Distributed Consensus Optimization

Ming Yan
Michigan State University, CMSE/Mathematics

September 14, 2018

why we need decentralized optimization?

Decentralized vehicles/aircrafts coordination [1]

Aircrafts formation

Flock of birds

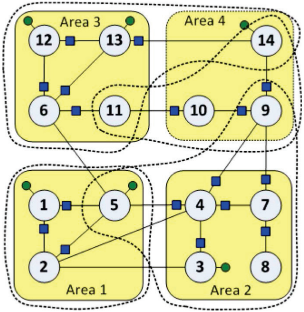
- Average consensus problem

$$\begin{array}{ll}\min_{\{x_{(i)}\}} & \|x_{(1)} - b_1\|_2^2 + \cdots + \|x_{(n)} - b_n\|_2^2 \\ \text{s.t.} & x_{(1)} = \cdots = x_{(n)}\end{array}$$

[1] Ren, Wei, Randal W. Beard, and Ella M. Atkins. "Information consensus in multivehicle cooperative control." IEEE Control Systems 27.2 (2007): 71-82.

why we need decentralized optimization?

Decentralized state estimation of smart grid [2]



- Least squares (Gaussian noise) + ℓ_1 norm (sparse anomalies)

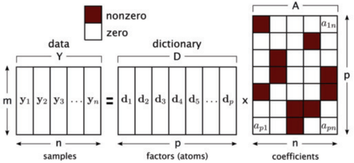
$$\begin{aligned} \min_{\{x_{(i)} \in \mathcal{X}_i, v_{(i)}\}} \quad & \sum_{i=1}^{n} f_i(x_{(i)}, v_{(i)}) \\ \text{s.t.} \quad & x_{(h)}[j] = x_{(j)}[h], \ \forall j \in \mathcal{N}_h, \ \forall h \end{aligned}$$

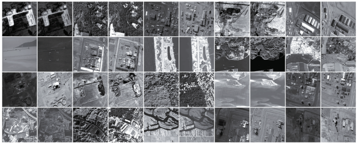
where $f_i(x_{(i)}, v_{(i)}) = \|z_i - H_i x_{(i)} - v_{(i)}\|_2^2 + \lambda \|v_{(i)}\|_1$, and the model parameter λ can be obtained through cross validation

[2] Kekatos, Vassilis, and Georgios B. Giannakis. "Distributed robust power system state estimation." IEEE Transactions on Power Systems 28.2 (2013): 1617-1626.

why we need decentralized optimization?

Decentralized dictionary learning [3]





- Matrix factorization + regularization

$$\min_{D,A} \quad \frac{1}{2}\sum_{j=1}^{c}\left(\|Y_{:,j} - DA_{:,j}\|_{\mathrm{F}}^2 + \lambda\|A_{:,j}\|_1\right) + \gamma\|D\|_{\mathrm{F}}^2, \text{ where}$$

$Y \in \mathbb{R}^{m\times n}$ – training data distributed over c agents
$D \in \mathbb{R}^{m\times p}$ – dictionary that constitutes Y
$A \in \mathbb{R}^{p\times n}$ – sparse coefficient vectors that encode Y, divided into n parts

[3] Wai, Hoi-To, Tsung-Hui Chang, and Anna Scaglione. "A consensus-based decentralized algorithm for non-convex optimization with application to dictionary learning." Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.

why we need decentralized optimization?

Decentralized data/signal processing [4]

- cost/risk minimization: $\min \bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

- Communication and computation balance, robust
- Privacy preservation: Exchange f_i? No! Exchange $x_{(i)}^k$
- ......

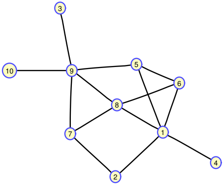
- Unmanned vehicles coordination, in-vehicle networking
- Smart grid management, power station management
- Decentralized recommender systems, multi-group cooperation
- Decentralized network utility maximization
- Decentralized resource allocation
- ......

[4] Ren, Wei, Randal W. Beard, and Ella M. Atkins. "Information consensus in multivehicle cooperative control." IEEE Control Systems 27.2 (2007): 71-82.

what is decentralized optimization?

Decentralized consensus optimization

$$x^* = \arg\min_{x\in\mathcal{C}\subseteq\mathbb{R}^p} \quad \bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \qquad (1)$$



- involves multiple agents
- connected network
- messaging 1-hop neighbors
- each agent owns private objective
- each agent makes local decision
- optimize overall obejective
- all agents reach consensus

- Compared to centralized system: robustness, computation balanced, computation balanced, privacy preserving
- Related topics: in-vehicle networking, internet of things, cloud computing, big data

simplest decentralized consensus problem: averaging

decentralized averaging

One iteration:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^{k},$$

where $\mathbf{x} = [x_1, \ldots, x_n]^\top$.

decentralized averaging

One iteration:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k,$$

where $\mathbf{x} = [x_1, \ldots, x_n]^\top$.

- $\mathbf{W}$ encodes the network; nonzero entries correspond to edges; we assume that $\mathbf{W}$ is symmetric (for undirected networks).

$$\mathbf{W} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} 2/3 & 1/3 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 2/3 \end{bmatrix}$$

decentralized averaging

One iteration:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k,$$

where $\mathbf{x} = [x_1, \ldots, x_n]^\top$.

- $\mathbf{W}$ encodes the network; nonzero entries correspond to edges; we assume that $\mathbf{W}$ is symmetric (for undirected networks).

$$\mathbf{W} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} 2/3 & 1/3 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 2/3 \end{bmatrix}$$

- $\mathbf{1}_{n\times 1}$ is a fixed point, i.e., $\mathbf{W}$ has an eigenvalue 1; row/column sum equals one.

decentralized averaging

One iteration:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^{k},$$

where $\mathbf{x} = [x_1, \ldots, x_n]^\top$.

- $\mathbf{W}$ encodes the network; nonzero entries correspond to edges; we assume that $\mathbf{W}$ is symmetric (for undirected networks).

$$\mathbf{W} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} 2/3 & 1/3 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 2/3 \end{bmatrix}$$

- $\mathbf{1}_{n\times 1}$ is a fixed point, i.e., $\mathbf{W}$ has an eigenvalue 1; row/column sum equals one.
- Any fixed point is consensus, i.e., $\mathbf{x}^* = c\mathbf{1}_{n\times 1}$.

decentralized averaging

One iteration:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k,$$

where $\mathbf{x} = [x_1, \ldots, x_n]^\top$.

- $\mathbf{W}$ encodes the network; nonzero entries correspond to edges; we assume that $\mathbf{W}$ is symmetric (for undirected networks).

$$\mathbf{W} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} 2/3 & 1/3 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 2/3 \end{bmatrix}$$

- $\mathbf{1}_{n\times 1}$ is a fixed point, i.e., $\mathbf{W}$ has an eigenvalue 1; row/column sum equals one.
- Any fixed point is consensus, i.e., $\mathbf{x}^* = c\mathbf{1}_{n\times 1}$.
- $\mathbf{W}$ has all other eigenvalues in $(-1, 1)$.

decentralized averaging

One iteration:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^{k},$$

where $\mathbf{x} = [x_1, \ldots, x_n]^\top$.

- $\mathbf{W}$ encodes the network; nonzero entries correspond to edges; we assume that $\mathbf{W}$ is symmetric (for undirected networks).

$$\mathbf{W} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} 2/3 & 1/3 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 2/3 \end{bmatrix}$$

- $\mathbf{1}_{n\times 1}$ is a fixed point, i.e., $\mathbf{W}$ has an eigenvalue 1; row/column sum equals one.
- Any fixed point is consensus, i.e., $\mathbf{x}^* = c\mathbf{1}_{n\times 1}$.
- $\mathbf{W}$ has all other eigenvalues in $(-1, 1)$.
- $\mathbf{1}^\top\mathbf{x}^{k+1} = \mathbf{1}^\top\mathbf{W}\mathbf{x}^{k} = \mathbf{1}^\top\mathbf{x}^{k}$; the sum is fixed during the iteration.

decentralized averaging

One iteration:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k,$$

where $\mathbf{x} = [x_1, \ldots, x_n]^\top$.

