EP3260: Machine Learning Over Networks

Lecture 6: Alternating Direction Method of Multipliers

Euhanna Ghadimi

Ericsson, Stockholm, Sweden

Division of Network and Systems Engineering
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology, Stockholm, Sweden

https://sites.google.com/view/mlons/home

February 2019

What is this lecture about?

ADMM: a low complexity and robust optimization method with large variety of applications, including

- decentralized problems in which devices, parallel processors, or agents coordinate to solve large-scale problems while imposing little communication overheads (messages)
- machine learning and statistical problems with huge data sizes
- effective heuristic method in many non-convex problems

Outline

Outline

1. Preliminaries
   Dual decomposition
   Method of multipliers

2. ADMM

3. Application examples

4. References

Dual problem

Convex problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array} \qquad (1)$$

the Lagrangian is defined as

$$L(x, y) = f(x) + y^\top (Ax - b),$$

where y is the dual variable (Lagrange multiplier);
the dual problem is then defined as

$$\text{maximize} \quad \{g(y) = \inf_x L(x, y)\}$$

One recovers a (not necessarily unique) primal optimal $x^\star$ from a dual optimal $y^\star$

$$x^\star = \operatorname{argmin}_x \quad L(x, y^\star)$$

Dual ascent

In the dual ascent method, one solves the dual problem using gradient ascent.

1. perform an x-minimization step

$$x^{k+1} = \operatorname*{argmin}_x L(x, y^k)$$

2. calculate $\nabla g(y) = Ax^{k+1} - b$ (assuming g is differentiable), then update the dual variable using gradient method

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

- Under proper choice of step-size parameters α^k and some other conditions x^k converges to an optimal primal and y^k converges to an optimal dual point.
- The method called dual ascent because, under proper choice of α^k, the dual function converges monotonically $g(y^{k+1}) > g(y^k)$.
- If g is non-differentiable then $Ax^{k+1} - b$ is a sub-gradient of $-g$ and convergence is non-monotonic.

Dual decomposition

Suppose $f(x) = f_1(x_1) + \cdots + f_N(x_N)$, $x = (x_1, \ldots, x_N)$, with f_i and x_i corresponding to a partition of the original problem (variable)

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{i=1}^{N} f_i(x_i) \\ \text{subject to} & \sum_{i=1}^{N} A_i x_i = b \end{array}$$

then the Lagrangain is also separable (in x)

$$L(x, y) = \sum_{i=1}^{N} L_i(x_i, y) = \sum_{i=1}^{N} f_i(x_i) + \sum_{i=1}^{N} y^\top A_i x_i - y^\top b$$

the x-minimization step in dual ascent method splits into N parallel minimization

$$x_i^{k+1} = \underset{x_i}{\text{argmin}}\ L_i(x_i, y^k)$$

Dual decomposition

The dual decomposition method (dates back to 60's)

$$x_i^{k+1} = \underset{x_i}{\text{argmin}}\, L_i(x_i, y^k), \quad i = 1, \ldots, N,$$

$$y^{k+1} = y^k + \alpha^k (\sum_{i=1}^{N} A_i x_i^{k+1} - b)$$

A potentially large-scale problem is solved iteratively in a distributed fashion

- perform parallel independent x_i-minimization step
- gather residuals $A_i x_i^{k+1}$ to calculate global dual variable y^{k+1}; then broadcast it to distributed workers

Under several assumptions, it converges; however, often slowly!

Method of multipliers

Uses so called augmented Lagrangian (with penalty parameter $\rho > 0$)

$$L_\rho(x, y) = f(x) + y^\top (Ax - b) + \frac{\rho}{2}\|Ax - b\|_2^2$$

It can be viewed as Lagrangian of the problem

$$\begin{array}{ll} \text{minimize} & f(x) + \frac{\rho}{2}\|Ax - b\|_2^2 \\ \text{subject to} & Ax = b \end{array}$$

which clearly is equivalent to original problem (for any feasible point the additional term in objective diminishes). Forming the associated dual function $g_\rho(y) = \inf_x L_\rho(x, y)$ and applying dual ascent

$$\begin{aligned} x^{k+1} &= \underset{x}{\operatorname{argmin}}\, L_\rho(x, y^k), \\ y^{k+1} &= y^k + \rho(Ax^{k+1} - b) \end{aligned}$$

Note that x-minimization step uses the augmented Lagrangian, and the penalty parameter ρ is used as the step size α^k.

