)\| \le \epsilon^r,$$

good/accurate solutions only needed in final stages

    where the error $\epsilon_r \to 0$; e.g. pick $\epsilon^r = 1/r$ or $1/r^2$.
  - **Outer loop**: Update λ^r, c^r as follows: increase c^r to ∞ and keep λ^r bounded.
    e.g. $c^r = r^2$ or 1.5^r.

Theoretical Guarantee: every regular limit point of $\{x^r\}$ is a stationary point of (P).

Quadratic Penalty Method

Let's review the theory/algorithm so far.

Problem: (P) $\min_x f(x)$, subject to $x \in X,\ h(x) = 0$.
Quadratic Penalty Method:

- ◇ Define $L_c(x, \lambda) = f(x) + \lambda^T h(x) + \frac{c}{2}\underbrace{h(x)^2}_{\geq 0}$.
- ◇ Pick initial λ^0, c^0.
- ◇ For $r = 0, 1, 2, \ldots$
  - – **Inner loop**: Solve $\min_x L_{c^r}(x, \lambda^r)$, to obtain x^r s.t.

$$\|\nabla L_{c^r}(x^r, \lambda^r)\| \leq \epsilon^r,$$

    where the error $\epsilon_r \to 0$; e.g. pick $\epsilon^r = 1/r$ or $1/r^2$.
  - – **Outer loop**: Update λ^r, c^r as follows: increase c^r to ∞ and keep λ^r bounded.
    e.g. $c^r = r^2$ or 1.5^r.

Theoretical Guarantee: every regular limit point of $\{x^r\}$ is a stationary point of (P).

Question $\epsilon^r = 0.01, \forall r$?
Possibly works; but final error unknown,
maybe 0.1, maybe 100.

- In practice, set an outerloop
stopping error at 0.01,
then you can let
$\epsilon^r = 0.01$.
It works
in practice.

Theoretical Issue

- The theory only provides some guaranteee. It may fail due to:
  - ⋆ Inner loop failure: x^r with $\nabla_x L_{c^r}(x^r, \lambda^r) \approx 0$ cannot be found.
    - – Happen when L is unbounded below
  - ⋆ Outer loop failures:
    - ⋄ no limit point (x^r can be unbounded)
    - ⋄ no regular limit point
    - – Happen often when the problem is infeasible (algorithm converges to infeasible vector)
- Success case: A sequence $\{x^r\}$ with $\nabla_x L_{c^r}(x^r, \lambda^r) \approx 0$ is found and it has a regular limit point x^*.
  - ⋄ x^* together with λ^* [the corresponding limit point of $\{\lambda^r + c^r h(x^r)\}$] satisfies the first-order necessary conditions.

Theoretical Issue

- The theory only provides some guaranteee. It may fail due to:
  - ⋆ Inner loop failure: x^r with $\nabla_x L_{c^r}(x^r, \lambda^r) \approx 0$ cannot be found.
    - – Happen when L is unbounded below
  - ⋆ Outer loop failures:
    - ◇ no limit point (x^r can be unbounded)
    - ◇ no regular limit point

      related to homework 5
    - – Happen often when the problem is infeasible (algorithm converges to infeasible vector)

- Success case: A sequence $\{x^r\}$ with $\nabla_x L_{c^r}(x^r, \lambda^r) \approx 0$ is found and it has a regular limit point x^*.
  - ◇ x^* together with λ^* [the corresponding limit point of $\{\lambda^r + c^r h(x^r)\}$] satisfies the first-order necessary conditions.

