Privacy-Aware Compression for Federated Learning Through Numerical Mechanism Design

Chuan Guo [1]   **Kamalika Chaudhuri** [1]   **Pierre Stock** [1]   **Mike Rabbat** [1]

Abstract

In private federated learning (FL), a server aggregates differentially private updates from a large number of clients in order to train a machine learning model. The main challenge in this setting is balancing privacy with both classification accuracy of the learnt model as well as the number of bits communicated between the clients and server. Prior work has achieved a good trade-off by designing a privacy-aware compression mechanism, called the minimum variance unbiased (MVU) mechanism, that numerically solves an optimization problem to determine the parameters of the mechanism. This paper builds upon it by introducing a new interpolation procedure in the numerical design process that allows for a far more efficient privacy analysis. The result is the new Interpolated MVU mechanism that is more scalable, has a better privacy-utility trade-off, and provides SOTA results on communication-efficient private FL on a variety of datasets.

1. Introduction

Federated learning (FL; McMahan et al. (2017)) is a distributed solution for machine learning on sensitive data and has been the topic of much recent interest. In federated learning, raw client data resides entirely on the client device, and a server coordinates the collaborative training of a global model through transmissions of model updates. Direct transmission of raw updates (pseudo-gradients) can lead to privacy leaks (Zhu et al., 2019; Geiping et al., 2020; Zhao et al., 2020; Yin et al., 2021; Jeon et al., 2021), and an effective method is to use a rigorous privacy solution such as differential privacy (DP; Dwork et al. (2006)) to transmit privacy-preserving updates. Since federated learning typically involves low-bandwidth clients, a requirement for

success is that the amount of communication between the clients and server should be relatively low. Thus a major challenge in this space is ensuring a good privacy-accuracy-communication trade-off—in other words, balancing the predictive accuracy of the model learnt with the privacy parameter as well as the number of bits transmitted between the client and server.

Previous work in this area (Kairouz et al., 2021a; Agarwal et al., 2021; Triastcyn et al., 2021; Chen et al., 2022b) abstracted this problem as privacy-aware compression for distributed mean estimation (DME). Here, the goal is to design a DP mechanism whose output recovers the input in expectation (*i.e.*, is unbiased) and can be represented using a small number of bits. In particular, recent work (Chaudhuri et al., 2022) introduces the *minimum variance unbiased* (MVU) mechanism, which solves a numerical optimization problem to find the parameters of a mechanism that ensures unbiasedness, privacy and communication efficiency while minimizing the variance of the output. They also experimentally demonstrate that, for a given bit budget, this can lead to better privacy-utility trade-offs than many other methods that independently privatize and then compress.

Unfortunately, a problem with the MVU mechanism (and many other privacy-aware compression mechanisms such as Kairouz et al. (2021a); Agarwal et al. (2021)) is that it does not scale well to high-dimensional vectors. This is a crucial drawback for its application to private FL since model updates are often high-dimensional. A major underlying reason for this drawback is the MVU mechanism's insistence on unbiasedness. This is achieved using *randomized dithering* to ensure that the mechanism is unbiased for all continuous-valued inputs, but leads to inefficiencies in the mechanism's computation and privacy accounting for high-dimensional vectors. Importantly, this drawback is in fact superficial since unbiasedness is not required for FL, and prior work has shown that even crude compressors such as SignSGD (Jin et al., 2020) can perform remarkably well.

In this work, we propose the *interpolated MVU* (I-MVU) mechanism, which extends MVU using an interpolation procedure that relaxes the unbiasedness assumption. By its discrete nature, the MVU mechanism can be viewed as sampling from a particular categorical distribution, and

hence can be expressed in exponential family form. The proposed I-MVU mechanism handles continuous-valued inputs by *interpolating the natural exponential family parameters*, rather than directly interpolating the probabilities as in dithering. We introduce a new analysis technique, and by further exploiting special properties of the exponential family, we obtain a tight *numerical* privacy analysis for the vector extension under both L_1 and L_2 geometry.

Experimentally, we find that under both client-level and sample-level DP settings and across various benchmark datasets, the I-MVU mechanism provides a better privacy-utility trade-off than SignSGD (Jin et al., 2020) and MVU (Chaudhuri et al., 2022) at an extremely low communication budget of *one bit per gradient dimension*. Moreover, I-MVU achieves close to the same performance as the standard non-compressed Laplace and Gaussian mechanisms (Abadi et al., 2016) for similar levels of (ϵ, δ)-DP, leading to new state-of-the-art results for private communication-efficient FL.

2. Background

2.1. Differential Privacy

Differential privacy (Dwork et al., 2006) is a cryptographically-motivated definition of privacy that has emerged as the gold standard in privacy-preserving data analysis, and provides rigorous guarantees of privacy leakage induced by a mechanism $\mathcal{M}$.

Definition 1 (Differential privacy). *We say that a mechanism $\mathcal{M}$ that applies to datasets $\mathcal{D}$ is (ϵ, δ)-differentially private, denoted (ϵ, δ)-DP, if for any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ[1] in a single individual's private value, and any output set O, we have:*

$$\mathbb{P}(\mathcal{M}(\mathcal{D}) \in O) \leq e^\epsilon \mathbb{P}(\mathcal{M}(\mathcal{D}') \in O) + \delta.$$

More generally, the framework of DP seeks to bound the difference in distribution between $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D}')$ when $\mathcal{D}$ and $\mathcal{D}'$ differ by a single record that corresponds to one person's private value. Enforcing this property ensures that a single person's data does not substantially change the probability of any event that comes out of using the output of the mechanism $\mathcal{M}$.

Rényi DP and composition. A useful variant of DP is Rényi differential privacy (RDP) (Mironov, 2017), which instead bounds the Rényi divergence (Rényi et al., 1961) between $\mathcal{M}(\mathcal{D})$ and $\mathcal{M}(\mathcal{D}')$ by some ϵ for neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ by a single person's private value. Formally, we say that $\mathcal{M}$ is (α, ϵ)-RDP if:

$$\max(D_\alpha(\mathcal{M}(\mathcal{D}) \,||\, \mathcal{M}(\mathcal{D}')), D_\alpha(\mathcal{M}(\mathcal{D}') \,||\, \mathcal{M}(\mathcal{D}))) \leq \epsilon,$$

[1]We adopt the leave-one-out neighboring notion.

where D_α denotes the order-α Rényi divergence.

An important property of Rényi DP is that it supports easy "sequential composition" of mechanisms—in other words, it makes it relatively simple to measure how the privacy guarantee decays as we release the outputs of multiple differentially private mechanisms on the same dataset $\mathcal{D}$. Specifically, we have that if $\mathcal{M}_1, \ldots, \mathcal{M}_T$ are mechanisms with $\mathcal{M}_t$ being (α, ϵ_t)-RDP for $t = 1, \ldots, T$, then the composition of the T mechanisms is $(\alpha, \sum_{t=1}^{T} \epsilon_t)$-RDP. Another useful property of RDP is its conversion to (ϵ, δ)-DP (Balle et al., 2020): If $\mathcal{M}$ is $(\alpha, \epsilon_\alpha)$-RDP for $\alpha > 1$ then it is also (ϵ, δ)-DP for any $0 < \delta < 1$ with

$$\epsilon = \epsilon_\alpha + \log\left(\frac{\alpha - 1}{\alpha}\right) - \frac{\log \delta + \log \alpha}{\alpha - 1}. \qquad (1)$$

2.2. Federated Learning with Differential Privacy

Federated learning (FL) (McMahan et al., 2017; Kairouz et al., 2021b) allows distributed training of ML models across multiple clients without centralized data storage. A server coordinates training by acquiring model updates from clients, aggregating them, and then transmitting an updated model back to the clients, and the process repeats until convergence. One promise of FL is data privacy since the updates are computed locally on each client using their own data, and hence no client data is ever explicitly transmitted to the server (or anyone else) throughout training.

In spite of this, a recent line of work showed that it is possible to reconstruct training samples from the model updates in a process called *gradient inversion* (Zhu et al., 2019; Geiping et al., 2020; Zhao et al., 2020). To truly preserve the privacy of training data in FL, differential privacy (DP; Dwork et al. (2006) can be applied to the model updates (Geyer et al., 2017) to provide provable guarantees against data reconstruction from the its output (Balle et al., 2022; Guo et al., 2022; Stock et al., 2022). To apply DP to FL training, given a DP mechanism $\mathcal{M}$ and a client update $\mathbf{x}$, the client instead sends $\mathcal{M}(\mathbf{x})$ to the server. For a given round, the client's privacy leakage can be computed in terms of *local DP* if the privatized update $\mathcal{M}(\mathbf{x})$ is revealed to the server, or *global DP* if secure aggregation (Bonawitz et al., 2017) is applied to aggregate the privatized updates before revealing it to the server. The total privacy leakage throughout training can then be computed via RDP composition and conversion to (ϵ, δ)-DP via Equation 1.

2.3. MVU Mechanism

Privacy-aware compression. Since model updates in FL are high-dimensional vectors of size equal to the number of model parameters, it is also important in practice to compress these updates for communication efficiency. Prior work studied this *privacy-aware compression* problem un-

der the abstraction of distributed mean estimation (DME), where clients aim to compute the mean of a set of vectors in a distributed fashion under communication constraint. Formally stated, consider the problem of transmitting a vector $\mathbf{x} \in \mathbb{R}^d$ differentially privately using at most bd bits, with $b \geq 1$ small enough so that the entire vector $\mathbf{x}$ can be transmitted efficiently. One can reduce this problem to a scalar one by considering how to privately compress $x \in [0, 1]$ using at most b bits, and then scaling the vector $\mathbf{x}$ appropriately and applying the scalar mechanism coordinate-wise.

