Decentralized Parallel Algorithm for Training Generative Adversarial Nets

Mingrui Liu[†], **Youssef Mroueh**[‡], **Wei Zhang**[‡], **Xiaodong Cui**[‡], **Jerret Ross**[‡], **Tianbao Yang**[†], **Payel Das**[‡]
[†]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA
[‡]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
mingrui-liu@uiowa.edu

Abstract

Generative Adversarial Networks (GANs) are powerful class of generative models in the deep learning community. Current practice on large-scale GAN training [1] utilizes large models and distributed large-batch training strategies, and is implemented on deep learning frameworks (e.g., TensorFlow, PyTorch, etc.) designed in a centralized manner. In the centralized network topology, every worker needs to communicate with the central node. However, when the network bandwidth is low or network latency is high, the performance would be significantly degraded. Despite recent progress on decentralized algorithms for training deep neural networks, it remains unclear whether it is possible to train GANs in a decentralized manner. In this paper, we design a decentralized algorithm for solving a class of non-convex non-concave min-max problem with provable guarantee. Experimental results on GANs demonstrate the effectiveness of the proposed algorithm.

1 Introduction

Generative Adversarial Networks (GANs) [5] are very effective in modeling high dimensional data such as images, but are notoriously known to be difficult to train. The recent progress on large-scale GAN training [1] suggests using distributed large-batch training techniques on large models. Their algorithm is based on centralized network topology, in which each worker computes the local stochastic gradient based on its local data and send it to the central node, and the central node aggregates the local stochastic gradients together, updates the model parameters by first-order methods and then sends the parameters back to each worker. The central node is the busiest node since it needs to communicate with each worker concurrently, which is the main bottleneck of centralized algorithms since it could lead to communication traffic jam when the network bandwidth is low or network latency is high. To address this issue, decentralized algorithms are usually considered as a surrogate when the cost of centralized communication is prohibitively expensive. In decentralized algorithms, instead, each worker only communicates with its neighbors and no central node is needed.

Decentralized algorithms are well-studied in the literature [21, 24, 25, 11]. However, all of them are designed for solving convex or non-convex minimization problems, and none of them are directly applicable for non-convex non-concave min-max problems such as GANs. This motivates us to consider the following question:

Is it possible to design decentralized algorithms for solving non-convex non-concave problems with provable guarantee?

In this paper, we give a *positive* answer to this question by designing the first decentralized algorithm for solving a class of non-convex non-concave min-max problem with theoretical guarantee, which is also verified by numerical experiments.

Our problem of interest is to solve the following stochastic optimization problem:

$$\min_{\mathbf{u}\in\mathcal{U}}\max_{\mathbf{v}\in\mathcal{V}} F(\mathbf{u},\mathbf{v}) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[f(\mathbf{u},\mathbf{v};\xi)\right], \tag{1}$$

where $\mathcal{U}$, $\mathcal{V}$ are convex and compact sets, $F(\mathbf{u},\mathbf{v})$ is possibly non-convex in $\mathbf{u}$ and non-concave in $\mathbf{v}$, ξ is a random variable following an unknown distribution $\mathcal{D}$. In the context of GANs, $\mathbf{u}$ and $\mathbf{v}$ stand for the parameters for generator and discriminator respectively.

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada.

If $F(\mathbf{u}, \mathbf{v})$ is convex in $\mathbf{u}$ and concave in $\mathbf{v}$, there is a series of work [18, 16, 19, 17, 8, 4] providing algorithms and establishing non-asymptotic convergence to saddle point, i.e. the point $(\mathbf{u}_*, \mathbf{v}_*)$ such that $F(\mathbf{u}_*, \mathbf{v}) \leq F(\mathbf{u}_*, \mathbf{v}_*) \leq F(\mathbf{u}, \mathbf{v}_*)$ for any $\mathbf{u} \in \mathcal{U}$ and any $\mathbf{v} \in \mathcal{V}$. However, when $F(\mathbf{u}, \mathbf{v})$ is non-convex in $\mathbf{u}$ and non-concave in $\mathbf{v}$, finding the saddle point is NP-hard in general. In this case, recent works commonly resort to finding an ϵ-stationary point if the function is smooth, i.e., to find a point $(\mathbf{u}, \mathbf{v})$ such that $\|\mathbf{g}(\mathbf{u}, \mathbf{v})\| \leq \epsilon$, where $\mathbf{g}(\mathbf{u}, \mathbf{v}) = [\nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}), -\nabla_{\mathbf{v}} F(\mathbf{u}, \mathbf{v})]^\top$. Note that this is a necessary condition for finding a (local) saddle point. There are several works [7, 13, 23] establishing non-asymptotic convergence to ϵ-stationary point for non-convex non-concave min-max problems under various assumptions, and others [3, 4, 14, 2] focus on GAN training and get good empirical performance. However, all of them focus on single-machine setting. In contrast, our proposed algorithm is suitable for multiple machines with decentralized network communication. Our main contributions are:

- We design and analyze an algorithm called Decentralized Parallel Optimistic Stochastic Gradient (DPOSG) and establish its non-asymptotic convergence.
- Our empirical studies demonstrate the effectiveness of the proposed algorithm. We demonstrate that DPOSG enjoys speedup compared with a single machine baseline when training WGAN-GP on CIFAR10 .

2 Preliminaries and Notations

We use $\|\cdot\|$ to denote the Euclidean norm. At every point $\mathbf{x} \in \mathcal{X}$, we only have access to a noisy observation of $\mathbf{g}$, i.e., $\mathbf{g}(\mathbf{x}; \xi) = [\nabla_{\mathbf{u}} f(\mathbf{u}, \mathbf{v}; \xi), -\nabla_{\mathbf{v}} f(\mathbf{u}, \mathbf{v}; \xi)]^\top$, where ξ is a random variable. In the subsequent of this paper, we use the term *stochastic gradient* to stand for $\mathbf{g}(\mathbf{x}; \xi)$.

Throughout the paper, we make the following assumption:

Assumption 1. *(i).* $\mathbf{g}$ *is L-Lipschitz continuous, i.e.* $\|\mathbf{g}(\mathbf{x}_1) - \mathbf{g}(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$ *for* $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$.

(ii). *For* $\forall \mathbf{x} \in \mathcal{X}$, $\mathbb{E}\left[\mathbf{g}(\mathbf{x}; \xi)\right] = \mathbf{g}(\mathbf{x})$, $\mathbb{E}\left\|\mathbf{g}(\mathbf{x}; \xi) - \mathbf{g}(\mathbf{x})\right\|^2 \leq \sigma^2$.

(iii). $\|\mathbf{g}(\mathbf{x})\| \leq G$ *for* $\forall \mathbf{x} \in \mathcal{X}$.

(iv). *There exists* $\mathbf{x}_*$ *such that* $\langle \mathbf{g}(\mathbf{x}), \mathbf{x} - \mathbf{x}_* \rangle \geq 0$.

Remark: The Assumptions (i), (ii), (iii) are usually made in optimization literature and are standard. The Assumption (iv) is used frequently in previous works for analyzing algorithms for solving non-monotone variational inequalities [7, 14]. For nonconvex minimization problem, it has been shown that this assumption holds when using SGD for learning neural networks [10, 9, 28].

3 Decentralized Parallel Optimistic Stochastic Gradient

In this section, inspired by the algorithm in [7, 3] in the single-machine setting, we propose an algorithm named Decentralized Parallel Optimistic Stochastic Gradient (DPOSG), which only allows decentralized communications between workers and there is no central node as in the centralized setting which requires communication with each node concurrently in each iteration. Instead, information is only exchanged between neighborhood nodes in the decentralized setting.

Suppose we have M machines. Denote $W \in \mathbb{R}^{M \times M}$ by a doubly stochastic matrix which satisfies $0 \leq W_{ij} \leq 1$, $W^\top = W$, $\sum_{j=1}^{M} W_{ij} = 1$ for $i, j = 1, \ldots, M$. In distributed optimization literature, W is used to characterize the decentralized communication topology, in which W_{ij} characterizes the degree of how node j is able to affect node i, and $W_{ij} = 0$ means node i and j are disconnected.

Denote $\lambda_i(\cdot)$ by the i-th largest eigenvalue of W, then we know that $\lambda_1(W) = 1$. In addition, we assume that $\max\left(|\lambda_2(W)|, |\lambda_M(W)|\right) < 1$.

