A Survey of Compressed Communication and Related Techniques in Distributed Machine Learning

www.surveyx.cn

Abstract

In distributed machine learning and signal processing, compressed communication techniques, including quantization compression, distributed adaptive filtering, consensus algorithms, gradient sparsification, and federated learning, are vital for enhancing efficiency, scalability, and privacy. This survey explores their integration, emphasizing their role in mitigating communication bottlenecks and optimizing resource usage. Techniques like Deep Gradient Compression (DGC) and error feedback mechanisms reduce bandwidth requirements while maintaining accuracy, crucial for scalable distributed training. Federated learning frameworks such as FLARE and FedCPU demonstrate how compressed communication optimizes model performance under constraints. Innovations like ternary quantization and adaptive compression strategies further enhance efficiency, particularly in non-IID scenarios. Decentralized frameworks, including d-SVD and DAGC, improve convergence and robustness, addressing challenges in decentralized and federated learning environments. These techniques collectively advance the capabilities of distributed learning systems, ensuring robust performance despite communication and computational constraints. As the field evolves, ongoing research is essential to refine these methods and expand their applicability, paving the way for more efficient and resilient learning frameworks. This survey underscores the critical role of these techniques in overcoming challenges related to communication efficiency, resource constraints, and system heterogeneity, contributing to the development of scalable and effective distributed machine learning applications.

1 Introduction

1.1 Context and Motivation

The deployment of compressed communication techniques in distributed machine learning and signal processing is driven by the necessity to alleviate communication bottlenecks, enhance scalability, and ensure data privacy. In distributed systems, communication overhead significantly hinders scalability, necessitating innovative cost-reduction approaches. The high communication costs associated with training extensive deep neural networks using data-parallel distributed optimization underscore the urgent need for advanced compressed communication techniques. Research indicates that traditional gradient exchanges in distributed training can be redundant, with methods like Deep Gradient Compression (DGC) achieving compression ratios of 270x to 600x without compromising accuracy. Furthermore, novel adaptive gradient methods that integrate gradient compression have shown improved convergence rates and have been successfully implemented in scalable systems such as BytePS-Compress, enhancing training efficiency across various models, including ResNet50 and BERT-base. These advancements mitigate communication overhead and enable large-scale distributed training on lower-cost network infrastructures [1, 2].

In federated learning, the challenge of transmitting model updates across numerous devices without sharing raw data is intensified by bandwidth limitations and latency issues inherent in wireless

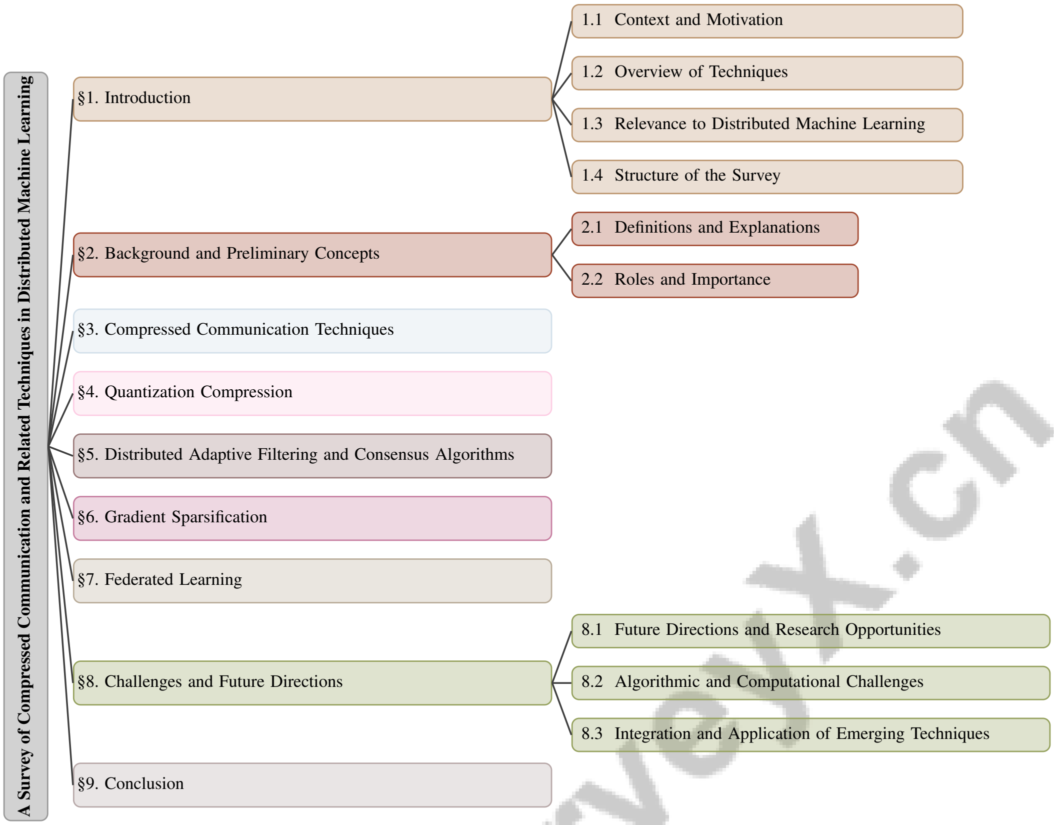

A Survey of Compressed Communication and Related Techniques in Distributed Machine Learning
§1. Introduction
1.1 Context and Motivation
1.2 Overview of Techniques
1.3 Relevance to Distributed Machine Learning
1.4 Structure of the Survey
§2. Background and Preliminary Concepts
2.1 Definitions and Explanations
2.2 Roles and Importance
§3. Compressed Communication Techniques
§4. Quantization Compression
§5. Distributed Adaptive Filtering and Consensus Algorithms
§6. Gradient Sparsification
§7. Federated Learning
§8. Challenges and Future Directions
8.1 Future Directions and Research Opportunities
8.2 Algorithmic and Computational Challenges
8.3 Integration and Application of Emerging Techniques
§9. Conclusion

Figure 1: chapter structure

networks [3]. This scenario is further complicated by heterogeneous data distribution among clients, potentially decreasing model performance. Compressed communication techniques are vital in optimizing resource usage while preserving data integrity and privacy [4]. The proliferation of IoT and the demand for intelligent applications highlight the impracticality of traditional centralized AI algorithms due to privacy concerns and inefficiencies in data transmission [5].

Gradient sparsification, which involves transmitting only significant updates, has emerged as a promising solution to enhance stochastic gradient descent efficiency, particularly when full gradient transmission is impractical due to high communication costs. This technique is effective in improving accuracy and efficiency in distributed mean estimation. Additionally, the limitations of conventional communication systems in addressing IoT applications emphasize the importance of transmitting only relevant information for decision-making [6].

In mobile environments, compressed communication techniques are critical for alleviating communication bottlenecks and improving convergence rates, especially in non-IID scenarios. The substantial communication and computation demands imposed by federated learning on battery-constrained devices within mobile edge networks necessitate innovative energy-efficient solutions. This is particularly important as these devices engage in continuous local training and periodic global synchronization, which can drain battery life significantly. Recent research has focused on optimizing energy efficiency through techniques such as flexible communication compression, balancing local computation demands with wireless communication. By addressing challenges posed by heterogeneous device capabilities and unreliable network connections, these advancements aim to enhance federated learning performance while minimizing energy consumption [7, 8, 9]. Collectively, these factors drive the development and implementation of compressed communication techniques, which are essential for enhancing the efficiency, scalability, and robustness of distributed machine learning systems, particularly in bandwidth-constrained environments with high privacy and security demands.

1.2 Overview of Techniques

In distributed machine learning and signal processing, several pivotal techniques have been developed to address communication efficiency and scalability challenges. Compressed communication, a cornerstone technique, aims to reduce data volume transmitted across networks, alleviating communication bottlenecks. Methods such as Deep Gradient Compression (DGC) compress gradients to minimize bandwidth usage while maintaining model accuracy [1]. Similarly, AdaQuantFL employs an adaptive quantization strategy that adjusts quantization levels during training, optimizing communication efficiency and maintaining a low error floor [10].

Quantization compression is crucial for reducing data representation precision, effectively lowering communication costs. Data-Aware Gradient Compression (DAGC) exemplifies this approach, tailoring compression ratios based on individual worker data volumes to enhance communication efficiency [11]. These methods are essential for deploying deep neural networks in resource-constrained environments, underscoring the need for compression [12].

Gradient sparsification techniques are vital for enhancing communication efficiency. For instance, the SAPS-PSGD algorithm transmits only significant gradient updates, reducing communication overhead while preserving convergence performance [13]. Additionally, Dual-Way Gradient Sparsification (DGS) optimizes communication by leveraging model difference tracking [14]. SIGN SGD compresses gradient information to just its sign, sufficient for effective optimization and minimizing communication overhead [15].

In federated learning, integrating compressed communication techniques is essential for efficient model updates across decentralized networks. The compressed L2GD algorithm, employing bidirectional communication compression, exemplifies advancements in federated learning frameworks to enhance communication efficiency [16]. Fast blockchain-based federated learning frameworks, such as BCFL, further demonstrate the intersection of federated learning and communication efficiency by leveraging compression techniques to reduce transmitted data [17].

Moreover, the development of unbiased single-scale and multiscale quantizers compatible with all-reduce operations reflects efforts to maintain performance while reducing communication overhead [18]. The exploration of digital and analog communication schemes in wireless federated learning frameworks underscores the ongoing quest to optimize communication under stringent resource constraints [19].

Recent research has also focused on goal-oriented (GO) and semantic communications, particularly in data compression for IoT applications [20]. Techniques such as Shadowheart SGD introduce novel unbiased compression methods that improve time complexities in centralized methods [21]. MoTEF integrates communication compression with momentum tracking and error feedback to address challenges in decentralized optimization [22]. Meanwhile, AdaGossip dynamically adjusts consensus step-size based on compressed model differences, enhancing communication efficiency in decentralized learning [23].

