A PROOFS

Proof. (Of Lemma 5) Observe that in this case:

\[ \mathbb{E}(A_n(M(x_1), \ldots, M(x_n))) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(M(x_i)) = \frac{1}{n} \sum_{i=1}^{n} x_i = T(x_1, \ldots, x_n). \]

Additionally,

\[ \mathbb{E} \left[ \left( A_n(M(x_1), \ldots, M(x_n)) - \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}((M(x_i) - x_i)^2). \]

If \( M \) has bounded variance, then the variance of \( A_n(M(x_1), \ldots, M(x_n)) \) diminishes with \( n \). The rest of the lemma follows by an application of the Chebyshev’s inequality.

Proof. (Of Lemma 6) The proof generalizes the argument that the Laplace mechanism applied independently to each coordinate is differentially private for vectors with bounded \( L_1 \)-sensitivity. Let \( x \in \mathbb{R}^d \) with \( \|x\|_p \leq \Delta \), and let \( Q_0, Q_1 \) be density functions for the output distributions of \( M \) with or without the input \( x \). Then for any output value \( z \):

\[
\frac{Q_0(z)}{Q_1(z)} = \prod_{i=1}^{d} \frac{Q_0(x_i)}{Q_1(x_i)} \leq \prod_{i=1}^{d} \exp(\epsilon x_i^p) \quad \text{since } M \text{ is } \epsilon \text{-metric DP} \\
= \exp \left( \epsilon \sum_{i=1}^{d} x_i^p \right) \leq \exp(\epsilon \Delta^p).
\]

The reverse inequality can be derived similarly. Unbiasedness follows from the fact that \( M \) is unbiased for each dimension \( i = 1, \ldots, d \).

Proof. (Of Lemma 7) Let \( p \) be a feasible solution for (4). Let \( \odot \) and \( \odot^{-1} \) denote element-wise product and inverse, respectively. Then:

\[
\sum_{l=1}^{d} (D^a, p_l)_F = \sum_{l=1}^{d} (C \odot (D^a \odot C^\odot^{-1}), p_l)_F \\
\leq (D^a_j, C_j)_F \leq (D^a_j, C_j)_F (B - 1)^p \Delta^p \\
= d_0 D^a_j \ldots
\]

Theorem 1. Unbiased Bitwise Randomized Response satisfies \( \epsilon \)-local DP and is unbiased.

Proof. By standard proofs of the Randomized Response mechanism, transmitting bit \( j \) of \( z \) is \( \epsilon/b \)-differentially private. The entire procedure is thus \( \epsilon \)-differentially private by composition.

To show unbiasedness, first we observe that \( \mathbb{E}[z] = x \). Additionally, let us write \( z = \sum_{j=0}^{b-1} 2^{-j} z_j \). The transmitted number \( t \) is decoded as \( t = \sum_{j=0}^{b-1} 2^{-j} t_j \). Thus, if \( \mathbb{E}[t_j] = z_j \), then the entire algorithm is unbiased. Observe that:

\[
\mathbb{E}[t_j] = a_0 + (a_1 - a_0) \mathbb{E}[y_j] \\
= a_0 + (a_1 - a_0) \left( \frac{z_j}{1 + e^{-\epsilon/b}} + \frac{(1 - z_j) \cdot e^{-\epsilon/b}}{1 + e^{-\epsilon/b}} \right) \\
= z_j,
\]

where the last step follows from some algebra. The theorem follows by noting that \( \mathbb{E}[z] = x \) from properties of dithering.

Theorem 2. Unbiased Generalized Randomized Response satisfies \( \epsilon \)-local DP and is unbiased.
We show below that the increase in norm is not too high. While this leads to an unbiased solution, it might increase the norm of the vector.

Lemma 3. Let $v$ be a vector such that $\|v\| \leq 1$ and $v_i \in [-1, 1]$ for each coordinate $i$. Let $v'$ be the vector obtained by dithering each coordinate of $v$ to a grid of size $B$ (so that the difference between any two grid points is $\Delta = \frac{2}{B-1}$).

Then, with probability $\geq 1 - \delta$,

$$\|v'\|^2 \leq \|v\|^2 + \sqrt{2}\|v\|\Delta \log(4/\delta) + d\Delta^2/4 + \sqrt{2d\Delta \log(4/\delta)}.$$ 

Proof. Let $\Delta = \frac{2}{B-1}$ be the difference between any two grid points. For a coordinate $i$, let $v_i = \lambda_i + a_i$ where $\lambda_i$ is the closest grid point that is $\leq v_i$ and $a_i \geq 0$. We also let $v'_i = \lambda_i + Z_i$; observe that by the dithering algorithm, $Z_i \in (0, \Delta)$, with $\mathbb{E}[Z_i] = a_i$. Additionally, $\text{Var}(Z_i) \leq \Delta^2$.

Additionally, we observe that $\|v'\|^2 = \sum_i (\lambda_i + Z_i)^2 = \sum_i \lambda_i^2 + 2\lambda_i Z_i + Z_i^2$. By algebra, we get that:

$$\|v'\|^2 - \|v\|^2 = \sum_i (Z_i^2 - a_i^2) + \sum_i 2\lambda_i (Z_i - a_i)$$

We next bound these terms one by one. To bound the second term, we observe that $\mathbb{E}[Z_i] = a_i$ and apply Hoeffding’s inequality. This gives us:

$$\Pr(\sum_i \lambda_i Z_i \geq \sum_i \lambda_i a_i + t) \leq 2e^{-t^2/2\sum_i \lambda_i^2 \Delta^2}$$

Plugging in $t = \sqrt{2\sum_i \lambda_i^2 \Delta \log(4/\delta)}$ makes the right hand side $\leq \delta/2$. To bound the first term, we again use a Hoeffd-
We leave the design of more sophisticated vector dithering techniques that simultaneously preserve unbiasedness and norm bound for future work.

