Generalization of Decentralized Gradient Descent with Separable Data

Hossein Taheri HOSSEIN@UCSB.EDU
Department of Electrical and Computer Engineering, UC Santa Barbara.

Christos Thrampoulidis CTHRAMPO@UBC.CA
Department of Electrical and Computer Engineering, University of British Columbia.

Abstract

Decentralized learning offers privacy and communication efficiency when data are naturally distributed among agents communicating over an underlying graph. Motivated by overparameterized learning settings, in which models are trained to zero training loss, we study algorithmic and generalization properties of decentralized learning with gradient descent on separable data. Specifically, for decentralized gradient descent (DGD) and a variety of loss functions that asymptote to zero at infinity (including exponential and logistic losses), we derive novel finite-time generalization bounds. This complements a long line of recent work that studies the generalization performance and the implicit bias of gradient descent over separable data, but has thus far been limited to centralized learning scenarios. Notably, our generalization bounds match in order their centralized counterparts. Critical behind this, and of independent interest, is establishing novel bounds on the training loss and the rate-of-consensus of DGD for a class of self-bounded losses. Finally, we conduct numerical experiments which corroborate our theoretical results.

1. Introduction

Machine learning tasks often revolve around inference from data using empirical risk minimization (ERM):

$$\min_{w\in\mathbb{R}^d} \hat{F}(w) := \frac{1}{n}\sum_{i=1}^{n} f\left(w, x_i\right). \tag{1}$$

Here $f:\mathbb{R}^d\times\mathbb{R}^{d'}\to\mathbb{R}$ is a loss function and $x_i := y_i a_i$, where $(a_i, y_i)_{i=1}^n \overset{\text{iid}}{\sim} \mathcal{D}$ represent features and labels, sampled from a distribution $\mathcal{D}$. In large scale machine learning, due to privacy concerns and communication constraints, data points are often distributed on a set of local computing agents. Decentralized learning methods aim at minimizing the global loss function (1) while agents communicate their parameters on an underlying connected graph. The most ubiquitous of these algorithms is Decentralized Gradient Descent (DGD). Here the ℓth agent runs a step of gradient descent followed by an averaging step in which every agent replaces its parameter with the average of its neighbors [18]:

$$w_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} w_k^{(t)} - \eta_t \nabla \hat{F}_\ell(w_\ell^{(t)}). \tag{2}$$

The superscripts signify the iteration number and $A_{\ell k}$ refers to the averaging weights used by agent ℓ for the parameter of agent k. The global loss $\hat{F}$ is the average of local loss functions $\hat{F}_\ell$, $\ell \le N$, where each $\hat{F}_\ell$ is formed as the average empirical risk evaluated on the local training dataset $\mathcal{S}_\ell$ of the ℓth agent:

© H. Taheri & C. Thrampoulidis.

$$\hat{F}(w) = \frac{1}{N}\sum_{\ell=1}^{N} \hat{F}_\ell(w), \quad \hat{F}_\ell(w) = \frac{1}{n_\ell}\sum_{x_j \in \mathcal{S}_\ell} f(w, x_j), \tag{3}$$

where n_ℓ denotes the dataset size of agent ℓ. Convergence properties of the train loss $\hat{F}(\cdot)$ in DGD have been studied extensively in literature, e.g., [13, 16–18, 31]. The bulk of these studies build upon classical optimization theory [20] suited for studying the train loss per iteration. In particular, it is well-stablished in the literature that DGD converges at the rate $\frac{1}{T}\sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)}) - \hat{F}^\star = \mathcal{O}(\frac{1}{\sqrt{T}})$ for smooth convex functions [16]. Here $\bar{w}^{(t)}$ is the average of local parameters $w_\ell^{(t)}$. Our results in Sections 2.1-2.2 show a rate of $\hat{F}(\bar{w}^{(T)}) = \mathcal{O}(\frac{(\log T)^2}{T})$ and $\|W^{(T)} - \bar{W}^{(T)}\|_F^2 = \mathcal{O}(\frac{(\log T)^4}{T^2})$ for the training loss and consensus error of DGD over separable data with "exponentially tailed" losses.

The study of generalization performance of DGD algorithms in the literature is mostly limited to empirical observations e.g., [8, 9, 29], making the theory behind test error performance largely unexplored. Our first goal in this paper is to complement prior general results on the convergence of training loss in DGD by considering specific, but commonly encountered, settings in ERM over separable data. This includes the analysis of non-smooth objectives such as the exponential loss, analysis of logistic regression in the interpolating regime where the optimum is achieved at infinity, and analysis of objectives satisfying the PL condition. Our second goal is to study, for the first time, convergence rates of the DGD test loss $F(\bar{w}^{(t)}) := \mathbb{E}_{x\sim\mathcal{D}}[f(\bar{w}^{(t)}, x)]$. Finally, we leverage recent advances in the study of centralized learning with separable data to design fast algorithms for decentralized learning.

Contributions. Our contributions are summarized as follows,

- In Section 2.1 we derive convergence rates for the training and test loss of the decentralized gradient descent algorithm with separable data. Our results hold for convex losses satisfying realizability and self-boundedness, as well as, convex losses satisfying self-boundedness and the PL condition. In Section 2.2, we prove that under additional self-boundedness assumptions on the Hessian and gradient, which holds for exponentially tailed losses, the test loss bound can be improved to approximately match the test loss bounds of centralized GD. In Appendix A, we extend our results to losses satisfying the PL condition. Our numerical experiments in Appendix H verify our results.
- In Appendix G, we propose two algorithms for speeding up the convergence of decentralized learning with separable data. Notably, numerical experiments demonstrate that our proposed algorithms significantly improve the train error and test error of decentralized logistic regression.

Further related works

Decentralized learning. Over the last few years there have been numerous research works which consider the convergence of first order methods for decentralized learning; an incomplete list includes [2, 8, 10, 11, 13, 16, 18, 22, 27, 28, 30, 31]. The study of generalization of DGD was mostly limited to numerical experiments. While this paper was being written, we were notified of the recent work [26] which studies the generalization bounds of decentralized methods for Lipschitz convex losses. However, we consider the setting of separable data and exponentially tailed losses and show that in our settings the train/test loss of DGD is closely similar to centralized GD. Compared to this

work, we also propose improved algorithms for learning with separable data. We highlight that our rates on the training loss are comparable to [10, Theorem 2]. While [10] also derives convergence of DGD train loss on separable data, their analysis is valid only for bounded optimizers. In contrast, we derive training loss bounds which are true for the case of unbounded optimizers as is the case for logistic regression over separable data.

Implicit bias of GD. For centralized settings, a line of recent works [5, 6, 15, 23–25] studies the parameter convergence, as well as training and test loss convergence, of gradient descent on separable data, showing that for (a class of) monotonic losses the solution to ERM and the max-margin solution are the same in direction., i.e., $\|\hat{w}^{(t)} - \hat{w}_{\text{MM}}\| \to 0$. Here $\hat{w}^{(t)} := w^{(t)}/\|w^{(t)}\|$ and $\hat{w}_{\text{MM}} := w_{\text{MM}}/\|w_{\text{MM}}\|$, where the vector w_{MM} is the solution to the hard-margin support vector machine problem. Notably, [5, 25] characterized the rate of convergence for margin gap to be $\|\hat{w}^{(T)} - \hat{w}_{\text{MM}}\| = \mathcal{O}(1/\log(T))$ and for the training loss to be $\hat{F}(w^{(T)}) = \mathcal{O}(\frac{1}{\eta T})$. Recently, Shamir [24] and Schliserman and Koren [23] showed that the test loss of GD for logistic regression on linearly separable data satisfies $F(w^{(T)}) = \tilde{\mathcal{O}}(\frac{1}{\eta T} + \frac{1}{n})$ signifying that overfitting does not happen during the iterates of GD. In Section 2.2 (Remark 7), we show that the test loss of DGD with logistic regression on linearly separable data satisfies $\mathbb{E}[F(\bar{w}^{(T)})] = \tilde{\mathcal{O}}(\frac{1}{\eta T} + \frac{1}{n} + \eta^2)$. This signifies that, under the data separability assumption, the rate of decay for test loss is essentially identical for centralized and decentralized scenarios.

2. Main results

Throughout the paper we make the following standard assumption on the mixing matrix $A = [A_{ij}]_{N\times N}$ corresponding to the underlying connected network.

Assumption 1 (Mixing matrix) *The mixing matrix $A \in \mathbb{R}^{N\times N}$ is symmetric, doubly stochastic with bounded spectrum i.e., $|\lambda_i(A)| \in (0, 1]$ and $\lambda_2(A) < 1$.*

First, we state a lemma which relates the generalization loss of DGD at iteration t to its train loss and consensus error up to iteration t. The lemma is derived based on a stability analysis [3, 4, 12]. Specifically we use a self-boundedness and a realizability assumption [23] which makes the stability analysis feasible for settings such as logistic regression on separable data. Additionally, we assume convexity and L-smoothness of the loss function. For simplicity, we denote $f_i(w) := f(w, x_i)$ for $i \le n$.

Assumption 2 (Convexity) *The loss functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are convex and differentiable, satisfying,*

$$f_i(w) \le f_i(v) + \langle \nabla f_i(w), w - v\rangle.$$

Assumption 3 (Smoothness) *The loss functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are L-smooth and differentiable:*

$$f_i(w) \le f_i(v) + \langle \nabla f_i(v), w - v\rangle + \frac{L}{2}\|w - v\|^2.$$

Assumption 4 (Self-boundedness of the gradient) *The loss functions $f_i : \mathbb{R}^d \to \mathbb{R}$ satisfy the self-boundedness property with the parameters (c, α),*

$$\|\nabla f_i(w)\| \le c(f_i(w))^\alpha,$$

for $\alpha \in [\frac{1}{2}, 1]$.

We note that Assumption 4 is weaker than Assumption 3, since an L-smooth function f satisfies $\|\nabla f(w)\|^2 \leq 2L(f(w) - f^\star)$, where $f^\star := \inf_w f(w)$. However, we make use of the smoothness property whenever it suits the analysis, in particular in the training loss analysis.

Throughout the paper, we assume that the training data is separable.

Assumption 5 (Data separability) *The training loss $\hat{F}$ in Eq.(3) satisfies the data separability condition i.e., $\hat{F}^\star = 0$, where $\hat{F}^\star := \inf_{w\in\mathbb{R}^d} \hat{F}(w)$.*

Before stating our key lemma, we introduce a few necessary notations. We define the matrix $W^{(t)} \in \mathbb{R}^{N\times d}$ as the concatenation of all agents' parameters at iteration t, i.e., $W = [w_1^{(t)}, \cdots, w_N^{(t)}]^\top$. We also denote by $\bar{w}^{(t)}$ the average of local parameters, i.e., $\bar{w}^{(t)} := \frac{1}{N}\sum_{i=1}^N w_i^{(t)}$, and denote by $\bar{W}^{(t)} \in \mathbb{R}^{N\times d}$ its concatenated matrix $\bar{W}^{(t)} = [\bar{w}^{(t)}, \cdots, \bar{w}^{(t)}]$.

Lemma 1 (Key lemma, Informal version) *Consider the iterates of decentralized gradient descent in Eq.(2) with a fixed positive step-size $\eta \leq \frac{2}{L}$. Let Assumptions 1-4 hold. Then for the test loss F at iteration $T \geq 1$, it holds that*

$$\mathbb{E}[F(\bar{w}^{(T)})] = \tag{4}$$
$$\mathcal{O}\Big(\mathbb{E}[\hat{F}(\bar{w}^{(T)})] + \frac{\eta^2 L^2 c^2 T^2}{n^{3-2\alpha}}\mathbb{E}\Big[(\frac{1}{T}\sum_{t=1}^T \hat{F}(\bar{w}^{(t)}))^{2\alpha}\Big] + \frac{\eta^2 L^4}{N}\mathbb{E}\Big[(\sum_{t=1}^T \|W^{(t)} - \bar{W}^{(t)}\|_F)^2\Big]\Big),$$

where the expectation is over training samples.

The precise statement and the proof of Lemma 1 are deferred to Appendix B. As a result of Lemma 1, the test error is related to the test loss to the train loss and the consensus error. It is worth remarking that Eq. (4) is in fact valid for both DGD and the Decentralized Gradient Tracking (DGT) [19] algorithms. This is merely because the proof of Lemma 1 only relies on the updates of the "averaged" parameter $\bar{w}^{(t)}$ and moreover the update rule of $\bar{w}^{(t)}$ for both DGD and DGT is $\bar{w}^{(t)} = \bar{w}^{(t-1)} - \eta\frac{1}{N}\sum_{\ell=1}^N \nabla\hat{F}_\ell(w_\ell^{(t-1)})$. Thus by Eq.(4) one can also obtain test loss bounds of DGT after replacing appropriate bounds of DGT for the training loss and consensus error. We leave this to future work.

In the following sections, we discuss how Lemma 1 yields test loss bounds on DGD under different assumptions on the loss function.

2.1. Convergence with general convex losses

The upper-bound in Eq.(4) shows how the consensus error and train loss of DGD affect the test loss.

The next lemma bounds the training loss and consensus error of DGD for general convex losses. The proof is deferred to Appendix C.1

Lemma 2 (Training bounds for convex losses) *Under Assumptions 1-4, for any $w \in \mathbb{R}^d$ and for a fixed step-size $\eta < \min(\frac{1-\alpha_1}{L}, \frac{1}{L}\sqrt{\frac{1-\alpha_1}{2\alpha_2}})$, where $\alpha_1 \in (0,1), \alpha_2 > 0$ are parameters that depend only on the mixing matrix, the following holds for the train loss and consensus error of DGD (2):*

$$\frac{1}{T}\sum_{t=1}^T \hat{F}(\bar{w}^{(t)}) \leq \frac{2\|w\|^2}{\eta T} + 4\hat{F}(w), \tag{5}$$
$$\frac{1}{T}\sum_{t=1}^T \|W^{(t)} - \bar{W}^{(t)}\|_F^2 \leq \frac{\alpha_2\eta^2 L^2}{(1-\alpha_1)}(\frac{2\|w\|^2}{\eta T} + 4\hat{F}(w)).$$

To bound the training loss for loss functions $f(\cdot)$ where the optimum is attained at infinity we need a realizability assumption. In particular, we choose $w \in \mathbb{R}^d$ (of Lemma 2) using the following assumption.

Assumption 6 (Realizability) *The loss functions $f_i : \mathbb{R}^d \to \mathbb{R}$ satisfy the realizability condition, i.e. there exists a decreasing function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ such that for every $\varepsilon > 0$ there exists $\hat{w} \in \mathbb{R}^d$ with $\|\hat{w}\| \leq \rho(\varepsilon)$ that satisfies $f_i(\hat{w}) \leq \varepsilon$.*

The set of assumptions 2-6 covers logistic loss over linearly separable data, in addition to losses with other exponential-type tails $\exp(-w^r)$ and polynomial tail w^{-r}, for $r > 0$.

