ORIE7391 Augmented Lagrangian Method and Alternating Direction Method of Multiplier

Siyu Kong (sk3333)

Cornell University

February 21, 2022

Table of Contents

Question

- Question 1: Which of the following statements about l_1 penalty method is correct?
  (a). After adding the l_1 penalty function, we get a smooth objective function in penalty method.
  (b). Compared to quadratic penalty method, l_1 penalty method relies less on the choices of penalty parameter ρ.
  (c). Both of (a) and (b).
  (d). Neither of (a) or (b).
- Question 2: What is true about augmented Lagrangian method?
  (a). The choice of penalty parameters in augmented Lagrangian method is irrelevant to the convergence rate.
  (b). Under mild conditions augmented Lagrangian is smooth.
  (c). Both of (a) and (b).
  (d). Neither of (a) or (b).

Problem Formulation

Consider the following optimization problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & c_i(x) = 0 \text{ for } 1 \le i \le m \end{aligned} \tag{1}$$

where f and c_i are smooth functions.
In fact, (1) is equivalent to augmented form:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & Q(x;\rho) = f(x) + \frac{\rho}{2}\|c(x)\|_2^2 \\ \text{s.t.} \quad & c_i(x) = 0 \text{ for } 1 \le i \le m \end{aligned} \tag{2}$$

where $c(x) = (c_1(x), \dots, c_m(x))^T \in \mathbb{R}^m$ and $\|\cdot\|_2$ is l_2 norm.

Toy Algorithm - Quadratic Penalty Method (Theoretic Framework)

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & Q(x;\rho) = f(x) + \frac{\rho}{2}\|c(x)\|_2^2 \\ \text{s.t.} \quad & c_i(x) = 0 \text{ for } 1 \le i \le m \end{aligned}$$

Algorithm Idea: At each step, given ρ^k, solve for unconstrained optimization problem $x^k = \min_x Q(x;\rho^k)$, denote x^* as primal solution, then $Q(x^k;\rho^k) \le Q(x^*;\rho^k) = p$. Therefore usually x^k is not feasible. When $\rho^k \to \infty$, hope $x^k \to x^*$.

Algorithm Quadratic Penalty Method

Consider an increasing sequence of $\{\rho^k\}$ and a decreasing sequence $\{\epsilon^k\} \to 0$.
for $k = 0, 1, 2, ...,$
  find $x^k = \arg\min Q(x;\rho^k)$ (When $\|\nabla_x Q(x^k;\rho^k)\| \le \epsilon^k$ holds).
Output x^k when it achieve the convergence test .

Penalty Function and Penalty Parameter

In form (2), we introduce a quadratic penalty function $g_\rho(x)$ of the form

$$g_\rho(x) = \frac{\rho}{2}\|c(x)\|_2^2 \tag{3}$$

- **Good Property**: $Q(x;\rho)$ is smooth, $\nabla_x Q(x;\rho^k)$ easy to compute.
- **Bad Property**: (i). $\nabla_x g_\rho(\tilde{x}) = 0$ for any feasible $\tilde{x}$, need large ρ to guarantee convergence to feasible solution. (ii). $H = \nabla^2_{xx} Q(x;\rho)$, some eigenvalues of H approach constants, others are of order ρ, when $\rho \to \infty$, condition number tends to infinity.
- **Other Variant**: Take $g_\rho(x) = \dfrac{\rho}{2}\|c(x)\|_1$. Better convergence result but lack of smoothness.

Theoretical Convergence Result and Practical Implementation

- **Convergence**: When x^k is global minimization of $Q(x;\rho^k)$ and $\rho^k \to \infty$, $x^* = \lim_{k\to\infty} x^k$ is a global minimum solution.
- **Practical Implementation**: Highly adopted in applications due to
  (i). Simplicity
  (ii). Ill-conditioning problem can be solved by numerical scheme.

Problem and Goal

Problem: Any method to both guarantee the **smoothness** of objective function as well as good **convergence** result?
Answer: Yes. Combining quadratic penalty method with Lagrangian method. In literature we call it **method of multipliers** or **augmented Lagrangian method**.

