Optimization Theory and Algorithm II	**October 24, 2022**
Lecture 13	
Lecturer:Xiangyu Chang	*Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 Alternating Direction Method of Multipliers

This part is summarized from the article [1].

1.1 Three Related Algorithms

Algorithm 1: Dual Gradient Ascent

Consider

$$\min_{\mathbf{x}} f(\mathbf{x}),$$
$$s.t.\ A\mathbf{x} - \mathbf{b} = 0.$$

Lagrangian: $L(\mathbf{x}, \boldsymbol{\nu}) = f(\mathbf{x}) + \boldsymbol{\nu}^\top (A\mathbf{x} - \mathbf{b})$. Thus,

$$g(\boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\nu}) = L(\mathbf{x}^*(\boldsymbol{\nu}), \boldsymbol{\nu}).$$

The dual problem is

$$\max_{\boldsymbol{\nu}} g(\boldsymbol{\nu}).$$

Because we have

$$\nabla g(\boldsymbol{\nu}) = \frac{\partial L}{\partial \mathbf{x}^*} \frac{\partial \mathbf{x}^*}{\partial \boldsymbol{\nu}} + \frac{\partial L}{\partial \boldsymbol{\nu}} = (A\mathbf{x} - \mathbf{b}),$$

where $\frac{\partial L}{\partial \mathbf{x}^*} = 0$. Based on that, the dual gradient assent algorithm is

$$\textbf{Step 1: } \mathbf{x}^t = \arg\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\nu}^t), \tag{1}$$
$$\textbf{Step 2: } \boldsymbol{\nu}^{t+1} = \boldsymbol{\nu}^t + s_t(A\mathbf{x}^t - \mathbf{b}). \tag{2}$$

The dual variable $\boldsymbol{\nu}$ can be interpreted as a vector of prices, and $\boldsymbol{\nu}$-update is called a "price update" step.

Algorithm 2: Dual Decomposition

The major benefit of the dual ascent method is that it can lead to a decentralized algorithm if f is separable. We consider

$$\min_{\mathbf{x}} f(\mathbf{x}) = \sum_{k=1}^{K} f_k(\mathbf{x}_k),$$
$$s.t.\ A\mathbf{x} = \sum_{k=1}^{K} A_k \mathbf{x}_k = \mathbf{b},$$

where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_K)^\top \in \mathbb{R}^n, \mathbf{x}_k \in \mathbb{R}^{n_k}, \sum_{k=1}^{K} \mathbf{x}_k = n$.

For Lagrangian:

$$
\begin{aligned}
L(\mathbf{x}, \boldsymbol{\nu}) &= \sum_{k=1}^{K} f_k(\mathbf{x}_k) + \boldsymbol{\nu}^\top (\sum_{k=1}^{K} A_k \mathbf{x}_k - \mathbf{b}) \\
&= \sum_{k=1}^{K} \underbrace{\left\{ f_k(\mathbf{x}_k) + \boldsymbol{\nu}^\top (A_k \mathbf{x}_k - \mathbf{b}/K) \right\}}_{:= L_k(\mathbf{x}_k, \boldsymbol{\nu})}.
\end{aligned}
$$

Algorithm:

$$
\begin{cases}
\mathbf{x}_k^{t+1} & = \arg\min_{\mathbf{x}_k} L_k(\mathbf{x}_k, \boldsymbol{\nu}^t), \\
\boldsymbol{\nu}^{t+1} & = \boldsymbol{\nu}^t + s_t (A\mathbf{x}^{t+1} - \mathbf{b}).
\end{cases}
$$

So, first we broadcast $\boldsymbol{\nu}^t$ to all threads. Then they compute each $\mathbf{x}_k^{t+1}$. Second, aggregate all $\mathbf{x}_k^{t+1}$ to obtain $\mathbf{x}^{t+1}$.

Algorithm 3: Method of Multipliers.

