The Poisson Binomial Mechanism for Unbiased Federated Learning with Secure Aggregation

Wei-Ning Chen [1] **Ayfer Özgür** [1] **Peter Kairouz** [2]

Abstract

We introduce the Poisson Binomial mechanism (PBM), a discrete differential privacy mechanism for distributed mean estimation (DME) with applications to federated learning and analytics. We provide a tight analysis of its privacy guarantees, showing that it achieves the same privacy-accuracy trade-offs as the continuous Gaussian mechanism. Our analysis is based on a novel bound on the Rényi divergence of two Poisson binomial distributions that may be of independent interest.

Unlike previous discrete DP schemes based on additive noise, our mechanism encodes local information into a parameter of the binomial distribution, and hence the output distribution is discrete with bounded support. Moreover, the support does not increase as the privacy budget $\varepsilon \to 0$ as in the case of additive schemes which require the addition of more noise to achieve higher privacy; on the contrary, the support becomes smaller as $\varepsilon \to 0$. The bounded support enables us to combine our mechanism with secure aggregation (SecAgg), a multi-party cryptographic protocol, without the need of performing modular clipping which results in an unbiased estimator of the sum of the local vectors. This in turn allows us to apply it in the private FL setting and provide an upper bound on the convergence rate of the SGD algorithm. Moreover, since the support of the output distribution becomes smaller as $\varepsilon \to 0$, the communication cost of our scheme decreases with the privacy constraint ε, outperforming all previous distributed DP schemes based on additive noise in the high privacy or low communication regimes.

1. Introduction

The standard technique for ensuring differential privacy (DP) (Dwork et al., 2006b) of learning algorithms is to add noise either to the output of a function evaluated on the data (in the centralized setting) or locally to the data itself (in federated settings (Kairouz et al., 2019; McMahan et al., 2016)). Two commonly used distributions for noise are the Gaussian and Laplace distributions. While simple enough for mathematical reasoning and analysis, the continuous nature of these distributions presents a number of challenges. First, it is not possible to represent real samples on finite computers, making these mechanisms prone to numerical errors that can break privacy guarantees (Mironov, 2012). Second, they cannot be used in the federated setting where it may be desirable to first locally perturb the data (e.g. the local model update computed by stochastic gradient descent(SGD) iterations) and then use cryptographic primitives such as secure aggregation (SecAgg) (Bonawitz et al., 2016b) to allow the server to obtain a summary of the local data (such as the mean of local model updates) without having access to individual information. This combination of local DP and secure aggregation is desirable as it does not rely on the clients' full trust in the server, while potentially achieving the same utility-privacy trade-off as in the centralized case. However, secure aggregation is based on modular arithmetic which is not compatible with the real output from a privatization mechanism that relies on perturbing data with continuous noise. This has led to an increasing recent interest in mechanisms that perturb the data (or a function of it) with the addition of discrete noise, such as the binomial in (Dwork et al., 2006a; Agarwal et al., 2018), the discrete Gaussian in (Canonne et al., 2020; Kairouz et al., 2021), and Skellam noise in (Agarwal et al., 2021).

These additive discrete noise mechanisms however have a few of their own shortcomings. First, when the data itself is continuous, as in the case of local model updates obtained from SGD iterations in federated learning, it has to be discretized before the addition of discrete noise. This adds quantization noise and complicates analysis. Second, these distributions have discrete yet unbounded support which means that the privatized data has to go through modular clipping when combined with secure aggregation protocols

[1] Department of Electrical Engineering, Stanford University [2] Google Research. Correspondence to: Wei-Ning Chen <wnchen@stanford.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

which operate on a finite group. This is the approach in (Kairouz et al., 2021) and (Agarwal et al., 2021), which focus on developing differentially private federated learning algorithms by using these additive noise mechanisms locally at the clients and then feeding the privatized local updates to the secure aggregation protocol. Upon modular clipping, however, the discrete additive noise becomes no longer zero-mean, and hence the resulting estimator (of the mean of the local model updates) is *biased*. The bias makes it difficult to provide tight convergence guarantees for stochastic first-order optimization methods which rely on this estimate. In contrast, it is usually not difficult to provide a tight convergence analysis when the optimization method has access to a noisy but unbiased estimate of the true mean of the updates. Finally, all additive noise mechanism (continuous or discrete) share the common principle of adding more noise to achieve higher privacy, i.e. they require a higher noise variance when higher privacy is desired. This, however, has a direct impact on the communication cost in federated settings when the noise is added locally. Indeed, for all of the above mentioned schemes (Kairouz et al., 2021; Agarwal et al., 2021; 2018) the communication cost grows *inversely* with the privacy budget; when high privacy is desired nodes need a large bit budget to communicate the large noise they add to their updates. This contradicts the conclusion of (Chen et al., 2020) which shows that in a federated learning setting without secure aggregation the optimal communication cost can be made to *decrease* with the privacy budget; intuitively we can use less bits in the high privacy regime because we are required to communicate less information about the local data.

In this paper, we develop a novel differential privacy mechanism that does not rely on additive noise. This mechanism, which we call the multi-dimensional Poisson Binomial mechanism (PBM), takes a continuous input, encodes it into the parameter p of a binomial distribution $\mathsf{Binom}\left(m,p\right)$, and generates a sample from this distribution[1]. This results in a finite and discrete output in $\mathbb{Z}_m$ which can be easily combined with the integer modular arithmetic in SecAgg, without the need for quantization or modular clipping. As a result, the estimate (of the average model updates) obtained at the output of SecAgg is *unbiased* leading, to our knowledge, to the first unbiased privacy scheme compatible with SecAgg. Moreover, the communication cost of PBM decreases when the privacy budget ε decreases. This is because the first parameter m of the distribution $\mathsf{Binom}\left(m,p\right)$ is linear in the privacy budget ε and hence the logarithm of it, which dictates the communication budget, decreases to 1 as $\varepsilon \to 0$.

Our contributions. Our main technical contributions are summarized as follows.

- We introduce the multi-dimensional Poisson binomial mechanism, an unbiased and bounded discrete DP mechanism for distributed mean estimation (DME). We provide a tight analysis for its Rényi DP (RDP) guarantees showing that it provides the same utility-privacy trade-off as the continuous Gaussian mechanism. As a by-product, our analysis yields a novel bound on the Rényi divergence of two Poisson binomial distributions that can be useful in other applications.
- We show that the communication cost of our scheme (defined as the number of bits needed to achieve the accuracy of the centralized DP model) decreases with the privacy budget, as opposed to previous discrete DP schemes (Agarwal et al., 2018; Kairouz et al., 2021; Agarwal et al., 2021). Thus in the high-privacy regime, our scheme uses significantly less bandwidth while still achieving the right order of accuracy.
- We combine PBM with distributed SGD and SecAgg in a FL setting and analyze its convergence rate.

1.1. Problem setup and prerequisites

In this section, we present the distributed mean estimation (DME) (Suresh et al., 2017) problem under differential privacy and SecAgg. Note that DME is closely related to federated learning with SGD, where in each iteration, the server updates the global model by a noisy mean of the local model updates. This noisy estimate is typically obtained by using a DME scheme, and thus one can easily build a distributed DP-SGD scheme (and hence a private FL scheme) from a differentially private DME scheme.

Consider n clients each with local data $x_i \in \mathbb{R}^d$ that satisfies $\left\lVert x_i \right\rVert_2 \leq c$ (one can think of x_i as a clipped local gradient). A server wants to learn an estimate $\hat{\mu}$ of the mean $\mu \triangleq \frac{1}{n}\sum_i x_i$ after communicating with the n clients.

Secure aggregation. In order to fully leverage the distributed nature of FL to enhance clients' privacy, the honest-but-curious server collects local data through a *secure aggregation protocol*. More precisely, each client encodes x_i into a finite additive group $\mathcal{Z}$ by computing $Z_i \triangleq \mathcal{A}_{\mathsf{enc}}(x_i)$. The n clients and the server then participate in the SecAgg protocol, so that only $\sum_i Z_i$ can be revealed to the server. Finally, the server computes $\hat{\mu}$ based on $\sum_i Z_i$, an estimate of the true mean. The goal is to jointly design an encoder $\mathcal{A}_{\mathsf{enc}}$ and an estimator $\hat{\mu}$, such that

1. $\hat{\mu}$ satisfies a differential privacy constraint (see Definition 1.1 and Definition 1.2 for formal statements).