The techniques discussed collectively constitute a robust toolkit aimed at significantly improving the efficiency and scalability of distributed machine learning systems. This toolkit addresses critical challenges of communication overhead and resource limitations through innovative strategies such as gradient compression, dynamic server updates, adaptive optimization methods, and model pruning. For instance, the introduction of a novel adaptive gradient method with gradient compression enhances convergence rates for non-convex problems, while systems like BytePS-Compress optimize the communication process between workers and parameter servers, resulting in notable reductions in training time for complex models like ResNet50 and BERT without sacrificing accuracy. Additionally, frameworks like FedDUMAP leverage shared server data to boost training efficiency and accuracy, achieving remarkable performance improvements over traditional approaches. These advancements facilitate the handling of larger models and datasets in distributed environments, making them more practical and effective for real-world applications [2, 24, 25].

1.3 Relevance to Distributed Machine Learning

The integration of compressed communication techniques is pivotal in distributed machine learning, particularly in enhancing the efficiency and scalability of federated learning and other decentralized frameworks. Communication bottlenecks pose significant challenges to scalability, as existing meth-

AI-generated, for reference only.

ods often struggle to achieve linear speed-up with an increasing number of clients [22]. Techniques such as Sparse Binary Compression (SBC) have demonstrated their ability to drastically reduce communication costs while maintaining model performance, thereby enhancing the scalability of distributed training [26].

In federated learning, the need to transmit model updates across numerous devices without sharing raw data is compounded by bandwidth limitations and latency issues [3]. Methods like FedPAQ, which integrate periodic averaging, partial device participation, and quantized message-passing, effectively address these challenges, ensuring efficient communication and model convergence [3]. The significance of variance reduction and communication compression in enhancing Byzantine robustness is also crucial for effective distributed optimization [6].

Adaptive gradient compression strategies are essential for optimizing communication efficiency and enhancing model convergence in distributed settings characterized by diverse data distributions [4]. This adaptability is particularly beneficial in scenarios where data heterogeneity can adversely affect model performance [27]. Prox-LEAD, for instance, incorporates communication compression to efficiently solve decentralized stochastic composite optimization problems, including those with non-smooth components [27].

The significance of making distributed learning methods interpretable and comprehensible to human operators is underscored, particularly in scenarios requiring human oversight [5]. Techniques such as Deep Gradient Compression (DGC) and accelerated compressed gradient descent (ACGD) exemplify the fusion of compression with acceleration, enhancing convergence rates and communication efficiency [28]. Moreover, adaptive consensus step-size methods, like AdaGossip, allow for better handling of errors introduced by communication compression, further optimizing decentralized learning [23].

The strategies outlined in the referenced works collectively enhance the implementation of distributed machine learning systems in resource-constrained settings. They achieve this by ensuring robust performance while effectively tackling critical challenges such as communication efficiency, privacy concerns, and scalability. Notably, approaches like federated learning allow for collaborative model training using distributed data without compromising data security or privacy, adhering to legal regulations. Additionally, innovative methods such as interpretable data fusion improve human interpretability of complex machine learning processes, facilitating better human-machine interactions while maintaining competitive performance metrics [29, 5]. By optimizing resource utilization and preserving data integrity, compressed communication techniques are indispensable for advancing distributed machine learning, particularly in the context of federated learning and decentralized optimization.

1.4 Structure of the Survey

The survey is meticulously structured to provide an in-depth analysis of compressed communication techniques, specifically focusing on their application in distributed machine learning and signal processing, including innovative methods such as gradient compression and sparsification to mitigate communication overhead and enhance scalability in large-scale machine learning systems [2, 30]. We begin with an introduction that delineates the context and motivation behind these techniques, followed by an overview of the key methodologies and their relevance to distributed machine learning. This sets the stage for a detailed examination of background and preliminary concepts, where core definitions and their roles are elaborated.

The subsequent sections delve into specific techniques, starting with compressed communication methods, including quantization and gradient sparsification, and their impact on data transmission efficiency. We then explore event-triggered and bidirectional compression methods, as well as decentralized and peer-to-peer communication models. The focus then shifts to quantization compression, detailing its processes, benefits, and associated challenges.

Further, the survey investigates distributed adaptive filtering and consensus algorithms, assessing their contributions to decentralized model training and resource optimization. This is followed by an analysis of gradient sparsification techniques, emphasizing their role in reducing communication overhead and maintaining model performance. The section on federated learning highlights its advantages in privacy preservation and decentralized training, alongside the integration of compressed communication techniques within these frameworks.

The survey concludes with an in-depth discussion on the multifaceted challenges and prospective future directions in the implementation of federated learning techniques. It identifies key algorithmic and computational issues, such as the complexities of data communication, privacy concerns, and the need for efficient aggregation methods. Additionally, the survey explores opportunities for integrating emerging approaches, including interpretable data fusion methods that enhance human-machine interaction while maintaining data privacy and communication efficiency, thereby paving the way for advancements in collaborative machine learning across diverse and distributed environments [29, 31, 5, 32]. The conclusion synthesizes the key findings, underscoring the importance of these techniques in advancing distributed machine learning and signal processing.The following sections are organized as shown in Figure 1.

2 Background and Preliminary Concepts

2.1 Definitions and Explanations

Distributed machine learning involves key concepts that tackle communication efficiency and scalability issues. Federated Learning (FL) allows multiple devices to collaboratively train a shared model without exchanging raw data, thus ensuring privacy and enhancing data security [33]. This approach is crucial in privacy-constrained or regulated environments [34]. However, FL faces scalability challenges and high communication overhead due to frequent model updates between clients and a central server [34], compounded by the non-IID nature of client data, which can hinder model performance [35].

Decentralized Federated Learning (DFL) enhances FL by enabling devices to aggregate models without a central server, addressing issues from imperfect communication channels [36]. This is vital in environments with unreliable communication infrastructures, facilitating collaborative model training [37].

Compressed communication techniques reduce communication overhead while maintaining convergence rates, essential for efficiently transmitting large gradient vectors, a significant bottleneck in scaling deep learning training [26]. Techniques such as gradient sparsification and optimal encoding manage communication loads in resource-constrained settings [26].

Quantization reduces data representation precision to lower communication costs. Despite its non-differentiable nature, quantization is crucial in wireless FL frameworks, addressing uplink transmission issues in both digital and analog communication [38].

In decentralized learning, distributed adaptive filtering and consensus algorithms optimize large-scale convex optimization problems over networks, relying on sparse communication among agents with access to only local data. These algorithms facilitate efficient data processing and decision-making in decentralized environments [38].

Gradient sparsification reduces communication overhead by transmitting only significant gradient updates, particularly effective in optimizing non-convex loss functions within distributed frameworks [37]. Enhanced by error feedback mechanisms, this technique manages memory efficiently while maintaining performance [39]. AdaGossip, which adaptively adjusts consensus step sizes, exemplifies efforts to mitigate high communication overheads in decentralized learning [23].

These concepts establish foundational strategies for enhancing system efficiency, scalability, and robustness across distributed machine learning applications, addressing challenges such as communication efficiency, privacy, and model convergence. Decentralized stochastic composite optimization on a connected n-node network with heterogeneous local objectives, involving both smooth and non-smooth components, underscores these techniques' complexity and necessity [27].

2.2 Roles and Importance

Compressed communication, quantization compression, distributed adaptive filtering, consensus algorithms, gradient sparsification, and federated learning are pivotal in advancing distributed machine learning and signal processing. These concepts address challenges posed by data privacy, non-IID data distributions, and computational and communication constraints in wireless environments [40]. Federated learning and its decentralized variants enable collaborative model training while preserving data privacy and security, essential given legal restrictions on aggregating raw data from multiple sources [32].

Communication efficiency is crucial in distributed optimization, particularly for large neural network models, as communication bottlenecks can severely impact performance [25]. Techniques like gradient sparsification and quantization compression are vital in reducing communication overhead by transmitting only necessary information, optimizing resource usage [20]. This is especially important in decentralized systems characterized by client heterogeneity, communication resource limitations, and security vulnerabilities [41].

Partial node participation in federated learning can lead to inefficient communication and suboptimal convergence if not managed properly. Advanced methods employing adaptive strategies for communication and computation are essential to mitigate these issues [42]. Furthermore, distributed adaptive filtering and consensus algorithms optimize large-scale convex optimization problems over networks, facilitating efficient data processing and decision-making in decentralized environments [38].

While first-order methods are common, they often incur high communication costs and slow convergence rates dependent on the optimization problem's condition number [43]. Advancing second-order methods and other optimization techniques is vital for improving convergence rates and reducing communication load. Collectively, these techniques enhance privacy, reduce communication overhead, and enable efficient collaborative model training across distributed devices, thereby advancing distributed machine learning and signal processing [38].

3 Compressed Communication Techniques

Category	Feature	Method
Compressed Communication in Distributed Systems	Efficiency Enhancements Adaptive Strategies Data Representation	BVRM[6], FedPAQ[3] AG[23], EAFO[4] RLF[5]
Event-Triggered and Bidirectional Compression Methods	Selective Communication Dual-Direction Optimization	SCG[44], CC-DQM[45] 2D[46], EF21-P[47]
Decentralized and Peer-to-Peer Communication Models	Decentralized Efficiency	Ok-Topk[48], SAPS-PSGD[13], Prox-LEAD[27]

Table 1: This table provides a comprehensive overview of various compressed communication techniques employed in distributed machine learning systems. It categorizes these techniques into three primary areas: compressed communication in distributed systems, event-triggered and bidirectional compression methods, and decentralized and peer-to-peer communication models. Each category lists specific features and corresponding methods, highlighting their roles in enhancing communication efficiency and model performance.

In distributed machine learning, optimizing communication strategies is essential for enhancing performance across multiple collaborating nodes. This section explores various compressed communication techniques that alleviate communication bottlenecks, thus improving efficiency within distributed systems. As illustrated in Figure 2, the hierarchical structure of these compressed communication techniques can be categorized into three main areas: compressed communication in distributed systems, event-triggered and bidirectional compression methods, and decentralized and peer-to-peer communication models. Each of these categories encompasses specific methods and strategies, which play a pivotal role in enhancing communication efficiency and model performance. Understanding this structure forms a basis for comprehending the significance of these techniques within modern machine learning frameworks. Table 5 presents a detailed categorization of compressed communication techniques in distributed systems, illustrating key methods and strategies that address communication bottlenecks and improve system efficiency.

