ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096363

CPA: COMPRESSED PRIVATE AGGREGATION FOR SCALABLE FEDERATED LEARNING OVER MASSIVE NETWORKS

Natalie Lang, Elad Sofer, Nir Shlezinger, Rafael G. L. D'Oliveira, and Salim El Rouayheb

ABSTRACT

Federated learning (FL) allows a central server to train a model using remote users' data. FL faces challenges in preserving the local datasets privacy and in its communication overhead; which is considerably dominant in large-scale networks. These limitations are often mitigated individually by local differential privacy (LDP) mechanisms, compression, and user-selection techniques, which often come at the cost of accuracy. In this work we present *compressed private aggregation (CPA)*, which allows massive deployments to simultaneously communicate at extremely low bit-rates while achieving privacy, anonymity, and resilience to malicious users. CPA randomizes a code-book for compressing the data into a few bits, ensuring anonymity and robustness, with a subsequent perturbation to hold LDP. We provide both a theoretical analysis and a numerical study, demonstrating the performance gains of CPA compared with separate mechanisms for compression and privacy.

Index Terms— FL, LDP, anonymity, compression.

1. INTRODUCTION

The unprecedented success of deep learning relies highly on the availability of data, often gathered by edge devices (e.g., mobile phones and vehicles), and which is readily exposed to potential privacy threats. *Federated learning (FL)* [1, 2] is an emerging paradigm for a remote training that avoids sharing data. Learning is performed locally, with periodic centralized aggregations of the models orchestrated by a server. Nonetheless, it was recently shown that private data can be extracted from the exchanged models updates if these are not properly protected [3,4]. Additionally, the repeated exchange of highly parameterized models results in considerable bottlenecks and delays [5], especially noticeable in large networks.

Various methods have been proposed to face these challenges: to guarantee privacy, local differential privacy (LDP) framework is widely adopted, relying on randomly corrupting the model [6]; the communication overhead is relaxed by sparsifying [7], sub-sampling [8], or quantization [9]. Scalability is enabled via partial user selection [5] based on prioritizing more influential users or individual resource constraints [10]. Other studies jointly tackle compression and privacy in FL, where [11] utilizes dithered quantization followed by LDP perturbations, and [12] transforms randomized lattices quantization byproduct distortion into privacy preserving noise (PPN). The above techniques either tackle each objective individually with possible excess distortion; restrict the available resources hence ultimately affecting the learned model accuracy; or are not inherently scalable to suit massive systems and tolerate large groups of colluding users.

In this work, we present a novel privacy preserving scheme designed for robust large-scale FL. The method, coined *compressed private aggregation (CPA)*, dramatically reduces communications by conveying the model updates via one bit messages, while providing provable k-anonymity and LDP. It inspired by private multi-group aggregation [13] and geo-indistinguishability [14] schemes, involving massive systems where scalability and robustness are key factors.

We leverage a random code-book to encode a set of model entries into a single bit. This representation provides anonymity; is simply extendable to hold LDP; and by utterly limiting each user's influence, is consequently resilient to malicious ones. The decoding procedure translates the received bits from all different users into an empirical discrete histogram over the model update values, whose aggregated mean converges into the averaged global trained model. By doing so, the server does not reconstruct the individual model updates, notably facilitating the participation of numerous users. At last, we numerically validate that CPA outperforms conventional methodologies for private compressed FL.

The rest of this paper is organized as follows: Section 2 briefly reviews the FL system model and relevant preliminaries. CPA is presented in Section 3 and numerically evaluated in Section 4. Finally, Section 5 provides concluding remarks.

2. SYSTEM MODEL & PRELIMINARIES

2.1. Federated Learning

In FL, a server trains a model parameterized by $\boldsymbol{w} \in \mathbb{R}^m$ using datasets available at a group of K users, denoted $\mathcal{D}_1, \ldots, \mathcal{D}_K$, assumed to be private. FL involves distributing a global model to the users, which locally train it using their data, sending back the model updates for aggregation [2].

