Les Cahiers du GERAD ISSN: 0711–2440

Distributed stochastic gradient descent with quantized compressive sensing

D. Mitra,
A. Khisti

G–2020–23–EIW14

April 2020

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2020
– Bibliothèque et Archives Canada, 2020

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2020
– Library and Archives Canada, 2020

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Distributed stochastic gradient descent with quantized compressive sensing

Dipayan Mitra
Ashish Khisti

Department of Electrical & Computer Engineering,
University of Toronto, Toronto (Ontario) Canada,
M5S 1A1

mitradi2@ece.utoronto.ca
akhisti@ece.utoronto.ca

April 2020
Les Cahiers du GERAD
G–2020–23–EIW14

Abstract: *One of the major challenges in large-scale distributed machine learning involving stochastic gradient methods is the high cost of gradient communication over multiple nodes. Gradient quantization and sparsification have been studied to reduce the communication cost. In this work we bridge the gap between gradient sparsity and quantization. We propose a quantizatized compressive sensing-based approach to address the issue of gradient communication. Our approach compresses the gradients by a random matrix and apply 1-bit quantization to reduce the communication cost. We also provide a theoretical analysis on the convergence of our approach, under the gradient bound assumption.*

1 Introduction

The advent of internet-of-things (IoT) has changed the way data is collected and processed. Millions of connected users are generating huge amount of un-processed data, often sensitive. Distributed processing, exploiting data-parallelism, is often adopted to train a large-scale machine learning model [13, 18, 22]. Chen et al. proposed Synchronous stochastic gradient descent (Sync-SGD), which is often used as a preferred distributed optimization technique [5]. Sync-SGD consists of a centralised *parameter server* and a number of *workers* performing the following tasks:

- Parameter server communicates the model parameters with each worker.
- Each worker, in parallel, computes the gradients on a mini-batch of training data. Gradients are then sent to parameter server.
- Upon receiving gradient updates from all participating workers, parameter server performs a gradient aggregation. Global model parameters are updated based on the aggregated gradients and sent to each worker.

The above process is repeated until the model converges.

Although Sync-SGD performs well while the number of participating workers are scaled up, communication of gradients between the workers and the parameter server causes a bottleneck [10, 21]. In other words, the total time required for one complete iteration of Sync-SGD can be categorized as gradient computation and communication by each worker. Earlier studies have identified gradient communication cost to be more challenging over the gradient computation cost [25, 26, 27]. Hence compressing the gradients to reduce the communication overhead is widely studied.

In literature various techniques involving sparsity and quantization of deep neural networks (DNNs) have been explored [2, 8, 9, 12, 23, 24]. Stich et al. proposed gradient sparsification technique, where each worker sends top-k gradient parameters to the parameter server [15, 19]. Although such heuristic technique works well in practice, scaling up the number of workers leads to poor compression performance and divergence [20]. However, to the best of our knowledge, compressive sensing has not been used, exploiting sparsity of the gradient updates.

In this paper, we propose a quantized compressive sensing approach to reduce the gradient communication time. We use compressive sensing to acquire compressed measurements from the sparse gradient updates. Compressed measurements are further quantized using a 1-bit quantizer, for the ace of hardware implementation. Our work has two major contributions: (1) we propose a quantized compressive sensing-based approach to reduce the the gradient communication cost in the distributed learning, (2) we provide a theoretical analysis on the convergence of the proposed approach.

2 Motivation

Our work is motivated by the observation of sparsity, induced by Rectified Linear Unit (ReLU) activation layers, in DNNs. ReLU activation function takes the following form: $f(x) = \max(0, x)$ and

forces the gradients to be 0 $\forall x < 0$ [11]. As a result on average 44% of the operations performed in most of the modern DNNs, for example AlexNet, GoogLeNet etc., are *ineffective* [17]. In our work, compressive sensing is used to exploit this sparsity of DNNs.

3 Background

3.1 Compressive sensing

Compressive sensing is a sampling technique for signals which are sparse or compressible in some known basis [6]. Let us assume an N dimensional signal $\mathbf{x}$ which is K'-sparse [1], where $K' << N$. The sensing process can be defined as follows,

$$\mathbf{y}_{M\times 1} = \mathbf{\Phi}_{M\times N}\mathbf{x}_{N\times 1} \tag{1}$$

$$= \mathbf{\Phi}_{M\times N}\mathbf{\Psi}_{N\times N}\mathbf{s}_{N\times 1} \tag{2}$$

where $\mathbf{y}$ denotes M dimensional measurement vector, $\mathbf{\Phi}$ and $\mathbf{\Psi}$ denote measurement matrix and the sparsifying basis respectively. Here $\mathbf{s}$ denotes the sparse vector.

