Non-Convex Distributed Gradient Descent

Mingyi Hong

University Of Minnesota

M. Hong would like to thank Mr. Siliang Zeng for helping prepare the slides.

Outline

- DGD for Non-Convex Problems using constant stepsize.
- DGD for Non-Convex Problems using decreasing stepsize.
- Convergence analysis of Non-Convex DGD with constant stepsize.

Discussion on DGD

- As we introduced in the last lectures, a number of decentralized algorithms have been proposed for convex consensus optimization.
- However, to the behaviors or consensus nonconvex optimization, our understanding is more limited.
- In this lecture, we will introduce methods for non-convex problems.
- The paper is mainly based on [Zeng-Yin 18] [1]

[1] Zeng and Yin, "On Nonconvex Decentralized Gradient Descent", IEEE TSP, 2018

Discussion on DGD

- Just as DGD for convex problems, Non-Convex DGD with constant stepsize can only converge to a neighborhood of consensus stationary solution.

- When diminishing step sizes are used, convergence to a consensus stationary solution under some regular assumptions can be proved.

Non-Convex Multiagent-Optimization Problem

- We consider an undirected, connected network of m agents and the following consensus optimization problem defined on the network:

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & \sum_{i=1}^{m} f_i(x) \\ \text{subject to} \quad & x \in \mathbb{R} \end{aligned} \tag{1.1}$$

- where f_i is a differentiable function only known to the agent i.

Non-Convex Multiagent-Optimization Problem

- Consider a connected undirected network $\mathcal{G} = \{\mathcal{N}, \xi\}$, where $\mathcal{N}$ is a set of m nodes and ξ is the edge set.

- Any edge $(i, j) \in \xi$ represents a communication link between nodes i and j. Let $x_i \in \mathbb{R}^n$ denote the local copy of x at node i.

Non-Convex Multiagent-Optimization Problem

- We reformulate the consensus problem (1.1) into the equivalent problem:

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & f(]\boldsymbol{x}) := \sum_{i=1}^{m} f_i(x_i), \\ \text{subject to} \quad & x_i = x_j, \forall (i,j) \in \xi \end{aligned} \tag{1.2}$$

where $\boldsymbol{x} \in \mathbb{R}^m$, $f(\boldsymbol{x}) \in \mathbb{R}$ as we defined previously.

Algorithm: Non-Convex DGD

- The algorithm DGD for the non-convex objective (1.2) is described as follows.

- Pick an arbitrary $\boldsymbol{x}^0$. For $k = 0, 1, \cdots$, compute

$$\boldsymbol{x}^{k+1} \leftarrow W\boldsymbol{x}^k - \alpha_k \nabla f(\boldsymbol{x}^k) \tag{1.3}$$

  where W is a mixing matrix and $\alpha_k > 0$ is a step-size parameter.

Convergence Analysis of Non-Convex DGD

- To start the analysis of Non-Convex DGD, we first need to construct several important definitions and assumptions.
- Compared with the analysis of Convex DGD, the assumptions introduced in Non-Convex DGD are more specific.

Convergence Analysis of Non-Convex DGD

Definition 1.1

(Lipschitz differentiability): A function h is called Lipschitz differentiable if h is differentiable and its gradient ∇h is Lipschitz continuous, i.e., $\|\nabla h(x) - \nabla h(y)\| \leq L\|x - y\|$, $\forall x, y \in \mathrm{dom}(h)$, where $L > 0$ is its Lipschitz constant.

Lipschitz differentiability: a common condition.

Convergence Analysis of Non-Convex DGD

Definition 1.2

(Coercivity): A function h is called coercive if $\|u\| \to \infty$ implies $h(x) \to \infty$.

Coercivity is a new condition we introduce.

Convergence Analysis of Non-Convex DGD

- With these new definitions, now we are able to construct the assumptions we need.

