CSCI-6961 Project G3

Asynchronous Distributed ADMM for Consensus Optimization

Md Shamim Hussain, Joshua Kavner, Osama Minhas, Allie Solko-Breslin, Lingyu Zhang

Dec 2, 2020

Ruiliang Zhang and James Kwok, *ICML* 2014

Talk Outline

1. Motivation
2. Synchronous (Centralized) ADMM
3. Asynchronous (Centralized) ADMM
4. Network Topology Analysis
5. Empirical Evaluation
6. Discussion

Motivation

ADMM vs Mini-batch SGD

- Jointly optimizes data on multiple servers without exchanging data
- Consensus optimization with ADMM instead of Mini-batch SGD
- Parallelization of optimization using independent different workers

Distributed Asynchronous ADMM

- Avoids straggling by allowing workers to update at different speeds
- Bounds delay on slowest worker to encourage regular updating

Related Works

ADMM -- Gabay et al 1976

126 page survey by
Boyd et. al. 2011

O(1/t) convergence ADMM
[Bingsheng He, Xiaoming Yuan, SIAM 2011]

Decentralized ADMM
[Ermin Wei, Asuman Ozgander, IEEE 2012]

Asynchronous Decentralized ADMM
[Ermin Wei, Asuman Ozgander, IEEE 2013]

Async. Centralized ADMM
[Ruiliang Zhang, James Kwok, ICML 2014]

Main result of the paper

- Asynchronous distributed processing can be extended to ADMM
- Faster than synchronous ADMM in practice
- Models that can be formulated using ADMM can be learned efficiently
- O(1/t) convergence

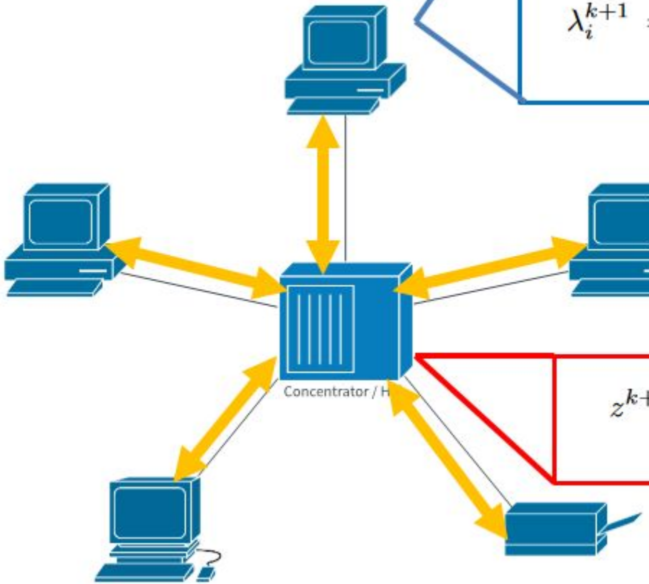
ADMM: Alternating Direction Method of Multipliers

Key	Name	Description	Stock	Price	LastOrdered
ARC1	Arc welder	250 Amps	8	119.00	25-Nov-2013
BRK8	Bracket	250mm	46	5.66	18-Nov-2013
BRK9	Bracket	400mm	82	6.98	1-Jul-2013
HOS8	Hose	1/2"	27	27.50	18-Aug-2013
WGT4	Widget	Green	16	13.99	3-Feb-2013
WGT6	Widget	Purple	76	13.99	31-Mar-2013

Key	Name	Description	Stock	Price	LastOrdered
ARC1	Arc welder	250 Amps	8	119.00	25-Nov-2013
BRK8	Bracket	250mm	46	5.66	18-Nov-2013
BRK9	Bracket	400mm	82	6.98	1-Jul-2013

Key	Name	Description	Stock	Price	LastOrdered
HOS8	Hose	1/2"	27	27.50	18-Aug-2013
WGT4	Widget	Green	16	13.99	3-Feb-2013
WGT6	Widget	Purple	76	13.99	31-Mar-2013

Synchronous (Centralized) ADMM



$$x_i^{k+1} = \arg\min_x f_i(x) + \langle \lambda_i^k, x\rangle + \frac{\beta}{2}\|x - z^k\|^2$$

$$\lambda_i^{k+1} = \lambda_i^k + \beta(x_i^{k+1} - z^{k+1})$$

$$z^{k+1} = \arg\min_z \sum_{i=1}^{N} -\langle \lambda_i^k, z\rangle + \frac{\beta}{2}\|x_i^{k+1} - z\|^2$$

ADMM: Alternating Direction Method of Multipliers

Suppose we have the following problem:

minimize $f(x) + g(z)$ such that $Ax + Bz = c$.