Lagrangian and Dual Problem

Still consider primal augmented form, denote optimal value of (4) as p:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & Q(x;\rho) = f(x) + \frac{\rho}{2}\|c(x)\|_2^2 \\ \text{s.t.} \quad & c_i(x) = 0 \text{ for } 1 \le i \le m \end{aligned} \tag{4}$$

- Lagrangian of $Q(x;\rho)$ is $L(x,y;\rho) = f(x) + \sum_{1\le i\le m} y_i c_i(x) + \frac{\rho}{2}\|c(x)\|_2^2$
- Dual Function: $g_\rho(y) = \inf_x L(x,y;\rho)$
- Dual Problem: $\max_y g_\rho(y)$, optimal value denoted as d
- Weak Duality: $d \le p$
- When strong duality holds, $p = d$, instead of minimizing primal problem, we can **maximize dual problem**.

Proof of Weak Duality

For any function $f(x)$, the weak duality $d \leq p$ holds.

Proof.

Denote x^* as primal optimal solution. For any y, we have $g_\rho(y) = \inf_x L(x, y; \rho)$. Particularly, $g_\rho(y) \leq L(x^*, y; \rho) = f(x^*) = p$. Then dual optimal value $d = \max_y g_\rho(y) \leq p$. □

Another point of view:

- For any function $f(x, y)$, $\inf_x \sup_y f(x, y) \geq \sup_y \inf_x f(x, y)$.
- Consider $L(x, y; \rho) = f(x) + \sum_{1 \leq i \leq m} y_i c_i(x) + \frac{\rho}{2}\|c(x)\|_2^2$, therefore $\inf_x Q(x; \rho) = \inf_x \sup_y L(x, y; \rho)$.

Algorithm Derivation

Still consider primal augmented form, denote optimal value of (4) as p:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & Q(x;\rho) = f(x) + \frac{\rho}{2}\|c(x)\|_2^2 \\ \text{s.t.} \quad & c_i(x) = 0 \text{ for } 1 \le i \le m \end{aligned} \tag{5}$$

- Lagrangian of $Q(x;\rho)$ is
  $L(x, y;\rho) = f(x) + \sum_{1\le i\le m} y_i c_i(x) + \dfrac{\rho}{2}\|c(x)\|_2^2$
- Dual Function: $g_\rho(y) = \inf_x L(x, y;\rho) \le p$
- Dual Problem: $\max_y g_\rho(y)$, optimal value denoted as d
- **Algorithm Idea**:
  (i). Given y^k, first compute $x^{k+1} = \arg\min L(x, y^k;\rho)$
  (ii). Approximate $\nabla g_\rho(y^k)$ by $\nabla_y L(x^{k+1}, y^k;\rho)$.

Algorithm Derivation

Algorithm Idea:
(i). Given y^k, first compute $x^{k+1} = \arg\min L(x, y^k; \rho)$
(ii). Approximate $\nabla g_\rho(y^k)$ by $\nabla_y L(x^{k+1}, y^k; \rho)$.
Analysis:

- $\nabla_x L(x, y; \rho) = \nabla f(x) + \sum_{1 \le i \le m}(y_i + \rho c_i(x))\nabla c_i(x)$.
- $\nabla_y L(x, y; \rho) = c(x)$

Algorithm Augmented Lagrangian Method

Consider a non-decreasing sequence of $\{\rho^k\}$.
for $k = 0, 1, 2, \ldots$, given a pair (x^k, y^k) at k-th iteration,
  find $x^{k+1} = \arg\min_x L(x, y^k; \rho)$,
  update $y^{k+1} = y^k + \rho^k c(x^{k+1})$.
Output (x^k, y^k) when convergence test is satisfied.

Comparison with Quadratic Penalty Method

Algorithm Augmented Lagrangian Method

Consider a non-decreasing sequence of $\{\rho^k\}$.
for $k = 0, 1, 2, \ldots$, given a pair (x^k, y^k) at k-th iteration,
  find $x^{k+1} = \arg\min_x L(x, y^k; \rho)$,
  update $y^{k+1} = y^k + \rho^k c(x^{k+1})$.
Output (x^k, y^k) when convergence test is satisfied.