Denote $\mathbf{z}_k^i \in \mathbb{R}^{d \times 1}$ (and $\mathbf{x}_k^i \in \mathbb{R}^{d \times 1}$) by the parameters in i-th machine at k-th iteration, and both $\mathbf{z}_k^i$ and $\mathbf{x}_k^i$ have the same shape of trainable parameter of neural networks (the trainable parameters of discriminator and generator are concatenated together in GAN setting). Define $Z_k = \left[\mathbf{z}_k^1, \ldots, \mathbf{z}_k^M\right] \in \mathbb{R}^{d \times M}$, $X_k = \left[\mathbf{x}_k^1, \ldots, \mathbf{x}_k^M\right] \in \mathbb{R}^{d \times M}$, $\mathbf{g}(Z_k) = \left[\mathbf{g}(\mathbf{z}_k^1), \ldots, \mathbf{g}(\mathbf{z}_k^M)\right] \in \mathbb{R}^{d \times M}$, $\widehat{\mathbf{g}}(\xi_k, Z_k) = \left[\mathbf{g}(\mathbf{z}_k^1; \xi_k^1), \ldots, \mathbf{g}(\mathbf{z}_k^M; \xi_k^M)\right] \in \mathbb{R}^{d \times M}$, where Z_k, X_k are concatenation of all local

variables, $\widehat{\mathbf{g}}(Z_k), \mathbf{g}(Z_k)$ are concatenations of all local stochastic gradients and their corresponding unbiased estimates. The Algorithm is presented Algorithm 1.

Algorithm 1 Decentralized Parallel Optimistic Stochastic Gradient (DPOSG)

1: **Input:** $Z_0 = X_0 = \mathbf{0}_{d\times M}$
2: **for** $k = 1, \ldots, N$ **do**
3: $\quad Z_k = X_{k-1}W^t - \eta \cdot \widehat{\mathbf{g}}(\xi_{k-1}, Z_{k-1})$
4: $\quad X_k = X_{k-1}W^t - \eta \cdot \widehat{\mathbf{g}}(\xi_k, Z_k)$
5: **end for**

Remark: In both line 3 and line 4, $X_{k-1}W^t$ is the weight averaging step, which can be implemented in parallel with stochastic gradient calculation step (evaluating $\widehat{\mathbf{g}}(\xi_{k-1}, Z_{k-1})$ and $\widehat{\mathbf{g}}(\xi_k, Z_k)$). When we encounter a large batch in training deep neural networks, the running time spent on stochastic gradient calculation usually dominates that on the weight averaging step in every iteration, so the elapsed time in this case is almost the same as the time spent on the gradient calculation step. This feature makes our algorithm practical and numerically attractive.

To establish our convergence result, we need the following assumption.

Assumption 2. $\|\mathbf{x}_*\| \leq \frac{D}{2}$, $\|\mathbf{z}_k^i\| \leq \frac{D}{2}$, $\|\mathbf{x}_k^i\| \leq \frac{D}{2}$ *for* $k = 1, \ldots, N$, $i = 1, \ldots, M$ *with some* $D > 0$.

Remark: Assumption 2 holds when we use normalization layers in the discriminators and the generator such as spectral normalization of weights [15], that will keep the norms of the weights bounded. Regularization techniques such as weight decay also ensures that the weights of the networks remain bounded throughout the training.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Denote* m *by the size of minibatch used in each machine to estimate stochastic gradient. Define* $\bar{\mathbf{z}}_k = \frac{1}{M}\sum_{i=1}^M \mathbf{z}_k^i$ *and* $\rho = \max(|\lambda_2(W)|, |\lambda_M(W)|) < 1$*, where* $\lambda_i(\cdot)$ *stands for the* i*-th largest eigenvalue. Run Algorithm 1 for* N *iterations, in which* $t \geq \log_{\frac{1}{\rho}}\left(1 + \frac{M\sqrt{mMG^2+\sigma^2}}{4\sigma}\right)$*. Take* $\eta \leq \min\left(\frac{1}{6\sqrt{2}L}, \frac{1-\rho^t}{\sqrt{18cML}}, \frac{\sqrt{1-\rho^t}}{2m^{1/2}M^{3/4}L}\right)$ *with* $c = 321$*. Then we have*

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 \leq 8\left(\frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2}{\eta^2 N} + \frac{20\sigma^2}{mM} + \frac{48(DL\sigma + \sigma^2)}{\sqrt{mM}}\right)$$

Remark: Our goal is to make sure $\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 \leq \epsilon^2$ and Theorem 1 establishes a non-asymptotic ergodic convergence. In the single-machine setting, to find the ϵ-stationary point, the algorithm in [7] requires the minibatch size to be dependent on ϵ. However, in practice, it is not reasonable to assume m to be dependent on ϵ in single-machine setting since the machine has a memory limit. Handling such a large minibatch could incur some significant system overhead. When m is a constant independent of ϵ, then the algorithm in [7] cannot be guaranteed to converge to ϵ-stationary point. In the multiple-machines setting, mM is the effective batch size, and we can choose m to be constant and M to be dependent on ϵ (since there is no restriction about choosing the number of machines M). For example, taking $m = O(1)$, $M = O(\epsilon^{-4})$, $N = O(\epsilon^{-8})$, then the algorithm has $O(\epsilon^{-12})$ total complexity to find ϵ-stationary point. Please note that our bound can be further improved, we consider doing so by using time-varying step-size. In addition, another important measure in distributed computing is the communication complexity on the busiest node [11]. DPOSG has communication complexity $O(t \times$ degree of the network$) = O(\log(1/\epsilon))$, while the complexity is $O(\epsilon^{-2})$ in the centralized counterpart Centralized Parallel Optimistic Stochastic Gradient (CPOSG). More details of CPOSG are presented in the next section.

4 Empirical Studies

Software and Hardware PyTorch 1.0.0 is the underlying DL framework. We use the CUDA 10.1 compiler, the CUDA-aware OpenMPI 3.1.1, and g++ 4.8.5 compiler to build our communication library, which connects with PyTorch via a Python-C interface. The decentralized communicatino scheme in DPOSG is similar to that in [11]. In addition, we overlap the gradients computation with the weights exchanging and averaging to further improve runtime performance. An implicit barrier is enforced at the end of each iteration so that every learner advance in a lock-step. To compare speedup, we also implemented CPOSG. In CPOSG, conceptually every worker runs Optimistic Stochastic

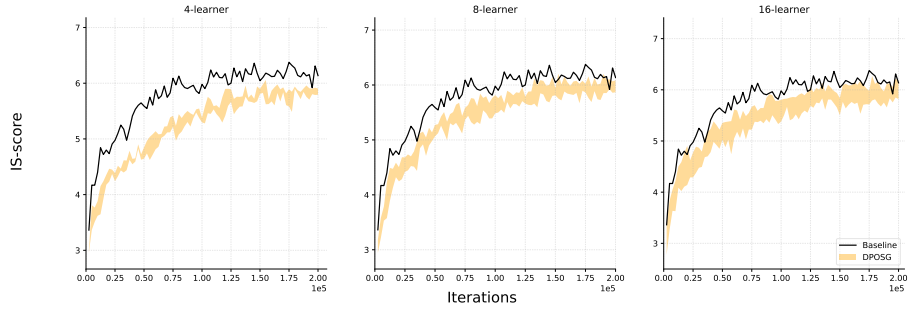

Figure 1: Convergence comparison between baseline and DPOSG. DPOSG matches the baseline convergence rate up to 16 learners. The batch size used in baseline is 256, we fix the total batch size as 256 for each set up.

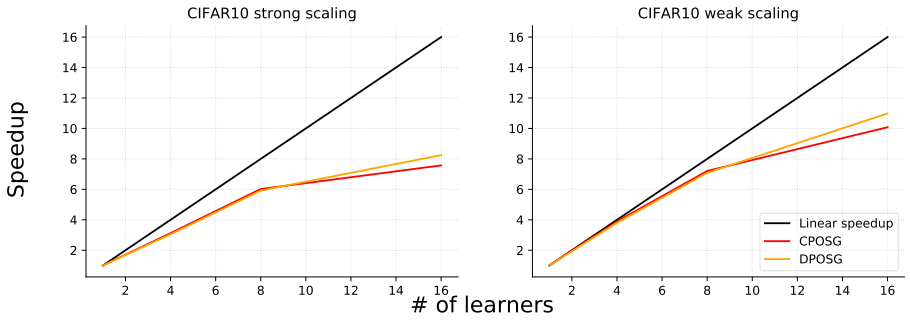

Figure 2: Speedup comparison up to 16 learners. DPOSG outperforms CPOSG, which is based on the state-of-the art Nvidia NCCL allreduce implementation.

Gradient [3] and sends the model parameters to the central node in each iteration, and the central node takes the average of the model parameters and sends it back to each worker. Nvidia NCCL [20], a state-of-the art allreduce implementation, is used as the communication mechanism in CPOSG. We develop and experiment our systems on a cluster which has 4 servers in total. Each server is equipped with 14-core Intel Xeon E5-2680 v4 2.40GHz processor, 1TB main memory, and 4 Nvidia P100 GPUs. GPUs and CPUs are connected via PCIe Gen3 bus, which has a 16GB/s peak bandwidth in each direction. The servers are connected with 100Gbit/s Ethernet.