[1] We note that the binomial mechanism proposed in (Agarwal et al., 2018) is an additive noise mechanism (it adds Binomial noise to the data), and while it has a finite output range, it does not provide any Rényi DP guarantees (which is the main focus of this paper as RDP allows for tightly accounting privacy loss across multiple rounds).

2. The per-client communication cost $b = \log |\mathcal{Z}|$ is small.
3. $\hat{\mu}$ is unbiased (i.e. $\mathbb{E}[\hat{\mu}] = \mu$) and has small mean squared error (MSE): $\mathbb{E}\left[\|\hat{\mu} - \mu\|_2^2\right]$.

Without loss of generality, we will set $\mathcal{Z}$ to be $(\mathbb{Z}_M)^l$ for some $l, M \in \mathbb{N}$ (where $\mathbb{Z}_M$ is the group of integers modulo M equipped with modulo M addition), so $b = l \log M$ is the total number of communicated bits. The summation is coordinate-wise modulo M addition, i.e.

$$\mathsf{SecAgg}\left(\mathcal{A}_{\mathsf{enc}}(x_1), ..., \mathcal{A}_{\mathsf{enc}}(x_n)\right) = \sum_i \mathcal{A}_{\mathsf{enc}}(x_i) \mathsf{mod} M.$$

Differential Privacy. Finally, we introduce the notion of differential privacy (Dwork et al., 2006b) and Rényi differential privacy (RDP) (Mironov, 2017). We are mostly interested in developing mechanisms that satisfy RDP, as it allows for tight privacy accounting across training iterations.

Definition 1.1 ((Approximate) Differential Privacy)**.** For $\varepsilon, \delta \geq 0$, a randomized mechanism M satisfies (ε, δ)-DP if for all neighboring datasets D, D' and all $\mathcal{S}$ in the range of M, we have that

$$\Pr\left(M(D) \in \mathcal{S}\right) \leq e^{\varepsilon} \Pr\left(M(D') \in \mathcal{S}\right) + \delta,$$

where D and D' are neighboring pairs if they can be obtained from each other by adding or removing all the records that belong to a particular user.

Definition 1.2 (Rényi Differential Privacy (RDP))**.** A randomized mechanism M satisfies (α, ε)-RDP if for any two neighboring datasets D, D', we have that $D_\alpha\left(P_{M(D)}, P_{M(D')}\right) \leq \varepsilon$ where $D_\alpha\left(P, Q\right)$ is the Rényi divergence between P and Q and is given by

$$D_\alpha\left(P, Q\right) \triangleq \frac{1}{\alpha} \log \left(\mathbb{E}_Q\left[\left(\frac{P(X)}{Q(X)}\right)^\alpha\right]\right).$$

Note that one can cast RDP to (approximate) DP. See Section B for details.

1.2. Related works

The closest works to ours are the distributed discrete DP mechanisms cpSGD (Agarwal et al., 2018), DDG (Kairouz et al., 2021), and Skellam (Agarwal et al., 2021). Unlike our proposed scheme, these mechanisms achieve differential privacy (DP) (Dwork et al., 2006b) by adding discrete noise that (1) has a distribution that asymptotically converges to a normal distribution, and (2) are (nearly) "closed" under addition. However, since the noise is asymptotically normal, in the high-privacy regimes where ε is small, the variance of the noise (and hence the communication cost) explodes. In addition, since the noise has infinite range (except for cpSGD[2]), one has to perform modular clipping in order to perform SecAgg. This leads to bias that can cause issues for the downstream tasks such as SGD.

In this paper, we combine our discrete DP mechanism with SecAgg (more precisely, single-server SecAgg) to achieve distributed DP without introducing bias. Single-server SecAgg is achieved via additive masking over a finite group (Bonawitz et al., 2016a; Bell et al., 2020). To achieve provable privacy guarantees, however, SecAgg is insufficient as the sum of local model updates may still leak sensitive information (Melis et al., 2019; Song & Shmatikov, 2019; Carlini et al., 2019; Shokri et al., 2017). To address this issue, DP-SGD or DP-FedAvg can be employed (Song et al., 2013; Bassily et al., 2014; Geyer et al., 2017; McMahan et al., 2017). In this work, we aim to provide privacy guarantees in the form of Rényi DP (Mironov, 2017) because it allows for tracking the end-to-end privacy loss tightly.

We also distinguish our distributed DP setup from the local DP setup (Kasiviswanathan et al., 2011; Evfimievski et al., 2004; Warner, 1965), where the data is perturbed on the client-side before it is collected by the server in the clear. Although both local DP and distributed DP with SecAgg do not rely on a fully trusted centralized server, the local DP model provides stronger privacy guarantees as it allows the server to observe individual (privatized) information, while distributed DP requires that the server executes the SecAgg protocol faithfully. Given its strong privacy guarantees, local DP naturally suffers from poor privacy-utility trade-offs (Kasiviswanathan et al., 2011; Duchi et al., 2013; Kairouz et al., 2016). That is why we focus on distributed DP via SecAgg in this paper.

Our scheme also makes use of Kashin's representation (Kashin, 1977; Lyubarskii & Vershynin, 2010), a powerful tool that enables us to transform the ℓ_2 geometry of the input data to an ℓ_∞ one in a lossless fashion. This facilitates the analysis and allows for decoupling the high-dimensional problem into 1-dimensional sub-tasks. Similar idea has been used in different settings; for instance, (Feldman et al., 2017; Caldas et al., 2018; Chen et al., 2020).

2. Main Results

We introduce the Poisson Binomial mechanism for DME with SecAgg and differential privacy. The proposed protocol (Algorithm 1) consists of three stages:

- Each client computes the Kashin's representation of local data x_i (denoted as y_i), which allows for optimally transforming the ℓ_2 geometry of the data into ℓ_∞.

[2] However, we note that cpSGD only satisfies approximate DP but not Renyi DP, so we can only use strong composition theorems (Dwork et al., 2010; Kairouz et al., 2016) to account privacy loss.

- The n clients apply the scalar Poisson Binomial mechanism (Algorithm 2) separately on each coordinate of y_i, and the server estimates $\mu_y \triangleq \frac{1}{n}\sum_i y_i$.
- The server reconstructs $\hat{\mu}_x$ from the Kashin representation of $\hat{\mu}_y$.

Note that Algorithm 1 builds on Algorithm 2, the scalar version of PBM, which we analyze in Section 3. Parameters (m, θ) determine privacy, communication cost, and MSE; in other words, the privacy-utility trade-offs of Algorithm 1 can be fully characterized by (m, θ), which we summarize in the following theorem.

Theorem 2.1. *Let $\|x_i\| \leq c$, $m \in \mathbb{N}$, and $\theta \in [0, \frac{1}{4}]$. Then with parameters m, θ, Algorithm 1:*

- *satisfies $(\alpha, \varepsilon(\alpha))$-RDP for any $\alpha > 1$ and $\varepsilon(\alpha) = \Omega\left(dm\theta^2\alpha/n\right)$,*
- *requires $O(d\left(\log m + \log n\right))$ bits of per-client communication,*
- *yields an unbiased estimator $\hat{\mu}$ with $O\left(\frac{c^2}{nm\theta^2}\right)$ MSE.*

Remark 2.2. Although in Theorem 2.1 we present an asymptotic result, we remark that (1) the MSE can be upper bound explicitly, and (2) the Rényi DP can be computed numerically (as shown in Section 2.1). Indeed, we show that when we pick θ small enough, the MSE of PBM converges to the (centralized continuous) Gaussian mechanism quickly.

Several observations are given in order. First, the privacy guarantee $\varepsilon(\alpha)$ can be written as a function of the variance (i.e., the MSE): $\varepsilon(\alpha) = \Omega\left(\frac{dc^2\alpha}{n^2\mathsf{MSE}(\hat{\mu})}\right)$. This privacy-accuracy trade-off matches that of the (centralized) Gaussian mechanism [3] given by $\varepsilon_{\mathsf{Gauss}}(\alpha) = \Omega\left(\frac{dc^2\alpha}{n^2\mathsf{MSE}(\hat{\mu}_{\mathsf{Gauss}})}\right)$ (which is obtained by bounding the sensitivity of the mean function by c^2/n^2). This implies that Algorithm 1 attains order-optimal errors. In Section 2.1 below, we numerically compute the MSE-privacy trade-offs of PBM and the Gaussian mechanism.