Method of multipliers

It converges under more general conditions than dual ascent, including cases when f is unbounded from above or it is not strictly convex.

under primal-dual feasibility (i.e., the optimality conditions of problem (1))

$$Ax^\star - b = 0, \quad \nabla f(x^\star) + A^\top y^\star = 0,$$

since x^{k+1} minimizes $L_\rho(x, y^k)$, one has

$$\begin{aligned} 0 &= \nabla_x L_\rho(x^{k+1}, y^k) \\ &= \nabla f(x^{k+1}) + A^\top (y^k + \rho(Ax^{k+1} - b)) \\ &= \nabla f(x^{k+1}) + A^\top y^{k+1} \end{aligned}$$

- So particular step-size choice $\alpha^k = \rho$ at dual-update ($y^{k+1} = y^k + \rho(Ax^{k+1} - b)$) makes (x^{k+1}, y^k) dual feasible!
- It remains to achieve primal optimality at limit

$$Ax^{k+1} - b \to 0.$$

Method of multipliers

- Method of multiplier enjoys greatly improved convergence properties compared to dual-ascent
- However, it comes at a cost. When f is separable, the augmented Lagrangian L_ρ is not separable, so the x-minimization step cannot be carried out separately in parallel for x_i's.

Next, we will study a method to resolve this issue!

Outline

Alternating direction method of multipliers

- Introduced in 70's!
- Enjoys improved convergence properties of method of multiplier
- Resolves the issue of non-decomposability of the augmented Lagrangian L_ρ
- It performs Gauss-seidel decomposition in the primal update! So the x-minimization step can be carried out separately in parallel for x_i's.

Alternating direction method of multipliers

ADMM canonical form (with convex f and g): with two sets of variables and separable costs

$$\begin{aligned} \text{minimize } & f(x) + g(z) \\ \text{subject to } & Ax + Bz = c \end{aligned}$$

Augmented Lagrangian:

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

And the ADMM iterates:

$$\begin{aligned} x^{k+1} &= \underset{x}{\text{argmin}}\, L_\rho(x, z^k, y^k), \\ z^{k+1} &= \underset{z}{\text{argmin}}\, L_\rho(x^{k+1}, z, y^k), \\ y^{k+1} &= y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \end{aligned}$$

Remarks

- Minimizing jointly over x and z falls back to method of multipliers!
- Once minimizing over x or z, the other variable is kept fixed (one pass of Gauss-Seidel)
- Recall from decomposition method: if problem variables are separable

$$\text{minimize } f(x) = \sum_{i\in[N]} f_i(x_i)$$
$$\text{subject to } \sum_{i\in[N]} A_i x_i = b$$

then, $L_\rho(x,y) = \sum_{i=1}^N f_i(x_i) + \sum_{i=1}^N y^\top A_i x_i - y^\top b + \frac{\rho}{2}\|\sum_{i=1}^N A_i x_i - b\|_2^2$
Applying ADMM, now we get separable x_i-updates

$$x_i^{k+1} = \underset{x_i}{\text{argmin}}\, L_\rho(x_i, x_j^k, y^k) = \underset{x_i}{\text{argmin}}\, f_i(x_i) + y^{k\top} A_i x_i + \frac{\rho}{2}\|A_i x_i - b\|_2^2 + \rho x_i^\top A_i^\top \sum_{i\neq j} A_j x_j^k$$

Optimality conditions

Primal-dual feasibility (i.e., the optimality conditions of problem (13))

$$Ax^\star + Bz^\star = c, \quad \nabla f(x^\star) + A^\top y^\star = 0, \quad \nabla g(z^\star) + B^\top y^\star = 0$$

since z^{k+1} minimizes $L_\rho(x^{k+1}, z, y^k)$, one has

$$\begin{aligned} 0 &= \nabla_z L_\rho(x^{k+1}, z, y^k) \\ &= \nabla g(z^{k+1}) + B^\top (y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)) \\ &= \nabla g(z^{k+1}) + B^\top y^{k+1} \end{aligned}$$