Dataset, Models and Experimental Results We conduct experiments on Wasserstein GAN with Gradient Penalty [6] on CIFAR10 data. The neural network architectures and hyperparameters are the same as in [6]. Figure 1 depicts the convergence rate of DPOSG, compared to the baseline using single machine (Optimistic Stochastic Gradient on a single machine [3]) . The x-axis is the number of iterations, and the y-axis is the Inception Score (IS) [22]. Baseline was trained with a batch size of 256. We ran experiments with 4 learners (batch size 64 per learner), 8 learners (batch size 32 per learner) and 16 learners (batch size 16 per learner). Since the models on each learner are different at any time, the orange band shows the lowest IS of a learner and the highest IS of a learner when measured. DPOSG largely matches the baseline convergence.

Figure 2 depicts the speedup comparison between CPOSG and DPOSG. We plot both the speedup for strong scaling where total problem size is fixed (i.e. total batch size is fixed to 256) and the speedup for weak scaling where problem size per learner is fixed (i.e., batch size 16 for all learners). DPOSG outperforms CPOSG when number of learners is large (e.g. 16). Our hardware platform runs on a very fast network (i.e., 100 Gbit/s Ethernet). DPOSG is expected to have even better performance than CPOSG when the network is slow [11, 12].

5 Conclusion

We present DPOSG, a decentralized algorithm for a class of min-max problems with provable guarantees and verified by empirical studies. In the future, we plan to run benchmarks on commodity cloud systems where network speed is not as fast to further compare DPOSG and CPOSG. Also, we leave the development of the asynchronous version of DPOSG as proposed in [12, 26, 27] to future work.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[2] Tatjana Chavdarova, Gauthier Gidel, Francois Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *arXiv preprint arXiv:1904.08598*, 2019.

[3] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[4] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[7] AN Iusem, Alejandro Jofré, Roberto I Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

[8] Anatoli Juditsky, Arkadi Nemirovski, Claire Tauvel, et al. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[9] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

[10] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[11] 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.

[12] Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. Asynchronous decentralized parallel stochastic gradient descent. In *ICML*, 2018.

[13] Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 2018.

[14] Panayotis Mertikopoulos, Houssam Zenati, Bruno Lecouat, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

[15] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[16] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[17] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[18] Arkadi Nemirovski and D Yudin. On cezari?s convergence of the steepest descent method for approximating saddle point of convex-concave functions. In *Soviet Math. Dokl*, volume 19, pages 258–269, 1978.

[19] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[20] Nvidia. *NCCL: Optimized primitives for collective multi-GPU communication*. Available at `https://github.com/NVIDIA/nccl`.

[21] S Sundhar Ram, A Nedić, and Venugopal V Veeravalli. Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 3581–3586. IEEE, 2009.

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[23] Maziar Sanjabi, Meisam Razaviyayn, and Jason D Lee. Solving non-convex non-concave min-max games under polyak-{\L} ojasiewicz condition. *arXiv preprint arXiv:1812.02878*, 2018.

[24] Feng Yan, Shreyas Sundaram, SVN Vishwanathan, and Yuan Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2012.

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

[26] Wei Zhang, Xiaodong Cui, Ulrich Finkler, Brian Kingsbury, George Saon, David Kung, and Michael Picheny. Distributed deep learning strategies for automatic speech recognition. In *ICASSP'2019*, May 2019.

[27] Wei Zhang, Xiaodong Cui, Ulrich Finkler, Abdullah Saon, George Kayi, Alper Buyuktosunoglu, Brian Kingsbury, David Kung, and Michael Picheny. A highly efficient distributed deep learning system for automatic speechrecognition. In *INTERSPEECH'2019*, Sept 2019.

[28] Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

A Proof of Theorem 1

Denote $\|\cdot\|$ by the ℓ_2 norm or the matrix spectral norm depending on the argument. Denote $\|\cdot\|_F$ by the matrix Frobenius norm. Define $\epsilon_k^i = \mathbf{g}(\mathbf{z}_k^i;\xi_k^i) - \mathbf{g}(\mathbf{z}_k^i)$, $\widehat{\mathbf{g}}(\epsilon_k^i, \mathbf{z}_k^i) = \mathbf{g}(\mathbf{z}_k^i;\xi_k^1)$, $\widehat{\mathbf{g}}(\epsilon_k, Z_k) = \left[\mathbf{g}(\mathbf{z}_k^1;\xi_k^1), \ldots, \mathbf{g}(\mathbf{z}_k^M;\xi_k^M)\right] \in \mathbb{R}^{d\times M}$, $\bar{\mathbf{z}}_k = \frac{1}{M}\sum_{i=1}^M \mathbf{z}_k^i$. Define $\mathbf{e}_i = [0,\ldots,1,\ldots,0]^\top$, where 1 appears at the i-th coordinate and the rest entries are all zeros.

A.1 Propositions

We first present several propositions which are useful for further analysis.

Proposition 1 (Lemma 4 in [11]). *For any doubly stochastic matrix W where $0 \le W_{ij} \le 1$, $W^\top = W$, $\sum_{j=1}^M W_{ij} = 1$ for $i, j = 1,\ldots,M$. Define $\rho = \max\left(|\lambda_2(W)|, |\lambda_M(W)|\right) < 1$. Then $\left\|\frac{1}{M}\mathbf{1}_M - W^t\mathbf{e}_i\right\| \le \rho^t$, for $\forall i \in \{1,\ldots,M\}$, $t \in \mathbb{N}$.*

Proposition 2 (Lemma 5 in [11]).

$$\mathbb{E}\left\|\mathbf{g}(Z_j)\right\|^2 \le \sum_{h=1}^M 3\mathbb{E}L^2 \left\|\frac{\sum_{i'=1}^M \mathbf{z}_j^i}{M} - \mathbf{z}_j^h\right\|^2 + 3\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2$$

A.2 Useful Lemmas

Inspired by [11], we introduce the Lemma 1 which is useful for our analysis.

Lemma 1.

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\frac{1}{M}\mathbf{g}(Z_k)\mathbf{1}_M - \mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 \le \frac{2\eta^2 ML^2\sigma^2}{(1-\rho^t)\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)} + \frac{18\eta^2ML^2}{(1-\rho^t)^2\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)}\cdot\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}\left(\bar{\mathbf{z}}_k\right)\right\|^2$$

Proof.

$$\mathbb{E}\left\|\frac{1}{M}\mathbf{g}(Z_k)\mathbf{1}_M - \mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 \le \frac{1}{M}\sum_{i=1}^M\mathbb{E}\left\|\mathbf{g}(\mathbf{z}_k^i) - \mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 \le \frac{L^2}{M}\sum_{i=1}^M\mathbb{E}\left\|\mathbf{z}_k^i - \bar{\mathbf{z}}_k\right\|^2$$

$$= \frac{L^2}{M}\sum_{i=1}^M\mathbb{E}\left\|\frac{1}{M}X_{k-1}W^t\mathbf{1}_M - \frac{\eta}{M}\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M - \left(X_{k-1}W^t\mathbf{e}_i - \eta\widehat{\mathbf{g}}(\epsilon_{k-1}, \mathbf{z}_{k-1})\mathbf{e}_i\right)\right\|^2$$

$$\overset{(a)}{=} \frac{L^2}{M}\sum_{i=1}^M\mathbb{E}\left\|\frac{1}{M}X_0\mathbf{1}_M - \frac{\eta}{M}\sum_{j=0}^{k-1}\widehat{\mathbf{g}}(\epsilon_j, Z_j)\mathbf{1}_M - \left(X_0W^{tk}\mathbf{e}_i - \eta\sum_{j=0}^{k-1}\widehat{\mathbf{g}}(\epsilon_j, Z_j)W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2$$

$$\overset{(b)}{=} \frac{L^2}{M}\sum_{i=1}^M\mathbb{E}\left\|\eta\sum_{j=0}^{k-1}\widehat{\mathbf{g}}(\epsilon_j, Z_j)\left(\frac{\mathbf{1}_M}{M} - W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2$$

$$\le \frac{L^2}{M}\cdot 2\eta^2\sum_{i=1}^M\left[\mathbb{E}\left\|\sum_{j=0}^{k-1}\left(\widehat{\mathbf{g}}(\epsilon_j, Z_j) - \mathbf{g}(Z_j)\right)\left(\frac{\mathbf{1}_M}{M} - W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2 + \mathbb{E}\left\|\sum_{j=0}^{k-1}\mathbf{g}(Z_j)\left(\frac{\mathbf{1}_M}{M} - W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2\right] \tag{2}$$