Convergence Analysis of Non-Convex DGD

- **Assumption 1** (Objective): The objective functions $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $i = 1, ..., n$, satisfy the following:
  - f_i is Lipschitz differentiable with constant $L_{f_i} > 0$.
  - f_i is proper (i.e., not everywhere infinite) and coercive.

Convergence Analysis of Non-Convex DGD

- According to Assumption 1, the sum $\sum_{i=1}^{m} f_i(x_i)$ is L_f-Lipschitz differentiable with $L_f := \max_i L_{f_i}$.
- In addition, each f_i is lower bounded following Part (2) of the assumption 1.
- Then we can construct the second assumption which is about the mixing matrix.

Convergence Analysis of Non-Convex DGD

- **Assumption 2** (Mixing matrix): The mixing matrix $W = [w_{ij} \in \mathbb{R}^{n \times n}]$ has the following properties:
  - (Graph) If $i \neq j$ and $(i, j) \notin \xi$, then $w_{ij} = 0$, otherwise, $w_{ij} > 0$.
  - (Symmetry) $W = W^T$.
  - (Null space property) null$\{I - W\}$ = span$\{\mathbf{1}\}$.
  - (Spectral property) $I \succeq W \succ -I$.

Convergence Analysis of Non-Convex DGD

- By Assumption 2, a solution $\boldsymbol{x}_{\text{opt}}$ to problem(1.2) satisfies

$$(I - W)\boldsymbol{x}_{\text{opt}} = 0$$

- Due to the symmetric assumption of W , its eigenvalues are real and can be sorted as

$$1 = \lambda_1(W) > \lambda_2(W) \geq \cdots \geq \lambda_n(W) > -1.$$

where $\lambda_i(W)$ denote the ith largest eigenvalue of W .

- Let ζ be the second largest magnitude eigenvalue of W. Then

$$\zeta = \max\{|\lambda_2(W)|, |\lambda_n(W)|\}. \tag{1.4}$$

Convergence Analysis of Non-Convex DGD

- Given those definitions and well-constructed assumptions, now we are able to analyze the convergence results of Non-Convex DGD.
- We consider the convergence of DGD with both a fixed step size and a sequence of decreasing step sizes.

Convergence results of DGD with a fixed step size

- The convergence result of DGD with a fixed step size (i.e., $\alpha_k \equiv \alpha$) is established based on the Lyapunov function:

$$\mathcal{L}_\alpha(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) + \frac{1}{2\alpha}\|\boldsymbol{x}\|_{I-W}^2 \tag{1.5}$$

Convexity is not assumed.

Convergence of DGD with constant step size

Lemma 1.3

(Gradient descent interpretation) The sequence $\{\boldsymbol{x}^k\}$ generated by the DGD iteration (1.3) *is the same sequence generated by applying gradient descent with the fixed step size α to the objective function $\mathcal{L}_\alpha(\boldsymbol{x})$.*

Proof:

$$
\begin{aligned}
\boldsymbol{x}^{k+1} &= W\boldsymbol{x}^k - \nabla f(\boldsymbol{x}^k) \\
&= \boldsymbol{x}^k - \alpha\left(\nabla f(\boldsymbol{x}^k) + \alpha^{-1}(I - W)\boldsymbol{x}^k\right) \\
&= \boldsymbol{x}^k - \alpha \nabla \mathcal{L}_\alpha(\boldsymbol{x}^k) \qquad (1.6)
\end{aligned}
$$

$\square$

DGD can be interpreted as a centralized descent of $\mathcal{L}_\alpha(\boldsymbol{x})$.