3.1 Compressed Communication in Distributed Systems

Compressed communication techniques are crucial in distributed systems for mitigating communication bottlenecks during model training. Key methods include gradient sparsification and quantization, which reduce the volume of exchanged data, thereby enhancing communication efficiency. Techniques like Sparse Binary Compression (SBC) transmit only significant updates, effectively reducing communication load while preserving model accuracy. For instance, MoTEF integrates momentum tracking to improve convergence rates in decentralized optimization [26, 22].

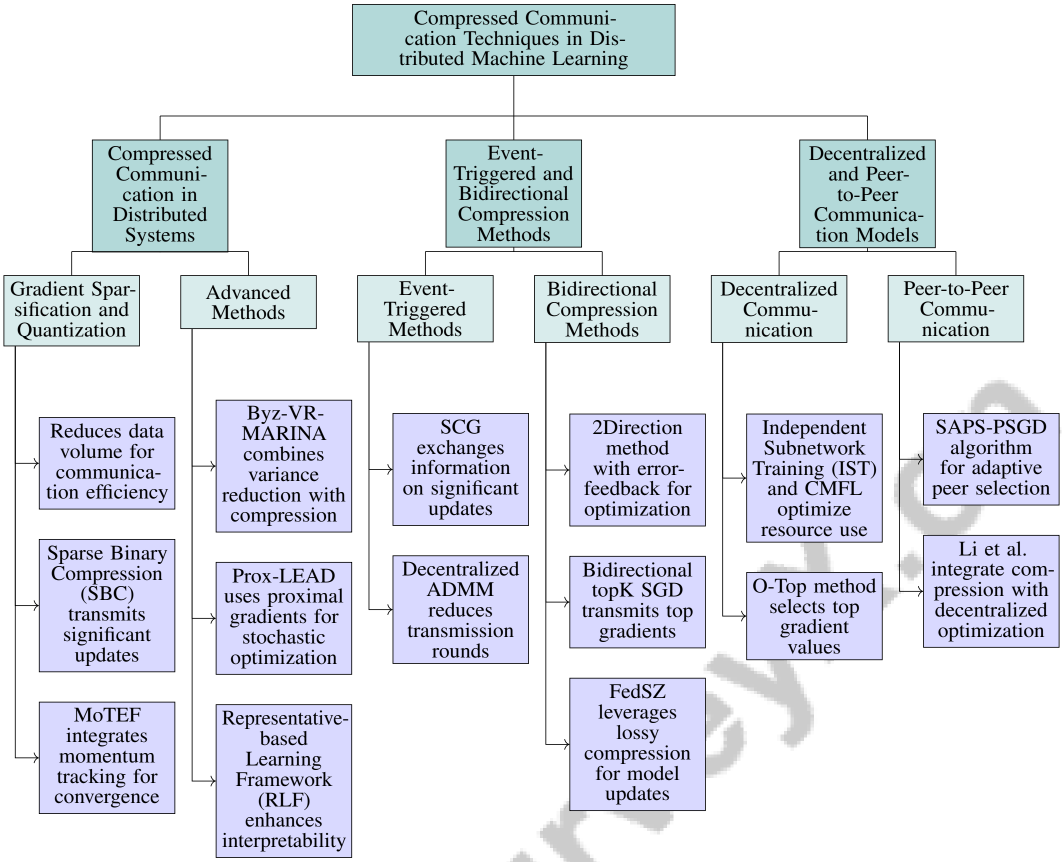

Compressed Communication Techniques in Distributed Machine Learning
Compressed Communication in Distributed Systems
Event-Triggered and Bidirectional Compression Methods
Decentralized and Peer-to-Peer Communication Models
Gradient Sparsification and Quantization
Advanced Methods
Event-Triggered Methods
Bidirectional Compression Methods
Decentralized Communication
Peer-to-Peer Communication
Reduces data volume for communication efficiency
Sparse Binary Compression (SBC) transmits significant updates
MoTEF integrates momentum tracking for convergence
Byz-VR-MARINA combines variance reduction with compression
Prox-LEAD uses proximal gradients for stochastic optimization
Representative-based Learning Framework (RLF) enhances interpretability
SCG exchanges information on significant updates
Decentralized ADMM reduces transmission rounds
2Direction method with error-feedback for optimization
Bidirectional topK SGD transmits top gradients
FedSZ leverages lossy compression for model updates
Independent Subnetwork Training (IST) and CMFL optimize resource use
O-Top method selects top gradient values
SAPS-PSGD algorithm for adaptive peer selection
Li et al. integrate compression with decentralized optimization

Figure 2: This figure illustrates the hierarchical structure of compressed communication techniques in distributed machine learning. It categorizes these techniques into three main areas: compressed communication in distributed systems, event-triggered and bidirectional compression methods, and decentralized and peer-to-peer communication models. Each category is further divided into specific methods and strategies, highlighting their roles in optimizing communication efficiency and model performance.

Method Name	**Communication Efficiency**	**Data Compression Techniques**	**Decentralized Optimization**
EAFO[4]	Parameter Compression	Parameter Compression Budgets	Adaptive Consensus Step-size
FedPAQ[3]	Periodic Averaging	Quantized Message-passing	Partial Participation
BVRM[6]	Communication Compression	Communication Compression	Decentralized Learning Environments
Prox-LEAD[27]	Communication Compression	Communication Compression	Decentralized Optimization
RLF[5]	Reduced Communication Overhead	Not Mentioned	Decentralized Devices
AG[23]	Adaptive Consensus Step-size	Compressed Model Differences	Dynamic Adjustment

Table 2: Overview of various methods for enhancing communication efficiency in distributed systems, detailing their specific approaches to data compression and decentralized optimization. This table highlights the strategies employed by each method, such as parameter compression and adaptive consensus step-size, to address communication bottlenecks in federated learning and decentralized environments.

Quantization further complements gradient sparsification by reducing data precision, thus lowering communication costs. Algorithms such as Efficient Adaptive Federated Optimization (EAFO) balance computation, communication, and precision by adjusting local updates and compression parameters [4]. The FedPAQ framework reduces communication costs by allowing multiple local updates before server synchronization, optimizing efficiency through update compression [3].

Advanced methods like Byz-VR-MARINA combine variance reduction with communication compression to maintain model integrity against Byzantine workers [6]. Prox-LEAD utilizes proximal gradients and communication compression for efficient stochastic composite optimization [27]. In federated learning, compressed communication is vital for efficient model updates, as seen in the

AI-generated, for reference only.

Representative-based Learning Framework (RLF), which enhances interpretability and data fusion [5]. AdaGossip employs gossip-error to compute an adaptive consensus step-size, further improving decentralized learning [23].

As illustrated in Figure 3, this figure depicts the hierarchical classification of compressed communication techniques in distributed systems, highlighting primary methods such as gradient sparsification, quantization, and advanced techniques. Each category encompasses specific algorithms that enhance communication efficiency and model performance. These techniques underscore the importance of compressed communication in distributed systems, facilitating efficient model training by optimizing data transmission and reducing communication costs. This is particularly relevant in federated learning, where privacy-preserving strategies are essential. Techniques like Deep Gradient Compression (DGC) achieve compression ratios of 270x to 600x while maintaining accuracy, addressing high bandwidth demands through methods such as momentum correction and local gradient clipping [28, 1, 49].

Table 2 provides a comprehensive comparison of different methods implemented to improve communication efficiency in distributed systems, focusing on their respective data compression techniques and decentralized optimization strategies.

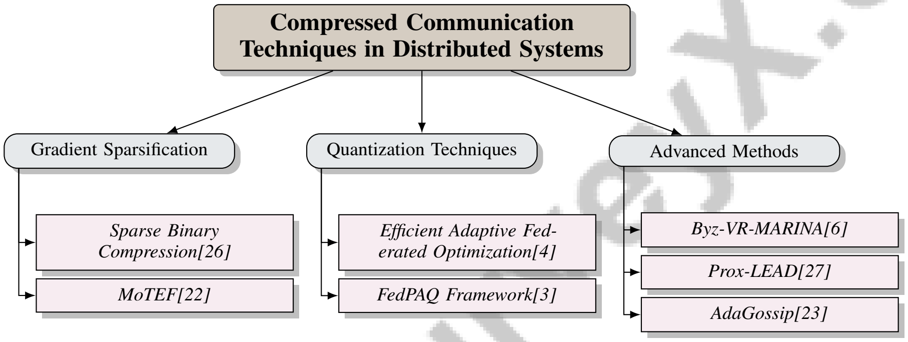


Figure 3: This figure illustrates the hierarchical classification of compressed communication techniques in distributed systems, highlighting primary methods such as gradient sparsification, quantization, and advanced techniques. Each category encompasses specific algorithms that enhance communication efficiency and model performance.

3.2 Event-Triggered and Bidirectional Compression Methods

Method Name	Communication Efficiency	Compression Techniques	Convergence and Performance
SCG[44]	Communication Efficiency	Compressed Messages	Linear Convergence Rates
CC-DQM[45]	Superior Communication Efficiency	Compressed Communication	Exact Linear Convergence
2D[46]	Bidirectional Compressed Communication	Bidirectional Communication Compression	Superior Communication Complexity
EF21-P[47]	Bidirectional Compression	Bidirectional Compression	State-of-the-art

Table 3: Comparison of various event-triggered and bidirectional compression methods in distributed machine learning, highlighting their communication efficiency, compression techniques, and convergence performance. The table provides insights into the effectiveness of each method in optimizing communication while maintaining or enhancing convergence rates and performance.

Event-triggered and bidirectional compression methods enhance communication efficiency in distributed machine learning by selectively transmitting information based on predefined criteria and facilitating two-way data exchange. Recent advancements include adaptive gradient methods achieving O(1/T) convergence for non-convex problems. BytePS-Compress enables two-way gradient compression, significantly reducing training times for models like ResNet50 and BERT without sacrificing accuracy. The 2Direction method introduces an error-feedback mechanism that accelerates distributed convex optimization [46, 2].

Event-triggered methods, such as SCG, involve agents exchanging compressed information only when significant updates occur, reducing unnecessary communication [44]. The decentralized ADMM framework exemplifies this by reducing transmission rounds and bits, enhancing scalability [45].