Algorithm 1 The Poisson Binomial Mechanism

```
Input: $x_1, ..., x_n \in \mathcal{B}_d(c)$, parameters $\theta \in [0, \frac{1}{4}]$, $m \in \mathbb{N}$, a tight frame $U$ associated with Kashin's representation at level $K > 0$
for each client $i$ do
    Set $y_i$ to be the Kashin's representation of $x_i$, so $y_i \in \mathbb{R}^{\Theta(d)}$ and $\|y_i\|_\infty \leq \frac{cK}{\sqrt{d}}$.
    for coordinate $j$ of $y_i$ do
        $Z_{ij} \leftarrow \mathsf{scalar\_PBM}\left(y_{ij}, m, \theta, c' = \frac{cK}{\sqrt{d}}\right)$
    end for
    Send $Z_i$ to the server via SecAgg
end for
(Server) Computes $\hat{\mu_y} = \frac{c'}{mn\theta}\left(\sum_i Z_i - \frac{mn}{2}\right)$
(Server) Computes $\hat{\mu} = U\hat{\mu_y}$
Return: $\hat{\mu}$
```

Algorithm 2 The (Scalar) Poisson Binomial Mechanism

```
Input: $c > 0$, $x_i \in [-c, c]$, $\theta \in [0, \frac{1}{4}]$, $m \in \mathbb{N}$
Re-scaling $x_i$: $p_i \triangleq \frac{\theta}{c}x_i + \frac{1}{2}$.
Privatization: $Z_i \triangleq \mathsf{Binom}\left(m, p_i\right) \in \mathbb{Z}_m$.
Return: $Z_i$
```

Note that in Theorem 2.1, both $\varepsilon(\alpha)$ and the the variance of the estimator depend on the parameters m and θ of the algorithm through the product $m\theta^2$. Hence, this leaves some freedom in the choice of m and θ if one is concerned only with privacy and MSE. However, the choice of m also dictates the communication cost. We next describe how one can pick (θ, m) to minimize the communication cost for the same $(\alpha, \varepsilon(\alpha))$-RDP constraint and MSE, where the latter is dictated by the first according to the above trade-off. Observe that the privacy budget $\varepsilon(\alpha)$ fixes the value of the product $m\theta^2$, so to minimize m, and hence the communication cost, we would like to pick θ as large as possible. However, θ is restricted to $[0, \frac{1}{4}]$. Therefore we can determine m and θ by the following two steps:

1. Set $m = 1$ and compute the corresponding θ such that the resulting privacy is $\varepsilon(\alpha)$. If $\theta > 1/4$, clip θ to 1/4. This leads to $\theta = O\left(\min\left(\frac{1}{4}, \sqrt{\frac{n\varepsilon(\alpha)}{d\alpha}}\right)\right)$.
2. Then, we adjust m again according to θ. If $\theta = 1/4$ (i.e. when θ clipped in the previous step), we set $m = O\left(\frac{n\varepsilon(\alpha)}{d\alpha}\right)$. Otherwise $m = 1$. Hence m is upper bounded by $\max\left(1, O\left(\frac{n\varepsilon(\alpha)}{d\alpha}\right)\right)$.

Plugging the above upper bound on m to Theorem 2.1, the communication cost becomes $O\left(d\left(\log\left(n + \frac{d\varepsilon(\alpha)}{n\alpha}\right)\right)\right)$.

Next, to compare the communication cost of our scheme with previous schemes, we convert it into $(\varepsilon_{\mathsf{DP}}(\delta), \delta)$-DP via Lemma B.2 and arrive at the following corollary:

Corollary 2.3 (Approximate DP of PBM)**.** *By setting $\theta = O\left(\min\left(\frac{1}{4}, \sqrt{\frac{n\log(1/\delta)}{d\varepsilon_{\mathsf{DP}}^2}}\right)\right)$ and $m = \left\lceil\frac{d\varepsilon_{\mathsf{DP}}^2}{n\log(1/\delta)}\right\rceil$, Algorithm 1 satisfies $(\varepsilon_{\mathsf{DP}}, \delta)$-approximate DP. Moreover, the (per-client) communication cost is $O\left(d\left(\log\left(n + \frac{d\varepsilon_{\mathsf{DP}}^2}{n\log(1/\delta)}\right)\right)\right)$, and $\hat{\mu}$ is unbiased with MSE at most $O_\delta\left(\frac{c^2d}{n^2\varepsilon_{\mathsf{DP}}^2}\right)$.*

[3] We notice that when the ℓ_2 sensitivity is c^2/n^2, a Gaussian mechanism that adds $N(0, \sigma^2\mathbb{I}_d)$ noise achieves RDP $\varepsilon_{\mathsf{Gauss}}(\alpha) = \frac{c^2\alpha}{2n^2\sigma^2}$, and the corresponding $\mathsf{MSE}(\hat{\mu}_{\mathsf{Gauss}}) = d\sigma^2$.

We remark that the communication cost of PBM decreases as $\varepsilon_{\mathsf{DP}}$) decreases, exhibiting the correct dependency on $\varepsilon_{\mathsf{DP}}$ in the high-privacy regime. The communication cost of other discrete DP mechanisms based on additive noise, such as (Agarwal et al., 2018; Kairouz et al., 2021; Agarwal et al., 2021) increase as ε gets smaller. For instance, the DDG mechanism (Kairouz et al., 2021) requires $O\left(d\left(\log\left(n+\frac{d}{\varepsilon_{\mathsf{DP}}^2}\right)\right)\right)$ bits of communication per-client, which becomes unbounded when $\varepsilon_{\mathsf{DP}} \to 0$. See Table 2 for a comparison.

	communication	MSE	bias
PBM	$O\left(d\left(\log\lceil\frac{d\varepsilon_{\mathsf{DP}}^2}{n}\rceil\right)\right)$	$O_\delta\left(\frac{c^2d}{n^2\varepsilon_{\mathsf{DP}}^2}\right)$	no
DDG	$O\left(d\left(\log\lceil\frac{d}{\varepsilon_{\mathsf{DP}}^2}\rceil\right)\right)$	$O_\delta\left(\frac{c^2d}{n^2\varepsilon_{\mathsf{DP}}^2}\right)$	yes
Skellam	$O\left(d\left(\log\lceil\frac{d}{\varepsilon_{\mathsf{DP}}^2}\rceil\right)\right)$	$O_\delta\left(\frac{c^2d}{n^2\varepsilon_{\mathsf{DP}}^2}\right)$	yes
Binomial	$O\left(d\left(\log\lceil\frac{d}{\varepsilon_{\mathsf{DP}}^2}\rceil\right)\right)$	$O_\delta\left(\frac{c^2d\log(d)}{n^2\varepsilon_{\mathsf{DP}}^2}\right)$	yes

Table 1. A comparison of the communication costs and MSEs of different discrete DP schemes. For the communication cost, we hide the dependency on $\log n$ since we are interested in high-dimensional regimes where $d \gg n$.

2.1. Numerical evaluation

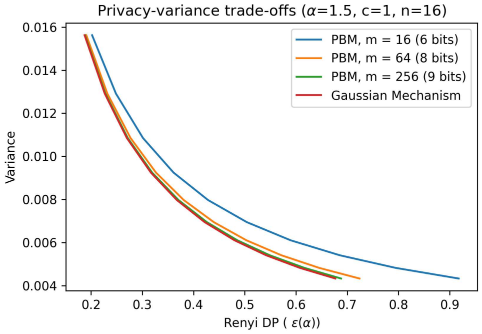

Figure 1. Privacy-MSE (variance) trade-offs of PBM and the Gaussian mechanism.

In figure 1, we numerically compute the privacy guarantee of Algorithm 1 and compared it with the Gaussian mechanism. For the PBM, we fix the communication cost (i.e. fix m), vary parameter θ, and compute the corresponding Rényi DP (i.e., $\varepsilon(\alpha)$) and MSE. We see that as m increases, the privacy-MSE curve approaches to that of the Gaussian mechanism, indicating that our scheme is also optimal in its leading constant. We present another numerical results in Section C, in wich we fix θ and vary m to get the trade-off curves.