Scalar MVU. The *minimum variance unbiased* (MVU) mechanism (Chaudhuri et al., 2022) solves the privacy-aware compression problem by first discretizing the interval $[0, 1]$ into B_{in} points $\mathcal{X} = \{x_1 = 0, x_2, \ldots, x_{B_{\text{in}}} = 1\}$ with $x_i := (i - 1)/(B_{\text{in}} - 1)$. If $x = x_i$, the mechanism samples $j \sim \text{Categorical}(\mathbf{p}_i)$ using a probability vector $\mathbf{p}_i \in \Delta^{B_{\text{out}}-1}$ and outputs $\mathcal{M}(x) = a_j \in \mathbb{R}$ where $\{a_1, \ldots, a_{B_{\text{out}}}\}$ is a pre-determined output alphabet. The probability vectors $\mathbf{p}_1, \ldots, \mathbf{p}_{B_{\text{in}}}$ and output alphabet $\{a_1, \ldots, a_{B_{\text{out}}}\}$ are numerically designed so that the mechanism satisfies the following three properties:

1. *ϵ-differential privacy*: $e^{-\epsilon} \mathbf{p}_{i',j} \leq \mathbf{p}_{i,j} \leq e^{\epsilon} \mathbf{p}_{i',j}$ for all $i \neq i'$ and all j.
2. *Unbiasedness*: $\sum_{j=1}^{B_{\text{out}}} a_j \mathbf{p}_{i,j} = x_i$ for all i.
3. *Minimum variance*: $\sum_{i=1}^{B_{\text{in}}} \text{Var}(\mathcal{M}(x_i))$ is minimal among all mechanisms satisfying 1 and 2.

The MVU mechanism can then be applied to all $x \in [0, 1]$ by *randomly dithering* x to the nearest x_i and x_{i+1} such that the dithering is unbiased in expectation, *i.e.*,

$$\mathcal{M}(x) \sim \begin{cases} \text{Categorical}(\mathbf{p}_i) & \text{w.p. } (x_{i+1} - x)/\Delta, \\ \text{Categorical}(\mathbf{p}_{i+1}) & \text{w.p. } (x - x_i)/\Delta, \end{cases}$$

where $\Delta = 1/(B_{\text{in}} - 1)$. One can also view this dithering procedure as linearly interpolating between $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$: $\mathcal{M}(x) \sim \text{Categorical}(\mathbf{p}(x))$ with

$$\mathbf{p}(x) = \left(\frac{x_{i+1} - x}{\Delta} \right) \mathbf{p}_i + \left(\frac{x - x_i}{\Delta} \right) \mathbf{p}_{i+1}. \quad (2)$$

It is straightforward to generalize the mechanism to any bounded x by scaling it to $[0, 1]$ and then applying the MVU mechanism.

Vector MVU. For a d-dimensional vector $\mathbf{x}$, the MVU mechanism can be applied independently to each coordinate and the privacy cost is $d\epsilon$ by composition if $\mathbf{x} \in [0, 1]^d$ (or in general, if $\|\mathbf{x}\|_\infty$ is bounded). The conversion from scalar to vector MVU for privatizing FL updates is more complex since DP training in FL commonly require bounds on $\|\mathbf{x}\|_1$ or $\|\mathbf{x}\|_2$ instead of $\|\mathbf{x}\|_\infty$ (Geyer et al., 2017). To handle these settings, Chaudhuri et al. (2022) proposed an

alternative design that replaced property 1 with a metric-DP (Chatzikokolakis et al., 2013) variant: Given a metric $d(x, x')$, the mechanism satisfies

$$\epsilon\text{-metric-DP: } e^{-\epsilon d(x_i, x_{i'})} \mathbf{p}_{i',j} \leq \mathbf{p}_{i,j} \leq e^{\epsilon d(x_i, x_{i'})} \mathbf{p}_{i',j}$$

for all $i \neq i'$ and all j. Designing MVU with metric-DP property with L_1 (resp. L_2) metric allows it to compose more gracefully across dimensions for L_1-bounded (resp. L_2-bounded) vectors.

3. Interpolated MVU Mechanism

Motivation. At a high level, distributed mean estimation in general and its application to FL have different considerations and requirements.

1. The unbiasedness requirement in DME is often unnecessary for FL, as the global model does not require a truly unbiased gradient in order to achieve high accuracy.
2. Model updates in FL are typically high-dimensional, which creates issues for the randomized dithering operation used in MVU and other privacy-aware compression mechanisms (Kairouz et al., 2021a; Agarwal et al., 2021). For vectors $\mathbf{x} \in \mathbb{R}^d$, dithering increases $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_2$ significantly when d is large, hence rejection sampling is needed to control the norm of $\mathbf{x}$ after dithering. This makes it computationally expensive to apply MVU and causing its privacy accounting to be pessimistic.

Interpolated MVU. To address these issues, we propose a new interpolation scheme for extending the MVU mechanism to continuous-valued vector inputs. Our new scheme relaxes the unbiasedness requirement in MVU and makes it more suitable for FL use cases. Our first insight is that the categorical distribution used in MVU can be expressed in *exponential family form* as follows:

$$\mathbb{P}(j|\boldsymbol{\eta}) = \exp(\mathbf{e}_j^\top \boldsymbol{\eta} - A(\boldsymbol{\eta})), \ A(\boldsymbol{\eta}) = \log \left(\sum_j \exp(\boldsymbol{\eta}_j) \right), \quad (3)$$

where $\mathbf{e}_j$ is the j-th standard basis vector. Note that if $\mathbf{p} \in \Delta^{B_{\text{out}}-1}$ then its natural parameter is $\boldsymbol{\eta} = \log \mathbf{p}$.

Expressing the MVU mechanism in exponential family form gives us an alternative way to interpolate the MVU mechanism across the grid points $\mathcal{X} = \{x_1, \ldots, x_{B_{\text{in}}}\}$: by interpolating linearly in the *natural parameter* space. In detail, consider first the scalar setting with input $x \in [0, 1]$. Let $\mathbf{p}_1, \ldots, \mathbf{p}_{B_{\text{in}}} \in \Delta^{B_{\text{out}}-1}$ be sampling probability vectors obtained from the MVU mechanism and let $\boldsymbol{\eta}_i = \log \mathbf{p}_i$ be the natural parameters. Given $x \in [0, 1]$, say $x \in [x_i, x_{i+1}]$ for some i, the *interpolated MVU* (I-MVU) mechanism samples $j \sim \mathbb{P}(\cdot|\boldsymbol{\eta}(x))$ according to Equation 3 and outputs a_j from the MVU output alphabet, where

$$\boldsymbol{\eta}(x) = \left(\frac{x_{i+1} - x}{\Delta} \right) \boldsymbol{\eta}_i + \left(\frac{x - x_i}{\Delta} \right) \boldsymbol{\eta}_{i+1} \quad (4)$$
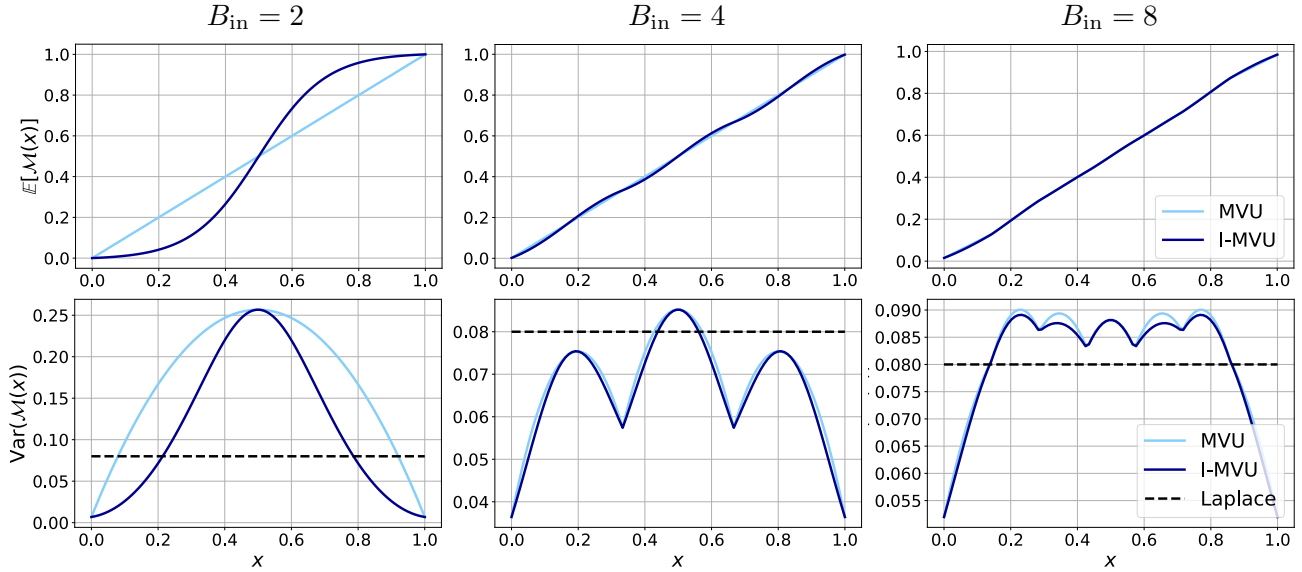
Figure 1: Comparison of MVU and I-MVU with $b = 3$ and L_1-metric-DP $\epsilon = 5$. (**Top**) Plot of the expected value of MVU and I-MVU mechanisms for a scalar input $x \in [0, 1]$. For $B_{\text{in}} = 2$, the expected value of MVU is close to a diagonal line with zero bias, but the I-MVU mechanism incurs a significant bias due to interpolating in the natural parameter space. Increasing the input grid size to $B_{\text{in}} = 4, 8$ reduces the bias of I-MVU drastically to near zero. (**Bottom**) Plot of the variance of MVU and I-MVU in comparison to that of the Laplace mechanism with equal ϵ. Both mechanisms achieve comparable variance to the Laplace mechanism when B_{in} is large enough.

and $\Delta = 1/(B_{\text{in}} - 1)$. In other words, instead of linearly interpolating between $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$ as in Equation 2 for the MVU mechanism, we linearly interpolate between the natural parameters $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_{i+1}$.

Controlling bias. For MVU, because the mechanism is designed so that it is unbiased at all grid points in $\mathcal{X}$, it is easy to show via Equation 2 and linearity of expectation that it is unbiased for all $x \in [0, 1]$. This is not true for I-MVU, and interpolating in the natural parameter space incurs some bias when $x \notin \mathcal{X}$. However, by increasing the number of grid points B_{in}, we can reduce this bias to arbitrarily small with very few drawbacks.