Convergence of DGD with constant step size

Lemma 1.4

(Sufficient descent of $\{\mathcal{L}_\alpha(\boldsymbol{x}^k)\}$) Let Assumptions 1 and 2 hold. Set the step size $0 < \alpha < \frac{1+\lambda_n(W)}{L_f}$. It holds that for all $k \in \mathbb{N}$

$$L_\alpha(\boldsymbol{x}^{k+1}) \leq L_\alpha(\boldsymbol{x}^k) - \frac{1}{2}\left(\alpha^{-1}(1+\lambda_n(W)) - L_f\right)\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2, \tag{1.7}$$

Proof: From $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha\nabla\mathcal{L}_\alpha(\boldsymbol{x}^k)$, it follows that

$$\langle \nabla L_\alpha(\boldsymbol{x}^k), \boldsymbol{x}^{k+1} - \boldsymbol{x}^k\rangle = -\frac{\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2}{\alpha}. \tag{1.8}$$

Convergence of DGD with constant step size

- Since $\sum_{i=1}^m \nabla f_i(\boldsymbol{x}_i)$ is L_f-Lipschitz, $\nabla \mathcal{L}_\alpha$ is Lipschitz with the constant

$$L^* \triangleq L_f + \alpha^{-1}\lambda_{\max}(I - W) = L_f + \alpha^{-1}\left(I - \lambda_n(W)\right).$$

- It implies

$$\begin{aligned} L_\alpha(\boldsymbol{x}^{k+1}) \leq & L_\alpha(\boldsymbol{x}^k) + \langle \nabla L_\alpha(\boldsymbol{x}^k), \boldsymbol{x}^{k+1} - \boldsymbol{x}^k \rangle \\ & + \frac{L^*}{2}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \end{aligned} \tag{1.9}$$

which is the desired result. $\square$

Convergence of DGD with constant step size

Lemma 1.5

(Boundedness). Under Assumptions 1 and 2, if $0 < \alpha < \frac{1+\lambda_n(W)}{L_f}$, *then the sequence* $\{\mathcal{L}_\alpha(\boldsymbol{x}^k)\}$ *is lower bounded, and the sequence* $\{\boldsymbol{x}^k\}$ *is bounded, i.e., there exists a constant* $\mathcal{B} > 0$ *such that* $\|\boldsymbol{x}^k\| < \mathcal{B}$ *for all* k.

Convergence of DGD with constant step size

Proof of Lemma (1.5)

- The lower boundedness of $\mathcal{L}_\alpha(\boldsymbol{x}^k)$ is due to the lower boundedness of each f_i as it is proper and coercive (Assumption 1 Part (2)).

- By Lemma (1.4) and the choice of α, $\mathcal{L}_\alpha(\boldsymbol{x}^k)$ is nonincreasing and upper bounded by $\mathcal{L}_\alpha(\boldsymbol{x}^0) < \infty$. Hence, $f(\boldsymbol{x}^k) \leq \mathcal{L}_\alpha(\boldsymbol{x}^0)$ implies that $\boldsymbol{x}^k$ is bounded due to the coercivity of $f(\boldsymbol{x})$ (Assumption 1 Part (2)). $\square$

Convergence of DGD with constant step size

- Utilizing Lemma (1.4) and (1.5), we can immediately obtain the following lemma:

Lemma 1.6

(ℓ_2^2-summable and asymptotic regularity): It holds that $\sum_{k=0}^{\infty} \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| < \infty$ and that $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \to 0$ as $k \to \infty$.

Convergence of DGD with constant step size

Lemma 1.7

(Gradient Bound): $\|\nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)\| \leq \alpha^{-1}\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|$.

- This Lemma is directly from the equation (1.6) $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha\nabla\mathcal{L}_\alpha(\boldsymbol{x}^k)$.
- Based on the above lemmas, we get the global convergence of DGD.

Convergence results of DGD with a fixed step size

Theorem 1.8

(Global convergence).

- *Let* $\{\boldsymbol{x}^k\}$ *be the sequence generated by DGD* (1.3) *with the step size* $0 < \alpha < \frac{1+\lambda_n(W)}{L_f}$. *Let Assumptions 1 and 2 hold. Then* $\{\boldsymbol{x}^k\}$ *has at least one accumulation point* $\boldsymbol{x}^*$, *and any such point is a stationary point of* $\mathcal{L}_\alpha(\boldsymbol{x})$.