- $\mathbf{W}$ encodes the network; nonzero entries correspond to edges; we assume that $\mathbf{W}$ is symmetric (for undirected networks).

$$\mathbf{W} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} 2/3 & 1/3 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/3 & 2/3 \end{bmatrix}$$

- $\mathbf{1}_{n\times 1}$ is a fixed point, i.e., $\mathbf{W}$ has an eigenvalue 1; row/column sum equals one.
- Any fixed point is consensus, i.e., $\mathbf{x}^* = c\mathbf{1}_{n\times 1}$.
- $\mathbf{W}$ has all other eigenvalues in $(-1, 1)$.
- $\mathbf{1}^\top\mathbf{x}^{k+1} = \mathbf{1}^\top\mathbf{W}\mathbf{x}^k = \mathbf{1}^\top\mathbf{x}^k$; the sum is fixed during the iteration.
- Convergence speed depends on the second largest eigenvalue of $\mathbf{W}$ in absolute value.

decentralized averaging as gradient descent

Decentralized averaging:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^{k}$$

decentralized averaging as gradient descent

Decentralized averaging:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k$$

Rewrite the iteration as:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k = \mathbf{x}^k - (\mathbf{I} - \mathbf{W})\mathbf{x}^k.$$

decentralized averaging as gradient descent

Decentralized averaging:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k$$

Rewrite the iteration as:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k = \mathbf{x}^k - (\mathbf{I} - \mathbf{W})\mathbf{x}^k.$$

It is equivalent to gradient descent with stepsize 1 for

$$\underset{\mathbf{x}}{\text{minimize}} \ \frac{1}{2}\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{x}\|_2^2.$$

decentralized averaging as gradient descent

Decentralized averaging:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k$$

Rewrite the iteration as:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k = \mathbf{x}^k - (\mathbf{I} - \mathbf{W})\mathbf{x}^k.$$

It is equivalent to gradient descent with stepsize 1 for

$$\underset{\mathbf{x}}{\text{minimize}}\ \ \frac{1}{2}\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{x}\|_2^2.$$

The final solution depends on the initial sum $\mathbf{1}^\top\mathbf{x}^0$.

decentralized averaging as gradient descent

Decentralized averaging:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k$$

Rewrite the iteration as:

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k = \mathbf{x}^k - (\mathbf{I} - \mathbf{W})\mathbf{x}^k.$$

It is equivalent to gradient descent with stepsize 1 for

$$\operatorname*{minimize}_{\mathbf{x}} \ \frac{1}{2}\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{x}\|_2^2.$$

The final solution depends on the initial sum $\mathbf{1}^\top \mathbf{x}^0$.

- The Lipschitz constant of $(\mathbf{I} - \mathbf{W})\mathbf{x}$ is smaller than 2, so we can choose stepsize 1.

decentralized gradient descent

Consider problem

$$\underset{\mathbf{x}}{\text{minimize}} \; f(\mathbf{x}) = \sum_{i=1}^{n} {\color{red} f_i(x_i)}, \quad \text{s.t. } {\color{blue} x_1 = x_2 = \cdots = x_n}.$$

decentralized gradient descent

Consider problem

$$\underset{\mathbf{x}}{\text{minimize}} \;\; f(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i), \quad \text{s.t. } x_1 = x_2 = \cdots = x_n.$$

Decentralized gradient descent (DGD) (Nedic-Ozdaglar '09)

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^{k} - \lambda \nabla f(\mathbf{x}^{k}).$$

decentralized gradient descent

Consider problem

$$\underset{\mathbf{x}}{\text{minimize}}\ f(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i), \quad \text{s.t. } x_1 = x_2 = \cdots = x_n.$$

Decentralized gradient descent (DGD) (Nedic-Ozdaglar '09)

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \lambda\nabla f(\mathbf{x}^k).$$

- Rewrite it as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - ((\mathbf{I} - \mathbf{W})\mathbf{x}^k + \lambda\nabla f(\mathbf{x}^k)).$$

  DGD is gradient descent with stepsize one of

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{1}{2}\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}).$$

decentralized gradient descent

Consider problem

$$\underset{\mathbf{x}}{\text{minimize}}\ f(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i), \quad \text{s.t. } x_1 = x_2 = \cdots = x_n.$$

Decentralized gradient descent (DGD) (Nedic-Ozdaglar '09)

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \lambda\nabla f(\mathbf{x}^k).$$

- Rewrite it as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - ((\mathbf{I} - \mathbf{W})\mathbf{x}^k + \lambda\nabla f(\mathbf{x}^k)).$$

  DGD is gradient descent with stepsize one of

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{1}{2}\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}).$$

- The solution is generally non consensus, i.e., $\mathbf{W}\mathbf{x}^* = \mathbf{x}^* + \lambda\nabla f(\mathbf{x}^*) \neq \mathbf{x}^*$.

decentralized gradient descent

Consider problem

$$\underset{\mathbf{x}}{\text{minimize}}\ f(\mathbf{x}) = \sum_{i=1}^{n} f_i(x_i), \quad \text{s.t. } x_1 = x_2 = \cdots = x_n.$$

Decentralized gradient descent (DGD) (Nedic-Ozdaglar '09)

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \lambda \nabla f(\mathbf{x}^k).$$

- Rewrite it as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - ((\mathbf{I} - \mathbf{W})\mathbf{x}^k + \lambda \nabla f(\mathbf{x}^k)).$$

  DGD is gradient descent with stepsize one of

$$\underset{\mathbf{x}}{\text{minimize}}\ \frac{1}{2}\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}).$$

- The solution is generally non consensus, i.e., $\mathbf{W}\mathbf{x}^* = \mathbf{x}^* + \lambda \nabla f(\mathbf{x}^*) \neq \mathbf{x}^*$.
- Diminishing stepsize, i.e., decreasing λ during the iteration.

constant stepsize?

constant stepsize?

- alternating direction method of multipliers (ADMM) (Shi et al. '14, Chang-Hong-Wang '15, Hong-Chang '17)

constant stepsize?

- alternating direction method of multipliers (ADMM) (Shi et al. '14, Chang-Hong-Wang '15, Hong-Chang '17)
- multi-consensus inner loops (Chen-Ozdaglar '12, Jakovetic-Xavier-Moura '14)

constant stepsize?

- alternating direction method of multipliers (ADMM) (Shi et al. '14, Chang-Hong-Wang '15, Hong-Chang '17)
- multi-consensus inner loops (Chen-Ozdaglar '12, Jakovetic-Xavier-Moura '14)
- EXTRA/PG-EXTRA (Shi et al. '15)

decentralized smooth optimization

Problem:

$$\underset{\mathbf{x}}{\text{minimize}}\ \ f(\mathbf{x}), \quad \text{s.t. } \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x} = \mathbf{0}.$$

decentralized smooth optimization

Problem:

$$\underset{\mathbf{x}}{\text{minimize}}\ \ f(\mathbf{x}), \quad \text{s.t. } \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}=\mathbf{0}.$$

- Lagrangian function

$$f(\mathbf{x})+\langle\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x},\mathbf{s}\rangle,$$

where $\mathbf{s}$ is the Lagrangian multiplier.

decentralized smooth optimization

Problem:

$$\operatorname*{minimize}_{\mathbf{x}} \; f(\mathbf{x}), \quad \text{s.t. } \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x} = \mathbf{0}.$$

- Lagrangian function

  $$f(\mathbf{x}) + \langle \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}, \mathbf{s} \rangle,$$

  where $\mathbf{s}$ is the Lagrangian multiplier.