where (a) holds since $W\mathbf{1}_M = \mathbf{1}_M$, (b) holds since $X_0 = \mathbf{0}_{d\times M}$. Note that

$$\mathbb{E}\left\|\sum_{j=0}^{k-1}\left(\widehat{\mathbf{g}}(\epsilon_j, Z_j) - \mathbf{g}(Z_j)\right)\left(\frac{\mathbf{1}_M}{M} - W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2 \overset{(a)}{=} \sum_{j=0}^{k-1}\mathbb{E}\left\|\left(\widehat{\mathbf{g}}(\epsilon_j, Z_j) - \mathbf{g}(Z_j)\right)\left(\frac{\mathbf{1}_M}{M} - W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2$$

$$\le \sum_{j=0}^{k-1}\mathbb{E}\left\|\left(\widehat{\mathbf{g}}(\epsilon_j, Z_j) - \mathbf{g}(Z_j)\right)\right\|^2\cdot\left\|\left(\frac{\mathbf{1}_M}{M} - W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2$$

$$\le \sum_{j=0}^{k-1}\mathbb{E}\left\|\left(\widehat{\mathbf{g}}(\epsilon_j, Z_j) - \mathbf{g}(Z_j)\right)\right\|_F^2\cdot\left\|\left(\frac{\mathbf{1}_M}{M} - W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2 \overset{(b)}{\le} \frac{M\sigma^2}{1-\rho^t}\sum_{j=0}^{k-1}\rho^{t(k-j-1)} \le \frac{M\sigma^2}{1-\rho^t} \tag{3}$$

where (a) holds since ϵ_i, ϵ_j with $i \neq j$ are mutually conditionally independent of each other, (b) holds because of Proposition 1. In addition, note that

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{j=0}^{k-1}\mathbf{g}(Z_j)\left(\frac{\mathbf{1}_M}{M}-W^{k-j-1}\mathbf{e}_i\right)\right\|^2 \\
&=\sum_{j=0}^{k-1}\mathbb{E}\left\|\mathbf{g}(Z_j)\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2+\sum_{j\neq j'}\mathbb{E}\left\langle\mathbf{g}(Z_j)\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right),\mathbf{g}(Z_{j'})\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j'-1)}\mathbf{e}_i\right)\right\rangle \\
&:=\mathbf{I}_1+\mathbf{I}_2
\end{aligned} \tag{4}$$

Then we try to bound $\mathbf{I}_1$ and $\mathbf{I}_2$ respectively.

$$\begin{aligned}
\mathbf{I}_1 &\leq \sum_{j=0}^{k-1}\mathbb{E}\left\|\mathbf{g}(Z_j)\right\|^2\cdot\left\|\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2 \\
&\overset{(a)}{\leq} 3\sum_{j=0}^{k-1}\left(\sum_{h=1}^{M}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i'}}{M}-\mathbf{z}_j^h\right\|^2+\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\right)\cdot\left\|\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2
\end{aligned}$$

where (a) comes from the Proposition 2. For $\mathbf{I}_2$, note that

$$\begin{aligned}
\mathbf{I}_2 &\leq \sum_{j\neq j'}\mathbb{E}\left\|\mathbf{g}(Z_j)\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right)\right\|\cdot\left\|\mathbf{g}(Z_{j'})\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j'-1)}\mathbf{e}_i\right)\right\| \\
&\leq \sum_{j\neq j'}\mathbb{E}\left\|\mathbf{g}(Z_j)\right\|\left\|\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right\|\left\|\mathbf{g}(Z_{j'})\right\|\left\|\frac{\mathbf{1}_M}{M}-W^{t(k-j'-1)}\mathbf{e}_i\right\| \\
&\leq \sum_{j\neq j'}\frac{1}{2}\mathbb{E}\left\|\mathbf{g}(Z_j)\right\|^2\left\|\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right\|\left\|\frac{\mathbf{1}_M}{M}-W^{t(k-j'-1)}\mathbf{e}_i\right\| \\
&\quad+\sum_{j\neq j'}\frac{1}{2}\mathbb{E}\left\|\mathbf{g}(Z_{j'})\right\|^2\left\|\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right\|\left\|\frac{\mathbf{1}_M}{M}-W^{t(k-j'-1)}\mathbf{e}_i\right\| \\
&\overset{(a)}{\leq} \sum_{j\neq j'}\frac{1}{2}\mathbb{E}\left(\left\|\mathbf{g}(Z_j)\right\|^2+\left\|\mathbf{g}(Z_{j'})\right\|^2\right)\rho^{t(2k-(j+j')-2)}=\sum_{j\neq j'}\mathbb{E}\left\|\mathbf{g}(Z_j)\right\|^2\rho^{t(2k-(j+j')-2)} \\
&\overset{(b)}{\leq} 3\sum_{j\neq j'}\left(\sum_{h=1}^{M}\mathbb{E}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i}}{M}-\mathbf{z}_j^h\right\|^2+\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\right)\rho^{t(2k-(j+j')-2)} \\
&= 6\sum_{j=0}^{k-1}\left(\sum_{h=1}^{M}\mathbb{E}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i}}{M}-\mathbf{z}_j^h\right\|^2+\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\right)\sum_{j'=j+1}^{k-1}\rho^{t(2k-(j+j')-2)} \\
&\leq 6\sum_{j=0}^{k-1}\left(\sum_{h=1}^{M}\mathbb{E}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i}}{M}-\mathbf{z}_j^h\right\|^2+\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\right)\frac{\rho^{t(k-j-1)}}{1-\rho^t}
\end{aligned}$$

where (a) holds by Proposition 1, (b) comes from Proposition 2.

Using the bound of $\mathbf{I}_1$, $\mathbf{I}_2$ and by (3) and (4), we know that

$$
\begin{aligned}
&2\eta^2\left[\mathbb{E}\left\|\sum_{j=0}^{k-1}\left(\widehat{\mathbf{g}}(\epsilon_j,Z_j)-\mathbf{g}(Z_j)\right)\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2+\mathbb{E}\left\|\sum_{j=0}^{k-1}\mathbf{g}(Z_j)\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2\right]\\
&\le \frac{2\eta^2M\sigma^2}{1-\rho^t}+6\eta^2\sum_{j=0}^{k-1}\sum_{h=1}^{M}\mathbb{E}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i'}}{M}-\mathbf{z}_j^h\right\|^2\cdot\left\|\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2\\
&\quad+6\eta^2\sum_{j=0}^{k-1}\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\cdot\left\|\left(\frac{\mathbf{1}_M}{M}-W^{t(k-j-1)}\mathbf{e}_i\right)\right\|^2\\
&\quad+12\eta^2\sum_{j=0}^{k-1}\left(\sum_{h=1}^{M}\mathbb{E}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i}}{M}-\mathbf{z}_j^h\right\|^2+\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\right)\frac{\rho^{t(k-j-1)}}{1-\rho^t}\\
&\overset{(a)}{\le}\frac{2\eta^2M\sigma^2}{1-\rho^t}+6\eta^2\sum_{j=0}^{k-1}\sum_{h=1}^{M}\mathbb{E}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i'}}{M}-\mathbf{z}_j^h\right\|^2\rho^{2t(k-j-1)}+6\eta^2\sum_{j=0}^{k-1}\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\rho^{2t(k-j-1)}\\
&\quad+12\eta^2\sum_{j=0}^{k-1}\left(\sum_{h=1}^{M}\mathbb{E}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i}}{M}-\mathbf{z}_j^h\right\|^2+\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\right)\frac{\rho^{t(k-j-1)}}{1-\rho^t}\\
&=\frac{2\eta^2M\sigma^2}{1-\rho^t}+6\eta^2\sum_{j=0}^{k-1}\sum_{h=1}^{M}\mathbb{E}L^2\left\|\frac{\sum_{i'=1}^{M}\mathbf{z}_j^{i'}}{M}-\mathbf{z}_j^h\right\|^2\left(\rho^{2t(k-j-1)}+\frac{2\rho^{t(k-j-1)}}{1-\rho^t}\right)\\
&\quad+6\eta^2\sum_{j=0}^{k-1}\mathbb{E}\left\|T\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\left(\rho^{2t(k-j-1)}+\frac{2\rho^{t(k-j-1)}}{1-\rho^t}\right)
\end{aligned}
\tag{5}
$$

where (a) comes from Proposition 1.