Practical Issue: Convergence Speed

- Ill-conditioning: The condition number of the Hessian $\nabla^2_{xx} L_{c^r}(x^r, \lambda^r)$ tends to increase with c^r.
  - Often the major issue why quadratic penalty method fails

$L = \underbrace{f(x)}_{I} + \langle \lambda, \underbrace{\underbrace{h(x)}_{x_1}}_{0} \rangle + \underbrace{h(x)^2}_{\begin{pmatrix} 1 & \\ & 0 \end{pmatrix}}$

- **Example**:

$$\begin{aligned} \text{minimize} \quad & f(x) = \frac{1}{2}(x_1^2 + x_2^2) \rightarrow \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \\ \text{subject to} \quad & x_1 = 1 \end{aligned}$$

$\nabla^2_{xx} L_c(x, \lambda) = I + c \cdot \text{diag}(1, 0) = \text{diag}(1 + c, 1).$

Condition number $1 + c \to \infty$ as $c \to \infty$.

- **Lesson**: Don't pick huge c initially!

Practical Issue: Convergence Speed

- Ill-conditioning: The condition number of the Hessian $\nabla^2_{xx} L_{c^r}(x^r, \lambda^r)$ tends to increase with c^r.
  - Often the major issue why quadratic penalty method fails

- **Example**:

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

$L_c(x, \lambda) = - + \frac{c}{2} h(x)^2$

$\nabla^2_{xx} L_c(x, \lambda) = I + c \cdot \text{diag}(1, 0) = \text{diag}(1 + c, 1)$.

Condition number $1 + c \to \infty$ as $c \to \infty$.

- **Lesson**: Don't pick huge c initially!

Overcome ill-conditioning

The theory Prop. 13.2 doesn't require warm-start, as it doesn't care about subproblem solver.
If we care about computation time, then warm-start is very important.

- One solution: warm-start
  - Idea: solution for c should be close to solution for $c + \epsilon$
  - So gradually change c, and use previous solution x^r as initial point of $(r+1)$-th inner loop
  - How to pick rate of increasing c^r and inner loop accuracy? Big question of any double-loop algorithm

- Traditional way to overcome ill-conditioning: Newton-like method
  - May not be scalable for large problems

Inequality Constraints

- Convert to equality case by squared slack variables
  - Convert $g_j(x) \leq 0$ to $g_j(x) + z_j^2 = 0$.

- The penalty method solves problems of the form

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

  for various values of μ and c.

- Trick: First minimize $\bar{L}_c(x, z, \lambda, \mu)$ w.r.t. z to get an expression

$$L_c(x, \lambda, \mu) = L_c(x, \lambda) + \frac{1}{2c} \sum_j \left\{ (\max\{0, \mu_j + c g_j\})^2 - \mu_j^2 \right\}.$$

- This is the new function to use. In primal step, minimize $L_c(x, \lambda, \mu)$ w.r.t. x.

Outline

Update Multipliers Based on Mechanism 2

- In Prop 13.2, the penalty method does NOT require explicit λ
- However as long as $x \to x^*$, we will be able to recover λ^* by

$$\lambda^r + c^r h(x^r) \to \lambda^*$$

- It may be a good idea to appropriately update the λ sequence as well, such as

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

  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 _______ loop)
  - ⋆ Faster convergence when λ^r is updated than when λ^r is kept constant. (faster _______ loop)

Update Multipliers Based on Mechanism 2

- In Prop 13.2, the penalty method does NOT require explicit λ
- However as long as $x \to x^*$, we will be able to recover λ^* by

$$\lambda^r + c^r h(x^r) \to \lambda^*$$

- It may be a good idea to appropriately update the λ sequence as well, such as

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

  This is the (1st order) method of multipliers.
- Key advantages to be shown: $c^r = c$
  - ⋆ Less ill-conditioning: It is not necessary that $c^r \to \infty$ (only that c^r exceeds some threshold). (faster inner loop)
  - ⋆ Faster convergence when λ^r is updated than when λ^r is kept constant. (faster outer loop) $c^r \to \infty$

Augmented Lagrangian Method (ALM)

- Consider the equality constrained problem

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

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

- The (1st order) multiplier method finds

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

  and updates λ^r using

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

History and Names

- First appeared in Hestenes-69 and Powell-69; and Haarhoff and Buys-70.