Figure 1 shows the expectation (top) and variance (bottom) of MVU and I-MVU for different values of $B_{\text{in}} = 2, 4, 8$. The mechanisms are constructed to output $b = 3$ bits with target L_1-metric-DP $\epsilon = 5$. For $B_{\text{in}} = 2$, we see that $\mathbb{E}[\mathcal{M}(x)] = x$ for all $x \in [0, 1]$ for MVU, while the same does not hold for I-MVU. As we increase the number of grid points to $B_{\text{in}} = 4, 8$, the expectation of I-MVU more closely matches that of MVU and the bias tends towards zero. In the bottom plots we show the variance of MVU and I-MVU for different $x \in [0, 1]$ and compare it with that of the Laplace mechanism for the same value of $\epsilon = 5$. Notably, for $B_{\text{in}} = 4, 8$, MVU and I-MVU attain comparable variance to that of Laplace while communicating only $b = 3$ bits.

3.1. Extensions

Vector I-MVU. We can extend scalar I-MVU to vector I-MVU by applying the scalar mechanism independently across the vector coordinates. As we will show in subsection 3.2, by utilizing special properties of exponential family distributions, we can more tightly analyze the privacy cost of I-MVU for L_1- and L_2-bounded vectors. In addition, because I-MVU does not need to use rejection sampling to control the gradient norm after randomized dithering, privatizing the input simply amounts to sampling a single categorical random variable per coordinate and is much more computationally efficient.

Input scaling. One way to extend I-MVU to arbitrary bounded ranges is to first scale the input to $[0, 1]$ and then apply the mechanism as usual. However, note that the interpolation scheme in Equation 4 is in fact well-defined for any $x \in \mathbb{R}$, and hence the scaled input does not need to be strictly in the range $[0, 1]$. We leverage this property by introducing a scaling factor $\beta \geq 0$ as follows. For $u \in [-C, C]$, the β-scaled I-MVU mechanism is defined as:

$$\mathcal{M}_\beta(u) = \mathcal{M}\left(\frac{1}{2} + \frac{\beta u}{2C}\right),$$

where $\mathcal{M}$ is the plain I-MVU mechanism. Note that this scaling effectively ensures that the input to $\mathcal{M}$ is in the range $[(1 - \beta)/2, (1 + \beta)/2]$, with $\beta = 1$ corresponding to

scaling the input to $[0, 1]$. For vectors $\mathbf{u}$ with $\|\mathbf{u}\|_2 \leq C$, the β-scaled input $\mathbf{x} = \frac{1}{2} + \frac{\beta \mathbf{u}}{2C}$ satisfies $\|\mathbf{x}\|_2 \leq \beta/2$.

One advantage for using β-scaling is that if the distribution of u is highly concentrated near zero, then scaling with $\beta > 1$ ensures that the input to $\mathcal{M}$ is more spread out in the range $[0, 1]$. Because the MVU mechanism is designed to minimize the average variance across the range $[0, 1]$ (property 3), doing so allows us to better utilize its minimum variance property. For L_1- and L_2-bounded vectors $\mathbf{u}$ this is especially true, where the distribution of coordinates of $\mathbf{u}$ is likely concentrated near zero, hence β-scaling with a large β is essential for achieving good performance.

3.2. Privacy Analysis

We analyze privacy leakage of the I-MVU mechanism in terms of both pure DP and Rényi DP (Mironov, 2017). We first present a general analysis for the I-MVU mechanism with any $B_{\text{in}} \geq 1$ applied to L_1-bounded vectors in Theorem 1. Then we give a more specialized analysis for I-MVU with $B_{\text{in}} = 2$ applied to L_2-bounded vectors in Theorem 2. Proofs of theorems and lemmas are given in Appendix A.

L_1**-bounded vector analysis.** Our strategy is to first analyze the scalar mechanism and express its max divergence for two differing inputs x and x' as a function of $|x' - x|$ (Lemma 1). Then, by independently applying the mechanism across coordinates, we can sum the max divergence across coordinates and upper bound the total DP ϵ as a function of $\|\mathbf{x}' - \mathbf{x}\|_1$ (Theorem 1).

Lemma 1. *Suppose* $x, x' \in [0, 1]$ *and the MVU mechanism satisfies* ϵ L_1*-metric-DP. Let* $\epsilon' = (B_{in} - 1) \times (\max_i \max_{\xi \in [x_i, x_{i+1}]} |\sigma(\boldsymbol{\eta}(\xi))^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i)|)$. *Then:*

$$D_\infty(\mathbb{P}(\cdot|\boldsymbol{\eta}(x)) \,||\, \mathbb{P}(\cdot|\boldsymbol{\eta}(x'))) \leq (\epsilon + \epsilon')|x' - x|.$$

Theorem 1. *Let* ϵ' *be the constant from Lemma 1. Suppose the MVU mechanism is* ϵ L_1*-metric-DP and that* $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ *satisfy* $\|\mathbf{x}' - \mathbf{x}\|_1 \leq C$, *then the I-MVU mechanism is* $(\epsilon + \epsilon')C$*-DP.*

The constant ϵ' can be computed numerically by maximizing the function $h(\xi) = |\sigma(\boldsymbol{\eta}(\xi))^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i)|$ over the range $\xi \in [x_i, x_{i+1}]$ for every i.

L_2**-bounded vector analysis.** Our privacy analysis for L_2-bounded vectors follows a similar strategy, but instead depends on a measure of information known as *Fisher information*, which we define below for completeness. Intuitively, Fisher information measures the curvature of the Rényi divergence, *i.e.*, the rate of change of divergence with respect to input x.

Definition 2. *Let* f *be the density function of a distribution parameterized by* $x \in \mathbb{R}$. *The Fisher information of* x

contained in a sample $Z \sim f(\cdot|x)$ is:

$$\mathcal{I}_Z(x) := \mathbb{E}_Z \left[\left(\frac{d}{dx} \log f(Z|x) \right)^2 \right]. \tag{5}$$

In our setting, the distribution $\mathbb{P}(\cdot|\boldsymbol{\eta}(x))$ is defined by the private data x, and Fisher information measures how much information is revealed about x through a sample $j \sim \mathbb{P}(\cdot|\boldsymbol{\eta}(x))$. It is noteworthy that such a reasoning has also been used to define Fisher information as a privacy metric (Hannun et al., 2021).

For Fisher information to be well-defined, it is necessary that the distribution has a differentiable log density function. For the I-MVU interpolation scheme this is only true if $B_{\text{in}} = 2$; otherwise, the log density may be non-differentiable when transitioning from an interval $[x_{i-1}, x_i]$ to $[x_i, x_{i+1}]$. Thus, Lemma 2 and Theorem 2 below are specialized to $B_{\text{in}} = 2$, which we find is sufficient for most FL applications with L_2-bounded vectors.

Lemma 2. *Let* $B_{in} = 2$ *and let* $M = \sup_{x \in \mathbb{R}} \mathcal{I}_Z(x)$ *be an upper bound on the Fisher information of the mechanism* $\mathcal{M}$. *Then for any* $x_1, x_2 \in \mathbb{R}$:

$$D_\alpha(\mathbb{P}(\cdot|\boldsymbol{\eta}(x_1)) \,||\, \mathbb{P}(\cdot|\boldsymbol{\eta}(x_2))) \leq \alpha M (x_2 - x_1)^2 / 2. \tag{6}$$

Applying Lemma 2 in a similar manner as in the L_1-bounded case by summing the Rényi divergence across coordinates as a function of $\|\mathbf{x}' - \mathbf{x}\|_2^2$ yields Theorem 2.

Theorem 2. *Let* $B_{in} = 2$ *and let* M *be the Fisher information constant from Lemma 2. Suppose that* $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ *satisfy* $\|\mathbf{x}_2 - \mathbf{x}_1\|_2 \leq C$. *Then the I-MVU mechanism is* $(\alpha, \alpha M C^2 / 2)$*-RDP for all* $\alpha > 1$.

Proof sketch. We first derive the Taylor series expression for the Rényi divergence between $\mathbb{P}(\cdot|\boldsymbol{\eta}(x_1))$ and $\mathbb{P}(\cdot|\boldsymbol{\eta}(x_2))$. Since Rényi divergence is minimized and is equal to 0 when $x_1 = x_2$, the zeroth-order and first-order terms in the Taylor series are 0. The coefficient for the second-order term is given by the Fisher information $\mathcal{I}_Z(x_1)$ (Haussler & Opper, 1997), and thus we give a numerical method to compute $M = \sup_{x \in \mathbb{R}} \mathcal{I}_Z(x)$ in Appendix B and use it in Equation 6 to bound the RDP ϵ.

4. Experiments

We evaluate the I-MVU mechanism for federated learning under the local DP setting, *i.e.*, clients transmit the privately compressed model update $\mathcal{M}(\mathbf{x})$ to the server *before* aggregation. We consider two different notions of adjacency for DP: *client-level*, where adjacency is defined as removal of any client, and *sample-level*, where adjacency is defined as removal of a single sample at one client.
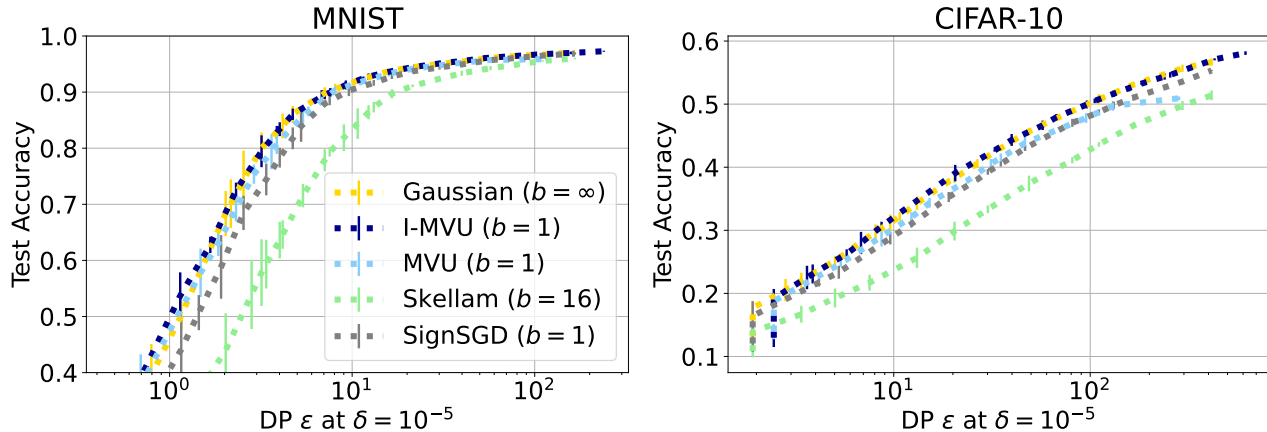
Figure 2: Privacy vs. accuracy plot for the client-level DP setting on MNIST (left) and CIFAR-10 (right) for L_2-bounded client updates. Each line shows the best privacy-utility curve across a grid of hyperparameters, and vertical bar shows standard deviation. I-MVU consistently outperforms all other baselines and achieves the same privacy-utility trade-off as the non-compressed Gaussian mechanism while requiring only one bit communication per parameter.
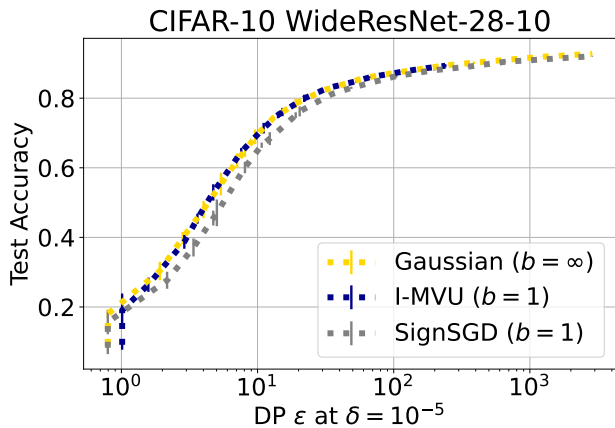