Once the compressive measurements are obtained, goal of the reconstruction is to find the sparsest solution from $\mathbf{y}$. Although the sparsest solution can be obtained by solving ℓ_0 optimization problem, it is computationally complex. Instead, in classical compressive sensing ℓ_1 minimization problem is solved to obtain the sparse solution, which is theoretically proven to be equivalent to minimizing ℓ_0 optimization problem [4, 7]. The reconstruction of the compressed measurements can be expressed as follows,

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{\Phi}\mathbf{x} \tag{3}$$

where $||.||_1$ represents ℓ_1 norm.

3.2 Quantized Compressive Sensing

Boufounos et al. introduced quantized compressive sensing (QCS) where the quantization is modeled as an additive measurement noise, shown in equation 4 [3].

$$\mathbf{y} \quad = \quad Q(\mathbf{\Phi}\mathbf{x}) = \mathbf{\Phi}\mathbf{x} + \mathbf{e} \tag{4}$$

where $Q(.)$ denotes the quantizer. Measurement noise $\mathbf{n}$ is bounded by the quantization interval Δ and the dimension of the compressed measurement (M) as follows [3],

$$||\mathbf{e}||^2 \leq \sqrt{\frac{M\Delta^2}{12}} = \epsilon \tag{5}$$

In QCS reconstructed signal can be obtained by solving,

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{y} - \Phi\mathbf{x}\|_2 \leq \epsilon \tag{6}$$

A LP-based reconstruction algorithm can be used to obtain $\widehat{\mathbf{x}}$ from the compressed measurements $\mathbf{y}$ [3, 4]. In this work, the reconstruction error β is considered to be a factor accounting both the quantization error (during measurement) and the LP-based reconstruction error. Equation 7 shows the aforementioned noise, which can be bounded by a positive quantity β.

$$||\widehat{\mathbf{x}} - \mathbf{x}||^2 = ||\mathbf{n}||^2 \leq \beta \tag{7}$$

In the next section we discuss the proposed QCS-based gradient compression in distributed learning.

[1] Although in literature sparsity is denoted as K, we use K' to denote sparsity for avoiding conflict with the total number of workers (denoted by K).

4 Proposed approach

4.1 Problem formulation

Let us consider K number of workers participating in a distributed learning process to evaluate parameters $\mathbf{w}$ on training samples $\mathbf{x}$, drawn (i.i.d.) from a probability distribution $dP(\mathbf{x})$. At t-th iteration, a mini-batch of training samples are split and evenly distributed among K workers. Each worker computes its local gradients $\mathbf{g}_t^{(k)}$ with respect to its training samples $\mathbf{x}_t^{(k)}$ and communicates the update with the parameter server to perform aggregation, following: $\mathbf{g}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{g}_t^{(k)}$. The aggregated global gradient $\mathbf{g}_t$ is used to evaluate the updated model parameter $\mathbf{w}_{t+1}$ as,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma \mathbf{g}_t \tag{8}$$

where γ denotes learning rate. Updated parameters are sent back to each worker to compute the gradients for the $(t+1)$-th iteration [14, 23]. The process is repeated untiall convergence is attained.

4.2 Proposed distributed learning approach

Based on the motivation (see Section 2), we use QCS to compress the sparse gradients and obtain the quantized compressed measurement $\mathbf{y}_t^{(k)}$ as follows,

$$\mathbf{y}_t^{(k)} = Q(\mathbf{\Phi}^{(k)}\mathbf{g}_t^{(k)}) \tag{9}$$

Each worker sends the compressed measurements or $\mathbf{y}_t^{(k)}$ to the parameter server. As the quantization is performed on the compressed gradients, our approach requires lower communication cost over standard gradient quantization approaches (where quantization is performed directly on the gradients).

At the parameter server the quantized compressed measurements are recovered to obtain $\tilde{\mathbf{g}}_t^{(k)}$. Parameter server performs the gradient aggregation $\tilde{\mathbf{g}}_t = \frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{g}}_t^{(k)}$ followed by the parameter update shown as,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma \tilde{\mathbf{g}}_t \tag{10}$$

where γ denotes the learning rate.