- Optimality condition (KKT):

  $$\begin{aligned} \mathbf{0} =& \nabla f(\mathbf{x}^*) + \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^*, \\ \mathbf{0} =& -\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^*. \end{aligned}$$

decentralized smooth optimization

Problem:

$$\operatorname*{minimize}_{\mathbf{x}} \ f(\mathbf{x}), \quad \text{s.t. } \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x} = \mathbf{0}.$$

- Lagrangian function

$$f(\mathbf{x}) + \langle \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}, \mathbf{s}\rangle,$$

  where $\mathbf{s}$ is the Lagrangian multiplier.

- Optimality condition (KKT):

$$\begin{aligned} \mathbf{0} =& \nabla f(\mathbf{x}^*) + \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^*, \\ \mathbf{0} =& -\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^*. \end{aligned}$$

- It is the same as

$$-\begin{bmatrix} \nabla f(\mathbf{x}^*) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \mathbf{s}^* \end{bmatrix}.$$

forward-backward

- The KKT system

$$
\begin{aligned}
& -\begin{bmatrix}\nabla f(\mathbf{x}^*)\\ \mathbf{0}\end{bmatrix}\\
= & \begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}}\\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}^*\\ \mathbf{s}^*\end{bmatrix}.
\end{aligned}
$$

forward-backward

- Using forward-backward in the KKT form

$$\begin{bmatrix} \alpha\mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{s}^k \end{bmatrix} - \begin{bmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{0} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha\mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix}.$$

forward-backward

- Using forward-backward in the KKT form

$$\begin{bmatrix} \alpha\mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{s}^k \end{bmatrix} - \begin{bmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{0} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha\mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix}.$$

- It reduces to

$$\begin{bmatrix} \alpha\mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{s}^k \end{bmatrix} - \begin{bmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \alpha\mathbf{I} & \mathbf{0} \\ -2\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix}.$$

forward-backward

- Using forward-backward in the KKT form

$$\begin{bmatrix} \alpha \mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{s}^k \end{bmatrix} - \begin{bmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{0} \end{bmatrix}$$
$$= \begin{bmatrix} \alpha \mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix}.$$

- It reduces to

$$\begin{bmatrix} \alpha \mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{s}^k \end{bmatrix} - \begin{bmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \alpha \mathbf{I} & \mathbf{0} \\ -2\sqrt{\mathbf{I}-\mathbf{W}} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix}.$$

- It is equivalent to

$$\begin{aligned} \alpha \mathbf{x}^k - \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^k - \nabla f(\mathbf{x}^k) =& \alpha \mathbf{x}^{k+1}, \\ -\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^k + \beta \mathbf{s}^k =& -2\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^{k+1} + \beta \mathbf{s}^{k+1}. \end{aligned}$$

forward-backward

- Using forward-backward in the KKT form

$$\begin{bmatrix} \alpha \mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{s}^k \end{bmatrix} - \begin{bmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{0} \end{bmatrix}$$
$$= \begin{bmatrix} \alpha \mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix}.$$

- It reduces to

$$\begin{bmatrix} \alpha \mathbf{I} & -\sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{s}^k \end{bmatrix} - \begin{bmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \alpha \mathbf{I} & \mathbf{0} \\ -2\sqrt{\mathbf{I}-\mathbf{W}} & \beta \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix}.$$

- It is equivalent to

$$\begin{aligned} \alpha \mathbf{x}^k - \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^k - \nabla f(\mathbf{x}^k) =& \alpha \mathbf{x}^{k+1}, \\ -\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^k + \beta \mathbf{s}^k =& -2\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^{k+1} + \beta \mathbf{s}^{k+1}. \end{aligned}$$

- For simplicity, let $\mathbf{t} = \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}$, and we have

$$\begin{aligned} \alpha \mathbf{x}^k - \mathbf{t}^k - \nabla f(\mathbf{x}^k) =& \alpha \mathbf{x}^{k+1}, \\ -(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \beta \mathbf{t}^k =& -2(\mathbf{I}-\mathbf{W})\mathbf{x}^{k+1} + \beta \mathbf{t}^{k+1}. \end{aligned}$$

EXact firsT-ordeR Algorithm (EXTRA)

- From the previous slide

$$
\begin{aligned}
\alpha\mathbf{x}^k - \mathbf{t}^k - \nabla f(\mathbf{x}^k) =& \alpha\mathbf{x}^{k+1},\\
-(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \beta\mathbf{t}^k =& -2(\mathbf{I}-\mathbf{W})\mathbf{x}^{k+1} + \beta\mathbf{t}^{k+1}.
\end{aligned}
$$

EXact firsT-ordeR Algorithm (EXTRA)

- From the previous slide

$$\begin{aligned}\alpha\mathbf{x}^k - \mathbf{t}^k - \nabla f(\mathbf{x}^k) =& \alpha\mathbf{x}^{k+1},\\ -(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \beta\mathbf{t}^k =& -2(\mathbf{I}-\mathbf{W})\mathbf{x}^{k+1} + \beta\mathbf{t}^{k+1}.\end{aligned}$$

- We have

$$\begin{aligned}\alpha\mathbf{x}^{k+1} =& \alpha\mathbf{x}^k - \mathbf{t}^k - \nabla f(\mathbf{x}^k)\\ =& \alpha\mathbf{x}^k - \frac{\mathbf{I}-\mathbf{W}}{\beta}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \mathbf{t}^{k-1} - \nabla f(\mathbf{x}^k)\\ =& \alpha\mathbf{x}^k - \frac{\mathbf{I}-\mathbf{W}}{\beta}(2\mathbf{x}^k - \mathbf{x}^{k-1}) + \alpha\mathbf{x}^k + \nabla f(\mathbf{x}^{k-1}) - \alpha\mathbf{x}^{k-1} - \nabla f(\mathbf{x}^k)\\ =& \Big(\alpha\mathbf{I} - \frac{\mathbf{I}-\mathbf{W}}{\beta}\Big)(2\mathbf{x}^k - \mathbf{x}^{k-1}) + \nabla f(\mathbf{x}^{k-1}) - \nabla f(\mathbf{x}^k).\end{aligned}$$

EXact firsT-ordeR Algorithm (EXTRA)

- From the previous slide

$$
\begin{aligned}
\alpha\mathbf{x}^k - \mathbf{t}^k - \nabla f(\mathbf{x}^k) =& \alpha\mathbf{x}^{k+1},\\
-(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \beta\mathbf{t}^k =& -2(\mathbf{I}-\mathbf{W})\mathbf{x}^{k+1} + \beta\mathbf{t}^{k+1}.
\end{aligned}
$$

- We have

$$
\begin{aligned}
\alpha\mathbf{x}^{k+1} =& \alpha\mathbf{x}^k - \mathbf{t}^k - \nabla f(\mathbf{x}^k)\\
=& \alpha\mathbf{x}^k - \frac{\mathbf{I}-\mathbf{W}}{\beta}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \mathbf{t}^{k-1} - \nabla f(\mathbf{x}^k)\\
=& \alpha\mathbf{x}^k - \frac{\mathbf{I}-\mathbf{W}}{\beta}(2\mathbf{x}^k - \mathbf{x}^{k-1}) + \alpha\mathbf{x}^k + \nabla f(\mathbf{x}^{k-1}) - \alpha\mathbf{x}^{k-1} - \nabla f(\mathbf{x}^k)\\
=& \left(\alpha\mathbf{I} - \frac{\mathbf{I}-\mathbf{W}}{\beta}\right)(2\mathbf{x}^k - \mathbf{x}^{k-1}) + \nabla f(\mathbf{x}^{k-1}) - \nabla f(\mathbf{x}^k).
\end{aligned}
$$