Augmented Lagrangian

The augmented Lagrangian is of the form:

$$L_\beta(x, z, \lambda) = f(x) + g(z) + \lambda^T(Ax + By - c) + (\beta/2)||(Ax + By - c)||_2^2.$$

Iterations

ADMM consists of the iterations:

$$x^{k+1} = \mathrm{argmin}_x\, L_\beta(x, z^k, \lambda^k)$$
$$z^{k+1} = \mathrm{argmin}_z\, L_\beta(x^k, z, \lambda^k)$$
$$\lambda^{k+1} = \lambda^k + \beta(Ax^{k+1} + Bz^{k+1} - c)$$

ADMM: Consensus Optimization

Let $(x_i)_{i=1}^{N} \in \mathbb{R}^n$. Then $x = [x_1 || ... || x_N]^T \in \mathbb{R}^{nN}$

$$\text{Substitute } A \leftarrow \text{diag}_N(\mathbb{I}_{n \times n}) = \begin{pmatrix} \mathbb{I} & 0 & \dots & 0 \\ 0 & \mathbb{I} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathbb{I} \end{pmatrix}$$

$$B \leftarrow -1 * \text{repeat}_N(\mathbb{I}_{n \times n}) = -\begin{pmatrix} \mathbb{I} & \mathbb{I} & \dots & \mathbb{I} \end{pmatrix}^T$$

$c \leftarrow 0$

$f(x) := \sum_{i=1}^{N} f_i(x_i)$

$g(z) := 0$

$$\boxed{\begin{array}{l} \text{minimize } f(x) + g(z) \\ \text{such that } Ax + Bz = c \end{array}}$$

ADMM: Consensus Optimization

$$\min_{x_1,\ldots,x_N,z} \sum_{i=1}^{N} f_i(x_i) \ : \ x_i = z, \ i = 1, 2, \ldots, N$$