Arguably the most common FL scheme is federated averaging (FedAvg) [1], where the global model is updated by averaging the local models. Letting $\boldsymbol{w}_t$ denote the global pa-

N. Lang, E. Sofer, and N. Shlezinger are with the School of ECE, Ben-Gurion University of the Negev, Be'er-Sheva, Israel (e-mails: langn, eladsofe@post.bgu.ac.il; nirshl@bgu.ac.il). R. G. L. D'Oliveira is with the School of Mathematical and Statistical Sciences, Clemson University, SC, (e-mail: rdolive@clemson.edu). S. El Rouayheb is with the Department of ECE, Rutgers University, Piscataway, NJ (e-mail: salim.elrouayheb@rutgers.edu).
The work of Salim El Rouayheb was supported in part by the National Science Foundation (NSF) under CNS Grants 1801630 and 2148182.

rameters at time step t, the server shares $\boldsymbol{w}_t$ with the users. The rth user performs τ training iterations using $\mathcal{D}_r$ to update $\boldsymbol{w}_t$ into $\boldsymbol{w}^r_{t+\tau}$. The updates $\boldsymbol{h}^r_{t+\tau} = \boldsymbol{w}^r_{t+\tau} - \boldsymbol{w}_t$ are shared with the server, that in turn sets the global model as

$$\boldsymbol{w}_{t+\tau} \triangleq \boldsymbol{w}_t + \sum_{r=1}^{K} \alpha_r \boldsymbol{h}^r_{t+\tau} = \sum_{r=1}^{K} \alpha_r \boldsymbol{w}^r_{t+\tau}. \tag{1}$$

As sharing $\boldsymbol{w}^r_{t+\tau}$ can possibly load the communication and leak private information, it motivates the integration of quantization and privacy enhancement techniques, discussed next.

2.2. Quantization Preliminaries

Vector quantization is the encoding of a set of continuous-amplitude quantities into a finite-bit representation [15]. The design of vector quantizers often relies on statistical modelling [16, Ch. 23], which is likely to be unavailable in FL [9]. Quantizers which are invariant of the underlying distribution are referred to as *universal*; a leading approach to implement such quantizers is based on lattice quantization [17]:

Definition 2.1 (Lattice Quantizer)**.** *A lattice quantizer of dimension $L \in \mathbb{Z}^+$ and generator $\boldsymbol{G} \in \mathbb{R}^{L\times L}$ maps $\boldsymbol{x} \in \mathbb{R}^L$ into a discrete $Q_{\mathcal{L}}(\boldsymbol{x})$ by selecting the nearest point in the lattice $\mathcal{L} \triangleq \{\boldsymbol{G}\boldsymbol{l} : \boldsymbol{l} \in \mathbb{Z}^L\}$, i.e., $Q_{\mathcal{L}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}\in\mathcal{L}} \|\boldsymbol{x}-\boldsymbol{z}\|$.*

To apply $Q_{\mathcal{L}}$ to a vector $\boldsymbol{x} \in \mathbb{R}^{ML}$, it is divided into $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M]^T$, and each sub-vector is quantized separately. A lattice $\mathcal{L}$ partitions $\mathbb{R}^L$ into cells centered around the lattice points. The number of lattice points in $\mathcal{L}$ is countable but infinite. To obtain a finite-bit representation, it is common to restrict $\mathcal{L}$ to include only points in a sphere of radius γ, and the number of points dictates the number of bits per sample R. For $L = 1$, $Q_{\mathcal{L}}(\cdot)$ specializes scalar uniform quantization.

2.3. Privacy Preliminaries

Privacy in FL is commonly quantified in terms of LDP [18], which provides guarantees in users-server settings.