- Let $\alpha\beta = 2$ and we have EXTRA (Shi et al. '15)

$$
\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})).
$$

convergence conditions for EXTRA: I

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

convergence conditions for EXTRA: I

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $f = 0$:

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} \mathbf{I}+\mathbf{W} & -\frac{\mathbf{I}+\mathbf{W}}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix}.$$

convergence conditions for EXTRA: I

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $f=0$:

$$\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{x}^k\end{bmatrix} = \begin{bmatrix}\mathbf{I}+\mathbf{W} & -\frac{\mathbf{I}+\mathbf{W}}{2}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{x}^{k-1}\end{bmatrix}.$$

- Let $\mathbf{I}+\mathbf{W} = \mathbf{U}\Sigma\mathbf{U}^\top$.

$$\begin{bmatrix}\mathbf{I}+\mathbf{W} & -\frac{\mathbf{I}+\mathbf{W}}{2}\\ \mathbf{I} & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{U} & \\ & \mathbf{U}\end{bmatrix}\begin{bmatrix}\Sigma & -\frac{\Sigma}{2}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}^\top & 0\\ & \mathbf{U}^\top\end{bmatrix}.$$

convergence conditions for EXTRA: I

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $f = 0$:

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} \mathbf{I}+\mathbf{W} & -\frac{\mathbf{I}+\mathbf{W}}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix}.$$

- Let $\mathbf{I}+\mathbf{W} = \mathbf{U}\Sigma\mathbf{U}^\top$.

$$\begin{bmatrix} \mathbf{I}+\mathbf{W} & -\frac{\mathbf{I}+\mathbf{W}}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{U} & \\ & \mathbf{U} \end{bmatrix} \begin{bmatrix} \Sigma & -\frac{\Sigma}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top & 0 \\ & \mathbf{U}^\top \end{bmatrix}.$$

- The iteration becomes

$$\begin{bmatrix} \mathbf{U}^\top\mathbf{x}^{k+1} \\ \mathbf{U}^\top\mathbf{x}^k \end{bmatrix} = \begin{bmatrix} \Sigma & -\frac{\Sigma}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top\mathbf{x}^k \\ \mathbf{U}^\top\mathbf{x}^{k-1} \end{bmatrix}.$$

convergence conditions for EXTRA: I

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $f = 0$:

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} \mathbf{I}+\mathbf{W} & -\frac{\mathbf{I}+\mathbf{W}}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix}.$$

- Let $\mathbf{I}+\mathbf{W} = \mathbf{U}\Sigma\mathbf{U}^\top$.

$$\begin{bmatrix} \mathbf{I}+\mathbf{W} & -\frac{\mathbf{I}+\mathbf{W}}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{U} & \\ & \mathbf{U} \end{bmatrix} \begin{bmatrix} \Sigma & -\frac{\Sigma}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top & 0 \\ & \mathbf{U}^\top \end{bmatrix}.$$

- The iteration becomes

$$\begin{bmatrix} \mathbf{U}^\top\mathbf{x}^{k+1} \\ \mathbf{U}^\top\mathbf{x}^k \end{bmatrix} = \begin{bmatrix} \Sigma & -\frac{\Sigma}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top\mathbf{x}^k \\ \mathbf{U}^\top\mathbf{x}^{k-1} \end{bmatrix}.$$

- The condition for $\mathbf{W}$ is $-2/3 < \lambda(\Sigma) = \lambda(\mathbf{W}+\mathbf{I}) \leq 2$, which is $-5/3 < \lambda(\mathbf{W}) \leq 1$.

convergence conditions for EXTRA: II

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

convergence conditions for EXTRA: II

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $\nabla f(\mathbf{x}^k) = \mathbf{x}^k - \mathbf{b}$:

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} \mathbf{I}+\mathbf{W}-\frac{1}{\alpha}\mathbf{I} & -\frac{\mathbf{I}+\mathbf{W}}{2}+\frac{1}{\alpha}\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix}.$$

convergence conditions for EXTRA: II

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $\nabla f(\mathbf{x}^k) = \mathbf{x}^k - \mathbf{b}$:

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} \mathbf{I}+\mathbf{W} - \frac{1}{\alpha}\mathbf{I} & -\frac{\mathbf{I}+\mathbf{W}}{2} + \frac{1}{\alpha}\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix}.$$

- Let $\mathbf{I}+\mathbf{W} = \mathbf{U}\Sigma\mathbf{U}^\top$.

$$\begin{bmatrix} \mathbf{I}+\mathbf{W} - \frac{1}{\alpha}\mathbf{I} & -\frac{\mathbf{I}+\mathbf{W}}{2} + \frac{1}{\alpha}\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{U} & \\ & \mathbf{U} \end{bmatrix} \begin{bmatrix} \Sigma - \frac{1}{\alpha}\mathbf{I} & -\frac{\Sigma}{2} + \frac{1}{\alpha}\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top & 0 \\ & \mathbf{U}^\top \end{bmatrix}.$$

convergence conditions for EXTRA: II

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $\nabla f(\mathbf{x}^k) = \mathbf{x}^k - \mathbf{b}$:

$$\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{x}^k\end{bmatrix} = \begin{bmatrix}\mathbf{I}+\mathbf{W}-\frac{1}{\alpha}\mathbf{I} & -\frac{\mathbf{I}+\mathbf{W}}{2}+\frac{1}{\alpha}\mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{x}^{k-1}\end{bmatrix}.$$

- Let $\mathbf{I}+\mathbf{W} = \mathbf{U}\Sigma\mathbf{U}^\top$.

$$\begin{bmatrix}\mathbf{I}+\mathbf{W}-\frac{1}{\alpha}\mathbf{I} & -\frac{\mathbf{I}+\mathbf{W}}{2}+\frac{1}{\alpha}\mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{bmatrix} = \begin{bmatrix}\mathbf{U} & \\ & \mathbf{U}\end{bmatrix}\begin{bmatrix}\Sigma-\frac{1}{\alpha}\mathbf{I} & -\frac{\Sigma}{2}+\frac{1}{\alpha}\mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}^\top & 0\\ & \mathbf{U}^\top\end{bmatrix}.$$

- The condition for $\mathbf{W}$ is $4/(3\alpha) - 2/3 < \lambda(\Sigma) = \lambda(\mathbf{W}+\mathbf{I}) \leq 2$, which is $4/(3\alpha) - 5/3 < \lambda(\mathbf{W}) \leq 1$. In addition, we have stepsize $1/\alpha < 2$.

conditions for general EXTRA

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})).$$

conditions for general EXTRA

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})).$$

Initial condition $(k = 0, 1)$:

$$\mathbf{x}^1 = \mathbf{x}^0 - \frac{1}{\alpha}\nabla f(\mathbf{x}^0).$$

conditions for general EXTRA

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})).$$

Initial condition ($k = 0, 1$):

$$\mathbf{x}^1 = \mathbf{x}^0 - \frac{1}{\alpha}\nabla f(\mathbf{x}^0).$$

Convergence condition (Li-Yan '17):

$$\begin{aligned} 4/(3\alpha) - 5/3 < &\ \lambda(\mathbf{W}) \leq 1, \\ 1/\alpha < &\ 2/L. \end{aligned}$$

conditions for general EXTRA

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})).$$

Initial condition ($k = 0, 1$):

$$\mathbf{x}^1 = \mathbf{x}^0 - \frac{1}{\alpha}\nabla f(\mathbf{x}^0).$$

Convergence condition (Li-Yan '17):

$$\begin{aligned} 4/(3\alpha) - 5/3 <&\ \lambda(\mathbf{W}) \leq 1, \\ 1/\alpha <&\ 2/L. \end{aligned}$$

