Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers

Alternating Direction Method of Multipliers

Presented by: Weiyu Li, Jincheng Pang

2018.05
Focus

a. Duality

b. Constraints

c. ADMM

c.1. Dual Ascent

c.2. Augmented Lagrangian Methods

c.3. Alternating Direction Method of Multipliers
Dual Ascent

Aim

\[
\begin{align*}
\text{minimize} & \quad f(x), \quad x \in \mathbb{R}^n \\
\text{subject to} & \quad Ax = b,
\end{align*}
\]

where \( A \in \mathbb{R}^{m \times n} \), and \( f \) is convex.

Lagrangean

- \( L(x, \lambda) = f(x) + \lambda^T(Ax - b) \).
- **Dual function** \( g(\lambda) = \inf_x L(x, \lambda) = -f^*(-A^T \lambda) - b^T \lambda \).
- **Dual problem**
  \[
  \text{maximize} \quad g(\lambda), \quad \lambda \in \mathbb{R}^m.
  \]

If the strong duality holds, and there’s only one minimizer of \( L(x, \lambda^*), \) then

\[
x^* = \arg \min_x L(x, \lambda^*).
\]
Algorithm 1: *Dual ascent method.*

1. Initialize $x^{(0)}$, $\lambda^{(0)}$, step size $\alpha^{(0)} > 0$.

2. Update until stopping criteria are met:

   $x^{(k+1)} = \arg \min_x L(x, \lambda^{(k)})$

   \hspace{1cm} (3)

   $\lambda^{(k+1)} = \lambda^{(k)} + \alpha^{(k)}(Ax^{(k+1)} - b)$.

   \hspace{1cm} (4)

Remarks

1. Update of $\lambda$ can be interpreted as *price* update.

2. Motivation for the update of $\lambda$: we can show that $\nabla g(\lambda) = Ax - b$.

3. This algorithm is called dual ascent since, with appropriate choice of $\alpha^{(k)}$, the dual function $g$ increases in each step.
Dual Decomposition

Suppose $f$ is *separable*, i.e., $f(x) = \sum_{i=1}^{N} f_i(x_i)$, where $x_i \in \mathbb{R}^{n_i}$ are subvectors of $x$, then the Lagrangian can be written as

$$L(x, \lambda) = \sum_{i=1}^{N} L_i(x_i, \lambda) := \sum_{i=1}^{N} \left( f_i(x_i) + \lambda^T A_i x_i - \frac{1}{N} \lambda^T b \right),$$

(5)

where $A_i$ are partitioned conformably with $x_i$.

**Algorithm 1**: *Dual ascent method.*

1. Initialize $x^{(0)}$, $\lambda^{(0)}$, step size $\alpha^{(0)} > 0$.

2. Update until stopping criteria are met:

   $$x_i^{(k+1)} = \arg \min_x L_i(x_i, \lambda^{(k)}), \quad i = 1, \ldots, N$$

   (6)

   $$\lambda^{(k+1)} = \lambda^{(k)} + \alpha^{(k)} (Ax^{(k+1)} - b).$$

   (7)

In general, each iteration requires a *broadcast* and a *gather* operation.
Augmented Lagrangian Methods

ALM brings robustness to the dual ascent method, and converges with less assumptions.

The augmented Lagrangian for (1) is

\[ L_\rho(x, \lambda) = f(x) + \lambda^T (Ax - b) + \frac{\rho}{2} \|Ax - b\|^2_2, \]  

where \( \rho > 0 \) is the penalty parameter.

Interpretation

The augmented Lagrangian can be regarded as the Lagrangian of the equivalent form to the original problem:

\[
\begin{align*}
\text{minimize} & \quad f(x) + \frac{\rho}{2} \|Ax - b\|^2_2 \\
\text{subject to} & \quad Ax = b.
\end{align*}
\]
Algorithm 2: Method of multipliers.

1. Initialize $x^{(0)}$, $\lambda^{(0)}$, $\rho > 0$.

2. Update until stopping criteria are met:

   \begin{align*}
   x^{(k+1)} &= \arg \min_x L_\rho(x, \lambda^{(k)}) \quad (10) \\
   \lambda^{(k+1)} &= \lambda^{(k)} + \rho(Ax^{(k+1)} - b). \quad (11)
   \end{align*}

Remark

Motivation for the update of $\lambda$:

\begin{align*}
0 &= \nabla_x L_\rho(x^{(k+1)}, \lambda^{(k)}) \\
  &= \nabla_x f(x^{(k+1)}) + A^T \lambda^{(k)} + \rho A^T (Ax^{(k+1)} - b) \\
  &= \nabla_x f(x^{(k+1)}) + A^T \lambda^{(k+1)}.
\end{align*}

Notice that the KKT conditions are:

\begin{align*}
Ax^* - b &= 0, \quad \nabla f(x^*) + A^T \lambda^* = 0. \quad (12)
\end{align*}

So using $\rho$ as the step size makes the pair $(x^{(k+1)}, \lambda^{(k+1)})$ dual feasible.
Pro & Cons of ALM

- ALM converges with less assumptions, for example, even when $f$ takes on the value $+\infty$ or is not strictly convex.
- However, when $f$ is separable, $L_\rho$ is not separable due to the augmented term.

Therefore, we introduce ADMM, which intends to blend the decomposability of dual ascent with the superior convergence properties.
Alternating Direction Method of Multipliers

Aim

\[
\text{minimize} \quad f(x) + g(z), \quad x \in \mathbb{R}^n, z \in \mathbb{R}^m \\
\text{subject to} \quad Ax + Bz = c,
\]

where \( A \in \mathbb{R}^{p \times n}, B \in \mathbb{R}^{p \times m}, c \in \mathbb{R}^p \), and \( f, g \) are convex.