Remark 3 (Training loss of DGD on separable data) *The realizability assumption appeared in [12, 23] (and implicitly used in [5, 24]) as the required assumption for bounding the training error under separable data. It can be checked that for γ denoting the margin of training data, loss functions with an exponential tail such as logistic loss satisfy this assumption with $\rho(\varepsilon) = \frac{1}{\gamma}\log(\frac{1}{\varepsilon})$ (e.g., see Proposition 28 and [23, Lemma 4]). Based on Lemma 2, this leads to the following bound for DGD training loss for all $\varepsilon > 0$,*

$$\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}) \leq \frac{2\log(1/\varepsilon)^2}{\eta T} + 4\varepsilon. \tag{6}$$

In particular, choosing $\varepsilon = 1/T$, gives a rate of $\mathcal{O}(\frac{(\log T)^2}{\eta T})$, surprisingly matching up to the corresponding rate for centralized GD derived in [5, Theorem 1.1].

Theorem 4 (Test loss with convex losses) *Under Assumptions 1-6, by choosing $\eta < \min(\frac{1-\alpha_1}{L\sqrt{T}}, \frac{1}{L}\sqrt{\frac{1-\alpha_1}{2\alpha_2 T}})$ where $\alpha_1 \in (0,1), \alpha_2 > 0$ are parameters that depend only on the mixing matrix and assuming $\varepsilon \leq \frac{\rho(\varepsilon)}{\eta T}$ the following bound holds for the test error of DGD for iteration T,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\bar{w}^{(t)})] = \mathcal{O}\Big(\frac{\rho(\varepsilon)^2}{\sqrt{T}} + \frac{L^2c^2\rho(\varepsilon)^{4\alpha}}{n^{3-2\alpha}}T^{1-\alpha} + \frac{L^4\rho(\varepsilon)^2}{\sqrt{T}}\Big), \tag{7}$$

where the expectation is over i.i.d. training samples.

Remark 5 (DGD with logistic regression never overfits) *The proof of Theorem 4 is delayed to Appendix C.2. As in Remark 3, we take logistic regression on separable data as our case study. For logistic regression (as well as other loss functions with an exponential tail), it can be verified that the self-boundedness assumption holds with $\alpha = 1$. Similar to Remark 3 it holds that $\rho(\varepsilon) = \frac{1}{\gamma}\log(\frac{1}{\varepsilon})$, thus choosing $\varepsilon = 1/\sqrt{T}$ results in a test loss rate $\tilde{\mathcal{O}}(\frac{1}{\sqrt{T}} + \frac{1}{n})$ by Eq.(7). This indicates that the upper-bound decreases at a rate of $\tilde{\mathcal{O}}(\frac{1}{\sqrt{T}})$ until after $T = n^2 \cdot (\max(\frac{1}{Lc}, \frac{L}{c}))^4$ iterations where the upper bound essentially reduces to $\tilde{\mathcal{O}}(\frac{L^2c^2}{n})$. Additionally, the fact that the upper-bound is decreasing proves that with appropriate choice of step-size, overfitting never happens along the path of DGD at any iteration.*

2.2. On the convergence of DGD with exponentially-tailed losses

We note that the bounds in Lemma 2 and Theorem 4 hold for the average loss across iterations $t \leq T$. It is straight-forward to see that if DGD is a descent algorithm i.e., $\hat{F}(\bar{w}^{(t+1)}) \leq \hat{F}(\bar{w}^{(t)})$ for all $t \leq T$, then $\hat{F}(\bar{w}^{(T)}) \leq \frac{1}{T}\sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)})$; thus implying that the upper-bounds on training and test loss hold for the last iterate of DGD. Here, we show that under certain conditions on the loss and step-size, DGD is a descent algorithm. Moreover, we show that the consensus error of 2 as well as the test loss bounds of Theorem 4 can be improved for this class of convex loss functions.

In particular, we use the following assumptions together with the self-boundedness gradient assumption (Assumption 4) with $\alpha = 1$ as well as the convexity assumption.

Assumption 7 (Self-bounded Hessian) *The local losses $\hat{F}_\ell : \mathbb{R}^d \to \mathbb{R}$ satisfy the following for the Hessian matrices $\nabla^2 \hat{F}_\ell$ and a positive constant h,*

$$\|\nabla^2 \hat{F}_\ell(w)\| \leq h\, \hat{F}_\ell(w).$$

Assumption 8 (Self-lowerbounded gradient) *The global loss satisfies for a constant τ that*

$$\|\nabla \hat{F}(w)\| \geq \tau \hat{F}(w).$$

Assumptions 2, 4, 7 and 8 considered in this section, include linear classification with non-smooth losses such as the exponential loss, losses with super-exponential tails ($\exp(-x^r), r > 1$) and the logistic loss; e.g., see Proposition 26 in the appendix.

Theorem 6 (Last iterate convergence of DGD) *Consider DGD with the loss functions and mixing matrix satisfying Assumptions 1,2,7,8 and Assumption 4 with $\alpha = 1$ and $c = h$. Assume that the step-size satisfies $\eta < \frac{\gamma}{\hat{F}(1)}$, for a constant γ depending only on the mixing matrix, τ and h. Then, the train loss and the consensus error of DGD at iteration T satisfy the following for all $w \in \mathbb{R}^d$,*

$$\hat{F}(\bar{w}^{(T)}) \leq 4\hat{F}(w) + \frac{2\|w\|^2}{\eta T},$$

$$\|W^{(T)} - \bar{W}^{(T)}\|_F^2 \leq \mathcal{O}(\eta^2 \hat{F}^2(w) + \frac{\|w\|^4}{T^2}).$$

Remark 7 (Improved rates) *We prove Theorem 6 in Appendix D.1. While similar to Lemma 2, for logistic regression we have $\hat{F}(\bar{w}^{(T)}) = \tilde{\mathcal{O}}(\frac{1}{\eta T} + \frac{1}{T})$, for the consensus error rate we have by applying Theorem 6 and noting that $\rho(\varepsilon) = \log(1/\varepsilon)/\gamma$,*

$$\|W^{(T)} - \bar{W}^{(T)}\|_F^2 \leq \mathcal{O}(\eta^2 \varepsilon^2 + \frac{(\log(1/\varepsilon))^4}{T^2}).$$

After choosing $\varepsilon = 1/T$, we have the improved rate $\|W^{(T)} - \bar{W}^{(T)}\|_F^2 = \tilde{\mathcal{O}}(\frac{1}{T^2})$, which is a significant improvement from $\tilde{\mathcal{O}}(\frac{1}{T})$ for general convex losses with constant η (Lemma 2). For the test loss, employing Lemma 1 with the new rates for the consensus error leads to the following rate for DGD with logistic regression,

$$\mathbb{E}[F(\bar{w}^{(T)})] = \tilde{\mathcal{O}}\Big(\frac{1}{\eta T} + \frac{1}{n} + \eta^2\Big).$$

We note that this is a superior rate for the test loss of logistic regression, compared to the rate of Remark 5. Notably, this is comparable to the test loss rate for "centralized" logistic regression on linearly separable data, derived as $\mathbb{E}[F(\bar{w}^{(T)})] = \tilde{\mathcal{O}}(\frac{1}{\eta T} + \frac{1}{n})$ in [23, 24].

3. Conclusion

We studied the behavior of train loss and test loss of decentralized gradient descent (DGD) methods when training dataset is separable. Notably, our results are the first rigorous guarantees for the generalization error of DGD. This line of work can be extended in several directions, with the most exciting one being the analysis of non-convex objectives.

References

[1] Uci wine data set, web address : https://archive.ics.uci.edu/ml/datasets/wine.

[2] Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Michael Rabbat. Stochastic gradient push for distributed deep learning. *arXiv preprint arXiv:1811.10792*, 2018.

[3] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[4] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[5] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[6] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

[7] Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

[8] Zhanhong Jiang, Aditya Balu, Chinmay Hegde, and Soumik Sarkar. Collaborative deep learning in fixed topology networks. *Advances in Neural Information Processing Systems*, 30, 2017.

[9] Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv:1907.09356*, 2019.

[10] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

[11] Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33:18342–18352, 2020.

[12] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

[13] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

[14] Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les equations aux derive es partielles*, 1963.

[15] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[16] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

[17] Angelia Nedić and Alex Olshevsky. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947, 2016.

[18] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[19] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[20] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[21] Boris Polyak. Gradient methods for the minimisation of functionals. *Ussr Computational Mathematics and Mathematical Physics*, 3:864–878, 1963.

[22] Shi Pu, Wei Shi, Jinming Xu, and Angelia Nedic. Push-pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 2020.

[23] Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. *arXiv preprint arXiv:2202.13441*, 2022.

[24] Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22(85):1–20, 2021.

[25] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[26] Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9756–9764, 2021.

[27] Mohammad Taha Toghani and César A Uribe. Communication-efficient distributed cooperative learning with compressed beliefs. *IEEE Transactions on Control of Network Systems*, 2022.

[28] Mohammad Taha Toghani and César A Uribe. Scalable average consensus with compressed communications. In *2022 American Control Conference (ACC)*, pages 3412–3417. IEEE, 2022.

[29] Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019.

[30] Ran Xin, Usman A Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 69:1842–1858, 2021.

[31] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Appendix

The appendix is organized as follows:

1. Convergence and generalization under the PL condition is discussed in Appendix A. The proofs are included in Appendix E.
2. The proof of Lemma 1 is included in Appendix B.
3. The proofs for Section 2.1 are included in Appendix C.
4. The proofs for Section 2.2 are included in Appendix D.
5. Auxiliary results on our assumptions are included in Appendix F.
6. Our proposed algorithms for improving decentralized learning with separable data are presented in Appendix G.
7. Finally, we conduct numerical experiments in Appendix H.

Notation

Throughout the appendix we use the following notations:

$$\bar{w} := \frac{1}{N}\sum_{\ell=1}^{N} w_\ell, \quad \bar{W} := [\bar{w}, \bar{w}, \cdots, \bar{w}]^\top \in \mathbb{R}^{N\times d},$$

$$W := [w_1, \cdots, w_N]^\top \in \mathbb{R}^{N\times d},$$

$$\hat{F}(W) := \frac{1}{N}\sum_{\ell=1}^{N} \hat{F}_\ell(w_\ell),$$

$$\nabla \hat{F}(w) := \frac{1}{n}\sum_{i=1}^{n} \nabla f(w, x_i),$$

$$\nabla \hat{F}(W) := [\nabla \hat{F}_1(w_1), \nabla \hat{F}_2(w_2), \cdots, \nabla \hat{F}_N(w_N)]^\top \in \mathbb{R}^{N\times d},$$

$$\bar{\nabla} \hat{F}(W) := \frac{1}{N}\sum_{\ell=1}^{N} \nabla \hat{F}_\ell(w_\ell),$$

$$\nabla \hat{F}_\ell(w_\ell) := \frac{N}{n}\sum_{x_j \in \mathcal{S}_\ell} \nabla f(w_\ell, x_j).$$

where recall that d is the dimension of ambient space, n is the total sample size, N is the number of agents and each agent has access to n/N samples.

Appendix A. Convergence under the PL condition

In this section, we extend our results to the case where the global loss satisfies the μ-PL condition. Formally, the PL condition [14, 21] is defined as follows,

Assumption 9 (PL condition) *The loss function $\hat{F} : \mathbb{R}^d \to \mathbb{R}$ satisfies the Polyak-Lojasiewic(PL) condition with the parameter μ:*

$$\|\nabla \hat{F}(w)\|^2 \geq 2\mu(\hat{F}(w) - \hat{F}^\star).$$

The next lemma shows that DGD enjoys an exponential rate under the PL condition and smoothness. See Appendix E.1 for a proof.

Lemma 8 (Train loss under the PL condition) *Let Assumptions 1,3 and 9 hold and let the step-size $\eta \leq \min(\frac{1-\alpha_1}{\mu}, \frac{1}{2L^2}\sqrt{\frac{(1-\alpha_1)\mu}{\alpha_2}}, \frac{1}{L})$, where the constants $\alpha_1 \in (0,1)$ and $\alpha_2 > 0$ depend only on the mixing matrix. Define $\zeta := 1 - \frac{\eta\mu}{2}$, then under the data separability assumption, the iterates of DGD satisfy for all $t \geq 1$,*

$$\hat{F}(\bar{w}^{(t)}) \leq \zeta^{t-1}\hat{F}(\bar{w}^{(1)}),$$
$$\|W^{(t)} - \bar{W}^{(t)}\|_F^2 \leq \frac{2\alpha_2\eta^2L^2\hat{F}(\bar{w}^{(1)})}{1-\alpha_1}\zeta^{t-1}.$$

We use this lemma combined with our key lemma 1 to obtain the test loss bound in the next theorem. The proof is provided in Appendix E.2.

Theorem 9 (Test loss under the PL condition) *Let Assumptions 1-4 and 9 hold, and let η and ζ be as in Lemma 8. Then the iterates of DGD satisfy for all $T \geq 1$,*

$$\mathbb{E}[F(\bar{w}^{(T)})] = \mathcal{O}\Big(\zeta^T + \frac{L^2c^2}{n^{3-2\alpha}\mu^{2\alpha}}(\eta T)^{2-2\alpha} + \frac{\eta^2L^4}{\mu^2}\Big). \tag{8}$$

Remark 10 *The bound in (8) is directly related to $(\eta T)^{2-2\alpha}$. When $\alpha < 1$, as in the case of smooth functions such as highly over-parameterized Least-squares $f(w,x) = \|1 - w^\top x\|^2$ where $d >> n$, the bound becomes vacuous as it is increasing with T. This indicates the existence of overfitting in DGD under such scenarios; with the optimal value of T achieved at the very early steps of gradient descent.*

Appendix B. Proof of Lemma 1

Lemma 11 (Formal statement of Lemma 1) *Consider the iterates of decentralized gradient descent in Eq.(2) with a fixed positive step-size $\eta \leq \frac{2}{L}$. Let Assumptions 1-4 hold. Then for the test loss F at iteration $T \geq 1$, it holds that*

$$\mathbb{E}[F(\bar{w}^{(T)})] \leq 4\mathbb{E}[\hat{F}(\bar{w}^{(T)})] + \frac{9L^2c^2\eta^2T^2}{n^{3-2\alpha}}\mathbb{E}[(\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}))^{2\alpha}] \tag{9}$$
$$+ \frac{9L^4\eta^2}{Nn}\sum_{i=1}^{n}\mathbb{E}[(\sum_{t=1}^{T}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F)^2] + \frac{9L^4\eta^2}{N}\mathbb{E}[(\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F)^2].$$

where the expectation is over training samples and $W_{\neg i}^{(t)}, \bar{W}_{\neg i}^{(t)}$ denotes the parameter matrix and averaged parameter matrix at iteration t for the DGD algorithm when the ith data sample is left out.