- When we find $x^{k+1} = \arg\min_x L(x, y^k; \rho)$, the starting point of the search is less sensitive, which can simply put as $x_s^{k+1} = x^k$.
- Step size ρ^k is not required to increase indefinitely (we can even set $\rho_k \equiv \rho$ for suitable ρ), and ill conditioning is less a problem.

Alternative Understanding of the Augmented Lagrangian Method

Apart from dual accent idea, we can also connect augmented Lagrangian method with KKT conditions as follow:

Lemma (Characterization of Strong Duality)

In convex optimization problem, when primal-dual pair (x^*, y^*) *satisfy KKT conditions:*

- $\nabla_x L(x^*, y^*; \rho) = 0$
- $c_i(x^*) = 0$ *for* $1 \leq i \leq m$

they are primal-dual optimal and strong duality holds. Particularly, $Q(x^*, \rho) = g_\rho(y^*)$*.*

Explanation: In augmented Lagrangian method, $\{x^k, y^k\}$ is a sequence that converges to a pair which satisfies KKT conditions.

Existence of Local Minima of the Augmented Lagrangian

Question Concerned: Whether local minima of the augmented Lagrangian exist? If so, how their distance from local minima of the original problem is affected by the values of the multiplier y^k and the penalty parameter ρ^k?
Answer: Theorem 17.6 in the reading NW04. (When (x^k, y^k) is close enough to (x^*, y^*), the local minima in the algorithm exists with suitable choice of ρ^k).

Existence of Local Minima of the Augmented Lagrangian

Assumption: Let x^* be a local minimum and LICQ holds, and f, c are C^2 functions on some open sphere centred at x^*. Further more x^* together with its associated Lagrange multiplier vector y^* satisfies KKT conditions and second-order conditions:

$$z^T \nabla^2_{xx} L(x^*, y^*; 0) z > 0 \tag{6}$$

for all $z \neq 0$ with $\nabla c_i(x^*)^T z = 0$ for any $1 \leq i \leq m$.
LICQ(Linear Independence Constraint Qualification): $\nabla c_i(x^*)$ for $1 \leq i \leq m$ are linearly independent vectors.

Theorem of Minima Existence

Theorem (17.6 in NW04)

Assume a pair (x^*, y^*) *satisfy the assumption, there exists a threshold* $\bar{\rho}$, *and positive scalars* δ, ϵ *and* M *such that*

- *(i). For all* y^k *and* ρ^k *satisfying* $\|y^k - y^*\| \le \rho^k \delta$, $\rho^k \ge \bar{\rho}$, *the problem* $\min_x L(x, y^k; \rho^k)$ *subject to* $\|x - x^*\| \le \epsilon$ *has a unique solution* x^{k+1}. *Moreover, we have* $\|x^{k+1} - x^*\| \le M\|y^k - y^*\|/\rho^k$.
- *(ii). Under the same condition as (i), we have* $\|y^{k+1} - y^*\| \le M\|y^k - y^*\|/\rho^k$.
- *(iii). Under the same condition as (i), the matrix* $\nabla^2_{xx} L(x^k, y^k; \rho^k)$ *is positive definite and the constraint gradient* $\nabla c_i(x^k)$, $1 \le i \le m$ *are linearly independent.*

Proof of Theorem 17.6

Proof.

(i). For $\rho > 0$, consider the following system of equations on $(x, \tilde{y}, y, \rho)$:

$$\nabla f(x) + A(x)^T \tilde{y} = 0, \quad c(x) + (y - \tilde{y})/\rho = 0 \tag{7}$$

where $A(x)^T = [\nabla c_i(x)]_{1 \leq i \leq m}$. Particularly, from definition of iteration steps, we have $(x^{k+1}, y^{k+1}, y^k, \rho^k)$ satisfy the above equations (7).

Now define $t \in \mathbb{R}^m, \gamma \in \mathbb{R}$ as

$$t = (y - y^*)/\rho, \quad \gamma = 1/\rho. \tag{8}$$

□

Proof of Theorem 17.6

Proof.

