ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 3-2: Decentralized Optimization for Learning

Jia (Kevin) Liu

Assistant Professor
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

Spring 2022

Outline

In this lecture:

- Key Idea of Decentralized Nonconvex Optimization for Learning
- Representative Techniques
- Convergence Results

Revisit the Distributed/Federated Learning Problem

- Consider the problem:

$$\min_{\mathbf{x}\in\mathbb{R}^m} f(\mathbf{x}) \triangleq \min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{x}),$$

where $f_i(\mathbf{x}) \triangleq \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F_i(\mathbf{x},\xi_i)]$ is nonconvex

- Distributed/Federated Learning: The "summation" in the mini-batched SGD, which implies a decomposable and distributed implementation:
  - Each stochastic gradient $\nabla F(\mathbf{x}_k, \xi_i)$ can be computed by a "worker/client" i
  - B_k workers can compute such stochastic gradients in parallel
  - A server collects the stochastic gradients returned by workers and aggregate

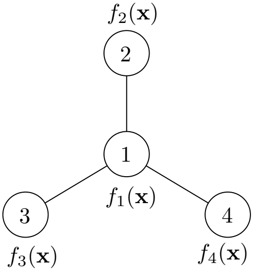
But what if we don't have a server?

Reasons for "Not Having a Server" in Distributed Learning

- Networks Having No Infrastructure
  - Networking protocols based on random access (CSMA, ALOHA, etc.)
  - Ad hoc sensor networks for environmental monitoring
  - Multi-agent systems (autonomous driving, UAVs/UGVs, robotics, etc.)
  - Autonomous swarms on battle field
  - In-situ disaster recovery
- Security/Robustness/Privacy Concerns
  - Avoid single point of failure
  - Avoid having a single target under cyber-attacks
  - Avoid communication/networking bottleneck
  - Need for information privacy preservation
  - Need for decentralization to avoid being controlled by a single party
- Economics Motivations
  - Competition/collaboration among entities
  - Build trust between multiple parties
  - Fairness guarantees
  - Promote personalization and diversity...

Decentralization Optimization for Learning: The Setup

- A network represented by a connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, with $|\mathcal{N}| = N$, $|\mathcal{L}| = L$
- $\mathbf{x} \in \mathbb{R}^d$: a global learning model
- Each node/agent i can only evaluate a local objective function $f_i(\mathbf{x}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[F_i(\mathbf{x}, \xi_i)]$
- Global objective function is: $\frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x})$
- Goal: To learn the global model collaboratively in a decentralized fashion (i.e., w/o needing any server)



Example: Decentralized Learning in Multi-Agent Systems

- A multi-agent system (drones, robots, soldiers, etc.). Each agent collects high-resolution images $\{\mathbf{u}_{ij}, \mathbf{v}_{ij}, \theta_{ij}\}_{j=1}^{N_i}$
- $\mathbf{u}_{ij}$, $\mathbf{v}_{ij}$, θ_{ij}: pixels, geographical information, ground-truth label of the j-th image at agent i.

- Agents collaboratively perform image regression based on linear model with parameters $\mathbf{x} = [\mathbf{x}_1^\top \ \mathbf{x}_2^\top]^\top$
- This problem can be written as: $\min_{\mathbf{x}} f(\mathbf{x}) \triangleq \min_{\mathbf{x}} \sum_{i=1}^{N} f_i(\mathbf{x})$, where $f_i(\mathbf{x}) \triangleq \frac{1}{N_i} \sum_{j=1}^{N_i} (\theta_{ij} - \mathbf{u}_{ij}^\top \mathbf{x}_1 - \mathbf{v}_{ij}^\top \mathbf{x}_2)^2$

Consensus Reformulation: The First Step

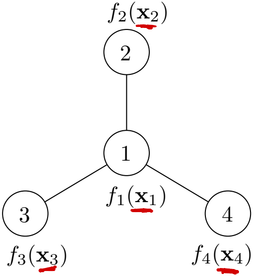
- Goal: To solve the following optimization problem distributively & collaboratively

$$\min_{x \in \mathbb{R}^d} f(\mathbf{x}) = \min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x})$$

- Clearly, this problem can be rewritten in a consensus form:

Local

$$\min_{\mathbf{x}_i \in \mathbb{R}^d, \forall i} \left\{ \frac{1}{N} \sum_{i=1}^{N} f_i(\underline{\mathbf{x}_i}) \,\middle|\, \underline{\mathbf{x}_i = \mathbf{x}_j}, \forall (i,j) \in \mathcal{L} \right\}$$



The consensus reformulation shares the same spirit with distributed/federated learning that each node maintains a local copy of the global model

Recall What We Did When We Have a Server

- In distributed/federated learning: Each node/client i computes

$$\mathbf{x}_{i,k+1} = \bar{\mathbf{x}}_k - s_k \mathbf{g}_{i,k}$$

where $\bar{\mathbf{x}}_k \triangleq \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{i,k}$ is the node/client average in iteration k

- This prompts the following natural idea for decentralized learning

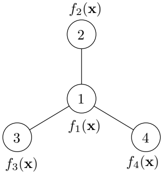
$$\mathbf{x}_{i,k+1} = \text{“Some approximation of } \bar{\mathbf{x}}_k\text{”} - s_k \mathbf{g}_{i,k}$$

- This idea turns out to the foundation of decentralized consensus optimization
  - Note: This is an insight in hindsight. Decentralized consensus optimization traces its roots to the seminal work [Tsitsiklis, Ph.D. Thesis@MIT, 1984]!

A Decentralized Method for Computing Average

Consider a consensus matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ that satisfies:

- Doubly stochastic: $\sum_{i=1}^{N} [\mathbf{W}]_{ij} = \sum_{j=1}^{N} [\mathbf{W}]_{ij} = 1$.
- Sparsity pattern defined by network topology: $[\mathbf{W}]_{ij} > 0$ for $\forall\ (i,j) \in \mathcal{L}$ and $[\mathbf{W}]_{ij} = 0$ otherwise
- Symmetric and hence real eigenvalues in $(-1, 1]$ (thus can be sorted). Moreover, easy to see that $\lambda_{\max} = 1$ with corresponding eigenvector $\mathbf{1}_N$.
- W.l.o.g., denote eigenvalues as $-1 < \lambda_N \leq \cdots \leq \lambda_1 = 1$. Let $\beta \triangleq \max\{|\lambda_2|, |\lambda_N|\}$ (i.e., 2nd-largest eigenvalue in magnitude).



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

A Decentralized Method for Computing Average

1. $k = 0$. Each node has initial value $\mathbf{x}_{i,0}$ to be averaged with other nodes

2. In k-th iteration: Each node shares its local copy to its neighbors.

3. Upon reception of all local copies from its neighbors, each node performs the local updates:

$$\mathbf{x}_{i,k+1} = \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{x}_{j,k},$$

   where $\mathcal{N}_i \triangleq \{j \in \mathcal{N} : (i,j) \in \mathcal{L}\}$.

4. Let $k \leftarrow k+1$ and go to Step 2

A Decentralized Method for Computing Average

- Define a stacked matrix of all local copies:

$$\mathbf{X}_k \triangleq \left[\begin{array}{cccc} \mathbf{x}_{1,k} & \mathbf{x}_{2,k} & \cdots & \mathbf{x}_{N,k} \end{array} \right] \in \mathbb{R}^{d \times N}.$$

$\mathbb{R}^d$

- Then the algorithm in the previous slide can be compactly written as

$$\mathbf{X}_{k+1} = \mathbf{X}_k \mathbf{W}, \qquad X_{k+1}^T = W X_k^T$$

(i.e., $\mathbf{X}_k = \mathbf{X}_0 \mathbf{W}^k$). Similar to a discrete-time finite-state Markov chain.

Perron-Frobenius Thm

- Fact: The stationary distribution of an irreducible aperiodic finite-state Markov chain is uniform iff its transition matrix is doubly stochastic.

- Convergence rate of "averaging": Let $\mathbf{W}^\infty = \lim_{k\to\infty} \mathbf{W}^k$. Then, we have $\mathbf{W}^\infty = \frac{1}{N}\mathbf{1}_N \mathbf{1}_N^\top$. Further, it holds that

$$\begin{bmatrix} \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & \ddots & \vdots \\ \frac{1}{N} & \cdots & \frac{1}{N} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \left\| \mathbf{W}^\infty \mathbf{e}_i - \mathbf{W}^k \mathbf{e}_i \right\|^2 \le \beta^{2k}, \quad \forall i \in \{1, \ldots, N\}, k \in \mathbb{N}. \qquad \frac{\lambda_2}{\lambda_1}$$

i-th basis vector in $\mathbb{R}^d$.

