Lagrangian Methods for Constrained Optimization

Yinyu Ye
Department of Management Science and Engineering
Stanford University
Stanford, CA 94305, U.S.A.

http://www.stanford.edu/~yyye

LY: Chapter 14

The Lagrangian Function and Method

We consider

$$f^* := \min \quad f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0},\ \mathbf{x} \in X. \tag{1}$$

Recall that the Lagrangian function:

$$L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \mathbf{y}^T \mathbf{h}(\mathbf{x}).$$

and the dual function:

$$\phi(\mathbf{y}) = \min_{\mathbf{x} \in X} L(\mathbf{x}, \mathbf{y}); \tag{2}$$

and the dual problem

$$(f^* \geq) \phi^* := \max \quad \phi(\mathbf{y}). \tag{3}$$

In many cases, one can find $\mathbf{y}^*$ of dual problem (3), a unconstrained optimization problem; then compute $\mathbf{x}^*$ from (2).

The Local Duality Theorem

Suppose $\mathbf{x}^*$ is a local minimizer, and consider the localized problem

$$f(\mathbf{x}^*) := \min \quad f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0},\ \mathbf{x} \in X,\ \|\mathbf{x} - \mathbf{x}^*\|^2 \leq \epsilon. \qquad (4)$$

Then, the localized Lagrangian function:

$$L_{\mathbf{x}^*}(\mathbf{x}, \mathbf{y}, \mu) = f(\mathbf{x}) + \mathbf{y}^T\mathbf{h}(\mathbf{x}) + \mu(\|\mathbf{x} - \mathbf{x}^*\|^2 - \epsilon).$$

and the localized dual function:

$$\phi_{\mathbf{x}^*}(\mathbf{y}, \mu) = \min_{\mathbf{x} \in X,\ \|\mathbf{x} - \mathbf{x}^*\|^2 \leq \epsilon} L_{\mathbf{x}^*}(\mathbf{x}, \mathbf{y}, \mu); \qquad (5)$$

and the localized dual problem

$$\max \quad \phi(\mathbf{y}, \mu \geq 0). \qquad (6)$$

Under certain constraint qualification, we must have $f(\mathbf{x}^*) = \phi(\mathbf{y}^*, \mu^* = 0)$ where the localization constraint is inactive.

The gradient and Hessian of ϕ

Let $\mathbf{x}(\mathbf{y})$ be a minimizer of (2). Then

$$\phi(\mathbf{y}) = f(\mathbf{x}(\mathbf{y})) + \mathbf{y}^T\mathbf{h}(\mathbf{x}(\mathbf{y}))$$

Thus,

$$\begin{aligned}\nabla\phi(\mathbf{y}) &= \nabla f(\mathbf{x}(\mathbf{y}))\nabla\mathbf{x}(\mathbf{y}) + \mathbf{y}^T\nabla\mathbf{h}(\mathbf{x}(\mathbf{y}))\nabla\mathbf{x}(\mathbf{y}) + \mathbf{h}(\mathbf{x}(\mathbf{y})) \\ &= (\nabla f(\mathbf{x}(\mathbf{y})) + \mathbf{y}^T\nabla\mathbf{h}(\mathbf{x}(\mathbf{y})))\nabla\mathbf{x}(\mathbf{y}) + \mathbf{h}(\mathbf{x}(\mathbf{y})) \\ &= \mathbf{h}(\mathbf{x}(\mathbf{y})).\end{aligned}$$

Similarly, we can derive

$$\nabla^2\phi(\mathbf{y}) = -\nabla\mathbf{h}(\mathbf{x}(\mathbf{y}))\left(\nabla_x^2 L(\mathbf{x}(\mathbf{y}),\mathbf{y})\right)^{-1}\nabla\mathbf{h}(\mathbf{x}(\mathbf{y}))^T,$$

where $\nabla_x^2 L(\mathbf{x}(\mathbf{y}),\mathbf{y})$ is the Hessian of the Lagrangian function that is assumed to be positive definite at the (local) minimizer in the whole space.