Proof We define $w_{\ell,\neg i}^{(t)}$ as the parameter of agent ℓ resulting from decentralized gradient descent at iteration t, when the ith training sample $i \leq n$, is left out during training. We emphasize that the ith sample may or may not belong to the dataset of agent ℓ.

We define $\bar{w}_{\neg i}^{(t)} \in \mathbb{R}^d$

$$\bar{w}_{\neg i}^{(t)} := \frac{1}{N}\sum_{j=1}^{N} w_{j,\neg i}^{(t)},$$

as the average of all agents' parameters at iteration t, when the ith sample is left of the algorithm. Thus, the parameter matrices $W_{\neg i}^{(t)}, \bar{W}_{\neg i}^{(t)} \in \mathbb{R}^{N\times d}$ are defined as follows,

$$\begin{aligned} W_{\neg i}^{(t)} &:= [w_{1,\neg i}^{(t)}, w_{2.\neg i}^{(t)}, \cdots, w_{N,\neg i}^{(t)}], \\ \bar{W}_{\neg i}^{(t)} &:= [\bar{w}_{\neg i}^{(t)}, \bar{w}_{\neg i}^{(t)}, \cdots, \bar{w}_{\neg i}^{(t)}]. \end{aligned}$$

The first step in the proof is to bound the term $\frac{1}{n}\sum_{i=1}^{n}\|\bar{w}^{(t)} - \bar{w}_{\neg i}^{(t)}\|^2$. By definition of DGD in Eq.(2), we have the following update rule for the averaged parameter,

$$\bar{w}^{(t+1)} = \bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(W^{(t)}).$$

Thus by adding and subtracting $\bar{\nabla}\hat{F}(\bar{W}^{(t)})$ and $\bar{\nabla}\hat{F}(\bar{W}_{\neg i}^{(t)})$, we have

$$\begin{aligned} &\|\bar{w}^{(t+1)} - \bar{w}_{\neg i}^{(t+1)}\| \\ &= \left\|\bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(W^{(t)}) - (\bar{w}_{\neg i}^{(t)} - \eta\bar{\nabla}\hat{F}(W_{\neg i}^{(t)}))\right\| \\ &= \Big\|\bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(\bar{W}^{(t)}) + \eta(\bar{\nabla}\hat{F}(W^{(t)}) - \bar{\nabla}\hat{F}(\bar{W}^{(t)})) \\ &\qquad\qquad - (\bar{w}_{\neg i}^{(t)} - \eta\nabla\hat{F}(\bar{W}_{\neg i}^{(t)})) + \eta\bar{\nabla}\hat{F}(W_{\neg i}^{(t)}) - \eta\nabla\hat{F}(\bar{W}_{\neg i}^{(t)})\Big\| \\ &\leq \left\|\bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(\bar{W}^{(t)}) - (\bar{w}_{\neg i}^{(t)} - \eta\bar{\nabla}\hat{F}(\bar{W}_{\neg i}^{(t)}))\right\| \\ &\qquad\qquad + \eta\left\|\bar{\nabla}\hat{F}(W_{\neg i}^{(t)}) - \nabla\hat{F}(\bar{W}_{\neg i}^{(t)})\right\| + \eta\left\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{W}^{(t)})\right\| \end{aligned}$$

For the last term, using smoothness, we can write

$$\begin{aligned}\left\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{W}^{(t)})\right\| &= \frac{1}{N}\Big\|\sum_{\ell=1}^{N}\nabla\hat{F}_\ell(w_\ell^{(t)}) - \nabla\hat{F}_\ell(\bar{w}^{(t)})\Big\| \\ &\le \frac{1}{N}\sum_{\ell=1}^{N}\left\|\nabla\hat{F}_\ell(w_\ell^{(t)}) - \nabla\hat{F}_\ell(\bar{w}^{(t)})\right\| \\ &\le \frac{L}{N}\sum_{\ell=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\| \\ &\le \frac{L}{\sqrt{N}}(\sum_{\ell=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\|^2)^{1/2} \\ &= \frac{L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F\end{aligned}$$

The second term is upper-bounded similarly. Using these bounds, splitting the gradient $\nabla\hat{F}(\bar{W}^{(t)}) = \frac{1}{n}\sum_{i'\neq i} f(\bar{w}^{(t)}, x_{i'}) + \frac{1}{n} f(\bar{w}^{(t)}, x_i)$, using smoothness and convexity $\|w+\eta\nabla f(w) - v - \eta\nabla f(v)\| \le \|w - v\|$ for $\eta \le 2/L$ [20] and employing Assumption 4 we can write

$$\begin{aligned}&\|\bar{w}^{(t+1)} - \bar{w}_{\neg i}^{(t+1)}\| \\ &\le \|\bar{w}^{(t)} - \eta\nabla\hat{F}(\bar{W}^{(t)}) - (\bar{w}_{\neg i}^{(t)} - \eta\nabla\hat{F}(\bar{W}_{\neg i}^{(t)}))\| \\ &\qquad + \frac{\eta L}{\sqrt{N}}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F \\ &\le \frac{1}{n}\sum_{i'\neq i}\|\bar{w}^{(t)} - \eta\nabla f(\bar{w}^{(t)}, x_{i'}) - \bar{w}_{\neg i}^{(t)} + \eta\nabla f(\bar{w}_{\neg i}^{(t)}, x_{i'})\| + \frac{1}{n}\|\bar{w}^{(t)} - \eta\nabla f(\bar{w}^{(t)}, x_i) - \bar{w}_{\neg i}^{(t)}\| \\ &\qquad + \frac{\eta L}{\sqrt{N}}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F \\ &\le \|\bar{w}^{(t)} - \bar{w}_{\neg i}^{(t)}\| + \frac{\eta}{n}\|\nabla f(\bar{w}^{(t)}, x_i)\| + \frac{\eta L}{\sqrt{N}}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F \\ &\le \|\bar{w}^{(t)} - \bar{w}_{\neg i}^{(t)}\| + \frac{c\eta}{n}\,(f(\bar{w}^{(t)}, x_i))^\alpha + \frac{\eta L}{\sqrt{N}}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F + \frac{\eta L}{\sqrt{N}}\|W^{(t)} - \bar{W}^{(t)}\|_F.\end{aligned}$$

By summing over $t \in [T]$,

$$\begin{aligned}\|\bar{w}^{(T+1)} - \bar{w}_{\neg i}^{(T+1)}\| \le \frac{c\eta}{n}\sum_{t=1}^{T}(f(\bar{w}^{(t)}, x_i))^\alpha &+ \frac{\eta L}{\sqrt{N}}\sum_{t=1}^{T}\|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F \\ &+ \frac{\eta L}{\sqrt{N}}\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F.\end{aligned}$$

We define for the ease of notation the following two consensus terms,

$$e^{(T)} := (\sum_{t=1}^{T} \|W^{(t)} - \bar{W}^{(t)}\|_F)^2$$
$$e_i^{(T)} := (\sum_{t=1}^{T} \|W_{\neg i}^{(t)} - \bar{W}_{\neg i}^{(t)}\|_F)^2$$

Thus the bound for the squared term can be written as follows

$$\|\bar{w}^{(T+1)} - \bar{w}_{\neg i}^{(T+1)}\|^2 \leq \frac{3c^2\eta^2}{n^2}(\sum_{t=1}^{T}(f(\bar{w}^{(t)}, x_i))^\alpha)^2 + \frac{3\eta^2L^2}{N}e^{(T)} + \frac{3\eta^2L^2}{N}e_i^{(T)}.$$

By averaging over $i \in [n]$ and noting that $\alpha \in [1/2, 1]$ so that x^α is concave, we conclude that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\bar{w}^{(T+1)} - \bar{w}_{\neg i}^{(T+1)}\|^2 \\
&\quad\leq \frac{3c^2\eta^2}{n^3}\sum_{i=1}^{n}(\sum_{t=1}^{T}(f(\bar{w}^{(t)}, x_i))^\alpha)^2 + \frac{3\eta^2L^2}{Nn}\sum_{i=1}^{n}e_i^{(T)} + \frac{3\eta^2L^2}{N}e^{(T)} \\
&\quad\leq \frac{3c^2\eta^2T^2}{n^3}\sum_{i=1}^{n}(\frac{1}{T}\sum_{t=1}^{T}f(\bar{w}^{(t)}, x_i))^{2\alpha} + \frac{3\eta^2L^2}{Nn}\sum_{i=1}^{n}e_i^{(T)} + \frac{3\eta^2L^2}{N}e^{(T)} \\
&\quad\leq \frac{3c^2\eta^2T^{2(1-\alpha)}}{n^{3-2\alpha}}(\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}))^{2\alpha} + \frac{3\eta^2L^2}{Nn}\sum_{i=1}^{n}e_i^{(T)} + \frac{3\eta^2L^2}{N}e^{(T)}.
\end{aligned}$$

Thus we have for iteration T:

$$\frac{1}{n}\sum_{i=1}^{n}\|\bar{w}^{(T)} - \bar{w}_{\neg i}^{(T)}\|^2 \leq \frac{3c^2\eta^2T^2}{n^{3-2\alpha}}(\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}))^{2\alpha} + \frac{3\eta^2L^2}{Nn}\sum_{i=1}^{n}e_i^{(T)} + \frac{3\eta^2L^2}{N}e^{(T)}. \quad (10)$$

Next we use [23, lemma 7], which states that for the L-smooth loss f, the test error for a $w \in \mathbb{R}^d$ as the output of an algorithm with dataset size n, satisfies the following,

$$\mathbb{E}[F(w)] \leq 4\mathbb{E}[\hat{F}(w)] + \frac{3L^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|w - w_{\neg i}\|^2],$$

where expectations are taken over the training set $(x_1, x_2, \cdots, x_n)$. We replace w with $\bar{w}^{(T)}$ and by using (10),

$$\begin{aligned}
\mathbb{E}[F(\bar{w}^{(T)})] &\leq 4\mathbb{E}[\hat{F}(\bar{w}^{(T)})] + \frac{3L^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|\bar{w}^{(T)} - \bar{w}_{\neg i}^{(T)}\|^2] \\
&\leq 4\mathbb{E}[\hat{F}(\bar{w}^{(T)})] + \frac{9L^2c^2\eta^2T^2}{n^{3-2\alpha}}\mathbb{E}[(\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}))^{2\alpha}] \\
&\qquad + \frac{9L^4\eta^2}{Nn}\sum_{i=1}^{n}\mathbb{E}[e_i^{(T)}] + \frac{9L^4\eta^2}{N}\mathbb{E}[e^{(T)}].
\end{aligned}$$

This leads to (9) and completes the proof. ■

Finally, we note that we can combine the two terms for the the consensus error $e_i^{(T)}, e^{(T)}$, as the consensus error is not affected (order-wise) by removing a sample from the dataset. This can be seen by noting that our bounds on the consensus are not affected by the sample-size. Rather they depend crucially on the smoothness parameter L which upper-bounds the smoothness parameter of leave-one-out loss. This leads to the informal version of the lemma.

Appendix C. Proofs for Section 2.1

Lemma 12 (Recursions for the consensus error) *Let the step-size $\eta \leq (1-\lambda)/4L$ where $\lambda := \max((|\lambda_2(A)|, |\lambda_N(A)|))^2$. The consensus error of DGD under Assumptions 1,3 satisfies the following:*

$$\|W^{(t)} - \bar{W}^{(t)}\|_F^2 < \alpha_1 \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + \alpha_2 \eta^2 L^2 \hat{F}(\bar{w}^{(t-1)}), \tag{11}$$

where $\alpha_1 := \frac{3+\lambda}{4}, \alpha_2 := 4(\frac{2}{1-\lambda} - 1)$.

Proof Denoting $A^\infty := \lim_{t\to\infty} A^t = \frac{1}{N}\mathbf{1}\mathbf{1}^T$, it holds by Assumption 1,

$$\begin{aligned}
\|AW - \bar{W}\|_F^2 &= \|(A - A^\infty)(W - \bar{W})\|_F^2 \\
&= \sum_{i=1}^N \|(A - A^\infty)(W_i - \bar{W}_i)\|^2 \\
&\leq \sum_{i=1}^N \|A - A^\infty\|^2 \|W_i - \bar{W}_i\|^2 \\
&\leq \max(\lambda_2^2(A), \lambda_N^2(A)) \cdot \|W - \bar{W}\|_F^2,
\end{aligned} \tag{12}$$

where W_i is the i th column of W. By Assumption 1, $\lambda = \max((|\lambda_2(A)|, |\lambda_N(A)|))^2 < 1$. For the consensus error, we can write,

$$\begin{aligned}
\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &= \|W^{(t)} - \bar{W}^{(t-1)} - \bar{W}^{(t)} + \bar{W}^{(t-1)}\|_F^2 \\
&\leq \|W^{(t)} - \bar{W}^{(t-1)}\|_F^2 \\
&= \|AW^{(t-1)} - \eta \nabla \hat{F}(W^{(t-1)}) - \bar{W}^{(t-1)}\|_F^2 \\
&\leq (1+\beta)\|AW^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + (1+\beta^{-1})\eta^2 \|\nabla \hat{F}(W^{(t-1)})\|_F^2,
\end{aligned}$$

where the second step is due to $\|X - \bar{X}\|_F \leq \|X\|_F$ [9, 10]. The last line holds for any $\beta > 0$, due to $\|a+b\|^2 \leq (1+\beta^{-1})\|a\|^2 + (1+\beta^{-1})\|b\|^2$.

Based on this inequality and by noting (12) and using the $L-$smoothness assumption, we can deduce that,

$$\begin{aligned}
\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &\leq (1+\beta)\lambda \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + (1+\beta^{-1})\eta^2 \|\nabla \hat{F}(W^{(t-1)})\|_F^2 \\
&\leq (1+\beta)\lambda \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + 2(1+\beta^{-1})\eta^2 \|\nabla \hat{F}(W^{(t-1)}) - \nabla \hat{F}(\bar{W}^{(t-1)})\|_F^2 \\
&\qquad + 2(1+\beta^{-1})\eta^2 \|\nabla \hat{F}(\bar{W}^{(t-1)})\|_F^2 \qquad (13) \\
&\leq (1+\beta)\lambda \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + 2(1+\beta^{-1})\eta^2 L^2 \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 \\
&\qquad + 4(1+\beta^{-1})\eta^2 L \hat{F}(\bar{w}_{t-1}),
\end{aligned}$$

where the last step is due to $L-$smoothness and the non-negativity of $\hat{F}_\ell$.

$$\begin{aligned}\|\nabla \hat{F}(\bar{W}^{(t-1)})\|_F^2 &= \frac{1}{N^2}\sum_{\ell=1}^{N}\|\nabla \hat{F}_\ell(\bar{w}^{(t-1)})\|^2 \\ &\leq \frac{2L}{N^2}\sum_{\ell=1}^{N}(\hat{F}_\ell(\bar{w}^{(t-1)}) - \hat{F}_\ell^\star) \\ &\leq 2L\hat{F}(\bar{w}^{(t-1)}).\end{aligned}$$

Thus,

$$\begin{aligned}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 < ((1+\beta)\lambda + 2(1+\beta^{-1})\eta^2 L^2)\|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 \\ + 4(1+\beta^{-1})\eta^2 L\,\hat{F}(\bar{w}^{(t-1)}).\end{aligned}$$

choose $\beta = (1-\lambda)/(2\lambda)$, then by the assumption $\eta \leq (1-\lambda)/4L$, we can derive the following two inequalities,

$$\begin{aligned}&(1+\beta)\lambda + 2(1+\beta^{-1})\eta^2 L^2 < (3+\lambda)/4, \\ &4(1+\beta^{-1}) < 4(2/(1-\lambda) - 1).\end{aligned}$$

This concludes the lemma. ■

By telescoping summation over the iterates $t = 1, \cdots, T$ of the consensus error in Eq.(11), we end up with the consensus error at iteration T. The final expression is stated in the next lemma.