WTS: $\| \underbrace{\underline{W}^{\infty}}_{=\frac{1}{N}\mathbb{1}_N} \underline{e}_i - \underline{W}^k \underline{e}_i \|^2 \leq \beta^{2k}$ ← Lemma.

Proof. $\| \underline{W}^{\infty} \underline{e}_i - \underline{W}^k \underline{e}_i \|^2 = \| (\underline{W}^{\infty} - \underline{W}^k) \underline{e}_i \|^2$

$$\leq \underbrace{\| \underline{W}^{\infty} - \underline{W}^k \|^2}_{\text{induced norm}} \cdot \| \underline{e}_i \|^2 = \| \underline{W}^{\infty} - \underline{W}^k \|^2 \qquad (1).$$

Note that $\underline{W}$ is symmetric $\Rightarrow$ It has real eigenvalues.

$\underline{W} = \underline{U} \underline{\Lambda} \underline{U}^T$, where $\underline{\Lambda} = \begin{bmatrix} \lambda_1 = 1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}$, $\underline{U} = \begin{bmatrix} | & | & & | \\ \underline{u}_1 & \underline{u}_2 & \cdots & \underline{u}_N \\ | & | & & | \end{bmatrix}$

unitary. $\underline{U}^T \underline{U} = \underline{U} \underline{U}^T = \underline{I}$

So, $\underline{W}^k = \underbrace{\underline{U} \underline{\Lambda} \cancel{\underline{U}^T} \cdot \cancel{\underline{U}} \underline{\Lambda} \underline{U}^T \cdots \cancel{\underline{U}} \underline{\Lambda} \underline{U}^T}_{k \text{ terms}} = \underline{U} \underline{\Lambda}^k \underline{U}^T$, $\quad \underline{\Lambda}^k = \begin{bmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_N^k \end{bmatrix}$

Also, $\underline{W}^{\infty} = \frac{1}{N} \mathbb{1} \mathbb{1}_N^T$. clearly, it has one eigenvalue 1 and eigenvector $\mathbb{1}_N$. $\underline{W}^{\infty} = \underline{U}^T \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \underline{U}^T \qquad \sum_{i=1}^{N} \lambda_i \underline{u}_i \underline{u}_i^T$

$$(1) = \left\| \underline{U} \left(\begin{bmatrix} 1 & & \\ & 0 & \\ & & \ddots \\ & & & 0 \end{bmatrix} - \begin{bmatrix} 1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_N \end{bmatrix} \right) \underline{U}^T \right\|^2 = \left\| \sum_{i=2}^{N} \lambda_i \underline{u}_i \underline{u}_i^T \right\|^2$$

$$\leq \beta^{2k} \left\| \sum_{i=1}^{N} \underline{u}_i \underline{u}_i^T \right\|^2 = \beta^{2k} \| \underbrace{\underline{U} \underline{U}^T}_{=\underline{I}} \|^2 = \beta^{2k}. \qquad \blacksquare$$

replace $\lambda_i, \forall i$ by β, factor it out, adding $\beta \underline{u}_1 \underline{u}_1^T$

Decentralized Stochastic Gradient Descent (DSGD)

The DSGD algorithm [Nedic and Ozdaglar, TAC'09]:

DGD
PGD P-SGD } DSGD.

1. Initialization: Let $k = 1$. Choose initial values for $\mathbf{x}_{i,1}$ and step-size α_1.

2. In k-th iteration: Each node sends its local copy to its neighbors.

3. Upon reception of all local copies from its neighbors, each node updates its local copy:

"some approx. $\bar{x}_k$" run this mult. steps.

$$\mathbf{x}_{i,k+1} = \underbrace{\sum\nolimits_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{x}_{j,k}}_{\text{Avg consensus step}} - \underbrace{s_k \nabla F_i(\mathbf{x}_{i,k}, \xi_{i,k})}_{\text{Local SGD step}},$$

where $\mathcal{N}_i \triangleq \{ j \in \mathcal{N} : (i,j) \in \mathcal{L} \}$.

DGD^t

can be done t "rounds".

4. Let $k \leftarrow k + 1$ and go to Step 2

Convergence Results of DSGD

Assumptions:

- $f_i(\cdot)$, $\forall i$ are L-smooth
- Unbiased stochastic gradients: $\mathbb{E}_{\xi_{i,k}\sim\mathcal{D}_i}[\nabla F_i(\mathbf{x}_{i,k},\xi_{i,k})] = \nabla f_i(\mathbf{x}_{i,k})$, $\forall i,k$
- Bounded local stochastic gradient variance:

$$\mathbb{E}[\|\nabla F_i(\mathbf{x},\xi) - \nabla f_i(\mathbf{x})\|^2] \le \sigma^2, \quad \forall i, \mathbf{x}$$

- Bounded gradient dissimilarity: means $\nabla f_i(x)$ still follows D_i

$$\mathbb{E}_{i\sim\mathcal{U}([n])}[\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2] \le \zeta^2, \quad \forall \mathbf{x}$$

- Start from $\mathbf{0}$: $\mathbf{X}_0 = \mathbf{0}$ (not necessary, but simplifies the proof w.l.o.g.)

Convergence Results of DSGD

- Let $s_k = s$, $\forall k$, and define two constants:

$$D_1 := \left(\frac{1}{2} - \frac{9s^2L^2N}{(1-\beta)^2 D_2}\right), \text{ and } D_2 := \left(1 - \frac{18s^2}{(1-\beta)^2} N L^2\right)$$

Theorem 1 ([Lian et al. NeurIPS'17])

$$\begin{bmatrix} \nabla f_1(x_{1,k}) & \cdots & \nabla f_N(x_{N,k}) \end{bmatrix}_{d\times N}$$

Under the stated assumptions, the following convergence rate holds for DSGD:

$$\|\nabla f(\bar{x}_k)\|^2$$

$$\frac{1}{K}\left(\frac{1-sL}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\partial f(\mathbf{X}_k)\mathbf{1}_N}{N}\right\|^2\right] + D_1 \sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\frac{\mathbf{X}_k\mathbf{1}_N}{N}\right)\right\|^2\right]\right)$$

$$\leq \underbrace{\frac{f(\mathbf{0}) - f^*}{sK} + \frac{sL}{2N}\sigma^2 + \frac{s^2L^2N\sigma^2}{(1-\beta^2)D_2} + \frac{9s^2L^2N\zeta^2}{(1-\beta)^2 D_2}}$$

$$\begin{bmatrix} x_{1,k} & \cdots & x_{N,k} \end{bmatrix}_{d\times N}$$

$= O(\frac{1}{K})$ to some error ball.

$$\bar{x}_k \triangleq \frac{1}{N}\sum_{i=1}^{N} x_{i,k}$$

Convergence Results of DSGD

Corollary 2 ([Lian et al. NeurIPS'17])

Under the same assumptions as in Theorem 1, if $s = \frac{1}{2L+\sigma\sqrt{K/N}}$, *then DSGD achieves the following convergence rate:*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f\left(\overset{\bar{x}_k}{\frac{\mathbf{X}_k\mathbf{1}_N}{N}}\right)\right\|^2\right] \le \frac{8(f(\mathbf{0})-f^*)}{K} + \frac{(8f(\mathbf{0})-8f^*+4L)}{\sqrt{KN}}.$$

Remark 1

If K *is sufficiently large such that*

(Δ)

$$K \ge \frac{4L^4N^5}{\sigma^2(f(\mathbf{0})-f^*+L)^2}\left(\frac{\sigma^2}{1-\beta^2}+\frac{9\zeta^2}{(1-\beta)^2}\right) \text{ and } K \ge \frac{72L^2N^2}{\sigma^2(1-\beta)^2},$$

then the convergence rate of DSGD is $O\left(\frac{1}{K}+\frac{1}{\sqrt{NK}}\right)$. ← linear speedup

Convergence Results of DSGD

Theorem 3 ([Lian et al. NeurIPS'17])

With $s = \frac{1}{2L+\sigma\sqrt{K/N}}$ *and under the same assumptions in Theorem 1, it holds that*

$$\frac{1}{KN}\mathbb{E}\left[\sum_{k=0}^{K-1}\sum_{i=1}^{N}\left\|\frac{\sum_{i'=1}^{N}\mathbf{x}_{i',k}}{N}-\mathbf{x}_{i,k}\right\|^2\right] \leq Ns^2\frac{A}{D_2},$$

where the constant A *is defined as:*

$$A := \frac{2\sigma^2}{1-\beta^2}+\frac{18\zeta^2}{(1-\beta)^2}+\frac{L^2}{D_1}\left(\frac{\sigma^2}{1-\beta^2}+\frac{9\zeta^2}{(1-\beta)^2}\right) + \frac{18}{(1-\beta)^2}\left(\frac{f(\mathbf{0})-f^*}{sK}+\frac{sL\sigma^2}{2ND_1}\right).$$

