OPT2022: 14th Annual Workshop on Optimization for Machine Learning

On Generalization of Decentralized Learning with Separable Data

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

Christos Thrampoulidis CTHRAMPO@ECE.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 approximately 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, on the algorithmic front, we design improved gradient-based routines for decentralized learning with separable data and empirically demonstrate orders-of-magnitude of speed-up in terms of both training and generalization performance.

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 [20]:

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

. The long version of the paper including additional results and all the proofs is available at [29].

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 $\hat{F}_\ell$, $\ell \leq N$, where each $\hat{F}_\ell$ is formed as the average empirical risk evaluated on the local training dataset $\mathcal{S}_\ell$ of the ℓ th agent:

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

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., [15, 18–20, 34]. The bulk of these studies build upon classical optimization theory [22] 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 [18]. 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., [9, 11, 32], 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. We extend our results to losses satisfying the PL condition (in the long version of this work [29, Sec. 2.3]). Numerical experiments in verify our theoretical results [29, Sec. 3.2].

- We propose two algorithms (see [29, Sec. 2.4]) for speeding up the convergence of decentralized learning with separable data. Notably, numerical experiments [29, Sec. 3.1] 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 [3, 9, 12, 13, 15, 18, 20, 24, 30, 31, 33, 34]. 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 [28] 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 [12, Theorem 2]. While [12] 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 [6, 7, 17, 25–27] 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, [6, 27] 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 [26] and Schliserman and Koren [25] 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 9), 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 [4, 5, 14]. Specifically we use a self-boundedness and a realizability assumption [25] 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. Formally, we assume the following, where for simplicity, we use the short-hand $f_x(w) := f(w, x)$ for the loss incurred at a generic $x \in \mathcal{D}$ in the data distribution $\mathcal{D}$.

Assumption 2 (Convexity) *The loss functions $f_x : \mathbb{R}^d \to \mathbb{R}$ are convex and differentiable, satisfying, $f_x(w) \le f_x(v) + \langle \nabla f_x(w), w - v\rangle$.*

Assumption 3 (Smoothness) *The loss functions $f_x : \mathbb{R}^d \to \mathbb{R}$ are L-smooth and differentiable, i.e. $f_x(w) \le f_x(v) + \langle \nabla f_x(v), w - v\rangle + \frac{L}{2}\|w - v\|^2$.*

Assumption 4 (Self-boundedness of the gradient) *The loss functions $f_x : \mathbb{R}^d \to \mathbb{R}$ satisfy the self-boundedness property with the parameters $c > 0$ and $\alpha \in [\frac{1}{2}, 1]$, i.e.,*

$$\|\nabla f_x(w)\| \leq c\,(f_x(w))^{\alpha}\,.$$

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.

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.

Lemma 1 bounds the test loss with respect to the train loss and the consensus error. In the following sections, we show how Lemma 1 yields test loss bounds on DGD by establishing bounds on the train loss and consensus errors under different assumptions on the loss function.

It is worth remarking that Eq. (4) is in fact valid not only for DGD, but also for Decentralized Gradient Tracking (DGT). DGT is another popular algorithm for distributed learning that can accelerate train error convergence over DGD by modifying the update in Eq.(2) such that each agent keeps a running estimate of the global gradient [21]. The reason why Eq.(4) continues to hold for DGD is that the proof of Lemma 1 only relies on the updates of the "averaged" parameter $\bar{w}^{(t)} := \frac{1}{N}\sum_{\ell=1}^N w_\ell$ and that the update rule of $\bar{w}^{(t)}$ for both DGD and DGT is derived as $\bar{w}^{(t)} = \bar{w}^{(t-1)} - \eta\frac{1}{N}\sum_{\ell=1}^N \nabla\hat{F}_\ell(w_\ell^{(t-1)})$. Thus, starting with 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.

Lemma 2 (Training bounds for convex losses) *Under Assumptions 1-4, for any $w \in \mathbb{R}^d$ and for a fixed step-size $\eta < \frac{1}{L}\min(1-\alpha_1, \sqrt{\frac{1-\alpha_1}{2\alpha_2}})$, where $\alpha_1 \in (3/4, 1), \alpha_2 > 4$ are parameters that* *depend only on the mixing matrix, the following holds for the train loss and consensus error of DGD (2):*

$$\frac{1}{T}\sum_{t=1}^{T}\hat{F}(\bar{w}^{(t)}) \leq \frac{2\|w\|^2}{\eta T} + 4\hat{F}(w), \tag{5}$$

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

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 5 (Realizability) *The loss functions $f_x : \mathbb{R}^d \to \mathbb{R}$ satisfy the realizability condition, i.e. $\exists$ 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 $f_x(\hat{w}) \leq \varepsilon$.*

The set of assumptions 2-5 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 $r > 0$; See [29, Prop. 24].