An Example

Consider a toy problem

$$\begin{aligned} \text{minimize} \quad & (x_1-1)^2+(x_2-1)^2 \\ \text{subject to} \quad & x_1+2x_2-1=0, \\ & 2x_1+x_2-1=0. \end{aligned}$$

$$L(\mathbf{x},\mathbf{y})=(x_1-1)^2+(x_2-1)^2+y_1(x_1+2x_2-1)+y_2(2x_1+x_2-1).$$

$$x_1=-0.5y_1-y_2+1, \quad x_2=-y_1-0.5y_2+1.$$

$$\phi(\mathbf{y})=-1.25y_1^2-1.25y_2^2-2y_1y_2+2y_1+2y_2.$$

$$\nabla \phi(\mathbf{y}) = \begin{pmatrix} -2.5y_1 - 2y_2 + 2 \\ -2y_1 - 2.5y_2 1 + 2 \end{pmatrix},$$

$$\nabla^2 \phi(\mathbf{y}) = -\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}^T = -\begin{pmatrix} 2.5 & 2 \\ 2 & 2.5 \end{pmatrix}$$

The Augmented Lagrangian Function

In both theory and practice, we actually consider an augmented Lagrangian function (ALF)

$$L_a(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \mathbf{y}^T \mathbf{h}(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{h}(\mathbf{x})\|^2,$$

which corresponds to an equivalent problem of (1):

$$f^* := \min \quad f(\mathbf{x}) + \frac{\beta}{2}\|\mathbf{h}(\mathbf{x})\|^2 \quad \text{s.t.} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0},\ \mathbf{x} \in X.$$

Note that, although at feasibility the additional square term in objective is redundant, it helps to improve strict convexity of the Lagrangian function.

The Augmented Lagrangian Dual

Now the dual function:

$$\phi_a(\mathbf{y}) = \min_{\mathbf{x} \in X} L_a(\mathbf{x}, \mathbf{y}); \tag{7}$$

and the dual problem

$$(f^* \geq) \phi_a^* := \max \quad \phi_a(\mathbf{y}). \tag{8}$$

Note that the dual function satisfies $\frac{1}{\beta}$-Lipschitz condition (see Chapter 14 of LY).

For the convex optimization case, $\mathbf{h}(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$, we have

$$\nabla^2 L_a(\mathbf{x}, \mathbf{y}) = \nabla^2 f(\mathbf{x}) + \beta(A^T A).$$

The Augmented Lagrangian Method

The augmented Lagrangian method (ALM) is:

Start from any $(\mathbf{x}^0 \in X, \mathbf{y}^0)$, we compute a new iterate pair

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in X} L_a(\mathbf{x}, \mathbf{y}^k), \text{ and } \mathbf{y}^{k+1} = \mathbf{y}^k + \beta \mathbf{h}(\mathbf{x}^{k+1}).$$

The calculation of $\mathbf{x}$ is used to compute the gradient vector of $\phi_a(\mathbf{y})$, which is a steepest ascent direction.

The method converges just like the SDM, because the dual function satisfies $\frac{1}{\beta}$-Lipschitz condition.

Other SDM strategies may be adapted to update $\mathbf{y}$ (the BB ...).

Analysis of the Augmented Lagrangian Method

Consider the convex optimization case $\mathbf{h}(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$. Since $\mathbf{x}^{k+1}$ makes KKT condition:

$$
\begin{aligned}
\mathbf{0} &= \nabla f(\mathbf{x}^{k+1}) + A^T\mathbf{y}^k + \beta A^T(A\mathbf{x}^{k+1} - \mathbf{b}) \\
&= \nabla f(\mathbf{x}^{k+1}) + A^T(\mathbf{y}^k + \beta(A\mathbf{x}^{k+1} - \mathbf{b})) \\
&= \nabla f(\mathbf{x}^{k+1}) + A^T\mathbf{y}^{k+1},
\end{aligned}
$$

we only need to concern about whether or not $\|A\mathbf{x}^k - \mathbf{b}\|$ converges to zero and how fast it converges. First, from the convexity of $f(\mathbf{x})$, we have