Note: In this work, we focused on providing a convergence analysis for a general setup where each worker uses different measurement matrix (which is available to the parameter server a priori). As a result, parameter server needs to perform QCS recovery K times. Whereas, if each user uses same measurement matrix, parameter server would require to perform QCS recovery only once. Following measurement vector would be considered in the aforementioned case: $\mathbf{y}_t = Q(\mathbf{\Phi}[\frac{1}{K}\sum_{k=1}^{K}\mathbf{g}_t^{(k)}])$. Aggregated gradient $\tilde{\mathbf{g}}_t$ can be obtained by performing recovery only once, as opposed to K times. The convergence rate can be modified accordingly.

4.3 Convergence analysis

In this section we analyse the convergence of the proposed approach in the non-convex setting, which is typical in most of the deep learning systems. For this analysis we follow the standard assumptions of the stochastic optimization summarized by Allen et al. [1].

Assumption 1 *$\forall \mathbf{w}$ and some constant f^*, global objective function $f(\mathbf{w}) > f^*$.*

Above assumption guarantees the convergence of the global objective function to a stationary point.

Assumption 2 *Let $\bar{\mathbf{g}}(\mathbf{w})$ denote $\nabla\mathbf{f}(\mathbf{w})$ evaluated at $\mathbf{w} = [w_1, w_2, \ldots, w_d]^T$. Then $\forall \mathbf{w}$, $\mathbf{\Theta} = [\theta_1, \theta_2, \ldots, \theta_d]^T$ and a non-negative constant vector $\mathbf{L} = [l_1, l_2, \ldots, l_d]^T$,*

$$|f(\Theta) - [f(\mathbf{w}) + \bar{\mathbf{g}}(\mathbf{w})^T(\Theta - \mathbf{w})]| \leq \frac{1}{2}\sum_{i=1}^{d} l_i(\theta_i - w_i)^2$$

Above assumption acts as a smoothness criteria. We define $l' = ||\mathbf{L}||_\infty$.

Assumption 3 *Stochastic gradient* $\mathbf{g}(\mathbf{w})$ *is an unbiased estimate having bounded coordinate variance* $\mathbb{E}[\mathbf{g}(\mathbf{w})] = \bar{\mathbf{g}}(\mathbf{w})$ *and,*

$$\mathbb{E}[(\mathbf{g^{(k)}}(\mathbf{w})_i - \bar{\mathbf{g}}(\mathbf{w})_i)^2] \leq \sigma_i^2$$

for some non-negative constant vector $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_d]^T$.

Assumption 4 *Let* $\bar{\mathbf{n}}_t = \mathbb{E}[\mathbf{n}_t]$ *and there exists a non-negative* μ *such that (for* $\mu < 1$*),*

$$||\bar{\mathbf{n}}_t|| \leq \mu ||\bar{\mathbf{g}}_t||$$

Under assumptions {1, 2, 3, 4} we have the following convergence rate:

Theorem 1 *Let* T *be the total number of iterations and learning rate* $\gamma = \frac{1}{l' K \sqrt{T}}$ *and* f_0 *be the initial objective value. Then,*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} ||\bar{\mathbf{g}}_t||^2\right] \leq \frac{1}{\sqrt{T}}\left[\frac{l' K^2 (f_0 - f^*) + ||\boldsymbol{\sigma}||^2 + \beta}{1-\mu}\right]$$

Proof. From assumption 2 we can write,

$$f_{t+1} - f_t \leq \bar{\mathbf{g}}_t^T (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{1}{2}\sum_{i=1}^{d} l_i (\mathbf{w}_{t+1} - \mathbf{w}_t)_i^2 \tag{11}$$

where f_t denotes the global objective at t-th iteration and the gradient of which is denoted by $\bar{\mathbf{g}}_t$.