Bidirectional compression methods optimize data exchange between nodes and central servers. The 2Direction method enhances distributed optimization through bidirectional compressed communication, significantly reducing communication costs [46]. Integrating bidirectional compression with probabilistic communication protocols in personalized federated learning further improves efficiency [16].

Bidirectional topK SGD reduces communication volume while maintaining convergence rates by transmitting only top-ranked gradients [50]. Methods like DASHA and MoTEF use unbiased compressors to minimize communication costs in nonconvex optimization, achieving linear speed-up without strict conditions on data heterogeneity [47].

Event-triggered and bidirectional compression techniques form a comprehensive framework for optimizing communication efficiency in distributed machine learning. They significantly reduce information exchanged between clients and servers while maintaining model performance and convergence rates. Methods like FedSZ leverage lossy compression to minimize model update sizes without compromising accuracy, achieving compression ratios up to 12.61 times. The 2Direction method enhances efficiency through fast bidirectional communication and error-feedback, improving communication complexity benchmarks [46, 51, 52, 2]. These advancements alleviate bottlenecks and enable scalable, efficient distributed training across various models and datasets. Table 3 presents a comparative analysis of event-triggered and bidirectional compression methods, illustrating their impact on communication efficiency and performance in distributed machine learning environments.

3.3 Decentralized and Peer-to-Peer Communication Models

Method Name	Communication Efficiency	Scalability and Robustness	Bandwidth Management
Ok-Topk[48]	Gradient Compression	Direct Node Communication	Top-k Selection
SAPS-PSGD[13]	Adaptive Peer Selection	Direct Node Communication	Communication Compression
Prox-LEAD[27]	Communication Compression	Decentralized Optimization	Communication Compression

Table 4: Comparison of Decentralized Communication Methods in Distributed Machine Learning: The table outlines three methods—Ok-Topk, SAPS-PSGD, and Prox-LEAD—highlighting their communication efficiency, scalability, robustness, and bandwidth management strategies. Each method employs distinct techniques such as gradient compression, adaptive peer selection, and communication compression to enhance distributed learning performance.

Decentralized and peer-to-peer communication models are pivotal in distributed machine learning, especially when centralized coordination is impractical. These models enable direct node communication, reducing reliance on a central server, thus enhancing system robustness and scalability. Techniques like Independent Subnetwork Training (IST) and Communication-Mitigated Federated Learning (CMFL) address communication overhead and optimize resource utilization [53, 25].

Table 4 provides a comparative analysis of various decentralized communication methods, focusing on their efficiency, scalability, and bandwidth management in the context of distributed machine learning. The O-Top method exemplifies efficient decentralized communication by selecting top-gradient values, integrating them into an SGD optimizer, and managing bandwidth constraints [48]. This reduces communication overhead while maintaining model performance.

The SAPS-PSGD algorithm enhances communication efficiency by adaptively selecting peers based on bandwidth, reducing network traffic [13]. This is beneficial in heterogeneous networks, ensuring optimal resource use and scalability.

Incorporating communication compression into decentralized frameworks further improves performance. Li et al. combine decentralized optimization with communication compression, addressing both smooth and non-smooth optimization problems [27]. This approach balances communication cost and optimization accuracy, enabling efficient training in diverse environments.

Decentralized and peer-to-peer models emphasize efficient communication strategies in distributed machine learning. By employing techniques like top- selection, adaptive peer selection, and communication compression, these models tackle bandwidth limitations, scalability, and system robustness. Gradient compression enhances communication efficiency in distributed training, allowing substantial reductions in communication overhead while maintaining convergence performance. The BytePS-Compress system notably improves training speed for models like ResNet50 and BERT-base without sacrificing accuracy, while decentralized algorithms optimize bandwidth usage, alleviating bottlenecks in low-bandwidth environments [2, 13]. As distributed learning environments evolve, developing such models will be essential for advancing the field and enabling efficient, scalable machine learning applications.

Feature	Compressed Communication in Distributed Systems	Event-Triggered and Bidirectional Compression Methods	Decentralized and Peer-to-Peer Communication Models
Communication Strategy	Gradient Sparsification	Selective Transmission	Direct Node Communication
Optimization Focus	Data Transmission	Two-way Data Exchange	Bandwidth Management
Efficiency Impact	Reduced Communication Costs	Improved Convergence Rates	Enhanced Scalability

Table 5: This table provides a comparative analysis of communication strategies in distributed systems, highlighting the key features of compressed communication, event-triggered and bidirectional compression methods, and decentralized communication models. It emphasizes the optimization focus and efficiency impact of each approach, offering insights into how these methods address communication bottlenecks and enhance system performance.

4 Quantization Compression

Quantization compression is essential in distributed machine learning for enhancing communication efficiency and reducing data transmission. Vector quantization consolidates parameters into single codes, improving compression efficiency in fields like computer vision and natural language processing. Examining the interaction between these techniques, model architecture, and optimization strategies reveals their effectiveness in reducing the computational and memory demands of deep neural networks, particularly in resource-constrained environments [12, 54, 55, 56, 57]. This analysis highlights their role in federated learning, offering insights into their implementation and impacts.

4.1 Quantization and Vector Quantization Techniques

Quantization techniques enhance communication efficiency in distributed machine learning by reducing data precision, thereby lowering transmission volumes. This is critical in federated learning, where bandwidth is limited. The FedPAQ framework exemplifies this by using periodic averaging and quantization to minimize communication load while maintaining performance [3]. Such strategies are vital for balancing communication efficiency, model accuracy, and convergence.

In decentralized learning, Sparse Binary Compression (SBC) sparsifies weight updates by retaining significant changes, improving communication efficiency without compromising model performance [26]. This approach illustrates quantization trade-offs, as selective compression can yield better convergence rates than full precision models.

Advanced techniques like PowerSGD demonstrate the advantages of approximating gradients through lightweight processes, enhancing communication efficiency [58]. The HEROES framework optimizes model training by adaptively assigning tensor blocks and local update frequencies based on client capabilities, further enhancing federated learning efficiency [59].

Induced compressors support quantization compression by maintaining low memory requirements while ensuring effective communication [39]. Balancing computation and communication times is crucial, as shown by methods like Shadowheart SGD, which ensure optimization efficiency under varying conditions [21]. Collectively, these methods highlight the importance of quantization in efficient distributed learning frameworks while addressing interpretability and privacy concerns [5].

This figure illustrates key quantization and vector quantization techniques in distributed machine learning, focusing on communication efficiency, model optimization, and privacy and interpretability. As shown in Figure 4, each category highlights significant methods and frameworks that contribute to enhancing federated learning and decentralized learning environments. Graph (a) shows the Rényi Differential Privacy (RDP) bound for Gaussian and quantized Gaussian distributions with = 1, plotted against log2(k), highlighting quantization's impact on data privacy. Graph (b) features curves for different values (3, 7, and 15), showing the effect of varying quantization parameters. Graph (c) tracks the parameter across epochs, illustrating its dynamic behavior during learning. These examples underscore the diverse applications and complexities of quantization techniques in data processing [60, 54, 61].

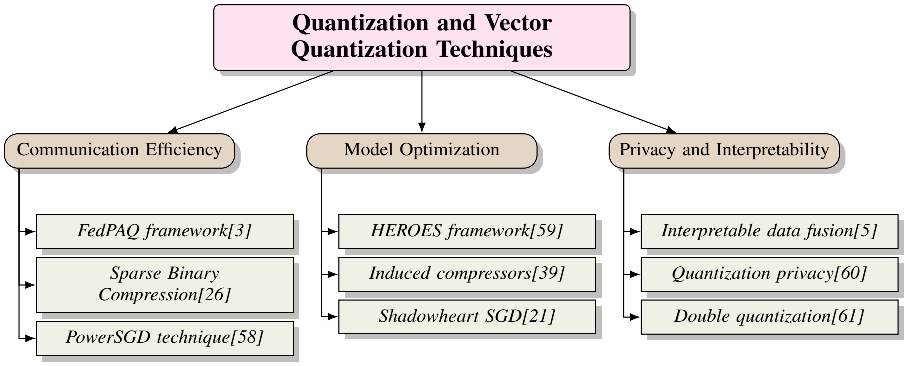


Figure 4: This figure illustrates key quantization and vector quantization techniques in distributed machine learning, focusing on communication efficiency, model optimization, and privacy and interpretability. Each category highlights significant methods and frameworks that contribute to enhancing federated learning and decentralized learning environments.

4.2 Challenges in Quantization

Implementing quantization compression in distributed machine learning systems presents challenges affecting convergence and performance. A key issue is divergence due to varying adaptive learning rates across nodes, which can hinder convergence and complicate synchronization [62]. This problem is pronounced in environments with extreme data heterogeneity and fluctuating network conditions, where uniform quantization may not enhance communication efficiency [63].

Another challenge is the computational demands of optimization, especially during fine-tuning. The computational intensity of methods like permute quantization may limit their feasibility in resource-constrained scenarios [55]. This is exacerbated by irregular network topologies, where communication efficiency can decline, impacting linearly convergent algorithms [64].

High compression ratios in quantization can slow convergence rates, requiring a balance between compression and convergence to maintain performance [65]. This balance is challenging in scenarios with heterogeneous device capabilities or extreme network conditions, affecting overall performance [66].

Moreover, treating pruning and quantization as separate processes can lead to suboptimal outcomes, often requiring retraining to recover accuracy post-compression, complicating implementation [56]. Assumptions in frameworks like CMFL that global updates can be estimated from previous iterations may not hold universally, leading to inaccuracies and inefficiencies [53].

Future research should focus on optimizing sparsification, exploring alternative measurement matrices, and developing strategies to enhance client coordination to address these challenges [67]. Such efforts are vital for advancing quantization compression techniques and ensuring their robust application across diverse distributed learning environments.

5 Distributed Adaptive Filtering and Consensus Algorithms

Consensus algorithms are integral to distributed adaptive filtering, optimizing resource utilization and model convergence in decentralized machine learning systems. These algorithms enable coordination among distributed agents without centralized control, forming the backbone of decentralized training frameworks and enhancing resource efficiency.

