ADMM, Fast ADMM and some applications

Zhe Wang

Applied Mathematics

西安交通大學
XI'AN JIAOTONG UNIVERSITY

2018年9月28日

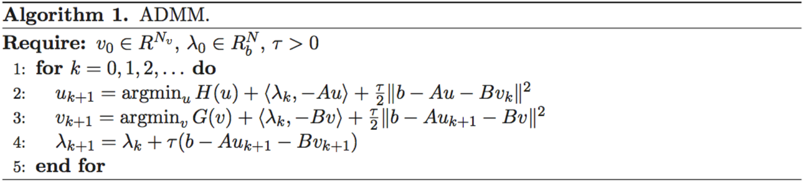
ADMM: Alternating direction method of multipliers

Optimization：

$$\min_{u,v} \; H(u) + G(v) \qquad \text{s.t.} \;\; Au + Bv = b \tag{1}$$

H, G are all convex.

Algorithm：



```
Algorithm 1. ADMM.
Require: $v_0 \in R^{N_v}$, $\lambda_0 \in R_b^N$, $\tau > 0$
for $k = 0, 1, 2, \dots$ do
    $u_{k+1} = \mathrm{argmin}_u \, H(u) + \langle \lambda_k, -Au \rangle + \frac{\tau}{2}\|b - Au - Bv_k\|^2$
    $v_{k+1} = \mathrm{argmin}_v \, G(v) + \langle \lambda_k, -Bv \rangle + \frac{\tau}{2}\|b - Au_{k+1} - Bv\|^2$
    $\lambda_{k+1} = \lambda_k + \tau(b - Au_{k+1} - Bv_{k+1})$
end for
```

Notice：ADMM is suitable for such kind of optimization problems owing to it's simplicity. But if you have a high requirement for precision or the property of H, G is not so good, ADMM may not be so ideal.

Review： When strong duality holds, point (u, v) is the optimal solution for primal and dual problems iff it satisfies KKT condition.

$$\begin{aligned} b - Au^* - Bv^* &= 0 \\ 0 \in \partial H(u^*) - A^T\lambda^* \\ 0 \in \partial G(v^*) - B^T\lambda^* \end{aligned} \tag{2}$$

Interesting fact

It is not a easy calculation to check if a point satisfies KKT condition. Thus, the most magical part of ADMM is that the last property of KKT is satisfied automatically in every iteration. As for the first and second property, the distance to the convergence point for every iteration can also be controlled.

Convergence Rate

If H and G are strongly convex , the convergence rate for ADMM is $O(1/k)$

According to the update of ν

$$\begin{aligned} 0 &\in \partial G(\nu_{k+1}) - B^T\lambda_k - \tau B^T(b - Au_{k+1} - B\nu_{k+1}) \\ &= \partial G(\nu_{k+1}) - B^T\lambda_{k+1} \end{aligned} \tag{3}$$

According to the update of u

$$\begin{aligned} 0 &\in \partial H(u_{k+1}) - A^T\lambda_k - \tau A^T(b - Au_{k+1} - B\nu_k) \\ &= \partial H(u_{k+1}) - A^T\lambda_{k+1} - \tau A^T B(\nu_{k+1} - \nu_k) \end{aligned} \tag{4}$$

It is equivalent to

$$\tau A^T B(\nu_{k+1} - \nu_k) \in \partial H(u_{k+1}) - A^T\lambda_{k+1} \tag{5}$$

If KKT holds , $0 \in \partial H(u^*) - A^T\lambda^*$. Thus, primal residual and dual residual are defined as:

$$\begin{aligned} r_k &= b - Au_k - B\nu_k \\ d_k &= \tau A^T B(\nu_k - \nu_{k-1}) \end{aligned} \tag{6}$$

Definition

F is strongly convex, if $\exists \sigma_H > 0$ for all x and y, the following inequality holds:

$$\lambda H(x) + (1-\lambda)H(y) \geq H(\lambda x + \lambda y) + \lambda(1-\lambda)\sigma_H ||x-y||^2 \tag{7}$$

where σ_H is called moduli

Assumption

Assume H and G are all strongly convex, and their moduli are σ_H and σ_G, respectively.

