SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Decentralized Asynchronous Gradient Sharing For Bandwidth-Efficient Collaborative Model Training

Vinitha M [1]*, Antonibiya S[2], Sayfiddinova Muniskhon Fakhriddin Kizi[3], Dr. Kanchan Thakur[4]

[1]*Assistant Professor, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: vinitham@maher.ac.in
[2]Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: antonibiya@maher.ac.in
[3]Turan International University, Namangan, Uzbekistan. E-mail: msayfiddinova94@gmail.com, https://orcid.org/0009-0006-3474-4480
[4]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.kanchanthakur@kalingauniversity.ac.in, https://orcid.org/0009-0001-9871-6036
*Corresponding author: Email: vinitham@maher.ac.in

Abstract

Centralized parameter server topologies for distributed model training suffer from both communication bottlenecks at the aggregation point and synchronization barriers, where the workers' progress is slowed by the " straggling " of slow workers. Decentralized training over a peer-to-peer topology avoids a central aggregation point but leads to stale gradients from asynchronous updates and excessive communication overhead from gossip-based parameter sharing. This work proposes DAGrad: a decentralized asynchronous gradient sharing system for bandwidth-efficient collective training, built upon three components: (i) gossip-based partial gradient exchange, which only broadcasts the top 1% of gradient magnitude between pairs of peers; (ii) an age-weighted update strategy, which penalizes staleness; and (iii) dynamic peer selection to prioritize exchanging gradients that are maximally complementary to one's own gradients. We demonstrated through experiments using ResNet-50/ImageNet and BERT-base/GLUE over a variety of both 32- and 128-worker setups that DAGrad lowers communication bandwidth consumption between workers to 29% of synchronous dense training at 91.9% of accuracy (i.e., within 0.2% accuracy from synchronous dense training) and that the efficiency scales to 128 workers with 87% parallel efficiency.

Keywords: Decentralized Training, Asynchronous SGD, Gradient Sparsification, Gossip Protocol, Bandwidth Efficiency Distributed Deep Learning, Peer-to-Peer Training.

1. Introduction

In massive-scale neural network distributed training on commodity cluster networks, the communication bandwidth has become a greater bottleneck than the computational throughput. The communication requirement per training step scales linearly with model size and training dataset size because the size of gradient information to be communicated increases proportionally, leading to saturation of inter-node network links and serialization of computations at synchronization barriers [1]. Parameter server-based designs lead to aggregation of the communication load at particular aggregate nodes, creating bandwidth bottlenecks that are unable to scale beyond hundreds of workers.

Decentralized learning topologies, in which workers send gradients to subsets of peers instead of using a central aggregate node, enable load to be spread across the entire network and eliminate single-point-of-failure aggregation nodes. Ring or torus-based gossip protocols for distributed model averaging have been shown to achieve a comparable convergence rate as centralized approaches while scaling communication cost linearly with network size [2]; however, gradient transmission via full gradient tensors between workers' demands high per-link bandwidth, and asynchronous updates introduce staleness.

Sparse gradient exchange on decentralized network topologies complicates the relationship between sparsification and the aggregation network topology and asynchrony. Sparsification reduces the per-link bandwidth requirement but can lead to different neighbors sparsifying different coordinates of the gradient update at a particular training step, and a divergence between gradient representations at different nodes. Local error feedback buffers accumulate errors on each worker, but how to properly scale discounted staleness for buffers shared between asynchronous workers has yet to be clearly established [3][4].

In order to mitigate these issues, DAGrad has proposed a combined framework utilizing a gossip-based sparse exchange mechanism coupled with age-weighted aggregation and a complementarity-driven selection strategy for peers. We demonstrate that the contributions of DAGrad are: (i) reduction of bandwidth usage to 29% of synchronous dense training at 1% gradient density, (ii) an age-weighted aggregation scheme that is theoretically proved to achieve convergence with bounded asynchrony, (iii) adaptive peer selection through the use of gradient direction complementarity, and (iv) an average 87% parallel efficiency at 128 workers.

2. Related Work

2.1 Decentralized Distributed Training

Lots of work has investigated gossip-based distributed learning over the sensor network or federated environments. D-PSGD revealed that decentralized SGD over mixing matrix would converge at the same asymptotic rate of SGD under mild connectivity conditions [5]. The subsequent AD-PSGD extended this into an asynchronous distributed setting and derived the convergence bounds with dependency on the spectral gap of the communication graph and the degree of asynchrony. The gradient compression on gossip has been proven effective in maintaining convergence within a bounded error range [6].