5.1 Consensus Algorithms and Resource Usage

In distributed machine learning, consensus algorithms optimize resource usage and ensure model convergence by facilitating agreement among distributed agents. The Byz-VR-MARINA framework exemplifies robust aggregation methods that filter Byzantine influences, enhancing efficiency and resilience in adversarial environments [6]. Federated learning faces challenges in balancing local updates and parameter compression, affecting efficiency and convergence [4]. Adaptive approaches are required to dynamically adjust to distributed systems' varying conditions, ensuring effective resource utilization while maintaining robust model performance.

As illustrated in Figure 5, the hierarchical organization of consensus algorithms and resource usage in distributed machine learning highlights key aspects such as robust aggregation methods, federated learning efficiency, and decentralized federated learning techniques. Gradient sparsification techniques effectively reduce communication overhead in large-scale distributed systems without compromising convergence speed. These techniques are crucial in communication-constrained environments, reducing data exchange while preserving convergence rates. Adaptive gradient methods incorporating gradient compression achieve convergence rates of O(1/T) for non-convex problems. Frameworks like Global Momentum Fusion (GMF) enhance communication efficiency in federated learning, maintaining model accuracy even with non-IID data distributions. Flexible communication compression strategies optimize energy consumption in federated learning over mobile edge networks, enabling devices to balance local computation and communication demands without sacrificing performance [68, 69, 2, 9]. ADOTA-FL improves robustness and efficiency in federated learning by mitigating channel fading and interference through adaptive stepsize adjustments, optimizing resource usage in wireless environments.

Decentralized Federated Learning (DFL) enhances convergence rates and model consensus by enabling direct client communication, eliminating the need for a central server. This significantly reduces communication overhead and improves efficiency in decentralized model training, addressing challenges such as privacy protection, distribution shifts, and communication bottlenecks while supporting various optimization techniques across heterogeneous data environments [7, 24, 41, 70, 71]. Soft-DSGD leverages all available communication links to facilitate efficient resource usage and model convergence, achieving performance levels comparable to reliable communication conditions.

Consensus algorithms are essential for the efficient operation of distributed machine learning systems, enabling multiple computational nodes to collaboratively train models by coordinating updates over unreliable networks. They address challenges such as communication bottlenecks and data privacy while maintaining convergence performance [72, 73, 5, 13, 32]. By tackling key issues like communication efficiency, client heterogeneity, and Byzantine robustness, these algorithms enhance the capabilities of decentralized training frameworks, paving the way for scalable and resilient distributed learning applications.

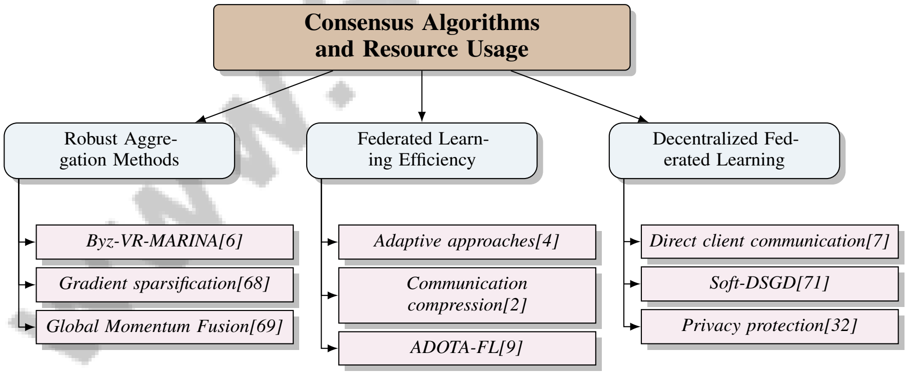

Consensus Algorithms and Resource Usage
Robust Aggregation Methods
Federated Learning Efficiency
Decentralized Federated Learning
Byz-VR-MARINA[6]
Gradient sparsification[68]
Global Momentum Fusion[69]
Adaptive approaches[4]
Communication compression[2]
ADOTA-FL[9]
Direct client communication[7]
Soft-DSGD[71]
Privacy protection[32]

Figure 5: This figure illustrates the hierarchical organization of consensus algorithms and resource usage in distributed machine learning, highlighting robust aggregation methods, federated learning efficiency, and decentralized federated learning techniques.

5.2 Decentralized Model Training Frameworks

Decentralized model training frameworks are crucial for addressing scalability, resource constraints, and data privacy in distributed machine learning. These frameworks enable collaborative learning across devices without relying on a central server, enhancing system robustness and efficiency. The FAT-Clipping framework exemplifies this by applying clipping to client updates in federated learning, effectively managing fat-tailed noise and optimizing resource usage while ensuring model convergence [69].

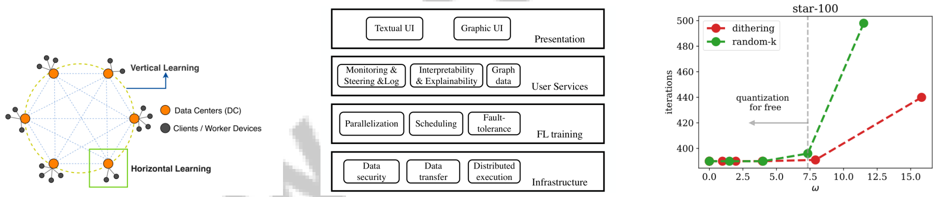
AI-generated, for reference only.

In scenarios with heterogeneous data distribution, frameworks like FL+HC utilize hierarchical clustering to create specialized models for groups of clients with similar data distributions. This approach improves overall model accuracy and reduces communication costs by tailoring updates to specific client groups [74]. By focusing on unique data characteristics, FL+HC enables more efficient model training in decentralized environments.

The Average Consensus Algorithm with Interference (ACA-I) illustrates the potential of decentralized frameworks by iteratively computing the average of initial sensor observations through message passing among nodes, accounting for interference and network geometry [75]. This method enhances the convergence speed of consensus algorithms, crucial for maintaining model performance in decentralized systems with complex communication dynamics.

Semi-decentralized approaches, such as TT-HF, support local model updates and consensus among devices before aggregating at the server, balancing the benefits of decentralized and centralized training paradigms [76]. This hybrid approach allows for more flexible and efficient model training, particularly in environments with varying network conditions and resource availability.

These decentralized model training frameworks underscore the critical need for adaptive and resource-efficient strategies in distributed machine learning, addressing challenges such as low training efficiency, limited computational resources, and communication bottlenecks. For instance, the Fed-DUMAP framework enhances training efficiency by leveraging shared server data and implementing dynamic updates, while adaptive optimization methods and layer-specific pruning ensure superior accuracy and reduced computational costs. Consensus-driven learning techniques facilitate model training across unreliable networks with local datasets, and communication-efficient algorithms alleviate bandwidth constraints by enabling dynamic peer selection and model sparsification, ultimately improving convergence performance and efficiency in decentralized learning environments [13, 73, 24, 72]. By leveraging techniques such as hierarchical clustering, clipping, and consensus algorithms, these frameworks address key challenges related to data heterogeneity, communication efficiency, and system scalability, paving the way for robust and scalable distributed learning applications.



(a) Vertical and Horizontal Learning in a Networked System[77]

(b) The image depicts a layered structure of a system, with each layer representing a different aspect of the system's functionality.[32]

(c) Quantization for Free: A Comparative Study of Dithering and Random-k Methods in Star-100[64]

Figure 6: Examples of Decentralized Model Training Frameworks

As shown in Figure 6, the exploration of distributed adaptive filtering and consensus algorithms is pivotal in enhancing the efficiency and scalability of machine learning systems. The examples in Figure 6 highlight three significant aspects of decentralized model training frameworks. Firstly, "Vertical and Horizontal Learning in a Networked System" illustrates a model where data centers and client devices collaborate through a network, emphasizing the dual paradigms of vertical and horizontal learning to optimize information flow and processing. Secondly, the layered structure depicted in the second image highlights the multifaceted functionality of decentralized systems, with distinct layers such as presentation and user services, contributing to overall system efficiency and user experience. Lastly, the "Quantization for Free" study compares dithering and random-k methods within the Star-100 framework, providing insights into their convergence behaviors and guiding the selection of appropriate quantization techniques in decentralized environments. Together, these examples underscore the complexity and potential of decentralized model training frameworks in advancing machine learning technologies [77, 32, 64].

AI-generated, for reference only.

6 Gradient Sparsification

6.1 Gradient Sparsification and Top-k Techniques

Gradient sparsification is pivotal in distributed optimization for reducing communication costs by transmitting only significant gradient components, thereby minimizing overhead while maintaining convergence. Techniques like DGS prioritize essential updates, enhancing convergence without local gradient accumulation [78]. The TopK SGD method exemplifies this by selecting the top K gradient components based on magnitude and accumulating others locally to ensure convergence [79]. Addressing the computational load of sorting in TopK SGD, innovations such as GaussianK offer tighter bounds for the TopK operator, reducing computational demands while preserving convergence [80]. These advancements are crucial for improving the efficiency of sparsification techniques.

Adaptive methods like FAB-topk enable clients to contribute a minimal number of gradient elements to the global sparse gradient, balancing communication and computation [81]. The hard-threshold sparsifier surpasses TopK in communication efficiency and convergence speed by minimizing total error during distributed training [82]. In federated learning, integrating sparsification with compression pipelines and scaling factors balances sparsity and learning progress, optimizing outcomes. The impact of various compression methods on convergence highlights the need for understanding these effects, as different techniques can yield diverse results [83]. For instance, FedPR uses prototypes to guide local training, improving accuracy and convergence [35].

The AdaGossip method excels in reducing communication overhead while maintaining model accuracy, marking a significant advancement in sparsification techniques [23]. These strategies are vital for optimizing communication efficiency, enhancing convergence rates, and sustaining model performance in distributed machine learning. By selectively transmitting critical updates, gradient sparsification and top-k techniques offer a compelling alternative to traditional compression methods, addressing scalability, resource constraints, and data privacy in modern machine learning applications.