Definition 2.2 (ϵ-LDP)**.** *A randomized mechanism $\mathcal{M}$ satisfies ϵ-LDP if for any pairs of input values v, v' in the domain of $\mathcal{M}$ and for any possible output y, it holds that*

$$\Pr[\mathcal{M}(v) = y] \leq e^{\epsilon} \Pr[\mathcal{M}(v') = y]. \tag{2}$$

Definition 2.2 can be interpreted as a bundle between stochasticity and privacy: if two different inputs are probable (up to a privacy budget) to be associated with the same output, then privacy is preserved, as each sample is not uniquely distinguishable. A principle method for achieving ϵ-LDP is the *randomized response (RR) mechanism* [19]. In RR, a user who possesses a private bit transmits it correctly with probability $p > 1/2$. By (2), it can be shown that RR satisfies $\log\left(\frac{p}{1-p}\right)$-LDP [20] and can be viewed as a PPN-adding mechanism to obfuscate individual data.

Although LDP is a preferable privacy measure, it often requires the introduction of a dominant PPN to be guaranteed. Alternative privacy measures, which are not inherently bundled with stochasticity, are based on anonymization, e.g.,

Definition 2.3 (k-anonymity [21])**.** *A deterministic mechanism $\mathcal{M}$ holds k-anonymity if for every input v in the domain of $\mathcal{M}$ there are at least $k-1$ different inputs $\{v'_i\}_{i=1}^{k-1}$, where*

$$\mathcal{M}(v) = \mathcal{M}(v'_i), \qquad \forall i \in \{1, \ldots, k-1\}. \tag{3}$$

If $\mathcal{M}$ satisfies k-anonymity, any observer of an output of $\mathcal{M}$ is unable to discriminate between at least k possible inputs.

2.4. Problem Formulation

Our goal is to design a privacy mechanism for FL which provides compression and is scalable. We are interested in obtaining a mapping $\boldsymbol{h}^r_t \mapsto \boldsymbol{w}_t$ of the local updates at the rth user into the global model available at the server, which meets the henceforth requirements. The scheme must be:

- *R1* *Private*: holding k-anonymity and ϵ-LDP with respect to the private dataset $\mathcal{D}_r$, for a given anonymity degree k and privacy budget ϵ, respectively.
- *R2* *Compressed*: communications to the server should involve at most 1-bit per sample.
- *R3* *Universal*: invariant to the distribution of $\boldsymbol{h}^r_t$.
- *R4* *Scalable*: operable with possibly millions of participants.
- *R5* *Robust*: resilient to malicious adversarial participants and tolerate a large group of colluding users.

Notice that we are focusing on achieving LDP in each time instance, which is known to enable privacy enhancement in multi-round FL training procedures [6].

Requirements *R1*-*R3* can be satisfied by first perturbing the data to meet *R1*, followed by universal quantization to satisfy *R2*-*R3*, as both techniques are invariant to the distribution of $\boldsymbol{h}^r_t$. However, the server decoding in these separate schemes requires individual reconstruction, which may result in violating *R4* while not accounting for *R5*. Furthermore, both privacy and quantization can be modelled as corrupting the model updates, motivating a joint design tailored for FL.

3. COMPRESSED PRIVATE AGGREGATION

3.1. The CPA Algorithm

We design CPA based on *R1*-*R5* by extending the recent schemes of [13] and [14] to FL settings. Broadly speaking, CPA leverages FL repeated communication to generate a random code-book and encode the data with the aid of a lattice quantizer (holding *R3*). The generated code enables the transmission of a set of model updates entries with a single bit (by *R2*), which guarantees k-anonymity of the data, and supporting LDP by applying RR to the conveyed bits (satisfying *R1*). In the decoding procedure, the received bits are translated into an empirical histogram over the model update values, rather than recovering each model update separately (holding *R4*). The aggregated mean over this histogram converges into the FedAvg trained model, inherently limiting the influence of potential malicious participating users as they can, at most, flip one-bit (assuring *R5*). These steps, illustrated in Fig. 1 and summarized as Algorithm 1, are described next in detail.

