

A APPENDIX

A.1 PROOFS

Proof of theorem 1. Let $u \in \mathcal{B}^d$ and $U \sim \text{Unif}(\mathcal{B}^d)$. The vector $V \in \mathcal{B}^d$ sampled as in algorithm 1, has p.m.f.

$$p(v | u) \propto \begin{cases} 1/\mathbb{P}(\mathcal{M}(U, u) > \kappa) & \text{if } \mathcal{M}(v, u) > \kappa \\ 1/\mathbb{P}(\mathcal{M}(U, u) < \kappa) & \text{if } \mathcal{M}(v, u) \leq \kappa. \end{cases}$$

The event that $\mathcal{M}(U, u) = \kappa$ when $\frac{d+\kappa+1}{2} \in \mathbb{Z}$ implies that U and u match in exactly $\frac{d+\kappa+1}{2}$ coordinates; the number of such matches is $\binom{d}{(d+\kappa+1)/2}$. Computing the binomial sum, we have

$$\begin{aligned} \mathbb{P}(\mathcal{M}(U, u) > \kappa) &= \frac{1}{K^d} \sum_{\ell=\lceil \frac{d+\kappa+1}{2} \rceil}^d \binom{d}{\ell} (K-1)^{(d-\ell)} \\ \mathbb{P}(\mathcal{M}(U, u) \leq \kappa) &= \frac{1}{K^d} \sum_{\ell=0}^{\lceil \frac{d+\kappa+1}{2} \rceil - 1} \binom{d}{\ell} (K-1)^{(d-\ell)} \end{aligned} \tag{1}$$

Now we show unbiasedness via showing $\mathbb{E}[V | u = u] = m \cdot u$. We have

$$\mathbb{E}[V | u = u] = p\mathbb{E}[U | \mathcal{M}(U, u) > \kappa] + (1-p)\mathbb{E}[U | \mathcal{M}(U, u) \leq \kappa]$$

By rotational symmetry, it suffices to show:

$$\mathbb{E}[V^1 | u = u] = p \underbrace{\mathbb{E}[U^1 | \mathcal{M}(U, u) > \kappa]}_{\Upsilon_1} + (1-p) \underbrace{\mathbb{E}[U^1 | \mathcal{M}(U, u) \leq \kappa]}_{\Upsilon_2}$$

For Υ_1 , we have:

$$\begin{aligned} \Upsilon_1 &= \frac{1}{K^d \mathbb{P}(\mathcal{M}(U, u) > \kappa)} \times \sum_{\ell=\tau}^d \left(u^1 \binom{d}{\ell} (K-1)^{d-\ell} - \sum_{w \in \mathcal{B} \setminus u^1} w \binom{d}{\ell} (K-1)^{d-\ell-1} \right) \\ &= \frac{u^1}{K^d \mathbb{P}(\mathcal{M}(U, u) > \kappa)} \times \binom{d-1}{\tau-1} (K-1)^{(d-\tau)} \end{aligned}$$

Where in the second equality we used the fact that $\sum_{w \in \mathcal{B}} w = 0$. Similar calculations yield:

$$\Upsilon_2 = -\frac{u^1}{K^d \mathbb{P}(\mathcal{M}(U, u) \leq \kappa)} \times \binom{d-1}{\tau-1} (K-1)^{(d-\tau)}$$

Combining the preceding display with (1), we have:

$$\begin{aligned} \mathbb{E}[V^1 | u = u] &= \left(p \frac{\binom{d-1}{\tau-1} (K-1)^{d-\tau}}{\sum_{\ell=\tau}^d \binom{d}{\ell} (K-1)^{d-\ell}} - (1-p) \frac{\binom{d-1}{\tau-1} (K-1)^{d-\tau}}{\sum_{\ell=0}^{\tau-1} \binom{d}{\ell} (K-1)^{d-\ell}} \right) \times u^1 \\ &= mu^1 \end{aligned}$$

Next we show privacy guarantee. As $\mathbb{P}(\mathcal{M}(U, u) > \kappa)$ is decreasing in κ for any $u, u' \in \mathcal{B}^d$ and $v \in \mathcal{B}^d$ we have

$$\frac{p(v | u)}{p(v | u')} \leq \frac{p}{1-p} \cdot \frac{\mathbb{P}(\mathcal{M}(U, u') \leq \kappa)}{\mathbb{P}(\mathcal{M}(U, u) > \kappa)} = \frac{p}{1-p} \times \frac{\sum_{\ell=0}^{\tau-1} \binom{d}{\ell} (K-1)^{(d-\ell)}}{\sum_{\ell=\tau}^d \binom{d}{\ell} (K-1)^{(d-\ell)}}$$

The result follows by relation (5) \square

A.2 ON THE PRACTICALITY OF PRIVACY-PRESERVING SELECTION METHODS

Performing top- k selection with local privacy constraints could be done via two approaches. The first take is iteratively running the exponential mechanism (Dwork et al., 2014) for k times, each time selecting a single index with gumble noise. The bottleneck of this take is that it requires sampling