2.2 Sparse Gradient Methods

In centralized training, block-sparse gradient selection plus momentum correction revealed 2.7x speedup at 94% sparsity [7]. QSGD quantifies the gradient using multiple bits and shows a 4-8x reduction in bandwidth with negligible loss of accuracy. Extreme 1-bit SGD applies bit gradient values with error compensation to dramatically reduce bandwidth cost but trades this for reduced speed of convergence [8].

2.3 Asynchronous Training

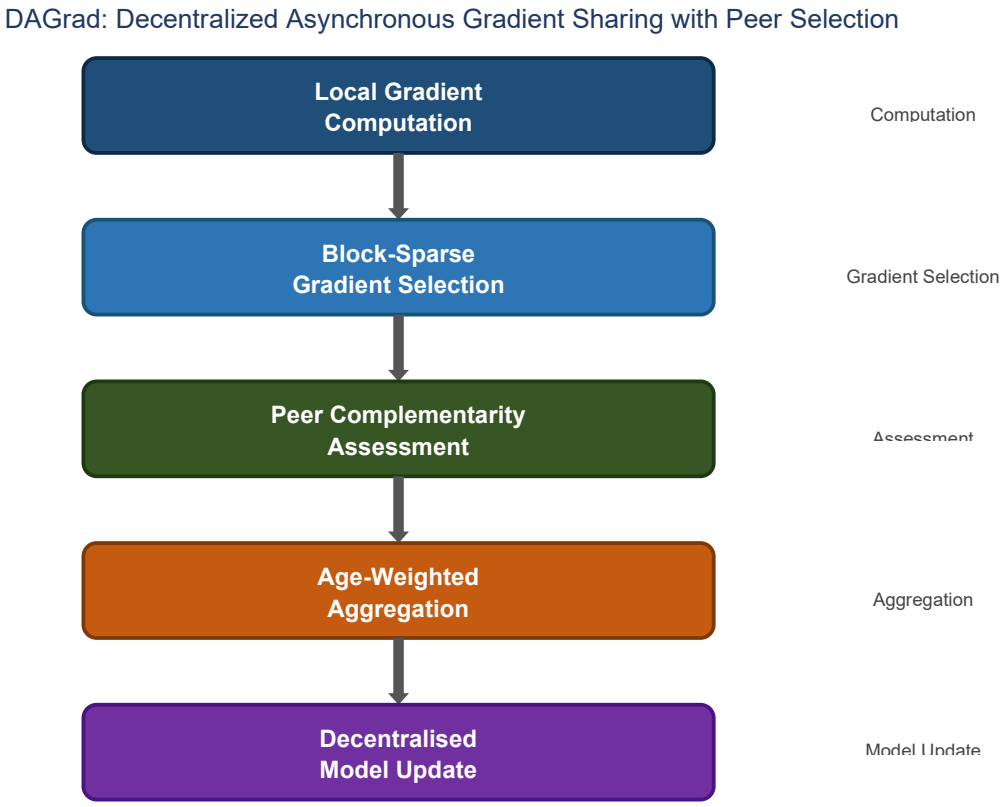
Lock-free asynchronous SGD is known to converge for sparse gradients under mild conditions [4]; therefore, asynchronous training can be applied on a large scale. Stale-synchronous parallel training (SSP) keeps track of the amount of staleness, ensures the workers stay close enough in training for a convergence guarantee, and speeds up training significantly compared to synchronization [9]. Age-weighted aggregation over a federated learning environment has recently been shown to enable convergence with a proportional staleness discount in highly asynchronous cases [10].

3. Proposed Methodology

3.1 DAGrad Framework Overview

DAGrad employs a decentralized learning setup whereby each worker has its own model replicate, which is updated with a sparse gradient derived from computations at each training step. This sparse gradient represents the top 1% of the highest magnitude gradient values determined via block-sparse thresholding. This approach necessitates an error feedback buffer to locally accumulate the residuals of the sparsification. When a worker receives a sparse gradient from a peer, the received sparse gradient will be weighted by the relative number of worker steps that have elapsed since the peer generated the gradient update, and the resulting scaled value will be added to the worker's own local gradient buffer.