3. The Scalar Poisson Binomial Mechanism

In this section, we analyze the utility and privacy guarantees of the scalar version of PBM (i.e., with $d = 1$). Recall that when $d = 1$, each x_i in the DME task (see the formulation in Section 1.1) becomes a bounded real number with $|x_i| < c$ eliminating the Kashin step. Under this special case, Algorithm 2 encodes each x_i into a parameter of a binomial distribution by 1) first mapping x_i into $\left[\frac{1}{2}-\theta, \frac{1}{2}+\theta\right]$ by $p_i \triangleq \frac{1}{2}+\frac{\theta}{c}x_i$, and then 2) generating a binomial random variable $Z_i \sim \mathsf{Binom}(m, p_i)$.

Notice that from each Z_i, one can obtain an unbiased estimator on x_i by computing $\hat{x_i} = \frac{c}{\theta}\left(\frac{1}{m}Z_i - \frac{1}{2}\right)$. Therefore, upon collecting $\sum_i Z_i$ from SecAgg protocol, the server can estimate μ by $\hat{\mu}\left(\sum_i Z_i\right) \triangleq \frac{c}{nm\theta}\left(\sum_i Z_i - \frac{mn}{2}\right)$ (recall that the server can only learn $\sum_i Z_i$ but not individual Z_i).

Remark 3.1. As discussed before, m and θ can be chosen to achieve a desired privacy-utility-communciation trade-off. Intuitively, with larger m, one can reduce the variance of the estimator while weakening the privacy guarantees; similarly with smaller θ, one would get a better privacy guarantee by trading off the accuracy.

Utility of the scalar PBM. As mentioned above, $\hat{\mu}$ yields an *unbiased* estimate on μ, and the variance can be calculated as $\mathsf{Var}\left(\hat{\mu}\right) = \frac{c^2}{m^2\theta^2}\sum_i \mathsf{Var}\left(Z_i\right) \leq \frac{c^2}{4nm\theta^2}$.

On the other hand, since $Z_i \leq m$, $\sum_i Z_i \leq nm$. Thus to avoid overflow, we will set $M = nm$, where recall that M the size of the finite group SecAgg operates on. Therefore, the communication cost of Algorithm 2 is $\log M = \log n + \log m$ bits per client.

Privacy of the scalar PBM. Next, the privacy guarantee (in an RDP form), is summarized in the following corollary.

Corollary 3.2. *Let $m \in \mathbb{N}$ and $\theta \in \left[0, \frac{1}{4}\right]$ be parameters of Algorithm 2. Then Algorithm 2 satisfies $(\alpha, \varepsilon(\alpha))$-RDP for any $\alpha > 1$ and*

$$\varepsilon(\alpha) \geq C_0\left(\frac{\theta^2}{(1-2\theta)^4}\right)\frac{\alpha m}{n}, \tag{1}$$

where $C_0 > 0$ is an universal constant.

3.1. Analysis of the RDP

To analyze the privacy loss of Algorithm 1, let $Y_i \sim \mathsf{Binom}(m, p_i)$ and $Y_1' \sim \mathsf{Binom}(m, p_1')$, where recall that $p_1, ..., p_n, p_1' \in \left[\frac{1}{2}-\theta, \frac{1}{2}+\theta\right]$. For any $\alpha > 1$, the $\varepsilon(\alpha)$ is given by

$$\max_{p_1,\dots,p_n,p_1'} D_\alpha\left(P_{Y_1+Y_2+\dots+Y_n} \middle\| P_{Y_1'+Y_2+\dots+Y_n}\right), \tag{2}$$

with the maximum taken over $\left[\frac{1}{2}-\theta, \frac{1}{2}+\theta\right]^{n+1}$. Our main technical contribution is the following (orderwise) tight upper bound on the Rényi divergence of two Poisson binomial distributions, which then characterizes the privacy loss of our scheme.

Theorem 3.3. *Let $\alpha > 1$ and $p_1, ..., p_n, p_1' \in [\frac{1}{2}-\theta, \frac{1}{2}+\theta]$. Let $Y_i \sim \mathsf{Binom}(m, p_i)$ and $Y_1' \sim \mathsf{Binom}(m, p_1')$. Then it holds that*

$$\begin{aligned} & D_\alpha\left(P_{Y_1+Y_2+...+Y_n} \middle\| P_{Y_1'+Y_2+...+Y_n}\right) \\ & \leq C_0 \frac{\theta^2}{(1-2\theta)^4}\left(\min\left(4, \frac{\alpha^2}{\alpha-1}\right)\right)\frac{m}{n}, \end{aligned}$$

where $C_0 > 0$ is an universal constant.

Remark 3.4. Although we present an asymptotic result here, using the quasi-convexity of Rényi divergence, one can show that the worst-case scenario is attained by the extremal points (i.e., when $p_i \in \left\{\frac{1}{2}-\theta, \frac{1}{2}+\theta\right\}$), as shown in Lemma 3.5. This allows us to efficiently compute the privacy loss *exactly*, as shown in Section 2.1.

An immediate corollary of Theorem 3.3 is the RDP guarantee of the proposed PBM (summarized in Corollary 3.2).

In the rest of this section, we provide a proof of Theorem 3.3.

Step 0: decomposing Y_i. To begin with, observe that since $Y_i \sim \mathsf{Binom}(m, p_i)$, we can decompose it into sum of m independent and identical copies of $\mathsf{Ber}(p_i)$, i.e., $Y_i = \sum_{j=1}^m X_i^{(j)}$, where $X_i^{(j)} \overset{\text{i.i.d.}}{\sim} \mathsf{Ber}(p_i)$ for $j \in [m]$. Therefore

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \sum_{j=1}^m X_i^{(j)} = \sum_{j=1}^m \underbrace{\left(\sum_{i=1}^n X_i^{(j)}\right)}_{\triangleq Z_j},$$

and similarly we can write $Y_1' + \sum_{i=2}^n Y_i = \sum_{j=1}^m Z_j'$, where $Z_j' = X_1^{'(j)} + \sum_{i=2}^n X_i^{(j)}$.

Grouping the summation of $X_i^{(j)}$ according to $j \in [m]$ and applying the data processing inequality for Rényi divergence, we upper bound (2) by

$$\begin{aligned} & \max_{p_1,...,p_n,p_1'} D_\alpha\left(P_{Y_1+Y_2+...+Y_n} \middle\| P_{Y_1'+Y_2+...+Y_n}\right) \\ = & \max_{p_1,...,p_n,p_1'} D_\alpha\left(P_{Z_1+...+Z_m} \middle\| P_{Z_1'...+Z_m'}\right) \\ \leq & \max_{p_1,...,p_n,p_1'} m D_\alpha\left(P_{Z_1} \middle\| P_{Z_1'}\right) \\ = & \max_{p_1,...,p_n,p_1'} m D_\alpha\left(P_{X_1+X_2+...+X_n} \middle\| P_{X_1'+X_2+...+X_n}\right), \end{aligned} \tag{3}$$

where $X_i \sim \mathsf{Ber}(p_i)$ and $X_1' \sim \mathsf{Ber}(p_1')$.