Figure 3: Privacy vs. accuracy plot for the client-level DP setting on CIFAR finetuning with an ImageNet-pretrained WideResNet-28-10 model. Each line shows the best privacy-utility curve across a grid of hyperparameters, and vertical bar shows standard deviation.

Baselines. We compare I-MVU with several privacy-aware compression mechanisms that have been proposed for private FL training:

1. *MVU mechanism* (Chaudhuri et al., 2022) for L_2-bounded client updates.

2. *Skellam mechanism* (Agarwal et al., 2021), which discretizes the input using randomized dithering and adds noise from the Skellam distribution to achieve privacy-aware compression of L_2-bounded client updates.

3. *SignSGD* (Jin et al., 2020) is a strong simple baseline that first applies a non-compressed DP mechanism to

the gradient and then transmits its sign coordinate-wise. Since it operates as a post-processing step on top of a non-compressed DP mechanism, it inherits the privacy guarantee of the non-compressed mechanism while only outputting one bit per parameter. We evaluate SignSGD with both Laplace and Gaussian mechanisms for L_1-bounded and L_2-bounded client updates.

4. *Laplace and Gaussian mechanisms* serve as the non-compressed baselines for L_1-bounded and L_2-bounded client updates. We transmit the private updates using floating point with 32 bits per parameter.

A good privacy-aware compression mechanism should achieve a desirable privacy-utility trade-off (*i.e.*, high test accuracy with low DP ϵ) while transmitting the client updates in a communication-efficient manner. Since our application domain is FL, we did not consider many generic DP distributed mean estimation mechanisms such as PrivUnit (Bhowmick et al., 2018) and Poisson binomial mechanism (Chen et al., 2022b). These mechanisms are designed primarily for privately compressing low-dimensional vectors in an unbiased manner; in contrast, model updates in FL are high-dimensional L_2-bounded vectors and do not need to be unbiased.

4.1. Client-level DP

We first evaluate under the *client-level DP* setting on MNIST and CIFAR-10 (Krizhevsky et al., 2009). Here, the privacy analysis guarantees that the learning algorithm is differentially private with respect to the removal of any client. We divide the training set among the clients with client sample size 1. Each client performs a single local gradient update in every FL round. This setting is equivalent to DP-SGD train-
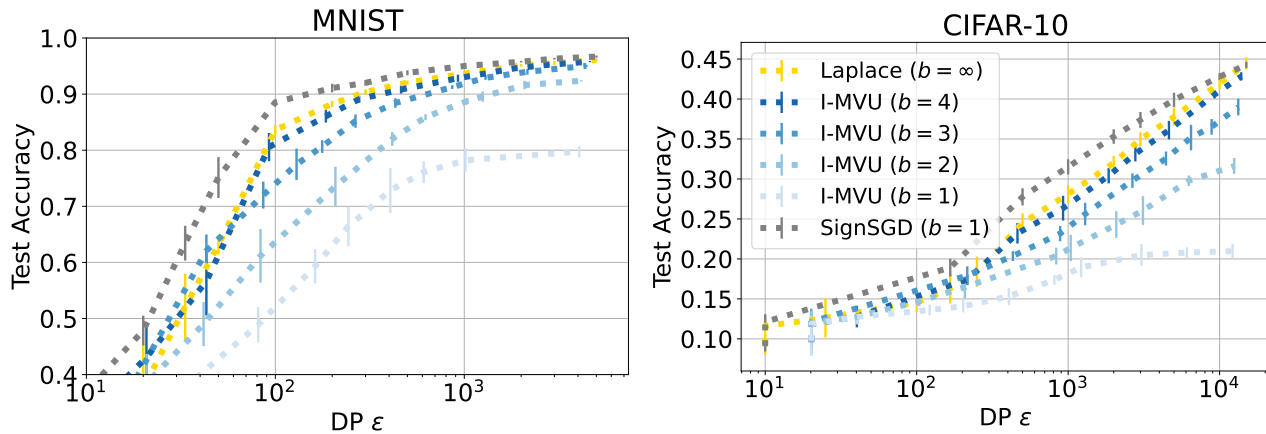
Figure 4: Privacy vs. accuracy plot for the client-level DP setting on MNIST (left) and CIFAR-10 (right) for L_1-bounded client updates. Each line shows the best privacy-utility curve across a grid of hyperparameters, and vertical bar shows standard deviation. I-MVU with $b = 4$ achieves the same privacy-utility trade-off as the non-compressed Laplace mechanism, while decreasing the communication budget b worsens the trade-off. Surprisingly, SignSGD consistently outperforms the non-compressed Laplace baseline, possibly due to the variance reduction (and bias-introducing) effect of sign compression.

ing (Abadi et al., 2016) but with the Gaussian mechanism replaced by a communication-efficient private mechanism.

Training details. Following (Chaudhuri et al., 2022), we train a linear model on top of ScatterNet features (Tramer & Boneh, 2020). This training recipe remains highly competitive under the central DP setting for MNIST and CIFAR-10 without leveraging any public data, hence we adopt it for FL training under local DP. Following Abadi et al. (2016), we apply L_1 and L_2 gradient norm clipping to control the gradient sensitivity and then apply a privacy-aware compression mechanism to transmit the clipped gradient privately. We perform a grid search over hyperparameters such as number of update rounds, step size, gradient norm clip, and mechanism parameters σ (for Gaussian and SignSGD) and ϵ (for MVU and I-MVU); see Appendix C for details.

Result for L_2-bounded client updates. We first consider the more common approach of L_2-norm clipping. Figure 2 shows the privacy vs. test accuracy curve on MNIST (left) and CIFAR-10 (right). Privacy is measured in terms of (ϵ, δ)-DP at $\delta = 10^{-5}$. For each hyperparameter setting, we compute the average test accuracy over five training runs to produce a single ϵ-accuracy pair, and plot the Pareto frontier as a dotted line. The yellow line corresponds to the standard Gaussian mechanism without compression, which attains the best test accuracy at any given privacy budget ϵ. Both SignSGD and MVU are competitive, achieving close to the same level of accuracy as the Gaussian mechanism, but a non-negligible gap remains, especially on CIFAR-10. In contrast, I-MVU attains nearly the same performance as the Gaussian mechanism at all values of ϵ on both MNIST and CIFAR-10. Since MVU and I-MVU are near-identical mechanisms, we argue that the performance gain comes

primarily from the tight privacy analysis for L_2 geometry using Fisher information (Section 3.2).

In Figure 3 we repeat the experiment with an ImageNet pre-trained WideResNet-28-10 (Zagoruyko & Komodakis, 2016) model. The model is pre-trained on down-scaled 32×32 resolution ImageNet (Deng et al., 2009) samples and then finetuned on CIFAR-10 using either the Gaussian mechanism, signSGD or I-MVU. We did not apply the MVU or Skellam mechanisms because the dithering operation is prohibitively expensive on such a large model. We observe a similar result where I-MVU matches the performance of the Gaussian mechanism across the board.

Result for L_1-bounded client updates. For completeness, we also show the result for compressing updates with L_1-norm clipping. This approach is uncommon in private FL training as the noise variance is much larger, but can provide an ϵ-DP guarantee as opposed to (ϵ, δ)-DP with L_2-norm clipping. Figure 4 shows the privacy (in terms of ϵ-DP) vs. test accuracy Pareto frontier curve on MNIST (left) and CIFAR-10 (right). It can be seen that I-MVU with $b = 4$ attains almost the same privacy-utility trade-off as the non-compressed Laplace mechanism, while I-MVU with lower b attain a worse trade-off. Surprisingly, SignSGD outperforms both Laplace and I-MVU with a communication budget of $b = 1$. We suspect this is due to the inherent variance reduction effect of sign compression and that FL training does not require the gradient to be unbiased.

4.2. Sample-level DP

Next, we evaluate under the *sample-level DP* setting on the FEMNIST dataset (Caldas et al., 2018) for classifying written characters into 62 distinct classes. Privacy analy-
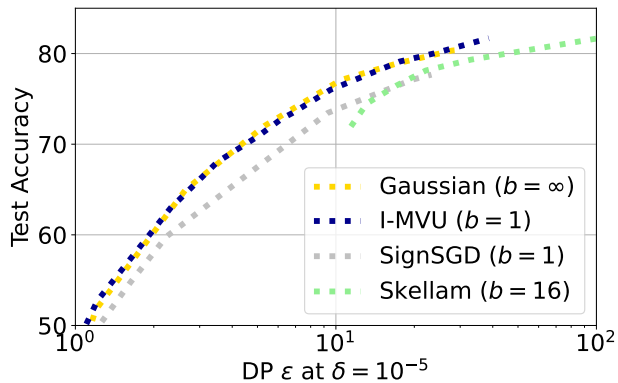
Figure 5: Privacy vs. accuracy plot for the sample-level DP scenario on FEMNIST. I-MVU with one-bit communication budget per coordinate consistently performs better than SignSGD and is competitive with the non-compressed Gaussian baseline across the entire range of ϵ. Finally, I-MVU outperforms the Skellam mechanism while using 16 times less communication bandwidth.

sis guarantees that the learning algorithm is differentially private with respect to the removal of *any training sample from a client*. The dataset has a pre-defined train split with $3,500$ clients, from which we randomly select $3,150$ clients for training and the remaining 350 clients for testing. A set of 5 clients is selected in each training round. Each client performs full batch gradient descent for a single local gradient update to compute the update vector. The update vector is privatized using a communication-efficient private mechanism and transmitted to the server.