Linear convergence condition:

- $f(\mathbf{x})$ is strongly convex. (Li-Yan '17)

conditions for general EXTRA

EXTRA:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})).$$

Initial condition ($k = 0, 1$):

$$\mathbf{x}^1 = \mathbf{x}^0 - \frac{1}{\alpha}\nabla f(\mathbf{x}^0).$$

Convergence condition (Li-Yan '17):

$$\begin{aligned} 4/(3\alpha) - 5/3 < \ & \lambda(\mathbf{W}) \leq 1, \\ 1/\alpha < \ & 2/L. \end{aligned}$$

Linear convergence condition:

- $f(\mathbf{x})$ is strongly convex. (Li-Yan '17)
- weaker condition on $f(\mathbf{x})$ but more restrict condition for both parameters. (Shi et al. '15)

large stepsize as centralized ones?

decentralized smooth optimization

Problem:

$$\underset{\mathbf{x}}{\text{minimize}} \;\; f(\mathbf{x}), \quad \text{s.t. } \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x} = \mathbf{0}.$$

decentralized smooth optimization

Problem:

$$\underset{\mathbf{x}}{\text{minimize}} \;\; f(\mathbf{x}), \quad \text{s.t. } \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x} = \mathbf{0}.$$

- Lagrangian function

$$f(\mathbf{x}) + \langle \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}, \mathbf{s} \rangle,$$

where $\mathbf{s}$ is the Lagrangian multiplier.

decentralized smooth optimization

Problem:

$$\underset{\mathbf{x}}{\text{minimize}} \;\; f(\mathbf{x}), \quad \text{s.t. } \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x} = \mathbf{0}.$$

- Lagrangian function

$$f(\mathbf{x}) + \langle \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}, \mathbf{s}\rangle,$$

  where $\mathbf{s}$ is the Lagrangian multiplier.

- Optimality condition (KKT):

$$\begin{aligned} \mathbf{0} =& \nabla f(\mathbf{x}^*) + \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^*, \\ \mathbf{0} =& -\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^*. \end{aligned}$$

decentralized smooth optimization

Problem:

$$\underset{\mathbf{x}}{\text{minimize}} \ f(\mathbf{x}), \quad \text{s.t. } \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x} = \mathbf{0}.$$

- Lagrangian function

$$f(\mathbf{x}) + \langle \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}, \mathbf{s} \rangle,$$

  where $\mathbf{s}$ is the Lagrangian multiplier.

- Optimality condition (KKT):

$$\begin{aligned} \mathbf{0} =& \nabla f(\mathbf{x}^*) + \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^*, \\ \mathbf{0} =& -\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^*. \end{aligned}$$

- It is the same as

$$-\begin{bmatrix} \nabla f(\mathbf{x}^*) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \mathbf{s}^* \end{bmatrix}.$$

forward-backward

- The KKT system

$$
\begin{aligned}
&-\begin{bmatrix}\nabla f(\mathbf{x}^*)\\ \mathbf{0}\end{bmatrix}\\
=&\begin{bmatrix}\mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}}\\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}^*\\ \mathbf{s}^*\end{bmatrix}.
\end{aligned}
$$

forward-backward

- Using forward-backward in the KKT form

$$
\begin{aligned}
&\begin{bmatrix} \alpha\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \beta\mathbf{I} - \frac{1}{\alpha}(\mathbf{I}-\mathbf{W}) \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{s}^k \end{bmatrix} - \begin{bmatrix} \nabla f(\mathbf{x}^k) \\ \mathbf{0} \end{bmatrix} \\
=&\begin{bmatrix} \alpha\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \beta\mathbf{I} - \frac{1}{\alpha}(\mathbf{I}-\mathbf{W}) \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \sqrt{\mathbf{I}-\mathbf{W}} \\ -\sqrt{\mathbf{I}-\mathbf{W}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{s}^{k+1} \end{bmatrix}.
\end{aligned}
$$

forward-backward

- Combine the right hand side:

$$\begin{bmatrix}\alpha\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \beta\mathbf{I}-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{s}^k\end{bmatrix}-\begin{bmatrix}\nabla f(\mathbf{x}^k)\\ \mathbf{0}\end{bmatrix}$$

$$=\begin{bmatrix}\alpha\mathbf{I} & \sqrt{\mathbf{I}-\mathbf{W}}\\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I}-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\end{bmatrix}\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{s}^{k+1}\end{bmatrix}.$$

forward-backward

- Combine the right hand side:

$$\begin{bmatrix}\alpha\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \beta\mathbf{I}-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{s}^k\end{bmatrix}-\begin{bmatrix}\nabla f(\mathbf{x}^k)\\ \mathbf{0}\end{bmatrix}$$
$$=\begin{bmatrix}\alpha\mathbf{I} & \sqrt{\mathbf{I}-\mathbf{W}}\\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I}-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\end{bmatrix}\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{s}^{k+1}\end{bmatrix}.$$

- Apply Gaussian elimination:

$$\begin{bmatrix}\alpha\mathbf{I} & \mathbf{0}\\ \sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I}-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{s}^k\end{bmatrix}-\begin{bmatrix}\nabla f(\mathbf{x}^k)\\ \frac{1}{\alpha}\sqrt{\mathbf{I}-\mathbf{W}}\nabla f(\mathbf{x}^k)\end{bmatrix}$$
$$=\begin{bmatrix}\alpha\mathbf{I} & \sqrt{\mathbf{I}-\mathbf{W}}\\ \mathbf{0} & \beta\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{s}^{k+1}\end{bmatrix}.$$

forward-backward

- Combine the right hand side:

$$\begin{bmatrix}\alpha\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \beta\mathbf{I}-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{s}^k\end{bmatrix}-\begin{bmatrix}\nabla f(\mathbf{x}^k)\\ \mathbf{0}\end{bmatrix}$$
$$=\begin{bmatrix}\alpha\mathbf{I} & \sqrt{\mathbf{I}-\mathbf{W}}\\ -\sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I}-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\end{bmatrix}\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{s}^{k+1}\end{bmatrix}.$$

- Apply Gaussian elimination:

$$\begin{bmatrix}\alpha\mathbf{I} & \mathbf{0}\\ \sqrt{\mathbf{I}-\mathbf{W}} & \beta\mathbf{I}-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{s}^k\end{bmatrix}-\begin{bmatrix}\nabla f(\mathbf{x}^k)\\ \frac{1}{\alpha}\sqrt{\mathbf{I}-\mathbf{W}}\nabla f(\mathbf{x}^k)\end{bmatrix}$$
$$=\begin{bmatrix}\alpha\mathbf{I} & \sqrt{\mathbf{I}-\mathbf{W}}\\ \mathbf{0} & \beta\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{s}^{k+1}\end{bmatrix}.$$

- It is equivalent to

$$\begin{aligned}\alpha\mathbf{x}^k-\nabla f(\mathbf{x}^k)-\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^{k+1}&=\alpha\mathbf{x}^{k+1},\\ \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^k+\beta\Big(\mathbf{I}-\frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{s}^k-\frac{1}{\alpha}\sqrt{\mathbf{I}-\mathbf{W}}\nabla f(\mathbf{x}^k)&=\beta\mathbf{s}^{k+1}.\end{aligned}$$

NIDS (Li-Shi-Yan '17)