If the assumption holds, the global convergence rate for ADMM is $O(1/k)$

Notation:

$$\begin{aligned} u^+ &= \arg\min_u H(u) + < \lambda, -Au > + \frac{\tau}{2}||b - Au - Bv||^2 \\ \nu^+ &= \arg\min_v G(v) + < \lambda, -Bv > + \frac{\tau}{2}||b - Au^+ - Bv||^2 \\ \lambda^+ &= \lambda + \tau(b - Au^+ - B\nu^+) \end{aligned} \tag{8}$$

Clearly, the duality of the primal problem is

$$\max D(\lambda) = -H^*(A^T\lambda) + < \lambda, b > - G^*(B^T\lambda) \tag{9}$$

Remarks:

$$\begin{aligned}\Psi(\lambda) &= A\nabla H^*(A^T\lambda)\\ \Phi(\lambda) &= B\nabla G^*(B^T\lambda)\end{aligned} \tag{10}$$

Our aim :

$$b \in \Psi(\lambda^*) + \Phi(\lambda^*) \tag{11}$$

Since H, G are strongly convex, it can be concluded $L(\Psi) \leq \frac{\rho(A^TA)}{\sigma_H}$, $L(\Phi) \leq \frac{\rho(B^TB)}{\sigma_G}$

tips: If f is strongly convex with moduli σ_f, the gradients of the conjugate function f is lipschitz continuous with constant $1/\sigma_f$

Lemma

Define:

$$\begin{aligned}\lambda^{1/2} &= \lambda + \tau(b - Au^+ - B\nu)\\ \lambda^+ &= \lambda + \tau(b - Au^+ - B\nu^+)\end{aligned} \tag{12}$$

Thus,

$$Au^+ = A\nabla H^*(A^T\lambda^{1/2}) = \Phi(\lambda^{1/2}),\ \ B\nu^+ = B\nabla G^*(B^T\lambda^+) = \Phi(\lambda^+)$$

Lemma

Suppose $\tau^3 \leq \frac{\sigma_H \sigma_G^2}{\rho(A^T A)\rho^2(B^T B)}$, *and that* $B\nu = \Phi(\lambda)$, *then for any* γ

$$D(\lambda^+) - D(\gamma) \geq \tau^{-1} < \gamma - \lambda, \lambda - \lambda^+ > + \frac{1}{2\tau} ||\lambda - \lambda^+||^2 \tag{13}$$

Here comes the theorem

Theorem

If H and G are strongly convex, and that $\tau^3 \leq \frac{\sigma_H \sigma_G^2}{\rho(A^T A)\rho^2(B^T B)}$ *, then for* $k \geq 1$ *, the dual variable* $\{\lambda_k\}$ *satisfies*:

$$D(\lambda^*) - D(\lambda_k) \leq \frac{||\lambda^* - \lambda_1||^2}{2\tau(k-1)} \tag{14}$$

In lemma 2 , let $\gamma = \lambda^*$, $(\nu, \lambda) = (\nu_k, \lambda_k)$, then $\lambda^+ = \lambda_{k+1}$, what's more,

$$\begin{aligned} 2\tau(D(\lambda_{k+1}) - D(\lambda^\star)) &\geq \|\lambda_{k+1} - \lambda_k\|^2 + 2\langle \lambda_k - \lambda^\star, \lambda_{k+1} - \lambda_k \rangle \\ &= \|\lambda^\star - \lambda_{k+1}\|^2 - \|\lambda^\star - \lambda_k\|^2. \end{aligned}$$

sum over $k = 1, 2, \cdots, n-1$ yields

$$2\tau\left(-(n-1)D(\lambda^\star) + \sum_{k=1}^{n-1} D(\lambda_{k+1})\right) \geq \|\lambda^\star - \lambda_n\|^2 - \|\lambda^\star - \lambda_1\|^2.$$

Based on lemma2 again , with $\gamma = \lambda = \lambda_k$ and $\nu = \nu_k$

$$2\tau\left(D(\lambda_{k+1}) - D(\lambda_k)\right) \geq \|\lambda_k - \lambda_{k+1}\|^2.$$

Multiply both sides by $k-1$, sum over $k = 1, 2, \cdots, n-1$ to abtain

$$2\tau \sum_{k=1}^{n-1} (kD(\lambda_{k+1}) - D(\lambda_{k+1}) - (k-1)D(\lambda_k)) \geq \sum_{k=1}^{n-1} k\|\lambda_k - \lambda_{k+1}\|^2.$$