Training details. We train a simple 4-layer convolutional network with 46k parameters for classification. The model achieves 84% accuracy when trained non-privately with FL. The client optimizer is SGD with a learning rate of 0.1 and no momentum. The server implements FedAvg (McMahan et al., 2017) with a momentum of 0.9. We perform a grid search on the server learning rate, the clipping factor, the noise multiplier σ for Gaussian and SignSGD, noise scale μ for Skellam, and the L_1-metric DP ϵ and scale hyperparameters for I-MVU. The hyperparameter ranges are given in Tables 5, 4 and 6 in the appendix. In particular, SignSGD uses a much lower server-side learning rates since the updates (in $\{\pm 1\}$) have higher magnitude.

Result. We show the privacy-accuracy trade-off for FEMNIST in Figure 5. The Pareto frontier represents the optimal privacy-accuracy trade-off. The DP privacy budget ϵ is given at $\delta = 10^{-5}$. We observe that I-MVU (blue dashed line) performs better than SignSGD (silver dashed line) for the same communication budget of one bit per update coordinate across the entire range of considered privacy budgets ϵ. Moreover, I-MVU performs on par with the non-compressed Gaussian baseline (yellow dashed line), where clients per-

form local DP-SGD without compressing model updates. I-MVU also outperforms the Skellam mechanism (green dashed line) using 16 times less bandwidth. Additionally, to demonstrate the performance of I-MVU across a wider bit range, we display the performance of the proposed method for various communication budgets in Figure 6.

5. Related Work

Privacy-aware compression. The problem of privacy-aware compression for distributed mean estimation (DME) has received much attention in recent years. Prior work proposed notable mechanisms such as PrivUnit (Bhowmick et al., 2018; Asi et al., 2022), discrete Gaussian (Canonne et al., 2020; Kairouz et al., 2021a), Skellam (Agarwal et al., 2021), MVU (Chaudhuri et al., 2022) and others (Chen et al., 2020; Girgis et al., 2021; Shah et al., 2022; Chen et al., 2022b; Lang & Shlezinger, 2022) for computing the mean of a set of vectors under communication constraint. Feldman & Talwar (2021) provides a general procedure for converting a non-compressed mechanism to a compressed one that is computationally differentially private under the assumption that the pseudo-random generator in the privacy mechanism is hard to break.

There are also other approaches for privacy-aware compression in FL that do not rely on solving the DME problem. For example, Jin et al. (2020) first showed that a simple post-processing of non-compressed mechanisms by taking the sign is a strong baseline that does not attain unbiasedness but works well in practice. Other mechanisms such as DP-REC (Triastcyn et al., 2021) have also leveraged special properties of FL to design more efficient privacy-aware compression schemes for transmitting model updates.

Non-private compression in FL. Previous work has also studied non-private compression in FL, with a focus on characterizing how noise due to (lossy) compression impacts optimization dynamics. Methods have been proposed to only compress up-link (from client to server) messages (Basu et al., 2019; Reisizadeh et al., 2020; Albasyoni et al., 2020; Haddadpour et al., 2021; Mitra et al., 2021; Sadiev et al., 2022), as well as both up- and down-link messages (Philippenko & Dieuleveut, 2020; Condat et al., 2022). Although none of these provides privacy, they do illustrate that federated training is possible with biased compression methods. Note that, in private FL, the downlink messages can be considered non-private, so only the uplink would need to make use of privacy-aware compression.

Mechanism design through optimization. Classical DP mechanisms are mostly derived from simple parameterized distributions and whose privacy is analyzed mathematically. For more specialized applications such as FL where the

requirements are complex, it is often helpful to design custom mechanisms by optimizing a certain objective under DP constraints. Geng et al. (2015) proposed the staircase mechanism and showed that it is variance-optimal in the low-to-medium privacy regime for privatizing L_1-bounded vectors. Bhowmick et al. (2018); Asi et al. (2022) proposed the PrivUnit mechanism and its variant PrivUnitG that are variance-optimal for privatizing unit L_2 vectors under a communication constraint. More similar to our work, Bordenabe et al. (2014) designed an optimal mechanism for location privacy by numerically solving a linear program.

6. Conclusion

We proposed the Interpolated MVU (I-MVU) mechanism that drastically reduces the amount of uplink communication in cross-device FL while providing differential privacy guarantees. I-MVU can be readily applied to DP training of FL models with L_1- and L_2-bounded vectors, with performing matching that of the Laplace and Gaussian mechanisms while communicating as few as one bit per parameter.

Limitations. We discuss several limitations of I-MVU and suggest potential directions for future work.

1. The mechanism currently operates under the local DP setting, which does not leverage the power of secure aggregation to further reduce the privacy cost. Since I-MVU outputs samples from the categorical distribution, their sum forms a so-called Poisson multinomial distribution, whose properties need to be analyzed to understand the privacy cost of I-MVU under the central DP setting.

2. Our work focused on extending MVU through better interpolation and privacy analysis and did not modify the underlying mechanism design problem. Choices such as the set of grid points $\mathcal{X}$, variance objective, and DP constraint can be adapted to better leverage the structure of the new interpolation scheme.

3. Our scalar-to-vector extension of I-MVU can achieve at best a one bit per parameter compression rate. Compressing further to below one bit per parameter may be necessary for applications of FL to large-scale models. Combining I-MVU with recent work on sketching-based compression (Chen et al., 2022a) can be a promising solution.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Agarwal, N., Kairouz, P., and Liu, Z. The skellam mechanism for differentially private federated learning. *Advances in Neural Information Processing Systems*, 34: 5052–5064, 2021.

Albasyoni, A., Safaryan, M., Condat, L., and Richtárik, P. Optimal gradient compression for distributed federated learning. *arXiv preprint arXiv:2010.03246*, 2020.

Asi, H., Feldman, V., and Talwar, K. Optimal algorithms for mean estimation under local differential privacy. *arXiv preprint arXiv:2205.02466*, 2022.

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., and Sato, T. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 2496–2506. PMLR, 2020.

Balle, B., Cherubin, G., and Hayes, J. Reconstructing training data with informed adversaries. *arXiv preprint arXiv:2201.04845*, 2022.

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-Local-SGD: Distributed SGD with quantization, sparsification, and local computations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G., and Rogers, R. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. Optimal geo-indistinguishable mechanisms for location privacy. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 251–262, 2014.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Canonne, C. L., Kamath, G., and Steinke, T. The discrete gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020.

Chatzikokolakis, K., Andrés, M. E., Bordenabe, N. E., and Palamidessi, C. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pp. 82–102. Springer, 2013.

Chaudhuri, K., Guo, C., and Rabbat, M. Privacy-aware compression for federated data analysis. *arXiv preprint arXiv:2203.08134*, 2022.

Chen, W.-N., Kairouz, P., and Ozgur, A. Breaking the communication-privacy-accuracy trilemma. *Advances in Neural Information Processing Systems*, 33:3312–3324, 2020.

Chen, W.-N., Choo, C. A. C., Kairouz, P., and Suresh, A. T. The fundamental price of secure aggregation in differentially private federated learning. In *International Conference on Machine Learning*, pp. 3056–3089. PMLR, 2022a.

Chen, W.-N., Ozgur, A., and Kairouz, P. The poisson binomial mechanism for unbiased federated learning with secure aggregation. In *International Conference on Machine Learning*, pp. 3490–3506. PMLR, 2022b.

Condat, L., Agarský, I., and Richtárik, P. Provably doubly accelerated federated learning: The first theoretically successful combination of local training and compressed communication. *arXiv preprint arXiv:2210.13277*, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Feldman, V. and Talwar, K. Lossless compression of efficient private local randomizers. In *International Conference on Machine Learning*, pp. 3208–3219. PMLR, 2021.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

Geng, Q., Kairouz, P., Oh, S., and Viswanath, P. The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184, 2015.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Girgis, A., Data, D., Diggavi, S., Kairouz, P., and Suresh, A. T. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2521–2529. PMLR, 2021.

Guo, C., Karrer, B., Chaudhuri, K., and van der Maaten, L. Bounding training data reconstruction in private (deep) learning. *arXiv preprint arXiv:2201.12383*, 2022.

Haddadpour, F., Kamani, M. M., Mokhtari, A., and Mahdavi, M. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, volume PMLR 130, pp. 2350–2358, 2021.

Hannun, A., Guo, C., and van der Maaten, L. Measuring data leakage in machine-learning models with fisher information. In *Uncertainty in Artificial Intelligence*, pp. 760–770. PMLR, 2021.

Haussler, D. and Opper, M. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997.

Jeon, J., Lee, K., Oh, S., Ok, J., et al. Gradient inversion with generative image prior. *Advances in Neural Information Processing Systems*, 34:29898–29908, 2021.

Jin, R., Huang, Y., He, X., Dai, H., and Wu, T. Stochastic-sign sgd for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

Kairouz, P., Liu, Z., and Steinke, T. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *International Conference on Machine Learning*, pp. 5201–5212. PMLR, 2021a.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Felix, X. Y., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021b.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lang, N. and Shlezinger, N. Joint privacy enhancement and quantization in federated learning. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2040–2045. IEEE, 2022.

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

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14606–14619, 2021.

Nielsen, F. and Nock, R. On r\'enyi and tsallis entropies and divergences for exponential families. *arXiv preprint arXiv:1105.3259*, 2011.

Philippenko, C. and Dieuleveut, A. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: Tight convergence guarantees. *arXiv preprint arXiv:2006.14591*, 2020.

Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031, 2020.

Rényi, A. et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. Berkeley, California, USA, 1961.

Sadiev, A., Malinovsky, G., Gorbunov, E., Sokolov, I., Khaled, A., Burlachenko, K., and Richtárik, P. Federated optimization algorithms with random reshuffling and gradient compression. *arXiv preprint arXiv:202206.07021*, 2022.