6.2 Adaptive and Data-Aware Compression Techniques

Adaptive and data-aware compression techniques are crucial for enhancing communication efficiency in distributed machine learning, particularly through gradient sparsification, which strategically reduces exchanged data. This approach alleviates communication overhead, often limiting distributed systems' scalability. Recent advancements include adaptive gradient methods with compression, achieving a convergence rate of $O(1/\sqrt{T})$ for non-convex problems. Systems like BytePS-Compress improve training times for models like ResNet50, VGG16, and BERT-base by significant margins while maintaining accuracy. A convex optimization formulation for sparsification minimizes communication costs by randomly omitting gradient coordinates, ensuring unbiased remaining data and optimizing efficiency in large-scale machine learning [2, 30]. These techniques dynamically adjust compression levels based on data characteristics and training dynamics, enhancing training efficiency and convergence rates.

The DEFT framework overcomes traditional methods' limitations by eliminating gradient build-up and balancing gradient selection across workers, improving performance and efficiency [84]. The High Performance Gradient Sparsification (HPGS) method exemplifies the evolution of sparsification strategies by eliminating the need for sorting gradients, enhancing training efficiency for large-scale distributed learning [85]. TopK SGD remains a foundational technique, significantly reducing communication overhead while maintaining accuracy [86]. Its efficiency in large-scale distributed training is well-documented, highlighting its relevance in resource-constrained environments.

Adaptive techniques like MiCRO emphasize precise parameter tuning for optimal performance across varying training settings [87]. For example, MiCRO's initial threshold requires careful adjustment to ensure effective sparsification without compromising outcomes.

Integrating adaptive and data-aware strategies into sparsification techniques marks a significant evolution, enhancing communication efficiency and reducing computational overhead. These advancements address critical federated learning challenges, especially with non-i.i.d. local datasets. By employing adaptive gradient sparsity, these methods ensure equitable client contributions and optimize the communication-computation trade-off, leading to substantial improvements in accuracy—demonstrated by up to a 40

AI-generated, for reference only.

7 Federated Learning

7.1 Integration of Compressed Communication Techniques

In federated learning (FL), integrating compressed communication techniques is vital for alleviating communication bottlenecks and addressing resource constraints prevalent in distributed machine learning. These techniques enhance system robustness and communication efficiency, enabling effective model training even in resource-limited settings. Methods such as gradient compression, adaptive sparsification, and data-aware model selection significantly reduce communication overhead, accelerate training, and improve scalability. BytePS-Compress and Salient Grads exemplify substantial enhancements in training performance and communication efficiency, particularly under high-dimensional and non-IID data conditions, without sacrificing model accuracy [5, 2, 30, 88, 89].

Gradient compression frameworks reduce communication costs by compressing and reconstructing gradient updates, ensuring efficient FL system operations. Combining random sparsification with gradient perturbation enhances privacy while maintaining model accuracy, demonstrating the synergy between privacy-preserving mechanisms and communication efficiency [90].

Liu et al. propose a four-layer functional architecture for FL, emphasizing the integration of communication-efficient techniques across presentation, user services, FL training, and infrastructure layers [29]. This structured approach supports scalable and robust FL deployments.

The induced compressor framework optimizes communication by involving only a subset of nodes in each FL round, effectively managing communication costs and improving convergence rates in heterogeneous environments [39]. ADOTA-FL leverages historical gradient data to mitigate channel disturbances, enhancing convergence and training stability [91]. This strategy underscores the potential of historical data in optimizing communication processes.

Efficient and secure FL frameworks balance communication efficiency with privacy concerns through secure aggregation techniques and adaptive client sampling, addressing statistical and system heterogeneity to reduce convergence time and enhance model performance [92]. Recent frameworks like FedDUMAP and FedDUAP demonstrate improved training speed and accuracy by utilizing shared data and adaptive pruning, while FAT-Clipping addresses challenges of fat-tailed noise, ensuring robust convergence across varied settings [69, 93, 94, 24].

7.2 Innovative Federated Learning Frameworks

Innovative federated learning (FL) frameworks advance distributed machine learning by overcoming traditional limitations and exploring new applications. These frameworks enhance interpretability, decentralized aggregation, and expand FL's applicability in intelligent systems, addressing challenges like system heterogeneity, efficient communication, and robust model performance [32].

Over-the-air aggregation methods significantly improve FL system accuracy and efficiency, particularly in wireless communication environments where efficient data transmission is crucial [95]. Adaptive and secure communication strategies balance efficiency with privacy and security, utilizing sophisticated aggregation methods and adaptive client sampling to manage statistical and system heterogeneity. Notably, some approaches achieve a 73

These innovative FL frameworks represent significant advances in distributed machine learning, addressing interpretability, decentralized aggregation, and communication efficiency. Continuous exploration and enhancement of frameworks like FedDUMAP and FedDUAP are essential for maximizing FL's effectiveness in intelligent systems, enhancing model accuracy, convergence speed, and computational efficiency [41, 93, 94, 24].

7.3 Challenges in Federated Learning

Federated learning (FL) faces challenges impacting scalability, efficiency, and privacy. A major issue is the inefficiency of a single global model across diverse user data, leading to suboptimal performance due to client data distribution heterogeneity [96]. This heterogeneity complicates convergence and necessitates strategies for managing diverse data characteristics.

Synchronous updates in cloud settings can delay FL system efficiency, requiring all clients to synchronize updates, leading to substantial waiting times, especially with variable network conditions and client participation rates [96]. Communication overhead of model updates necessitates efficient protocols.

Balancing compression and convergence speed is crucial, as compression reduces communication costs but may slow convergence if not managed carefully, especially in resource-constrained environments [90]. Privacy concerns require protecting sensitive user data while enabling effective model training. Integrating privacy-preserving techniques like differential privacy and secure communication protocols mitigates communication overhead and training variance due to data heterogeneity, preserving model utility. Joint privacy enhancement and quantization methods illustrate balancing privacy and model performance by augmenting updates with privacy-preserving noise [97, 98].

Addressing FL challenges, including data heterogeneity, communication bottlenecks, and privacy concerns, is crucial for developing scalable, efficient, and privacy-preserving solutions that accommodate diverse distributed learning environments while adhering to legal and regulatory frameworks [29, 93, 32].

8 Challenges and Future Directions

8.1 Future Directions and Research Opportunities

Advancing distributed machine learning systems presents numerous research opportunities to overcome existing challenges while enhancing scalability and effectiveness. A key focus is refining the Efficient Adaptive Federated Optimization (EAFO) algorithm to balance the trade-offs between communication, computation, and precision, extending its applicability across diverse machine learning frameworks [4]. In decentralized learning, further exploration of the AdaGossip algorithm, particularly its theoretical convergence properties and adaptation to complex graph structures, could address challenges in decentralized environments [23]. The Sparse Binary Compression (SBC) technique could benefit from adaptive strategies that dynamically adjust sparsity levels based on real-time communication constraints, enhancing distributed system effectiveness [26].

Research in decentralized federated learning should focus on optimizing communication strategies in environments with fluctuating link reliability. Enhancing the robustness of methods like Prox-LEAD under diverse conditions and extending their applicability to various optimization frameworks could significantly propel the field forward [27]. Techniques such as asynchronous communication or local steps could improve spectral gap dependencies in decentralized stochastic optimization, enhancing method efficiency [22]. Refining the Byz-VR-MARINA method for complex scenarios, including decentralized learning environments, presents another promising research direction [6]. In distributed learning over wireless networks, developing robust resource allocation methods, enhancing privacy, and integrating emerging technologies like edge computing are crucial [38].

Collectively, these research opportunities aim to enhance the robustness, efficiency, and scalability of distributed machine learning systems, addressing the diverse challenges of modern applications. By exploring innovative methodologies such as interpretable data fusion, consensus-driven learning, and advanced federated learning frameworks like FedDUAP and FedDUMAP, researchers can significantly improve the adaptability and effectiveness of distributed learning systems. These approaches enhance model accuracy and training efficiency, ensure data privacy, and facilitate intuitive human-machine interactions, paving the way for more robust and user-friendly learning frameworks [73, 5, 24, 94, 32].

8.2 Algorithmic and Computational Challenges

Implementing compressed communication techniques in distributed machine learning presents algorithmic and computational challenges that can hinder system efficiency and scalability. A critical issue is the assumption of bounded gradients, which may not hold in practical scenarios, adversely affecting algorithm performance like SPARQ-SGD [99]. This limitation necessitates developing more robust methods capable of handling unbounded gradients effectively. High communication costs associated with gradient aggregation in low bandwidth networks underscore the need for layer-wise sparsification techniques to enhance communication efficiency [100].

Algorithmic challenges also arise in federated learning, where transmitting dense model weights can lead to inefficiencies. While Salient Grads seeks to address these inefficiencies, it may struggle in environments with frequently changing network topologies or inadequate peer selection that does not account for dynamic bandwidth variations. Moreover, the PoSw algorithm is susceptible to collusion attacks, where compromised models could skew consensus, presenting a significant challenge in federated learning [101]. The trade-off between compression and convergence speed remains a pressing issue in federated learning. Techniques such as FedPAQ can complicate matters by extending the local update period, potentially introducing noise and suboptimal convergence if not managed carefully [3]. Furthermore, existing differential privacy mechanisms introduce noise proportional to model size, which can degrade accuracy [90]. These privacy-preserving strategies add complexity, impacting model performance and increasing vulnerability to data poisoning and inference attacks.

The design of coding and control mechanisms in dynamic environments poses additional challenges, as current frameworks must adapt to evolving goals [20]. Implementing joint privacy enhancement and quantization methods relies on specific assumptions about model update distributions, requiring meticulous parameter tuning for optimal performance [97]. Moreover, convergence guarantees for gradient sparsification techniques, especially in non-convex smooth objectives, remain a theoretical challenge [79]. Extreme data heterogeneity and managing step size variance in federated learning techniques continue to present significant challenges [33]. These issues emphasize the need for ongoing innovation and research in distributed machine learning, aiming to develop more efficient, scalable, and resilient systems that can effectively meet the diverse demands of modern applications.

8.3 Integration and Application of Emerging Techniques

The integration and application of emerging techniques in distributed machine learning, particularly within federated learning frameworks, are crucial for advancing the field. A significant focus is optimizing the balance between privacy and accuracy, which is essential for maintaining data integrity while ensuring robust model performance. Advanced differential privacy techniques are vital for enhancing privacy measures without substantially compromising accuracy, especially in the context of device heterogeneity [102]. Exploring resilient optimization methods, such as the RTTS-LSGD algorithm, presents another promising avenue. Future research could investigate relaxing operational conditions or adapting it to various adversarial settings, thereby broadening its applicability [103]. Additionally, the development of resource-aware asynchronous frameworks like PAO-Fed highlights the potential for refining weight-decreasing systems for delayed updates, which could be further explored in complex environments to optimize resource usage and model convergence [104].