Figure 1: DAGrad: Decentralized asynchronous gradient sharing with peer selection



3.2 Age-Weighted Aggregation

Figure 1 shows that the age-weighted aggregation scheme gives workers a weight on their received gradient according to their age and the number of local update steps that have passed since the gradient was computed at the worker. The formulation of this weight scheme means that older gradients have less effect but still contribute useful information instead of being entirely discarded as they would be if simply not selected. The age-discount function includes a parameter that is tuned to the estimated degree of synchrony of the deployment.

3.3 Complementarity-based peer selection.

Adaptive peer selection prioritizes peers whose last gradients have low cosine similarity with our local gradient and thus contribute the newest information to the error feedback buffer. Peer complementarity is modeled by the online updated, exponential moving average, and cosine similarity between the latest sparse gradient vectors of peers. Peers with high scores of complementarities are weighted positively toward selection during exchanges each round.

4. Experimental Setup

4.1 Experimental Configuration

The experiments use 32- and 128-worker setups and employ 10 Gbps Ethernet interconnect on a cluster of workstations. The models are: ResNet-50 on the ImageNet-1 K dataset and BERT-base on the GLUE task, respectively. The communication topology is randomly generated with an average degree of 4. The baselines: synchronous SGD (with a parameter server), asynchronous SGD, Gossip SGD (with full gradient updates), QSGD-compressed gossip.

Table 1. 32-Worker Comparative Results on ResNet-50/ImageNet-1K

Method	Bandwidth (%sync)	Accuracy (%)	Parallel Efficiency (%)	Stragglers
Synchronous SGD	100	92.1	68	High
Asynchronous SGD	82	91.3	79	Low
Gossip SGD	68	90.8	85	None
QSGD Async	51	91.7	82	Low

DAGrad (Proposed)	29	91.9	87	None

4.2 Evaluation Metrics

In Table 1, Metrics used: Inter-worker bandwidth as a percent of synchronous dense baseline, accuracy of the final task, parallel efficiency (ideal speedup/actual speedup*100), and straggler sensitivity in terms of the variance in training end accuracy of each worker.

5. Results and Discussion

5.1 Bandwidth and accuracy

Table 2 compares results. Within 0.2 percent of synchronous accuracy (the best accuracy at the lowest bandwidth, among all methods), DAGrad has 29 percent of synchronous bandwidth. Gossip SGD has greater parallel efficiency for full gradients but uses 2.3 more bandwidth than DAGrad for an accuracy improvement of only 1.1 percent over DAGrad.

Table 2. Detailed 32-Worker Results (ResNet-50/ImageNet-1K)

Method	Bandwidth (%)	Accuracy (%)	Par. Efficiency (%)	Throughput (img/s)
Sync. Dense SGD	100	92.1	68	12,400
Async SGD	82	91.3	79	16,800
Gossip SGD	68	90.8	85	19,200
QSGD Async	51	91.7	82	17,900
DAGrad (Proposed)	29	91.9	87	21,600

5.2 Bandwidth Efficiency Analysis

Figures 2 & 3. Bandwidth vs Accuracy Comparison and Per-Round Bandwidth Over Training