B.2 CONNECTION BETWEEN DP AND COMPRESSION

We highlight an interesting effect on the required communication budget as a result of adding differentially private noise. Figure 1 shows the optimized sampling probability matrix \( P \) for the MVU mechanism with a fixed input quantization level \( b_\text{in} = 5 \) and various values of \( b_\text{out} \). As \( b_\text{out} \) increases, the overall structure in the matrix \( P \) remains nearly the same but becomes more refined. Moreover, in the bottom right plot, it is evident that the marginal benefit to MSE becomes lower as \( b_\text{out} \) increases. This observation suggests that for a given \( \epsilon \), having more communication budget is eventually not beneficial to aggregation accuracy since the amount of information in the data becomes obscured by the DP mechanism and hence requires fewer bits to communicate.

B.3 DISTRIBUTED MEAN ESTIMATION

For the vector distributed mean estimation experiment in Section 5.1, the different private compression mechanisms used different values of the communication budget \( b \). We justify the choice of \( b \) as follows.

### Figure 2: DP-SGD training of a small convolutional network on MNIST with Gaussian mechanism, stochastic signSGD and MVU mechanism. Each point corresponds to a single hyperparameter setting, and dashed line shows Pareto frontier of privacy-utility trade-off.

#### L1-sensitivity setting.

CLDP outputs a total number of \( \log_2(d) + 1 = 8 \) bits, which is lower than that of both Skellam and MVU and cannot be tuned. Skellam performs truncation to the range \( \{-2^{b-1}, 2^{b-1} - 1\} \) after perturbing the quantized input with Skellam noise, and hence requires a value of \( b \) that is large enough to prevent truncation error. We intentionally afforded Skellam a large budget of \( b = 16 \) so that truncation error rarely occurs, and show that even in this setting MVU can outperform Skellam in terms of estimation MSE. For MVU, we chose \( b_\text{in} = 9 \), which is the minimum value required to avoid a large quantization error, and \( b = b_\text{out} = 3 \).

#### L2-sensitivity setting.

CLDP uses a communication budget of \( b = \log_2(d) + 1 = 8 \) per coordinate and is not tunable. We used the same \( b = 16 \) budget for Skellam as in the L1-sensitivity setting. For MVU, we chose \( b_\text{in} = 5 \) and \( b = b_\text{out} = 3 \) for both the L1- and L2-metric DP versions, which results in a communication budget that is lower than both CLDP and Skellam. For the L1-metric DP version, we found that optimizing MVU to satisfy \((\epsilon/2)\)-metric DP with respect to the L1 metric results in an \((\epsilon', \delta)\)-DP mechanism with \( \epsilon' \approx \epsilon \) and \( \delta = 1/(n + 1) \) after optimal RDP conversion.

B.4 PRIVATE SGD

In Section 5.2, we trained a linear model on top of features extracted by a scattering network on the MNIST dataset. In addition, we consider a convolutional network with tanh activation, which has been found to be more suitable for...
Layer Parameters

ScatterNet Scale $J = 2$, $L = 8$ angles, depth 2
GroupNorm \cite{Wu2018} 6 groups of 24 channels each
Fully connected 10 units

Table 1: Architecture for scatter + linear model.

Layer Parameters

Convolution + tanh 16 filters of $8 \times 8$, stride 2, padding 2
Average pooling $2 \times 2$, stride 1
Convolution + tanh 32 filters of $4 \times 4$, stride 2, padding 0
Average pooling $2 \times 2$, stride 1
Fully connected + tanh 32 units
Fully connected + tanh 10 units

Table 2: Architecture for convolutional network model.

Hyperparameter Values

Batch size 600
Momentum 0.5
# Iterations $T$ 500, 1000, 2000, 3000, 5000
Noise multiplier $\sigma$ for Gaussian and signSGD 0.5, 1, 2, 3, 5
$L_1$-metric DP parameter $\epsilon$ for MVU 0.25, 0.5, 0.75, 1, 2, 3, 5
Step size $\rho$ 0.01, 0.03, 0.1
Gradient norm clip $C$ 0.25, 0.5, 1, 2, 4, 8

Table 3: Hyperparameters for DP-SGD on MNIST.

Hyperparameter Values

Batch size 500
Momentum 0.5
# Iterations $T$ 1000, 2000, 3000, 5000, 10000, 15000
Noise multiplier $\sigma$ for Gaussian and signSGD 0.5, 1, 2, 3, 5
$L_1$-metric DP parameter $\epsilon$ for MVU 0.25, 0.5, 0.75, 1, 2, 3, 5
Step size $\rho$ 0.01, 0.03, 0.1
Gradient norm clip $C$ 0.25, 0.5, 1, 2, 4, 8

Table 4: Hyperparameters for DP-SGD on CIFAR-10.

DP-SGD \cite{Papernot2020}. We give the architecture details of both models in Tables 1 and 2.

Hyperparameters. DP-SGD has several hyperparameters, and we exhaustive test all setting combinations to produce the scatter plots in Figures 4 and 2. Tables 3 and 4 give the choice of values that we considered for each hyperparameter.

Result for convolutional network. Figure 2 shows the comparison of DP-SGD training with Gaussian mechanism, stochastic signSGD, and MVU mechanism with $b = 1$. The experimental setting is identical to that of Figure 4 except for the model being a small convolutional network trained end-to-end. We observe a similar result that MVU recovers the performance of signSGD at equal communication budget of $b = 1$.

References