The integration of signal processing techniques with federated learning remains a significant area for innovation. Addressing unresolved questions regarding this integration could yield more robust methods capable of resisting adversarial attacks, enhancing the resilience and effectiveness of federated learning systems [31]. Future work should focus on privacy enhancements, communication complexity analysis, and extending these techniques to decentralized federated learning setups to broaden their applicability and impact [105]. Adaptive federated dropout techniques, which involve dynamically selecting sub-models, represent another exciting research direction. Investigating their impact on fairness in federated learning could provide insights into optimizing performance across heterogeneous client populations, ensuring equitable outcomes and efficient resource utilization [106].

These emerging techniques collectively offer substantial opportunities for enhancing the scalability, efficiency, and robustness of distributed machine learning systems. By prioritizing the integration of advanced privacy measures, such as differential privacy and random sparsification, alongside resilient optimization techniques and adaptive frameworks, researchers can significantly improve the effectiveness and versatility of federated learning applications. This approach addresses critical challenges, including data privacy, communication efficiency, and the statistical heterogeneity of decentralized data, ultimately leading to more robust and efficient machine learning models across diverse applications, particularly within IoT networks [90, 93, 94, 102].

9 Conclusion

The survey highlights the essential role of compressed communication techniques in advancing distributed machine learning and signal processing. Techniques such as Deep Gradient Compression (DGC) and error feedback mechanisms like EF-SGD effectively reduce communication bandwidth while maintaining model accuracy, crucial for scalable distributed learning. These methods address communication bottlenecks and resource constraints, enabling more efficient learning frameworks.

In federated learning, frameworks such as FLARE and FedCPU demonstrate the ability to sustain or enhance model accuracy under communication limitations, showcasing the effectiveness of compressed communication in optimizing federated applications. The use of ternary quantization methods, as implemented in T-FedAvg, further reduces communication costs while achieving competitive performance on benchmark datasets. These advancements are vital for optimizing federated learning in wireless networks, where efficient data transmission is paramount.

Furthermore, frameworks like DAGC significantly decrease the training iterations needed to reach target accuracies, illustrating the success of adaptive compression strategies in non-IID scenarios. The ProxSkip method achieves linear speedup in decentralized optimization, enhancing communication efficiency and exhibiting robustness against data heterogeneity. These strategies highlight the potential of communication-efficient methods to improve both the efficiency and resilience of distributed learning systems.

The survey also points to the potential of decentralized frameworks such as d-SVD, which improve communication efficiency and convergence in decentralized environments, addressing scalability challenges in ELAA systems. The application of d-regular expander graphs in DFL accelerates convergence and strengthens robustness to client failures, validating the theoretical benefits of the proposed overlay networks.

These findings underscore the critical role of compressed communication techniques in overcoming challenges related to communication bottlenecks, resource constraints, and system heterogeneity in distributed machine learning and signal processing. These innovations pave the way for more efficient, scalable, and resilient distributed learning systems capable of operating effectively in diverse and resource-constrained settings. Ongoing research and innovation will be crucial to further optimize these techniques and broaden their applicability to increasingly complex and dynamic learning environments.

References

[1] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.

[2] Yuchen Zhong, Cong Xie, Shuai Zheng, and Haibin Lin. Compressed communication for distributed training: Adaptive methods and system, 2021.

[3] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization, 2020.

[4] Zunming Chen, Hongyan Cui, Ensen Wu, and Yu Xi. Efficient adaptive federated optimization of federated learning for iot, 2022.

[5] Mengchen Fan, Baocheng Geng, Keren Li, Xueqian Wang, and Pramod K. Varshney. Interpretable data fusion for distributed learning: A representative approach via gradient matching, 2024.

[6] Eduard Gorbunov, Samuel Horváth, Peter Richtárik, and Gauthier Gidel. Variance reduction is an antidote to byzantines: Better rates, weaker assumptions and communication compression as a cherry on the top, 2023.

[7] Li Chou, Zichang Liu, Zhuang Wang, and Anshumali Shrivastava. Efficient and less centralized federated learning, 2021.

[8] Pavana Prakash, Jiahao Ding, Maoqiang Wu, Minglei Shu, Rong Yu, and Miao Pan. To talk or to work: Delay efficient federated learning over mobile edge devices, 2021.

[9] Liang Li, Dian Shi, Ronghui Hou, Hui Li, Miao Pan, and Zhu Han. To talk or to work: Flexible communication compression for energy efficient federated learning over heterogeneous mobile edge devices. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

[10] Divyansh Jhunjhunwala, Advait Gadhikar, Gauri Joshi, and Yonina C. Eldar. Adaptive quantization of model updates for communication-efficient federated learning, 2021.

[11] Rongwei Lu, Yutong Jiang, Yinan Mao, Chen Tang, Bin Chen, Laizhong Cui, and Zhi Wang. Data-aware gradient compression for fl in communication-constrained mobile computing, 2024.

[12] Haichuan Yang, Shupeng Gui, Yuhao Zhu, and Ji Liu. Automatic neural network compression by sparsity-quantization joint learning: A constrained optimization-based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2178–2188, 2020.

[13] Zhenheng Tang, Shaohuai Shi, and Xiaowen Chu. Communication-efficient decentralized learning with sparsification and adaptive peer selection, 2020.

[14] Zijie Yan, Danyang Xiao, Mengqiang Chen, Jieying Zhou, and Weigang Wu. Dual-way gradient sparsification for asynchronous distributed deep learning. In *Proceedings of the 49th International Conference on Parallel Processing*, pages 1–10, 2020.

[15] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.

[16] El Houcine Bergou, Konstantin Burlachenko, Aritra Dutta, and Peter Richtárik. Personalized federated learning with communication compression, 2022.

[17] Laizhong Cui, Xiaoxin Su, and Yipeng Zhou. A fast blockchain-based federated learning framework with compressed communications, 2022.

[18] S Vineeth. Unbiased single-scale and multi-scale quantizers for distributed optimization, 2022.

AI-generated, for reference only.

[19] Jiacheng Yao, Wei Xu, Zhaohui Yang, Xiaohu You, Mehdi Bennis, and H. Vincent Poor. Wireless federated learning over resource-constrained networks: Digital versus analog transmissions, 2024.

[20] Chao Zhang, Hang Zou, Samson Lasaulce, Walid Saad, Marios Kountouris, and Mehdi Bennis. Goal-oriented communications for the iot and application to data compression. *IEEE Internet of Things Magazine*, 5(4):58–63, 2023.

[21] Alexander Tyurin, Marta Pozzi, Ivan Ilin, and Peter Richtárik. Shadowheart sgd: Distributed asynchronous sgd with optimal time complexity under arbitrary computation and communication heterogeneity, 2024.

[22] Rustem Islamov, Yuan Gao, and Sebastian U. Stich. Towards faster decentralized stochastic optimization with communication compression, 2024.

[23] Sai Aparna Aketi, Abolfazl Hashemi, and Kaushik Roy. Adagossip: Adaptive consensus step-size for decentralized deep learning with communication compression, 2024.

[24] Ji Liu, Juncheng Jia, Hong Zhang, Yuhui Yun, Leye Wang, Yang Zhou, Huaiyu Dai, and Dejing Dou. Efficient federated learning using dynamic update and adaptive pruning with momentum on shared server data, 2024.

[25] Egor Shulgin and Peter Richtárik. Towards a better theoretical understanding of independent subnetwork training, 2024.

[26] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Sparse binary compression: Towards distributed deep learning with minimal communication, 2018.

[27] Yao Li, Xiaorui Liu, Jiliang Tang, Ming Yan, and Kun Yuan. Decentralized composite optimization with compression, 2021.

[28] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training, 2020.

[29] Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, 64(4):885–917, 2022.

[30] Gradient sparsification for comm.

[31] Tomer Gafni, Nir Shlezinger, Kobi Cohen, Yonina C. Eldar, and H. Vincent Poor. Federated learning: A signal processing perspective, 2021.

[32] Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. From distributed machine learning to federated learning: A survey, 2022.

[33] Langming Liu and Dingxuan Zhou. Analysis of regularized federated learning, 2024.

[34] Bin Wang, Jun Fang, Hongbin Li, Xiaojun Yuan, and Qing Ling. Confederated learning: Federated learning with decentralized edge servers, 2022.

[35] Yu Qiao, Huy Q. Le, and Choong Seon Hong. Boosting federated learning convergence with prototype regularization, 2023.

[36] Weicai Li, Tiejun Lv, Wei Ni, Jingbo Zhao, Ekram Hossain, and H. Vincent Poor. Decentralized federated learning over imperfect communication channels, 2024.

[37] Hao Ye, Le Liang, and Geoffrey Li. Decentralized federated learning with unreliable communications, 2021.

[38] Mingzhe Chen, Deniz Gündüz, Kaibin Huang, Walid Saad, Mehdi Bennis, Aneta Vulgarakis Feljan, and H. Vincent Poor. Distributed learning in wireless networks: Recent progress and future challenges, 2021.

AI-generated, for reference only.

[39] Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning, 2021.

[40] Solmaz Niknam, Harpreet S Dhillon, and Jeffrey H Reed. Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Communications Magazine*, 58(6):46–51, 2020.

[41] Liangqi Yuan, Ziran Wang, Lichao Sun, Philip S. Yu, and Christopher G. Brinton. Decentralized federated learning: A survey and perspective, 2024.

[42] Alexander Tyurin and Peter Richtárik. A computation and communication efficient method for distributed nonconvex problems in the partial participation setting, 2024.

[43] Rustem Islamov, Xun Qian, and Peter Richtárik. Distributed second order methods with fast rates and compressed communication, 2021.

[44] Mohammad Taha Toghani and César A. Uribe. Scalable average consensus with compressed communications, 2021.

