Introduction to Alternating Direction Method of Multipliers

Yale Chang

Machine Learning Group Meeting

September 29, 2016
Formulation

\[ \min_{x,z} f(x) + g(z) \quad \text{s.t.} \quad Ax + Bz = c \]  

(1)

- where \( x \in \mathbb{R}^n, z \in \mathbb{R}^m, A \in \mathbb{R}^{p \times n}, B \in \mathbb{R}^{p \times m}, c \in \mathbb{R}^p \)
- \( f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) and \( g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\} \) are closed, proper, and convex, or equivalently, the epigraph of \( f, g \) are closed nonempty convex sets (epigraph \( f = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} | f(x) \leq t\} \))
- \( f, g \) can be non-differentiable and can take value \(+\infty\) (e.g. indicator function \( \mathbb{I}(x) = 0 \) if \( x \in C \) and \( \mathbb{I}(x) = +\infty \) otherwise)
- We also assume the strong duality condition holds (Slater’s condition)
The Augmented Lagrangian of problem (1) is

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

(2)

The ADMM iterations are

\[ x^{k+1} = \arg \min_x L_\rho(x, z^k, y^k) \]  

(3)

\[ z^{k+1} = \arg \min_z L_\rho(x^{k+1}, z, y^k) \]  

(4)

\[ y^{k+1} = y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \]  

(5)

The algorithm state is \((z^k, y^k)\)
Let $r = Ax + Bz - c$, $u = y/\rho$, the augmented Lagrangian can be written as

$$L_\rho(x, z, y) = f(x) + g(z) + y^T r + \rho/2 ||r||_2^2$$
$$= f(x) + g(z) + \rho/2 ||r + y/\rho||_2^2 - 1/(2\rho)||y||_2^2$$
$$= f(x) + g(z) + \rho/2 ||r + u||_2^2 - \rho/2 ||u||_2^2$$

The ADMM iterations become

$$x^{k+1} = \arg \min_x f(x) + \rho/2 ||Ax + Bz^k - c + u^k||_2^2$$  \hspace{1cm} (6)$$
$$z^{k+1} = \arg \min_z g(z) + \rho/2 ||Ax^{k+1} + Bz - c + u^k||_2^2$$  \hspace{1cm} (7)$$
$$u^{k+1} = u^k + Ax^{k+1} + Bz^{k+1} - c$$  \hspace{1cm} (8)$$

In practice, we prefer the scaled form because it’s shorter and easier to work with.
Stopping Criteria

The following inequality always holds

\[ f(x^k) + g(z^k) - p^* \leq -(y^k)^T r^k + (x^k - x^*)^T s^k \quad (9) \]

- where \( p^* = \inf \{ f(x) + g(z) \mid Ax + Bz = c \} \), \( x^* \) is the optimal value of \( x \) when \( f(x) + g(z) \) achieves the optimal objective \( p^* \)
- \( r^k = Ax^k + Bz^k - c \) is called \textit{primal residual}, \( s^k = \rho A^T B(z^k - z^{k-1}) \) is called \textit{dual residual}
- If \( \|r^k\|_2 \leq \epsilon^{\text{primal}} \) and \( \|s^k\|_2 \leq \epsilon^{\text{dual}} \) both hold, the iterations can stop
- There are some suggestions on how to set \( \epsilon^{\text{primal}} \) and \( \epsilon^{\text{dual}} \) in practice, you can refer to Boyd’s tutorial for details
- We will skip the proof of convergence for now
How to solve each individual step in the iterations? The $x$-update step can be expressed as

$$x^+ = \arg \min_x f(x) + (\rho/2)\|Ax - v\|_2^2$$  \hspace{1cm} (10)$$

where $v = -Bz + c - u$ is a known constant in this step.

**Case 1: Proximal Operator:** assume $A = I$, which appears frequently in examples, we have

$$x^+ = \arg \min_x f(x) + (\rho/2)\|x - v\|_2^2 = \text{prox}_{f,\rho}(v)$$

where $\text{prox}_{f,\rho}$ is called the proximal operator of $f$ with penalty $\rho$, when the function $f$ is simple enough, the proximal operator can be evaluated analytically.
**Indicator Function:** \( f(x) = 0 \) if \( x \in C \) and \( f(x) = +\infty \) otherwise, where \( C \) is a closed nonempty convex set, then the \( x \)-update is

\[
x^+ = \text{prox}_{f,\rho}(v) = \Pi_C(v)
\]

where \( \Pi_C \) denotes the projection onto set \( C \) (in the Euclidean norm).