Lemma 13 *Under the assumptions of Lemma 12, it holds for $T > 1$ that,*

$$\|W^{(T)} - \bar{W}^{(T)}\|_F^2 < \alpha_1^{T-1}\|W_1 - \bar{W}_1\|_F^2 + (\alpha_2\eta^2 L)\sum_{t=1}^{T-1}\alpha_1^{t-1}\hat{F}(\bar{w}_{T-t}).$$

Lemma 14 *Under the assumptions of Lemma 12 and the zero initialization assumption for all agents, the average consensus error satisfies,*

$$\frac{1}{T}\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 \leq \frac{\alpha_2\eta^2 L}{(1-\alpha_1)T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}_t).$$

Proof By Lemma 13 and the zero initialization assumption, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &\leq \frac{\alpha_2\eta^2 L}{T}\sum_{t=2}^{T}\sum_{i=1}^{t-1}\alpha_1^{i-1}\hat{F}(\bar{w}_{t-i}) \\ &\leq \frac{\alpha_2\eta^2 L}{T}\sum_{i=1}^{T-1}\alpha_1^{i-1}\sum_{t=1}^{T-i}\hat{F}(\bar{w}^{(t)}) \\ &\leq \frac{\alpha_2\eta^2 L}{T}\sum_{i=1}^{T-1}\alpha_1^{i-1}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \\ &\leq \frac{\alpha_2\eta^2 L}{(1-\alpha_1)T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}_t).\end{aligned}$$

■

Lemma 15 *Under Assumptions 2,3 and for any $w \in \mathbb{R}^d$, the following recursions hold for the Decentralized gradient descent updates:*

$$\frac{2\eta - 4L\eta^2}{T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\|\bar{w}^{(1)} - w\|^2}{T} + 2\eta\hat{F}(w) + \frac{2L^2\eta^2 + \eta L}{T}\sum_{t=1}^{T-1}\|W^{(t)} - \bar{W}^{(t)}\|_F^2$$

Proof Starting by expanding, we get

$$\begin{aligned}\|\bar{w}^{(t+1)} - w\|^2 &= \|\bar{w}^{(t)} - \eta\bar{\nabla}\hat{F}(W^{(t)}) - w\|^2 \\ &= \|\bar{w}^{(t)} - w\|^2 + \eta^2\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 - 2\eta\langle\bar{w}^{(t)} - w, \bar{\nabla}\hat{F}(W^{(t)})\rangle\end{aligned}$$

For the second term, using the $L-$smoothness property and non-negativity of loss, we obtain the following:

$$\begin{aligned}\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 &= \|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{w}^{(t)}) + \nabla\hat{F}(\bar{w}^{(t)})\|^2 \\ &\leq 2\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{w}^{(t)})\|^2 + 2\|\nabla\hat{F}(\bar{w}^{(t)})\|^2 \\ &\leq \frac{2L^2}{N}\sum_{i=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\|^2 + 4L\hat{F}(\bar{w}^{(t)}).\end{aligned}$$

For the third term by using $L-$smoothness and convexity properties we can write,

$$\begin{aligned}\langle\bar{w}^{(t)} - w, \bar{\nabla}\hat{F}(W^{(t)})\rangle &= \frac{1}{N}\sum_{\ell=1}^{N}\langle\bar{w}^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\ &= \frac{1}{N}\sum_{\ell=1}^{N}\langle\bar{w}^{(t)} - w_\ell^{(t)}, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle + \frac{1}{n}\sum_{\ell=1}^{N}\langle w_\ell^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\ &\geq \frac{1}{N}\sum_{\ell=1}^{N}\hat{F}_\ell(\bar{w}^{(t)}) - \hat{F}_\ell(w_\ell^{(t)}) - \frac{L}{2}\|w_\ell^{(t)} - \bar{w}^{(t)}\|^2 + \frac{1}{N}\sum_{\ell=1}^{N}\langle w_\ell^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\ &\geq \frac{1}{N}\sum_{\ell=1}^{N}\hat{F}_\ell(\bar{w}^{(t)}) - \hat{F}_\ell(w_\ell^{(t)}) - \frac{L}{2}\|w_\ell^{(t)} - \bar{w}^{(t)}\|^2 + \frac{1}{N}\sum_{\ell=1}^{N}\hat{F}_\ell(w_\ell^{(t)}) - \hat{F}_\ell(w) \\ &= \hat{F}(\bar{w}^{(t)}) - \hat{F}(w) - \frac{L}{2}\|W^{(t)} - \bar{W}^{(t)}\|_F^2.\end{aligned}$$

Combining these inequalities we derive the following:

$$\begin{aligned}\|\bar{w}^{(t+1)} - w\|^2 \leq \|\bar{w}^{(t)} - w\|^2 &+ \eta^2(2L^2\|W^{(t)} - \bar{W}^{(t)}\|_F^2 + 4L\hat{F}(\bar{w}^{(t)})) \\ &- 2\eta(\hat{F}(\bar{w}^{(t)}) - \hat{F}(w) - \frac{L}{2}\|W^{(t)} - \bar{W}^{(t)}\|_F^2).\end{aligned}$$

Summing these equations for $t = 1, 2, ..., T-1$,

$$\|\bar{w}^{(T)} - w\|^2 \leq \|\bar{w}^{(1)} - w\|^2 + \sum_{t=1}^{T-1}(2L^2\eta^2 + \eta L)\|W^{(t)} - \bar{W}^{(t)}\|_F^2 + \sum_{t=1}^{T-1}(4L\eta^2 - 2\eta)\hat{F}(\bar{w}^{(t)}) + 2\sum_{t=1}^{T-1}\eta\hat{F}(w)$$

$$\frac{2\eta - 4L\eta^2}{T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\|\bar{w}^{(1)} - w\|^2}{T} + \frac{2L^2\eta^2 + \eta L}{T}\sum_{t=1}^{T-1}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 + 2\eta\hat{F}(w).$$

■

C.1. Proof of Lemma 2

Lemma 16 (Restatement of Lemma 2) *Under Assumptions 1-4 and zero initialization, for any $w \in \mathbb{R}^d$ and for a fixed step-size $\eta < \min(\frac{1-\alpha_1}{L}, \frac{1}{L}\sqrt{\frac{1-\alpha_1}{2\alpha_2}})$, where $\alpha_1 \in (0,1), \alpha_2 > 0$ are parameters that depend only on the mixing matrix, the following holds for the train loss and consensus error of DGD (2):*

$$\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}) \leq \frac{2\|w\|^2}{\eta T} + 4\hat{F}(w), \tag{14}$$

$$\frac{1}{T}\sum_{t=1}^{T}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 \leq \frac{\alpha_2\eta^2 L}{(1-\alpha_1)}\left(\frac{2\|w\|^2}{\eta T} + 4\hat{F}(w)\right). \tag{15}$$

Proof We deduce from Lemma 15 that,

$$\frac{1}{T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\|w\|^2}{\eta T} + 2\hat{F}(w) + \frac{2L^2\eta + L}{T}\sum_{t=1}^{T-1}\|W^{(t)} - \bar{W}^{(t)}\|_F^2$$

By Lemma 14,

$$\frac{1}{T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\|w\|^2}{\eta T} + 2\hat{F}(w) + \frac{(2L^2\eta + L)\alpha_2\eta^2 L}{T(1-\alpha_1)}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}) \leq \frac{\|w\|^2}{\eta T} + 2\hat{F}(w) + \frac{1}{2T}\sum_{t=1}^{T-1}\hat{F}(\bar{w}^{(t)}).$$

where the condition on η ensures that $(2L^2\eta + L)\alpha_2\eta^2 L/(1-\alpha_1) < 1/2$. This gives the statement of the lemma for the training loss in Eq.(14). Appealing again to Lemma 14 for the consensus error yields (15). ■

C.2. Proof of Theorem 4

We are ready to prove Theorem 4 by combining our results from Lemmas 2 and 1. We state the proof for general choice of step-size η. In particular, Theorem 4 follows by the next theorem after choosing $\eta = \mathcal{O}(1/\sqrt{T})$.

Theorem 17 (Theorem 4 for general η) *Consider DGD under Assumptions 1-6, and choose $\eta < \min(\frac{1-\alpha_1}{L}, \frac{1}{L}\sqrt{\frac{1-\alpha_1}{2\alpha_2}})$. The following bound holds for the averaged test error of DGD with separable data up to iteration T, assuming $\varepsilon \leq \rho(\varepsilon)^2/\eta T$,*

$$\frac{1}{T}\sum_{t=1}^{T} F(\bar{w}^{(t)}) = \mathcal{O}(\frac{\rho(\varepsilon)^2}{\eta T} + \frac{L^2c^2\rho(\varepsilon)^{4\alpha}}{n^{3-2\alpha}}(\eta T)^{2-2\alpha} + \frac{L^4\rho(\varepsilon)^4}{N}\eta^3 T).$$

Proof *By Lemma 1,*

$$\begin{aligned}&\mathbb{E}[F(\bar{w}^{(t)})]\\ &\quad= \mathcal{O}(\mathbb{E}[\hat{F}(\bar{w}^{(t)})] + \frac{L^2c^2\eta^2t^2}{n^{3-2\alpha}}\mathbb{E}[(\frac{1}{t}\sum_{i=1}^{t}\hat{F}(\bar{w}^{(i)}))^{2\alpha}] + \frac{L^4\eta^2}{N}\mathbb{E}[(\sum_{i=1}^{t}\|W^{(i)} - \bar{W}^{(i)}\|)^2]).\end{aligned}$$

Thus, by Lemma 2,

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\bar{w}^{(t)})]\\ &= \mathcal{O}(\frac{\|w\|^2}{\eta T} + \hat{F}(w) + \frac{L^2c^2\eta^2}{n^{3-2\alpha}}\frac{1}{T}\sum_{t=1}^{T}t^2(\frac{\|w\|^2}{\eta t} + \hat{F}(w))^{2\alpha} + \frac{L^4\eta^4}{N}\frac{1}{T}\sum_{t=1}^{T}t^2(\frac{\|w\|^2}{\eta t} + \hat{F}(w))).\end{aligned}$$

By Assumption 6 and assuming $\varepsilon \leq \rho(\varepsilon)^2/\eta T$, the statement of the theorem follows. ■

Appendix D. Proofs for Section 2.2

Lemma 18 (Iterates of consensus error) *Consider DGD with the loss functions and mixing matrix satisfying Assumptions 1,2, 7 and Assumption 4 with $\alpha = 1$ and $c = h$. By choosing $\eta_{t-1} \leq \frac{1-\lambda}{4hM_{(t-1)}}$ the consensus error at iteration $t > 1$ satisfies*

$$\|W^{(t)} - \bar{W}^{(t)}\|_F^2 < \beta_1\|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 + \beta_2\eta_{t-1}^2h^2\hat{F}^2(\bar{w}^{(t-1)}) \tag{16}$$

where $\beta_1 := (3+\lambda)/4, \beta_2 := (4/(1-\lambda) - 2)$, $\lambda = \max(\lambda_2(A)^2, \lambda_N(A)^2)$ and $M_{(t-1)} := \max(\hat{F}(W^{(t-1)}), \hat{F}(\bar{w}^{(t-1)}))$.

Proof By Lemma 12 and the inequality (13), the consensus error satisfies for any $\beta > 0$,

$$\begin{aligned}\|W^{(t)} - \bar{W}^{(t)}\|_F^2 &< (1+\beta)\lambda\|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 \\ &+ 2(1+\beta^{-1})\eta_{t-1}^2\|\nabla\hat{F}(W^{(t-1)}) - \nabla\hat{F}(\bar{W}^{(t-1)})\|_F^2 + 2(1+\beta^{-1})\eta_{t-1}^2\|\nabla\hat{F}(\bar{W}^{(t-1)})\|_F^2\end{aligned} \tag{17}$$

For the second term in (17), we have the following chain of inequalities,

$$
\begin{aligned}
&\|\nabla \hat{F}(W^{(t-1)}) - \nabla \hat{F}(\bar{W}^{(t-1)})\|_F^2 \\
&\quad = \frac{1}{N^2}\sum_{\ell=1}^{N} \|\nabla \hat{F}_\ell(w_i^{(t-1)}) - \nabla \hat{F}_\ell(\bar{w}^{(t-1)})\|^2 \\
&\quad \leq \frac{1}{N^2}\sum_{\ell=1}^{N} \max_{v_\ell \in [w_\ell^{(t-1)}, \bar{w}^{(t-1)}]} \|\nabla^2 \hat{F}_\ell(v_\ell)\|^2 \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \qquad (18) \\
&\quad \leq \frac{h^2}{N^2}\sum_{\ell=1}^{N} \max_{v_\ell \in [w_\ell^{(t-1)}, \bar{w}^{(t-1)}]} (\hat{F}_\ell(v_\ell))^2 \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \qquad (19) \\
&\quad = \frac{h^2}{N^2}\sum_{\ell=1}^{N} (\max_{v_\ell \in [w_\ell^{(t-1)}, \bar{w}^{(t-1)}]} \hat{F}_\ell(v_\ell))^2 \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \\
&\quad \leq \frac{h^2}{N^2}\sum_{\ell=1}^{N} \max(\hat{F}_\ell^2(w_\ell^{(t-1)}), \hat{F}_\ell^2(\bar{w}^{(t-1)})) \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \qquad (20) \\
&\quad \leq \frac{h^2}{N^2} \max(\max_{k \leq N} \hat{F}_k^2(w_k^{(t-1)}), \max_{k \leq N} \hat{F}_k^2(\bar{w}^{(t-1)})) \sum_{\ell=1}^{N} \|w_\ell^{(t-1)} - \bar{w}^{(t-1)}\|^2 \qquad (21) \\
&\quad \leq h^2 M_{(t-1)}^2 \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2.
\end{aligned}
$$

The Taylor's remainder theorem gives (18) and $v_i \in [w_i^{(t-1)}, \bar{w}^{(t-1)}]$ denotes a point that lies on the line connecting $w_i^{(t-1)}$ and $\bar{w}^{(t-1)}$. Also, (19) is valid due to the self-boundedness of the Hessian stated in Assumption 7. The inequality (20) follows by the assumption of convexity of $\hat{F}_i$, due to the fact that for a convex function $f : \mathbb{R}^d \to \mathbb{R}$ and any two points $w_1, w_2 \in \mathbb{R}^d$, it holds that $\max_{v \in [w_1, w_2]} f(v) \leq \max(f(w_1), f(w_2))$. To derive (22), we used $\max_{i \leq N} \hat{F}_i(w_i) \leq N\hat{F}(W)$ and $\max_{i \leq N} \hat{F}_i(\bar{w}) \leq N \cdot \hat{F}(\bar{w})$, which hold since the loss functions are non-negative.