Define $\lambda_k=\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{z}_k^i-\bar{\mathbf{z}}_k\right\|^2$. By (5) and (2), then we have

$$
\begin{aligned}
\mathbb{E}\left[\lambda_k\right]\le&\frac{2\eta^2M\sigma^2}{1-\rho^t}+6\eta^2\sum_{j=0}^{k-1}\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\left(\rho^{2t(k-j-1)}+\frac{2\rho^{t(k-j-1)}}{1-\rho^t}\right)\\
&+6\eta^2ML^2\sum_{j=0}^{k-1}\mathbb{E}\left[\lambda_j\right]\left(\rho^{2t(k-j-1)}+\frac{2\rho^{t(k-j-1)}}{1-\rho^t}\right).
\end{aligned}
\tag{6}
$$

Summing over $k=0,\dots,N-1$ on both sides of (6) yield

$$
\begin{aligned}
\sum_{k=0}^{N-1}\mathbb{E}\left[\lambda_k\right]\le&\frac{2\eta^2M\sigma^2N}{1-\rho^t}+6\eta^2\sum_{k=0}^{N-1}\sum_{j=0}^{k-1}\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_j\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\left(\rho^{2t(k-j-1)}+\frac{2\rho^{t(k-j-1)}}{1-\rho^t}\right)\\
&+6\eta^2ML^2\sum_{k=0}^{N-1}\sum_{j=0}^{k-1}\mathbb{E}\left[\lambda_j\right]\left(\rho^{2t(k-j-1)}+\frac{2\rho^{t(k-j-1)}}{1-\rho^t}\right)\\
\le&\frac{2\eta^2M\sigma^2N}{1-\rho^t}+6\eta^2\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_k\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2\left(\sum_{i=0}^{\infty}\rho^{2ti}+\frac{2\sum_{i=0}^{\infty}\rho^{ti}}{1-\rho^t}\right)\\
&+6\eta^2ML^2\sum_{k=0}^{N-1}\mathbb{E}\left[\lambda_k\right]\left(\sum_{i=0}^{\infty}\rho^{2ti}+\frac{2\sum_{i=0}^{\infty}\rho^{ti}}{1-\rho^t}\right)\\
\le&\frac{2\eta^2M\sigma^2N}{1-\rho^t}+\frac{18\eta^2}{(1-\rho^t)^2}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_k\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2+\frac{18\eta^2ML^2}{(1-\rho^t)^2}\sum_{k=0}^{N-1}\mathbb{E}\left[\lambda_k\right]
\end{aligned}
\tag{7}
$$

Rearrange the terms in (7), and we have

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\lambda_k\right] \leq \frac{2\eta^2 M\sigma^2}{(1-\rho^t)\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)} + \frac{18\eta^2}{(1-\rho^t)^2\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)}\cdot\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_k\mathbf{1}_M}{M}\right)\mathbf{1}_M^\top\right\|^2$$
$$= \frac{2\eta^2 M\sigma^2}{(1-\rho^t)\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)} + \frac{18\eta^2 M}{(1-\rho^t)^2\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)}\cdot\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}\left(\frac{Z_k\mathbf{1}_M}{M}\right)\right\|^2 \tag{8}$$

Combining (8) and (2) suffices to prove the lemma. □

Based on Lemma 1 and by carefully choose the stepsize in the algorithm, we have the following Lemma 2.

Lemma 2. *By taking* $\eta = \min\left(\frac{1-\rho^t}{\sqrt{18cML}}, \frac{\sqrt{1-\rho^t}}{2m^{1/2}M^{3/4}L}\right)$ *with* $c \geq 2$*, we have*

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\frac{1}{M}\mathbf{g}(Z_k)\mathbf{1}_M - \mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 \leq \frac{\sigma^2}{\sqrt{mM}} + \frac{1}{c-1}\cdot\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}\left(\bar{\mathbf{z}}_k\right)\right\|^2$$

Proof. By $\eta^2 \leq \frac{(1-\rho^t)^2}{18cML^2} \leq \frac{(1-\rho^t)^2}{36ML^2}$, we know that $1-\frac{18\eta^2ML^2}{(1-\rho^t)^2} \geq \frac{1}{2}$. In addition, since $\eta^2 \leq \frac{1-\rho^t}{4\sqrt{mM^3}L^2}$, we know that

$$\frac{2\eta^2ML^2\sigma^2}{(1-\rho^t)\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)} \leq \frac{4\eta^2ML^2\sigma^2}{1-\rho^t} \leq \frac{\sigma^2}{\sqrt{mM}} \tag{9}$$

Note that $\frac{18\eta^2ML^2}{(1-\rho^t)^2\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)} = \frac{18ML^2}{\frac{(1-\rho^t)^2}{\eta^2}-18ML^2}$ is monotonically increasing in terms of η^2, and $\eta^2 \leq \frac{(1-\rho^t)^2}{18cML^2}$, and hence we have

$$\frac{18\eta^2ML^2}{(1-\rho^t)^2\left(1-\frac{18\eta^2ML^2}{(1-\rho^t)^2}\right)} \leq \frac{1}{c-1}. \tag{10}$$

Combining (9), (10) and Lemma 1 suffices to show the result. □

Lemma 3.

$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\frac{1}{M}\widehat{\mathbf{g}}(\epsilon_{k-1},Z_{k-1})\mathbf{1}_M - \widehat{\mathbf{g}}(\epsilon_{k-1},Z_{k-1})\mathbf{e}_i\right\|\right] \leq \frac{2\sigma}{\sqrt{mM}} + 2\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{g}(\mathbf{z}_{k-1}^i) - \mathbf{g}(\bar{\mathbf{z}}_{k-1})\right\|\right],$$

Proof.

$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\frac{1}{M}\widehat{\mathbf{g}}(\epsilon_{k-1},Z_{k-1})\mathbf{1}_M - \widehat{\mathbf{g}}(\epsilon_{k-1},Z_{k-1})\mathbf{e}_i\right\|\right] = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\sum_{j=1}^{M}\frac{\widehat{\mathbf{g}}(\epsilon_{k-1}^j,\mathbf{z}_{k-1}^i) - \widehat{\mathbf{g}}(\epsilon_{k-1}^i,\mathbf{z}_{k-1}^j)}{M}\right\|\right]$$
$$= \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\sum_{j=1}^{M}\frac{\widehat{\mathbf{g}}(\epsilon_{k-1}^j,\mathbf{z}_{k-1}^j) - \mathbf{g}(\mathbf{z}_{k-1}^j) + \mathbf{g}(\mathbf{z}_{k-1}^j) - \mathbf{g}(\bar{\mathbf{z}}_{k-1}) + \mathbf{g}(\bar{\mathbf{z}}_{k-1}) - \mathbf{g}(\mathbf{z}_{k-1}^i) + \mathbf{g}(\mathbf{z}_{k-1}^i) - \widehat{\mathbf{g}}(\epsilon_{k-1}^i,\mathbf{z}_{k-1}^i)}{M}\right\|\right]$$
$$\overset{(a)}{=} \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\sum_{j=1}^{M}\frac{\epsilon_{k-1}^j + \mathbf{g}(\mathbf{z}_{k-1}^j) - \mathbf{g}(\bar{\mathbf{z}}_{k-1}) + \mathbf{g}(\bar{\mathbf{z}}_{k-1}) - \mathbf{g}(\mathbf{z}_{k-1}^i) - \epsilon_{k-1}^i}{M}\right\|\right]$$
$$\overset{(b)}{\leq} \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{j=1}^{M}\left(\epsilon_{k-1}^j - \epsilon_{k-1}^i\right)\right\| + \frac{1}{M}\sum_{j=1}^{M}\left\|\mathbf{g}(\mathbf{z}_{k-1}^j) - \mathbf{g}(\bar{\mathbf{z}}_{k-1})\right\| + \left\|\mathbf{g}(\mathbf{z}_{k-1}^i) - \mathbf{g}(\bar{\mathbf{z}}_{k-1})\right\|\right]$$
$$\leq \frac{2\sigma}{\sqrt{mM}} + 2\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{g}(\mathbf{z}_{k-1}^i) - \mathbf{g}(\bar{\mathbf{z}}_{k-1})\right\|\right]$$

where (a) holds by the definition of ϵ_{k-1}^i and ϵ_{k-1}^j, (b) holds by the triangle inequality of norm, (c) holds since ϵ_{k-1}^j and ϵ_{k-1}^i with $i \neq j$ are conditionally mutually independent of each other and the fact that $\mathbb{E}\|\mathbf{x}\| \leq \sqrt{\mathbb{E}\|\mathbf{x}\|^2}$. □

Lemma 4. *Define* $\mu_k = \frac{1}{M}\sum_{i=1}^M \left\|\mathbf{g}(\mathbf{z}_k^i) - \mathbf{g}(\bar{\mathbf{z}}_k)\right\|$. *Suppose* $\eta < \frac{1}{4L}$ *and* $t \geq \log_{\frac{1}{\rho}}\left(1 + \frac{M\sqrt{mMG^2+\sigma^2}}{4\sigma}\right)$. *We have*

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\mu_k\right] < \frac{8\eta L\sigma}{\sqrt{mM}(1-4\eta L)}.$$