- *Furthermore, the rates of the sequences* $\{\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|\}$, *and* $\{\|\nabla \mathcal{L}_\alpha(\boldsymbol{x})\|^2\}$, *and* $\{\|\frac{1}{n}\mathbf{1}^T \nabla f(\boldsymbol{x}^k)\|^2\}$ *are* $o(\frac{1}{k})$. *The convergence rate of the sequence* $\{\frac{1}{K}\sum_{k=0}^{K-1} \|\frac{1}{n}\mathbf{1}^T \nabla f(\boldsymbol{x}^k)\|^2\}$ *is* $\mathcal{O}(\frac{1}{K})$.

Convergence of DGD with constant step size

Proof sketch of DGD with constant stepsize:

- Step 1: DGD is interpreted as the gradient descent algorithm applied to the Lyapunov function $\mathcal{L}_\alpha$.

- Step 2: Sufficient descent, lower boundedness, and bounded gradients are established for the sequence $\{\mathcal{L}_\alpha(\boldsymbol{x}^k)\}$, giving subsequence convergence of the DGD iterates;

Convergence of DGD with constant step size

Proof of Theorem 1:

Recall the Theorem 1 (1.8), we are ready to prove its convergence.

- By Lemma (1.5), the sequence $\{\boldsymbol{x}^k\}$ is bounded, so there exist a convergent subsequence and a limit point, denoted by $\{\boldsymbol{x}^{k_s}_{s\in\mathbb{N}} \to \boldsymbol{x}^*\}$ as $s \to \infty$.

- By Lemmas 1.4 and 1.5, $\mathcal{L}_\alpha(\boldsymbol{x}^k)$ is monotonically nonincreasing and lower bounded, and therefore $\mathcal{L}_\alpha(\boldsymbol{x}^k) \to \mathcal{L}^*$ for some $\mathcal{L}^*$ and $\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\| \to 0$ as $k \to \infty$.

Convergence of DGD with constant step size

- Based on Lemma 1.7, $\|\nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)\| \to 0$ as $k \to \infty$. In particular, $\|\nabla \mathcal{L}_\alpha(\boldsymbol{x}^{k_s})\| \to 0$ as $s \to \infty$.
- Hence, we have $\nabla \mathcal{L}_\alpha(\boldsymbol{x}^*) = 0$.

Convergence of DGD with constant step size

- The running best rate of the sequence $\{\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2\}$ is $o(\frac{1}{k})$ according to Theorem 3.3.1 in K. Knopp-1956 [2].

- Therefore, by Lemma 1.7, the running best rate of the sequence $\{\|\nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)\|^2\}$ is $o(\frac{1}{k})$.

[2] Knopp, Konrad. Infinite sequences and series. Courier Corporation, 1956.

Convergence of DGD with constant step size

- By (1.5), we know $\nabla \mathcal{L}_\alpha(\boldsymbol{x}^k) = \nabla f(\boldsymbol{x}^k) + \alpha^{-1}(I - W)\boldsymbol{x}^k$, which implies $\frac{1}{n}\mathbf{1}^T \nabla f(\boldsymbol{x}^k) = \frac{1}{n}\mathbf{1}^T \nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)$ due to $\frac{1}{n}\mathbf{1}^T(I - W) = 0$.

- Thus, we obtain

$$\|\frac{1}{n}\mathbf{1}^T \nabla f(\boldsymbol{x}^k)\|^2 = \|\frac{1}{n}\mathbf{1}^T \nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)\|^2 \leq \|\nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)\|^2,$$

  which implies the running best rate of $\{\|\frac{1}{n}\mathbf{1}^T \nabla f(\boldsymbol{x}^k)\|^2\}$ is also $o(\frac{1}{k})$.