In order to derive an upper-bound on the last term in (17), we use Assumption 4,

$$
\begin{aligned}
\|\nabla \hat{F}(\bar{W}^{(t-1)})\|_F^2 &= \frac{1}{N^2}\sum_{\ell=1}^{N} \|\nabla \hat{F}_\ell(\bar{w}^{(t-1)})\|^2 \\
&\leq \frac{h^2}{N^2}\sum_{\ell=1}^{N} (\hat{F}_\ell(\bar{w}^{(t-1)}))^2 \\
&\leq h^2 \hat{F}^2(\bar{w}^{(t-1)}).
\end{aligned}
$$

Replacing the upper-bounds back in (17), we conclude

$$
\begin{aligned}
\|W^{(t)} - \bar{W}^{(t)}\|_F^2 < ((1+\beta)\lambda + 2(1+\beta^{-1})\eta_{t-1}^2 h^2 M_{(t-1)}^2) \|W^{(t-1)} - \bar{W}^{(t-1)}\|_F^2 \\
+ 2(1+\beta^{-1})\eta_{t-1}^2 h^2 \hat{F}(\bar{w}^{(t-1)}).
\end{aligned}
$$

Choose $\beta = \frac{1-\lambda}{2\lambda}$. Then by the assumption of lemma $\eta_{t-1} \leq \frac{1-\lambda}{4hM_{(t-1)}}$, it can be verified that the following two inequalities hold,

$$(1+\beta)\lambda + 2(1+\beta^{-1})\eta_{t-1}^2 h^2 M_{(t-1)}^2 < (3+\lambda)/4,$$
$$2(1+1/\beta) < 2(2/(1-\lambda)-1).$$

This concludes the proof. ■

The next lemma bounds the consensus error at iteration T, by recursively evaluating (16).

Lemma 19 (Last iterate consensus error) *Under the assumptions and notations of Lemma 18, the consensus error at iteration T satisfies*

$$\|W^{(T)} - \bar{W}^{(T)}\|_F^2 < \beta_1^T \|W^{(1)} - \bar{W}^{(1)}\|_F^2 + \beta_2 h^2 \sum_{t=1}^{T-1} \beta_1^{t-1} \eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}).$$

Lemma 20 *Under the assumptions and notations of Lemma 18, with zero initialization $W^{(1)} = \bar{W}^{(1)} = 0$ and by choosing $\eta_t \leq \frac{(1-\lambda)\sqrt{1-\beta_1}}{8h^2 M_{(t)}\sqrt{\beta_2}}$ for $t \in \{1, \cdots, T-1\}$, it holds at iteration T that*

$$\frac{1}{2}\hat{F}(\bar{w}^{(T)}) \leq \hat{F}(W^{(T)}) \leq 2\hat{F}(\bar{w}^{(T)}). \tag{22}$$

Proof *First, we prove $\hat{F}(W^{(T)}) \leq 2\hat{F}(\bar{w}^{(T)})$. If $\hat{F}(W^{(T)}) \leq \hat{F}(\bar{w}^{(T)})$, the desired statement is proved. Assume $\hat{F}(W^{(T)}) \geq \hat{F}(\bar{w}^{(T)})$, then by applying Taylor's remainder theorem, the self-boundedness assumption 4 with $c = h, \alpha = 1$, convexity of $\hat{F}$, Lemma 19 and the restriction on the step-size, in respective order, we have the following inequalities,*

$$\begin{aligned}
\hat{F}(W^{(T)}) &\leq |\hat{F}(W^{(T)}) - \hat{F}(\bar{w}^{(T)})| + \hat{F}(\bar{w}^{(T)}) \\
&\leq \max_{v \in [\bar{W}^{(T)}, W^{(T)}]} \|\nabla \hat{F}(v)\| \cdot \|W^{(T)} - \bar{W}^{(T)}\| + \hat{F}(\bar{w}^{(T)}) \\
&\leq h \cdot \max_{v \in [\bar{W}^{(T)}, W^{(T)}]} \hat{F}(v) \cdot \|W^{(T)} - \bar{W}^{(T)}\| + \hat{F}(\bar{w}^{(T)}) \\
&\leq h \cdot \max(\hat{F}(W^{(T)}), \hat{F}(\bar{w}^{(T)})) \cdot \|W^{(T)} - \bar{W}^{(T)}\| + \hat{F}(\bar{w}^{(T)}) \\
&\leq \hat{F}(W^{(T)})(\beta_2 h^4 \sum_{t=1}^{T-1} \beta_1^{t-1} \eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}))^{1/2} + \hat{F}(\bar{w}^{(T)}) \\
&\leq \frac{1}{2}\hat{F}(W^{(T)}) + \hat{F}(\bar{w}^{(T)}).
\end{aligned}$$

Thus $\hat{F}(W^T) \leq 2\hat{F}(\bar{w}^{(T)})$. By exchanging $W^{(T)}$ and $\bar{w}^{(T)}$ and in a similar style we derive $\hat{F}(W^T) \geq \frac{1}{2}\hat{F}(\bar{w}^{(T)})$. This completes the proof of the lemma. ■

We are ready to prove Theorem 6. First, we prove that DGD is a descent algorithm in the next lemma.

Lemma 21 (descent lemma) *Consider DGD under the assumptions and notations of Lemma 18.Moreover, let Assumption 8 hold, then by choosing* $\eta_t \leq \frac{\gamma}{\hat{F}(\bar{w}^{(t)})}$ *for* $t \leq T$*, where* $\gamma := (\max(\frac{4h^3}{\tau^2}, h^2, \frac{6h^2\beta_2}{1-\beta_1}, \frac{4h^2\sqrt{\beta_2}}{\tau(1-\beta_1)}))^{-1}$*, DGD is a descent algorithm i.e.,*

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}).$$

Proof With the self-boundedness assumption on the Hessian (Assumption 7) and applying the Taylor's remainder theorem for step $T+1$ of DGD, we obtain the following,

$$\begin{aligned}
&\hat{F}(\bar{w}^{(T+1)}) \\
&\quad \leq \hat{F}(\bar{w}^{(T)}) + \langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{w}^{(t+1)} - \bar{w}^{(T)} \rangle + \frac{1}{2} \max_{v \in [\bar{w}^{(T)}, \bar{w}^{(T+1)}]} \|\nabla^2 \hat{F}(v)\| \|\bar{w}^{(T+1)} - \bar{w}^{(T)}\|^2 \\
&\quad \leq \hat{F}(\bar{w}^{(T)}) - \eta_T \langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{\nabla} \hat{F}(W^{(T)}) \rangle + \frac{\eta_T^2}{2} \max_{v \in [\bar{w}^{(T)}, \bar{w}^{(T+1)}]} \|\nabla^2 \hat{F}(v)\| \|\bar{\nabla} \hat{F}(W^{(T)})\|^2 \\
&\quad \leq \hat{F}(\bar{w}^{(T)}) - \eta_T \langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{\nabla} \hat{F}(W^{(T)}) \rangle + \frac{h\eta_T^2}{2} \max_{v \in [\bar{w}^{(T)}, \bar{w}^{(T+1)}]} \hat{F}(v) \|\bar{\nabla} \hat{F}(W^{(T)})\|^2 \\
&\quad \leq \hat{F}(\bar{w}^{(T)}) - \eta_T \langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{\nabla} \hat{F}(W^{(T)}) \rangle + \frac{h\eta_T^2}{2} \max(\hat{F}(\bar{w}^{(T)}), \hat{F}(\bar{w}^{(T+1)})) \|\bar{\nabla} \hat{F}(W^{(T)})\|^2,
\end{aligned} \tag{23}$$

where for the third step we used

$$\begin{aligned}
\|\nabla^2 \hat{F}(w)\| &\leq \frac{1}{N} \sum_{\ell=1}^{N} \|\nabla^2 \hat{F}_\ell(w)\| \\
&\leq \frac{h}{N} \sum_{\ell=1}^{N} \hat{F}_\ell(w) \\
&= h\, \hat{F}(w).
\end{aligned}$$

In the next step of the proof, we upper-bound the second and third terms in (23). For the second term, by noting that $2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, we can write

$$\langle \nabla \hat{F}(\bar{w}^{(T)}), \bar{\nabla} \hat{F}(W^{(T)}) \rangle = \frac{1}{2} \|\nabla \hat{F}(\bar{w}^{(T)})\|^2 + \frac{1}{2} \|\bar{\nabla} \hat{F}(W^{(T)})\|^2 - \frac{1}{2} \|\nabla \hat{F}(\bar{w}^{(T)}) - \bar{\nabla} \hat{F}(W^{(T)})\|^2. \tag{24}$$

By recalling (22) and Lemma 19, we find an upper-bound the last term in (24) as follows,

$$\begin{aligned}
\|\nabla \hat{F}(\bar{w}^{(T)}) - \bar{\nabla} \hat{F}(W^{(T)})\|^2 &\leq h^2 M_{(T)}^2 \|W^{(T)} - \bar{W}^{(T)}\|_F^2 \\
&\leq h^4 M_{(T)}^2 \beta_2 \sum_{t=1}^{T-1} \beta_1^{t-1} \eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}).
\end{aligned} \tag{25}$$

Returning back to (23), thus far we have derived the following,

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}) - \frac{\eta_T}{2}\|\nabla \hat{F}(\bar{w}^{(T)})\|^2 - \frac{\eta_T}{2}\|\bar{\nabla}\hat{F}(W^{(T)})\|^2 + \frac{1}{2}\eta_T h^4 M_{(T)}^2 \beta_2 \sum_{t=1}^{T-1} \beta_1^{t-1}\eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}) + \frac{h\eta_T^2}{2}\|\bar{\nabla}\hat{F}(W^{(T)})\|^2 \cdot \max(\hat{F}(\bar{w}^{(T)}), \hat{F}(\bar{w}^{(T+1)})). \tag{26}$$

We aim to prove that $\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)})$ for all $T \geq 1$. If $\hat{F}(\bar{w}^{(T+1)}) > \hat{F}(\bar{w}^{(T)})$, applying (26) with the assumption $\eta_t < \frac{\gamma}{\hat{F}(\bar{w}^{(t)})}$ yields,

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}) - \frac{\eta_T}{2}\|\nabla \hat{F}(\bar{w}^{(T)})\|^2 + \frac{1}{2(1-\beta_1)}\eta_T \gamma^2 h^4 M_{(T)}^2 \beta_2 + \frac{h\eta_T^2}{2}\|\bar{\nabla}\hat{F}(W^{(T)})\|^2 \cdot \hat{F}(\bar{w}^{(T+1)}). \tag{27}$$

Note that it holds due to (25) that,

$$\|\bar{\nabla}\hat{F}(W^{(T)})\|^2 \leq 2\|\nabla \hat{F}(\bar{w}^{(T)})\|^2 + 2\|\bar{\nabla}\hat{F}(W^{(T)}) - \nabla \hat{F}(\bar{w}^{(T)})\|^2 \leq 2\|\nabla \hat{F}(\bar{w}^{(T)})\|^2 + \frac{2}{1-\beta_1}\gamma^2 h^4 M_{(T)}^2 \beta_2.$$

Replacing this in (27) and noting that $M_{(T)} \leq 2\hat{F}(\bar{w}^{(T)})$ by Lemma 20, we can simplify the inequality (27) as follows,

$$\begin{aligned}
&\hat{F}(\bar{w}^{(T+1)}) \\
&\leq \hat{F}(\bar{w}^{(T)}) + \eta_T\|\nabla \hat{F}(\bar{w}^{(T)})\|^2(h\eta_T \hat{F}(\bar{w}^{(T+1)}) - \frac{1}{2}) + C' h^2 \gamma^2 \eta_T \hat{F}^2(\bar{w}^{(T)})\,(1 + \eta_T \hat{F}(\bar{w}^{(T+1)})) \\
&\leq \hat{F}(\bar{w}^{(T)}) + \eta_T h^2 \hat{F}^2(\bar{w}^{(T)})((h\eta_T + \gamma^2\eta_T C')\hat{F}(\bar{w}^{(T+1)}) - \frac{\tau^2}{2h^2} + \gamma^2 C'),
\end{aligned}$$

where for the ease of notation we define $C' := 4h^2\beta_2(1-\beta_1)^{-1}$. Recalling $\eta_T < \frac{\gamma}{\hat{F}(\bar{w}^{(T)})}$ and noting that by the assumption of the lemma $\gamma \leq \frac{\tau^2}{4h^3}$ and $\gamma < \frac{1}{h^2}$ and $\gamma < \frac{\tau}{2h\sqrt{C'}}$ we conclude that,

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}) + \frac{\tau}{2h}\hat{F}(\bar{w}^{(T)})(\frac{\hat{F}(\bar{w}^{(T+1)})}{\hat{F}(\bar{w}^{(T)})} - 1).$$

Dividing both sides by $\hat{F}(\bar{w}^{(T)})$ leads to the contradiction due to the fact that $\tau \leq h$ and thus $\tau/2h < 1$. Thus $\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)})$. This completes the proof. ■

D.1. Proof of Theorem 6

Theorem 22 (Restatement of Theorem 6) *Consider DGD with the loss functions and mixing matrix satisfying Assumptions 1, 2, 7, 8 and Assumption 4 with $\alpha = 1$ and $c = h$. Assume that the*

step-size satisfies $\eta < \frac{\gamma}{\hat{F}(1)}$*, for a constant* $\gamma := (\max(\frac{4h^3}{\tau^2}, h^2, \frac{6h^2\beta_2}{1-\beta_1}, \frac{4h^2\sqrt{\beta_2}}{\tau(1-\beta_1)}))^{-1}$*, where* β_1, β_2 *are defined same as in Lemma 18. Then DGD is a descent algorithm i.e, for all* $T \geq 1$ *it holds that*

$$\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)}).$$

Moreover, the train loss and the consensus error of DGD at iteration T *satisfy the following for all* $w \in \mathbb{R}^d$*,*

$$\hat{F}(\bar{w}^{(T)}) \leq 4\hat{F}(w) + \frac{2\|w\|^2}{\eta T}, \tag{28}$$

$$\|W^{(T)} - \bar{W}^{(T)}\|_F^2 = \mathcal{O}(\eta^2 \hat{F}^2(w) + \frac{\|w\|^4}{T^2}). \tag{29}$$

Proof First, we note that by Lemma 21, under the assumption $\eta_t \leq \gamma/\hat{F}(\bar{w}^{(t)})$ for $t \leq T$, we have $\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)})$. Thus fixing $\eta \leq \gamma/\hat{F}(\bar{w}^{(1)})$, ensures that $\hat{F}(\bar{w}^{(T+1)}) \leq \hat{F}(\bar{w}^{(T)})$, for all T.