- So particular step-size choice $\alpha^k = \rho$ at dual-update ($y^{k+1} = y^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$) satisfies second dual feasibility condition!
- It remains to achieve primal and first dual feasibility at limit

$$Ax^k + Bz^k - c \to 0, \quad \nabla f(x^k) + A^\top y^k \to 0$$

ADMM with scaled dual variables

Combine linear and quadratic terms in augmented Lagrangian

$$
\begin{aligned}
L_\rho(x, z, y) &= f(x) + g(z) + y^\top(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2 \\
&= f(x) + g(z) + \frac{\rho}{2}\|Ax + Bz - c + u\|_2^2 + \text{constant terms.}
\end{aligned}
$$

with $u = (1/\rho)y^k$.

And the ADMM iterates (in scaled form):

$$
\begin{aligned}
x^{k+1} &= \operatorname*{argmin}_x \; f(x) + \frac{\rho}{2}\|Ax + Bz^k - c + u^k\|_2^2, \\
z^{k+1} &= \operatorname*{argmin}_z \; g(z) + \frac{\rho}{2}\|Ax^{k+1} + Bz - c + u^k\|_2^2, \\
u^{k+1} &= u^k + Ax^{k+1} + Bz^{k+1} - c
\end{aligned}
$$

Proximal operators

- Consider x-update (when $A = I$)

$$x^{k+1} = \underset{x}{\text{argmin}}\ f(x) + \frac{\rho}{2}\|x + \underbrace{Bz^k - c + u^k}_{\text{const.}\triangleq -v}\|_2^2,$$

$$= \underset{x}{\text{argmin}}\ f(x) + \frac{\rho}{2}\|x - v\|_2^2 = \text{prox}_{f,\rho}(v),$$

- Example 1: indicator function of set C: $f = I_C$. then $x^{k+1} = \Pi_C(v)$, or projection onto C.

$$f(x) = I_{\geq b}(x) = \begin{cases} 0 & \text{if } x \geq b, \\ \infty & \text{otherwise.} \end{cases} \quad \text{then } x^{k+1} = \max(v, b).$$

- Example 2: ℓ_1-norm (Lasso problem): minimize $(1/2)\|Ax - b\|_2^2 + \lambda\|x\|_1$. Formulating the ADMM form, resulting z-update is called soft thresholding: $z_i^{k+1} = S_{\lambda/\rho}(x_i^{k+1} + u_i^k)$,

$$s_a(v) = \begin{cases} v - a & v > a \\ 0 & |v| \leq a \\ v + a & v < -a \end{cases}$$

Convergence

Somewhat most general case holds under two assumptions:

- The extended real valued functions $f \in \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g \in \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are closed, proper and convex.
  - it can be stated as: function f satisfies the assumption iff its epigraph

    $$\mathrm{epi}\, f = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} |\, f(x) \leq t\}$$

    is a closed nonempty convex set.
  - the assumption says there are x and z (not necessarily unique) that minimize the augmented Lagrangian (sub-problems in x and z-update are solvable).
- The Lagrangian has a saddle point: there exist $(x^\star, z^\star, y^\star)$ such that the following holds;

  $$L_0(x^\star, z^\star, y) \leq L_0(x^\star, z^\star, y^\star) \leq L_0(x, z, y^\star) \quad \forall x, z, y.$$

- Assumption 1 and 2 imply that strongly duality holds; no explicit assumptions on A and B.

Convergence

Under Assumptions 1 and 2, the ADMM iterates satisfy:

- Residual convergence (primal): $r^k = Ax^k + Bz^k - c \to 0$ as $k \to \infty$.
- Objective convergence $f(x^k) + g(z^k) \to f(x^\star) + g(z^\star)$ as $k \to \infty$.
- Dual variable convergence: $y^k \to y^\star$ as $k \to \infty$.
- Note that (under these assumptions only) x^k and z^k does not necessarily converge to optimal values.