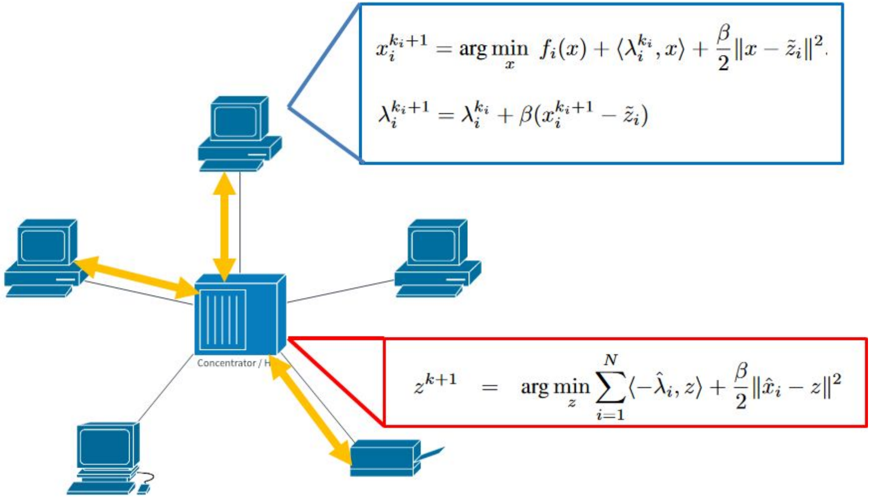
Augmented Lagrangian

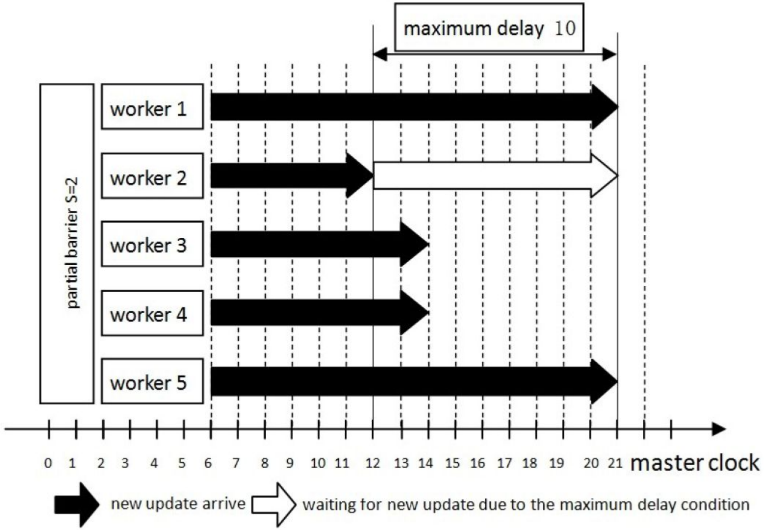
$$L(\{x_i\}, z,) = \sum_{i=1}^{N} f_i(x_i) + \langle \lambda_i, x_i - z \rangle + \frac{\beta}{2} \|x_i - z\|^2$$

Iterations

$$\begin{aligned} x_i^{k+1} &= \arg\min_x f_i(x) + \langle \lambda_i^k, x \rangle + \frac{\beta}{2} \|x - z^k\|^2 \\ z^{k+1} &= \arg\min_z \sum_{i=1}^{N} -\langle \lambda_i^k, z \rangle + \frac{\beta}{2} \|x_i^{k+1} - z\|^2, \\ \lambda_i^{k+1} &= \lambda_i^k + \beta(x_i^{k+1} - z^{k+1}). \end{aligned}$$

Asynchronous (Centralized) ADMM





Partial Barrier and Bounded Delay

Asynchronous (Centralized) ADMM: Algorithm

Master
Repeat:
- Wait for
  - A minimum of S updates from different workers
  - No update delayed more than τ steps
- Update consensus variable z
- Broadcast consensus variable z to workers

Worker i
Repeat:
- Update local variable x_i
- Send local variable x_i and multiplier λ_i to master
- Wait for updated consensus variable z
- Update multiplier λ_i

Asynchronous (Centralized) ADMM: Analysis

Theorem 4.2 Let (x^*, z^*) be the optimal (primal) solutions of the minimization problem (2), and $\{\lambda_i^*\}_{i=1}^N$ the corresponding optimal dual solution. Consider $\bar{x}_i = \frac{1}{T_i}\sum_{t=0}^{T_i-1} x_i^t$ and $\bar{z} = \frac{1}{T}\sum_{t=0}^{T-1} z^t$. Then

$$\mathbb{E}[\sum_{i=1}^{N} \underline{f_i(\bar{x}_i) - f_i(x^*)} + \langle \lambda_i^*, \underline{\bar{x}_i - \bar{z}} \rangle] \leq$$

$$\frac{N\tau}{2TS}[\sum_{i=1}^{N} \beta||z_i^0 - z^*||^2 + \frac{1}{\beta}||\lambda_i^0 - \lambda_i^*||^2]$$

where z_i^0 and λ_i^0 are the initial values of z_i and λ_i respectively, at worker i.

Comparative Performance Guarantees

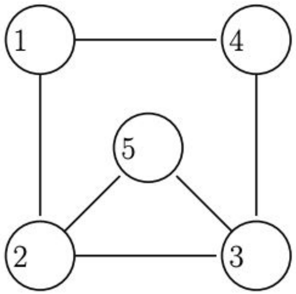
	Synchronous	**Asynchronous**
Centralized	O(1/t) [He, Yuan, 2011]	**O(1/t)** **[Zhang, Kwok, 2014]**
Decentralized	O(1/t) [Wei, Ozdaglar, 2012]	O(1/t) [Wei, Ozdaglar, 2013]

From centralized to decentralized ADMM

- Centralized ADMM requires a master node.
  - High computation cost
  - Risk of losing data if master node fails
  - Crowding of messages
- Extending standard ADMM to a decentralized setting:
  - Lower computation cost.
  - Data privacy.
- How do different topologies affect model performance?

Represent network topology as a matrix

- Represent network topology as a matrix [1].
  - N nodes and M edges.
  - Each node represents an agent.
  - Each node has an associated cost function.
  - Final optimization: To minimize the sum of the loss functions from the N nodes.
  - Incidence matrix is used as a constraint in the optimization task.