Step 1: maximum achieved by extremal points. Next, since $(P, Q) \mapsto D_\alpha(P\|Q)$ is quasi-convex (Van Erven & Harremos, 2014, Thoerem 13), we claim that (3) is maximized at extreme points:

Lemma 3.5. (3) *is maximized at extreme points i.e., when $p_1, ..., p_n, p_1' \in \left\{\frac{1}{2}-\theta, \frac{1}{2}+\theta\right\}$.*

This implies (3) can be upper bounded by the following binomial form:

$$\max_{k\in[n-1]} D_\alpha\Big(P_{\mathsf{Binom}\left(1+k,\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k-1,\frac{1}{2}+\theta\right)} \Big\| P_{\mathsf{Binom}\left(k,\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k,\frac{1}{2}+\theta\right)}\Big). \tag{4}$$

Step 2: applying data processing inequality. Next, we simplify (4) by carefully applying the data processing inequality. Let $k^* \in [n-1]$ maximize (4). If $k^* \leq \frac{n}{2}$, we apply the data processing inequality to discard the first half of common binomial random variables (see Figure 2 for an illustration), i.e.,

$$\mathsf{Binom}\left(k^*, \frac{1}{2}-\theta\right) + \mathsf{Binom}\left(n'-k^*, \frac{1}{2}+\theta\right),$$

where $n' \triangleq \lceil \frac{n-1}{2} \rceil$. On the other hand, if $k^* \geq \frac{n}{2}$, then we apply the data processing inequality to remove the second half of common parts, i.e.,

$$\mathsf{Binom}\left(n'+k^*, \frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*-1, \frac{1}{2}+\theta\right).$$

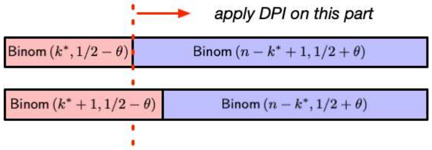
apply DPI on this part
Binom $(k^*, 1/2-\theta)$
Binom $(n-k^*+1, 1/2+\theta)$
Binom $(k^*+1, 1/2-\theta)$
Binom $(n-k^*, 1/2+\theta)$

Figure 2. An illustration of applying data processing inequality, where under the scenario of $k^* \leq \frac{1}{2}$, we discard the first half of common binomial sum.

This leads to the following lemma:

Lemma 3.6. (4) *is upper bounded by the maximum of the following two quantities:*

(a) $D_\alpha\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n',\frac{1}{2}+\theta\right)} \middle\| P_{\mathsf{Binom}\left(n'+1,\frac{1}{2}+\theta\right)}\right)$,

(b) $D_\alpha\left(P_{\mathsf{Binom}\left(n'+1,\frac{1}{2}-\theta\right)} \middle\| P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+\mathsf{Binom}\left(n',\frac{1}{2}-\theta\right)}\right)$.

Step 3: bounding the Rényi divergence via MGF. Finally, we upper bound each of the two terms in (7) separately. We start with the following simple but useful lemma, which bounds the Rényi divergence of two distributions by the sub-Gaussian norm of their likelihood ratio (LR).

Lemma 3.7. *Let P, Q be two probability measures on $\mathcal{X}$ and let $\frac{dP}{dQ}(x)$ be the Radon-Nikodym derivative. Let $X \sim$*

Q. Then for any $\alpha > 1$,

$$D_\alpha\left(P \| Q\right) \leq C_0 \frac{\alpha^2}{\alpha - 1} \left\| \frac{dP}{dQ}\left(X\right) - 1 \right\|_{\psi_2}^2,$$

where $\|Z\|_{\psi_2}$ denotes the sub-Gaussian norm of Z and $C_0 > 0$ is a universal constant.

To apply Lemma 3.7 to control (a) and (b) in Lemma 3.6, we need to compute and bound the sub-Gaussian norms of the likelihood ratio (LRs) of random variables in (a) and (b) of Lemma 3.6, respectively.

To this end, let us define $R(i) \triangleq \frac{P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}+\theta\right)}(i)}{P_{\mathsf{Binom}\left(n'+1, \frac{1}{2}+\theta\right)}(i)}$. Then the LRs corresponding to random variables of (a) and (b) in Lemma 3.6 are $R(I)$ and $1/R(I')$ respectively, where $I \sim P_{\mathsf{Binom}\left(n'+1, \frac{1}{2}+\theta\right)}$ and $I' \sim P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}+\theta\right)}$.

It turns out that $R(i)$ is a linear function of i, and since both I and I' are sum of binary random variables, one can control their sub-gaussian norms (and hence that of $R(I)$ and $1/R(I')$). We summarize the bound in the following lemma and defer the proof to Section D.5.

Lemma 3.8. *Let $R(i)$ be defined as above and let $I \sim P_{\mathsf{Binom}\left(n'+1, \frac{1}{2}+\theta\right)}$ and $I' \sim P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}+\theta\right)}$. Then*

- $\|R(I) - 1\|_{\psi_2}^2 \leq C_1 \frac{\theta^2}{(1-4\theta^2)^2(n'+1)}$,
- $\|1/R(I') - 1\|_{\psi_2}^2 \leq C_2 \frac{\theta^2}{(1-2\theta)^4} \frac{1}{n'+1}$,

for some $C_1, C_2 > 0$.

Step 4: putting everything together. Combining Lemma 3.8, Lemma 3.7, and Lemma 3.6, we obtain that

$$(4) \leq C_3 \frac{\theta^2}{(1-2\theta)^4}\left(\frac{\alpha^2}{\alpha-1}\right)\frac{1}{n'+1}.$$

Together with Lemma 3.5 and (3), we conclude that

$$\begin{aligned}\max_{p_1,\ldots,p_n,p_1'} & D_\alpha\left(P_{Y_1+Y_2+\ldots+Y_n} \middle\| P_{Y_1'+Y_2+\ldots+Y_n}\right) \\ & \leq C_7 \left(\frac{\theta^2}{(1-2\theta)^4}\right)\left(\frac{\alpha^2}{\alpha-1}\right)\frac{m}{n'+1} \\ & \leq C_0 \left(\frac{\theta^2}{(1-2\theta)^4}\right)\left(\frac{\alpha^2}{\alpha-1}\right)\frac{m}{n}, \qquad (5)\end{aligned}$$

for some $C_0 > 0$ large enough. Finally, since Rényi divergence is increasing with α, we also have for $\alpha < 2$,

$$\begin{aligned}& \max_{p_1,\ldots,p_n,p_1'} D_\alpha\left(P_{Y_1+Y_2+\ldots+Y_n} \middle\| P_{Y_1'+Y_2+\ldots+Y_n}\right) \\ & \leq \max_{p_1,\ldots,p_n,p_1'} D_2\left(P_{Y_1+Y_2+\ldots+Y_n} \middle\| P_{Y_1'+Y_2+\ldots+Y_n}\right) \\ & \leq C_0 \left(\frac{\theta^2}{(1-2\theta)^4}\right)\frac{4m}{n}. \qquad (6)\end{aligned}$$

Combining (5) and (6), we establish Theorem 3.3.

4. The Multi-dimensional PBM

Next, we extend the scalar PBM into the multi-dimensional setting, where $x_i \in \mathbb{R}^d$ and $\|x_i\|_2 \leq c$. The description of multi-dimensional PBM is given in Algorithm 1. The key step that allows us to cast the multi-dimensional DME into the scalar one is via Kashin's representation, which transforms the ℓ_2 geometry of the data into an ℓ_∞ geometry and hence enables us to decompose the problem into scalar sub-tasks.

4.1. Kashin's representation

We first introduce the idea of a tight frame in Kashin's representation. A tight frame is a set of vectors $\{u_j\}_{j=1}^D \in \mathbb{R}^d$ that satisfy Parseval's identity, i.e. $\|x\|_2^2 = \sum_{j=1}^D \langle u_j, x\rangle^2$ for all $x \in \mathbb{R}^d$.

A frame can be viewed as a generalization of the notion of an orthogonal basis in $\mathbb{R}^d$ for $N > d$. To increase robustness, we wish the information to be spread evenly across different coefficients, which motivates the following definition of a Kashin's representation:

Definition 4.1 (Kashin's representation(Kashin, 1977))**.** For a set of vectors $\{u_j\}_{j=1}^D$, we say the expansion

$$x = \sum_{j=1}^D a_j u_j, \text{ with } \max_j |a_j| \leq \frac{K}{\sqrt{D}} \|x\|_2$$

is a Kashin's representation of vector x at level K .

By Theorem 3.5 and Theorem 4.1 in (Lyubarskii & Vershynin, 2010), we have the following lemma:

Lemma 4.2 (Uncertainty principle)**.** *There exists a tight frame $U = [u_1, ..., u_D]$ with (1) $D = \Theta(d)$ and (2) Kashin's level $K = O\left(1\right)$.*

Lemma 4.2 implies that for each $x_i \in \mathbb{R}^d$ such that $\|x_i\|_2 \leq 1$, one can always represent each x_i with coefficients $y_i \in [-\gamma_0/\sqrt{d}, \gamma_0/\sqrt{d}]^{\gamma_1 d}$ for some $\gamma_0, \gamma_1 > 0$ and $x_i = U y_i$.