In practice, it is possible to construct examples where ADMM converges very slow. However, often it is easy to tune ADMM to converge to modest accuracy after few tens of iterations.

Optimal step-size

For standard QP, with positive definite Hessian

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^\top Qx + q^\top x \\ \text{subject to} & Ax \le c, \end{array} \rightarrow \begin{array}{ll} \text{minimize} & \frac{1}{2}x^\top Qx + q^\top x + I_+(z), \\ \text{subject to} & Ax + z = c, \end{array}$$

The ADMM iterates converge linearly: in terms of some residuals, e.g., $\|r^k\|_2 \le \delta\gamma^k\|r^0\|_2$ for some $\delta \in \mathbb{R}_+, \gamma \in (0,1)$

Moreover, the optimum choice of step-size ρ and corresponding convergence factor can be found as

$$\rho^\star = \frac{1}{\sqrt{\lambda_{\min}\lambda_{\max}}}, \quad \gamma^\star = \frac{\kappa - 1}{\kappa + 1}, \quad \kappa = \frac{\lambda_{\max}}{\lambda_{\min}}, \quad \lambda_i \triangleq \lambda(AQ^{-1}A^\top).$$

Although the linear rate is established for general QPs, the optimal parameter holds when A is full-row rank (or invertible).

Optimal step-size

The result can be generalized for full-row rank A and L-smooth and μ-strongly convex f!

- a convex $f()$ is L-smooth and μ-strongly convex if these two are also convex

$$f(x) - \frac{\mu}{2}\|x\|_2^2, \quad \frac{L}{2}\|x\|_2^2 - f(x)$$

- the ADMM iterates converge linearly with optimal step-size and convergence factor

$$\rho^\star = \sqrt{\frac{L\mu}{\lambda_{\min}(AA^\top)\lambda_{\max}(AA^\top)}}, \quad \gamma^\star = \frac{\bar{\kappa}-1}{\bar{\kappa}+1}, \quad \bar{\kappa} = \frac{L\lambda_{\max(}AA^\top)}{\mu\lambda_{\min}(AA^\top)}$$

Other ways of parameter tuning

- Relaxation: In z and y-updates the term Ax^{k+1} can be replaced with

$$\alpha^k Ax^{k+1} - (1-\alpha^k)(Bz^k - c), \quad \alpha^k \in (0,2)$$

  for $a^k > 1$ it is called over-relaxation, and for $\alpha^k < 1$ is under-relaxation

  theoretical best value: $\alpha^\star = 2$ for L-smooth and μ-strongly convex f.

  empirical results suggest $\alpha^k \in [1.5, 1.8]$ for improved convergence!

- Varying step-size ρ^k: empirically, convergence improvement can be achieved by changing ρ^k at each iteration to balance the primal and the dual residuals (not for cases with analytical $\rho^\star$).

$$\rho^{k+1} = \begin{cases} \tau^{\text{incr}} \rho^k & \text{if } \frac{\|r^k\|_2}{\|s^k\|_2} > \delta \\ \tau^{\text{decr}} \rho^k & \text{if } \frac{\|s^k\|_2}{\|r^k\|_2} > \delta, \\ \rho^k & \text{otherwise.} \end{cases}, \qquad \begin{aligned} r^k &= Ax^k + Bz^k - c \\ s^k &= \rho^k A^\top B(z^k - z^{k-1}) \end{aligned}$$

  where $\delta > 0, \tau^{\text{incr}} > 1, \tau^{\text{decr}} < 1$. Typical example $\delta = 10$, $\tau^{\text{incr}} = 2, \tau^{\text{decr}} = 0.5$.