By taking the expected improvement conditioned on $\mathbf{w}_t$ we get,

$$\mathbb{E}[f_{t+1} - f_t | \mathbf{w}_t] \leq \overbrace{\mathbb{E}[\bar{\mathbf{g}}_t^T (\mathbf{w}_{t+1} - \mathbf{w}_t) | \mathbf{w}_t]}^{\mathbf{I}} + \overbrace{\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{d} l_i (\mathbf{w}_{t+1} - \mathbf{w}_t)_i^2 | \mathbf{w}_t\right]}^{\mathbf{II}} \tag{12}$$

Considering part **I** of equation 12 we can write,

$$\begin{aligned}
\mathbb{E}[\bar{\mathbf{g}}_t^T (\mathbf{w}_{t+1} - \mathbf{w}_t) | \mathbf{w}_t] &= -\mathbb{E}[\bar{\mathbf{g}}_t^T \gamma \frac{1}{K}\sum_{k=1}^{K} \tilde{\mathbf{g}}_t^{(k)} | \mathbf{w}_t] \\
&= -\gamma \bar{\mathbf{g}}_t^T \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} (\mathbf{g}_t^{(k)} - \mathbf{n}_t^{(k)}) | \mathbf{w}_t\right] \\
&= -\gamma \bar{\mathbf{g}}_t^T \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \mathbf{g}_t^{(k)} | \mathbf{w}_t\right] + \gamma \bar{\mathbf{g}}_t^T \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \mathbf{n}_t^{(k)} | \mathbf{w}_t\right] \\
&= -\gamma \bar{\mathbf{g}}_t^T \bar{\mathbf{g}}_t + \gamma \bar{\mathbf{g}}_t^T \bar{\mathbf{n}}_t
\end{aligned} \tag{13}$$

Considering Assumptions 3 and 4, Equation 13 can be simplified as below,

$$\begin{aligned}
\mathbb{E}[\bar{\mathbf{g}}_t^T (\mathbf{w}_{t+1} - \mathbf{w}_t) | \mathbf{w}_t] &\leq -\gamma ||\bar{\mathbf{g}}_t||^2 + \gamma ||\bar{\mathbf{g}}_t|| ||\bar{\mathbf{n}}_t|| \\
&\leq -\gamma ||\bar{\mathbf{g}}_t||^2 + \gamma \mu ||\bar{\mathbf{g}}_t||^2
\end{aligned} \tag{14}$$

Considering **II** of Equation 12,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{d} l_i(\mathbf{w}_{t+1}-\mathbf{w}_t)_i^2|\mathbf{w}_t\right] \leq & \mathbb{E}\left[\frac{1}{2}l'||\mathbf{w}_{t+1}-\mathbf{w}_t||^2|\mathbf{w}_t\right] \\ = & \mathbb{E}\left[\frac{1}{2}l'\left|\left|\frac{\gamma}{K}\sum_{k=1}^{K}\tilde{\mathbf{g}}_t^{(k)}\right|\right|^2\middle|\mathbf{w}_t\right] \\ \leq & \frac{l'\gamma^2}{2K}\sum_{k=1}^{K}\mathbb{E}[||\tilde{\mathbf{g}}_t^{(k)}||^2|\mathbf{w}_t] \\ = & \frac{l'\gamma^2}{2K}\sum_{k=1}^{K}\mathbb{E}[||\mathbf{g}_t^{(k)}-\mathbf{n}_t^{(k)}||^2|\mathbf{w}_t] \\ \leq & \frac{l'\gamma^2}{2K}\left[\sum_{k=1}^{K}2\mathbb{E}\Big[||\mathbf{g}_t^{(k)}||^2|\mathbf{w}_t\Big]+\sum_{k=1}^{K}2\mathbb{E}\Big[||\mathbf{n}_t^{(k)}||^2|\mathbf{w}_t\Big]\right]\end{aligned} \tag{15}$$

From the variance bound of Assumption 3 we can write,

$$\mathbb{E}\Big[||\mathbf{g}_t^{(k)}-\bar{\mathbf{g}}_t||^2|\mathbf{w}_t\Big] \leq ||\boldsymbol{\sigma}||^2 \tag{16}$$

Equation 16 can be re-written as,

$$\begin{aligned}||\boldsymbol{\sigma}||^2 \geq & \mathbb{E}\Big[||\mathbf{g}_t^{(k)}-\bar{\mathbf{g}}_t||^2\Big] \\ = & \mathbb{E}\Big[||\mathbf{g}_t^{(k)}||^2-2\bar{\mathbf{g}}_t^T\mathbf{g}_t^{(k)}+||\bar{\mathbf{g}}_t||^2\Big] \\ = & \mathbb{E}\Big[||\mathbf{g}_t^{(k)}||^2\Big]-2\bar{\mathbf{g}}_t^T\mathbb{E}\Big[||\mathbf{g}_t^{(k)}||\Big]+||\bar{\mathbf{g}}_t||^2 \\ = & \mathbb{E}\Big[||\mathbf{g}_t^{(k)}||^2\Big]-2\bar{\mathbf{g}}_t^T\bar{\mathbf{g}}_t+||\bar{\mathbf{g}}_t||^2 \\ = & \mathbb{E}\Big[||\mathbf{g}_t^{(k)}||^2\Big]-||\bar{\mathbf{g}}_t||^2\end{aligned} \tag{17}$$