Remark 2

The local copies achieve consensus at the rate $O(1/K)$

Preparation:

$$\underline{X}_k \triangleq \begin{bmatrix} \vdots & & \vdots \\ x_{1,k} & \cdots & x_{N,k} \\ \vdots & & \vdots \end{bmatrix}_{d\times N}, \quad \underline{W} \triangleq \begin{bmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & & \\ w_{N1} & \cdots & w_{NN} \end{bmatrix}_{N\times N}, \quad \partial \underline{F}(\underline{X}_k, \underline{\xi}_k) \triangleq \begin{bmatrix} \vdots & & \vdots \\ \nabla F_1(x_{1,k},\xi_{1,k}) & \cdots & \nabla F_N(x_{N,k},\xi_{N,k}) \\ \vdots & & \vdots \end{bmatrix}_{d\times N}$$

$$\underline{\xi}_k = \begin{bmatrix} \xi_{1,k} \\ \vdots \\ \xi_{N,k} \end{bmatrix}$$

Recall: $x_{i,k+1} = \sum_{\substack{j=1 \\ j\in N(i)}}^{N} [\underline{W}]_{ij}\, x_{j,k} - s\nabla F_i(x_{i,k}, \xi_{i,k})$, $\forall i$

Concatenating $x_{i,k+1}$, $\forall i$, we have: $\quad \underline{X}_{k+1} = \underline{X}_k \underline{W}$

$$\begin{bmatrix} \vdots & & \vdots \\ x_{1,k+1} & \cdots & x_{N,k+1} \\ \vdots & & \vdots \end{bmatrix} = \begin{bmatrix} \vdots & & \vdots \\ x_{1,k} & \cdots & x_{N,k} \\ \vdots & & \vdots \end{bmatrix} \begin{bmatrix} w_{11} & \cdots & w_{1N} \\ \vdots & & \\ w_{N1} & \cdots & w_{NN} \end{bmatrix} - s \begin{bmatrix} \vdots & & \vdots \\ \nabla F_1(x_{1,k},\xi_{1,k}) & \cdots & \nabla F_N(x_{N,k},\xi_{N,k}) \\ \vdots & & \vdots \end{bmatrix}$$

In matrix form: $\underline{X}_{k+1} = \underline{X}_k \underline{W} - s\,\partial \underline{F}(\underline{X}_k, \underline{\xi}_k)$.

$$\Rightarrow \frac{1}{N}\underline{X}_{k+1}\underline{1}_N = \frac{1}{N}\underline{X}_k \underbrace{\underline{W}\,\underline{1}_N}_{\underline{1}_N} - \frac{s}{N}\partial \underline{F}(\underline{X}_k, \underline{\xi}_k)\,\underline{1}_N$$

$$\Rightarrow \bar{x}_{k+1} = \bar{x}_k - \frac{s}{N}\sum_{i=1}^{N} \nabla F_i(x_{i,k}, \xi_{i,k}).$$

Proof of Thm 1: From descent lemma,

$$\mathbb{E}[f(\bar{x}_{k+1})] = \mathbb{E}\Big[f\Big(\bar{x}_k - \overbrace{\frac{s}{N}\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k})}^{\text{"G"}}\Big)\Big]$$

$$\leq \mathbb{E}[f(\bar{x}_k)] - \frac{s}{N}\mathbb{E}\Big[\nabla f(\bar{x}_k)^T \sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k})\Big] + \underline{\frac{s^2 L}{2}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k})\Big\|^2\Big]}$$

Consider the quad term:

$$\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k})\Big\|^2\Big] \overset{+\&-}{=} \mathbb{E}\Big[\Big\|\frac{1}{N}\Big(\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k}) - \sum_{i=1}^{N}\nabla f_i(x_{i,k})\Big) + \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\Big\|^2\Big]$$

$$= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\right\|^2\right] + \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\right\|^2\right]$$

$$\Rightarrow \mathbb{E}\left[f(\bar{x}_{k+1})\right] \le \mathbb{E}\left[f(\bar{x}_k)\right] - \frac{s}{N}\mathbb{E}\left[\nabla f(\bar{x}_k)^T \sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k})\right] +$$

$$+ \frac{s^2L}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\right\|^2\right] + \frac{s^2L}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\right\|^2\right]$$

For 2nd last term:

$$\frac{s^2L}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\right\|^2\right]$$

$$= \frac{s^2L}{2N^2}\sum_{i=1}^{N}\mathbb{E}\Big[\underbrace{\left\|\nabla F_i(x_{i,k},\xi_{i,k}) - \nabla f_i(x_{i,k})\right\|^2}_{\le \sigma^2}\Big]$$ (unbiasedness).

$$\le \frac{s^2L}{2N}\sigma^2$$

$$\mathbb{E}\left[\nabla f(\bar{x}_k)^T \cdot \frac{1}{N}\sum_{i=1}^{N}\nabla f(x_{i,k})\right] \quad \left(\text{iter. law of } \mathbb{E}[\cdot]\right).$$

$$\mathbb{E}\left[f(\bar{x}_{k+1})\right] \le \mathbb{E}\left[f(\bar{x}_k)\right] - s\,\mathbb{E}\left[\nabla f(\bar{x}_k)^T \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k})\right] + \frac{s^2L\sigma^2}{2N}$$

$$+ \frac{s^2L}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\right\|^2\right]$$

$$a^Tb = \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 - \frac{1}{2}\|a-b\|^2$$

$$= \mathbb{E}\left[f(\bar{x}_k)\right] - \frac{s - s^2L}{2}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\right\|^2\right] - \frac{s}{2}\mathbb{E}\left[\left\|\cancel{\sum_{i=1}^{N}\nabla F_i(x_{i,k},\xi_{i,k})}\ \nabla f(\bar{x}_k)\right\|^2\right]$$

$$+ \frac{s^2L\sigma^2}{2N} + \frac{s}{2}\underbrace{\mathbb{E}\left[\left\|\left[\frac{1}{N}\sum_{i=1}^{N}\cancel{\nabla F_i(x_{i,k},\xi_{i,k})}\ \nabla f(x_{i,k})\right] - \nabla f(\bar{x}_k)\right\|^2\right]}_{T_1}$$

Now, we bound T_1

$$\mathbb{E}\left[\left\|\left[\frac{1}{N}\sum_{i=1}^{N} \overset{\nabla f(x_{i,k})}{\cancel{\nabla F_i(x_{i,k},\xi_{i,k})}}\right]-\nabla f(\bar{x}_k)\right\|^2\right]$$

$$=\frac{1}{N^2}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(\nabla f(\bar{x}_k)-\overset{\nabla f(x_{i,k})}{\cancel{\nabla F_i(x_{i,k},\xi_{i,k})}}\right)\right\|^2\right]$$

$$\left(\begin{array}{l}\mathbb{E}\left[\|z_1+\dots+z_n\|^2\right] \quad (*) \\ \le n\,\mathbb{E}\left[\|z_1\|^2+\dots+\|z_n\|^2\right] \\ \text{w/ } n=N.\end{array}\right)$$

$$\le \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla f(\bar{x}_k)-\overset{\nabla f(x_{i,k})}{\cancel{\nabla F_i(x_{i,k},\xi_{i,k})}}\right\|^2\right]$$

$$\cancel{\le}\,= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla f(\bar{x}_k)-\nabla f_i(x_{i,k})+\cancel{\nabla f_i(x_{i,k})-\nabla F_i(x_{i,k},\xi_{i,k})}\right\|^2\right]$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left[\underbrace{\mathbb{E}\left[\left\|\nabla f(\bar{x}_k)-\nabla f_i(x_{i,k})\right\|^2\right]}_{\le L^2\|\bar{x}_k-x_{i,k}\|^2}+\cancel{\underbrace{\mathbb{E}\left[\left\|\nabla f_i(x_{i,k})-\nabla F_i(x_{i,k},\xi_{i,k})\right\|^2\right]}_{\le\sigma^2}}\right]$$

$$\le \frac{L^2}{N}\underbrace{\sum_{i=1}^{N}\mathbb{E}\left[\left\|\bar{x}_k-x_{i,k}\right\|^2\right]}_{\text{agent-drift.}}$$