Shah, A., Chen, W.-N., Balle, J., Kairouz, P., and Theis, L. Optimal compression of locally differentially private mechanisms. In *International Conference on Artificial Intelligence and Statistics*, pp. 7680–7723. PMLR, 2022.

Stock, P., Shilov, I., Mironov, I., and Sablayrolles, A. Defending against reconstruction attacks with r\'enyi differential privacy. *arXiv preprint arXiv:2202.07623*, 2022.

Tramer, F. and Boneh, D. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020.

Triastcyn, A., Reisser, M., and Louizos, C. Dp-rec: Private & communication-efficient federated learning. *arXiv preprint arXiv:2111.05454*, 2021.

Yin, H., Mallya, A., Vahdat, A., Alvarez, J. M., Kautz, J., and Molchanov, P. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16337–16346, 2021.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhao, B., Mopuri, K. R., and Bilen, H. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A. Proofs

Lemma 1. *Suppose $x, x' \in [0, 1]$ and the MVU mechanism satisfies ϵ L_1-metric-DP. Let $\epsilon' = (B_{in} - 1) \cdot (\max_i \max_{\xi \in [x_i, x_{i+1}]} |\sigma(\boldsymbol{\eta}(\xi))^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i)|)$. Then:*

$$D_\infty(\mathbb{P}(\cdot|\boldsymbol{\eta}(x)) \,||\, \mathbb{P}(\cdot|\boldsymbol{\eta}(x'))) \le (\epsilon + \epsilon')|x' - x|.$$

Proof. Suppose first that $x, x' \in [x_i, x_{i+1}]$ for some i. Recall that

$$\boldsymbol{\eta}(x) = \left(\frac{x_{i+1} - x}{\Delta}\right) \boldsymbol{\eta}_i + \left(\frac{x - x_i}{\Delta}\right) \boldsymbol{\eta}_{i+1}.$$

Using Equation 3, we get that for any j:

$$\begin{aligned}
|\log \mathbb{P}(j|\boldsymbol{\eta}(x)) - \log \mathbb{P}(j|\boldsymbol{\eta}(x'))| &= \left| \left(\frac{x' - x}{\Delta}\right) \mathbf{e}_j^\top \boldsymbol{\eta}_i + \left(\frac{x - x'}{\Delta}\right) \mathbf{e}_j^\top \boldsymbol{\eta}_{i+1} - A(\boldsymbol{\eta}(x)) + A(\boldsymbol{\eta}(x')) \right| \\
&= \left| \left(\frac{x' - x}{\Delta}\right) \mathbf{e}_j^\top (\boldsymbol{\eta}_i - \boldsymbol{\eta}_{i+1}) - A(\boldsymbol{\eta}(x)) + A(\boldsymbol{\eta}(x')) \right|. \quad (7)
\end{aligned}$$

Since $A'(\boldsymbol{\eta}(x)) = \sigma(\boldsymbol{\eta}(x))$, we can apply Taylor expansion to A to get:

$$A(\boldsymbol{\eta}(x')) = A(\boldsymbol{\eta}(x)) + \left(\frac{x' - x}{\Delta}\right) \sigma(\boldsymbol{\eta}(\xi))^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i),$$

where $\xi \in [x_i, x_{i+1}]$. Substituting into Equation 7 gives:

$$\begin{aligned}
|\log \mathbb{P}(j|\boldsymbol{\eta}(x)) - \log \mathbb{P}(j|\boldsymbol{\eta}(x'))| &= \left| \left(\frac{x' - x}{\Delta}\right) \mathbf{e}_j^\top (\boldsymbol{\eta}_i - \boldsymbol{\eta}_{i+1}) + \left(\frac{x' - x}{\Delta}\right) \sigma(\boldsymbol{\eta}(\xi))^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i) \right| \\
&= \left| \left(\frac{x' - x}{\Delta}\right) (\sigma(\boldsymbol{\eta}(\xi)) - \mathbf{e}_j)^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i) \right| \\
&\le (B_{in} - 1) \cdot \left(|\sigma(\boldsymbol{\eta}(\xi))^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i)| + |\mathbf{e}_j^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i)| \right) |x' - x|.
\end{aligned}$$

Since the MVU mechanism is ϵ L_1-metric-DP, we have $|\mathbf{e}_j^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i)| = |\boldsymbol{\eta}_{i+1,j} - \boldsymbol{\eta}_{i,j}| \le \epsilon|x_{i+1} - x_i| = \epsilon/(B_{in} - 1)$. Thus:

$$|\log \mathbb{P}(j|\boldsymbol{\eta}(x)) - \log \mathbb{P}(j|\boldsymbol{\eta}(x'))| \le \left((B_{in} - 1) \cdot |\sigma(\boldsymbol{\eta}(\xi))^\top (\boldsymbol{\eta}_{i+1} - \boldsymbol{\eta}_i)| + \epsilon \right) |x' - x| \le (\epsilon + \epsilon')|x' - x|,$$

as desired. Suppose now that $x \in [x_i, x_{i+1}]$ and $x' \in [x_k, x_{k+1}]$ with $i \le k$. We can write:

$$\begin{aligned}
\log \mathbb{P}(j|\boldsymbol{\eta}(x)) - \log \mathbb{P}(j|\boldsymbol{\eta}(x')) &\le |\log \mathbb{P}(j|\boldsymbol{\eta}(x)) - \log \mathbb{P}(j|\boldsymbol{\eta}(x_{i+1}))| \\
&\quad + \sum_{l=i+1}^{k-1} |\log \mathbb{P}(j|\boldsymbol{\eta}(x_l)) - \log \mathbb{P}(j|\boldsymbol{\eta}(x_{l+1}))| + |\log \mathbb{P}(j|\boldsymbol{\eta}(x_k)) - \log \mathbb{P}(j|\boldsymbol{\eta}(x'))| \\
&\le (\epsilon + \epsilon') \left(|x - x_{i+1}| + \sum_{l=i+1}^{k-1} |x_l - x_{l+1}| + |x_k - x'| \right) \\
&= (\epsilon + \epsilon')|x - x'|.
\end{aligned}$$

A similar argument applies when $i > k$. $\qquad\square$

Theorem 1. *Let ϵ' be the constant from Lemma 1. Suppose the MVU mechanism is ϵ L_1-metric-DP and that $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ satisfy $\|\mathbf{x}' - \mathbf{x}\|_1 \le C$, then the I-MVU mechanism is $(\epsilon + \epsilon')C$-DP.*

Proof. Let $\mathbf{a} = \mathcal{M}(\mathbf{x}) \in \{a_1, \ldots, a_{B_{\text{out}}}\}^d$ be the output of the vector I-MVU mechanism that independently applies the scalar mechanism to each coordinate. Then:

$$
\begin{aligned}
D_\infty(\mathcal{M}(\mathbf{x}) \,\|\, \mathcal{M}(\mathbf{x}')) &= \max_{\mathbf{a} \in \{a_1, \ldots, a_{B_{\text{out}}}\}^d} \left| \sum_{k=1}^{d} \log \mathbb{P}(\mathbf{a}_k | \boldsymbol{\eta}(\mathbf{x}'_k)) - \sum_{k=1}^{d} \log \mathbb{P}(\mathbf{a}_k | \boldsymbol{\eta}(\mathbf{x}_k)) \right| \\
&\leq \sum_{k=1}^{d} \max_{\mathbf{a}_k \in \{a_1, \ldots, a_{B_{\text{out}}}\}} |\log \mathbb{P}(\mathbf{a}_k | \boldsymbol{\eta}(\mathbf{x}'_k)) - \log \mathbb{P}(\mathbf{a}_k | \boldsymbol{\eta}(\mathbf{x}_k))| \\
&= \sum_{k=1}^{d} D_\infty(\mathbb{P}(\cdot | \boldsymbol{\eta}(\mathbf{x}_k)) \,\|\, \mathbb{P}(\cdot | \boldsymbol{\eta}(\mathbf{x}'_k))) \\
&\leq \sum_{k=1}^{d} \tilde{\Delta}(\boldsymbol{\eta}) |\mathbf{x}'_k - \mathbf{x}_k| \quad \text{by Lemma 1} \\
&= \tilde{\Delta}(\boldsymbol{\eta}) \|\mathbf{x}' - \mathbf{x}\|_1.
\end{aligned}
$$

$\square$

Lemma 2. *Let $B_{in} = 2$ and let $M = \sup_{x \in \mathbb{R}} \mathcal{I}_Z(x)$ be an upper bound on the Fisher information of the mechanism $\mathcal{M}$. Then for any $x, x' \in \mathbb{R}$:*

$$
D_\alpha(\mathbb{P}(\cdot | \boldsymbol{\eta}(x)) \,\|\, \mathbb{P}(\cdot | \boldsymbol{\eta}(x'))) \leq \alpha M (x' - x)^2 / 2.
$$

Proof. We first derive an explicit form for the Fisher information. Let $f(\mathbf{z}; x)$ denote the pmf in Equation 3 for any $\mathbf{z} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_{B_{\text{out}}}\}$. The log pmf is:

$$
\log f(\mathbf{z}; x) = \mathbf{z}^\top \boldsymbol{\eta}(x) - A(\boldsymbol{\eta}(x)) \tag{8}
$$

Taking derivative with respect to x gives:

$$
\begin{aligned}
\frac{d}{dx} \log f(\mathbf{z}; x) &= (\mathbf{z} - \sigma(\boldsymbol{\eta}(x)))^\top (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1), \\
\left(\frac{d}{dx} \log f(\mathbf{z}; x) \right)^2 &= (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)^\top (\mathbf{z} - \sigma(\boldsymbol{\eta}(x)))(\mathbf{z} - \sigma(\boldsymbol{\eta}(x)))^\top (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1),
\end{aligned}
$$

where σ denotes the sigmoid function. Taking expectation over $\mathbf{z}$ gives the Fisher information:

$$
\mathcal{I}_Z(x) = (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)^\top U (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1), \tag{9}
$$

with $U = \text{diag}(\sigma(\boldsymbol{\eta}(x))) - \sigma(\boldsymbol{\eta}(x))\sigma(\boldsymbol{\eta}(x))^\top$.