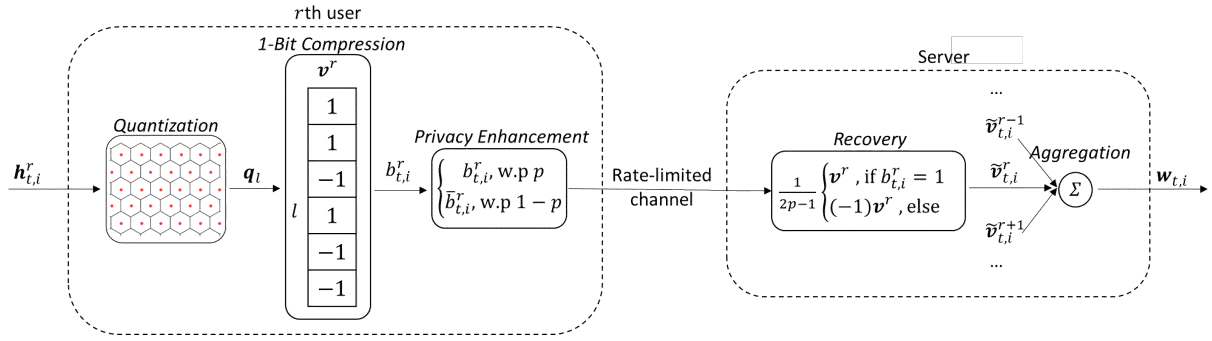

Fig. 1. Overview of CPA. The left dashed box represents the rth user encoding while the right describes the server decoding.

Initialization: At the beginning, the privacy parameters k and ϵ are set, and so does the compression lattice $\mathcal{L}$, i.e., fixing the lattice dimension L; its generator matrix $\boldsymbol{G}$; and radius γ (determined by the quantization rate R [22, Ch. 2], where R satisfies $k \leq 2^{LR-1}$). Moreover, a common seed s_r is shared between each user and the server, used for randomizing a code-word $\boldsymbol{v}^r$, which is uniformly distributed over all words in $\{-1,1\}^{2^{LR}}$ having equal amount of 1's and -1's.

Encoding: Since encoding is identical for all the users, we focus on the rth user, who is ready to transmit $\boldsymbol{h}_t^r$.

Quantization: CPA first maps the updates into a finite discrete index. To do so, $\boldsymbol{h}_t^r$ is divided into $M \triangleq \lceil \frac{d}{L} \rceil$ distinct $L \times 1$ vectors $\{\boldsymbol{h}_{t,i}^r\}_{i=1}^M$, and quantized by applying an L-dimensional lattice quantizer (Def. 2.1) to each $\boldsymbol{h}_{t,i}^r$.

1-Bit Compression: To proceed, we write $Q_{\mathcal{L}}(\boldsymbol{h}_{t,i}^r) = \boldsymbol{q}_l$ where $q_l \in \mathbb{R}^L$ is the lth lattice point. The user conveys a bit to the server based on the lth entry of $\boldsymbol{v}^r$, $\left[\boldsymbol{v}^r\right]_l$, where $b_{t,i}^r \triangleq 1$ if $\left[\boldsymbol{v}^r\right]_l = 1$; and $b_{t,i}^r \triangleq -1$ otherwise.

Privacy Enhancement: k-anonymity (Def. 2.3) directly follows from the design of $\boldsymbol{v}^r$, and to maintain ϵ-LDP, RR is applied to $b_{t,i}^r$: the true value is broadcasted with probability $p = \frac{e^\epsilon}{1+e^\epsilon}$, and its compliment with $1-p$.

Decoding: The server obtains $\{b_{t,i}^r\}_{r=1}^K$ and maps them into an empirical histogram, used to update the global model.

Recovery: Using the shared seed s_r, the server knows $\boldsymbol{v}^r$ and is able to restore each user code-word via

$$\tilde{\boldsymbol{v}}_{t,i}^r = \frac{1}{2p-1} \begin{cases} \boldsymbol{v}^r & \text{if } b_{t,i}^r = 1; \\ (-1) \cdot \boldsymbol{v}^r & \text{otherwise.} \end{cases} \tag{4}$$