To bound the "agent-drift": $\mathbb{E}\left[\|\bar{x}_k-x_{i,k}\|^2\right]\triangleq Q_{i,k}$

$$Q_{i,k}=\mathbb{E}\left[\|\bar{x}_k-x_{i,k}\|^2\right]=\mathbb{E}\left[\left\|\frac{1}{N}\mathbf{X}_k\mathbf{1}_N-\mathbf{X}_k\mathbf{e}_i\right\|^2\right]$$

$$\overset{\text{def}}{=}\mathbb{E}\left[\left\|\frac{1}{N}\left(\mathbf{X}_{k-1}\underbrace{\mathbf{W}\mathbf{1}_N}_{=\mathbf{1}_N}-s\,\partial\mathbf{F}(\mathbf{X}_{k-1},\xi_{k-1})\mathbf{1}_N\right)-\left(\mathbf{X}_{k-1}\mathbf{W}-s\,\partial\mathbf{F}(\mathbf{X}_{k-1},\xi_{k-1})\right)\mathbf{e}_i\right\|^2\right]$$

$$=\mathbb{E}\left[\left\|\frac{1}{N}\left(\mathbf{X}_{k-1}\mathbf{1}_N-s\,\partial\mathbf{F}(\mathbf{X}_{k-1},\xi_{k-1})\mathbf{1}_N\right)-\left(\mathbf{X}_{k-1}\mathbf{W}\mathbf{e}_i-s\,\partial\mathbf{F}(\mathbf{X}_{k-1},\xi_{k-1})\mathbf{e}_i\right)\right\|^2\right]$$

recursion

$$=\mathbb{E}\left[\left\|\frac{1}{N}\left(\overset{=0}{\cancel{\mathbf{X}_0\mathbf{1}_N}}-s\sum_{i=0}^{k-1}\partial\mathbf{F}(\mathbf{X}_i,\xi_i)\mathbf{1}_N\right)-\left(\overset{=0}{\cancel{\mathbf{X}_0\mathbf{W}^k\mathbf{e}_i}}-s\sum_{j=0}^{k-1}\partial\mathbf{F}(\mathbf{X}_j,\xi_j)\mathbf{W}^{k-j-1}\mathbf{e}_i\right)\right\|^2\right]$$

$$= \mathbb{E}\left[\left\| \sum_{j=0}^{k-1} s\,\partial \underline{\underline{F}}(\underline{\underline{X}}_j, \xi_j)\left(\frac{1}{N}\underline{1}_N - \underline{\underline{W}}^{k-j-1}\underline{e}_i\right)\right\|^2\right]$$

$$= s^2\,\mathbb{E}\left[\left\| \sum_{j=0}^{k-1} \partial \underline{\underline{F}}(\underline{\underline{X}}_j, \xi_j)\left(\frac{1}{N}\underline{1}_N - \underline{\underline{W}}^{k-j-1}\underline{e}_i\right)\right\|^2\right]$$

$$\partial f(\underline{\underline{X}}_j) = \begin{bmatrix} \vdots & & \vdots \\ \nabla f_1(\underline{x}_{1,j}) & \cdots & \nabla f_N(\underline{x}_{N,j}) \\ \vdots & & \vdots \end{bmatrix}_{d\times N}$$

$$\overset{+\&-}{=} s^2\,\mathbb{E}\left[\left\| \sum_{j=0}^{k-1} \left(\partial \underline{\underline{F}}(\underline{\underline{X}}_j, \xi_j) - \partial f(\underline{\underline{X}}_j) + \partial f(\underline{\underline{X}}_j)\right)\left(\frac{1}{N}\underline{1}_N - \underline{\underline{W}}^{k-j-1}\underline{e}_i\right)\right\|^2\right]$$

(use (*) w/ $n=2$)

$$\le \underbrace{2s^2\,\mathbb{E}\left[\left\| \sum_{j=0}^{k-1} \left(\partial \underline{\underline{F}}(\underline{\underline{X}}_j, \xi_j) - \partial f(\underline{\underline{X}}_j)\right)\left(\frac{1}{N}\underline{1}_N - \underline{\underline{W}}^{k-j-1}\underline{e}_i\right)\right\|^2\right]}_{T_2}$$

$$+ \underbrace{2s^2\,\mathbb{E}\left[\left\| \sum_{j=0}^{k-1} \partial f(\underline{\underline{X}}_j)\left(\frac{1}{N}\underline{1}_N - \underline{\underline{W}}^{k-j-1}\underline{e}_i\right)\right\|^2\right]}_{T_3}$$

Now, we bound T_2:

$$T_2 = \mathbb{E}\left[\left\| \sum_{j=0}^{k-1} \left(\partial \underline{\underline{F}}(\underline{\underline{X}}_j, \xi_j) - \partial f(\underline{\underline{X}}_j)\right)\left(\frac{1}{N}\underline{1}_N - \underline{\underline{W}}^{k-j-1}\underline{e}_i\right)\right\|^2\right]$$

$$\overset{\text{unbiasedness, indep.}}{=} \sum_{j=0}^{k-1} \mathbb{E}\left[\left\| \left(\partial \underline{\underline{F}}(\underline{\underline{X}}_j, \xi_j) - \partial f(\underline{\underline{X}}_j)\right)\left(\frac{1}{N}\underline{1}_N - \underline{\underline{W}}^{k-j-1}\underline{e}_i\right)\right\|^2\right]$$

$$\overset{\text{cachy-schwartz}}{\le} \sum_{j=0}^{k-1} \mathbb{E}\left[\underbrace{\left\| \partial \underline{\underline{F}}(\underline{\underline{X}}_j, \xi_j) - \partial f(\underline{\underline{X}}_j)\right\|_2^2}_{l_2\text{-induced norm} \le \|\cdot\|_F = \sum_{i=1}^{N}\|\nabla F_i(\underline{X}_i, \xi_i) - \nabla f_i(\underline{X}_i)\|^2 \le N\sigma^2}\left\|\frac{1}{N}\underline{1}_N - \underline{\underline{W}}^{k-j-1}\underline{e}_i\right\|^2\right]$$

$\underline{A} \in \mathbb{R}^{m\times n}$ (l_p-induced norm):

$$\|\underline{A}\|_p = \sup_{\underline{x}\ne \underline{0}} \frac{\|\underline{A}\underline{x}\|_p}{\|\underline{x}\|_p}$$

$\|\underline{A}\|_2 = \sqrt{\lambda_{max}(\underline{A}^T\underline{A})} = \sigma_{max}(\underline{A})$.

$\|\underline{A}\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|$ (max abs. col. sum).

$\|\underline{A}\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|$ (max abs. row sum).

"entry-wise" $\|\underline{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\mathrm{Tr}(\underline{A}^T\underline{A})} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(\underline{A})}$

$$\|\underline{A}\|_2 = \sigma_{max}(\underline{A}) \le \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(\underline{A})} = \|\underline{A}\|_F$$

(Continue to bound T_2).

$$T_2 \le \sum_{j=1}^{k-1} \mathbb{E}\left[\underbrace{\|\partial \underline{F}(\underline{X}_j;\underline{\xi}_j) - \partial f(\underline{X}_j)\|_F^2}_{\le N\sigma^2} \underbrace{\|\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\|^2}_{\le \beta^{2(k-j-1)}}\right]$$

$$\le N\sigma^2 \underbrace{\sum_{j=0}^{k-1}}_{k \text{ terms}} \beta^{2(k-j-1)} \le \frac{N\sigma^2}{1-\beta^2}$$

Now, we bound T_3:

$$T_3 = \mathbb{E}\left[\left\|\sum_{j=0}^{k-1} \partial f(\underline{X}_j)\left(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\right)\right\|^2\right]$$

$$\overset{\text{expand}}{=} \underbrace{\sum_{j=0}^{k-1} \mathbb{E}\left[\|\partial f(\underline{X}_j)(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i)\|^2\right]}_{T_4}$$

$$+ \underbrace{\sum_{j=0}^{k-1}\sum_{j'=0\ne j}^{k-1} \mathbb{E}\left[\left\langle \partial f(\underline{X}_j)(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i), \partial f(\underline{X}_{j'})(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j'-1}\underline{e}_i)\right\rangle\right]}_{T_5}$$

To bound T_3, we need to further bound T_4 & T_5.

$$T_4 = \sum_{j=0}^{k-1} \mathbb{E}\left[\|\partial f(\underline{X}_j)(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i)\|^2\right]$$

$$\overset{\text{Cachy-Schwatz}}{\le} \sum_{j=0}^{k-1} \underbrace{\mathbb{E}\left[\|\partial f(\underline{X}_j)\|^2\right]}_{\text{introduce a Lemma}} \|\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\|^2 \qquad (2).$$