The left reduces to

$$2\tau\left((n-1)D(\lambda_n) - \sum_{k=1}^{n-1} D(\lambda_{k+1})\right) \geq \sum_{k=1}^{n-1} k\|\lambda_k - \lambda_{k+1}\|^2.$$

combine the second and the last inequality to obtain：

$$2(n-1)\tau\left(D(\lambda_n)-D(\lambda^\star)\right) \geq \|\lambda^\star-\lambda_n\|^2-\|\lambda^\star-\lambda_1\|^2+\sum_{k=1}^{n-1}k\|\lambda_k-\lambda_{k+1}\|^2 \geq -\|\lambda^\star-\lambda_1\|^2$$

based on similar process

$$\begin{aligned} ||r_k||^2 &\leq O(1/K) \\ ||d_k||^2 &\leq O(1/K) \end{aligned} \tag{15}$$

It is a beautiful result, some papers also focus on controlling $D(\lambda)-D(\lambda^*)$ by r_k, d_k, but the result are almost same. For ADMM, the convergence rate is proved to be $O(1/K)$.

Examples

$$\begin{aligned} &\min \ ||z||_1 \\ &\text{s.t. } \ Ax - z = b \end{aligned} \tag{16}$$

$$x^{k+1} = (A^T A)^{-1} A^T (b + z^k - u^k) \tag{17}$$

$$z^{k+1} = S_{1/\rho}(Ax^{k+1} - b + u^k) \tag{18}$$

$$u^{k+1} = u^k + Ax^{k+1} - z^{k+1} - b \tag{19}$$

The update of $x^{k=1}$ is actually a fitting process for $Ax = b + z^k - u^k$, the second step is a proximal operator, while the last step is gradient ascent.

l1 regularized problem

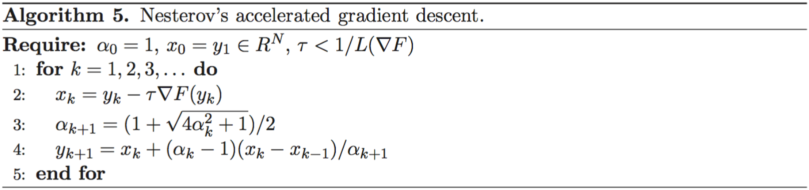
$$\min l(x) + \lambda ||x||_1 \tag{20}$$

Expressed it in the form of ADMM

$$\begin{aligned} &\min l(x) + g(z) \\ &\text{s.t. } x - z = 0 \end{aligned} \tag{21}$$

$$\begin{aligned} x^{k+1} &= \arg\min_x (l(x) + (\rho/2)||x - z^k + u^k||_2^2) \\ z^{k+1} &= S_{\lambda/\rho}(x^{k+1} + u^k) \\ u^{k+1} &= u^k + x^{k+1} - z^{k+1} \end{aligned} \tag{22}$$

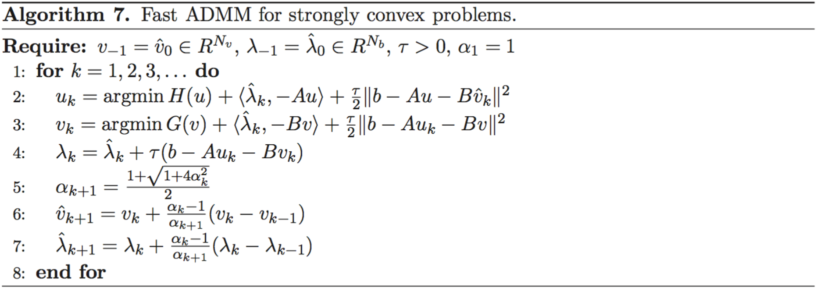
The first subproblem can be solved by proximal operator, if *f* is smooth and differentiable, it can be solve by standard Newton's method. By decomposing this problem, the iteration stages are clear and easy to perform.



Algorithm 5. Nesterov's accelerated gradient descent.

```
Require: $\alpha_0 = 1$, $x_0 = y_1 \in R^N$, $\tau < 1/L(\nabla F)$
for $k = 1, 2, 3, \ldots$ do
    $x_k = y_k - \tau \nabla F(y_k)$
    $\alpha_{k+1} = (1 + \sqrt{4\alpha_k^2 + 1})/2$
    $y_{k+1} = x_k + (\alpha_k - 1)(x_k - x_{k-1})/\alpha_{k+1}$
end for
```

Fast ADMM add a momentum term to accelerate compared with ADMM.