Proof. Define $\mathbf{e}_i = [0, \ldots, 1, \ldots, 0]^\top$, where 1 appears at the i-th coordinate and the rest entries are all zeros. Then we have

$$\begin{aligned}
\mu_k &\overset{(a)}{\leq} \frac{L}{M}\sum_{i=1}^M \left\|\mathbf{z}_k^i - \bar{\mathbf{z}}_k\right\| \\
&\overset{(b)}{=} \frac{L}{M}\sum_{i=1}^M \left\|\frac{1}{M}X_{k-1}\mathbf{1}_M - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M - \left(X_{k-1}W^t\mathbf{e}_i - \eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{e}_i\right)\right\| \\
&= \frac{L}{M}\sum_{i=1}^M \left\|\frac{1}{M}\left(X_0 - \eta\sum_{j=0}^{k-1}\widehat{\mathbf{g}}(\epsilon_j, Z_j) - \eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\right)\mathbf{1}_M \right. \\
&\qquad \left. - \left(X_0W^{tk} - \eta\sum_{j=0}^{k-1}\widehat{\mathbf{g}}(\epsilon_j, Z_j)W^{(k-j-1)t} - \eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\right)\mathbf{e}_i\right\| \\
&= \frac{L}{M}\sum_{i=1}^M \left\|\eta\sum_{j=0}^{k-2}\widehat{\mathbf{g}}(\epsilon_j, Z_j)\left(\frac{1}{M}\mathbf{1}_M - W^{(k-j-1)t}\mathbf{e}_i\right)\right\| + \frac{2\eta L}{M}\sum_{i=1}^M \left\|\frac{1}{M}\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M - \widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{e}_i\right\| \\
&\overset{(c)}{\leq} \frac{L\eta}{M}\sum_{i=1}^M \left\|\sum_{j=0}^{k-2}\widehat{\mathbf{g}}(\epsilon_j, Z_j)\rho^{(k-j-1)t}\right\| + \frac{2\eta L}{M}\sum_{i=1}^M \left\|\frac{1}{M}\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M - \widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{e}_i\right\| \\
&\leq \frac{L\eta}{M}\sum_{i=1}^M\sum_{j=0}^{k-2}\rho^{(k-j-1)t}\left\|\widehat{\mathbf{g}}(\epsilon_j, Z_j)\right\|_F + \frac{2\eta L}{M}\sum_{i=1}^M \left\|\frac{1}{M}\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M - \widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{e}_i\right\|
\end{aligned} \tag{11}$$

where (a) holds by the L-Lipschitz continuity of $\mathbf{g}$, (b) holds by the update and $W\mathbf{1}_M = \mathbf{1}_M$, (c) holds by Proposition 1.

Taking expectation over $\epsilon_0, \ldots, \epsilon_{k-1}$ on both sides of (11) yields

$$\begin{aligned}
\mathbb{E}\left[\mu_k\right] &\overset{(a)}{\leq} \frac{\eta L\rho^t}{1-\rho^t}M\sqrt{G^2 + \frac{\sigma^2}{m}} + \frac{2\eta L}{M}\sum_{i=1}^M \mathbb{E}\left[\left\|\frac{1}{M}\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M - \widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{e}_i\right\|\right] \\
&\overset{(b)}{\leq} \frac{\eta L\rho^t}{1-\rho^t}M\sqrt{G^2 + \frac{\sigma^2}{m}} + 2\eta L\left(2\mathbb{E}\left[\mu_{k-1}\right] + \frac{2\sigma}{\sqrt{Mm}}\right) \\
&= \frac{\eta L\rho^t}{1-\rho^t}M\sqrt{G^2 + \frac{\sigma^2}{m}} + 4\eta L\left(\mathbb{E}\left[\mu_{k-1}\right] + \frac{\sigma}{\sqrt{mM}}\right)
\end{aligned} \tag{12}$$

where (a) holds since $\mathbb{E}\left[\mathbf{g}(\mathbf{x};\xi)\right] = \mathbf{g}(\mathbf{x})$, $\mathbb{E}\left\|\mathbf{g}(\mathbf{x};\xi) - \mathbf{g}(\mathbf{x})\right\|^2 \leq \sigma^2$, $\|\mathbf{g}(\mathbf{x})\| \leq G$, and (b) holds by invoking Lemma 3.

Define $\delta = \frac{\rho^t}{4(1-\rho^t)} M\sqrt{G^2 + \frac{\sigma^2}{m}}$. By taking $t \geq \log_{\frac{1}{\rho}} \left(1 + \frac{M\sqrt{mMG^2+\sigma^2}}{4\sigma}\right)$, we can show that $\delta \leq \frac{\sigma}{\sqrt{mM}}$. Hence we can rewrite (12) as

$$\mathbb{E}\left[\mu_k\right] \leq 4\eta L \left(\mathbb{E}\left[\mu_{k-1}\right] + \frac{2\sigma}{\sqrt{mM}}\right).$$

Define $b_k = \mathbb{E}\left[\mu_k\right] + \frac{2\sigma}{\sqrt{mM}}$. Then we have $b_k \leq 4\eta L b_{k-1} + \frac{2\sigma}{\sqrt{mM}}$, which implies that $b_k + \frac{\frac{2\sigma}{\sqrt{mM}}}{4\eta L - 1} \leq 4\eta L \left(b_{k-1} + \frac{\frac{2\sigma}{\sqrt{mM}}}{4\eta L - 1}\right)$. Note that $b_0 = \mathbb{E}\left[\mu_0\right] + \frac{2\sigma}{\sqrt{mM}} = \frac{2\sigma}{\sqrt{mM}}$ and $4\eta L < 1$, and hence we have

$$\frac{1}{N}\sum_{k=0}^{N-1}\left(b_k + \frac{\frac{2\sigma}{\sqrt{mM}}}{4\eta L - 1}\right) \leq \frac{\frac{2\sigma}{\sqrt{mM}}\left(1 + \frac{1}{4\eta L - 1}\right)}{N(1-4\eta L)} < 0.$$

So we know that

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\mu_k\right] = \frac{1}{N}\sum_{k=0}^{N-1} b_k - \frac{2\sigma}{\sqrt{mM}} < \frac{1}{N}\sum_{k=0}^{N-1} b_k < \frac{\frac{2\sigma}{\sqrt{mM}} \cdot 4\eta L}{1 - 4\eta L} = \frac{8\eta L\sigma}{\sqrt{mM}(1-4\eta L)}.$$

Here completes the proof. □

A.3 Main Proof of Theorem 1

Proof. Define $\mathbf{1}_M = [1, \ldots, 1]^\top \in \mathbb{R}^{M\times 1}$, $\bar{\mathbf{z}}_k = \frac{1}{M} Z_k \mathbf{1}_M$, $\bar{\mathbf{x}}_k = \frac{1}{M} X_k \mathbf{1}_M$, $\bar{\epsilon}_k = \frac{1}{M}\sum_{i=1}^M \epsilon_k^i$. By the property of W, we know that $W\mathbf{1}_M = \mathbf{1}_M$.

Noting that for $\forall \mathbf{x} \in \mathcal{X} = \mathbb{R}^d$, we have

$$\begin{aligned}
&\left\|\frac{1}{M}X_k\mathbf{1}_M - \mathbf{x}\right\|^2 = \left\|\frac{1}{M}\left(X_{k-1}W^t - \eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\right)\cdot\mathbf{1}_M - \mathbf{x}\right\|^2 \\
&= \left\|\frac{1}{M}\left(X_{k-1}W^t - \eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\right)\cdot\mathbf{1}_M - \mathbf{x}\right\|^2 - \left\|\frac{1}{M}\left(X_{k-1}W^t - \eta\widehat{\mathbf{g}}(\epsilon_k, Z_k) - X_k\right)\cdot\mathbf{1}_M\right\|^2 \\
&= \left\|\bar{\mathbf{x}}_{k-1} - \mathbf{x}\right\|^2 - \left\|\bar{\mathbf{x}}_{k-1} - \bar{\mathbf{x}}_k\right\|^2 + 2\left\langle \mathbf{x} - \bar{\mathbf{x}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M\right\rangle \\
&= \left\|\bar{\mathbf{x}}_{k-1} - \mathbf{x}\right\|^2 - \left\|\bar{\mathbf{x}}_{k-1} - \bar{\mathbf{x}}_k\right\|^2 + 2\left\langle \mathbf{x} - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M\right\rangle + 2\left\langle \bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M\right\rangle \\
&= \left\|\bar{\mathbf{x}}_{k-1} - \mathbf{x}\right\|^2 - \left\|\bar{\mathbf{x}}_{k-1} - \bar{\mathbf{z}}_k + \bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\right\|^2 + 2\left\langle \mathbf{x} - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M\right\rangle + 2\left\langle \bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M\right\rangle \\
&= \left\|\bar{\mathbf{x}}_{k-1} - \mathbf{x}\right\|^2 - \left\|\bar{\mathbf{x}}_{k-1} - \bar{\mathbf{z}}_k\right\|^2 - \left\|\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\right\|^2 + 2\left\langle \mathbf{x} - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M\right\rangle \\
&\quad + 2\left\langle \bar{\mathbf{x}}_k - \bar{\mathbf{z}}_k, \bar{\mathbf{x}}_{k-1} - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M - \bar{\mathbf{z}}_k\right\rangle
\end{aligned} \tag{13}$$