From Equation 17 we get,

$$\mathbb{E}\Big[||\mathbf{g}_t^{(k)}||^2\Big] \leq ||\boldsymbol{\sigma}||^2+||\bar{\mathbf{g}}_t||^2 \tag{18}$$

Substituting Equations 7 and 18 into Equation 15 we can write,

$$\mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{d} l_i(\mathbf{w}_{t+1}-\mathbf{w}_t)_i^2|\mathbf{w}_t\right] \leq \gamma^2 l'\left[||\boldsymbol{\sigma}||^2+||\bar{\mathbf{g}}_t||^2+\beta\right] \tag{19}$$

Combining Equation 14 and 19 Equation 12 can be simplified as shown below,

$$\mathbb{E}[f_{t+1}-f_t|\mathbf{w}_t] \leq -\gamma||\bar{\mathbf{g}}_t||^2+\gamma\mu||\bar{\mathbf{g}}_t||^2+\gamma^2 l'\left[||\boldsymbol{\sigma}||^2+||\bar{\mathbf{g}}_t||^2+\beta\right] \tag{20}$$

Substituting the value of γ in Equation 20,

$$\begin{aligned}\mathbb{E}[f_{t+1}-f_t|\mathbf{w}_t] \leq & ||\bar{\mathbf{g}}_t||^2\Big(\frac{1}{l'TK^2}-\frac{1-\mu}{l'\sqrt{T}K}\Big)+\frac{1}{l'TK^2}(||\boldsymbol{\sigma}||^2+\beta) \\ \leq & -\frac{1-\mu}{l'\sqrt{T}K^2}||\bar{\mathbf{g}}_t||^2+\frac{1}{l'TK^2}(||\boldsymbol{\sigma}||^2+\beta)\end{aligned}$$

Let us further extend the expectation over randomness in the trajectory and perform a telescoping sum over all the iterations. We obtain,

$$\begin{aligned}
f_0 - f^* \geq & f_0 - \mathbb{E}[f_T] \\
= & \mathbb{E}\left[\sum_{t=0}^{T-1}(f_t - f_{t+1})\right] \\
\geq & \frac{1}{l'}\mathbb{E}\sum_{t=0}^{T-1}\left[\frac{(1-\mu)||\bar{\mathbf{g}}_t||^2}{\sqrt{T}K^2} - \frac{||\boldsymbol{\sigma}||^2+\beta}{TK^2}\right] \\
= & \mathbb{E}\left[\frac{\sqrt{T}(1-\mu)}{l'K^2}||\bar{\mathbf{g}}_t||^2 - \frac{||\sigma||^2+\beta}{l'K^2}\right] \\
= & \frac{1}{l'K^2}\left\{\sqrt{T}(1-\mu)\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}||\bar{\mathbf{g}}_t||^2\right] - (||\sigma||^2+\beta)\right\}
\end{aligned}$$

By rearranging the above inequality we can write,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}||\bar{\mathbf{g}}_t||^2\right] \leq \frac{1}{\sqrt{T}}\left[\frac{l'K^2(f_0 - f^*) + ||\boldsymbol{\sigma}||^2 + \beta}{1-\mu}\right]$$

This completes the proof. ■

Note that the asymptotic convergence rate of the proposed approach is $\mathbf{O}\left(\frac{\beta}{\sqrt{T}}\right)$. In comparison, SGD has the same asymptotic convergence rate of $\mathbf{O}\left(\frac{\beta}{\sqrt{T}}\right)$. Earlier work on error-compensated DoubleSqueeze admits the same convergence rate of $\mathbf{O}\left(\frac{\beta}{\sqrt{T}}\right)$ [16].