To derive the upper bound, we first define a function F_α for the Rényi divergence of the mechanism for a fixed x and varying x':

$$
F_\alpha(x'; x) = D_\alpha(\mathbb{P}(\cdot | \boldsymbol{\eta}(x)) \,\|\, \mathbb{P}(\cdot | \boldsymbol{\eta}(x'))). \tag{10}
$$

By Taylor's theorem, we can express F_α as:

$$
F_\alpha(x'; x) = F_\alpha(x; x) + (x' - x) F'_\alpha(x; x) + (x' - x)^2 F''_\alpha(\xi; x)/2
$$

for some $\xi \in [x, x']$. Note that $F_\alpha(x; x) = 0$ and $F'_\alpha(x; x) = 0$ (since x is the global minimizer of $F_\alpha(\cdot; x)$), so F_α is locally a quadratic function:

$$
F_\alpha(x'; x) = (x' - x)^2 F''_\alpha(\xi; x)/2. \tag{11}
$$

Since f is the pmf of an exponential family distribution, we can use the closed form expression (Nielsen & Nock, 2011) for Rényi divergence of exponential family distributions to express F_α and its derivatives:

$$
\begin{aligned}
F_\alpha(\xi; x) &= \frac{1}{\alpha - 1} [A(\alpha \boldsymbol{\eta}(x) + (1 - \alpha)\boldsymbol{\eta}(\xi)) - \alpha A(\boldsymbol{\eta}(x)) - (1 - \alpha) A(\boldsymbol{\eta}(\xi))] \\
F'_\alpha(\xi; x) &= (\sigma(\boldsymbol{\eta}(\xi)) - \sigma(\alpha \boldsymbol{\eta}(x) + (1 - \alpha)\boldsymbol{\eta}(\xi)))^\top (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1) \\
F''_\alpha(\xi; x) &= (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)^\top (V + (\alpha - 1)W)(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)
\end{aligned}
$$

13

where $V = \text{diag}(\sigma(\boldsymbol{\eta}(\xi))) - \sigma(\boldsymbol{\eta}(\xi))\sigma(\boldsymbol{\eta}(\xi))^\top$, $W = \text{diag}(\sigma(\boldsymbol{\eta}(x''))) - \sigma(\boldsymbol{\eta}(x''))\sigma(\boldsymbol{\eta}(x''))^\top$, and $x'' = \alpha x + (1 - \alpha)\xi$. Hence $F_\alpha''(\xi; x) = \mathcal{I}_Z(\xi) + (\alpha - 1)\mathcal{I}_Z(x'')$ by Equation 9. Upper bounding $\mathcal{I}_Z(\xi)$ and $\mathcal{I}_Z(x'')$ by $M := \sup_{x \in \mathbb{R}} \mathcal{I}_Z(x)$ and combining with Equation 11 gives the desired result. $\qquad\square$

Theorem 2. *Let $B_{in} = 2$ and let M be the Fisher information constant from Lemma 2. Suppose that $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ satisfy $\|\mathbf{x}' - \mathbf{x}\|_2 \leq C$, then the I-MVU mechanism is $(\alpha, \alpha M C^2/2)$-RDP for all $\alpha > 1$.*

Proof. Let $\mathbf{a} = \mathcal{M}(\mathbf{x}) \in \{a_1, \ldots, a_{B_{out}}\}^d$ be the output of the vector I-MVU mechanism that independently applies the scalar mechanism to each coordinate. Then:

$$
\begin{aligned}
D_\alpha(\mathcal{M}(\mathbf{x}) \| \mathcal{M}(\mathbf{x}')) &= \frac{1}{\alpha - 1} \log \sum_{\mathbf{a} \in \{a_1, \ldots, a_{B_{out}}\}^d} \prod_{k=1}^d \frac{\mathbb{P}(\mathbf{a}_k | \boldsymbol{\eta}(\mathbf{x}_k'))^\alpha}{\mathbb{P}(\mathbf{a}_k | \boldsymbol{\eta}(\mathbf{x}_k))^{\alpha - 1}} \\
&= \sum_{k=1}^d \frac{1}{\alpha - 1} \log \sum_{\mathbf{a}_k \in \{a_1, \ldots, a_{B_{out}}\}} \frac{\mathbb{P}(\mathbf{a}_k | \boldsymbol{\eta}(\mathbf{x}_k'))^\alpha}{\mathbb{P}(\mathbf{a}_k | \boldsymbol{\eta}(\mathbf{x}_k))^{\alpha - 1}} \quad \text{by independence} \\
&= \sum_{k=1}^d D_\alpha(\mathbb{P}(\cdot | \boldsymbol{\eta}(\mathbf{x}_k)) \| \mathbb{P}(\cdot | \boldsymbol{\eta}(\mathbf{x}_k'))) \\
&\leq \sum_{k=1}^d \alpha M(\mathbf{x}_k' - \mathbf{x}_k)^2/2 \quad \text{by Lemma 2} \\
&= \alpha M \|\mathbf{x}' - \mathbf{x}\|_2^2/2.
\end{aligned}
$$

$\qquad\square$

B. Computing Fisher Information

In this section we describe a method for computing $M = \sup_{x \in \mathbb{R}} \mathcal{I}_Z(x)$. We first define a condition for $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ that allows us to reduce this problem to maximizing $\mathcal{I}_Z(x)$ over a bounded range $[x_{\min}, x_{\max}]$.

Definition 3. *Two vectors $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathbb{R}^B$ are said to be* anadromic *if for all $j = 1, \ldots, B$, we have $(\boldsymbol{\eta}_1)_j = (\boldsymbol{\eta}_2)_{B-j+1}$.*

The following technical lemma proves several useful properties that hold when $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are anadromic.

Lemma 3. *Suppose that $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathbb{R}^B$ are anadromic. Let $\boldsymbol{\theta} = \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1$ and suppose that $j^+ = \arg\max_j \boldsymbol{\theta}_j, j^- = \arg\min_j \boldsymbol{\theta}_j$ are unique. Then the following hold:*

  (i) $\boldsymbol{\theta}_j = -\boldsymbol{\theta}_{B-j+1}$ for all j, and hence $j^- = B - j^+ + 1$.
  (ii) $\boldsymbol{\eta}(x)_j = \boldsymbol{\eta}(1-x)_{B-j+1}$ for all j.
  (iii) $\sigma(\boldsymbol{\eta}(x)) \to \mathbf{e}_{j^+}$ as $x \to \infty$ and $\sigma(\boldsymbol{\eta}(x)) \to \mathbf{e}_{j^-}$ as $x \to -\infty$.
  (iv) $\mathcal{I}_Z(x) = \mathcal{I}_Z(1-x)$ for all $x \in \mathbb{R}$.
  (v) $x = 1/2$ is a stationary point for $\mathcal{I}_Z(x)$.
  (vi) If $\sigma(\boldsymbol{\eta}(x))_{j^+} \geq 1/2$ then $\mathcal{I}_Z(x) \leq 4\boldsymbol{\theta}_{j^+}^2 \sigma(\boldsymbol{\eta}(x))_{j^+}(1 - \sigma(\boldsymbol{\eta}(x))_{j^+})$.

Proof. (i) Since $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are anadromic,

$$
\boldsymbol{\theta}_j = (\boldsymbol{\eta}_2)_j - (\boldsymbol{\eta}_1)_j = (\boldsymbol{\eta}_2)_j - (\boldsymbol{\eta}_2)_{B-j+1} = -((\boldsymbol{\eta}_2)_{B-j+1} - (\boldsymbol{\eta}_1)_j) = -\boldsymbol{\theta}_{B-j+1}.
$$

In particular, $\arg\max_j \boldsymbol{\theta}_j = B - (\arg\min_j \boldsymbol{\theta}_j) + 1$.

(ii) $\boldsymbol{\eta}(x)_j = (1-x)(\boldsymbol{\eta}_1)_j + x(\boldsymbol{\eta}_2)_j = (1-x)(\boldsymbol{\eta}_1)_{B-j+1} + x(\boldsymbol{\eta}_2)_{B-j+1} = \boldsymbol{\eta}(1-x)_{B-j+1}$.

(iii) Let $\bar{\boldsymbol{\eta}} = (\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)/2$ so that $\boldsymbol{\eta}(x) = \bar{\boldsymbol{\eta}} + (x - 1/2)\boldsymbol{\theta}$ for all $x \in \mathbb{R}$. It is clear that $\sigma(\boldsymbol{\eta}(x)) \to \mathbf{e}_{j^+}$ as $x \to \infty$ since j^+ is unique. A similar argument shows that $\sigma(\boldsymbol{\eta}(x)) \to \mathbf{e}_{j^-}$ as $x \to -\infty$.

(iv) Using the expression of $\mathcal{I}_Z(x)$ in the proof of Lemma 2, we get

$$
\begin{aligned}
\mathcal{I}_Z(x) &= \sum_j \boldsymbol{\theta}_j^2 \sigma(\boldsymbol{\eta}(x))_j - \left(\sum_j \boldsymbol{\theta}_j \sigma(\boldsymbol{\eta}(x))_j \right)^2 \\
&= \sum_j \boldsymbol{\theta}_{B-j+1}^2 \sigma(\boldsymbol{\eta}(1-x))_{B-j+1} - \left(\sum_j \boldsymbol{\theta}_{B-j+1} \sigma(\boldsymbol{\eta}(1-x))_{B-j+1} \right)^2 \quad \text{by (i) and (ii)} \\
&= \mathcal{I}_Z(1-x).
\end{aligned}
$$

(v) Differentiating $\mathcal{I}_Z(x)$ and using the above argument gives:

$$
\begin{aligned}
\mathcal{I}_Z'(x) &= (\boldsymbol{\theta}^3)^\top \sigma(\boldsymbol{\eta}(x)) - 3\boldsymbol{\theta}^\top \sigma(\boldsymbol{\eta}(x))(\boldsymbol{\theta}^2)^\top \sigma(\boldsymbol{\eta}(x)) + 2(\boldsymbol{\theta}^\top \sigma(\boldsymbol{\eta}(x)))^3 \\
&= -(\boldsymbol{\theta}^3)^\top \sigma(\boldsymbol{\eta}(1-x)) + 3\boldsymbol{\theta}^\top \sigma(\boldsymbol{\eta}(1-x))(\boldsymbol{\theta}^2)^\top \sigma(\boldsymbol{\eta}(1-x)) - 2(\boldsymbol{\theta}^\top \sigma(\boldsymbol{\eta}(1-x)))^3 \\
&= -\mathcal{I}_Z'(1-x).
\end{aligned}
$$

Then $\mathcal{I}_Z'(1/2) = -\mathcal{I}_Z'(1/2)$, so $\mathcal{I}_Z'(1/2) = 0$ and $x = 1/2$ is a stationary point.