1) Hyper-rectangle \( C = \{x \mid l \leq x \leq u\} \) (Reall the unit ball of \( \ell_\infty \))

\[
\implies (\Pi_C(v))_k = \begin{cases} l_k, & v_k \leq l_k \\ v_k, & l_k \leq v_k \leq u_k \\ u_k, & v_k \geq u_k \end{cases}
\]

2) Affine set \( C = \{x \in \mathbb{R}^n \mid Ax = b\} \) (\( A \in \mathbb{R}^{m \times n} \))

\[
\implies \Pi_C(v) = v - A^\dagger (Av - b) \quad (A^\dagger \text{ is pseudoinverse of } A)
\]

3) Positive semidefinite cone \( C = S^n_+ \)

\[
\implies \Pi_C(V) = \sum_{i=1}^n (\lambda_i) u_i u_i^T, \text{ where } V = \sum_{i=1}^n \lambda_i u_i u_i^T
\]

More examples can be found from “*Proximal Algorithms*” by Parikh and Boyd.
General Patterns: Quadratic Objective

Suppose \( f(x) = \frac{1}{2}x^T Px + q^T x + r \) \((P \in \mathbb{S}_+^n)\), the \( x \)-update becomes

\[
x^+ = \arg \min_x \frac{1}{2}x^T Px + q^T x + r + \frac{\rho}{2}||Ax - v||^2_2
\]

\[
= \arg \min_x \frac{1}{2}x^T (P + \rho A^T A)x - (\rho A^T v - q)^T x
\]

\[
= (P + \rho A^T A)^{-1}(\rho A^T v - q)
\]

Here we assume \( F = P + \rho A^T A \) is invertible. Computing \( x^+ \) requires solving for the linear system \( Fx = g \) \((g = \rho A^T v - q)\).

- **Exploiting Sparsity:** when \( F \) is sparse, the computation cost can be effectively reduced
- **Caching:** when multiple linear systems with the same coefficient matrix \( F \) need to be solved, the factorization of \( F \) can be computed once and saved for later use
When $f$ is smooth, general iterative methods can be used to carry out the $x$-minimization step, examples include gradient descent, conjugate gradient method and L-BFGS etc. The presence of the quadratic penalty term tends to improve the convergence of the iterative algorithms. There are a few techniques to speed up the convergence of ADMM iterations:

- **Early Termination**: early termination in the $x$- or $z$-updates can result in more ADMM iterations, but lower cost per iteration, giving an overall improvement in efficiency.

- **Warm Start**: initialize the iterative method using the solution obtained from the previous iteration.
**Block Separability:** Suppose \( x \in \mathbb{R}^n \) can be partitioned into \( N \) subvectors and that \( f \) is separable w.r.t. this partition, i.e.

\[
f(x) = f_1(x_1) + \cdots + f_N(x_N)
\]

where \( x_i \in \mathbb{R}^{n_i} \) and \( \sum_{i=1}^{N} n_i = n \). If the quadratic term \( ||Ax||^2 \) is also separable w.r.t. the partition, then the augmented Lagrangian \( L_\rho \) is separable. Then the \( x \)-update can be carried out in parallel.

One example is \( f(x) = \lambda ||x||_1 \) \((\lambda > 0)\) and \( A = I \), in this case the \( x_i \)-update is \( x_i^+ = \arg \min_x \lambda |x_i| + \frac{\rho}{2}(x_i - v_i)^2 \). Using subdifferential calculus, the solution is \( x_i^+ = S_{\lambda/\rho}(v_i) \), where the soft thresholding operator \( S \) is defined as

\[
S_\kappa(a) = \begin{cases} 
a - \kappa & a > \kappa \\
0 & |a| \leq \kappa \\
a + \kappa & a < -\kappa 
\end{cases}
\]
Apply ADMM to Constrained Convex Optimization

The generic constrained convex optimization problem is

\[
\min_x f(x) \quad \text{s.t.} \quad x \in C
\]  

(11)

where \( x \in \mathbb{R}^n \), \( f \) and \( C \) are convex. Its ADMM form can be written as

\[
\min_{x,z} f(x) + g(z) \quad \text{s.t.} \quad x - z = 0
\]  

(12)

where \( g \) is the indicator function of \( C \). The augmented Lagrangian is

\[
L_\rho(x, z, u) = f(x) + g(z) + \rho/2 ||x - z + u||_2^2
\]

so the scaled form of ADMM is

\[
x^{k+1} = \arg \min_x f(x) + \rho/2 ||x - z^k + u^k||_2^2 = \text{prox}_{f, \rho}(z^k - u^k)
\]

\[
z^{k+1} = \Pi_C(x^{k+1} + u^k)
\]

\[
u^{k+1} = u^k + x^{k+1} - z^{k+1}
\]

Note here \( f \) need not be smooth.
Example: Convex Feasibility

Find a point in the intersection of two closed nonempty convex sets $C$ and $D$. In ADMM, the problem can be written as