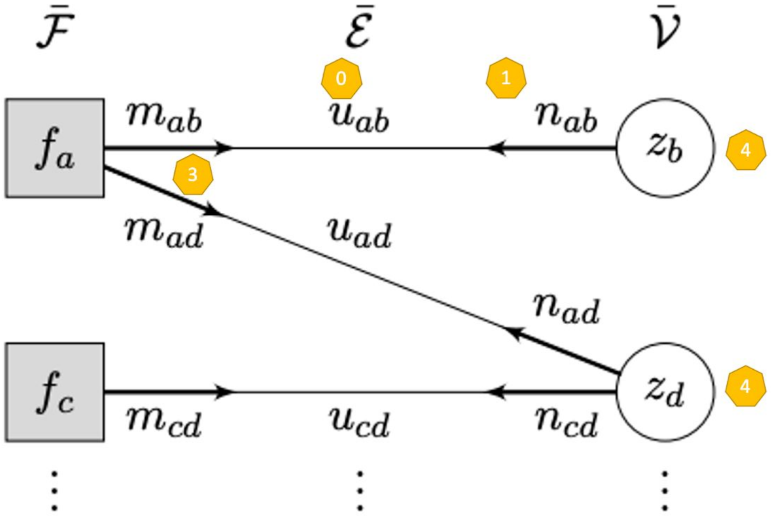
$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

$$\min_{x} \sum_{i=1}^{N} f_i(x_i) \quad s.t. \quad Ax = 0,$$

[1] Wei, Ermin, and Asuman Ozdaglar. "Distributed alternating direction method of multipliers." *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE, 2012.

Formulate ADMM as decentralized version.

- Message passing task [1].



0 $u^t_{ab} \leftarrow u^{t-1}_{ab} + \gamma x^t_{ab} - z^t_b + (1-\gamma) z^{t-1}_b$ for all $(a,b) \in \bar{\mathcal{E}}$

1 $\boxed{n^t_{ab} \leftarrow z^t_b - u^t_{ab}}$ Message passing variable for all $(a,b) \in \bar{\mathcal{E}}$

2 $\boldsymbol{x}^{t+1}_a \leftarrow \arg\min_{\boldsymbol{x}_a} \left\{ f_a(\boldsymbol{x}_a) + \frac{\rho}{2} \sum_{b \in N_a} (x_{ab} - n^t_{ab})^2 \right\}$ for all $a \in \bar{\mathcal{F}}$

3 $\boxed{m^{t+1}_{ab} \leftarrow \gamma x^{t+1}_{ab} + u^t_{ab}}$ Message passing variable for all $(a,b) \in \bar{\mathcal{E}}$

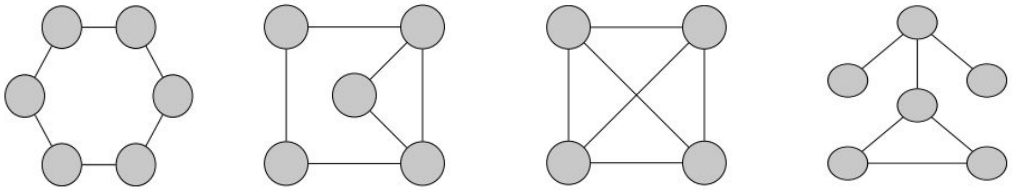
4 $z^{t+1}_b \leftarrow (1-\gamma) z^t_b + \frac{1}{|N_b|} \sum_{a \in N_b} m^{t+1}_{ab}$ for all $b \in \bar{\mathcal{V}}$

0 $u^{t+1}_{ab} \leftarrow u^t_{ab} + \gamma x^{t+1}_{ab} - z^{t+1}_b + (1-\gamma) z^t_b$ for all $(a,b) \in \bar{\mathcal{E}}$

[1]França, Guilherme, and José Bento. "How is distributed ADMM affected by network topology?." arXiv preprint arXiv:1710.00889 (2017).

Main factors in network topologies

- Main factors in network topologies:
    - Graph has at least one cycle of even length.
    - Graph has a cycle, but not with an even length.
    - Graphs without cycles.
    - Example: periodic grid, ring, k-hop lattice, graph sampled from the Erdos-Renyi model.