The augmented Lagrangian

\[
L_\rho(x, z, \lambda) = f(x) + g(z) + \lambda^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2.
\]
Algorithm 3: *ADMM*.

1. Initialize $x^{(0)}$, $z^{(0)}$, $\lambda^{(0)}$, $\rho > 0$.

2. Update until stopping criteria are met:

\[
x^{(k+1)} = \arg \min_x L_\rho(x, z^{(k)}, \lambda^{(k)})
\]
\[
z^{(k+1)} = \arg \min_x L_\rho(x^{(k+1)}, z, \lambda^{(k)})
\]
\[
\lambda^{(k+1)} = \lambda^{(k)} + \rho(Ax^{(k+1)} + Bz^{(k+1)} - c).
\]

Remarks

1. ALM can be regarded as

\[
(x^{(k+1)}, z^{(k+1)}) = \arg \min_{x,z} L_\rho(x, z, \lambda^{(k)})
\]
\[
\lambda^{(k+1)} = \lambda^{(k)} + \rho(Ax^{(k+1)} + Bz^{(k+1)} - c).
\]

2. The roles of $x$ and $z$ are not quite symmetric.
Defining the residual \( r := Ax + Bz - c \), we have

\[
\lambda^T r + \frac{\rho}{2} \| r \|_2^2 = \frac{\rho}{2} \| r + u \|_2^2 - \frac{\rho}{2} \| u \|_2^2,
\]  

(18)

where \( u := \frac{1}{\rho} \lambda \) is the \textit{scaled dual variable}.

Hence the update can be expressed as

\[
x^{(k+1)} = \arg \min_x \left( f(x) + \frac{\rho}{2} \| Ax + Bz^{(k)} - c + u^{(k)} \|_2^2 \right)
\]

(19)

\[
z^{(k+1)} = \arg \min_x \left( g(z) + \frac{\rho}{2} \| Ax^{(k+1)} + Bz - c + u^{(k)} \|_2^2 \right)
\]

(20)

\[
u^{(k+1)} = u^{(k)} + r^{(k+1)} = u^{(0)} + \sum_{j=1}^{k+1} r^{(j)}.
\]

(21)
For simplicity, suppose $f$ and $g$ are differentiable.

KKT for the ADMM problem

\[
Ax^* + Bz^* - c = 0 \quad \text{(primal feasibility)}
\]
\[
\nabla f(x^*) + A^T \lambda^* = 0, \quad \nabla g(z^*) + B^T \lambda^* = 0. \quad \text{(dual feasibility)}
\] (22)

By definition of $z^{(k+1)}$, we have that

\[
0 = \nabla g(z^{(k+1)}) + B^T \lambda^{(k)} + \rho B^T (Ax^{(k+1)} + Bz^{(k+1)} - c)
\]
\[
= \nabla g(z^{(k+1)}) + B^T \lambda^{(k+1)}.
\]

Analogously, for $x^{(k+1)}$ we have that

\[
0 = \nabla f(x^{(k+1)}) + A^T \lambda^{(k)} + \rho A^T (Ax^{(k+1)} + Bz^{(k)} - c)
\]
\[
= \nabla f(x^{(k+1)}) + A^T \lambda^{(k+1)} + \rho A^T B(z^{(k)} - z^{(k+1)}).
\]

Define the dual residual $s^{(k+1)} := \rho A^T B(z^{(k+1)} - z^{(k)})$, and primal residual $r^{(k+1)} := Ax^{(k+1)} + Bz^{(k+1)} - c$. These two residuals converge to 0 as ADMM proceeds. Therefore, stopping criteria can be derived from the residuals.
Extensions and Variations

• Varying penalty parameter $\rho$ in each iterations.

• General Augmenting terms: replacing $\frac{\rho}{2}\|r\|^2_2$ with $\frac{1}{2}r^TPr$.
  (Or replace the constraints $r = 0$ with $Fr = 0$, where $F^TF = P$.)

• Inexact minimization.

• Update ordering.
Example

Lasso

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|z\|_1 \\
\text{subject to} & \quad x - z = 0.
\end{align*}
\]  

(23)

Update:

\[
\begin{align*}
x^{(k+1)} &= (A^T A + \rho I)^{-1} \left( A^T b + \rho (z^{(k)} - u^{(k)}) \right) \\
z^{(k+1)} &= S_{\lambda/\rho}(x^{(k+1)} + u^{(k)}) \\
u^{(k+1)} &= u^{(k)} + x^{(k+1)} - z^{(k+1)}.
\end{align*}
\]  

(24)

(25)

(26)

where \( S \) is the \textit{soft thresholding operator} \( S_\kappa(a) = \begin{cases} a - \kappa, & a > \kappa \\ 0, & |a| \leq \kappa \\ a + \kappa, & a < -\kappa \end{cases} \).
Best Subset Selection

\[
\begin{align*}
\text{minimize} \quad & \|Ax - b\|_2^2 \\
\text{subject to} \quad & \text{card}(x) \leq K.
\end{align*}
\] (27)

Update:

\[
\begin{align*}
x^{(k+1)} &= \arg\min_x \left( \|Ax - b\|_2^2 + \frac{\rho}{2} \|x - z^{(k)} + u^{(k)}\|_2^2 \right) \\
z^{(k+1)} &= \Pi_K(x^{(k+1)} + u^{(k)}) \\
u^{(k+1)} &= u^{(k)} + x^{(k+1)} - z^{(k+1)}.
\end{align*}
\] (28, 29, 30)

where \(\Pi_K(v)\) keeps the \(K\) largest magnitude elements and zeroes out the rest.