Convergence of DGD with constant step size

- By Lemmas 1.4 and 1.7, it holds that

$$\|\nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)\|^2 \le \frac{2}{\alpha\left(1+\lambda_n(W)-\alpha L_f\right)}\left(\mathcal{L}_\alpha(\boldsymbol{x}^k)-\mathcal{L}_\alpha(\boldsymbol{x}^{k+1})\right),$$

which implies

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)\|^2 \le \frac{2\left(\mathcal{L}_\alpha(\boldsymbol{x}^0)-\mathcal{L}^*\right)}{\alpha\left(1+\lambda_n(W)-\alpha L_f\right)K}.$$

- Moreover, we note that $\|\frac{1}{n}\mathbf{1}^T\nabla f(\boldsymbol{x}^k)\|^2 \le \|\nabla \mathcal{L}_\alpha(\boldsymbol{x}^k)\|^2$. Thus, the convergence rate of $\{\frac{1}{K}\sum_{k=0}^{K-1}\|\frac{1}{n}\mathbf{1}^T\nabla f(\boldsymbol{x}^k)\|^2\}$ is $\mathcal{O}(\frac{1}{K})$.

$\square$

Convergence of DGD with constant step size

- Next, we derive the bound D on the gradient sequence $\{\nabla f(\boldsymbol{x}^k)\}$.

Convergence of DGD with constant step size

Lemma 1.9

Under Assumption 1, there exists a point $\boldsymbol{y}^*$ *satisfying* $\nabla f(\boldsymbol{y}^*) = 0$*, and the following bound holds*

$$\|\nabla f(\boldsymbol{x}^k)\| \leq D \triangleq L_f(\mathcal{B} + \|\boldsymbol{y}^*\|)\|, \forall k \in \mathbb{N}, \tag{1.10}$$

where $\mathcal{B}$ *is the bound of* $\|\boldsymbol{x}^k\|$ *given in Lemma 1.5.*

Convergence of DGD with constant step size

Proof:

- By the lower boundedness assumption (Assumption 1 Part (2)), the minimizer of $f(\boldsymbol{y})$ exists. Let $\boldsymbol{y}^*$ be a minimizer.
- Then by Lipschitz differentiability of each f_i (Assumption 1), we have that $\nabla f(\boldsymbol{y}^*) = 0$.
- Then for any k, we have

$$\begin{aligned}
\|\nabla f(\boldsymbol{x}^k)\| &= \|\nabla f(\boldsymbol{x}^k) - \nabla f(\boldsymbol{y}^*)\| \\
&\leq L_f \|\boldsymbol{x}^k - \boldsymbol{y}^*\| \\
&\leq L_f (\mathcal{B} + \|\boldsymbol{y}^*\|)
\end{aligned}$$

- Therefore, we proven this lemma. $\square$

Convergence results of DGD with a fixed step size

- According to Theorem 1.8, the sequence $\{\boldsymbol{x}^k\}$ can converge to $\boldsymbol{x}^*$ which is a stationary point of $\mathcal{L}_\alpha(\boldsymbol{x})$.

- Therefore, we get

$$\nabla \mathcal{L}_\alpha(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*) + \frac{1}{\alpha}(I - W)\boldsymbol{x}^* = 0. \tag{1.11}$$

$$\mathbf{1}^T \nabla \mathcal{L}_\alpha(\boldsymbol{x}^*) = \mathbf{1}^T \nabla f(\boldsymbol{x}^*) + \frac{1}{\alpha}\mathbf{1}^T(I - W)\boldsymbol{x}^* = 0. \tag{1.12}$$

- Since $\mathbf{1}^T(I - W) = 0$, (1.12) yields $\mathbf{1}^T \nabla f(\boldsymbol{x}^*) = 0$, indicating that $\boldsymbol{x}^*$ is also a stationary point to the separable function $\sum_{i=1}^m f_i(\boldsymbol{x}_i)$.