4.2. Proof of Theorem 2.1

With Lemma 4.2, we are well-prepared to analyze the performance of Algorithm 1. Recall that the three main steps in the multi-dimensional PBM are:

1. (Clients) compute a Kashin's representation of x_i with respect to a (common) tight frame U (denoted as y_i).
2. (Clients) sequentially transmit each coordinate of y_i via the scalar PBM.
3. (Server) reconstructs μ by $\hat{\mu} = U\hat{\mu}_y$.

Let us denote $y_i(j)$ as the j-th coordinate of y_i for $j \in [\gamma_1 d]$, and $\mu_y(j)$ as the j-th coordinate of $\mu_y = \frac{1}{n}\sum_{i=1}^n y_i$. Due to the property of the Kashin's representation, we know that if $\|x_i\|_2 \leq c$, $\|y_i\|_\infty \leq \frac{\gamma_0 c}{\sqrt{d}}$.

Therefore, using the scalar PBM (with parameters θ, m) for coordinate j and applying Theorem 3.3, it holds that

- the privacy loss is $\varepsilon_{\mathrm{j}}(\alpha) = C_0\left(\frac{\theta^2}{(1-2\theta)^4}\right)\frac{\alpha m}{n}$;
- $\mathbb{E}\left[\hat{\mu}_y(j)\right] = \hat{\mu}_y(i)$ and $\mathbb{E}\left[\left(\hat{\mu}_y(j) - \mu_j\right)^2\right] \leq \frac{\gamma_0^2 c^2}{4dnm\theta}$;
- The communication cost is $O\left(\log n + \log m\right)$ bits.

Repeating for $j = 1, .., \gamma_1 d$ and accounting the overall privacy loss via the composition theorem of RDP (Mironov, 2017, Proposition 1), the end-to-end RDP guarantee of Algorithm 2 becomes $\varepsilon(\alpha) = \sum_j \varepsilon_j(\alpha) = \gamma_1 C_0 \frac{d\alpha m\theta^2}{(1-2\theta)^4 n}$. Similarly, the communication cost is $\gamma_1 d\left(\log n + \log m\right)$ bits.

Finally, we control the ℓ_2 estimation error $\mathbb{E}\left[\|\mu - \hat{\mu}\|_2^2\right]$. Note that since $x_i = Uy_i$ for all $i = 1, ..., n$, we have $\mu_x = \frac{1}{n}\sum_i Uy_i = U\mu_y$. Also,

$$\mathbb{E}\left[\left(\hat{\mu} - \mu\right)^2\right] = \mathbb{E}\left[\left\|\sum\nolimits_j \left(\hat{\mu}_y(j) - \mu_y(j)\right) u_j\right\|_2^2\right]$$

$$\overset{\text{(a)}}{\leq} \mathbb{E}\left[\sum\nolimits_j \left(\hat{\mu}_y(j) - \mu_y(j)\right)^2\right] = \mathbb{E}\left[\|\hat{\mu}_y - \mu_y\|_2^2\right],$$

where (a) is due to the Cauchy–Schwarz inequality. Hence to bound the MSE of $\hat{\mu}$, it suffices to bound $\mathbb{E}\left[\|\hat{\mu}_y - \mu_y\|_2^2\right] \leq \gamma_1 \gamma_0^2 \frac{c^2}{4mn\theta} = O\left(\frac{c^2}{4mn\theta}\right)$.

5. Empirical evaluations

In this section, we evaluate PBM on the distributed mean estimation (DME) task. We follow the set-up of Kairouz et al. (2021); Agarwal et al. (2021): we generate $n = 1000$ client vectors with dimension $d = 250$, i.e., $x_1, ..., x_n \in \mathbb{R}^{250}$. Each local vector has bounded ℓ_2 and ℓ_∞ norms, i.e. $\|x_i\|_2 \leq 1$ and $\|x_i\|_\infty \leq \frac{1}{\sqrt{d}}$[4]. Our goal is to demonstrate that the utility of PBM matches that of the continuous Gaussian mechanism under the same privacy guarantees when given sufficient communication budget.

In Figure 3, we apply PBM with different m's (which dictate the communication cost) under a given privacy requirement ε. Note that once m is determined, the field that SecAgg operates on will be $\mathbb{Z}_{2^{\lceil \log_2(n\cdot m)\rceil}}$ and hence the communication cost becomes $\lceil \log_2(n \cdot m)\rceil$ bits.

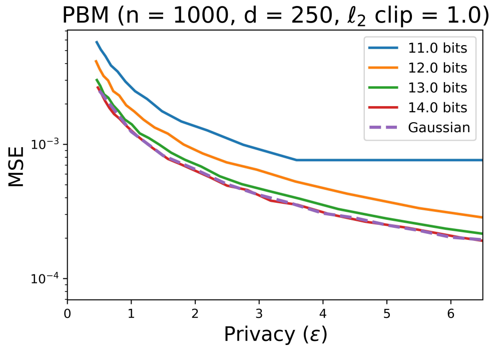

Figure 3. A comparison of PBM with the continuous Gaussian mechanism. We set $m = \{2, 4, 6, 16\}$, and the corresponding communication costs (i.e., the logarithmic of the field size that SecAgg operates on) are $B = \{11, 12, 13, 14\}$.

Reducing communication via modular clipping. Note that since the sum of encoded messages (i.e., $\sum Z_i$ in Algorithm 2) follows a Poisson-binomial distribution with $p_i \in \left[\frac{1}{2} - \theta, \frac{1}{2} + \theta\right]$, with high-probability, $\sum Z_i \in \left[\frac{nm(1-\theta)}{2} - c\sqrt{\frac{nm}{4}}, \frac{nm(1+\theta)}{2} + c\sqrt{\frac{nm}{4}}\right]$ (where c is a parameter used to control the failure probability). Therefore one can (modularly) clip local message in that range to further reduce and hence communication cost the field size to $nm\theta + c\sqrt{nm}$ (this, however, will incur *bias* to the final estimator). We report the result in Figure 5 of the appendix.

6. Application to private SGD

In this section, we apply PBM to distributed SGD. In each round, the server samples n out of N clients randomly, each (sampled) client computes a local gradient from its data, and the server aggregates the mean of the local gradients via PBM. Since PBM ensures distributed DP, we call the resulting scheme DDP-SGD.

We summarize DDP-SGD in Algorithm 3, in which we use $\mathsf{PBM}_{\mathsf{enc}}$ to denote the clients' procedure in Algorithm 1 (which includes computing Kashin's representation, sequentially applying the scalar PBM, and performing SecAgg) and $\mathsf{PBM}_{\mathsf{dec}}$ to denote the server's procedure.

Analysis of convergence rates. Next, we analyze the convergence rate of Algorithm 3. Due to the unbiased nature of PBM, one can easily control the convergence rate of DDP-SGD by the variance of PBM through the following lemma (which originates from (Ghadimi & Lan, 2013) but we use a version adapted from (Agarwal et al., 2018)):

Lemma 6.1 (Corollary 1 in (Agarwal et al., 2018))**.** *Assume $F(w) \triangleq \frac{1}{N}\sum_{i=1}^{N} \ell(w; d_i)$, where $\ell(\cdot, d)$ is an L-smooth and c-Lipschitz function for all $d \in \mathcal{D}$. Let w_0 satisfies $F(w_0) - F(w^*) \leq D_F$. Let $\hat{\mu}_g^t$ be the noisy model updates at round t and let the learning rate $\gamma \triangleq$*

[4] We perform the experiments under an ℓ_∞ geometry. Under ℓ_2 geometry, one can either apply random rotation and ℓ_∞ clipping, or compute the Kashin's representation, as discussed in Section 4.