Aggregation: The server then constructs with an aggregated mean of all $\{\tilde{\boldsymbol{v}}_{t,i}^r\}_{k=1}^K$, i.e. $\boldsymbol{v}_{t,i} \triangleq \frac{1}{K}\sum_{r=1}^K \tilde{\boldsymbol{v}}_{t,i}^r$. Practically, $\boldsymbol{v}_{t,i}$ is a discrete normalized histogram, representing how many users quantized their updated vector ith entry as a certain dictionary word. The histogram is utilized for updating the global model, replacing the FedAvg update in (1) by

$$\boldsymbol{w}_{t,i} = \boldsymbol{w}_{t-\tau,i} + \sum_{l=1}^{2^R} \left[\boldsymbol{v}_{t,i}\right]_l \cdot \boldsymbol{q}_l, \tag{5}$$

where $\boldsymbol{q}_l$ is the lth lattice point of $Q_{\mathcal{L}}$. The resulting global model is then obtained by stacking the sub-vectors $\{\boldsymbol{w}_{t,i}\}_{i=1}^M$.

3.2. Privacy Analysis

CPA aims to jointly support compression and privacy over large-scale networks: the former directly follows as each user conveys merely 1-bit per sample, i.e., M in a round; and the latter holds in both senses of LDP and k-anonymity:

Proposition 1. *CPA is ϵ-LDP with respect to $\mathcal{D}_r$.*

Proof. ϵ-LDP is obtained from the usage of RR combined with the LDP post-processing property [23]. □

Proposition 2. *CPA preserves k-anonymity with respect to the lattice quantization of $\boldsymbol{h}_{t,i}^r$.*

Proof. Since $\boldsymbol{v}^r$ has at least k identical entries for both ± 1, then $b_{t,i}^r$ can originate from k candidates for $Q_{\mathcal{L}}(\boldsymbol{h}_{t,i}^r)$. □

While Proposition 2 formulates the anonymity degree of each sub-vector, Corollary 1 reveals the higher degree of anonymity achieved with respect to the complete model:

Corollary 1. *CPA preserves k^M anonymity with respect to the lattice quantization of $\boldsymbol{h}_t^r$.*

3.3. Discussion

The ability of CPA to use 1-bit transmissions without notably affecting the trained model utility is a direct consequence of utilizing the unique characteristics of FL. These include sharing a source of common randomness and the aggregation nature of local updates, permitting to account on averages. Its operation is inherently robust, as any user manipulation is merely changing 1-bit, and scalable. In fact, for the latter, the learned model accuracy grows with the number of participating users, thus facilitating FL over large-scale networks.

Our framework can be extended to meet other privacy levels. E.g., using a uniform quantizer motivates a "continuous" type of k-anonymity, due to mapping a whole decision area into the same value; alternatively, adding stochasticity to the quantization can be utilized to hold other LDP additive noise mechanisms [12]. Considering communication, transmitting more than 1-bit from each user is also a possibility, that would change the length of the constructed histograms and therefore also the global update. Finally, the discrete histograms construction can account only on estimates of $\tilde{\boldsymbol{v}}_{t,i}^r$ crossing a certain threshold; shown to decrees noise [14] and allow a Byzantine robust training [24]. We leave the study of the aforementioned extensions for future investigation.

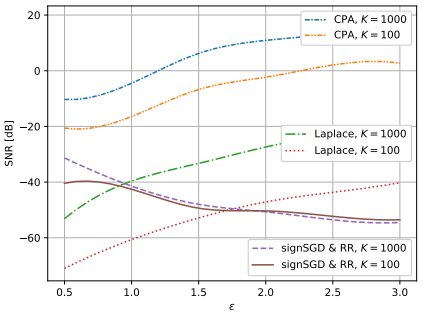

Fig. 2. SNR in received models versus ϵ for $K \in \{100, 1000\}$ edge users.

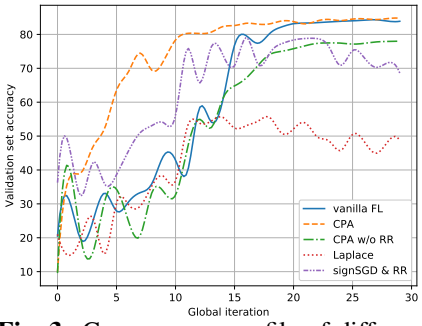

Fig. 3. Convergence profile of different FL schemes with $K = 1000$ edge users.