Note that

$$
\begin{aligned}
&2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M \right\rangle = 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\mathbf{g}(Z_k)\mathbf{1}_M \right\rangle + 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\sum_{i=1}^{M}\epsilon_k^i \right\rangle \\
&= 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \eta\mathbf{g}(\bar{\mathbf{z}}_k) \right\rangle + 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \eta\frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{g}(\mathbf{z}_k^i) - \mathbf{g}(\bar{\mathbf{z}}_k)\right) \right\rangle + 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\sum_{i=1}^{M}\epsilon_k^i \right\rangle \\
&\overset{(a)}{\leq} 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \eta\frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{g}(\mathbf{z}_k^i) - \mathbf{g}(\bar{\mathbf{z}}_k)\right) \right\rangle + 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\sum_{i=1}^{M}\epsilon_k^i \right\rangle \\
&\overset{(b)}{\leq} 2\eta D\left\| \frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{g}(\mathbf{z}_k^i) - \mathbf{g}(\bar{\mathbf{z}}_k)\right) \right\| + 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\sum_{i=1}^{M}\epsilon_k^i \right\rangle \\
&\leq 2\eta D\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{g}(\mathbf{z}_k^i) - \mathbf{g}(\bar{\mathbf{z}}_k)\right\| + 2\left\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\sum_{i=1}^{M}\epsilon_k^i \right\rangle
\end{aligned} \tag{14}
$$

where (a) holds since $\langle \mathbf{x}_* - \bar{\mathbf{z}}_k, \eta\mathbf{g}(\bar{\mathbf{z}}_k)\rangle \leq 0$, (b) holds by Cauchy-Schwarz inequality and $\|\bar{\mathbf{z}}_k - \mathbf{x}_*\| \leq D$. Note that

$$
\begin{aligned}
&2\left\langle \bar{\mathbf{x}}_k - \bar{\mathbf{z}}_k, \bar{\mathbf{x}}_{k-1} - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M - \bar{\mathbf{z}}_k \right\rangle \\
&= 2\left\langle \bar{\mathbf{x}}_k - \bar{\mathbf{z}}_k, \bar{\mathbf{x}}_{k-1} - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M - \bar{\mathbf{z}}_k \right\rangle \\
&\quad + 2\left\langle \bar{\mathbf{x}}_k - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M \right\rangle \\
&\overset{(a)}{=} 2\left\langle \bar{\mathbf{x}}_k - \bar{\mathbf{z}}_k, \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M \right\rangle \\
&\overset{(b)}{\leq} 2\left\| \left(\bar{\mathbf{x}}_{k-1} - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M\right) - \left(\bar{\mathbf{x}}_{k-1} - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M\right) \right\| \\
&\cdot \left\| \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M - \frac{1}{M}\eta\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M \right\| \leq 2\eta^2\left\| \frac{1}{M}\widehat{\mathbf{g}}(\epsilon_k, Z_k)\mathbf{1}_M - \frac{1}{M}\widehat{\mathbf{g}}(\epsilon_{k-1}, Z_{k-1})\mathbf{1}_M \right\|^2 \\
&\overset{(c)}{\leq} 6\eta^2\left\| \frac{1}{M}\left(\mathbf{g}(Z_k) - \mathbf{g}(Z_{k-1})\right)\mathbf{1}_M \right\|^2 + 6\eta^2\|\bar{\epsilon}_k\|^2 + 6\eta^2\|\bar{\epsilon}_{k-1}\|^2 \\
&\overset{(d)}{\leq} 18\eta^2\left\| \frac{1}{M}\mathbf{g}(Z_k)\mathbf{1}_M - \mathbf{g}(\bar{\mathbf{z}}_k) \right\|^2 + 18\eta^2\left\|\mathbf{g}(\bar{\mathbf{z}}_k) - \mathbf{g}(\bar{\mathbf{z}}_{k-1})\right\|^2 + 18\eta^2\left\| \frac{1}{M}\mathbf{g}(Z_{k-1})\mathbf{1}_M - \mathbf{g}(\bar{\mathbf{z}}_{k-1}) \right\|^2 \\
&\qquad + 6\eta^2\|\bar{\epsilon}_k\|^2 + 6\eta^2\|\bar{\epsilon}_{k-1}\|^2 \\
&\overset{(e)}{\leq} 18\eta^2\left\| \frac{1}{M}\mathbf{g}(Z_{k-1})\mathbf{1}_M - \mathbf{g}(\bar{\mathbf{z}}_{k-1}) \right\|^2 + 18\eta^2\left\| \frac{1}{M}\mathbf{g}(Z_k)\mathbf{1}_M - \mathbf{g}(\bar{\mathbf{z}}_k) \right\|^2 + 18\eta^2L^2\left\|\bar{\mathbf{z}}_{k-1} - \bar{\mathbf{z}}_k\right\|^2 \\
&\qquad + 6\eta^2\|\bar{\epsilon}_{k-1}\|^2 + 6\eta^2\|\bar{\epsilon}_k\|^2
\end{aligned} \tag{15}
$$

where (a) holds by the update of the algorithm and the fact that $W^t\mathbf{1}_M = \mathbf{1}_M$, (b) holds by utilizing Cauchy-Schwarz inequality and, (c) and (d) hold since $\|\mathbf{a}+\mathbf{b}+\mathbf{c}\|^2 \leq 3\|\mathbf{a}\|^2 + 3\|\mathbf{b}\|^2 + 3\|\mathbf{c}\|^2$, (e) holds by the L-Lipschitz continuity of $\mathbf{g}$.

Note that

$$\begin{aligned}
&\eta^2\left\|\mathbf{g}\left(\frac{1}{M}Z_k\mathbf{1}_M\right)\right\|^2=\|\bar{\mathbf{z}}_k-(\bar{\mathbf{z}}_k-\eta\mathbf{g}(\bar{\mathbf{z}}_k))\|^2=\|\bar{\mathbf{z}}_k-\bar{\mathbf{x}}_k+(\bar{\mathbf{x}}_k-(\bar{\mathbf{z}}_k-\eta\mathbf{g}(\bar{\mathbf{z}}_k)))\|^2\\
&\leq 2\left\|\bar{\mathbf{z}}_k-\bar{\mathbf{x}}_k\right\|^2+2\left\|(\bar{\mathbf{x}}_k-(\bar{\mathbf{z}}_k-\eta\mathbf{g}(\bar{\mathbf{z}}_k)))\right\|^2\\
&=2\left\|\bar{\mathbf{z}}_k-\bar{\mathbf{x}}_k\right\|^2+2\left\|\left(\frac{1}{M}\left(X_{k-1}W-\eta\widehat{\mathbf{g}}(\epsilon_k,Z_k)\right)\mathbf{1}_M-(\bar{\mathbf{z}}_k-\eta\mathbf{g}(\bar{\mathbf{z}}_k))\right)\right\|^2\\
&\leq 2\left\|\bar{\mathbf{z}}_k-\bar{\mathbf{x}}_k\right\|^2+4\left\|\frac{1}{M}\left(X_{k-1}W-Z_k\right)\mathbf{1}_M\right\|^2+4\eta^2\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\widehat{\mathbf{g}}(\epsilon_k,Z_k)-\mathbf{g}(\bar{\mathbf{z}}_k)\right)\right\|^2\\
&\leq 2\left\|\bar{\mathbf{z}}_k-\bar{\mathbf{x}}_k\right\|^2+4\left\|\bar{\mathbf{x}}_{k-1}-\mathbf{z}_k\right\|^2+4\eta^2\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{g}(\mathbf{z}_k^i)-\mathbf{g}(\bar{\mathbf{z}}_k)\right)\right\|^2+\frac{4\eta^2}{M^2}\sum_{i=1}^{M}\|\epsilon_k^i\|^2
\end{aligned}\tag{16}$$