Convergence results of DGD with a fixed step size

- Since the rows of $\boldsymbol{x}^*$ are not necessarily identical, we cannot say $\boldsymbol{x}^*$ is a stationary point to objective (1.2).
- However, the differences between the rows of $\boldsymbol{x}^*$ can be bounded.
- We show the bound in our next result. The result is adapted from K. Yuan-2016 [3].

[3] Yuan, Kun, Qing Ling, and Wotao Yin. "On the convergence of decentralized gradient descent." SIAM Journal on Optimization 26.3 (2016): 1835-1854.

Convergence results of DGD with a fixed step size

Proposition 1 (Consensual bound on $\boldsymbol{x}^*$):

- For each iteration k, define $\bar{x}^k := \frac{1}{n}\sum_{i=1}^m x_i^k$. Then, it holds for each node i that

$$\|x_i^k - \bar{x}^k\| \leq \frac{\alpha D}{1-\zeta}, \tag{1.13}$$

  where D is a universal bound of $\|\nabla f(\boldsymbol{x}^k)\|$ defined in Lemma 1.9

- As $k \to \infty$, (1.13) yields the consensual bound

$$\|x_i^* - \bar{x}^*\| \leq \frac{\alpha D}{1-\zeta},$$

  where $\bar{x}^* := \frac{1}{m}\sum_{i=1}^m x_i^*$.

Convergence results of DGD with a fixed step size

Proof of Proposition 1:

- According to the update (1.3), we obtain that

$$\boldsymbol{x}^k = W^k \boldsymbol{x}^0 - \alpha \sum_{j=0}^{k-1} W^{k-1-j} \nabla f(\boldsymbol{x}^j).$$

- Moreover, we denote that $\bar{x}^k = \frac{1}{m}\mathbf{1}^T \boldsymbol{x}^k$ and $\bar{\boldsymbol{x}}^k = \frac{1}{m}\mathbf{1}\mathbf{1}^T \boldsymbol{x}^k$.

Convergence results of DGD with a fixed step size

- As a result,

$$
\begin{aligned}
&\|\boldsymbol{x}_i^k - \bar{x}^k\| \\
\leq &\|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^k\| \\
= &\|\boldsymbol{x}^k - \frac{1}{m}\mathbf{1}\mathbf{1}^T\boldsymbol{x}^k\| \\
= &\| \left(I - \frac{1}{m}\mathbf{1}\mathbf{1}^T\right)\left(W^k\boldsymbol{x}^0 - \alpha\sum_{j=0}^{k-1} W^{k-1-j}\nabla f(\boldsymbol{x}^j)\right)\| \\
\leq &\| \left(W - \frac{1}{m}\mathbf{1}\mathbf{1}^T\right)^k \| \cdot \|\boldsymbol{X}^0\| + \alpha\sum_{j=0}^{k-1}\|W^{k-1-j} - \frac{1}{m}\mathbf{1}\mathbf{1}^T\|\|\nabla f(\boldsymbol{x}^j)\| \\
\leq &\zeta^k\|\boldsymbol{x}^0\| + \alpha D\sum_{j=0}^{k-1}\zeta^{k-1-j}
\end{aligned}
$$

Convergence results of DGD with a fixed step size

- As $k \to \infty$, $\zeta^k \|\boldsymbol{x}^0\|$ converges to 0.
- Moreover, $\alpha D \sum_{j=0}^{k-1} \zeta^{k-1-j}$ is bounded by $\frac{\alpha D}{1-\zeta}$.
- Hence, we completes the proof.

$\square$

Convergence results of DGD with a fixed step size

- Up to now, we see that using fixed step sizes, our results are limited.
- The stationary point $\boldsymbol{x}^*$ of $\mathcal{L}_\alpha$ is not a stationary point of the original problem.
- To address this issue, decreasing step sizes is used and better convergence results are obtained!.