Algorithm 7. Fast ADMM for strongly convex problems.

```
Require: $v_{-1} = \hat{v}_0 \in R^{N_v}$, $\lambda_{-1} = \hat{\lambda}_0 \in R^{N_b}$, $\tau > 0$, $\alpha_1 = 1$
for $k = 1, 2, 3, \ldots$ do
    $u_k = \text{argmin}\, H(u) + \langle \hat{\lambda}_k, -Au \rangle + \frac{\tau}{2}\|b - Au - B\hat{v}_k\|^2$
    $v_k = \text{argmin}\, G(v) + \langle \hat{\lambda}_k, -Bv \rangle + \frac{\tau}{2}\|b - Au_k - Bv\|^2$
    $\lambda_k = \hat{\lambda}_k + \tau(b - Au_k - Bv_k)$
    $\alpha_{k+1} = \frac{1+\sqrt{1+4\alpha_k^2}}{2}$
    $\hat{v}_{k+1} = v_k + \frac{\alpha_k - 1}{\alpha_{k+1}}(v_k - v_{k-1})$
    $\hat{\lambda}_{k+1} = \lambda_k + \frac{\alpha_k - 1}{\alpha_{k+1}}(\lambda_k - \lambda_{k-1})$
end for
```

First come the Conclusion:

If H and G are all strongly convex，Fast ADMM have a convergence rate of $O(1/k^2)$. For general problems, Fast ADMM with restart is proposed, it can converge globally, but the rate is not known.

Skip it if you are not interested in mathematics

if the H and G are strongly convex, and their moduli are σ_H, σ_G, respectively.

Notation:

$$s_k = a_k\lambda_k - (a_k - 1)\lambda_{k-1} - \lambda^* \tag{24}$$

Lemma

$$s_{k+1} = s_k + a_{k+1}(\lambda_{k+1} - \hat{\lambda}_{k+1}) \tag{25}$$

Lemma

Suppose H, G are strongly convex, and that G is quadratic, if $\tau^3 \leq \frac{\sigma_H \sigma_G^2}{\rho(A^T A)\rho^2(B^T B)}$, then the sequence $\{s_k\}$ satisfies

$$||s_{k+1}||^2 - ||s_k||^2 \leq 2a_k^2\tau(D(\lambda^*) - D(\lambda_k)) - 2a_{k+1}^2\tau(D(\lambda^*) - D(\lambda_{k+1})) \tag{26}$$

Theorem

If strong duality holds, The sequence $\{\lambda_k\}$ *generated by ADMM satisfies*

$$D(\lambda^*) - D(\lambda_k) \leq \frac{2||\hat{\lambda}_1 - \lambda^*||^2}{\tau(k+2)^2} \tag{27}$$

From the lemma2

$$\begin{aligned} 2a_{k+1}^2\tau\left(D(\lambda^\star) - D(\lambda_{k+1})\right) &\leq \|s_k\|^2 - \|s_{k+1}\|^2 \\ &\quad + 2a_k^2\tau\left(D(\lambda^\star) - D(\lambda_k)\right) \\ &\leq \|s_k\|^2 + 2a_k^2\tau\left(D(\lambda^\star) - D(\lambda_k)\right). \end{aligned}$$

Again according to lemma2

$$\|s_{k+1}\|^2 + 2a_{k+1}^2\tau\left(D(\lambda^\star) - D(\lambda_{k+1})\right) \leq \|s_k\|^2 + 2a_k^2\tau\left(D(\lambda^\star) - D(\lambda_k)\right)$$

which implies by induction that

$$\|s_k\|^2 + 2a_k^2\tau\left(D(\lambda^\star) - D(\lambda_k)\right) \leq \|s_1\|^2 + 2a_1^2\tau\left(D(\lambda^\star) - D(\lambda_1)\right).$$

Because of lemma 2：

$$D(\lambda_1) - D(\lambda^\star) \geq \frac{1}{2\tau}\|\lambda_1 - \hat{\lambda}_1\|^2 + \frac{1}{\tau}\langle \hat{\lambda}_1 - \lambda^\star, \lambda_1 - \hat{\lambda}_1\rangle = \frac{1}{2\tau}\left(\|\lambda_1 - \lambda^\star\|^2 - \|\hat{\lambda}_1 - \lambda^\star\|^2\right).$$

Combine the two inequlity

$$\begin{aligned} 2a_{k+1}^2\tau\left(D(\lambda^\star) - D(\lambda_{k+1})\right) &\leq \|s_1\|^2 + 2\tau\left(D(\lambda^\star) - D(\lambda_1)\right) \\ &= \|\lambda_1 - \lambda^\star\|^2 + 2\tau\left(D(\lambda^\star) - D(\lambda_1)\right) \\ &\leq \|\lambda_1 - \lambda^\star\|^2 + \|\hat{\lambda}_1 - \lambda^\star\|^2 - \|\lambda_1 - \lambda^\star\|^2 \\ &= \|\hat{\lambda}_1 - \lambda^\star\|^2, \end{aligned}$$