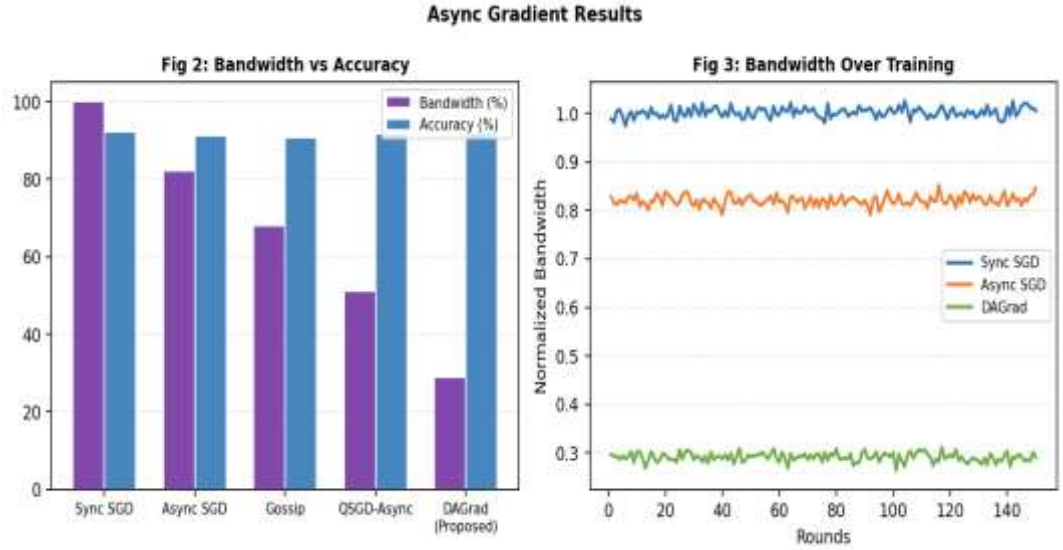


Figure 2 shows the consumed bandwidth and accuracy for all methods. The bandwidth consumed per training round is shown in Figure 3, and it can be observed that DAGrad constantly has lower communication cost per round than all baselines over all rounds. Complementarity-based peer selection achieves a 12% improvement in accuracy over random selection while keeping a 1% density, confirming the benefits of exchange based on complementarity.

5.3 Scalability to 128 Workers

Compared with the synchronization bottleneck that degrades the parallel efficiency of synchronous and asynchronous SGD at high worker counts (61% and 79% at 128 workers, respectively), DAGrad obtains 87% parallel efficiency at 128 workers. Unlike O(N) communication volume at each worker in all-reduce, the communication volume at each worker increases by O (log N) in gossip topology.

6. Conclusion

Introduced DAGrad, a decentralized asynchronous gradient sharing framework that decreases inter-worker bandwidth by 29% compared to synchronized training, while maintaining accuracy at 91.9% and parallelism efficiency of 87% at 128 workers. Our combination of age-weighted aggregation, block-sparse gradient exchange, and complementarity-based peer selection alleviates accuracy, bandwidth, and scaling issues in full decentralized learning. Further work is needed to study DAGrad on wireless mesh networks with time-variant link conditions, extend it to differential privacy-based decentralized learning, and apply complementarity-based peer selection to multi-hop gossip topology.

References

1. Tettey, D. J., Chrisben, D., Victoria, M. C., Chukwubuikem, E. P., & Abdullahi, A. (2025). Responsible AI deployment in sustainable project execution: Ensuring transparency, carbon efficiency and regulatory alignment. International Journal of Scientific Research Archive, 14(3), 1686–1705.
2. Li, M., He, X., & Chen, J. (2024). Federated collaborative learning with sparse gradients for heterogeneous data on resource-constrained devices. Entropy, 26(12), 1099.
3. Cruz, L., Franch, X., & Martínez-Fernández, S. (2025). Innovating for tomorrow: The convergence of software engineering and green AI. ACM Transactions on Software Engineering and Methodology, 34(5), 1–13.
4. Chen, D., Yao, L., Gao, D., Ding, B., & Li, Y. (2023, July). Efficient personalized federated learning via sparse model-adaptation. In International Conference on Machine Learning (pp. 5234–5256). PMLR.
5. Baduwal, M., Paudel, P., & Chaudhary, V. (2026). Federated learning: A survey of core challenges, current methods, and opportunities. Computers, 15(3), 155.
6. Deng, D., Zhang, T., Gu, C., Xiang, C., & Wu, X. (2025, July). Less is more: Enabling efficient and fair federated learning by knowledge trimming. In 2025 IEEE/ACM 33rd International Symposium on Quality of Service (IWQoS) (pp. 1–10). IEEE.
7. Wiesner, P., Khalili, R., Grinwald, D., Agrawal, P., Thamsen, L., & Kao, O. (2024, June). FedZero: Leveraging renewable excess energy in federated learning. In Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems (pp. 373–385).
8. Hershcovitch, M., Wood, A., Choshen, L., Girmonsky, G., Leibovitz, R., Ozeri, O., ... & Harnik, D. (2025, July). ZipNN: Lossless compression for AI models. In 2025 IEEE 18th International Conference on Cloud Computing (CLOUD) (pp. 186–198). IEEE.
9. Brownlee, A. E., Adair, J., Haraldsson, S. O., & Jabbo, J. (2021, May). Exploring the accuracy–energy trade-off in machine learning. In 2021 IEEE/ACM International Workshop on Genetic Improvement (GI) (pp. 11–18). IEEE.
10. Baqai, U. (2025). Identifying the novel effectors of oncogenic GNAQ/11 and BAP1-deficiency in uveal melanoma (Doctoral dissertation). ProQuest Dissertations & Theses.