Lemma 1: $\mathbb{E}\left[\|\partial f(\underline{X}_j)\|^2\right] \le \sum_{h=1}^{N} 3\,\mathbb{E}\left[L^2 \overbrace{\|\bar{x}_j - x_{h,j}\|^2}^{Q_{j,h}}\right] + 3N\varsigma^2 + 3\,\mathbb{E}\left[\|\nabla f(\bar{x}_j)\,\underline{1}_N^T\|^2\right]$

Proof: $\mathbb{E}\left[\|\partial f(\underline{X}_j)\|^2\right] \overset{+\&-}{=} \mathbb{E}\Big[\|\partial f(\underline{X}_j) - \partial f(\bar{x}_j \underline{1}_N^T) + \cancel{\partial f(\bar{x}_j \underline{1}_N^T)}$

$- \nabla f(\bar{x}_j)\underline{1}_N^T + \nabla f(\bar{x}_j)\underline{1}_N^T\|^2\Big]$

$\partial f(\underline{X}_j) = [\nabla f_1(x_{1,j}) \cdots \nabla f_N(x_{N,j})]_{d\times N}$

$\partial f(\bar{x}_j \underline{1}_N^T) = [\nabla f_1(\bar{x}_j) \cdots \nabla f_N(\bar{x}_j)]_{d\times N}$

$\nabla f(\bar{x}_j)\underline{1}_N^T = [\nabla f(\bar{x}_j) \cdots \nabla f(\bar{x}_j)]_{d\times N}.$

$\le 3\,\mathbb{E}\left[\|\partial f(\underline{X}_j) - \partial f(\bar{x}_j \underline{1}_N^T)\|^2\right] + 3\,\mathbb{E}\left[\|\partial f(\bar{x}_j \underline{1}_N^T) - \nabla f(\bar{x}_j)\underline{1}_N^T\|^2\right] + 3\,\mathbb{E}\left[\|\nabla f(\bar{x}_j)\underline{1}_N^T\|^2\right]$

$\le 3\,\mathbb{E}\left[\|\partial f(\underline{X}_j) - \partial f(\bar{x}_j \underline{1}_N^T)\|_F^2\right] + 3\,\mathbb{E}\left[\|\partial f(\bar{x}_j \underline{1}_N^T) - \nabla f(\bar{x}_j)\underline{1}_N^T\|_F^2\right] + 3\,\mathbb{E}\left[\|\nabla f(\bar{x}_j)\underline{1}_N^T\|^2\right]$

$[\nabla f_1(x_{1,j}) - \nabla f_1(\bar{x}_j), \cdots \nabla f_N(x_{N,j}) - \nabla f_N(\bar{x}_j)]_{d\times N}$

$[\nabla f_1(\bar{x}_j) - \nabla f(\bar{x}_j), \cdots \nabla f_N(\bar{x}_j) - \nabla f(\bar{x}_j)]_{d\times N}$

$= 3\,\mathbb{E}\left[\sum_{h=1}^{N} \underbrace{\|\nabla f_h(x_{h,j}) - \nabla f_h(\bar{x}_j)\|^2}_{\le L^2\|x_{h,j} - \bar{x}_j\|^2}\right] + 3\,\mathbb{E}\left[\sum_{h=1}^{N} \underbrace{\|\nabla f_h(\bar{x}_j) - \nabla f(\bar{x}_j)\|^2}_{\le \varsigma^2}\right] + 3\,\mathbb{E}\left[\|\nabla f(\bar{x}_j)\underline{1}_N^T\|^2\right]$

non-i.i.d.

$\le \sum_{h=1}^{N} 3\,\mathbb{E}\left[L^2\|x_{h,j} - \bar{x}_j\|^2\right] + 3N\varsigma^2 + 3\,\mathbb{E}\left[\|\nabla f(\bar{x}_j)\underline{1}_N^T\|^2\right].$ ▨

(Continue on bounding T_4): Using Lemma 1 in (2):

$T_4 \le \sum_{j=0}^{k-1} \sum_{h=1}^{N} 3\,\mathbb{E}\left[L^2\|x_{h,j} - \bar{x}_j\|^2\right] + 3N\varsigma^2 + 3\,\mathbb{E}\left[\|\nabla f(\bar{x}_j)\underline{1}_N^T\|^2\right] \times$

$\underbrace{\left\|\frac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\right\|^2}_{\le \beta^{2(k-j-1)}}$

$$\leq \frac{3N\varsigma^2}{1-\beta^2} + \sum_{j=0}^{k-1} \sum_{h=1}^{N} 3\,\mathbb{E}\Big[L^2 \underbrace{\|\underline{x}_{h,j} - \bar{\underline{x}}_j\|^2}_{Q_{h,j}}\Big] \Big\|\frac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\Big\|^2$$

$$+ \sum_{j=0}^{k-1} \sum_{h=1}^{N} 3\,\mathbb{E}\Big[\|\nabla f(\bar{\underline{x}}_j)\,\underline{1}_N^T\|^2\Big] \Big\|\frac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\Big\|^2 \qquad (3).$$

Now, we bnd T_5:

$$T_5 = \sum_{j=0}^{k-1} \sum_{j'=0 \neq j}^{k-1} \mathbb{E}\Big[\Big\langle \partial f(\underline{X}_j)\big(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\big),\ \partial f(\underline{X}_{j'})\big(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j'-1}\underline{e}_i\big)\Big\rangle\Big]$$

Cachy-Schwarz

$$\leq \sum_{j=0}^{k-1} \sum_{j'=0 \neq j}^{k-1} \mathbb{E}\Big[\big\|\partial f(\underline{X}_j)\big(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\big)\big\| \cdot \big\|\partial f(\underline{X}_{j'})\big(\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j'-1}\underline{e}_i\big)\big\|\Big]$$

Cachy-Schwatz

$$\leq \sum_{j=0}^{k-1} \sum_{j'=0 \neq j}^{k-1} \mathbb{E}\Big[\|\partial f(\underline{X}_j)\| \cdot \big\|\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\big\| \cdot \|\partial f(\underline{X}_{j'})\| \cdot \big\|\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j'-1}\underline{e}_i\big\|\Big]$$

Young's

$ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$

$$\leq \sum_{j=0}^{k-1} \sum_{j'=0 \neq j}^{k-1} \mathbb{E}\Big[\tfrac{1}{2}\|\partial f(\underline{X}_j)\|^2 \cdot \underbrace{\big\|\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\big\|}_{\leq \beta^{(k-j-1)}} \cdot \underbrace{\big\|\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j'-1}\underline{e}_i\big\|}_{\leq \beta^{(k-j'-1)}}\Big]$$

$$+ \sum_{j=0}^{k-1} \sum_{j'=0 \neq j}^{k-1} \mathbb{E}\Big[\tfrac{1}{2}\|\partial f(\underline{X}_{j'})\|^2 \cdot \big\|\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j-1}\underline{e}_i\big\| \cdot \big\|\tfrac{1}{N}\underline{1}_N - \underline{W}^{k-j'-1}\underline{e}_i\big\|\Big]$$

$$\leq \underbrace{\sum_{j=0}^{k-1} \sum_{j'=0 \neq j}^{k-1}}_{k^2 \text{ terms}} \mathbb{E}\Big[\Big(\tfrac{1}{2}\|\partial f(\underline{X}_j)\|^2 + \tfrac{1}{2}\|\partial f(\underline{X}_{j'})\|^2\Big)\beta^{2(k-\frac{j+j'}{2}-1)}\Big]$$

$$= \sum_{j=0}^{k-1} \sum_{j'=0 \neq j}^{k-1} \mathbb{E}\Big[\underbrace{\|\partial f(\underline{X}_j)\|^2}_{\text{Lemma 1}} \beta^{2(k-\frac{j+j'}{2}-1)}\Big]$$

$$\leq \sum_{j=0}^{k-1}\sum_{j'=0\neq j}^{k-1}\left[\sum_{h=1}^{N} 3\mathbb{E}\left[L^2\underbrace{\|x_{h,j}-\bar{x}_j\|^2}_{Q_{h,j}}\right] + 3N\varsigma^2 + 3\mathbb{E}\left[\|\nabla f(\bar{x}_j)\mathbb{1}_N^T\|^2\right]\right] \times \beta^{2(k-\frac{j+j'}{2}-1)}$$

$$= \underbrace{\sum_{j=0}^{k-1}\sum_{j'=0\neq j}^{k-1} 3N\varsigma^2 \beta^{2(k-\frac{j+j'}{2}-1)}}_{T_6} + \underbrace{3\sum_{j=0}^{k-1}\sum_{j'=0\neq j}^{k-1}\left[\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right] + \mathbb{E}\left[\|\nabla f(\bar{x}_j)\mathbb{1}_N\|^2\right]\right] \times \beta^{2(k-\frac{j+j'}{2}-1)}}_{T_7}$$