Outline

Consensus optimization

Problem with N objective terms: minimize $\sum_{i=1}^{N} f_i(x)$;

f_i might be the loss function for i-th block of training data $\{(a_i, b_i)\}_{i=1}^{N}$ in a statistical optimization problem: e.g.,

- LASSO

$$\text{minimize}_{x\in\mathbb{R}^m} \quad \frac{1}{N}\sum_{i=1}^{N}(b_i - x^\top a_i)^2 + \lambda\|x\|_1$$

- Classification: $a_i \in \mathbb{R}^{m_i}, b \in \{-1, +1\}$, $l(\cdot)$ a loss function (hinge, logistic, ...), $r(\cdot)$ a regularization function ($\ell_1, \ell_2, \dots$)

$$\text{minimize}_{x\in\mathbb{R}^m, w\in\mathbb{R}} \quad \frac{1}{N}\sum_{i=1}^{N} l(b_i(a_i^\top x + w)) + r(x)$$

- Support vector machine (SVM)

$$\text{minimize}_{x\in\mathbb{R}^m, w\in\mathbb{R}} \quad \frac{1}{N}\sum_{i=1}^{N} \max\{0, 1 - b_i(x^\top a_i + w)\} + \lambda\|x\|_2^2$$

Consensus optimization

Problem with N objective terms: minimize $\sum_{i=1}^{N} f_i(x)$;

ADMM iterates:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{N} f_i(x_i) \\ \text{subject to} & x_i - z = 0, \end{array}$$

- x_i are local variables
- z is global variable (kept in a central node)
- $x_i - z = 0$ are consistency or *consensus* variables
- can add regularization by adding $g(z)$ term into objective

Consensus optimization- ADMM formulation

$$L_\rho(x, y, z) = \sum_{i=1}^N \left(f_i(x_i) + y_i^\top (x_i - z) + \rho/2 \|x_i - z\|_2^2 \right)$$

ADMM form:

$$\begin{aligned} x_i^{k+1} &= \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + y_i^{k\top} (x_i - z^k) + \rho/2 \|x_i - z^k\|_2^2 \right) \\ z^{k+1} &= \frac{1}{N} \sum_{i=1}^N \left(x_i^{k+1} + 1/\rho y_i^k \right) \\ y_i^{k+1} &= y_i^k + \rho(x_i^{k+1} - z^{k+1}) \end{aligned}$$

Under regularization, the averaging term in z-update is replaced with $\text{prox}_{g,\rho}$.

Consensus optimization- ADMM formulation

One can check $\sum_{i=1}^{N} y_i^k = 0$, which further simplifies the ADMM algorithm to

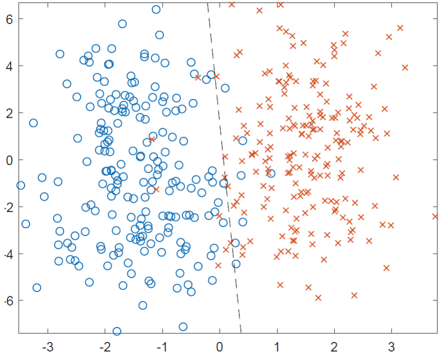
$$
\begin{aligned}
x_i^{k+1} &= \operatorname*{argmin}_{x_i} \left(f_i(x_i) + y_i^{k\top}(x_i - \bar{x}^k) + \rho/2 \|x_i - \bar{x}^k\|_2^2 \right) \\
y_i^{k+1} &= y_i^k + \rho(x_i^{k+1} - \bar{x}^{k+1})
\end{aligned}
$$

where $\bar{x}^k = (1/N) \sum_{i=1}^{N} x_i^k$,

At each iteration

- local processors compute y_i^k and x_i^k in parallel using global variable $\bar{x}^{k-1}$
- central node gathers x_i^k from local processors and computes the average $\bar{x}^k$
- central node scatters the average $\bar{x}^k$ to processors

Performance of ADMM- Consensus SVM



- minimize$_{x\in\mathbb{R}^m, w\in\mathbb{R}}$ $\quad \frac{1}{N}\sum_{i=1}^{N} \max\{0, 1 - b_i(x^\top a_i + w)\} + \lambda\|x\|_2^2$
- A simple example with $a_i \in \mathbb{R}^{2\times 400}$ samples partitioned in the worst way (2 subsystems that each holds only positive $b_i = +1$ or negative $b_i = -1$ examples)
- After 60 iterations of consensus ADMM, primal and dual residuals decay to 10^{-2}