From the previous slide:

$$\begin{aligned}
\alpha\mathbf{x}^k - \nabla f(\mathbf{x}^k) - \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^{k+1} =& \alpha\mathbf{x}^{k+1}, \\
\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^k + \beta\Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{s}^k - \frac{1}{\alpha}\sqrt{\mathbf{I}-\mathbf{W}}\nabla f(\mathbf{x}^k) =& \beta\mathbf{s}^{k+1}.
\end{aligned}$$

NIDS (Li-Shi-Yan '17)

From the previous slide:

$$\begin{aligned}\alpha\mathbf{x}^k-\nabla f(\mathbf{x}^k)-\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}^{k+1}=&\alpha\mathbf{x}^{k+1},\\ \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{x}^k+\beta\Big(\mathbf{I}-\frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{s}^k-\frac{1}{\alpha}\sqrt{\mathbf{I}-\mathbf{W}}\nabla f(\mathbf{x}^k)=&\beta\mathbf{s}^{k+1}.\end{aligned}$$

Let $\mathbf{t}=\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}$:

$$\begin{aligned}\alpha\mathbf{x}^k-\nabla f(\mathbf{x}^k)-\mathbf{t}^{k+1}=&\alpha\mathbf{x}^{k+1},\\ -(\mathbf{I}-\mathbf{W})\mathbf{x}^k+\beta\Big(\mathbf{I}-\frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{t}^k-\frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\nabla f(\mathbf{x}^k)=&\beta\mathbf{t}^{k+1}.\end{aligned}$$

NIDS (Li-Shi-Yan '17)

Let $\mathbf{t} = \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}$:

$$\begin{aligned} \alpha\mathbf{x}^k - \nabla f(\mathbf{x}^k) - \mathbf{t}^{k+1} =& \alpha\mathbf{x}^{k+1}, \\ -(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \beta\Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{t}^k - \frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\nabla f(\mathbf{x}^k) =& \beta\mathbf{t}^{k+1}. \end{aligned}$$

NIDS (Li-Shi-Yan '17)

Let $\mathbf{t} = \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}$:

$$\begin{aligned} \alpha\mathbf{x}^k - \nabla f(\mathbf{x}^k) - \mathbf{t}^{k+1} =& \alpha\mathbf{x}^{k+1}, \\ -(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \beta\Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{t}^k - \frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\nabla f(\mathbf{x}^k) =& \beta\mathbf{t}^{k+1}. \end{aligned}$$

We have

$$\begin{aligned} & \alpha\mathbf{x}^{k+1} \\ =& \alpha\mathbf{x}^k - \nabla f(\mathbf{x}^k) - \mathbf{t}^{k+1} \\ =& \alpha\mathbf{x}^k - \nabla f(\mathbf{x}^k) - \Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{t}^k - \frac{1}{\beta}(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\nabla f(\mathbf{x}^k) \\ =& \Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)(\alpha\mathbf{x}^k - \mathbf{t}^k - \nabla f(\mathbf{x}^k)) \\ =& \Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)(\alpha\mathbf{x}^k + \alpha\mathbf{x}^k - \alpha\mathbf{x}^{k-1} + \nabla f(\mathbf{x}^{k-1}) - \nabla f(\mathbf{x}^k)). \end{aligned}$$

NIDS (Li-Shi-Yan '17)

Let $\mathbf{t} = \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{s}$:

$$\begin{aligned}
\alpha\mathbf{x}^k - \nabla f(\mathbf{x}^k) - \mathbf{t}^{k+1} =& \alpha\mathbf{x}^{k+1},\\
-(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \beta\Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{t}^k - \frac{1}{\alpha}(\mathbf{I}-\mathbf{W})\nabla f(\mathbf{x}^k) =& \beta\mathbf{t}^{k+1}.
\end{aligned}$$

We have

$$\begin{aligned}
&\alpha\mathbf{x}^{k+1}\\
=&\alpha\mathbf{x}^k - \nabla f(\mathbf{x}^k) - \mathbf{t}^{k+1}\\
=&\alpha\mathbf{x}^k - \nabla f(\mathbf{x}^k) - \Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)\mathbf{t}^k - \frac{1}{\beta}(\mathbf{I}-\mathbf{W})\mathbf{x}^k + \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\nabla f(\mathbf{x}^k)\\
=&\Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)(\alpha\mathbf{x}^k - \mathbf{t}^k - \nabla f(\mathbf{x}^k))\\
=&\Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)(\alpha\mathbf{x}^k + \alpha\mathbf{x}^k - \alpha\mathbf{x}^{k-1} + \nabla f(\mathbf{x}^{k-1}) - \nabla f(\mathbf{x}^k)).
\end{aligned}$$

Thus

$$\mathbf{x}^{k+1} = \Big(\mathbf{I} - \frac{1}{\alpha\beta}(\mathbf{I}-\mathbf{W})\Big)(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))).$$

convergence conditions for NIDS

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I} + \mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

convergence conditions for NIDS

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $f = 0$ (same as EXTRA): The condition for $\mathbf{W}$ is $-5/3 < \lambda(\mathbf{W}) \leq 1$.

convergence conditions for NIDS

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $f = 0$ (same as EXTRA): The condition for $\mathbf{W}$ is $-5/3 < \lambda(\mathbf{W}) \leq 1$.
- If $\nabla f(\mathbf{x}^k) = \mathbf{x}^k - \mathbf{b}$:

$$\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{x}^k\end{bmatrix} = \begin{bmatrix}(2-\frac{1}{\alpha})\frac{\mathbf{I}+\mathbf{W}}{2} & -(1-\frac{1}{\alpha})\frac{\mathbf{I}+\mathbf{W}}{2}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{x}^{k-1}\end{bmatrix}$$

convergence conditions for NIDS

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $f = 0$ (same as EXTRA): The condition for $\mathbf{W}$ is $-5/3 < \lambda(\mathbf{W}) \leq 1$.
- If $\nabla f(\mathbf{x}^k) = \mathbf{x}^k - \mathbf{b}$:

$$\begin{bmatrix}\mathbf{x}^{k+1}\\ \mathbf{x}^k\end{bmatrix} = \begin{bmatrix}(2-\frac{1}{\alpha})\frac{\mathbf{I}+\mathbf{W}}{2} & -(1-\frac{1}{\alpha})\frac{\mathbf{I}+\mathbf{W}}{2}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}^k\\ \mathbf{x}^{k-1}\end{bmatrix}$$

- Let $\mathbf{I}+\mathbf{W} = \mathbf{U}\Sigma\mathbf{U}^\top$.

$$\begin{bmatrix}\mathbf{U}^\top\mathbf{x}^{k+1}\\ \mathbf{U}^\top\mathbf{x}^k\end{bmatrix} = \begin{bmatrix}(2-\frac{1}{\alpha})\frac{\Sigma}{2} & -(1-\frac{1}{\alpha})\frac{\Sigma}{2}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}^\top\mathbf{x}^k\\ \mathbf{U}^\top\mathbf{x}^{k-1}\end{bmatrix}$$

convergence conditions for NIDS

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- If $f = 0$ (same as EXTRA): The condition for $\mathbf{W}$ is $-5/3 < \lambda(\mathbf{W}) \leq 1$.
- If $\nabla f(\mathbf{x}^k) = \mathbf{x}^k - \mathbf{b}$:

$$\begin{bmatrix} \mathbf{x}^{k+1} \\ \mathbf{x}^k \end{bmatrix} = \begin{bmatrix} (2-\frac{1}{\alpha})\frac{\mathbf{I}+\mathbf{W}}{2} & -(1-\frac{1}{\alpha})\frac{\mathbf{I}+\mathbf{W}}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^k \\ \mathbf{x}^{k-1} \end{bmatrix}$$

- Let $\mathbf{I}+\mathbf{W} = \mathbf{U}\Sigma\mathbf{U}^\top$.