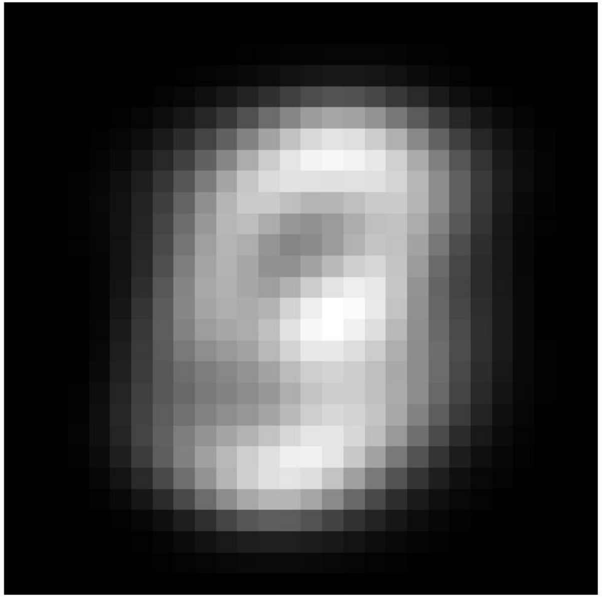


[1]França, Guilherme, and José Bento. "How is distributed ADMM affected by network topology?." arXiv preprint arXiv:1710.00889 (2017).

Empirical Evaluation: Data

- Implemented in Python
- TCP communication
- **N=16,** 16 Workers and a Master, each a physical process
- Approximate average image (pixel-by-pixel) over sample from MNIST
- Objective

$$\min_x f(x) = \sum_{i=1}^{N} ||x - \theta_i||^2$$



Empirical Evaluation: Partial Barrier

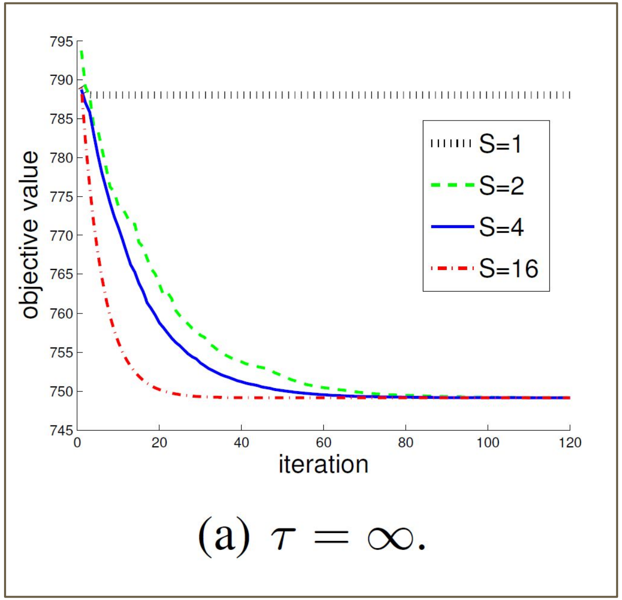
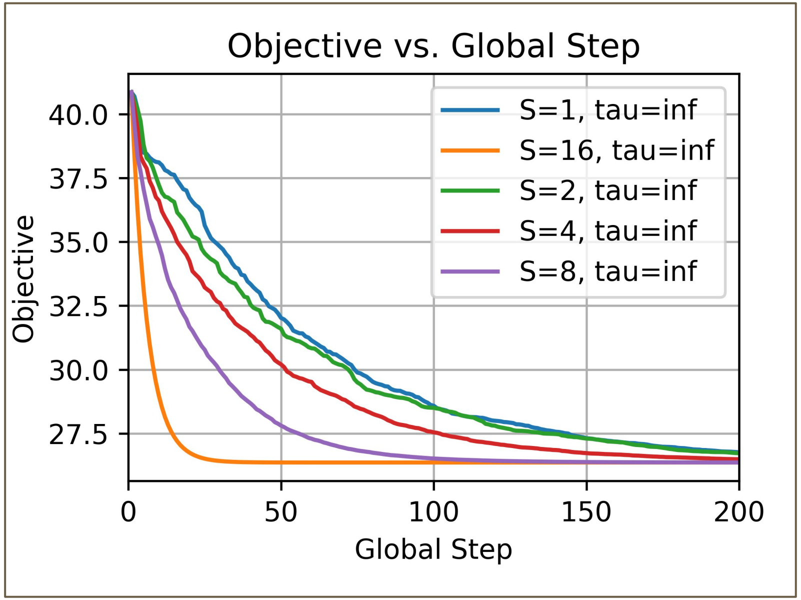