Next, we derive the train loss and consensus error under the assumptions of the theorem.

$$\|\bar{w}^{(t+1)} - w\|^2 = \|\bar{w}^{(t)} - w\|^2 + \eta^2\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 - 2\eta\langle \bar{w}^{(t)} - w, \bar{\nabla}\hat{F}(W^{(t)})\rangle \tag{30}$$

For the second term, we can write,

$$\begin{aligned}
\|\bar{\nabla}\hat{F}(W^{(t)})\| &= \frac{1}{n}\|\sum_{i=1}^{n} \nabla\hat{F}_i(w_i^{(t)})\| \\
&\leq \frac{h}{n}\sum_{i=1}^{n} \hat{F}_i(w_i^{(t)}) \\
&= h\hat{F}(W^{(t)}).
\end{aligned}$$

For the third term in (30), we have,

$$\begin{aligned}
-\langle \bar{w}^{(t)} - w, \bar{\nabla}\hat{F}(W^{(t)})\rangle &= -\frac{1}{N}\sum_{\ell=1}^{N}\langle \bar{w}^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\
&= -\frac{1}{N}\sum_{\ell=1}^{N}\langle \bar{w}^{(t)} - w_\ell^{(t)}, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle - \frac{1}{N}\sum_{\ell=1}^{N}\langle w_\ell^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\
&\leq \frac{1}{N}\sum_{\ell=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\|\|\nabla\hat{F}_i(w_\ell^{(t)})\| - \frac{1}{N}\sum_{\ell=1}^{N}\langle w_\ell^{(t)} - w, \nabla\hat{F}_\ell(w_\ell^{(t)})\rangle \\
&\leq \frac{1}{N}\sum_{\ell=1}^{N}\|w_\ell^{(t)} - \bar{w}^{(t)}\|\|\nabla\hat{F}_\ell(w_\ell^{(t)})\| + \frac{1}{N}\sum_{\ell=1}^{N}(\hat{F}_\ell(w) - \hat{F}_\ell(w_\ell^{(t)})) \qquad (31) \\
&\leq \frac{h}{N}\sum_{\ell=1}^{N}\hat{F}_\ell(w_\ell^{(t)})\|w_\ell^{(t)} - \bar{w}^{(t)}\| + \frac{1}{N}\sum_{\ell=1}^{N}(\hat{F}_\ell(w) - \hat{F}_\ell(w_\ell^{(t)})) \qquad (32) \\
&\leq h\hat{F}(W^{(t)})\|W^{(t)} - \bar{W}^{(t)}\|_F \ + \ \hat{F}(w) - \hat{F}(W^{(t)}).
\end{aligned}$$

Here (31) follows by convexity of $\hat{F}_i$, and (32) follows by the assumption on self-boundedness of the gradient.

Thus, the inequality (30) can be written as follows,

$$\|\bar{w}^{(t+1)} - w\|^2 \leq \|\bar{w}^{(t)} - w\|^2 + \eta^2 h^2 \hat{F}^2(W^{(t)}) + 2\eta h \hat{F}(W^{(t)}) \|W^{(t)} - \bar{W}^{(t)}\|_F + 2\eta \hat{F}(w) - 2\eta \hat{F}(W^{(t)}). \tag{33}$$

Moreover, by Lemma 19 and the assumption on η,

$$\|W^{(t)} - \bar{W}^{(t)}\|_F \leq (\beta_2 h^2 \sum_{t=1}^{T-1} \beta_1^{t-1} \eta_{T-t}^2 \hat{F}^2(\bar{w}_{T-t}))^{1/2} \leq \frac{1}{4h}$$

and

$$\eta \hat{F}(W^{(t)}) \leq \frac{1}{2h^2}.$$

Thus (33) changes into,

$$\|\bar{w}^{(t+1)} - w\|^2 \leq \|\bar{w}^{(t)} - w\|^2 - \eta \hat{F}(W^{(t)}) + 2\eta \hat{F}(w)$$

Telescoping sum leads to

$$\frac{1}{T} \sum_{t=1}^{T} \hat{F}(W^{(t)}) \leq 2\hat{F}(w) + \frac{\|\bar{w}^{(1)} - w\|^2}{\eta T}. \tag{34}$$

By Lemma 20, we have $\hat{F}(\bar{w}^{(t)}) \leq 2\hat{F}(W^{t)})$. Finally, as we proved in the beginning, DGD is a descent algorithm, implying

$$\hat{F}(\bar{w}^{(T)}) \leq \frac{1}{T} \sum_{t=1}^{T} \hat{F}(\bar{w}^{(t)})$$

In view of (34), this yields the claim of the theorem for the train loss (28). Appealing to Lemma 19, gives (29). This completes the proof of the theorem. ■

Appendix E. Proofs for Section A

Lemma 23 (Train loss under PL condition) *Let Assumptions 1,3 and 9 hold, and let* $\eta \leq \min(1/\mu, \sqrt{\frac{(1-\alpha_1)\mu}{4L^4\alpha_2}}, 1/L)$ *and* $\bar{\zeta} := \max(\frac{1+\alpha_1}{2}, 1-\eta\mu/2)$*, where* $\alpha_1 := \frac{3+\lambda}{4}, \alpha_2 := 4(\frac{2}{1-\lambda} - 1)$ *same as in Lemma 12, Then for* $t \geq 1$

$$\hat{F}(\bar{w}^{(t)}) \leq \bar{\zeta}^{t-1} \hat{F}(\bar{w}^{(1)}) \tag{35}$$

Proof by $L-$ smoothness we have

$$\begin{aligned}\hat{F}(\bar{w}^{(t+1)}) &\leq \hat{F}(\bar{w}^{(t)}) - \eta\langle\nabla\hat{F}(\bar{w}^{(t)}), \bar{\nabla}\hat{F}(W^{(t)})\rangle + \frac{\eta^2 L}{2}\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 \\ &= \hat{F}(\bar{w}^{(t)}) - \frac{\eta-\eta^2 L}{2}\|\bar{\nabla}\hat{F}(W^{(t)})\|^2 - \frac{\eta}{2}\|\nabla\hat{F}(\bar{w}^{(t)})\|^2 + \frac{\eta}{2}\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{w}^{(t)})\|^2 \\ &\leq \hat{F}(\bar{w}^{(t)}) - \frac{\eta}{2}\|\nabla\hat{F}(\bar{w}^{(t)})\|^2 + \frac{\eta}{2}\|\bar{\nabla}\hat{F}(W^{(t)}) - \nabla\hat{F}(\bar{w}^{(t)})\|^2 \\ &\leq \hat{F}(\bar{w}^{(t)}) - \frac{\eta}{2}\|\nabla\hat{F}(\bar{w}^{(t)})\|^2 + \eta L^2\|W^{(t)} - \bar{W}^{(t)}\|_F^2\end{aligned}$$

By $\mu-$PL condition we have,

$$\hat{F}(\bar{w}^{(t+1)}) \leq (1-\eta\mu)\hat{F}(\bar{w}^{(t)}) + \eta L^2\|W^{(t)} - \bar{W}^{(t)}\|_F^2.$$

By Lemma 12,

$$\|W^{(t)} - \bar{W}^{(t)}\|_F^2 < \alpha_2\eta^2 L\sum_{i=1}^{t-1}\alpha_1^{i-1}\hat{F}(\bar{w}^{(t-i)}) \tag{36}$$

which results in,

$$\hat{F}(\bar{w}^{(t+1)}) \leq (1-\eta\mu)\hat{F}(\bar{w}^{(t)}) + \alpha_2 L^3\eta^3\sum_{i=1}^{t-1}\alpha_1^{i-1}\hat{F}(\bar{w}^{(t-i)})$$

By induction assume $\hat{F}(\bar{w}^{(t)}) \leq \bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)})$ then using the assumptions on $\bar{\zeta}$ and η yield the following inequalities,

$$\begin{aligned}\hat{F}(\bar{w}^{(t+1)}) &\leq (1-\eta\mu)\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) + \alpha_2\eta^3 L^3\hat{F}(\bar{w}^{(1)})\sum_{i=1}^{t-1}\alpha_1^{i-1}\bar{\zeta}^{t-i-1} \\ &\leq (1-\eta\mu)\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) + \alpha_2\eta^3 L^3\hat{F}(\bar{w}^{(1)})\frac{\bar{\zeta}^{t-2}}{1-\alpha_1/\bar{\zeta}} \\ &= (1-\eta\mu+\alpha_2\eta^3 L^3/(\bar{\zeta}-\alpha_1))\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) \\ &\leq (1-\eta\mu+2\alpha_2\eta^3 L^3/(1-\alpha_1))\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) \\ &\leq (1-\eta\mu+\eta\mu/2)\bar{\zeta}^{t-1}\hat{F}(\bar{w}^{(1)}) \\ &\leq \bar{\zeta}^{t}\hat{F}(\bar{w}^{(1)}).\end{aligned}$$

This completes the proof of the lemma. ■

E.1. Proof of Lemma 8

Lemma 24 (Restatement of Lemma 8) *Let Assumptions 1,3 and 9 hold and let the step-size* $\eta \leq \min(\frac{1-\alpha_1}{\mu}, \frac{1}{2L^2}\sqrt{\frac{(1-\alpha_1)\mu}{\alpha_2}}, \frac{1}{L})$*, where the constants* $\alpha_1 \in (0,1)$ *and* $\alpha_2 > 0$ *are defined same as in*

Lemma 23. Define $\zeta := 1 - \frac{\eta\mu}{2}$, then under the data separability assumption, the iterates of DGD satisfy for all $t \geq 1$,

$$\hat{F}(\bar{w}^{(t)}) \leq \zeta^{t-1}\hat{F}(\bar{w}^{(1)}),$$
$$\|W^{(t)} - \bar{W}^{(t)}\|_F^2 \leq \frac{2\alpha_2\eta^2 L^2 \hat{F}(\bar{w}^{(1)})}{1-\alpha_1}\zeta^{t-1}.$$

Proof The bound on the train loss follows directly by Lemma 23, after noting that $\eta \leq \frac{1-\alpha_1}{\mu}$ implies $\frac{1+\alpha_1}{2} \leq 1 - \eta\mu/2$. The consensus error is derived by (36) and using the bound on $\hat{F}(\bar{w}^{(t)})$. ■

E.2. Proof of Theorem 9

Theorem 25 (Restatement of Theorem 9) *Let Assumptions 1-4 and 9 hold, and let η and ζ be as in Lemma 23. Then the iterates of DGD under the data separability assumption satisfy for all $T \geq 1$,*

$$\mathbb{E}[F(\bar{w}^{(T)})] = \mathcal{O}\Big(\zeta^T + \frac{L^2c^2}{n^{3-2\alpha}\mu^{2\alpha}}(\eta T)^{2-2\alpha} + \frac{\eta^2 L^4}{\mu^2 N}\Big).$$

Proof By simplifying Lemma 1 using the convergence bounds in Lemma 24, we end up with the following,

$$\mathbb{E}[F(\bar{w}^{(T)})] = \mathcal{O}\Big(\zeta^T + \frac{L^2c^2\eta^2}{n^{3-2\alpha}(1-\zeta)^{2\alpha}}T^{2-2\alpha} + \frac{\eta^2 L^4}{(1-\sqrt{\zeta})^2 N}\Big).$$

Based on the definition of ζ, we have $(\frac{\eta}{1-\zeta})^{2\alpha} = (\frac{2}{\mu})^{2\alpha}$ and $\frac{\eta^2}{(1-\sqrt{\zeta})^2} \leq \frac{4}{\mu^2}$. This proves the statement of the theorem. ■

Appendix F. Auxiliary Results

Proposition 26 (Bounds on the exponential loss) *Consider linear classification with the exponential loss $f(w,(a,y)) = \exp(-y \cdot w^\top a)$ over linearly separable dataset $(a_i, y_i)_{i=1}^n$ with binary labels y_i and with $\max_i \|a_i\| \leq r$ for a constant r. The training loss in this case satisfies for all $w \in \mathbb{R}^d$,*

$$\|\nabla\hat{F}(w)\| \in [c'F(w), cF(w)], \quad \|\nabla^2\hat{F}(w)\| \leq hF(w), \tag{37}$$

for constants c, c' and h independent of w.