$$\begin{bmatrix} \mathbf{U}^\top\mathbf{x}^{k+1} \\ \mathbf{U}^\top\mathbf{x}^k \end{bmatrix} = \begin{bmatrix} (2-\frac{1}{\alpha})\frac{\Sigma}{2} & -(1-\frac{1}{\alpha})\frac{\Sigma}{2} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top\mathbf{x}^k \\ \mathbf{U}^\top\mathbf{x}^{k-1} \end{bmatrix}$$

- Therefore, one sufficient condition is $-5/3 < \lambda(\mathbf{W}) \leq 1$ and $1/\alpha < 2$.

conditions of NIDS for general smooth functions

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

conditions of NIDS for general smooth functions

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

Initial condition ($k = 0, 1$):

$$\mathbf{x}^1 = \mathbf{x}^0 - \frac{1}{\alpha}\nabla f(\mathbf{x}^0).$$

conditions of NIDS for general smooth functions

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

Initial condition ($k = 0, 1$):

$$\mathbf{x}^1 = \mathbf{x}^0 - \frac{1}{\alpha}\nabla f(\mathbf{x}^0).$$

Convergence condition (Li-Yan '17):

$$-5/3 < \lambda(\mathbf{W}) \leq 1,$$
$$1/\alpha < 2/L.$$

conditions of NIDS for general smooth functions

NIDS (with $\alpha\beta = 2$):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

Initial condition ($k = 0, 1$):

$$\mathbf{x}^1 = \mathbf{x}^0 - \frac{1}{\alpha}\nabla f(\mathbf{x}^0).$$

Convergence condition (Li-Yan '17):

$$\begin{aligned} -5/3 <&\ \lambda(\mathbf{W}) \leq 1, \\ 1/\alpha <&\ 2/L. \end{aligned}$$

Linear convergence condition:

- $f(\mathbf{x})$ is strongly convex and $-1 < \lambda(\mathbf{W}) \leq 1$ (Li-Shi-Yan '17):

$$O\left(\max\left(1 - \frac{\mu}{L}, 1 - \frac{1-\lambda_2(\mathbf{W})}{1-\lambda_n(\mathbf{W})}\right)\right).$$

NIDS vs EXTRA

EXTRA

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})))$$

NIDS vs EXTRA

EXTRA

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})))$$

- The difference is in the data to be communicated.

NIDS vs EXTRA

EXTRA

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})))$$

- The difference is in the data to be communicated.
- But NIDS has a larger range for parameters than EXTRA.

NIDS vs EXTRA

EXTRA

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1}) - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1})))$$

- The difference is in the data to be communicated.
- But NIDS has a larger range for parameters than EXTRA.
- NIDS is faster than EXTRA.

advantages of NIDS

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

advantages of NIDS

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- The stepsize is large and does not depend on the network topology.

$$\frac{1}{\alpha} < \frac{2}{L}.$$

advantages of NIDS

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- The stepsize is large and does not depend on the network topology.

$$\frac{1}{\alpha} < \frac{2}{L}.$$

- Individual stepsizes can be included.

$$\frac{1}{\alpha_i} < \frac{2}{L_i}.$$

advantages of NIDS

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

- The stepsize is large and does not depend on the network topology.

$$\frac{1}{\alpha} < \frac{2}{L}.$$

- Individual stepsizes can be included.

$$\frac{1}{\alpha_i} < \frac{2}{L_i}.$$

- The linear convergence rate from the functions and the network are separated.

$$O\left(\max\left(1-\frac{\mu}{L}, 1-\frac{1-\lambda_2(\mathbf{W})}{1-\lambda_n(\mathbf{W})}\right)\right).$$

It matches the results for gradient descent and decentralized averaging without acceleration.

D^2: stochastic NIDS (Huang et al. '18)

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

D^2: stochastic NIDS (Huang et al. '18)

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

NIDS-stochastic (D^2: Decentralized Training over Decentralized Data):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k, \xi_k) - \nabla f(\mathbf{x}^{k-1}, \xi_{k-1}))$$

- $\nabla f(\mathbf{x}^k, \xi_k)$ is a stochastic gradient by sampling ξ_t from distribution $\mathcal{D}$.

D^2: stochastic NIDS (Huang et al. '18)

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

NIDS-stochastic (D^2: Decentralized Training over Decentralized Data):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k, \xi_k) - \nabla f(\mathbf{x}^{k-1}, \xi_{k-1}))$$

- $\nabla f(\mathbf{x}^k, \xi_k)$ is a stochastic gradient by sampling ξ_t from distribution $\mathcal{D}$.
- $\mathbb{E}_{\xi\sim\mathcal{D}}\nabla f(\mathbf{x};\xi) = \nabla f(\mathbf{x}), \quad \forall \mathbf{x}$.

D^2: stochastic NIDS (Huang et al. '18)

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

NIDS-stochastic (D^2: Decentralized Training over Decentralized Data):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k, \xi_k) - \nabla f(\mathbf{x}^{k-1}, \xi_{k-1}))$$

- $\nabla f(\mathbf{x}^k, \xi_k)$ is a stochastic gradient by sampling ξ_t from distribution $\mathcal{D}$.
- $\mathbb{E}_{\xi\sim\mathcal{D}} \nabla f(\mathbf{x}; \xi) = \nabla f(\mathbf{x}), \quad \forall \mathbf{x}$.
- $\mathbb{E}_{\xi\sim\mathcal{D}} \|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{x})\|^2 \leqslant \sigma^2, \quad \forall \mathbf{x}$.

D^2: stochastic NIDS (Huang et al. '18)

NIDS:

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^{k-1}))$$

NIDS-stochastic (D^2: Decentralized Training over Decentralized Data):

$$\mathbf{x}^{k+1} = \frac{\mathbf{I}+\mathbf{W}}{2}(2\mathbf{x}^k - \mathbf{x}^{k-1} - \frac{1}{\alpha}(\nabla f(\mathbf{x}^k, \xi_k) - \nabla f(\mathbf{x}^{k-1}, \xi_{k-1}))$$

- $\nabla f(\mathbf{x}^k, \xi_k)$ is a stochastic gradient by sampling ξ_t from distribution $\mathcal{D}$.
- $\mathbb{E}_{\xi\sim\mathcal{D}} \nabla f(\mathbf{x};\xi) = \nabla f(\mathbf{x}), \quad \forall \mathbf{x}$.
- $\mathbb{E}_{\xi\sim\mathcal{D}} \|\nabla f(\mathbf{x};\xi) - \nabla f(\mathbf{x})\|^2 \leqslant \sigma^2, \quad \forall \mathbf{x}$.
- Convergence result: if the stepsize is small enough (in the order of $(c + \sqrt{T/n})^{-1}$), the convergence rate is

$$O\left(\frac{\sigma}{\sqrt{nT}} + \frac{1}{T}\right).$$

numerical experiments

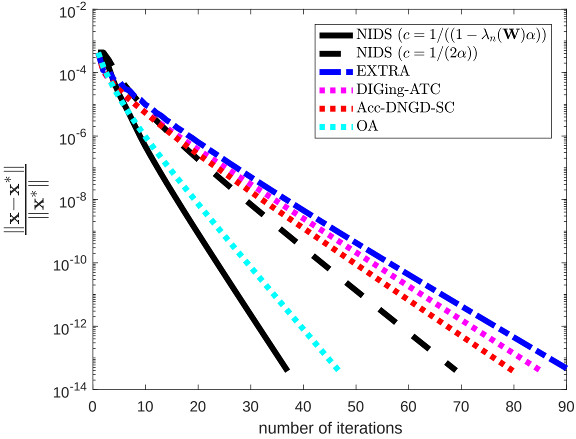
compared algorithms

- NIDS
- EXTRA/PG-EXTRA
- DIGing-ATC (Nedic et al. '16):