Figure 2a
Avg loss: random numbers
[Zhang and Kwok 14]



200 empirical iterations
Avg loss: MINST sample

Empirical Evaluation: Partial Barrier

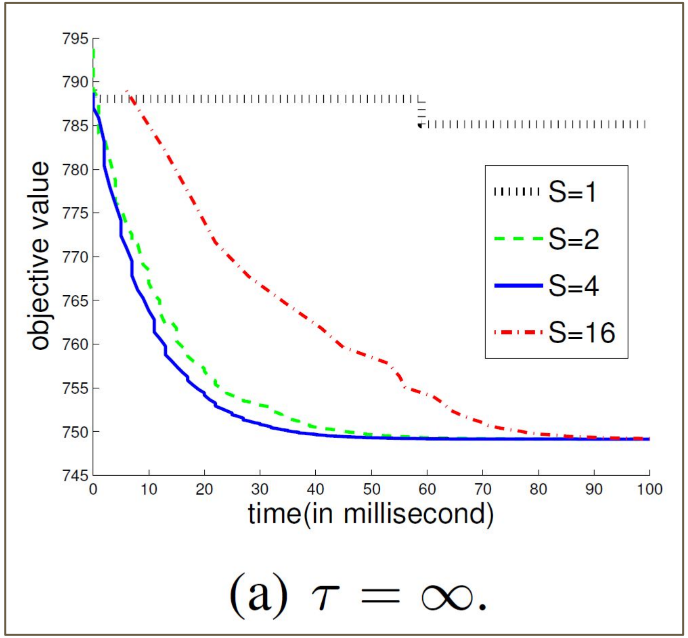
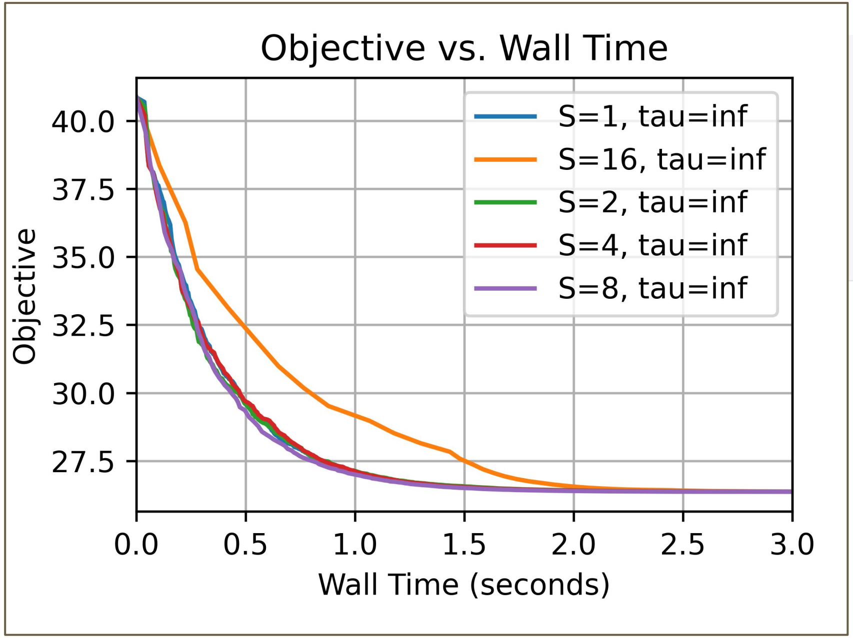


Figure 3a
Avg loss: random numbers
[Zhang and Kwok 14]