- Bertsekas's 1982 book "Constrained optimization and lagrange multiplier methods" gave a detailed analysis and many references (current version published in 2015)

- Initially called "multiplier method" or "method of multipliers"
  - to highlight the extra multiplier update, compared to penalty method

- Also called "augmented Lagrangian method (ALM)"
  - to highlight augmented Lagrangian used in the primal update

to compare with ________ method
which is based on Lagrangian, not augmented Lagrangian

Section 2.2. Another motivation of ALM

Dual Ascent (Linear Constraint Case)

- Consider the problem

$$\min_x f(x), \text{subject to } Ax = b,$$

where f is strictly convex.

- The Lagrangian $L(x, \lambda) = f(x) + \lambda^T(Ax - b)$.

- Consider the dual function

$$q(\lambda) = \min_x L(x, \lambda).$$

We used dual coordinated ascent to solve SVM. Can we generalize to other problems?

- Instead of $\min f$, we maximize the dual: $\max_\lambda q(\lambda)$
  - Same value since strong duality holds

Dual Ascent (cont'd)

- Knowledge: f is strictly convex, then q is differentiable
- Thus $\max_\lambda q(\lambda)$ is to maximize a differentiable concave function
- One option: apply gradient ascent to solve it:

$$\underline{\text{DA:}}\quad \lambda \leftarrow \lambda + \alpha \nabla q(\lambda).$$

- What is the gradient of the dual function $q(\lambda) = \min_x L(x, \lambda)$?
  - As L is strictly convex in x, for given λ it has unique minimizer $x^*(\lambda)$

    $L(x^*,\lambda) = f(x^*) + \langle \lambda, Ax-b\rangle + \frac{c}{2}\|Ax-b\|^2.$

  - Claim: $\nabla q(\lambda) = \partial_\lambda L(x^*, \lambda) =$ $\underline{Ax^*-b}$

    Do not consider $\frac{\partial x^*}{\partial \lambda}$ here.

    result from convex analysis

$q(\lambda) = \min_x (x+\lambda)^2.$

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

$\frac{\partial L(x^*,\lambda)}{\partial \lambda} = \frac{\partial (x^*+\lambda)^2}{\lambda} = 2(x^*+\lambda).$

- Dual ascent method:

$$x \leftarrow \operatorname*{argmin}_x L(x, \lambda),$$

$$\lambda \leftarrow \lambda + \alpha(Ax - b).$$

Dual Ascent (cont'd)

- Knowledge: f is strictly convex, then q is differentiable
- Thus $\max_\lambda q(\lambda)$ is to maximize a differentiable concave function
- One option: apply gradient ascent to solve it:

$$\lambda \leftarrow \lambda + \alpha \nabla q(\lambda).$$

- What is the gradient of the dual function $q(\lambda) = \min_x L(x, \lambda)$?
  - As L is strictly convex in L, for given λ it has unique minimizer $x^*(\lambda)$
  - Claim: $\nabla q(\lambda) = \partial_\lambda L(x^*, \lambda) =$ ____________
- **Dual ascent method**:

DA $\Leftrightarrow$

$$\begin{cases} x \leftarrow \underset{x}{\operatorname{argmin}}\, L(x, \lambda), \\ \lambda \leftarrow \lambda + \alpha(Ax - b). \end{cases}$$

Augmented Lagrandian Method (ALM)

- Issue of Dual Ascent: when f is not strictly convex, q may not be differentiable

- Consider an auxiliary problem

$$\min_x \overbrace{f(x) + \frac{c}{2}\|Ax-b\|^2}^{\tilde{f}(x)}, \text{subject to } Ax = b,$$

  - With extra $\|Ax-b\|^2$ term, under mild conditions one can show q_c is differentiable
  - $L_c(x,\lambda) =$ $\underbrace{f(x) + \frac{c}{2}\|Ax-b\|^2}_{\tilde{f}(x)} + \langle \lambda, Ax-b \rangle$, and dual $q_c(\lambda) =$ $\min_x L_c(x,\lambda)$.