We can rewrite (7) as

$$\nabla f(x) + A(x)^T \tilde{y} = 0, \quad c(x) + t + \gamma y^* - \gamma \tilde{y} = 0. \tag{7}$$

For $t = 0$ and $\gamma \in [0, 1/\bar{\rho}]$, from KKT conditions, we know (7) has the solution $x = x^*$ and $\tilde{y} = y^*$. The Jacobian w.r.t. $(x, \tilde{y})$ at such a solution is

$$\begin{bmatrix} \nabla^2_{xx} L_0(x^*, y^*) & A(x^*)^T \\ A(x^*) & -\gamma I \end{bmatrix} \tag{8}$$

In fact the matrix (8) is invertible for all $\gamma \in [0, 1/\bar{\rho}]$.

□

Proof of Theorem 17.6

Proof.

Denote $x(t,\gamma) := \min_x L(x, y^* + \rho t, \rho)$ and $y(t,\gamma)$ as the next y^{k+1} when starting at y^k indexed by t ($\rho = 1/\gamma$). From implicit function theorem, there exist ϵ and $\delta > 0$ such that, for $(x(t,\gamma), y(t,\gamma)) \in B((x^*, y^*), \epsilon)$ and $(t,\gamma) \in B(K,\delta)$ where $K := \{(0,\gamma) : \gamma \in [0, 1/\bar{\rho}]\}$

$$
\begin{aligned}
\nabla f(x(t,\gamma)) + A(x(t,\gamma))^T y(t,\gamma) &= 0 \\
c(x(t,\gamma)) + t + \gamma y^* - \gamma y(t,\gamma) &= 0
\end{aligned}
$$

Differentiate the two equations,

$$
\begin{bmatrix} \nabla_t x(t,\gamma)^T & \nabla_\gamma x(t,\gamma)^T \\ \nabla_t y(t,\gamma)^T & \nabla_\gamma y(t,\gamma)^T \end{bmatrix} = A(t,\gamma) \begin{bmatrix} 0 & 0 \\ -I & y(t,\gamma) - y^* \end{bmatrix} \tag{7}
$$

Proof of Theorem 17.6

Proof.

where

$$A(t,\gamma) = \begin{bmatrix} \nabla^2_{xx} L_0(x(t,\gamma), y(t,\gamma)) & A(x(t,\gamma))^T \\ A(x(t,\gamma)) & -\gamma I \end{bmatrix}. \tag{7}$$

Notice $A(t,\gamma)$ is uniformly bounded on $\{(t,\gamma) : |t| < \delta, \gamma \in [0, 1/\bar{\rho}]\}$. Now by applying fundamental theorem of calculus algebraic computation,

$$(|x(t,\gamma) - x^*|^2 + |y(t,\gamma) - y^*|^2)^{1/2} \leq 2\mu|t| \tag{8}$$

where μ is some number that is larger than the upper bound of $|A(t,\gamma)|$. Result of $\|x^{k+1} - x^*\| \leq M\|y^k - y^*\|/\rho^k$ and $\|y^{k+1} - y^*\| \leq M\|y^k - y^*\|/\rho^k$ follows. The factor $1/\rho^k$ comes with definition that $t = (y - y^*)/\rho$. $\square$

Convergence Result

From $\|y^{k+1} - y^*\| \leq M\|y^k - y^*\|/\rho^k$, we know that:

- When $\{\rho^k\}$ is chosen to increase and diverge to infinity, augmented Lagrangian method has a **superlinear** convergence rate.
- When $\{\rho^k\}$ is bounded (for example, a constant sequence), then augmented Lagrangian method has a **linear** convergence rate.

Alternating Direction Method of Multipliers

Key Idea: ADMM is an algorithm intended to blend the decomposability of dual ascent with the superior convergence properties of the method of multipliers.

Problem Formulation in ADMM

Consider the following problem

$$\begin{aligned} \min_{x,z} \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c \end{aligned} \tag{9}$$

with variables $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^m$, $A \in \mathbb{R}^{p\times n}$, $B \in \mathbb{R}^{p\times m}$ and $c \in \mathbb{R}^p$. Assume f and g are convex.
If we want to use augmented Lagrangian method, we have the following augmented Lagrangian:

$$L_\rho(x, z, y) = f(x)+g(z)+y^T(Ax+Bz-c)+\frac{\rho}{2}\|Ax+Bz-c\|_2^2 \tag{10}$$

Augmented Lagrangian Method v.s. ADMM

- In augmented Lagrangian method, need to solve

$$\min_{x,z} L_\rho(x, z, y)$$

Challenge: x, z in quadratic augmented term is not separable! Cannot divide the problem into smaller pieces.
Requirement: However, we want to keep this quadratic term because of its fast convergence rate.