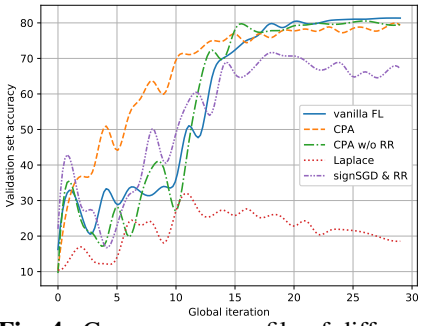

Fig. 4. Convergence profile of different FL schemes with $K = 100$ edge users.

Algorithm 1: CPA at time step t

Initialization:
  Shared seed s_r, degree of anonymity k, privacy budget ϵ, and lattice $\mathcal{L}$;
  Generate local code-word $\boldsymbol{v}^r$;
Encode (at the rth user side, for each i):
  Map $\boldsymbol{h}_{t,i}^r$ into $\boldsymbol{q}_l$ via the quantizer $Q_{\mathcal{L}}$;
  Set $b_{t,i}^r = 1$ if $\left[\boldsymbol{v}^r\right]_l = 1$ or $b_{t,i}^r = -1$ otherwise;
  Augment $b_{t,i}^r$ via RR and convey it to the server;
Decode (at the server side, for each i):
  Recover $\{\tilde{\boldsymbol{v}}_{t,i}^r\}_{r=1}^K$ via (4) and mean them all;
  Compute $\boldsymbol{w}_{t,i}$ using (5)

Result: The updated ith global model entry, $\boldsymbol{w}_{t,i}$;

4. EXPERIMENTAL STUDY

We consider the federated training[1] of a handwritten digit classification model using the MNIST dataset. A random subset of it is distributed among K users, who each possesses five samples for training a linear regression model with a softmax output layer, using local stochastic gradient descent (SGD) and a learning rate of 0.1. We evaluate the performance of CPA with a scalar quantizer, i.e., $L = 1$, bit-rate $R = 3$, and radius $\gamma = 0.05$, compared with: *vanilla FL* (without privacy or compression); *CPA without RR* (demonstrating compression without privacy); *Laplace* that is the perturbation of FL local updates with a Laplacian PPN, realizing Laplace mechanism [25] (the complement scheme of satisfying only privacy); and the common *signSGD* [26], which also utilizes 1-bit representations by replacing data with its sign, followed by *RR* (the straightforward separated design satisfying *R1*-*R5*).

We first validate that CPA reduces the overall noise better than its counterparts. To that aim, we evaluate for $K = \{100, 1000\}$ the observed SNR, defined as $\mathrm{Var}(\boldsymbol{w}_t)/\mathrm{Var}(\boldsymbol{w}_t - \tilde{\boldsymbol{w}}_t)$, versus different ϵ values and report that in Fig. 2. Evidently, CPA achieves the highest SNR which also, likewise Laplace, grow with looser privacy constraints and/or more users participating; while signSGD & RR demonstrates neither. This can be attributed to the coarse sign operation, whose distortion is so dominant such that it is sometimes reduced by privacy, and barely influenced by the number of edge users taking part in the FL training.

Next, we evaluate how the reduced excess distortion of CPA translates into an improved learning, and depict in Fig. 3 the validation set learning curves of all referenced methods, tested with $\epsilon = 1$ and $K = 1000$. Fig. 3 demonstrates that CPA achieves equivalent performance compared to vanilla FL which satisfies neither *R1* nor *R2*, while simultaneously satisfying both. Whereas our method is oriented for large-scale deployments, it is also valid with less edge uses, under relatively the same quality, as can be seen in Fig. 4, repeating the aforementioned scenario for $K = 100$. We further observe that the straightforward signSGD & RR suffers from excessive distortion that deteriorates its performance regardless of K, in line with similar findings in Fig. 2; as a result of using distinct quantization and privacy mechanism.