5 Conclusion

In this work, we introduce a novel quantized compressive sensing-based gradient compression approach. We exploit the gradient sparsity induced by the activation layers in DNNs to reduce the communication cost between the workers and the parameter server in a distributed learning framework. We also provide a theoretical analysis on the convergence of the proposed approach.

Further studies would involve validation with experimental results. Comparison would be made against the state of the art quantization-based gradient compression technique. Suitability of a low complexity QCS recovery algorithm would be investigated. The work would further be extended into de-centralized setting exploiting joint sparsity.

6 Acknowledgement

We thank Nikhil Krishnan and Erfan Hosseini for interesting discussions and feedback in providing the convergence analysis.

References

[1] Zeyuan Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In ICML, 2017.

[2] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. Signsgd: Compressed optimisation for non-convex problems. ArXiv, abs/1802.04434, 2018.

[3] Petros Boufounos and Richard Baraniuk. 1-bit compressive sensing. In 42nd Annual Conference on Information Sciences and Systems, pages 16–21, March 2008.

[4] Emmanuel J. Candès. Theory of signals/mathematical analysis. Comptes rendus - Mathematique, 346(9-10):589–592, 2008.

[5] Jianmin Chen, Rajat Monga, Samy Bengio, and Rafal Józefowicz. Revisiting distributed synchronous sgd. ArXiv, abs/1604.00981, 2017.

[6] D Donoho. Compressed sensing. IEEE Transactions on Information Theory, 52(4):1289–1306, April 2006.

[7] David Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Comm. Pure Appl. Math, 59:797–829, 2004.

[8] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. CoRR, abs/1510.00149, 2015.

[9] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In Advances in Neural Information Processing Systems 29, pages 4107–4115. Curran Associates, Inc., 2016.

[10] Mu Li, David Andersen, Alexander Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In Advances in Neural Information Processing Systems 27, pages 19–27. Curran Associates, Inc., 2014.

[11] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In ICML, pages 807–814. Omnipress, 2010.

[12] Jongsoo Park, Sheng R. Li, Wei Wen, Ping Tak Peter Tang, Hai Li, Yiran Chen, and Pradeep Dubey. Faster cnns with direct sparse convolutions and guided pruning. In ICLR, 2016.

[13] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In Advances in Neural Information Processing Systems 24, pages 693–701. Curran Associates, Inc., 2011.

[14] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In CCS 2015 - Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pages 1310–1321, 10 2015.

[15] Sebastian Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In Advances in Neural Information Processing Systems 31, pages 4447–4458. Curran Associates, Inc., 2018.

[16] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. `DoubleSqueeze`: Parallel stochastic gradient descent with double-pass error-compensated compression. In Proceedings of the 36th International Conference on Machine Learning, volume 97, pages 6155–6165, Long Beach, California, USA, 09–15 Jun 2019.

[17] Albericio et al. Cnvlutin: Ineffectual-neuron-free deep neural network computing. In 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), pages 1–13, June 2016.

[18] Chen et al. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. ArXiv, abs/1512.01274, 2015.

[19] Dan et al. The convergence of sparsified gradient methods. In Advances in Neural Information Processing Systems 31, pages 5973–5983. Curran Associates, Inc., 2018.

[20] Ivkin et al. Communication-efficient distributed sgd with sketching. In NeurIPS, 2019.

[21] Li et al. Scaling distributed machine learning with the parameter server. In 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14), pages 583–598, Broomfield, CO, October 2014. USENIX Association.

[22] Martín et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. ArXiv, abs/1603.04467, 2015.

[23] Wei et al. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In Advances in Neural Information Processing Systems 30, pages 1509–1519. Curran Associates, Inc., 2017.

[24] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In Advances in Neural Information Processing Systems 29, pages 2074–2082. Curran Associates, Inc., 2016.

[25] X. Yao, C. Huang, and L. Sun. Two-stream federated learning: Reduce the communication costs. In 2018 IEEE Visual Communications and Image Processing (VCIP), pages 1–4, Dec 2018.

[26] Xin Yao, Chaofeng Huang, and Lifeng Sun. Two-stream federated learning: Reduce the communication costs. 2018 IEEE Visual Communications and Image Processing (VCIP), pages 1–4, 2018.

[27] Xin Yao, Tianchi Huang, Chenglei Wu, Rui-Xiao Zhang, and Lifeng Sun. Federated learning with additional mechanisms on clients to reduce communication costs. ArXiv, abs/1908.05891, 2019.