Augmented Lagrangian Method v.s. ADMM

- In augmented Lagrangian method, need to solve

$$\min_{x,z} L_\rho(x, z, y)$$

Solution: Do x-minimization step and z-minimization step separately:

$$\begin{aligned} x^{k+1} &:= \arg\min_x L_\rho(x, z^k, y^k) \\ z^{k+1} &:= \arg\min_z L_\rho(x^{k+1}, z, y^k) \\ y^{k+1} &:= y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \end{aligned} \tag{11}$$

This method is called ADMM.

Convergence Result of ADMM

- Assumption 1: The (extended-real valued) functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are closed, proper can convex. Equivalently, $epi(f)$ and $epi(g)$ are closed nonempty convex set.
- Assumption 2: The unaugmented Lagrangian L_0 has a saddle point.

Theorem (Convergence Result of ADMM)

Under Assumptions 1 and 2, ADMM iterates satisfy the following

- *Residual convergence:* $r^k \to 0$ *as* $k \to$, *where* $r^k := Ax^k + Bz^k - c$. *Equivalently the iterates approach feasibility.*
- *Objective convergence:* $f(x^k) + g(z^k) \to p^*$ *as* $k \to \infty$.
- *Dual variable convergence:* $y^k \to y^*$ *as* $k \to \infty$ *where* y^* *is a dual optimal point.*

Example of ADMM: LASSO

Notice that

$$min_x f(x) + g(x) \Leftrightarrow min_{x,z} f(x) + g(z) \text{ s.t. } x - z = 0 \tag{12}$$

Now consider lasso problem: Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, want

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{13}$$

We can rewrite it as

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\alpha\|_1 \text{ s.t. } \beta - \alpha = 0. \tag{14}$$

ADMM Steps in LASSO

ADMM steps:

$$\begin{aligned} \beta^k &= (X^TX + \rho I)^{-1}(X^Ty + \rho(\alpha^{k-1} - y^{k-1})) \\ \alpha^k &= S_{\lambda/\rho}(\beta^k + y^{k-1}) \\ y^k &= y^{k-1} + \beta^k - \alpha^k \end{aligned} \tag{15}$$

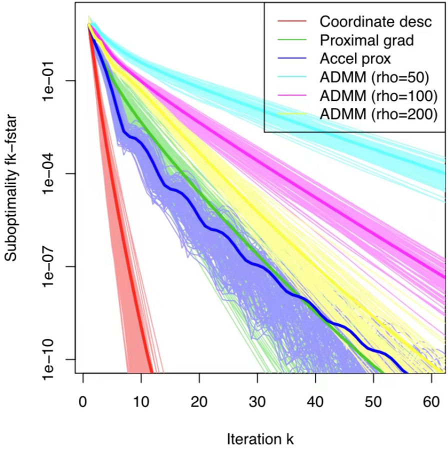
- The α update applies the soft-thresholding operator S_t defined as

$$[S_t(x)]_j = \begin{cases} x_j - t & \text{if } x_j - t \geq 0 \\ 0 & \text{if } -t < x_j < t \\ x_j + t & \text{if } x_j < -t \end{cases} \tag{16}$$

- Matrix $(X^TX + \rho I)$ is always invertible. Compute factorization (e.f. Cholesky) in $O(p^3)$ flops, then each β update takes $O(p^2)$ flops.

Convergence Result Compared to Other Method

An experiment with $n = 200$, $p = 50$ and 100 instances.



Practical Implementation

- Practical in application. In high dimensional data analysis, where we decompose the large-scale problem in a form of decomposition-coordination procedure, and then tackle the small local subproblems.
- Usually ADMM converges to modest accuracy within a few tens of iterations.
- However ADMM is slow to get high accuracy result.
- Such properties satisfy requirement from engineering view of point. For example, large scale machine learning problems require high convergence rate but is not demanding in a specific highly accurate result.

Thank You!