Note T_6 can bnded as:

$$T_6 = \sum_{j=0}^{k-1}\sum_{j'=0\neq j}^{k-1} 3N\varsigma^2 \beta^{2(k-\frac{j+j'}{2}-1)} = 6N\varsigma^2\sum_{j=0}^{k-1}\sum_{j'>j}^{k-1}\beta^{2(k-\frac{j+j'}{2}-1)}$$

$$= 6N\varsigma^2\sum_{j=0}^{k-1}\beta^{k-j-1}\sum_{j'>j}^{k-1}\beta^{k-j'-1} = 6N\varsigma^2\sum_{j=0}^{k-1}\beta^{k-j-1}\left[1+\beta+\cdots+\beta^{k-j-2}\right]$$

$$= 6N\varsigma^2\sum_{j=0}^{k-1}\beta^{k-j-1}\frac{1-\beta^{k-j-1}}{1-\beta} = \frac{6N\varsigma^2}{1-\beta}\left[\underbrace{\sum_{j=0}^{k-1}\beta^{k-j-1}}_{k\text{ terms.}} - \underbrace{\sum_{j=0}^{k-1}\beta^{2(k-j-1)}}_{k\text{ terms}}\right]$$

$$= \frac{6N\varsigma^2}{1-\beta}\left[\frac{1-\beta^k}{1-\beta} - \frac{1-\beta^{2k}}{1-\beta^2}\right] = 6N\varsigma^2\frac{(\overbrace{1-\beta^k}^{\leq 1})(\overbrace{\beta-\beta^k}^{\leq 1})}{(1-\beta)^2\underbrace{(1+\beta)}_{>1}} \leq \frac{6N\varsigma^2}{(1-\beta)^2}$$

Now, we bnd T_7.

$$T_7 = 3\sum_{j=0}^{k-1}\sum_{j'=0\neq j}^{k-1}\left[\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right] + \mathbb{E}\left[\|\nabla f(\bar{x}_j)\mathbb{1}_N\|^2\right]\right]\beta^{2(k-\frac{j+j'}{2}-1)}$$

$$= 6\sum_{j=0}^{k-1}\left[\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right] + \mathbb{E}\left[\|\nabla f(\bar{x}_j)\mathbb{1}_N\|^2\right]\right]\underbrace{\sum_{j'=j+1}^{k-1}}_{k-j-1\text{ terms.}}\beta^{2(k-\frac{j+j'}{2}-1)}$$

$$\leq 6\sum_{j=0}^{k-1}\left[\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right]+\mathbb{E}\left[\left\|\nabla f(\bar{x}_j)\,\mathbb{1}_N\right\|^2\right]\right]\frac{\beta^{k-j-1}}{1-\beta}$$

Plug T_6, T_7 into T_5, and then plugging T_5 & T_4 into T_3 yields:

$$T_3 \leq 3\sum_{j=0}^{k-1}\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right]\left\|\frac{1}{N}\mathbb{1}_N - \underline{W}^{k-j-1}\underline{e}_i\right\|^2$$

$$+3\sum_{j=0}^{k-1}\sum_{h=1}^{N}\mathbb{E}\left[\left\|\nabla f(\bar{x}_j)\,\mathbb{1}_N^T\right\|^2\right]\left\|\frac{1}{N}\mathbb{1}_N - \underline{W}^{k-j-1}\underline{e}_i\right\|^2$$

$$+6\sum_{j=0}^{k-1}\left[\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right]+\mathbb{E}\left[\left\|\nabla f(\bar{x}_j)\,\mathbb{1}_N\right\|^2\right]\right]\frac{\beta^{k-j-1}}{1-\beta}$$

$$+\boxed{\frac{3N\varsigma^2}{\underbrace{1-\beta^2}_{(1-\beta)(1+\beta)\geq(1-\beta)^2}}+\frac{6N\varsigma^2}{(1-\beta)^2}} \leq \frac{9N\varsigma^2}{(1-\beta)^2}$$

Putting the bnds for T_2 & T_3 back to $Q_{i,k}$:

$$Q_{i,k} \leq \underbrace{\frac{2s^2N\sigma^2}{1-\beta^2}}_{2s^2T_2} + 6s^2\sum_{j=0}^{k-1}\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right]\underbrace{\left\|\frac{1}{N}\mathbb{1}_N - \underline{W}^{k-j-1}\underline{e}_i\right\|^2}_{\leq \beta^{2(k-j-1)}}$$

$$+6s^2\sum_{j=0}^{k-1}\sum_{h=1}^{N}\mathbb{E}\left[\left\|\nabla f(\bar{x}_j)\,\mathbb{1}_N^T\right\|^2\right]\underline{\left\|\frac{1}{N}\mathbb{1}_N - \underline{W}^{k-j-1}\underline{e}_i\right\|^2}$$

$$+12s^2\sum_{j=0}^{k-1}\left[\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right]+\mathbb{E}\left[\left\|\nabla f(\bar{x}_j)\,\mathbb{1}_N\right\|^2\right]\right]\frac{\beta^{k-j-1}}{1-\beta}+\frac{18s^2N\varsigma^2}{(1-\beta)^2}$$

$$\leq \frac{2s^2N\sigma^2}{1-\beta^2}+\frac{18s^2N\varsigma^2}{(1-\beta)^2}+6s^2\sum_{j=0}^{k-1}\sum_{h=1}^{N}\mathbb{E}\left[L^2 Q_{h,j}\right]\beta^{2(k-j-1)}$$

$$+6s^2 \sum_{j=0}^{k-1} \sum_{h=1}^{N} \mathbb{E}\left[\left\|\nabla f(\bar{x}_j) \mathbb{1}_N^T\right\|^2\right] \beta^{2(k-j-1)}$$

$$+12s^2 \sum_{j=0}^{k-1} \left[\sum_{h=1}^{N} \mathbb{E}\left[L^2 Q_{h,j}\right] + \mathbb{E}\left[\left\|\nabla f(\bar{x}_j) \mathbb{1}_N\right\|^2\right]\right] \frac{\beta^{k-j-1}}{1-\beta}$$

$$= \frac{2s^2 N \sigma^2}{1-\beta^2} + \frac{18 s^2 N g^2}{(1-\beta)^2} + 6s^2 \sum_{j=0}^{k-1} \sum_{h=1}^{N} \mathbb{E}\left[L^2 Q_{h,j}\right]\left(\beta^{2(k-j-1)} + \frac{2\beta^{k-j-1}}{1-\beta}\right)$$

$$+6s^2 \sum_{j=0}^{k-1} \sum_{h=1}^{N} \mathbb{E}\left[\left\|\nabla f(\bar{x}_j) \mathbb{1}_N^T\right\|^2\right]\left(\beta^{2(k-j-1)} + \frac{2\beta^{k-j-1}}{1-\beta}\right).$$

Thus: $T_1 \leq \frac{L^2}{N} \sum_{i=1}^{N} \mathbb{E}\left[\|\bar{x}_k - x_{i,k}\|^2\right] = \frac{L^2}{N} \sum_{i=1}^{N} \mathbb{E}\left[Q_{i,k}\right]$

Recall:

$$\mathbb{E}\left[f(\bar{x}_{k+1})\right] \leq \mathbb{E}\left[f(\bar{x}_k)\right] - \frac{s-s^2L}{2} \mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{i,k})\right\|^2\right]$$

$$- \frac{s}{2} \mathbb{E}\left[\left\|\nabla f(\bar{x}_k)\right\|^2\right] + \frac{s^2 L \sigma^2}{2N} + \frac{s}{2} \mathbb{E}\left[T_1\right] \qquad (4)$$

Summing $k=0, \dots, K-1$ yields:

$$\frac{s-s^2L}{2} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_{i,k})\right\|^2\right] + \frac{s}{2} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla f(\bar{x}_k)\right\|^2\right]$$

$$\leq f(0) - f^* + \sum_{k=0}^{K-1} \frac{s^2 L \sigma^2}{2N} + \frac{s}{2} \sum_{k=0}^{K-1} \underbrace{\mathbb{E}\left[T_1\right]}_{\leq \frac{L^2}{N} \sum_{i=1}^{N} \mathbb{E}[Q_{i,k}] = L^2 \mathbb{E}[M_k]}$$

$M_k \triangleq \frac{1}{N} \sum_{i=1}^{N} Q_{i,k}$.