Proof Recall that,

$$\hat{F}(w) = \frac{1}{n}\sum_{i=1}^n \exp(-y_i \cdot w^\top a_i).$$

One can deduce that,

$$\nabla \hat{F}(w) = -\frac{1}{n}\sum_{i=1}^{n} y_i a_i \exp(-y_i\, w^\top a_i),$$
$$\nabla^2 \hat{F}(w) = \frac{1}{n}\sum_{i=1}^{n} a_i a_i^\top \exp(-y_i\, w^\top a_i).$$

Therefore it holds that,

$$\begin{aligned}
\|\nabla \hat{F}(w)\| &= \frac{1}{n}\|\sum_{i=1}^{n} y_i a_i \exp(-y_i w^\top a_i)\| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \|y_i a_i \exp(-y_i w^\top a_i)\| \\
&= \frac{1}{n}\sum_{i=1}^{n} \|y_i a_i\| \exp(-y_i w^\top a_i) \\
&\leq r\hat{F}(w).
\end{aligned}$$

A similar approach for the Hessian of $\hat{F}$ results in the following inequality,

$$\|\nabla^2 \hat{F}(w)\| \leq r^2 \hat{F}(w).$$

Moreover, due to linear separability there exists a $w^\star \in \mathbb{R}^d$ such that,

$$\frac{y_i {w^\star}^\top a_i}{\|w^\star\|} \geq \gamma, \;\; \forall i \in [N],$$

where $\gamma > 0$ denotes the margin. Therefore, using the supremum definition of norm we can write,

$$\begin{aligned}
\|\nabla \hat{F}(w)\| &= \frac{1}{n}\left\|\sum_{i=1}^{n} y_i a_i \exp(-y_i w^\top a_i)\right\| \\
&= \sup_{\substack{v\in\mathbb{R}^d \\ \text{s,t. } \|v\|=1}} \left\langle \frac{1}{n}\sum_{i=1}^{n} y_i a_i \exp(-y_i w^\top a_i), v \right\rangle \\
&\geq \left\langle \frac{1}{n}\sum_{i=1}^{n} y_i a_i \exp(-y_i w^\top a_i), \frac{w^\star}{\|w^\star\|} \right\rangle \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \gamma \cdot \exp(-y_i w^\top a_i) \\
&= \gamma \hat{F}(w).
\end{aligned}$$

This completes the proof. ■

Proposition 27 (Bounds on the logistic loss) *Consider linear classification with the logistic loss* $f(w,(a,y)) = \log(1+\exp(-y\cdot w^\top a))$ *over linearly separable dataset* $(a_i,y_i)_{i=1}^n$ *with binary labels* y_i *and with* $\max_i \|a_i\| \le r$ *for a constant* r*. The training loss in this case satisfies for all* $w\in\mathbb{R}^d$*,*

$$\|\nabla\hat{F}(w)\| \in [c'\Phi(w), cF(w)],\ \ \|\nabla^2\hat{F}(w)\| \le hF(w),$$

for $\Phi(w) := \frac{1}{n}\sum_{i=1}^n \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i\, w^\top a_i)}$ *and constants* c, c' *and* h *independent of* w*.*

Proof The training loss in this case is the following:

$$\hat{F}(w) = \frac{1}{n}\sum_{i=1}^n \log(1+\exp(-y_i\cdot w^\top a_i)).$$

Thus,

$$\nabla\hat{F}(w) = \frac{1}{n}\sum_{i=1}^n (-y_i a_i)\frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i\cdot w^\top a_i)},$$
$$\nabla^2\hat{F}(w) = \frac{1}{n}\sum_{i=1}^n a_i a_i^\top \frac{\exp(-y_i\, w^\top a_i)}{(1+\exp(-y_i\, w^\top a_i))^2}.$$

By considering the norm and noting that $\exp(t)/(1+\exp(t)) \le \log(1+\exp(t))$,

$$\begin{aligned}\|\nabla\hat{F}(w)\| &= \frac{1}{n}\left\|\sum_{i=1}^n (-y_i a_i)\frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i\cdot w^\top a_i)}\right\| \\ &\le \frac{1}{n}\sum_{i=1}^n \|y_i a_i\|\,\frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i\cdot w^\top a_i)} \\ &\le \frac{r}{n}\sum_{i=1}^n \log(1+\exp(-y_i w^\top a_i)) \\ &= r\hat{F}(w).\end{aligned}$$

Likewise, since $\exp(t)/(1+\exp(t))^2 \le 2\log(1+\exp(t))$, we can conclude that the operator norm of the Hessian satisfies,

$$\nabla^2\hat{F}(w) \le 2r^2\hat{F}(w).$$

This completes the proof of upper-bounds for the gradient and Hessian. For the lower-bound on gradient note that by using the supremum definition of norm and recalling the max-margin separator

satisfies $\frac{y_i w^{\star\top} a_i}{\|w^\star\|} \geq \gamma$ for the margin $\gamma > 0$ and all $i \in [n]$, we can write,

$$\begin{aligned}
\|\nabla \hat{F}(w)\| &= \frac{1}{n}\left\|\sum_{i=1}^{n} y_i a_i \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}\right\| \\
&= \sup_{\substack{v\in\mathbb{R}^d \\ \text{s,t. } \|v\|=1}} \left\langle \frac{1}{n}\sum_{i=1}^{n} y_i a_i \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}, v \right\rangle \\
&\geq \left\langle \frac{1}{n}\sum_{i=1}^{n} y_i a_i \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}, \frac{w^\star}{\|w^\star\|} \right\rangle \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \gamma \frac{\exp(-y_i\, w^\top a_i)}{1+\exp(-y_i \cdot w^\top a_i)}.
\end{aligned}$$

This yields the lower bound $\gamma\Phi(w)$ on the norm of gradient and completes the proof. ■

Proposition 28 (Realizability of the exponential and logistic loss [23]) *The exponential loss function satisfies the realizability assumption (Assumption 6) with $\rho(\varepsilon) = -\frac{1}{\gamma}\log(\varepsilon)$, where γ denotes the margin. Moreover, the logistic loss function satisfies the realizability assumption with $\rho(\varepsilon) = -\frac{1}{\gamma}\log(\exp(\varepsilon)-1)$.*

Appendix G. Improved algorithms for decentralized learning with separable data

In this section, we consider decentralized learning with exponential tail losses on separable data and propose algorithms for improving the convergence rates based on the normalized GD mechanism. Our proposal –Fast Distributed Logistic Regression(FDLR)– is summarized in Algorithm 1. Each agent keeps two local variables $w_\ell, v_\ell \in \mathbb{R}^d$ which are also communicated to neighbor agents in each round. In matrix notation, Algorithm 1 has the following two steps:

$$\begin{aligned}
W^{(t+1)} &= A(W^{(t)} - \eta \widetilde{V}^{(t)}), \\
V^{(t+1)} &= A\, V^{(t)} + \nabla \hat{F}(W^{(t+1)}) - \nabla \hat{F}(W^{(t)}).
\end{aligned}$$

Here $A \in \mathbb{R}^{N\times N}$ is the mixing matrix corresponding to an undirected network which satisfies the regularity conditions in Assumption 1. Furthermore $W^{(t)}, V^{(t)}, \nabla\hat{F}(W^{(t)}) \in \mathbb{R}^{N\times d}$ are formed by stacking $w_\ell^{(t)}, v_\ell^{(t)}$ and local gradients $\nabla \hat{F}_\ell(w_\ell^{(t)})$ for all ℓ, respectively. The matrix $\widetilde{V}^{(t)}$ is formed by concatenation of the vectors $v_\ell^{(t)}/\|v_\ell^{(t)}\|$. In step (2), every agent ℓ runs in parallel an update rule which resembles the distributed gradient descent update rule, with the exception that the local gradient $\nabla \hat{F}_\ell(w_\ell^{(t)})$ is replaced by $v_\ell^{(t)}/\|v_\ell^{(t)}\|$. In the next step, agents send their local parameters w_ℓ^t, v_ℓ^t to their neighbors. Step (4) is the consensus step in which agent ℓ computes a weighted average of $w_k^{(t+1/2)}$ sent from neighbor agent k, in order to update $w_\ell^{(t)}$. Step (5) uses the newly computed local gradient $\nabla \hat{F}_\ell(w_\ell^{(t+1)})$ and the gradient computed in the previous step to updates the local parameter $v_\ell^{(t)}$. This step is previously used in the gradient tracking algorithm

Algorithm 1: FDLR

Input: Initial values $w_\ell^{(1)}, v_\ell^{(1)} \in \mathbb{R}^d$ for all agents $\ell \in [N]$, step size η_t and mixing matrix $A = [A_{\ell k}]_{N\times N}$

for $t = 1, \dots, T$ *all agents* $\ell \in [N]$ *in parallel* **do**

$w_\ell^{(t+\frac{1}{2})} = w_\ell^{(t)} - \eta_t \frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|}$

send and receive local variables $w_\ell^{(t+\frac{1}{2})}$ and $v_\ell^{(t)}$

$w_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} w_k^{(t+\frac{1}{2})}$

$v_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} v_k^{(t)} + \nabla \hat{F}_\ell(w_\ell^{(t+1)}) - \nabla \hat{F}_\ell(w_\ell^{(t)})$

end

Algorithm 2: FDLR with Nesterov momentum

Input: Initial values $w_\ell^{(1)}, v_\ell^{(1)}, z^{(1)} \in \mathbb{R}^d$ for all agents $\ell \in [N]$, hyper-parameters η_t, γ_t and mixing matrix $A = [A_{\ell k}]_{N\times N}$

for $t = 1, \dots, T$ *all agents* $i \in [N]$ *in parallel* **do**

$z_\ell^{(t+1)} = \gamma_t (z_\ell^{(t)} + \frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|})$

$w_\ell^{(t+\frac{1}{2})} = w_\ell^{(t)} - \eta_t (z_\ell^{(t+1)} + \frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|})$

send and receive local variables $w_\ell^{(t+\frac{1}{2})}$ and $v_\ell^{(t)}$

$w_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} w_k^{(t+\frac{1}{2})}$

$v_\ell^{(t+1)} = \sum_{k\in\mathcal{N}_\ell} A_{\ell k} v_k^{(t)} + \nabla \hat{F}_\ell(w_\ell^{(t+1)}) - \nabla \hat{F}_\ell(w_\ell^{(t)})$

end

(e.g. see [19]). The purpose behind introducing the variable $v_\ell^{(t)}$ is to estimate the global loss. In particular, in the infinite time limit we can deduce that,

$$V^{(\infty)} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top \nabla \hat{F}(W^{(\infty)}).$$

While each $v_\ell^{(t)}$ is distinct among agents, each of them is estimating the global loss of distributed optimization i.e., $\sum_{\ell=1}^N \nabla \hat{F}_\ell(w_\ell^{(t)})$. With this, step (2) of FDLR is approximately implementing a normalized GD update, which provably speeds up the convergence in centralized logistic regression over separable data[15].

A Nesterov accelerated variant of normalized GD has already been studied by [7] for centralized logistic regression. In distributed case, the problem is more subtle as each agent has access only to local gradients. Algorithm 2 unifies FDLR with a Nesterov Momentum. In particular, in step (3) the local parameter $w_\ell^{(t)}$ is updated by a weighted average ($z_\ell^{(t+1)}$) of normalized gradients from previous iterations. While we do not prove convergence rates for the proposed algorithms, we study their asymptotic convergence. Notably, the next theorem shows that for a time-decaying step-size

$\eta = 1/\sqrt{t}$, the solution of FDRL converges in direction to the solution of centralized gradient descent. We defer numerical performance of both Algorithms to the next section.

Theorem 29 (Asymptotic convergence of FDLR) *Let the sequence $\{w_\ell^{(t)}\}$ be generated by FDRL(Algorithm 1) for a separable dataset and with $\eta_t = \mathcal{O}(1/\sqrt{t})$. Then for all $\ell \in [N]$,*

$$\lim_{t\to\infty} \frac{w_\ell^{(t)}}{\|w_\ell^{(t)}\|} = \frac{w_{\text{MM}}}{\|w_{\text{MM}}\|},$$

where w_{MM} is the solution to max-margin problem.

Proof Replace $\frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|}$ in step 2 of Algorithm 1 by arbitrary perturbations $\varepsilon_\ell^{(t)}$ of unit norm. Then note that the sequence $\{w_\ell^{(t)}\}$ generated by step 2 is identical to decentralized GD with $\eta\|\varepsilon_\ell^{(t)}\| \to 0$. Thus by [16, Lemma 1], consensus is asymptotically achieved for all $\ell \in [N]$, i.e.,

$$\lim_{t\to\infty} \|w_\ell^{(t)} - \bar{w}^{(t)}\| = 0.$$

Thus

$$\lim_{t\to\infty} \|w_\ell^{(t+1)} - w_\ell^{(t)}\| = \lim_{t\to\infty} \|\bar{w}^{(t+1)} - \bar{w}^{(t)}\| = \lim_{t\to\infty} \eta\|\bar{\varepsilon}^{(t)}\| = 0.$$

This implies that for all $i \in [N]$ we have $\lim_{t\to\infty} \|\nabla \hat{F}_\ell(w_\ell^{(t+1)}) - \nabla \hat{F}_\ell(w_\ell^{(t)})\| = 0$, thus by appealing again to [16, Lemma 1] and applying it to step (5) of Algorithm 1, we find that,

$$\lim_{t\to\infty} \|v_\ell^{(t)} - \bar{v}^{(t)}\| = 0.$$

Aggregations of gradients in step (5) imply that $\bar{v}^{(t)} = \frac{\mathbf{1}^\top \nabla \mathcal{L}(W^{(t)})}{N} \to \nabla \hat{F}(\bar{w}^{(t)})$. Thus step (2) of FDLR for every agent i converges to $\bar{w}^{(t)} - \eta \frac{\nabla \hat{F}(\bar{w}^{(t)})}{\|\nabla \hat{F}(\bar{w}^{(t)})\|}$, i.e.,

$$\left\| \left(w_\ell^{(t)} - \eta \frac{v_\ell^{(t)}}{\|v_\ell^{(t)}\|} \right) - \left(\bar{w}^{(t)} - \eta \frac{\nabla \hat{F}(\bar{w}^{(t)})}{\|\nabla \hat{F}(\bar{w}^{(t)})\|} \right) \right\| \overset{t\to\infty}{\Longrightarrow} 0.$$

Thus for all ℓ, the sequence $\{w_\ell^{(t)}\}$ converges to the solution of normalized GD, i.e., the max-margin separator w_{MM}, for linearly separable datasets ([15, Theorem 5]). This leads to the statement of the theorem. ■

Appendix H. Numerical Experiments

In this section, we present numerical experiments to verify our theoretical results and demonstrate the benefits of our proposed algorithms. We begin with a numerical study of the performance of FDLR.

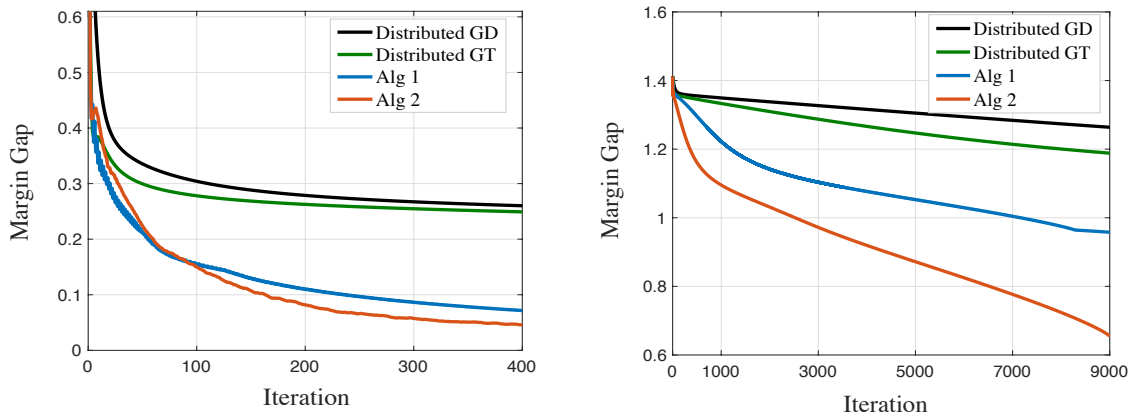

Figure 1: Margin gap for our proposed algorithms (Algorithms 1-2) compared to the vanilla distributed gradient descent and gradeint tracking algorithms on synthetic data $y = \texttt{sign}(a^\top w^\star)$ (Left) and on two classes from the UCI WINE dataset (Right).