- Apply dual ascent to solve the auxiliary problem:

$$\lambda \leftarrow \lambda + \alpha \nabla q_c(\lambda).$$

- **ALM** = "augmented version" of dual ascent, with stepsize c $(= \frac{1}{L})$.

$$x \leftarrow \operatorname*{argmin}_x L_c(x,\lambda) = f(x) + \lambda^T(Ax-b) + \frac{c}{2}\|Ax-b\|^2,$$

$$\lambda \leftarrow \lambda + c(Ax-b).$$

Dual View for General h (Reading)

- Consider the problem

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

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

$$q_c(\lambda) = \min_{\|x - x^*\| < \varepsilon} L_c(x(\lambda, c), \lambda),$$

$$\nabla q_c(\lambda) = \nabla_\lambda x(\lambda, c)\nabla_x L_c(x(\lambda, c), \lambda) + h(x(\lambda, c) = h(x(\lambda, c))$$

We have $\nabla q_c(\lambda^*) = h(x^*) = 0$ and $\nabla^2 q_c(\lambda^*) \prec 0$.

- ALM = "gradient ascent" for augmented dual q_{c^r} with special stepsize c^r

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

Convex Example

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

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

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

$$\underbrace{\lambda^{r+1} - \lambda^*}_{\text{dual error}} = \frac{\lambda^r - \lambda^*}{c^r+1} \rightarrow \text{dual error.}$$

$$e^{r+1} = \frac{e^r}{c^r+1}.$$

When does $\{e^r\}$ converge?

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

Bigger c, faster outer loop;
(slower inner loop)

$c^r = 99$, $e^{r+1} = \frac{e^r}{100}$, faster.

$c^r = 0.1$, $e^{r+1} = \frac{e^r}{1.1}$. slower

Nonconvex Example

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

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

provided $c^r > 1$ (otherwise the min does not exist, opt goes to $-\infty$)

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

$$\lambda^{r+1} - \lambda^* = -\frac{\lambda^r - \lambda^*}{c^r - 1}$$

$e^{r+1} = \frac{-e^r}{c^r - 1}$, need $c^r - 1 > 1$, i.e. $c^r > 2$.

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

Computational Aspects

- Key issue is how to select $\{c^r\}$, which should become larger than the "threshold" of the given problem.
    - c^0 should not be so large as to cause ill-conditioning initially
    - c^r should not be increased so fast that too much ill-conditioning in early stage
      inner loop slow
    - c^r should not be increased so slowly that the dual iteration converges slowly
      outer loop slow.
- 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. $1 \sim 3$. (if use Newton method for subproblem $\beta = 5 \sim 10$)
- In practice the minimization of $L_{c^r}(x, \lambda^r)$ is typically inexact (usually exact asymptotically).
    - In some variants of the method, only one Newton step per minimization is used (with safeguards).
- See more at Sec. 4.2.2.

Computational Aspects

- Key issue is how to select $\{c^r\}$, which should become larger than the "threshold" of the given problem.
  - c^0 should not be so large as to cause ill-conditioning initially
  - c^r should not be increased so fast that too much ill-conditioning in early stage
  - c^r should not be increased so slowly that the dual iteration converges slowly
- 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.
- In practice the minimization of $L_{c^r}(x, \lambda^r)$ is typically inexact (usually exact asymptotically).
  - In some variants of the method, only one Newton step per minimization is used (with safeguards).
- See more at Sec. 4.2.2.

Summary

In this lecture, we learned the following (think yourself before reading):

- A double-loop algorithm framework based on augmented Lagrangian
- Quadratic penalty method and convergence mechanism 1
  - Results that justify quadratic penalty method
- ALM (multiplier method), motivated from convergence mechanism 2
  - Another motivation: “augmented version” of dual ascent

Summary

In this lecture, we learned the following (think yourself before reading):

- A double-loop algorithm framework based on augmented Lagrangian

- Quadratic penalty method and convergence mechanism 1
  - Results that justify quadratic penalty method

- ALM (multiplier method), motivated from convergence mechanism 2
  - Another motivation: “augmented version” of dual ascent