Now, need to bnd $\mathbb{E}[M_k]$:

$$\mathbb{E}[M_k] = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} Q_{i,k}\right]$$

$$\leq \frac{2s^2N\sigma^2}{1-\beta^2} + \frac{18s^2N\varsigma^2}{(1-\beta)^2} + 6s^2\sum_{j=0}^{k-1}\sum_{h=1}^{N}\mathbb{E}\left[\left\|\nabla f(\bar{x}_j)\mathbf{1}_N^T\right\|^2\right] \times$$

$$\left(\beta^{2(k-j-1)} + \frac{2\beta^{k-j-1}}{1-\beta}\right).$$

$$+ 6s^2NL\sum_{j=0}^{k-1}\mathbb{E}[M_j]\left(\beta^{2(k-j-1)} + \frac{2\beta^{k-j-1}}{1-\beta}\right)$$

Suming the above from $k=0$ to $K-1$, and rearranging:

$$\sum_{k=0}^{K-1}\mathbb{E}[M_k] \leq \frac{2s^2N\sigma^2}{(1-\beta^2)\underbrace{\left(1-\frac{18s^2NL^2}{(1-\beta)^2}\right)}_{D_2}}K + \frac{18s^2N\varsigma^2}{(1-\beta)^2\underbrace{\left(1-\frac{18s^2NL^2}{(1-\beta)^2}\right)}_{D_2}}K$$

$$+ \frac{18s^2N}{(1-\beta)^2\underbrace{\left(1-\frac{18s^2NL^2}{(1-\beta)^2}\right)}_{D_2}}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(\bar{x}_k)\right\|^2\right] \tag{5}$$

Plugging (5) into (4), we have:

$$\frac{s-s^2L}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_{i,k})\right\|^2\right] + \frac{s}{2}\sum_{k=1}^{K-1}\mathbb{E}\left[\left\|\nabla f(\bar{x}_k)\right\|^2\right]$$

$$\leq f(0) - f^* + \frac{s^2KL\sigma^2}{2N} + \frac{s^3L^2N\sigma^2}{(1-\beta^2)D_2} + \frac{9Ns^3L^2}{(1-\beta)^2D_2}$$

$$+ \frac{9Ns^3L^2}{(1-\beta)^2D_2}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(\bar{x}_k)\right\|^2\right]$$ and defining $D_1 = \frac{1}{2} - \frac{9sL^2N}{(1-\beta)^2D_2}$

Rearranging & dividing both sides by sK, arrives at stated result. ∎

$\left(\frac{1}{L}, \frac{2}{L}\right)$

Proof of Corollary 2: If $s = \frac{1}{2L+\sigma\sqrt{K/N}} < \frac{1}{L}$, then $\frac{1-sL}{2} > 0$

Dropping the term associated w/ $\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f(x_{i,k})\right\|^2$, we have:

$$\frac{D_1}{K}\sum_{i=1}^{K-1} \mathbb{E}\left[\|\nabla f(\bar{x}_k)\|^2\right] \le \frac{2(f(0)-f^*)L}{K} + \frac{(f(0)-f^*)\sigma}{\sqrt{KN}} + \frac{L\sigma^2}{\cancel{4NL}+2\sigma\sqrt{KN}}$$

$$+ \frac{L^2N}{(\cancel{2L}+\sigma\sqrt{K/N})^2 D_2}\left(\frac{\sigma^2}{1-\beta^2} + \frac{9\varsigma^2}{(1-\beta)^2}\right)$$

$$\le \frac{2(f(0)-f^*)L}{K} + \frac{(f(0)-f^*+L/2)\sigma}{\sqrt{KN}} + \frac{L^2N}{(\sigma\sqrt{K/N})^2 D_2}\left(\frac{\sigma^2}{1-\beta^2} + \frac{9\varsigma^2}{(1-\beta)^2}\right) \quad (6).$$

Recall $D_1 = \frac{1}{2} - \frac{9s^2L^2N}{(1-\beta)^2 D_2}$, $D_2 = 1 - \frac{18s^2}{(1-\beta)^2}NL^2$

If $s^2 \le \frac{(1-\beta)^2}{36NL^2} \Rightarrow D_2 \ge \frac{1}{2}$. $s^2 \le \frac{(1-\beta)^2}{72L^2N} \Rightarrow D_1 \ge \frac{1}{4}$. $\Bigg\} \Rightarrow$

Since $s = \frac{1}{2L+\sigma\sqrt{K/N}} \le \frac{1}{\sigma\sqrt{K/N}} \Rightarrow s^2 \le \frac{N}{\sigma^2 K}$

If $\frac{N}{\sigma^2 K} \le \min\left\{\frac{(1-\beta)^2}{36NL^2}, \frac{(1-\beta)^2}{72NL^2}\right\}$ ✓, then $D_2 \ge \frac{1}{2}$, $D_1 \ge \frac{1}{4}$.

Now, replacing D_1 & D_2 by $\frac{1}{4}$ & $\frac{1}{2}$, resp., in (6):

$$\frac{1}{\cancel{4}K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla f(\bar{x}_k)\|^2\right] \le \frac{\cancel{2}\,8(f(0)-f^*)L}{K} + \frac{4(f(0)-4f^*+L/\cancel{2}\,2)\sigma}{\sqrt{KN}}$$

$$+ \frac{2L^2N}{(\sigma\sqrt{K/N})^2}\left(\frac{\sigma^2}{1-\beta^2} + \frac{9\varsigma^2}{(1-\beta)^2}\right)$$

$$\le \frac{(4f(0)-4f^*+2L)\sigma}{\sqrt{KN}} \quad \text{if } (\Delta) \text{ is true}$$

Combine these two yields the stated result.

Proof of Thm 3: With $s = \dfrac{1}{2L + \sigma\sqrt{K/N}}$, we have from (5):

$$\frac{1}{K}\underbrace{\sum_{k=0}^{K-1} \mathbb{E}[M_k]}_{\text{avg agent drift.}} \le \frac{2s^2 N \sigma^2}{(1-\beta^2)\, D_2} + \frac{18 s^2 N \varsigma^2}{(1-\beta)^2\, D_2} + \frac{18 s^2 N}{(1-\beta)^2\, D_2\, K} \sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla f(\bar{x}_k)\|^2\right] \qquad (5)$$

Corollary 2

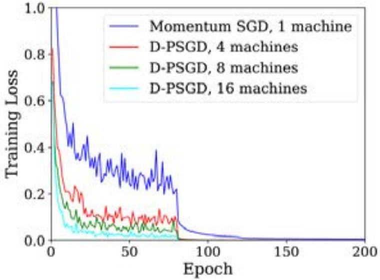
$$\le \frac{2s^2 N \sigma^2}{(1-\beta^2)\, D_2} + \frac{18 s^2 N \varsigma^2}{(1-\beta)^2\, D_2} + \frac{s^2 L^2 N}{D_1 D_2}\left(\frac{\sigma^2}{1-\beta^2} + \frac{9\varsigma^2}{(1-\beta)^2}\right) + \frac{18 s^2 N}{(1-\beta)^2 D_2}\left(\frac{f(0)-f^*}{sK} + \frac{sL\sigma^2}{2ND_1}\right)$$

$$\triangleq N s^2 \frac{A}{D_2}.$$

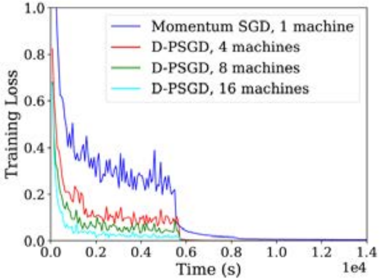
∎

Numerical Results of DSGD

- Linear Speedup Effect
  - 32-layer residual network and CIFAR-10 dataset
  - Up to 16 machines; each machine includes two Xeon E5-2680 8-core processors and a NVIDIA K20 GPU



(a) Iteration vs Training Loss



(b) Time vs Training Loss

A "Tug of War" in DSGD

Revisit the DSGD algorithm:

- The algorithmic update at each agent is:

$$\mathbf{x}_{i,k+1} = \underbrace{\sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{x}_{j,k}}_{\text{Avg consensus step}} - \underbrace{s_k \nabla F_i(\mathbf{x}_{i,k}, \xi_{i,k})}_{\text{Local SGD step}},$$

where $\mathcal{N}_i \triangleq \{j \in \mathcal{N} : (i,j) \in \mathcal{L}\}$.