$$
\begin{aligned}
\mathbf{0} \quad &\le (\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k))^T(\mathbf{x}^{k+1} - \mathbf{x}^k) \\
&= (-A^T\mathbf{y}^{k+1} + A^T\mathbf{y}^k)^T(\mathbf{x}^{k+1} - \mathbf{x}^k) \\
&= (-\mathbf{y}^{k+1} + \mathbf{y}^k)^T(A\mathbf{x}^{k+1} - A\mathbf{x}^k) \\
&= -\beta(A\mathbf{x}^{k+1} - \mathbf{b})(A\mathbf{x}^{k+1} - \mathbf{b} - (A\mathbf{x}^k - \mathbf{b})),
\end{aligned}
$$

which implies that

$$
\|A\mathbf{x}^{k+1} - \mathbf{b}\| \le \|A\mathbf{x}^k - \mathbf{b}\|,
$$

that is, the error is non-increasing.

Again, from the convexity, we have

$$\begin{aligned}
\mathbf{0} \;&\le (\nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^*))^T(\mathbf{x}^{k+1} - \mathbf{x}^*) \\
&= (-A^T\mathbf{y}^{k+1} + A^T\mathbf{y}^*)^T(\mathbf{x}^{k+1} - \mathbf{x}^*) \\
&= (-\mathbf{y}^{k+1} + \mathbf{y}^*)^T(A\mathbf{x}^{k+1} - A\mathbf{x}^*) = (-\mathbf{y}^{k+1} + \mathbf{y}^*)^T(A\mathbf{x}^{k+1} - \mathbf{b}) \\
&= \frac{1}{\beta}(\mathbf{y}^* - \mathbf{y}^{k+1})^T(\mathbf{y}^{k+1} - \mathbf{y}^k).
\end{aligned}$$

Thus, from the positivity of the cross product, we have

$$\begin{aligned}
\|\mathbf{y}^* - \mathbf{y}^k\|^2 \;&= \|\mathbf{y}^{k+1} - \mathbf{y}^k + \mathbf{y}^* - \mathbf{y}^{k+1}\|^2 \\
&\ge \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 + \|\mathbf{y}^* - \mathbf{y}^{k+1}\|^2 \\
&= \beta\|A\mathbf{x}^{k+1} - \mathbf{b}\|^2 + \|\mathbf{y}^* - \mathbf{y}^{k+1}\|^2.
\end{aligned}$$

Sum up from 0 to k of the inequality we have

$$\begin{aligned}\|\mathbf{y}^* - \mathbf{y}^0\|^2 &\geq \|\mathbf{y}^* - \mathbf{y}^{k+1}\|^2 + \beta \textstyle\sum_{l=0}^{k} \|A\mathbf{x}^{l+1} - \mathbf{b}\|^2 \\ &\geq \beta \textstyle\sum_{l=0}^{k} \|A\mathbf{x}^{l+1} - \mathbf{b}\|^2 \\ &\geq (k+1)\beta\|A\mathbf{x}^{k+1} - \mathbf{b}\|^2,\end{aligned}$$

where the last inequality from non-increasing property. Then, it gives the desired error bound:

$$\|A\mathbf{x}^{k+1} - \mathbf{b}\|^2 \leq \frac{1}{(k+1)\beta}\|\mathbf{y}^* - \mathbf{y}^0\|^2.$$

The Alternating Direction Method with Multipliers

For the ADMM method, we consider structured problem

$$\min \quad f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) \quad \text{s.t.} \quad A_1\mathbf{x}_1 + A_2\mathbf{x}_2 = \mathbf{b}.$$

Consider

$$L(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) = f_1(\mathbf{x}_1)+f_2(\mathbf{x}_2)+\mathbf{y}^T(A_1\mathbf{x}_1+A_2\mathbf{x}_2-\mathbf{b})+\frac{\beta}{2}\|A_1\mathbf{x}_1+A_2\mathbf{x}_2-\mathbf{b}\|^2.$$

Then, for any given $(\mathbf{x}_1^k, \mathbf{x}_2^k, \mathbf{y}^k)$, we compute a new iterate

$$\begin{aligned}
\mathbf{x}_1^{k+1} &= \arg\min_{\mathbf{x}_1} L(\mathbf{x}_1, \mathbf{x}_2^k, \mathbf{y}^k), \\
\mathbf{x}_2^{k+1} &= \arg\min_{\mathbf{x}_2} L(\mathbf{x}_1^{k+1}, \mathbf{x}_2, \mathbf{y}^k), \\
\mathbf{y}^{k+1} &= \mathbf{y}^k + \beta(A_1\mathbf{x}_1^{k+1} + A_2\mathbf{x}_2^{k+1} - \mathbf{b}).
\end{aligned}$$

Again, we can prove that the iterates converge with the same speed.