200 empirical iterations
Avg loss: MINST sample

Empirical Evaluation: Bounded Delay

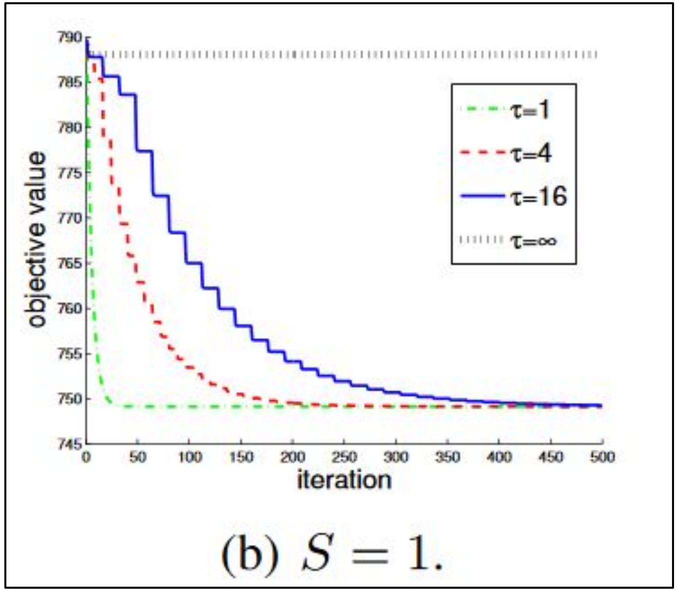
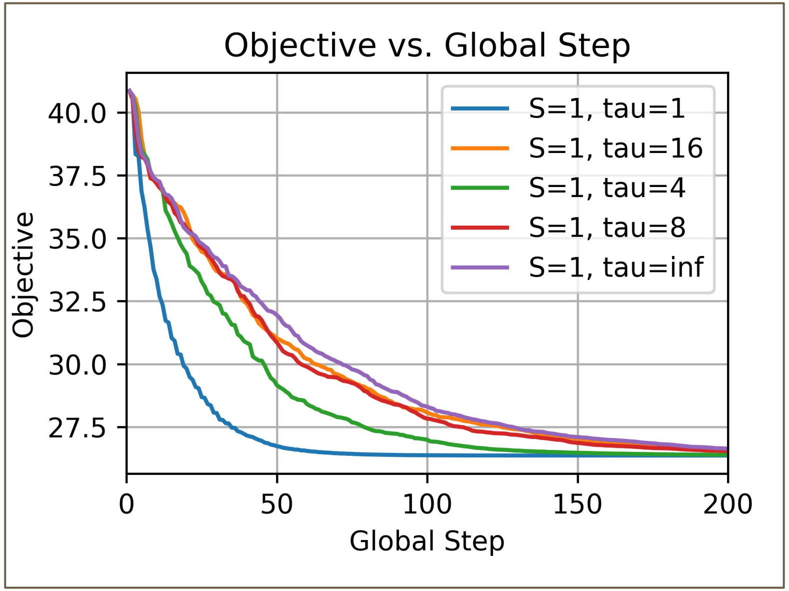
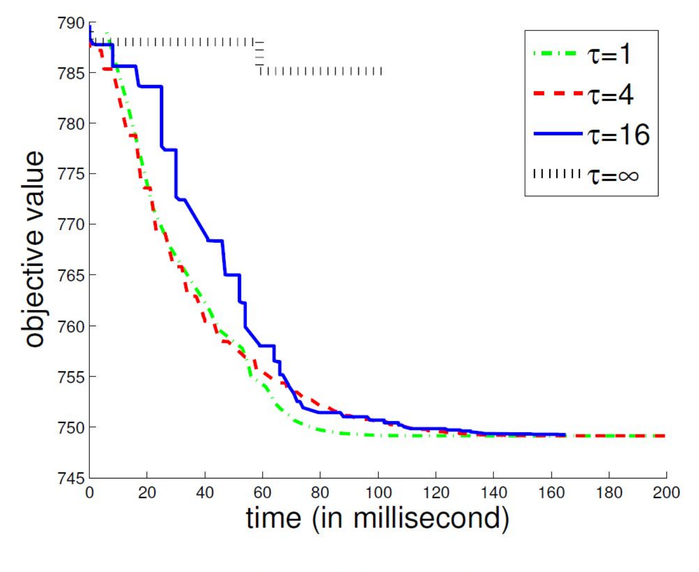


Figure 2b
Avg loss: random numbers
[Zhang and Kwok 14]