The average consensus step and the local SGD step "conflict" with each other. Can we do better?

The Gradient Tracking Idea

[Lorenzo-Scutari, TSIPN'16]
"full grad"

Gradient-Tracking DSGD: [Lu et al., DSW'19]:

1. Initialization: Let $k = 1$. Choose initial values for $\mathbf{x}_{i,1}$ and step-size s_1. Define an auxiliary variable $\mathbf{y}_{i,k}$ with $\mathbf{y}_{i,1} = \nabla F_i(\mathbf{x}_{i,1}, \xi_{i,1})$.
2. In k-th iteration: Each node sends its local copy and $y_{i,k}$ to its neighbors.
3. Upon reception of all local copies and $y_{i,k}$ from its neighbors, each node updates its local copy:

$$\mathbf{x}_{i,k+1} = \sum\nolimits_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{x}_{j,k} - s_k \mathbf{y}_{i,k},$$

tracking avg stochastic grad.

$$\mathbf{y}_{i,k+1} = \sum\nolimits_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{y}_{j,k} + \nabla F_i(\mathbf{x}_{i,k+1}, \xi_{i,k+1}) - \nabla F_i(\mathbf{x}_{i,k}, \xi_{i,k}).$$

where $\mathcal{N}_i \triangleq \{j \in \mathcal{N} : (i,j) \in \mathcal{L}\}$.

4. Let $k \leftarrow k + 1$ and go to Step 2

Notation:

$$\underline{\underline{X}}_k = \begin{bmatrix} | & & | \\ x_{1,k} & \cdots & x_{N,k} \\ | & & | \end{bmatrix}_{d\times N}, \quad \underline{\underline{Y}}_k = \begin{bmatrix} | & & | \\ y_{1,k} & \cdots & y_{N,k} \\ | & & | \end{bmatrix}_{d\times N},$$

$$\partial \underline{\underline{F}}(\underline{\underline{X}}_k, \xi_k) = \begin{bmatrix} | & & | \\ \nabla F_1(x_{1,k}, \xi_{1,k}) & \cdots & \nabla F_N(x_{N,k}, \xi_{N,k}) \\ | & & | \end{bmatrix}_{d\times N}.$$

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{i,k}, \quad \bar{y}_k = \frac{1}{N}\sum_{i=1}^{N} y_{i,k}.$$

In matrix form:

$$\begin{cases} \underline{\underline{X}}_{k+1} = \underline{\underline{X}}_k \underline{\underline{W}} - s_k \underline{\underline{Y}}_k \\ \underline{\underline{Y}}_{k+1} = \underline{\underline{Y}}_k \underline{\underline{W}} + \partial \underline{\underline{F}}(\underline{\underline{X}}_{k+1}, \xi_{k+1}) - \partial \underline{\underline{F}}(\underline{\underline{X}}_k, \xi_k). \end{cases}$$

Right multiply both eqns by $\frac{1}{N}\underline{1}_N$:

$$\Rightarrow \begin{cases} \frac{1}{N}\underline{\underline{X}}_{k+1}\underline{1}_N = \frac{1}{N}\underline{\underline{X}}_k \underline{\underline{W}}\underline{1}_N - \frac{s_k}{N}\underline{\underline{Y}}_k \underline{1}_N \\ \frac{1}{N}\underline{\underline{Y}}_{k+1}\underline{1}_N = \frac{1}{N}\underline{\underline{Y}}_k \underline{\underline{W}}\underline{1}_N + \frac{1}{N}\partial \underline{\underline{F}}(\underline{\underline{X}}_{k+1}, \xi_{k+1})\underline{1}_N - \frac{1}{N}\partial \underline{\underline{F}}(\underline{\underline{X}}_k, \xi_k)\underline{1}_N \end{cases}$$

$$\underline{\underline{W}}\underline{1}_N = \underline{1}_N$$

$$\Rightarrow \begin{cases} \bar{x}_{k+1} = \bar{x}_k - s_k \bar{y}_k & (1) \\ \bar{y}_{k+1} = \bar{y}_k + \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\cancel{\underline{\underline{X}}_{k+1}, \xi_{k+1}}\; x_{i,k+1}, \xi_{i,k+1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\cancel{\underline{\underline{X}}_k, \xi_k}\; x_{i,k}, \xi_{i,k}) & (2). \end{cases}$$

From (2), by recursion on $\bar{y}_k$:

$$\bar{y}_{\cancel{k+1}\,k} = \frac{1}{N}\sum_{i=1}^{N}\nabla F_i(\cancel{\underline{\underline{X}}_{k+1}, \xi_{k+1}}\; x_{i,k}, \xi_{i,k}).$$ From (1):

$$\bar{x}_{k+1} = \bar{x}_k - \frac{s_k}{N}\sum_{i=1}^{N}\nabla F_i(x_{i,k}, \xi_{i,k}).$$

Exactly the same as DSGD.

Convergence Results for GT-DSGD

$\underline{A} \in \mathbb{R}^{m\times n}$, $\underline{B} \in \mathbb{R}^{p\times q}$

$$\underline{A} \otimes \underline{B} = \begin{bmatrix} a_{11}\underline{B} & \cdots & a_{1n}\underline{B} \\ \vdots & & \\ a_{m1}\underline{B} & \cdots & a_{mn}\underline{B} \end{bmatrix} \in \mathbb{R}^{mp\times nq}$$

- Define $P^k \triangleq \mathbb{E}[f(\bar{\mathbf{x}}_k)] + \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}_N \otimes \bar{\mathbf{x}}_k\|^2] + Q\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}_N \otimes \bar{\mathbf{y}}_k\|^2]$

Theorem 4 (Convergence of Agent-Average [Lu et al. DSW'19])

If the step-size is set to $\frac{C_0}{\sqrt{T}}$, then it holds that:

kronecker product

$$C_1\mathbb{E}[\|\bar{\mathbf{y}}_k\|^2] + \frac{C_2}{C_0}\mathbb{E}[\|\mathbf{x}_t - \underline{\mathbf{1}_N \otimes \bar{\mathbf{x}}_t}\|^2] \leq \left(\frac{P^0 - P^*}{C_0} + C_4C_0\sigma^2\right)\frac{1}{\sqrt{T}}$$

$$\mathbf{x}_t = \begin{bmatrix} \underline{x}_{1,t} \\ \vdots \\ \underline{x}_{N,t} \end{bmatrix} \qquad \mathbf{1}_N \otimes \bar{\mathbf{x}}_t = \begin{bmatrix} \bar{\underline{x}}_t \\ \vdots \\ \bar{\underline{x}}_t \end{bmatrix}$$

[Zhang, Liu, Zhu, Bentley, Infocom'20]

$O(\frac{1}{T})$.

[— —— —— ——, MobiHoc'20]

[Liu, Zhang, Liu, Lu, NeurIPS'21] "MARL" "single-loop". $O(\frac{1}{T})$

Convergence Results for GT-GSGD

Theorem 5 (Contration of Consensus Gap [Lu et al. DSW'19])

Let ρ be some constant such that $(1+\rho)\beta^2 < 1$. It holds that:

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}_N \otimes \bar{\mathbf{x}}_{k+1}\|] \leq (1+\rho)\beta^2 \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}_N \otimes \bar{\mathbf{x}}_k\|^2]$$
$$+ 3\left(1+\frac{1}{\rho}\right) s^2 \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}_N \otimes \bar{\mathbf{y}}_k\|^2] + 6\left(1+\frac{1}{\rho}\right) s^2 \kappa\sigma^2,$$

$$\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}_N \otimes \bar{\mathbf{y}}_{k+1}\|] \leq \frac{4L^2 s^2}{N}\left(1+\frac{1}{\beta}\right)^2 \|\bar{\tilde{\mathbf{y}}}_k\|^2$$
$$+ \left(\frac{L^2}{N^2}\beta^2(1+\rho)\left(1+\frac{1}{\rho}\right) + \frac{4L^2}{N^2}\left(1+\frac{1}{\rho}\right)^2\right) \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}_N \otimes \bar{\mathbf{x}}_k\|^2]$$
$$+ \left((1+\rho)\beta^2 + \frac{4L^2 s^2}{N^2}\left(1+\frac{1}{\rho}\right)^2\right) \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}_N \otimes \bar{\mathbf{y}}_k\|^2]$$
$$\frac{4L^2 s^2}{N^2}\left(1+\frac{1}{\rho}\right)^2 \kappa\sigma^2.$$

Next Class

Zeroth-Order Methods