(vi) Using the fact that $0 \le \boldsymbol{\theta}_j^2 \le \boldsymbol{\theta}_{j+}^2$ for all j and

$$
\sum_j \boldsymbol{\theta}_j \sigma(\boldsymbol{\eta}(x))_j \ge \boldsymbol{\theta}_{j+} \sigma(\boldsymbol{\eta}(x))_{j+} + \boldsymbol{\theta}_{j-}(1 - \sigma(\boldsymbol{\eta}(x))_{j+}) = \boldsymbol{\theta}_{j+} \sigma(\boldsymbol{\eta}(x))_{j+} - \boldsymbol{\theta}_{j+}(1 - \sigma(\boldsymbol{\eta}(x))_{j+}) \ge 0,
$$

we have:

$$
\begin{aligned}
\mathcal{I}_Z(x) &= \sum_j \boldsymbol{\theta}_j^2 \sigma(\boldsymbol{\eta}(x))_j - \left(\sum_j \boldsymbol{\theta}_j \sigma(\boldsymbol{\eta}(x))_j \right)^2 \\
&\le \boldsymbol{\theta}_{j+}^2 - (\boldsymbol{\theta}_{j+} \sigma(\boldsymbol{\eta}(x))_{j+} - \boldsymbol{\theta}_{j+}(1 - \sigma(\boldsymbol{\eta}(x))_{j+}))^2 \\
&= \boldsymbol{\theta}_{j+}^2 (1 - (2\sigma(\boldsymbol{\eta}(x))_{j+} - 1)^2) \\
&= 4\boldsymbol{\theta}_{j+}^2 \sigma(\boldsymbol{\eta}(x))_{j+}(1 - \sigma(\boldsymbol{\eta}(x))_{j+}).
\end{aligned}
$$

$\square$

Algorithm. To use Lemma 3 to compute M, we first compute $I^* = \mathcal{I}_Z(1/2)$ since $x = 1/2$ is a stationary point by Lemma 3(v). By setting

$$
4\boldsymbol{\theta}_{j+}^2 \sigma(\boldsymbol{\eta}(x))_{j+}(1 - \sigma(\boldsymbol{\eta}(x))_{j+}) \le I^*
$$

and solving this quadratic equation for $\sigma(\boldsymbol{\eta}(x))_{j+}$, we can use the bound in Lemma 3(vi) to obtain that $\mathcal{I}_Z(x) \le I^*$ when $\sigma(\boldsymbol{\eta}(x))_{j+} \ge \left(1 + \sqrt{1 - I^*/\boldsymbol{\theta}_{j+}^2}\right)/2 \ge 1/2$. Since $\sigma(\boldsymbol{\eta}(x))_{j+} \to 1$ as $x \to \infty$ by (iii), we can determine the value $x_{\max}$ for which $\mathcal{I}_Z(x) \le I^*$ when $x \ge x_{\max}$. By (iv), $x_{\min} = 1 - x_{\max}$ satisfies $\mathcal{I}_Z(x) \le I^*$ when $x \le x_{\min}$. We can then do line search in $[x_{\min}, x_{\max}]$ (or equivalently, in $[1/2, x_{\max}]$ by Lemma 3(iv)) to obtain M.

C. Choice of Hyperparameters

The privacy vs. accuracy curves in the paper plot the Pareto frontier across a grid of hyperparameters. Table 1 gives the range of hyperparameter values used in Figure 2. Table 2 gives the range of hyperparameters used in Figure 3. Table 3 gives the range of hyperparameter values used in Figure 4. Tables 4, 5 and 6 give the range of hyperparameter values used in Figure 5.

D. Additional Results

The budget parameter b controls how many bits the I-MVU mechanism outputs. We performed ablation study on b to show that for the sample-level DP experiment in Section 4.2, I-MVU with $b = 1$ achieves nearly the same privacy-utility trade-off

Hyperparameter	Values
Batch size / cohort size	600
Momentum	0.5
# Iterations T	$100, 200, 300, 500, 1000, 2000, 3000, 5000$
Step size ρ	$0.001, 0.003, 0.01, 0.03, 0.1, 0.3$
Gradient norm clip C	$0.25, 0.5, 1, 2, 4, 8$
Noise multiplier σ for Gaussian and SignSGD	$0.5, 1, 2, 3, 5$
Noise scale μ for Skellam	$(\sigma C)^2$
L_1-metric DP parameter ϵ for MVU/I-MVU	$0.25, 0.5, 0.75, 1, 2, 3, 5$
β scaling for I-MVU	1

(a) MNIST

Hyperparameter	Values
Batch size / cohort size	500
Momentum	0.5
# Iterations T	$500, 1000, 2000, 3000, 5000, 10000, 15000$
Step size ρ	$0.001, 0.003, 0.01, 0.03, 0.1, 0.3$
Gradient norm clip C	$0.25, 0.5, 1, 2, 4, 8$
Noise multiplier σ for Gaussian and SignSGD	$0.5, 1, 2, 3, 5$
Noise scale μ for Skellam	$(\sigma C)^2$
L_1-metric DP parameter ϵ for MVU/I-MVU	$0.25, 0.5, 0.75, 1, 2, 3, 5$
β scaling for I-MVU	1

(b) CIFAR-10

Table 1: Hyperparameters for L_2-bounded client-level DP training

Hyperparameter	Values
Batch size / cohort size	500
Momentum	0.5
# Iterations T	$100, 200, 300, 500, 1000, 2000, 3000, 5000$
Step size ρ	$0.001, 0.003, 0.01, 0.03, 0.1, 0.3$
Gradient norm clip C	$0.25, 0.5, 1, 2, 4, 8$
Noise multiplier σ for Gaussian and SignSGD	$0.5, 1, 2, 3, 5$
Noise scale μ for Skellam	$(\sigma C)^2$
L_1-metric DP parameter ϵ for MVU/I-MVU	$0.25, 0.5, 0.75, 1, 2, 3, 5$
β scaling for I-MVU	1

Table 2: Hyperparameters for L_2-bounded client-level DP CIFAR-10 fine-tuning

Hyperparameter	Values
Batch size / cohort size	600
Momentum	0.5
# Iterations T	$100, 200, 300, 500, 1000, 2000, 3000, 5000$
Step size ρ	$0.01, 0.03, 0.1, 0.3, 1.0, 3.0$
Gradient norm clip C	$0.25, 0.5, 1, 2, 4, 8$
Noise multiplier μ for Laplace and SignSGD	$0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.3, 0.5$
L_1-metric DP parameter ϵ for I-MVU	$1, 2, 3, 5, 10, 20$
β scaling for I-MVU	8

(a) MNIST

Hyperparameter	Values
Batch size / cohort size	500
Momentum	0.5
# Iterations T	$500, 1000, 2000, 3000, 5000, 10000, 15000$
Step size ρ	$0.01, 0.03, 0.1, 0.3, 1.0, 3.0$
Gradient norm clip C	$0.25, 0.5, 1, 2, 4, 8$
Noise multiplier μ for Laplace and SignSGD	$0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.3, 0.5$
L_1-metric DP parameter ϵ for I-MVU	$1, 2, 3, 5, 10, 20$
β scaling for I-MVU	8

(b) CIFAR-10

Table 3: Hyperparameters for L_1-bounded client-level DP training

as with $b = 2, 3, 4$. Figure 6 shows the global model's test accuracy vs. DP ϵ for the Gaussian mechanism and I-MVU with $b = 1, 2, 3, 4$. Evidently, increasing b from 1 to 4 does not result in a significant increase in test accuracy for a fixed value of ϵ. This can be contrasted with the result in Figure 4, where increasing b from 1 to 4 *does* result in a significant boost in test accuracy.

Hyperparameter	Values
σ for Gaussian	0.6, 0.8, 1, 2, 4, 6, 8, 10, 16, 20, 24, 32, 40, 50, 60, 64, 70, 80, 90, 100, 110, 120, 128
μ for Skellam	$(\sigma C)^2$
Server-side learning rate	0.5, 1, 2
Clipping factor	0.1, 0.5, 1, 2, 10

Table 4: Hyperparameter range for Skellam and Gaussian on FEMNIST.

Hyperparameter	Values
ε	0.25, 0.5, 0.75, 1, 2, 3, 5, 6, 7, 8, 9, 10
Server-side learning rate	0.5, 1, 2
Scaling factor β	32, 64, 128

Table 5: Hyperparameter range for I-MVU on FEMNIST.

Hyperparameter	Values
σ	0.6, 0.8, 1, 2, 4, 6, 8, 10, 16, 32, 64, 128
Server-side learning rate	0.0001, 0.001, 0.01
Clipping factor	0.5, 1, 2

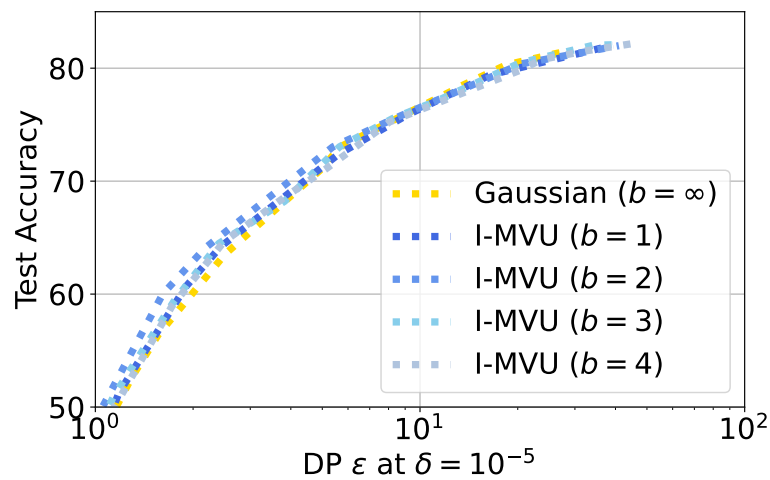Table 6: Hyperparameter range for SignSGD on FEMNIST.

Figure 6: Privacy/accuracy plot for the sample-level DP scenario on FEMNIST. Each point represents a single hyperparameter setting and the Pareto frontier is shown in dashed line. We display the performance of I-MVU with b bits communication budget per coordinate: it is competitive with the non-compressed Gaussian baseline across the entire range of ϵ.