Finally, we verify that CPA can indeed tolerate colluding malicious participants. Table 1 reports the test accuracy of the converged models under the manipulations of a subset of the users, out of $K = 1000$ ones. User manipulation is either sending its 1-bit data constantly as $'1'$; or randomly flipping it; referenced to the result achieved with None. In the worst case, Table 1 reveals a degrade of 5% in performance for 30% of malicious users.

Table 1. CPA's test accuracy with a subset of malicious users.

Subset \ Operation	None	'1's	Flipping
20%	0.86	0.85	0.84
30%	0.86	0.85	0.81

5. CONCLUSIONS

We proposed CPA, which realizes quantization and privacy in scalable and robust FL. It utilizes lattice quantization; a random code-book; a dedicated RR mechanism; and discrete histogram aggregations to yield provable desired privacy and anonymity levels while minimizing the commutation overhead and tolerating malicious users manipulations. We numerically demonstrated that the algorithm outcomes with less distorted and more reliable models compared with other applications of compressed and private FL, while approaching the performance achieved with none of these constraints.

[1]The source code used in our experimental study is available online at https://github.com/langnatalie/CPA.

6. REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[2] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated learning: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 14–41, 2022.

[3] L. Zhu and S. Han, "Deep leakage from gradients," in *Federated learning*. Springer, 2020, pp. 17–31.

[4] B. Zhao, K. R. Mopuri, and H. Bilen, "iDLG: Improved deep leakage from gradients," *arXiv preprint arXiv:2001.02610*, 2020.

[5] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, 2021.

[6] L. Sun, J. Qian, X. Chen, and P. S. Yu, "LDP-FL: Practical private aggregation in federated learning with local differential privacy," *arXiv preprint arXiv:2007.15789*, 2020.

[7] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," *arXiv preprint arXiv:1704.05021*, 2017.

[8] C. Hardy, E. Le Merrer, and B. Sericola, "Distributed deep learning on edge-devices in the parameter server model," in *Workshop on Decentralized Machine Learning, Optimization and Privacy*, 2017.

[9] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 500–514, 2020.

[10] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *IEEE international conference on communications (ICC)*, 2019.

[11] K. Chaudhuri, C. Guo, and M. Rabbat, "Privacy-aware compression for federated data analysis," in *Conference on Uncertainty in Artificial Intelligence*, 2022.

[12] N. Lang, E. Sofer, T. Shaked, and N. Shlezinger, "Joint privacy enhancement and quantization in federated learning," *IEEE Trans. Signal Process.*, 2023.

[13] C. Naim, R. G. L. D'Oliveira, and S. El Rouayheb, "Private multi-group aggregation," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 3, pp. 800–814, 2022.

[14] R. Thesmar, J. Thesmar, R. G. L. D'Oliveira, and M. Medard, "Cabdriver: Concentration to accurate boundaries while distorting randomly input variables to elude recognition," in *WSA 2021; 25th International ITG Workshop on Smart Antennas*, 2021, pp. 1–6.

[15] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[16] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes for 6.441 (MIT), ECE563 (University of Illinois Urbana-Champaign), and STAT 664 (Yale)*, 2012-2017.

[17] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 428–436, 1992.

[18] Y. Wang, Y. Tong, and D. Shi, "Federated latent dirichlet allocation: A local differential privacy based framework," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 6283–6290, 2020.

[19] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[20] M. Yang, L. Lyu, J. Zhao, T. Zhu, and K.-Y. Lam, "Local differential privacy and its applications: A comprehensive survey," *arXiv preprint arXiv:2008.03686*, 2020.

[21] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[22] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*. Springer Science & Business Media, 2013, vol. 290.

[23] X. Xiong, S. Liu, D. Li, Z. Cai, and X. Niu, "A comprehensive survey on local differential privacy," *Security and Communication Networks*, vol. 2020, 2020.

[24] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5650–5659.

[25] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," *Journal of Privacy and Confidentiality*, vol. 7, no. 3, pp. 17–51, 2016.

[26] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning*. PMLR, 2018, pp. 560–569.