Algorithm 3 Distributed DP-SGD

```
Input: Clients data set $d_1, ..., d_N \in \mathcal{D}$, PBM parameters $(\theta, m)$, loss function $\ell(d, w)$.
Goal: Compute $w_T \approx \arg\min_w \sum_{i=1}^N \ell\left(d_i, w\right)$.
Server generates an initial model weights $w_0 \in \mathbb{W}$
for iteration $t = 1, ..., T$ do
    Server samples a subset of $n$ clients $\mathcal{C}_t \subset [N]$ and broadcasts $w_{t-1}$ to them.
    for each client $i \in \mathcal{C}_t$ do
        Computes $g_i^t = \mathsf{Clip}_{\ell_2, c}\left(\nabla \ell\left(d_i, w_{t-1}\right)\right)$
        Computes $Z_i^t = \mathsf{PBM}_{\mathsf{enc}}\left(g_i^t\right)$
        Send $Z_i^t$ to the server via SecAgg
    end for
    (Server) decodes $\hat{\mu}_g^t = \mathsf{PBM}_{\mathsf{dec}}\left(\sum_i Z_i^t\right)$.
    (Server) updates the model by $w_t = w_{t-1} + \gamma \hat{\mu}_g^t$.
end for
Return: $w_T$
```

$\min\left\{L^{-1}, \sqrt{2D_F}\left(\sigma\sqrt{LT}\right)^{-1}\right\}$. *Then after T rounds,*

$$\mathbb{E}_{t\sim \mathsf{unif}(T)}\left[\left\|\nabla F(w_t)\right\|_2^2\right] \leq \frac{2D_F L}{T} + \frac{2\sqrt{2}\sigma\sqrt{LD_F}}{\sqrt{T}} + cB,$$

where $\sigma^2 = 2\Big(\max_{1\leq t\leq T} \mathbb{E}\left[\left\|\mu_g^t - \nabla F(w_t)\right\|_2^2\right]$

$$+ \max_{1\leq t\leq T} \mathbb{E}_Q\left[\left\|\mu_g^t - \tilde{\mu}_g^t\right\|_2^2\right]\Big),$$

and $B = \max_{1\leq t\leq T}\left\|\mathbb{E}_Q\left[\mu_g^t - \hat{\mu}_g^t\right]\right\|_2$.

Note that in each round, (1) μ_g^t (the true mean of (the sampled) clients' gradients) is an unbiased estimator of $\nabla F(w_t)$ (because clients are sampled uniformly at random), and (2) $\hat{\mu}_g^t$ is an unbiased estimator of μ_g^t since PBM is unbiased[5]. This implies $B = 0$ and $\sigma^2 = \max_t \mathsf{Var}\left(\mu_g^t\right) + \mathsf{Var}\left(\hat{\mu}_g^t \middle| \mu_g^t\right)$, where the first term is bounded by c^2, and applying Theorem 2.1, we can bound the second $\mathsf{Var}\left(\hat{\mu}_g^t \middle| \mu_g^t\right)$ by $\frac{c^2}{4nm\theta^2}$. Thus we arrive at the following conclusion:

Corollary 6.2 (Convergence of DDP-SGD)**.** *Under the same assumptions of Lemma 6.1, after $\tau \sim \mathsf{uniform}(T)$ iterations, the output of Algorithm 3 satisfies*

$$\mathbb{E}_\tau\left[\left\|\nabla F(w_\tau)\right\|_2^2\right] \leq \frac{LD_F}{T} + \frac{\sqrt{8c^2LD_F}}{\sqrt{T}}\sqrt{1 + \frac{1}{4nm\theta^2}}.$$

Remark 6.3. Note that due to the convergence guarantees of Lemma 6.1, in Corollary 6.2 we apply Algorithm 3 with a random stopping time.

Accounting for total privacy loss. To account for the total privacy loss, we first note that the per-round RDP guarantee is amplified by the sub-sampling of the clients with sampling rate $\kappa = \frac{n}{N}$. To quantify the tight amplification rate, one can apply (Wang et al., 2019, Theorem 9) and obtain a non-asymptotic upper bound on $\varepsilon_{\mathsf{sampled}}(\alpha)$. For instance, (Wang et al., 2019, Theorem 9) shows that when α is not too large (e.g., when $\alpha \leq 2$), $\varepsilon_{\mathsf{sampled}}(\alpha) = O\left(\kappa^2\varepsilon(\alpha)\right)$. Applying Theorem 2.1 and plugging $\varepsilon(\alpha) = \frac{dm\theta^2\alpha}{n}$ in this bound, we have $\varepsilon_{\mathsf{sampled}}(\alpha) = O\left(\frac{ndm\theta^2\alpha}{N^2}\right)$. To account for the privacy loss over all the T iterations, we apply the composition theorem for RDP (Mironov, 2017, Proposition 1), concluding that Algorithm 3 satisfies $(\alpha, \varepsilon_{\mathsf{final}})$ with $\varepsilon_{\mathsf{final}}(\alpha) = O\left(\frac{ndm\theta^2\alpha T}{N^2}\right)$ when $\alpha \leq 2$.

For high-privacy (where $nm\theta^2 \ll 1$) and small α regimes, we obtain a convergence rate-privacy trade-off:

$$\mathbb{E}_\tau\left[\left\|\nabla F(w_\tau)\right\|_2^2\right] \approx O\left(\sqrt{\frac{c^2 d\alpha}{N^2\varepsilon_{\mathsf{final}}(\alpha)}}\right).$$

This achieves the optimal rate of reported in (Bassily et al., 2014, Table 1) (though admittedly, our results only hold for high privacy regime and for $\alpha < 2$). Finally, we remark to obtain non-asymptotic trade-offs for full $\alpha > 1$ regimes, one has to resort to (Wang et al., 2019, Theorem 9).

7. Conclusion

In this paper, we present the Poisson Binomial mechanism, a discrete (Rényi) DP mechanism that can be combined with SecAgg for distributed mean estimation or federated learning/analytics. Unlike previous schemes, our mechanism is not based on additive noise, so in addition to achieving the optimal privacy-accuracy trade-off, it offers two extra advantages: (1) it results in an unbiased estimator, and (2) the communication cost with SecAgg decreases with the privacy budget ε. Leveraging the unbiasedness property, we propose a distributed DP-SGD algorithm and analyze its convergence rate. Several important open problems include deriving a lower bound on the communication cost for DME with DP and SecAgg, and evaluating our scheme on real-world FL datasets.

Acknowledgments

This work was supported in part by NSF Award # NeTS-1817205, a Google Faculty research award, and a National Semiconductor Corporation Stanford Graduate Fellowship. The authors would like to thank Thomas Steinke for inspiring discussion in the early stage of this work.

[5] Notice that the clipping step in Algorithm 3 does not increase bias since by the Lipschitz condition, $\left\|\nabla\ell\right\|_2 \leq c$.

References

Agarwal, N., Suresh, A. T., Yu, F. X. X., Kumar, S., and McMahan, B. cpsgd: Communication-efficient and differentially-private distributed sgd. In *Advances in Neural Information Processing Systems*, pp. 7564–7575, 2018.

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

Asoodeh, S., Liao, J., Calmon, F. P., Kosut, O., and Sankar, L. A better bound gives a hundred rounds: Enhanced privacy guarantees via f-divergences. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 920–925. IEEE, 2020.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

Bell, J. H., Bonawitz, K. A., Gascón, A., Lepoint, T., and Raykova, M. Secure single-server aggregation with (poly) logarithmic overhead. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1253–1269, 2020.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016a.

Bonawitz, K. A., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for federated learning on user-held data. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016b. URL https://arxiv.org/abs/1611.04482.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

Caldas, S., Konečny, J., McMahan, H. B., and Talwalkar, A. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018.

Canonne, C. L., Kamath, G., and Steinke, T. The discrete gaussian for differential privacy. *arXiv preprint arXiv:2004.00010*, 2020.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 267–284, 2019.

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

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006a.

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, 2006b.

Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE, 2010.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. Privacy preserving mining of association rules. *Information Systems*, 29(4):343–364, 2004.

Feldman, V., Guzman, C., and Vempala, S. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1265–1277. SIAM, 2017.

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

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Kairouz, P., Bonawitz, K., and Ramage, D. Discrete distribution estimation under local privacy. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 2436–2444, New York, New York, USA, 20–22 Jun 2016.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Kairouz, P., Liu, Z., and Steinke, T. The distributed discrete gaussian mechanism for federated learning with secure aggregation. *arXiv preprint arXiv:2102.06387*, 2021.