Consensus optimization- Networked formulation

Problem with N objective terms: minimize $\sum_{i=1}^{N} f_i(x)$;

$$
\begin{array}{ll}
\text{minimize} & \displaystyle\sum_{i=1}^{N} f_i(x_i) \\
\text{subject to} & x_i = z_{ij}, \quad \text{for } i = 1, \ldots, N,\ j \in \mathcal{N}_i, \\
& z_{ij} = z_{ji}, \quad \text{for } (i,j) \in \mathcal{E}.
\end{array}
$$

- x_i are local variables
- z_{ij} are auxiliary edge variables (kept locally)
- there are other ways to formulate consensus variables (node formulation, etc.)

Networked consensus optimization- ADMM

$L_\rho(x, y, z) =$
$\sum_{i=1}^N \left(f_i(x_i) + \sum_{i=1}^N \sum_{j \in \mathcal{N}_i} y_{ij}^\top (x_i - z_{ij}) + \frac{\rho}{2} \sum_{i=1}^N \sum_{j \in \mathcal{N}_i} \|x_i - z_{ij}\|_2^2 \right)$

ADMM form:

$$
\begin{aligned}
x_i^{k+1} &= \underset{x_i}{\operatorname{argmin}} \left(f_i(x_i) + \sum_{j \in \mathcal{N}_i} y_{ij}^{k\top} (x_i - z_{ij}^k) + \frac{\rho}{2} \sum_{j \in \mathcal{N}_i} \|x_i - z_{ij}^k\|_2^2 \right) \\
z_{ij}^{k+1} &= \frac{1}{2} \left(x_i^{k+1} + x_j^{k+1} + 1/\rho (y_{ij}^k + y_{ji}^k) \right) \\
y_{ij}^{k+1} &= y_{ij}^k + \rho (x_i^{k+1} - z_{ij}^{k+1})
\end{aligned}
$$

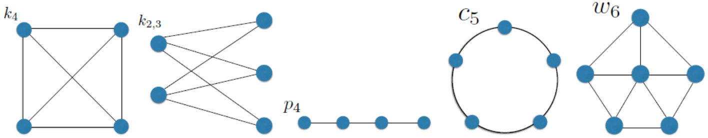
Under regularization, the averaging term in z_{ij}-update is replaced with $\mathrm{prox}_{g,\rho}$ of an average point!

Networked consensus optimization- Topology-dependent performance

After some algebra tricks, simplified decentralized ADMM form (with $g(z)=0$)

$$
\begin{aligned}
\partial f(x_i^{k+1}) + y_i^k + 2\rho|\mathcal{N}_i|x_i^{k+1} - \rho\left(|\mathcal{N}_i|x_i^k + \sum_{j\in\mathcal{N}_i} x_j^k\right) &= 0,\\
y_i^{k+1} = \ y_i^k + \rho\left(|\mathcal{N}_i|x_i^{k+1} - \sum_{j\in\mathcal{N}_i} x_j^{k+1}\right)&.
\end{aligned}
$$

- $y_i \in \mathbb{R}^{m_i}, i = 1, \ldots, N$ is the local Lagrange multipliers at agent i
- performance of decentralized ADMM also depends to underlying network topology (see [4-6] for convergence results and performance optimization)



Outline

Some references

- S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," FoT in Machine learning, 2011.

- E. Ghadimi, A. Teixeira, I. Shames, and M. johansson, "Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems," IEEE Transactions on Automatic Control, 2015.

- P. Giselsson, and S. Boyd, "Linear convergence and metric selection for Douglas-Rachford splitting and ADMM," IEEE Transactions on Automatic Control, 2017.

- W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," IEEE Transactions on Signal Processing, 2014.

- A. Teixeira, E. Ghadimi, I. Shames, H. Sandberg, and M. Johansson, "The ADMM algorithm for distributed quadratic problems: Parameter selection and constraint preconditioning," IEEE Transactions on Signal Processing. 2016.

- A. Makhdoumi, and A. Ozdaglar, "Convergence rate of distributed ADMM over networks," IEEE Transactions on Automatic Control, 2017.