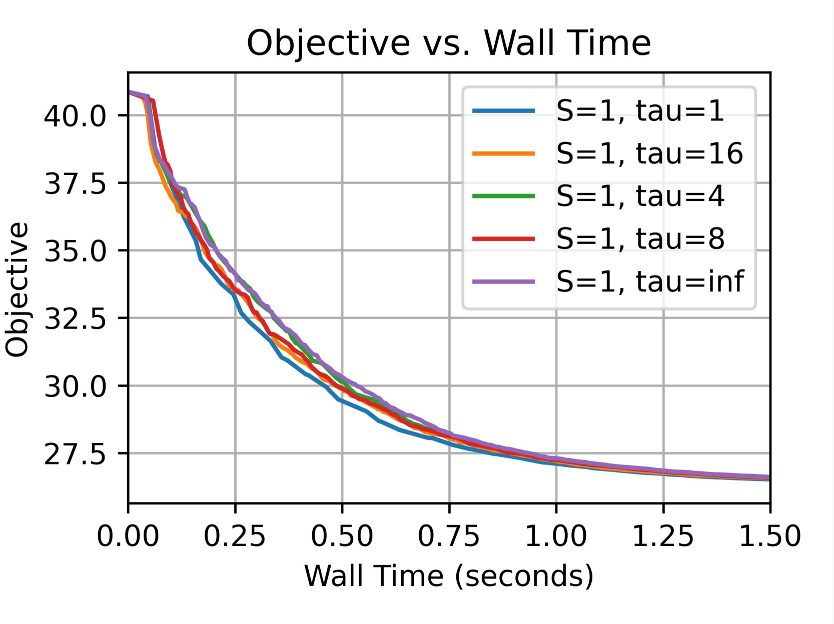
200 empirical iterations
Avg loss: MINST sample

Empirical Evaluation: Bounded Delay



(b) $S = 1.$

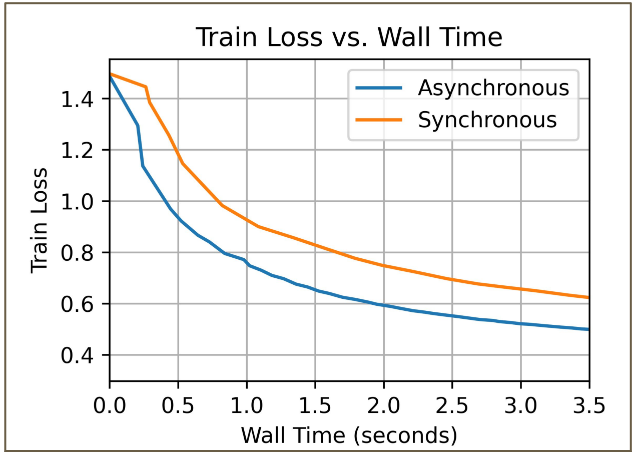
Figure 3b
Avg loss: random numbers
[Zhang and Kwok 14]

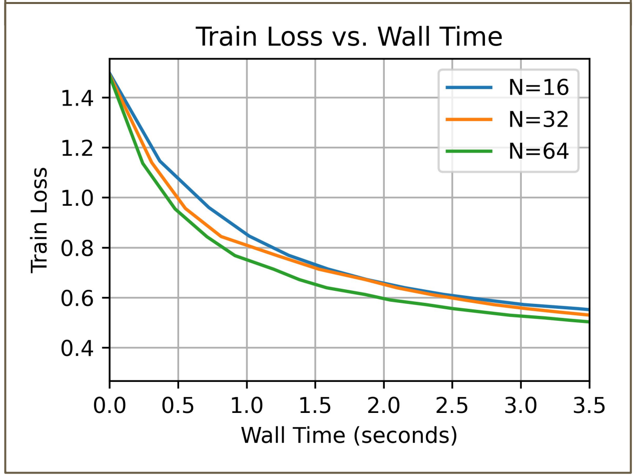


200 empirical iterations
Avg loss: MINST sample

Multiclass Logistic Regression - MNIST

- Each worker does gradient descent to minimize local loss
- Weights and biases as consensus variable
- Faster convergence than synchronous case
- Faster convergence when number of workers is increased





Advantages

- ADMM's flexible framework allows it to optimize a variety of objective functions and use numerous network topologies.
- By using nodes and distributed memory, ADMM has faster speed and lower computational complexity than other algorithms when working with large data sets.

Disadvantages

- Synchronous ADMM suffers from straggling - the algorithm is only as fast as its slowest component.
  - Asynchronous ADMM avoids this problem by avoiding the need to wait for the slower components.
- Centralized ADMM suffers from bottlenecking - all nodes communicate with the central node, which can get congested and slow the performance.
  - Decentralized ADMM avoids this problem by having nodes communicate with neighboring nodes instead of the central node.

Thank You