Consider

$$
\min_{\mathbf{x}} \ f(\mathbf{x}), \tag{3}
$$
$$
s.t. \ A\mathbf{x} - \mathbf{b} = 0. \tag{4}
$$

This is equivalent to

$$
\begin{aligned}
&\min_{\mathbf{x}} \ f(\mathbf{x}) + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{b}\|^2, \\
&s.t. \ A\mathbf{x} - \mathbf{b} = 0.
\end{aligned}
$$

The Lagrangian is called the **augmented Lagrangian** of (3). Denoted as

$$
L_\rho(\mathbf{x}, \boldsymbol{\nu}) = f(\mathbf{x}) + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{b}\|^2 + \boldsymbol{\nu}^\top (A\mathbf{x} - \mathbf{b}).
$$

Based on that, the dual gradient assent algorithm is

$$
\textbf{Step 1: } \mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \boldsymbol{\nu}^t), \tag{5}
$$
$$
\textbf{Step 2: } \boldsymbol{\nu}^{t+1} = \boldsymbol{\nu}^t + \rho (A\mathbf{x}^{t+1} - \mathbf{b}). \tag{6}
$$

Remark 1

- $\mathbf{x}$*-update adopts* L_ρ *is not* L*.*
- *Step size is* ρ *is not* s_t*.*
- *This is called "method of multiplers" (MM).*

Lemma 1 *Suppose that* $\mathbf{x}^{t+1}$ *is generated from MM via* $\boldsymbol{\nu}^t$*, then show that* $\mathbf{x}^{t+1}$ *is the stationary point of* $L(\mathbf{x}, \boldsymbol{\nu}^{t+1})$*.*

Proof 1 *We know that* $\mathbf{x}^{t+1}$ *minimizes* $L_\rho(\mathbf{x}, \boldsymbol{\nu}^t)$*, then*

$$
\begin{aligned}
\nabla_{\mathbf{x}} L_\rho(\mathbf{x}^{t+1}, \boldsymbol{\nu}^t) &= \nabla f(\mathbf{x}^{t+1}) + A^\top \boldsymbol{\nu}^t + \rho A^\top (A\mathbf{x}^{t+1} - \mathbf{b}) \\
&= \nabla f(\mathbf{x}^{t+1}) + A^\top (\boldsymbol{\nu}^t + \rho (A\mathbf{x}^{t+1} - \mathbf{b})) \\
&= \nabla f(\mathbf{x}^{t+1}) + A^\top \boldsymbol{\nu}^{t+1} = \nabla L(\mathbf{x}^{t+1}, \boldsymbol{\nu}^{t+1}) = 0.
\end{aligned}
$$

Q: When f is separable, then augmented Lagrangian L_ρ is not separable. So that $\mathbf{x}$-minimization step cannot be carried out separately in parallel for each $\mathbf{x}_i$. How to address this issue?

1.2 ADMM

Let us consider the following convex optimization problem:

$$\min_{\mathbf{x},\mathbf{z}} \; f(\mathbf{x}) + g(\mathbf{z}) \tag{7}$$

$$s.t. \; A\mathbf{x} + B\mathbf{z} = \mathbf{c}, \tag{8}$$

where $\mathbf{x} \in \mathbb{R}^{n_1}, \mathbf{z} \in \mathbb{R}^{n_2}, n_1 + n_2 = n, A \in \mathbb{R}^{m\times n_1}$ and $B \in \mathbb{R}^{m\times n_2}$. Further assume that f and g are convex.

The only difference form the general linear equality constrained problem is that the variables $\mathbf{x}, \mathbf{z}$ can be viewed splitted variable from a big one.

Example 1

$$\min_{\mathbf{x}} \; f_1(\mathbf{x}) + f_2(\mathbf{x}).$$

This is equivalent to

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{z}} \; f_1(\mathbf{x}) + f_2(\mathbf{z}), \\ &s.t. \; \mathbf{x} - \mathbf{z} = 0. \end{aligned}$$

Example 2

$$\min_{\mathbf{x}} \; f_1(\mathbf{x}) + f_2(A\mathbf{x}).$$

This is equivalent to

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{z}} \; f_1(\mathbf{x}) + f_2(\mathbf{z}), \\ &s.t. \; A\mathbf{x} - \mathbf{z} = 0. \end{aligned}$$

Example 3

$$\begin{aligned} &\min_{\mathbf{x}} \; f(\mathbf{x}), \\ &s.t. \; \mathbf{x} \in \mathcal{X}. \end{aligned}$$