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

where $f$ is the indicator function of $C$ and $g$ is the indicator function of $D$. The scaled form of ADMM is

$$x^{k+1} = \Pi_C(z^k - u^k)$$
$$z^{k+1} = \Pi_D(x^{k+1} + u^k)$$
$$u^{k+1} = u^k + x^{k+1} - z^{k+1}$$
Example: Linear and Quadratic Programming

\[
\min_{x} \frac{1}{2} x^T P x + q^T x \quad \text{s.t.} \ Ax = b, \ x \geq 0
\]

where \( x \in \mathbb{R}^n, P \in \mathbb{S}^n_+ \), when \( P = 0 \), this reduces to the linear programming. The ADMM form is

\[
\min_{x,z} f(x) + g(z) \quad \text{s.t.} \ x - z = 0
\]

where \( f(x) = \frac{1}{2} x^T P x + q^T x \), \( \text{dom} f = \{x \mid Ax = b\} \) and \( g \) is the indicator function of the nonnegative orthant \( \mathbb{R}^n_+ \). The scaled form of ADMM has iterations

\[
x^{k+1} = \arg \min_{x:Ax=b} f(x) + \frac{\rho}{2} \|x - z^k + u^k\|_2^2
\]

\[
z^{k+1} = \Pi_{\mathbb{R}^n_+}(x^{k+1} + u^k) = (x^{k+1} + u^k)_+
\]

\[
u^{k+1} = u^k + x^{k+1} - z^{k+1}
\]
Example: $\ell_1$-Norm Problems

A lot of machine learning algorithms involve minimizing a loss function with a regularization term or side constraints. Both the loss function and regularization term can be nonsmooth. We use $\ell_1$-norm problems to show how to apply ADMM to solve such problems.

$$\min_x l(x) + \lambda \|x\|_1$$

where $l$ is any convex loss function. The ADMM form is

$$\min_{x,z} l(x) + g(z) \quad \text{s.t. } x - z = 0$$

where $g(z) = \lambda \|z\|_1$, the ADMM iterations are

$$x_{k+1} = \arg \min_x l(x) + \rho/2 \|x - z^k + u^k\| = \text{prox}_{l,\rho}\|x - z^k + u^k\|_2$$

$$z^{k+1} = S_{\lambda/\rho}(x^{k+1} + u^k)$$

$$u^{k+1} = u^k + x^{k+1} - z^{k+1}$$
Given a dataset consisting of samples from a zero mean Gaussian distribution in $\mathbb{R}^n$: $a_i \approx \mathcal{N}(0, \Sigma)$, $i = 1, \ldots, N$,

- $(\Sigma^{-1})_{ij}$ is zero if and only if the $i$-th and $j$-th components of the random variable are conditionally independent given the other variables.
- Let $X = \Sigma^{-1}$ and the objective be the negative log-likelihood of data with $\ell_1$ regularization on $X$. The objective can be written as

$$\min_X \text{Tr}(SX) - \log |X| + \lambda \|X\|_1,$$ where $X \in S^n_+$, $\| \cdot \|_1$ is defined elementwise, the domain of $\log |X|$ is $S^n_{++}$, the set of symmetric positive definite $n \times n$ matrix.

The ADMM steps are

$$X^{k+1} = \arg \min_{X \in S^n_{++}} \text{Tr}(SX) - \log |X| + \rho/2 \|X - Z^K + U^K\|_F^2,$$

$$Z^{k+1} = \arg \min_{Z} \lambda \|Z\|_1 + \rho/2 \|X^{k+1} - Z + U^k\|_F^2,$$

$$U^{k+1} = U^k + X^{k+1} - Z^{k+1}.$$
The $Z$-minimization step can be solved by elementwise soft thresholding

$$Z_{ij}^{k+1} = S_{\lambda/\rho}(X_{ij}^{k+1} + U_{ij}^k)$$

The $X$-minimization step also has an analytical solution using the first-order optimality condition $S - X^{-1} + \rho(X - Z^k + U^k) = 0$. We will construct a matrix $X$ satisfying $\rho X - X^{-1} = \rho(Z^k - U^k) - S$ and $X \succ 0$. Let the RHS be factorized as $Q^T \Lambda Q (\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n), Q^T = QQ^T = I)$, then we have $\rho \hat{X} - \hat{X}^{-1} = \Lambda$, where $\hat{X} = Q^T X Q$. We can construct $X$ by finding numbers $\hat{X}_{ii}$ that satisfy $\rho \hat{X}_{ii} - 1/\hat{X}_{ii} = \lambda_i$, which leads to $\hat{X}_{ii} = (\lambda_i + \sqrt{\lambda_i^2 + 4\rho})/(2\rho)$.
What Have We Learned?

- If $f, g$ are smooth, the updates can always be solved using general iterative methods.
- If $f, g$ are so simple that an closed-form proximal operator evaluation is feasible, the updates can be carried out using analytical formulas. More examples of proximal operator evaluations are available at the paper “Proximal Algorithms” written by Parikh and Boyd.
- Techniques like early termination and warm start can be utilized to speed up convergence.
- We have been focusing on non-distributed versions of various problems, their distributed versions need to be studied for parallelization.