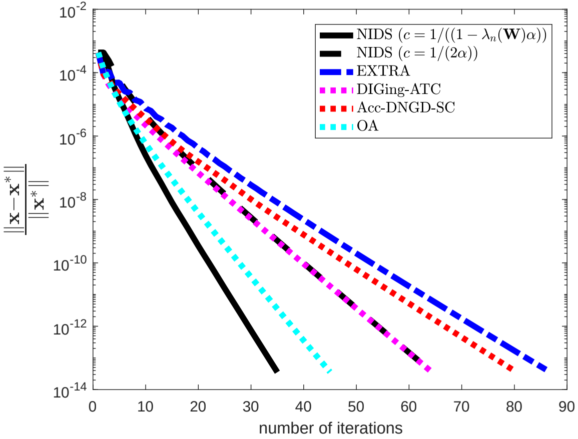
$$\begin{aligned}\mathbf{x}^{k+1} =& \mathbf{W}(\mathbf{x}^k - \alpha \mathbf{y}^k),\\ \mathbf{y}^{k+1} =& \mathbf{W}(\mathbf{y}^k + \nabla f(\mathbf{x}^{k+1}) - \nabla f(\mathbf{x}^k)).\end{aligned}$$

- accelerated distributed Nesterov gradient descent (Acc-DNGD-SC in (Qu-Li '17)
- dual friendly optimal algorithm (OA) for distributed optimization (Uribe et al. '17).

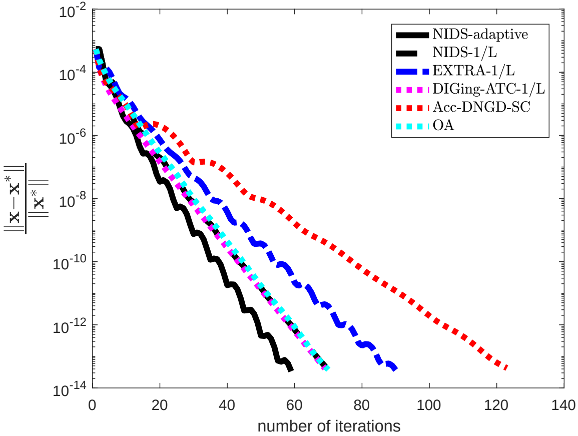
strongly convex: same stepsize



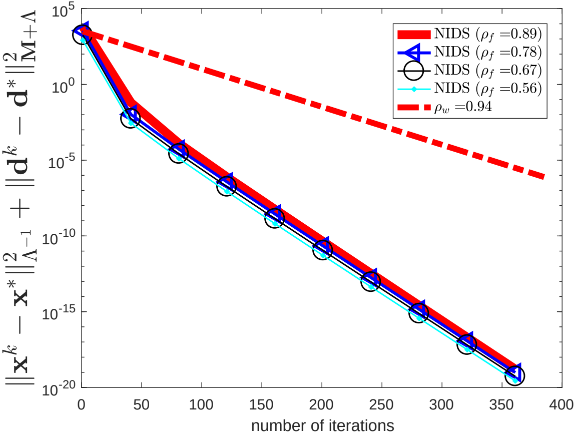
strongly convex: same stepsize



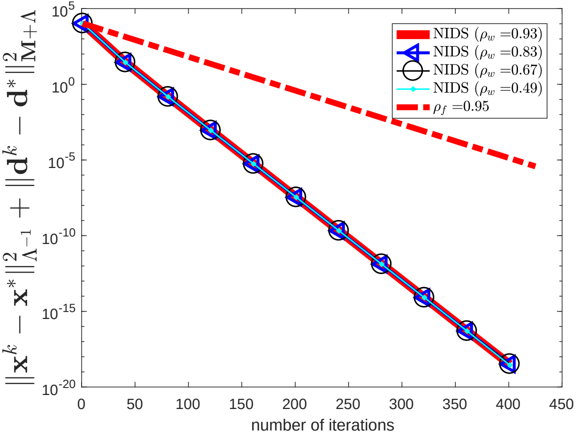
strongly convex: adaptive stepsize



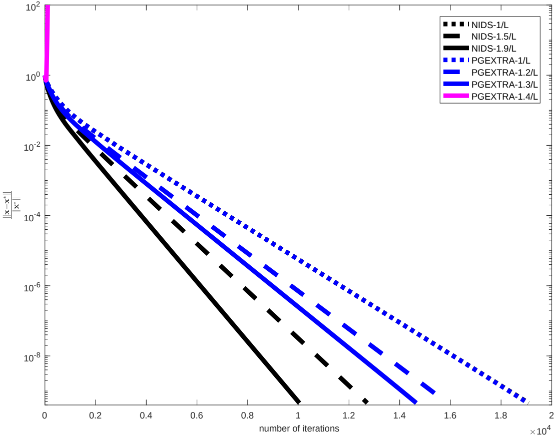
linear convergence rate bottleneck



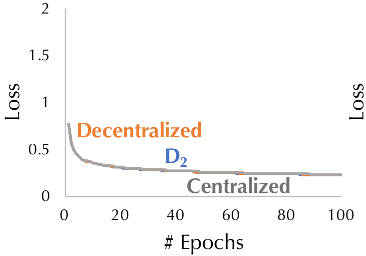
linear convergence rate bottleneck



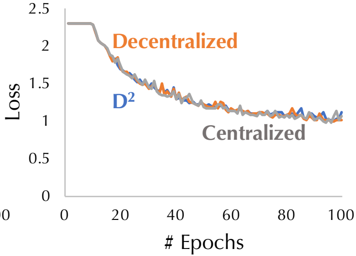
nonsmooth functions



stochastic case: shuffled

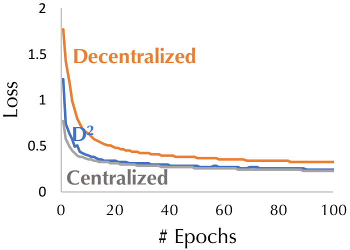


(a) TransferLearning

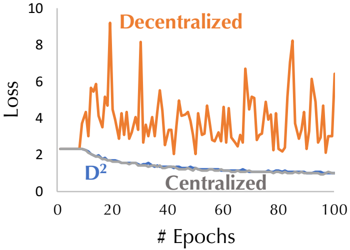


(b) LeNet

stochastic case: unshuffled



(a) TransferLearning



(b) LeNet

conclusion and open questions

conclusion

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA
- new algorithm NIDS

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA
- new algorithm NIDS

open questions

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA
- new algorithm NIDS

open questions

- network construction

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA
- new algorithm NIDS

open questions

- network construction
- preconditioning

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA
- new algorithm NIDS

open questions

- network construction
- preconditioning
- acceleration?

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA
- new algorithm NIDS

open questions

- network construction
- preconditioning
- acceleration?
- directed network?

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA
- new algorithm NIDS

open questions

- network construction
- preconditioning
- acceleration?
- directed network?
- dynamical network?

conclusion and open questions

conclusion

- optimal bounds for EXTRA/PG-EXTRA
- new algorithm NIDS

open questions

- network construction
- preconditioning
- acceleration?
- directed network?
- dynamical network?
- asynchronous?

Paper 1 Z. Li, W. Shi and M. Yan, A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates, arXiv:1704.07807

Code `https://github.com/mingyan08/NIDS`

Paper 2 Z. Li and M. Yan, A primal-dual algorithm with optimal stepsizes and its application in decentralized consensus optimization, arXiv:1711.06785

Paper 3 H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, D^2: decentralized training over decentralized data, ICML 2018, 4848-4856. `http://proceedings.mlr.press/v80/tang18a.html`

Thank You!