This is equivalent to

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{z}} \; f(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{z}), \\ &s.t. \; \mathbf{x} - \mathbf{z} = 0. \end{aligned}$$

Example 4 *Global consensus problem is*

$$\min_{\mathbf{x}} \sum_{j=1}^{J} f_j(\mathbf{x}).$$

This is equivalent to

$$\begin{aligned} &\min_{\mathbf{x}_i,\mathbf{x}} \; \sum_{j=1}^{J} f_j(\mathbf{x}_j)., \\ &s.t. \; \mathbf{x}_j - \mathbf{x} = 0. \end{aligned}$$

Actually, the problem (7) can be solved by MM. Its augmented Lagrangian is

$$L_\rho(\mathbf{x},\mathbf{z},\boldsymbol{\nu}) = f(\mathbf{x}) + g(\mathbf{z}) + \boldsymbol{\nu}^\top(A\mathbf{x}+B\mathbf{z}-\mathbf{c}) + \frac{\rho}{2}\|A\mathbf{x}+B\mathbf{z}-\mathbf{c}\|^2.$$

$$\begin{cases} (\mathbf{x}^{t+1},\mathbf{z}^{t+1}) = \arg\min_{\mathbf{x},\mathbf{z}} L_\rho(\mathbf{x},\mathbf{z},\boldsymbol{\nu}^t), \\ \boldsymbol{\nu}^{t+1} = \boldsymbol{\nu}^t + \rho(A\mathbf{x}^{t+1}+b\mathbf{z}^{t+1}-\mathbf{c}). \end{cases}$$

This formulation cannot be decomposed.

So, the ADMM algorithm is

$$\begin{cases} \mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} L_\rho(\mathbf{x},\mathbf{z}^t,\boldsymbol{\nu}^t), \\ \mathbf{z}^{t+1} = \arg\min_{\mathbf{z}} L_\rho(\mathbf{x}^{t+1},\mathbf{z},\boldsymbol{\nu}^t), \\ \boldsymbol{\nu}^{t+1} = \boldsymbol{\nu}^t + \rho(A\mathbf{x}^{t+1}+b\mathbf{z}^{t+1}-\mathbf{c}). \end{cases}$$

This is called "unscaled form". The corresponding "scaled form" is

$$\boldsymbol{\nu}^\top(A\mathbf{x}+B\mathbf{z}-\mathbf{c}) + \frac{\rho}{2}\|A\mathbf{x}+B\mathbf{z}-\mathbf{c}\|^2 = \frac{\rho}{2}\|A\mathbf{x}+B\mathbf{z}-\mathbf{c}+\boldsymbol{\nu}/\rho\|^2 - \frac{\rho}{2}\|\boldsymbol{\nu}/\rho\|^2.$$

Let $\mathbf{u} = \boldsymbol{\nu}/\rho$, then the so-called scaled form of ADMM is

$$\begin{cases} \mathbf{x}^{t+1} = \arg\min_{\mathbf{x}}(f(\mathbf{x}) + \frac{\rho}{2}\|A\mathbf{x}+B\mathbf{z}^t-\mathbf{c}+\mathbf{u}^t\|^2), \\ \mathbf{z}^{t+1} = \arg\min_{\mathbf{z}}(g(\mathbf{z}) + \frac{\rho}{2}\|A\mathbf{x}^{t+1}+B\mathbf{z}-\mathbf{c}+\mathbf{u}^t\|^2), \\ \mathbf{u}^{t+1} = \mathbf{u}^t + A\mathbf{x}^{t+1} + B\mathbf{z}^{t+1} - \mathbf{c}. \end{cases}$$

Example 5 *(LAD Regression)*

$$\min_{\mathbf{x}} \|A\mathbf{x}-\mathbf{b}\|_1.$$

This is equivalent to

$$\min_{\mathbf{x},\mathbf{z}} \ \|\mathbf{z}\|_1, \qquad s.t.\ A\mathbf{x}-\mathbf{z} = \mathbf{b}.$$