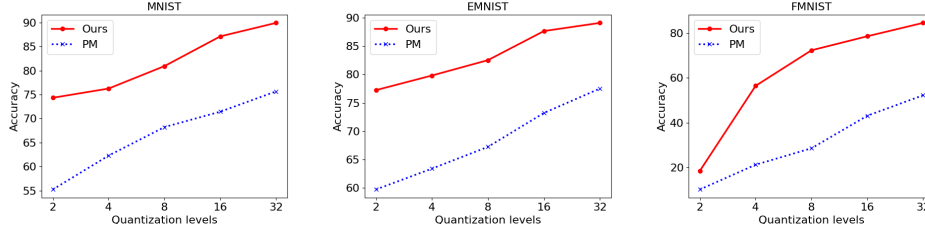


Figure 1: Comparison of sqSGD and sqSGD_{PTQ}

from a high dimensional distribution for k times, which is computationally heavy. Note that the privatization step is carried out on the client side device, which is usually assumed to be of limited computational power in FL settings (Kairouz et al., 2019), thus using iterative exponential mechanism is not practical for FL scenarios. The second take is the noisy top- k algorithm (Ding et al., 2019), which generalizes the report noisy max mechanism in Dwork et al. (2014). The algorithm requires adding Laplacian noise of scale $2Uk/\epsilon$ to each dimension of the gradient vector. In practice, k is typically chosen at the order of hundreds. Since most of the gradients are very small in magnitude, to ensure reasonable noise requires a high ϵ budget to allocate for the selection step. This would significantly affect the overall privacy level.

A.3 COMPARISONS OF sqSGD VS sqSGD_{PTQ}

We construct an instance of sqSGD_{PTQ} via utilizing the multi-dimensional piecewise mechanism PM detailed in Wang et al. (2019, Algorithm 4). As discussed in section 2.4, we need only slight modification to algorithm 2. In particular, we modify line 15 of algorithm 2 to be $Z_{s,t} = \text{PM}(\tilde{X}_{s,t})$. The sampling rate is chosen such that $\tilde{d} := \lfloor rd \rfloor < 2.5\epsilon$, as this guarantees that throughout the perturbation process of PM, there will be no additional subsampling performed, and the experimental results in section 4 do not necessarily implies the inferiority of the performance of PM. Throughout the training process, each client send gradient updates that are quantized to the range $\mathcal{B}^{\tilde{d}}$, with the radius chosen as $U = \frac{e^{\epsilon/2} + 1}{e^{\epsilon/2} - 1}$.

To compare the performance of sqSGD and sqSGD_{PTQ}, we evaluated both algorithms on MNIST, EMNIST and FMNIST datasets using the same federation scheme in section ?? . We fix the following hyperparameters: for the gradient update rule we set $\eta = 0.001$, $\alpha = \beta = 1.0$, for privacy level, we set $\epsilon = 400$ for MNIST and EMNIST, and $\epsilon = 2000$ for FMNIST. The sampling ratio is fixed at 0.5%. We compare the test set accuracy, using different quantization levels. The results are plotted in figure 1 The result suggests that sqSGD_{PTQ} is also valid for training large scale models, but to achieve comparable accuracy with sqSGD, a larger communication cost shall be sacrificed.

A.4 EXPERIMENTS ON THE EFFECT OF GRADIENT SUBSAMPLING

In this experiment, we investigate the effect of subsampling under the setup of training a ResNet110 model on the FMNIST dataset. We fix the privacy level at $\epsilon = 2000$ and quantization level at $K = 16$. We vary the subsampling ratio from the set $\{1, 5, 10, 50\} \times 10^{-3}$. The results are plotted in figure 2. It could be seen from the plot that using a high sampling ratio severely hurts the training performance, while also incurs more communication. It is thus necessary to perform subsampling. However, using a very low subsampling ratio also causes training failure. This phenomenon will be further explored in the next experiment.

A.5 EXPERIMENTS ON THE TRADE-OFF BETWEEN QUANTIZATION AND SAMPLING

In this experiment, we study the trade-off between quantization and sampling under the setup of training a LeNet-5 model on the MNIST dataset. We fix the privacy level at $\epsilon = 400$, and fix the total communication cost per client, measured using the product $r \log_2 K$. We vary the quantization level in the range $\{2, 8, 32, 128\}$ which corresponds to using $\{1, 3, 5, 7\}$ bits per client per dimension, and the sampling rate are adjusted accordingly. The results are shown in figure 3. The results suggest

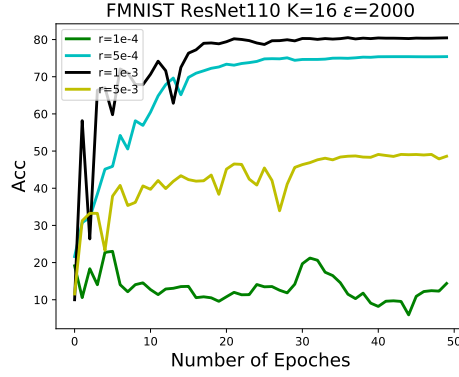


Figure 2: Study on the effect of gradient subsampling

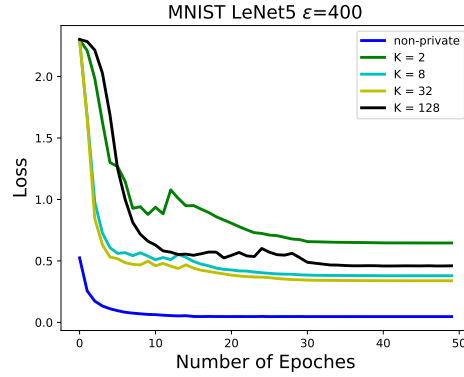


Figure 3: Study on the trade-off between quantization and sampling with fixed communication level

that increasing the quantization level may not monotonically increase training performance. This is mainly due to the random subsample scheme of sSGD, under which the structure of gradients is not fully explored.

REFERENCES

- Zeyu Ding, Yuxin Wang, Danfeng Zhang, and Daniel Kifer. Free gap information from the differentially private sparse vector and noisy max mechanisms, 2019.
- Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.
- Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D’Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrede Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2019.

Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 638–649. IEEE, 2019.