Remark 3 (Training loss of DGD on separable data) *The realizability assumption appeared in [14, 25] (and implicitly used in [6, 26]) 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 30 and [25, Lemma 4]). Based on Lemma 2, this leads to the following bound for DGD training loss for all $\varepsilon > 0$,*

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

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 [6, Theorem 1.1].

Remark 4 *The bounds of Lemma 2 are true for any dataset $\{x_i\}_{i\in[n]}$ provided that Assumptions 2 and 3 hold for all $f_x = f_{x_i} = f(w, x_i) := f_i(w), i \in [n]$. Similarly, (6) holds provided Assumption 5 is true over the training set (i.e. provided the training dataset is separable). However, bounding the test loss in Lemma 1, requires bounding the* expectation over all datasets *of the train/consensus errors. This is guaranteed by Assumptions 2-5 as they hold for any point x in the distribution.*

Theorem 5 (Test loss with convex losses) *Under Assumptions 1-5, by choosing $\eta < \frac{1}{L\sqrt{T}}\min(1-\alpha_1, \sqrt{\frac{1-\alpha_1}{2\alpha_2}})$ where $\alpha_1 \in (3/4, 1), \alpha_2 > 4$ 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 6 (DGD with logistic regression never overfits) *As in Remark 3, we take logistic regression on separable data with margin $\gamma > 0$ 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.*

Remark 7 (Log factors) *The attentive reader will have recognized in Remarks 3 and 6 that due to the "$\rho(\varepsilon) = \mathcal{O}(\log(T))$" factor, the upper bound on the test loss in Eq. (7) increases (very) slowly with $\log^4(T)$. Note that this term becomes dominant only when T is exponentially large with respect to the sample size n and the margin γ. Our experiments (in the long version of this work [29]) confirm this slow logarithmic increase late in the training phase. Analogous behavior, but for centralized GD training, are discussed in [25, 27].*

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

We note that the bounds in Lemma 2 and Theorem 5 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 5 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 6 (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 7 (Self-lowerbounded gradient) *The global loss satisfies for a constant τ that*

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

Assumptions 2, 4, 6 and 7 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.

Theorem 8 (Last iterate convergence of DGD) *Consider DGD with the loss functions and mixing matrix satisfying Assumptions 1,2,6,7 and Assumption 4 with $\alpha = 1$ and $c = h$. Assume that the step-size satisfies $\eta < \frac{\delta}{\hat{F}(1)}$, for a constant δ 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 $w \in \mathbb{R}^d$,*

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

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

Remark 9 (Improved rates) *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 8 and noting that* $\rho(\varepsilon) = \log(1/\varepsilon)/\gamma$ *,*

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

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,*

$$\mathbb{E}[F(\bar{w}^{(T)})] = \tilde{\mathcal{O}}\Big(\frac{1}{\eta T} + \frac{1}{n} + \eta^2\Big). \tag{8}$$

In accordance to Remark 4, we can conclude the above from Lemma 1 provided Assumptions 4 and 7. Thus, the bounds of Theorem 8 remain true for all training sets within the data distribution. We note that the resulting bound in Eq. (8) *is a superior rate for the test loss of logistic regression, compared to the rate of Remark 6. Concretely, setting* $\eta = 1/T^{1/3}$ *gives a rate of* $\tilde{\mathcal{O}}(1/T^{2/3} + 1/n)$*, faster than the* $\tilde{\mathcal{O}}(1/\sqrt{T} + 1/n)$ *rate in Remark 6. On the other hand, it is slightly slower compared to its centralized counterpart* $\tilde{\mathcal{O}}(1/T + 1/n)$ *in [25, 26]. As revealed by Lemma 1, the additional* η^2 *factor in Eq.* (8) *captures impact of the consensus term, which is unavoidable in decentralized learning.*

3. Conclusion

We studied the behavior of train loss and test loss of decentralized gradient descent (DGD) methods when training dataset is separable. To the best of our knowledge, this yields the first rigorous guarantees for the generalization error of DGD in such a setting. For the same setting, we also proposed fast algorithms and empirically verified that they accelarate both training and test accuracy. We believe our work opens several directions, with perhaps the most exciting one being the analysis of non-convex objectives. Generalization analysis of quantized DGD methods over time-varying and directed topologies [2] is another future direction. We are also interested in extending our results to other distributed settings such as federated learning [10] and Gradient Tracking [21].

References

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

[2] Sai Aparna Aketi, Amandeep Singh, and Jan Rabaey. Sparse-push: Communication-& energy-efficient decentralized distributed learning over directed & time-varying graphs with non-iid datasets. *arXiv preprint arXiv:2102.05715*, 2021.

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

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

[5] 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.

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

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

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

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

[10] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

[19] 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.

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

[21] 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.

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

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

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

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

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

[27] 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.

[28] 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.

[29] Hossein Taheri and Christos Thrampoulidis. Decentralized learning with separable data: Generalization and fast algorithms. *arXiv preprint arXiv:2209.07116*, 2022.

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

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

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

[33] 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.

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