Based on ADMM algorithm, it has

$$\begin{aligned} \mathbf{x}^{t+1} &= \arg\min_{\mathbf{x}} \frac{\rho}{2}\|A\mathbf{x}-\mathbf{z}^t-\mathbf{b}+\mathbf{u}^t\|^2 \\ &= (A^\top A)^{-1}A^\top(\mathbf{z}^t+\mathbf{b}-\mathbf{u}^t). \end{aligned}$$

$$\begin{aligned} \mathbf{z}^{t+1} &= \arg\min_{\mathbf{z}} \left\{\|\mathbf{z}\|_1 + \frac{\rho}{2}\|A\mathbf{x}^{t+1}-\mathbf{z}-\mathbf{b}+\mathbf{u}^t\|^2\right\} \\ &= S_{1/\rho}(A\mathbf{x}^{t+1}-\mathbf{b}+\mathbf{u}^t), \end{aligned}$$

where $S_{1/\rho}$ is the soft thresholding function. For $\mathbf{u}^{t+1}$,

$$\mathbf{u}^{t+1} = \mathbf{u}^t + A\mathbf{x}^{t+1} - \mathbf{z}^{t+1} - \mathbf{b}.$$

Example 6 *(LASSO)*

$$\min_{\mathbf{x}} \ \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1.$$

This is equivalent to

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2+\lambda\|\mathbf{z}\|_1,$$
$$s.t.\ \mathbf{x}-\mathbf{z}=0.$$

Based on ADMM algorithm, it has

$$\begin{aligned}\mathbf{x}^{t+1} &= \arg\min_{\mathbf{x}}\left\{\frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2+\frac{\rho}{2}\|\mathbf{x}-\mathbf{z}^t+\mathbf{u}^t\|^2\right\}\\ &= (A^\top A+\rho I)^{-1}(A^\top\mathbf{b}+\rho(\mathbf{z}^t-\mathbf{u}^t)).\end{aligned}$$

$$\begin{aligned}\mathbf{z}^{t+1} &= \arg\min_{\mathbf{z}}\left\{\lambda\|\mathbf{z}\|_1+\frac{\rho}{2}\|\mathbf{x}^{t+1}-\mathbf{z}+\mathbf{u}^t\|^2\right\}\\ &= S_{\lambda/\rho}(\mathbf{x}^{t+1}+\mathbf{u}^t),\end{aligned}$$

where $S_{\lambda/\rho}$ *is the soft thresholding function. For* $\mathbf{u}^{t+1}$,

$$\mathbf{u}^{t+1}=\mathbf{u}^t+\mathbf{x}^{t+1}-\mathbf{z}^{t+1}.$$

Example 7

$$\min_{\mathbf{x}} f(\mathbf{x}),$$
$$s.t.\ \mathbf{x}\in\mathcal{X}.$$

This is equivalent to

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x})+\delta_{\mathcal{X}}(\mathbf{z}),$$
$$s.t.\ \mathbf{x}-\mathbf{z}=0.$$

Based on ADMM algorithm, it has

$$\begin{aligned}\mathbf{x}^{t+1} &= \arg\min_{\mathbf{x}}\left\{f(\mathbf{x})+\frac{\rho}{2}\|\mathbf{x}-\mathbf{z}^t+\mathbf{u}^t\|^2\right\}\\ &= prox_{f/\rho}(\mathbf{z}^t-\mathbf{u}^t).\end{aligned}$$

$$\begin{aligned}\mathbf{z}^{t+1} &= \arg\min_{\mathbf{z}}\left\{\delta_{\mathcal{X}}(\mathbf{z})+\frac{\rho}{2}\|\mathbf{x}^{t+1}-\mathbf{z}+\mathbf{u}^t\|^2\right\}\\ &= \pi_{\mathcal{X}}(\mathbf{x}^{t+1}+\mathbf{u}^t),\end{aligned}$$