[45] Zhen Zhang, Shaofu Yang, and Wenying Xu. Decentralized admm with compressed and event-triggered communication, 2023.

[46] Alexander Tyurin and Peter Richtárik. 2direction: Theoretically faster distributed training with bidirectional communication compression, 2023.

[47] Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Ef21-p and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International Conference on Machine Learning*, pages 11761–11807. PMLR, 2023.

[48] Shigang Li and Torsten Hoefler. Near-optimal sparse allreduce for distributed deep learning, 2022.

[49] Lusine Abrahamyan, Yiming Chen, Giannis Bekoulis, and Nikos Deligiannis. Learned gradient compression for distributed deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7330–7344, 2021.

[50] William Zou, Hans De Sterck, and Jun Liu. Downlink compression improves topk sparsification, 2022.

[51] Berivan Isik, Francesco Pase, Deniz Gunduz, Sanmi Koyejo, Tsachy Weissman, and Michele Zorzi. Adaptive compression in federated learning via side information, 2024.

[52] Grant Wilkins, Sheng Di, Jon C. Calhoun, Zilinghan Li, Kibaek Kim, Robert Underwood, Richard Mortier, and Franck Cappello. Fedsz: Leveraging error-bounded lossy compression for federated learning communications, 2024.

[53] WANG Luping, WANG Wei, and LI Bo. Cmfl: Mitigating communication overhead for federated learning. In *2019 IEEE 39th international conference on distributed computing systems (ICDCS)*, pages 954–964. IEEE, 2019.

[54] Eirikur Agustsson and Lucas Theis. Universally quantized neural compression. *Advances in neural information processing systems*, 33:12367–12376, 2020.

[55] Julieta Martinez, Jashan Shewakramani, Ting Wei Liu, Ioan Andrei Bârsan, Wenyuan Zeng, and Raquel Urtasun. Permute, quantize, and fine-tune: Efficient compression of neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15699–15708, 2021.

[56] Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

[57] Wan Jiang, Gang Liu, Xiaofeng Chen, and Yipeng Zhou. Deep hierarchy quantization compression algorithm based on dynamic sampling, 2022.

[58] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[59] Jiaming Yan, Jianchun Liu, Shilong Wang, Hongli Xu, Haifeng Liu, and Jianhua Zhou. Heroes: Lightweight federated learning with neural composition and adaptive local update in heterogeneous edge networks, 2023.

[60] Tianqu Kang, Lumin Liu, Hengtao He, Jun Zhang, S. H. Song, and Khaled B. Letaief. The effect of quantization in federated learning: A rényi differential privacy perspective, 2024.

[61] Yue Yu, Jiaxiang Wu, and Longbo Huang. Double quantization for communication-efficient distributed optimization, 2019.

[62] Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method, 2021.

[63] Shayan Mohajer Hamidi and Ali Bereyhi. Rate-constrained quantization for communication-efficient federated learning, 2024.

[64] Dmitry Kovalev, Anastasia Koloskova, Martin Jaggi, Peter Richtarik, and Sebastian U. Stich. A linearly convergent algorithm for decentralized optimization: Sending less bits for free!, 2020.

[65] Shaohuai Shi, Zhenheng Tang, Qiang Wang, Kaiyong Zhao, and Xiaowen Chu. Layer-wise adaptive gradient sparsification for distributed deep learning with convergence guarantees. In *ECAI*, pages 1467–1474, 2020.

[66] Sihua Wang, Mingzhe Chen, Christopher G. Brinton, Changchuan Yin, Walid Saad, and Shuguang Cui. Performance optimization for variable bitwidth federated learning in wireless networks, 2023.

[67] Ema Becirovic, Zheng Chen, and Erik G Larsson. Optimal mimo combining for blind federated edge learning with gradient sparsification. In *2022 IEEE 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC)*, pages 1–5. IEEE, 2022.

[68] Chun-Chih Kuo, Ted Tsei Kuo, and Chia-Yu Lin. Improving federated learning communication efficiency with global momentum fusion for gradient compression schemes, 2022.

[69] Haibo Yang, Peiwen Qiu, and Jia Liu. Taming fat-tailed ("heavier-tailed” with potentially infinite variance) noise in federated learning, 2022.

[70] Wei Liu, Li Chen, and Wenyi Zhang. Decentralized federated learning: Balancing communication and computing costs, 2022.

[71] Yifan Shi, Li Shen, Kang Wei, Yan Sun, Bo Yuan, Xueqian Wang, and Dacheng Tao. Improving the model consistency of decentralized federated learning, 2023.

[72] Zhenheng Tang, Shaohuai Shi, and Xiaowen Chu. Communication-efficient decentralized learning with sparsification and adaptive peer selection. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 1207–1208. IEEE, 2020.

[73] Kyle Crandall and Dustin Webb. Consensus driven learning, 2020.

[74] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data, 2020.

[75] Sundaram Vanka, Martin Haenggi, and Vijay Gupta. Convergence speed of the consensus algorithm with interference and sparse long-range connectivity, 2010.

[76] Frank Po-Chen Lin, Seyyedali Hosseinalipour, Sheikh Shams Azam, Christopher G. Brinton, and Nicolo Michelusi. Semi-decentralized federated learning with cooperative d2d local model aggregations, 2021.

[77] Anirban Das and Stacy Patterson. Multi-tier federated learning for vertically partitioned data, 2021.

[78] Zijie Yan. Gradient sparification for asynchronous distributed training, 2019.

[79] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information Processing Systems*, 31, 2018.

[80] Shaohuai Shi, Xiaowen Chu, Ka Chun Cheung, and Simon See. Understanding top-k sparsification in distributed deep learning, 2019.

[81] Pengchao Han, Shiqiang Wang, and Kin K Leung. Adaptive gradient sparsification for efficient federated learning: An online learning approach. In *2020 IEEE 40th international conference on distributed computing systems (ICDCS)*, pages 300–310. IEEE, 2020.

[82] Rethinking gradient sparsificati.

[83] Linxuan Pan and Shenghui Song. Local sgd accelerates convergence by exploiting second order information of the loss function, 2023.

[84] Daegun Yoon and Sangyoon Oh. Deft: Exploiting gradient norm difference between model layers for scalable gradient sparsification, 2023.

[85] Daegun Yoon and Sangyoon Oh. Empirical analysis on top-k gradient sparsification for distributed deep learning in a supercomputing environment, 2022.

[86] Dan Alistarh, Torsten Hoefler, Mikael Johansson, Sarit Khirirat, Nikola Konstantinov, and Cédric Renggli. The convergence of sparsified gradient methods, 2018.

[87] Daegun Yoon and Sangyoon Oh. Micro: Near-zero cost gradient sparsification for scaling and accelerating distributed dnn training, 2024.

[88] Riyasat Ohib, Bishal Thapaliya, Pratyush Gaggenapalli, Jingyu Liu, Vince Calhoun, and Sergey Plis. Salientgrads: Sparse models for communication efficient and data aware distributed federated training, 2023.

[89] Md Zarif Hossain and Ahmed Imteaj. Fedavo: Improving communication efficiency in federated learning with african vultures optimizer, 2023.

[90] Rui Hu, Yanmin Gong, and Yuanxiong Guo. Federated learning with sparsification-amplified privacy and adaptive optimization. *arXiv preprint arXiv:2008.01558*, 2020.

[91] Chenhao Wang, Zihan Chen, Nikolaos Pappas, Howard H. Yang, Tony Q. S. Quek, and H. Vincent Poor. Adaptive federated learning over the air, 2024.

[92] Tao Liu, Zhi Wang, Hui He, Wei Shi, Liangliang Lin, Wei Shi, Ran An, and Chenhao Li. Efficient and secure federated learning for financial applications, 2023.

[93] Taki Hasan Rafi, Faiza Anan Noor, Tahmid Hussain, Dong-Kyu Chae, and Zhaohui Yang. A generalized look at federated learning: Survey and perspectives, 2023.

[94] Hong Zhang, Ji Liu, Juncheng Jia, Yang Zhou, Huaiyu Dai, and Dejing Dou. Fedduap: Federated learning with dynamic update and adaptive pruning using shared data on the server, 2022.

[95] Ema Becirovic, Zheng Chen, and Erik G. Larsson. Optimal mimo combining for blind federated edge learning with gradient sparsification, 2022.

[96] Ruiyuan Wu, Anna Scaglione, Hoi-To Wai, Nurullah Karakoc, Kari Hreinsson, and Wing-Kin Ma. Federated block coordinate descent scheme for learning global and personalized models, 2021.

[97] Natalie Lang, Elad Sofer, Tomer Shaked, and Nir Shlezinger. Joint privacy enhancement and quantization in federated learning, 2022.

[98] Nima Mohammadi, Jianan Bai, Qiang Fan, Yifei Song, Yang Yi, and Lingjia Liu. Differential privacy meets federated learning under communication constraints, 2021.

[99] Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. Sparq-sgd: Event-triggered and compressed communication in decentralized stochastic optimization, 2020.

[100] Shaohuai Shi, Qiang Wang, Kaiyong Zhao, Zhenheng Tang, Yuxin Wang, Xiang Huang, and Xiaowen Chu. A distributed synchronous sgd algorithm with global top-k sparsification for low bandwidth networks. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 2238–2247. IEEE, 2019.

[101] Ali Raza, Kim Phuc Tran, Ludovic Koehl, and Shujun Li. Proof of swarm based ensemble learning for federated learning applications, 2023.

[102] Sofia Zahri, Hajar Bennouri, and Ahmed M. Abdelmoniem. An empirical study of efficiency and privacy of federated learning algorithms, 2023.

[103] Amit Dutta and Thinh T. Doan. Resilient two-time-scale local stochastic gradient descent for byzantine federated learning, 2024.

[104] Francois Gauthier, Vinay Chakravarthi Gogineni, Stefan Werner, Yih-Fang Huang, and Anthony Kuh. Resource-aware asynchronous online federated learning for nonlinear regression, 2021.

[105] Mohammad Taha Toghani and César A. Uribe. Unbounded gradients in federated learning with buffered asynchronous aggregation, 2022.

[106] Nader Bouacida, Jiahui Hou, Hui Zang, and Xin Liu. Adaptive federated dropout: Improving communication efficiency and generalization for federated learning, 2020.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.