Then

$$D(\lambda^\star) - D(\lambda_k) \leq \frac{\|\hat{\lambda}_1 - \lambda^\star\|^2}{2\tau a_k^2}.$$

Still define the primal residual as

$$r_k = b - Au_k - B\nu_k \tag{28}$$

Change the definition of dual residual as

$$d_k = \tau A^T B(\nu_k - \hat{\nu}_k) \tag{29}$$

It can also be proved that

$$||r_k||^2 \leq O(1/k^2) \tag{30}$$

$$||d_k||^2 \leq O(1/k^2) \tag{31}$$

What' s more, there is another conclusion

$$H(u_k) + G(\nu_k) - p* \leq -\lambda_k^T r_k + (u^k - u^*)^T d_k \tag{32}$$

which can be used as a stopping criteria for the algorithm.

Examples

For MRI reconstruction or compressed sensing

$$\min_u |\nabla u| + \frac{\epsilon}{2}||\nabla u||^2 + \frac{\mu}{2}||RFu - f||^2 \tag{33}$$

in which, R is a diagonal matrix, f is the observation in Fourier domain , F is a Fourier transformation , tv term is added to enforce sparsity , the second order term for gradient is added to enforce the smoothness of solution, the last term is for fitting.

Take

$$\begin{aligned} H(u) &= \frac{\mu}{2}||RFu - f||^2 \\ G(\nu) &= |\nu| + \frac{\epsilon}{2}||\nu||^2 \\ A &= \nabla,\ B = I \end{aligned} \tag{34}$$

Considering the second step in ADMM:

$$\nu_k = \arg\min_\nu(|\nu| + \frac{\epsilon}{2}||\nu||^2 + < \lambda_k, \nu > + \frac{\tau}{2}||\nu - \nabla u_k||^2) \tag{35}$$

By Soft-thresholding Operator

$$\nu_k = shrink(\frac{\tau}{\tau+\epsilon}(\nabla u_k + \tau\hat{\lambda}_k), \frac{1}{\tau+\epsilon}) \tag{36}$$

Then, the first step in ADMM

$$(\tau\triangle + \mu F^T R' R F)u_{k+1} = \mu F^T R^T f + \nabla\cdot(\lambda_k + \tau\nu_k) \tag{37}$$

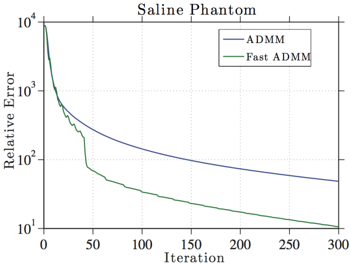
Discrete Laplace operator $\triangle$ is a convolution operator, can be expressed as in a diagonal matrix in Fourier domain

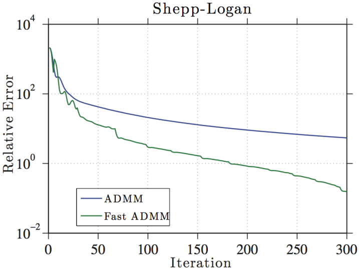
$$\triangle = F^T L F$$

Therefore

$$F^T(\tau + \mu R' R)F u_{k+1} = \mu F^T R^T f + \nabla\cdot(\lambda_k + \tau\nu_k) \tag{38}$$

$$u_{k+1} = F^T(\tau + \mu R' R)^{-1} F(\mu F^T R^T f + \nabla\cdot(\lambda_k + \tau\nu_k)) \tag{39}$$


Saline Phantom
Relative Error
10^4
10^3
10^2
10^1
0
50
100
150
200
250
300
Iteration
ADMM
Fast ADMM


Shepp-Logan
Relative Error
10^4
10^2
10^0
10^{-2}
0
50
100
150
200
250
300
Iteration
ADMM
Fast ADMM

For the deblur problem, K is blur kernel

$$\min_u |\nabla u| + \frac{\epsilon}{2}||\nabla u||^2 + \frac{\mu}{2}||K * u - f||^2 \tag{40}$$

linear convolution operator can be taken as diagonal transform in Fourier domain

$$K * u = F^T RFu \tag{41}$$

then the remaining are the same with the former slides