where $\pi_{\mathcal{X}}$ *is the projection function. For* $\mathbf{u}^{t+1}$,

$$\mathbf{u}^{t+1}=\mathbf{u}^t+\mathbf{x}^{t+1}-\mathbf{z}^{t+1}.$$

- *Non-negative Least Squares:* $f(\mathbf{x})=\frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2, \mathcal{X}=\{\mathbf{x}|\mathbf{x}\succeq 0\}$.
- *Ridge:* $f(\mathbf{x})=\frac{1}{2}\|A\mathbf{x}-\mathbf{b}\|^2, \mathcal{X}=\{\mathbf{x}|\|\mathbf{x}\|\le t\}$.
- *Basis Pursuit:* $f(\mathbf{x})=\|\mathbf{x}\|_1, \mathbf{X}=\{\mathbf{x}|A\mathbf{x}=\mathbf{b}\}$. *Then*

$$\begin{cases}\mathbf{x}^{t+1}=S_{1/\rho}(\mathbf{z}^t-\mathbf{u}^t),\\ \mathbf{z}^{t+1}=\pi_{\mathcal{X}}(\mathbf{x}^{t+1}+\mathbf{u}^t)=(I-A(AA^\top)^{-1}A)(\mathbf{x}^{t+1}+\mathbf{u}^t)+A^\top(AA^\top)^{-1}\mathbf{b}.\end{cases}$$

1.3 Optimality Conditions of ADMM

For the convex optimization problem (7), we have the necessary and sufficient optimality conditions for it as

$$\nabla f(\mathbf{x}^*) + A^\top \boldsymbol{\nu}^* = 0, \tag{9}$$
$$\nabla g(\mathbf{z}^*) + A^\top \boldsymbol{\nu}^* = 0, \tag{10}$$
$$A\mathbf{x}^* + B\mathbf{z}^* - \mathbf{c} = 0. \tag{11}$$

For (10), we know that $\mathbf{z}^{t+1}$ minimizes $L_\rho(\mathbf{x}^{t+1}, \mathbf{z}, \boldsymbol{\nu}^t)$, then

$$\begin{aligned} 0 &= \nabla g(\mathbf{z}^{t+1}) + B^\top \boldsymbol{\nu}^t + \rho B^\top (A\mathbf{x}^{t+1} + B\mathbf{z}^{t+1} - \mathbf{c}) \\ &= \nabla g(\mathbf{z}^{t+1}) + B^\top (\boldsymbol{\nu}^t + \rho(A\mathbf{x}^{t+1} + B\mathbf{z}^{t+1} - \mathbf{c})) \\ &= \nabla g(\mathbf{z}^{t+1}) + B^\top \boldsymbol{\nu}^{t+1}. \end{aligned}$$

So, $(\mathbf{z}^{t+1}, \boldsymbol{\nu}^{t+1})$ satisfies (10) in the KKT conditions.

For (9), we know that $\mathbf{x}^{t+1}$ minimizes $L_\rho(\mathbf{x}, \mathbf{z}^t, \boldsymbol{\nu}^t)$, then

$$\begin{aligned} 0 &= \nabla f(\mathbf{x}^{t+1}) + A^\top \boldsymbol{\nu}^t + \rho A^\top (A\mathbf{x}^{t+1} + B\mathbf{z}^t - \mathbf{c}) \\ &= \nabla f(\mathbf{x}^{t+1}) + A^\top (\boldsymbol{\nu}^t + \rho(A\mathbf{x}^{t+1} + B\mathbf{z}^{t+1} - \mathbf{c})) + \rho A^\top B(\mathbf{z}^t - \mathbf{z}^{t+1}) \\ &= \nabla f(\mathbf{x}^{t+1}) + A^\top \boldsymbol{\nu}^{t+1} + \rho A^\top B(\mathbf{z}^t - \mathbf{z}^{t+1}). \end{aligned}$$

Thus,

$$S^{t+1} := \rho A^\top B(\mathbf{z}^{t+1} - \mathbf{z}^t) = \nabla f(\mathbf{x}^{t+1}) + A^\top \boldsymbol{\nu}^{t+1},$$

this is called “dual residual”. Furthermore, define

$$R^{t+1} = A\mathbf{x}^{t+1} + B\mathbf{z}^t - \mathbf{c}$$

as “primal residual”.

The stopping conditions of ADMM should be

$$\|S^{t+1}\| \le \epsilon, \ \|R^{t+1}\| \le \epsilon. \tag{12}$$

When $\epsilon \to 0$, then KKT conditions are satisfied.

References

[1] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.