H.1. Experiments on FDLR

In Figure 1(Left), we compare the performance of FDLR and its momentum variant to distributed gradient descent(GD) and gradient tracking(GT) for exponential loss with signed measurements (i.e., $y = \texttt{sign}(a^\top w^\star)$ for samples a, labels y and the true vector of regressors $w^\star$) with $n = 100$, $d = 25$. The underlying graph is selected as an Erdos-Rènyi graph with $N = 50$ agents (each having access to 2 samples from training set) and connectivity probability $p_c = 0.3$. The $y-$axis, Margin Gap represents the distance of normalized $w_\ell^{(t)}$ to the normalized final solution for agent $\ell = 1$, i.e., $\|\frac{w_1^{(t)}}{\|w_1^{(t)}\|} - \frac{w_{\text{MM}}}{\|w_{\text{MM}}\|}\|$. Intuitively, Margin Gap can be thought as the difference in direction to the final solution. The hyper-parameters η_t, γ_t are fine-tuned for each algorithm to represent the best of each algorithm and the final values are $\eta_t = 0.1, 0.05, 0.5$ and 0.2 for Distributed GD, GT, Algorithm 1 and Algorithm 2, respectively and $\gamma_t = 0.8$ for Algorithm 2. Our algorithms significantly outperform the well-known distributed learning algorithms in directional convergence to the final solution. Regardless, in this case we noticed that the gain obtained by including the momentum is small. In Figure 1(Right), we consider a binary classification task on a real-world dataset (two classes of the UCI WINE dataset [1]) where $d = 13$ and $n = 107$. We compare the performance of FDLR (blue line) and its momentum variant (red line) with DGD and DGT on an Erdos-Rènyi graph with $N = 10$ and $p_c = 0.4$. The hyper-parameters are fine-tuned to $\eta_t = 12, 1, 0.9, 2$ for DGD, DGT, Algorithm 1 and Algorithm 2 respectively, and $\gamma_t = 0.88$ for Algorithm 2. Notably, while Algorithm 1 significantly outperforms both DGD and DGT, the benefits of adding the momentum are also significant in this case as Algorithm 2 demonstrates a faster rate of convergence than Algorithm 1.

Figure 2(Top) illustrates the train and test errors based on iteration for the same setting as Figure 1 with $n = 800$ and $d = 50$. Here the hyper-parameters are fine-tuned to be $\eta_t = 0.01, 0.01, 0.4, 0.5$ for DGD,DGT, Algorithm 1 and Algorithm 2, respectively and $\gamma_t = 0.5$ for Algorithm 2. Figure 2(Bottom) represents the train $\hat{F}(\bar{w}^{(t)})$ and test $F(\bar{w}^{(t)})$ losses. Here, we use the same dataset with $N = 10, p_c = 0.4$ and an exponential loss. The hyper-parameters are fine-tuned to $\eta_t =$

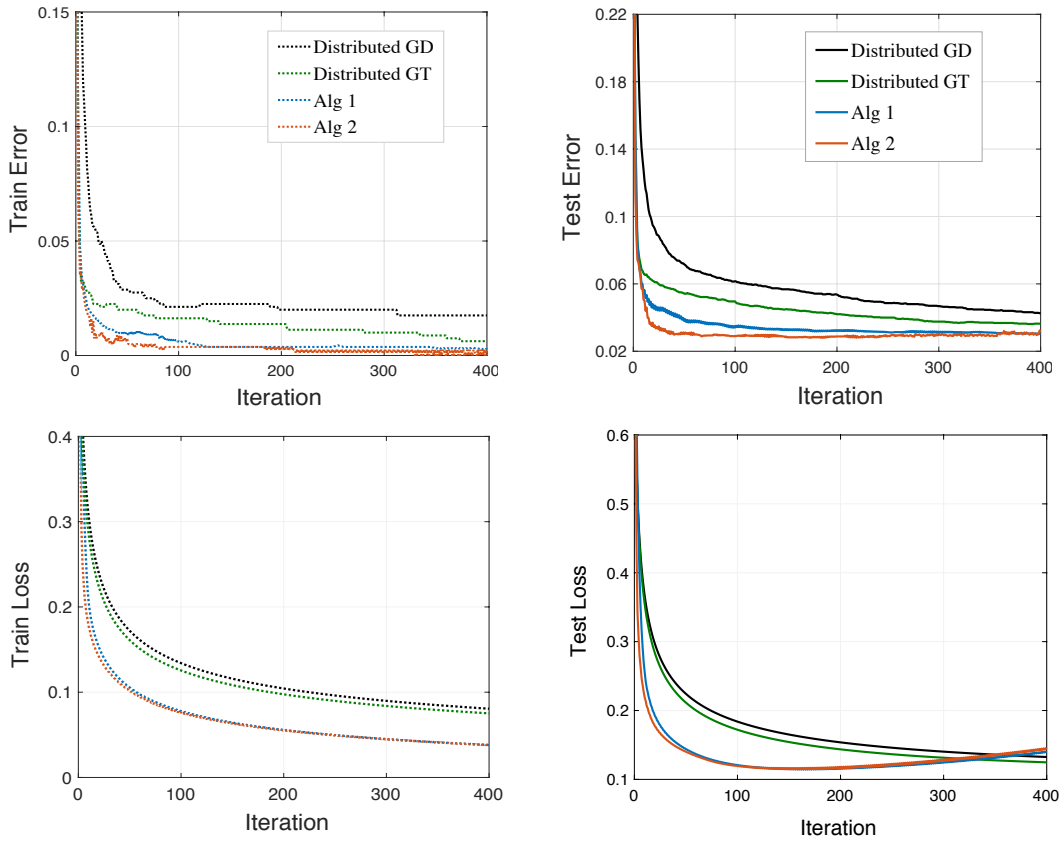

Figure 2: Train/test misclassification errors (Top) and train/test losses(Bottom) for our proposed algorithms compared to the decentralized gradient descent and gradient tracking algorithms.

$0.01, 0.012, 0.4, 0.6$ and $\gamma_t = 0.9$. Both of our algorithms outperform the commonly used methods gradient descent and gradient tracking, in training error and test error performance. We can see that FDLR and its momentum variant are faster in interpolating the training data. Also, the gains of adding the momentum are significant, since FDLR with Nesterov momentum reaches an approximation of its final test accuracy in 50 iterations, while the same happens for FDLR with approximately 300 iterations. An interesting phenomenon in Figure 2 (see also the bottom plot in Figure 3) is the behavior of "test loss": while during the starting phase the test loss is monotonically decreasing, after sufficient iterations the test loss starts increasing. This behavior of test loss is indeed captured by the bounds on the test loss of DGD in Theorems 4-6 and Remarks 5-7. In particular, the increase in test loss is observed in the bound for the test loss $\mathcal{O}(\frac{(\log T)^2}{\sqrt{T}} + \frac{(\log T)^2}{n})$ in Remark 5, where the presence of the term $\frac{(\log T)^2}{n}$ suggests that the bound after sufficient iterations starts to slowly increase.

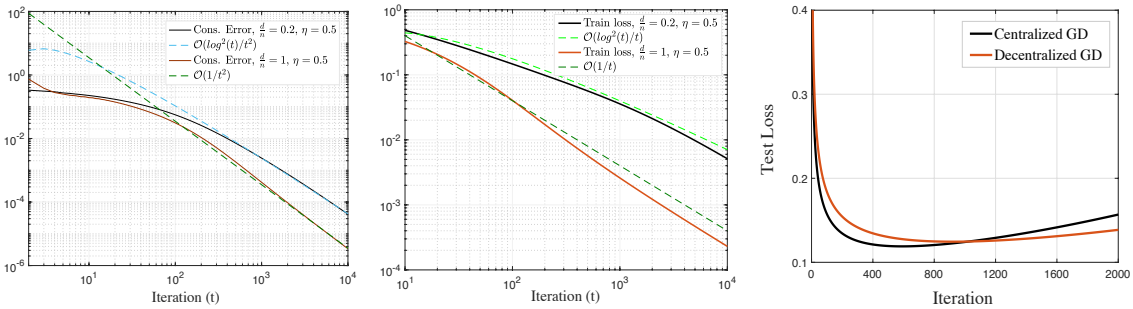

Figure 3: Consensus error, train loss and test loss for DGD with exponential loss and linearly separable data. Left and middle figures, verify the rates $\tilde{\mathcal{O}}(1/t^2)$ and $\tilde{\mathcal{O}}(1/t)$ (as predicted by Theorem 6 and Remark 7) for consensus error and training loss of DGD with logistic regression, respectively. The Right figure signifies that the test loss for DGD and GD show approximately similar convergence behavior under separable data.

H.2. Experiments on the convergence behavior of DGD

Next, we conduct numerical experiments on the convergence behavior of DGD for the train loss and the consensus error. We consider the same network topology, mixing matrix and data setup as in the last figure. Figure 3 (Left and middle) shows the consensus error $\frac{1}{N}\|W^{(t)} - \bar{W}^{(t)}\|_F^2$ and the train loss $\hat{F}(\bar{w}^{(t)})$ in solid lines, for two over-parameterization ratios d/n. We recall that d and n represent the dimension of the ambient space and the dataset size, respectively.

The dashed lines help to show for each solid line, the approximate rate of convergence after sufficient number of iterations. Notably, we observe that the convergence rates on the consensus error ($\tilde{\mathcal{O}}(1/t^2)$) and on the train error ($\tilde{\mathcal{O}}(1/t)$) stated in Remark 7 are attained in both cases (recall that $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors).

Figure 3 (Right) depicts the Test loss of DGD for $d/n = 0.05$. For comparison, the corresponding curve for centralized GD is also shown. Here step-sizes are fine-tuned to represent the best of each algorithm. In agreement with our findings in Remark 7, we observe the similar behavior between the convergence behavior of two algorithms. As a final remark, the slight increase in the curves of test loss are due to the logarithmic factors in the derived upper-bounds for the test loss.

H.3. Experiments on over-parameterized Least-squares

In Figure 4, we conduct experiments for highly over-parameterized Least-squares ($f(w, x) = \|1 - w^\top x\|^2$), where d is typically significantly larger than n to ensure perfect interpolation of dataset. Note that, the train loss is not strongly-convex in this case, instead it satisfies the PL condition(Assumption 9). Notably, as predicted by Lemma 8, we notice the linear convergence of the train loss and the consensus error in Figure 4 (Left). On the other hand, for the test loss, we observe its remarkably fast convergence (after approximately 50 iterations) to the optimal value, which is followed by a sharp increase in the subsequent iterations.

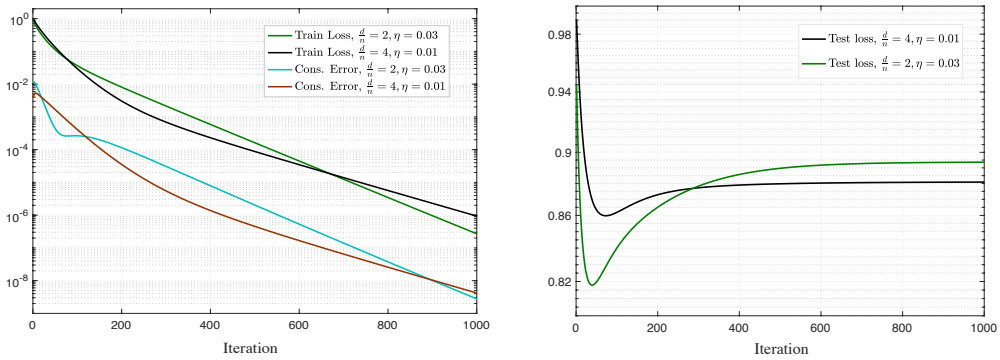

Figure 4: Consensus error, train loss and test loss for DGD with over-parameterized least-squares(square loss). The test loss achieves its optimum at the very early stages of DGD.

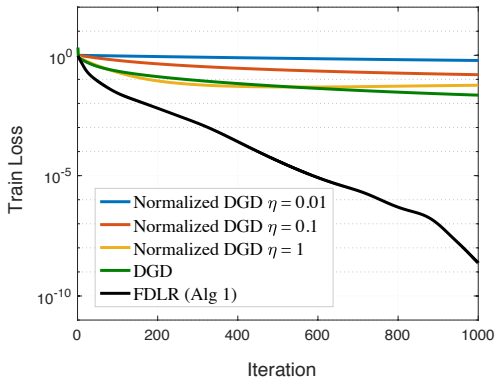

Figure 5: Normalized DGD with the update rule in Eq.(39) for different step-sizes η compared to DGD (Eq.(38)) and to FDLR (Alg 1). The step-sizes for DGD and FDLR are fine-tuned so that best of each algorithm is depicted. Normalized DGD cannot outperform DGD while FDLR is significantly faster than DGD. Here we consider linear classification with the exponential loss function and the dataset is generated according to signed measurements with Gaussian features and $n = 100, d = 50$.

H.4. On the update rule of FDLR

In the final section of the paper, we state a remark regarding the update rule of FDLR. Recall the update rule of DGD,

$$w_\ell^{(t+1)} = \sum_{k \in \mathcal{N}_\ell} A_{\ell k} w_k^{(t)} - \eta_t \nabla \hat{F}_\ell(w_\ell^{(t)}). \tag{38}$$

Notably, we expect FDLR to be perhaps the simplest approach for accommodating normalized gradients in decentralized learning setting since in DGD the agents only have access to their local gradients. In particular consider a Normalized DGD algorithm with the same update as in (38) but with $\nabla \hat{F}_i(w_i^{(t)})$ replaced by $\nabla \hat{F}_i(w_i^{(t)})/\|\nabla \hat{F}_i(w_i^{(t)})\|$, i.e.,

$$w_\ell^{(t+1)} = \sum_{k \in \mathcal{N}_\ell} A_{\ell k} w_k^{(t)} - \eta_t \frac{\nabla \hat{F}_\ell(w_\ell^{(t)})}{\|\nabla \hat{F}_\ell(w_\ell^{(t)})\|} \tag{39}$$

The Normalized DGD algorithm above does not lead to faster convergence. This is due to the fact that in DGD the local gradient norm $\|\nabla \hat{F}_i(w_i^{(t)})\|$ can be different than the global gradient norm $\|\nabla \hat{F}(w_i^{(t)})\|$. Thus even if with the update rule (39) the local parameters $w_i^{(t)}$ converge to the global optimal solution, still the update rule for the averaged parameter $\bar{w}^{(t)}$ is different than the update rule of centralized normalized GD. Our numerical experiment in Figure 5 demonstrates the incapability of Normalized DGD in speeding up DGD. In particular, we note that for any choice of step-size Normalized DGD does not lead to acceleration compared to DGD whereas FDLR massively outperforms DGD.