Convergence of DGD with decreasing step sizes

- In Proposition 1, we see the consensual error bound is proportional to the constant step size α.
- Therefore, it motivates the use of properly decreasing step size $\alpha_k = \mathcal{O}(\frac{1}{(k+1)^\epsilon})$ for some $0 < \epsilon \leq 1$, to diminish the consensual bound to 0.
- As a result, any accumulation point $\boldsymbol{x}^*$ becomes a stationary point of the original problem (1.2).

Convergence of DGD with decreasing step sizes

- To analyze DGD with decreasing step sizes, we add the following assumption.
- **Assumption 3** (Bounded gradient): For any k, $\nabla f(\boldsymbol{x}^k)$ is uniformly bounded by some constant $B > 0$, i.e., $\|\nabla f(\boldsymbol{x}^k)\| \le B$.
- This assumption is regular in the convergence analysis of decentralized gradient methods, though not required for centralized gradient descent.

Convergence of DGD with decreasing step sizes

- We take the step size sequence:

$$\alpha_k = \frac{1}{L_f(k+1)^\epsilon}, \quad 0 < \epsilon \leq 1. \tag{1.14}$$

- By iteratively applying iteration (1.3), we obtain the following expression

$$\boldsymbol{x}^k = W^k \boldsymbol{x}^0 - \sum_{j=0}^{k-1} \alpha_j W^{k-1-j} \nabla f(\boldsymbol{x}^j). \tag{1.15}$$

Convergence of DGD with decreasing step sizes

Proposition 3 (Asymptotic consensus rate). Let Assumptions 2 and 3 hold. Let DGD use (1.14). Let $\bar{\boldsymbol{x}}^k := \frac{1}{n}\mathbf{1}\mathbf{1}^T\boldsymbol{x}^k$. Then, $\|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^k\|$ converges to 0 at the rate of $\mathcal{O}(1/(k+1)^\epsilon)$.

Convergence of DGD with decreasing step sizes

- According to Proposition 3, decreasing step sizes can reach consensus asymptotically. (compared to a nonzero bound in the fixed step size case in Proposition 1)
- Moreover, with a larger ϵ, faster decaying step sizes generally imply a faster asymptotic consensus rate.

Convergence of DGD with decreasing step sizes

- Note that $(I - W)\bar{\boldsymbol{x}}^k = 0$ and thus

$$\|\boldsymbol{x}^k\|_{I-W}^2 = \|\boldsymbol{x}^k - \bar{\boldsymbol{x}}^k\|_{I-W}^2$$

. Then we can have the following result:

Corollary 1.10

Apply the setting of Proposition 3, $\|\boldsymbol{x}^k\|_{I-W}^2$ converges to 0 at the rate of $\mathcal{O}(\frac{1}{(k+1)^{2\epsilon}})$.

Convergence of DGD with decreasing step sizes

Theorem 1.11

(Final Convergence Results). Let Assumptions 1, 2 and 3 hold. Let DGD use step sizes (1.14)*. Then we obtain*

- $\{\mathcal{L}_{\alpha_{\|}}\}$ *and* $\{\mathbf{1}^T f(\boldsymbol{x}^k)\}$ *converge to the same limit;*
- $\lim_{k\to\infty} \mathbf{1}^T \nabla f(\boldsymbol{x}^k) = 0$*, and any limit point of* $\{\boldsymbol{X}^k\}$ *is a stationary point of problem* (1.2)*.*

Convergence of DGD with decreasing step sizes

- In the proof of Theorem (1.11), we will establish

$$\sum_{k=0}^{\infty} \left(\alpha_k^{-1}(1 + \lambda_n(W)) - L_f \right) \|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2 < \infty,$$

  which implies that the running best rate of the sequence $\{\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|\}$ is $o(1/k^{1+\epsilon})$.

- Theorem (1.11) shows that the objective sequence converges, and any limit point of $\{\boldsymbol{x}^k\}$ is a stationary point of the original problem.