Taking $\mathbf{x}=\mathbf{x}_*$ in (13), combining (14), (15) and noting that $W\mathbf{1}_M=\mathbf{1}_M$ yield

$$\begin{aligned}
&\|\bar{\mathbf{x}}_k-\mathbf{x}_*\|^2\leq\|\bar{\mathbf{x}}_{k-1}-\mathbf{x}_*\|^2-\|\bar{\mathbf{x}}_{k-1}-\bar{\mathbf{z}}_k\|^2-\|\bar{\mathbf{x}}_k-\bar{\mathbf{z}}_k\|^2+2\eta D\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{g}(\mathbf{z}_k^i)-\mathbf{g}(\bar{\mathbf{z}}_k)\right\|\\
&+2\left\langle\mathbf{x}_*-\bar{\mathbf{z}}_k,\frac{L}{M}\eta\sum_{i=1}^{M}\epsilon_k^i\right\rangle+18\eta^2\left\|\frac{1}{M}\mathbf{g}(Z_{k-1})\mathbf{1}_M-\mathbf{g}(\bar{\mathbf{z}}_{k-1})\right\|^2+18\eta^2\left\|\frac{1}{M}\mathbf{g}(Z_k)\mathbf{1}_M-\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2\\
&+18\eta^2L^2\left\|\bar{\mathbf{z}}_{k-1}-\bar{\mathbf{z}}_k\right\|^2+6\eta^2\|\bar{\epsilon}_k\|^2+6\eta^2\|\bar{\epsilon}_{k-1}\|^2
\end{aligned}\tag{17}$$

Noting that

$$\|\bar{\mathbf{z}}_{k-1}-\bar{\mathbf{z}}_k\|^2=\|\bar{\mathbf{z}}_{k-1}-\bar{\mathbf{x}}_{k-1}+\bar{\mathbf{x}}_{k-1}-\bar{\mathbf{z}}_k\|^2\leq 2\left\|\bar{\mathbf{z}}_{k-1}-\bar{\mathbf{x}}_{k-1}\right\|^2+2\left\|\bar{\mathbf{x}}_{k-1}-\bar{\mathbf{z}}_k\right\|^2,\tag{18}$$

Define $\Lambda_k=2\left\langle\mathbf{x}_*-\bar{\mathbf{z}}_k,\frac{L}{M}\eta\sum_{i=1}^{M}\epsilon_k^i\right\rangle$. We rearrange terms in (17) and combine (18) , which yield

$$\begin{aligned}
&\|\bar{\mathbf{x}}_{k-1}-\bar{\mathbf{z}}_k\|^2+\|\bar{\mathbf{x}}_k-\bar{\mathbf{z}}_k\|^2-18\eta^2L^2\left(2\left\|\bar{\mathbf{z}}_{k-1}-\bar{\mathbf{x}}_{k-1}\right\|^2+2\left\|\bar{\mathbf{x}}_{k-1}-\bar{\mathbf{z}}_k\right\|^2\right)\\
&\leq\|\bar{\mathbf{x}}_{k-1}-\mathbf{x}_*\|^2-\|\bar{\mathbf{x}}_k-\mathbf{x}_*\|^2+2\eta D\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{g}(\mathbf{z}_k^i)-\mathbf{g}(\bar{\mathbf{z}}_k)\right\|+\Lambda_k+6\eta^2\|\bar{\epsilon}_{k-1}\|^2+6\eta^2\|\bar{\epsilon}_k\|^2\\
&+18\eta^2\left\|\frac{1}{M}\mathbf{g}(Z_{k-1})\mathbf{1}_M-\mathbf{g}(\bar{\mathbf{z}}_{k-1})\right\|^2+18\eta^2\left\|\frac{1}{M}\mathbf{g}(Z_k)\mathbf{1}_M-\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2
\end{aligned}\tag{19}$$

Define $\mathbf{x}_{-1}^i=\mathbf{z}_{-1}^i=0$ for $\forall i\in\{1,\ldots,M\}$ and $\widehat{\mathbf{g}}(\epsilon_{-1},Z_{-1})=\mathbf{g}(Z_{-1})=0_{d\times M}$. Take summation over $k=0,\ldots,N-1$ in (19) and note that $\mathbf{z}_0^i=\mathbf{x}_0^i=0$ for $\forall i\in\{1,\ldots,M\}$, which yield

$$\begin{aligned}
&(1-36\eta^2L^2)\sum_{k=0}^{N-1}\|\bar{\mathbf{x}}_{k-1}-\bar{\mathbf{z}}_k\|^2+(1-36\eta^2L^2)\sum_{k=0}^{N-1}\|\bar{\mathbf{x}}_k-\bar{\mathbf{z}}_k\|^2\\
&\leq\|\bar{\mathbf{x}}_0-\mathbf{x}_*\|^2-\|\mathbf{x}_{N-1}-\mathbf{x}_*\|^2+12\eta^2\sum_{k=0}^{N-1}\|\bar{\epsilon}_k\|^2+\sum_{k=0}^{N-1}\Lambda_k\\
&+\sum_{k=0}^{N-1}2\eta D\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{g}(\mathbf{z}_k^i)-\mathbf{g}(\bar{\mathbf{z}}_k)\right\|+\sum_{k=0}^{N-1}36\eta^2\left\|\frac{1}{M}\mathbf{g}(Z_k)\mathbf{1}_M-\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2
\end{aligned}\tag{20}$$

By taking $\eta \leq \frac{1}{6\sqrt{2}L}$, we have $1 - 36\eta^2 L^2 \geq \frac{1}{2}$. Take expectation and divide N on both sides of (20), and then employing Lemma 2 and Lemma 4, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{2N}\left(\sum_{k=0}^{N-1}\|\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\|^2 + \sum_{k=0}^{N-1}\|\bar{\mathbf{x}}_{k-1} - \bar{\mathbf{z}}_k\|^2\right)\right] \\
&\leq \frac{\|\bar{\mathbf{x}}_0 - \mathbf{x}_*\|^2}{N} + 12\eta^2 \cdot \frac{\sigma^2}{mM} + \frac{16\eta^2 DL\sigma}{\sqrt{mM}(1-4\eta L)} + \frac{36\eta^2\sigma^2}{\sqrt{mM}} + \frac{36\eta^2}{c-1} \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2
\end{aligned} \tag{21}$$

By employing (16) and (21) and Lemma 2, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{k=0}^{N-1}\eta^2\mathbb{E}\left\|\mathbf{g}\left(\frac{1}{M}Z_k\mathbf{1}_M\right)\right\|^2 \leq \frac{4}{N}\mathbb{E}\left(\sum_{k=0}^{N-1}\|\bar{\mathbf{z}}_k - \bar{\mathbf{x}}_k\|^2 + \sum_{k=0}^{N-1}\|\bar{\mathbf{x}}_{k-1} - \bar{\mathbf{z}}_k\|^2\right) \\
&\quad + \frac{4\eta^2}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{g}(\mathbf{z}_k^i) - \mathbf{g}(\bar{\mathbf{z}}_k)\right)\right\|^2 + \frac{4\eta^2}{NM^2}\sum_{k=0}^{N-1}\sum_{i=1}^{M}\mathbb{E}\|\epsilon_k^i\|^2 \\
&\overset{(a)}{\leq} 8\left(\frac{\|\bar{\mathbf{x}}_0 - \mathbf{x}_*\|^2}{N} + 12\eta^2 \cdot \frac{\sigma^2}{mM} + \frac{16\eta^2 DL\sigma}{\sqrt{mM}(1-4\eta L)} + \frac{36\eta^2\sigma^2}{\sqrt{mM}} + \frac{36\eta^2}{c-1} \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2\right) \\
&\quad + \frac{4\eta^2\sigma^2}{mM} + \frac{4\eta^2}{c-1} \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 + \frac{4\eta^2\sigma^2}{mM}
\end{aligned} \tag{22}$$

where (a) holds by (21) and Lemma 2. Divide η^2 on both sides and by basic algebras, we have

$$\begin{aligned}
&\left(1 - \frac{320}{c-1}\right)\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 \leq 8\left(\frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2}{\eta^2 N} + \frac{20\sigma^2}{mM} + \frac{36\sigma^2}{\sqrt{mM}} + \frac{16DL\sigma}{\sqrt{mM}(1-4\eta L)}\right) \\
&\overset{(a)}{\leq} 8\left(\frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2}{\eta^2 N} + \frac{20\sigma^2}{mM} + \frac{48(DL\sigma + \sigma^2)}{\sqrt{mM}}\right)
\end{aligned} \tag{23}$$

where (a) holds since $1 - 4\eta L \geq \frac{1}{3}$ because of $\eta \leq \frac{1}{6\sqrt{2}L}$.

Taking $c = 321$, we know that

$$\frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left\|\mathbf{g}(\bar{\mathbf{z}}_k)\right\|^2 \leq 8\left(\frac{\|\mathbf{x}_0 - \mathbf{x}_*\|^2}{\eta^2 N} + \frac{20\sigma^2}{mM} + \frac{48(DL\sigma + \sigma^2)}{\sqrt{mM}}\right) \tag{24}$$

□