The ADMM method resembles the coordinate descent method ...

The ADMM method with three blocks

What about ADMM for

$$\min \quad f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_3(\mathbf{x}_3) \quad \text{s.t.} \quad A_1\mathbf{x}_1 + A_2\mathbf{x}_2 + A_3\mathbf{x}_3 = \mathbf{b},$$

where the Lagrangian function

$$\begin{aligned} L(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{y}) = \quad & f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_3(\mathbf{x}_3) + \mathbf{y}^T(A_1\mathbf{x}_1 + A_2\mathbf{x}_2 + A_3\mathbf{x}_3 - \mathbf{b}) \\ & + \frac{\beta}{2}\|A_1\mathbf{x}_1 + A_2\mathbf{x}_2 + A_3\mathbf{x}_3 - \mathbf{b}\|^2. \end{aligned}$$

Then, for any given $(\mathbf{x}_1^k, \mathbf{x}_2^k, \mathbf{x}_3^k, \mathbf{y}^k)$, we compute a new iterate

$$\begin{aligned} \mathbf{x}_1^{k+1} \quad & = \arg\min_{\mathbf{x}_1} L(\mathbf{x}_1, \mathbf{x}_2^k, \mathbf{x}_3^k, \mathbf{y}^k), \\ \mathbf{x}_2^{k+1} \quad & = \arg\min_{\mathbf{x}_2} L(\mathbf{x}_1^{k+1}, \mathbf{x}_2, \mathbf{x}_3^k, \mathbf{y}^k), \\ \mathbf{x}_3^{k+1} \quad & = \arg\min_{\mathbf{x}_3} L(\mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1}, \mathbf{x}_3, \mathbf{y}^k), \\ \mathbf{y}^{k+1} \quad & = \mathbf{y}^k + \beta(A_1\mathbf{x}^{k+1} + A_2\mathbf{x}_2^{k+1} + A_3\mathbf{x}_3^{k+1} - \mathbf{b}). \end{aligned}$$

Does it converges?

Consider the problem:

$$\min \quad 0\cdot x_1 + 0\cdot x_2 + 0\cdot x_3 \quad \text{s.t.} \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \mathbf{0},$$

The unique minimizer is $\mathbf{0}$.

Then, the ADMM with $\beta = 1$ would be a linear matrix mapping

$$\begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 4 & 6 & 0 & 0 & 0 & 0 \\ 5 & 7 & 9 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 2 & 0 & 1 & 0 \\ 1 & 2 & 2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \\ x_3^{k+1} \\ \mathbf{y}^{k+1} \end{pmatrix} = \begin{pmatrix} 0 & -4 & -5 & 1 & 1 & 1 \\ 0 & 0 & -7 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1^{k} \\ x_2^{k} \\ x_3^{k} \\ \mathbf{y}^{k} \end{pmatrix}.$$

which can be reduced to

$$\begin{pmatrix} x_2^{k+1} \\ x_3^{k+1} \\ \mathbf{y}^{k+1} \end{pmatrix} = M \begin{pmatrix} x_2^{k} \\ x_3^{k} \\ \mathbf{y}^{k} \end{pmatrix},$$

where

$$M = \frac{1}{162}\begin{pmatrix} 144 & -9 & -9 & -9 & 18 \\ 8 & 157 & -5 & 13 & -8 \\ 64 & 122 & 122 & -58 & -64 \\ 56 & -35 & -35 & 91 & -56 \\ -88 & -26 & -26 & -62 & 88 \end{pmatrix}.$$

But the spectral radius of the matrix is greater than 1, indicating the mapping is not a contraction.