Kashin, B. Section of some finite-dimensional sets and classes of smooth functions (in russian) izv. *Acad. Nauk. SSSR*, 41:334–351, 1977.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Lyubarskii, Y. and Vershynin, R. Uncertainty principles and vector quantization. *IEEE Transactions on Information Theory*, 56(7):3491–3501, 2010.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. Communication-efficient learning of deep networks from decentralized data (2016). *arXiv preprint arXiv:1602.05629*, 2016.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 691–706. IEEE, 2019.

Mironov, I. On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pp. 650–661, 2012.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017.

Song, C. and Shmatikov, V. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 196–206, 2019.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, 2013.

Suresh, A. T., Yu, F. X., Kumar, S., and McMahan, H. B. Distributed mean estimation with limited communication. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3329–3337. JMLR.org, 2017.

Van Erven, T. and Harremos, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1226–1235. PMLR, 2019.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

A. Societal Considerations

Harnessing distributed data holds the promise of impacting many facets of our lives. It could enable truly large scale smart infrastructure and IoT applications; having a profound and positive impact on power-grid efficiency, traffic, health-monitoring, medical diagnoses, carbon emissions, and many other areas. A foundational understanding of distributed learning and estimation can also benefit many different fields of study such as neuroscience, medicine, economics, and social networks, where statistical tools are often used to analyze information that is generated and processed in large networks.

While the above vision is expected to generate many disruptive business and social opportunities, it presents a number of unprecedented challenges. First, massive amounts of data need to be collected by, and transferred across, resource-constrained devices. Second, the collected data needs to be stored, processed, and analyzed at scales never previously seen. Third, serious concerns such as access control, data privacy, and security should be rigorously addressed.

Our work tackles the above challenges by examining the trade-off between communication, privacy, and accuracy by taking a holistic approach that examines all these constraints simultaneously and designing provably private FL mechanisms. Our work therefore serves as a stepping stone towards harnessing large-scale distributed data in a privacy-preserving and bandwidth efficient way.

B. Conversion of RDP to approximate DP

The following conversion lemma from (Asoodeh et al., 2020; Canonne et al., 2020; Bun & Steinke, 2016) relates RDP to $(\varepsilon_{\mathsf{DP}}(\delta), \delta)$-DP.

Lemma B.1. *If M satisfies $(\alpha, \varepsilon(\alpha))$-RDP for all $\alpha > 1$, then, for any $\delta > 0$, M satisfies $(\varepsilon_{\mathsf{DP}}(\delta), \delta)$-DP, where*

$$\varepsilon_{\mathsf{DP}}(\delta) = \inf_{\alpha>1} \varepsilon(\alpha) + \frac{\log\left(1/\alpha\delta\right)}{\alpha - 1} + \log(1 - 1/\alpha).$$

To compare the privacy-utility trade-off of our mechanism with previous works, it will be useful sometimes to convert the stronger RDP guarantees to an approximate DP guarantee. The following lemma, which is a simple application of Lemma B.1, provides a simple form for the resulting approximate DP guarantees.

Lemma B.2. *If M satisfies $(\alpha, \varepsilon(\alpha))$-RDP for all $\alpha > 1$, then, for any $\delta > 0$, M satisfies $(\varepsilon_{\mathsf{DP}}(\delta), \delta)$-DP, where*

$$\varepsilon_{\mathsf{DP}}(\delta) = \Theta\left(\sqrt{\left(\sup_{\alpha} \frac{\varepsilon(\alpha)}{\alpha}\right)\log(1/\delta)}\right).$$

C. Additional plots

Numerical evaluation of the scalar PBM. In figure 4, we compute the privacy guarantee of Algorithm 1 and compared it with the Gaussian mechanism. For the PBM, we fix θ, vary parameter m, and compute the corresponding Rényi DP (i.e., $\varepsilon(\alpha)$) and MSE. We see that as $\theta \to 0$, the privacy-MSE curve converges to that of the Gaussian mechanism fast.

Reducing communication via modular clipping. Since the sum of encoded messages (i.e., $\sum Z_i$ in Algorithm 2) follows a Poisson-binomial distribution with $p_i \in \left[\frac{1}{2} - \theta, \frac{1}{2} + \theta\right]$, with high-probability, $\sum Z_i \in \left[\frac{nm(1-\theta)}{2} - c\sqrt{\frac{nm}{4}}, \frac{nm(1+\theta)}{2} + c\sqrt{\frac{nm}{4}}\right]$ (where c is a parameter used to control the failure probability). Therefore one can (modularly) clip local message in that range to further reduce and hence communication cost the field size to $nm\theta + c\sqrt{nm}$ (this, however, will incur *bias* to the final estimator). In Figure 5, we perform PBM with modular clipping and set $c = \sqrt{30}$. Since c is large enough, we observe almost no impact to the errors while saving the (per-parameter) communication by 1 bit.

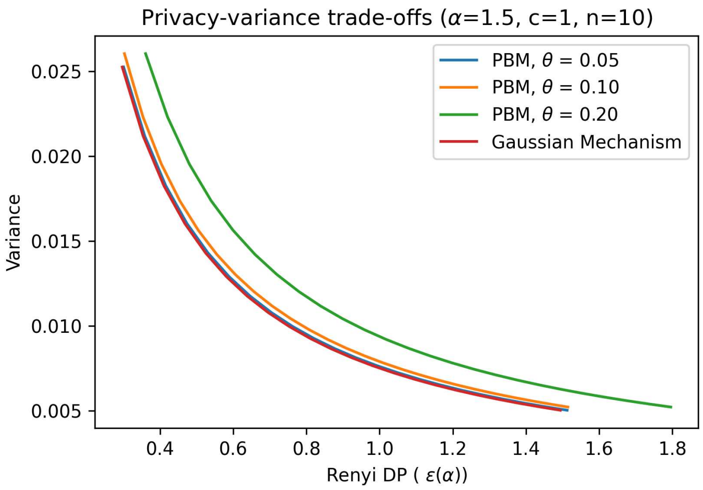

Figure 4. Privacy-MSE (variance) trade-offs of PBM and the Gaussian mechanism.

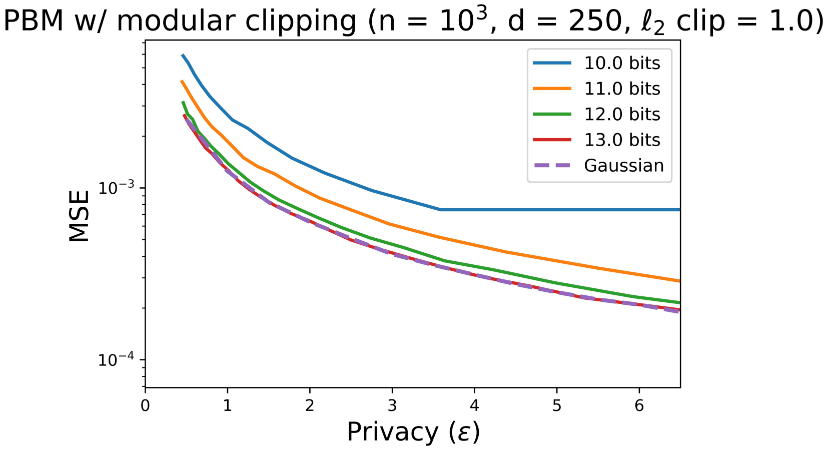

Figure 5. PBM with modular clipping, where $m = \{2, 4, 6, 16\}$. We set the modular clippiing parameter $c = \sqrt{30}$ and report the MSEs and the corresponding communication cost.

D. Additional Proofs

D.1. Proof of Lemma B.2

$$\begin{aligned}
\varepsilon_{\mathsf{DP}}(\delta) &\leq \varepsilon(\alpha) + \frac{\log\left(1/\alpha\delta\right)}{\alpha-1} + \log(1-1/\alpha) \\
&\leq \varepsilon(\alpha) + \frac{\log(1/\delta)}{\alpha-1} \\
&= \frac{\varepsilon(\alpha)}{\alpha} + \frac{\varepsilon(\alpha)}{\alpha}(\alpha-1) + \frac{\log(1/\delta)}{\alpha-1}.
\end{aligned}$$

Next, take $\alpha^* = 1 + \sqrt{\sup_\alpha \frac{\alpha}{\varepsilon(\alpha)}}$, we get

$$\varepsilon_{\mathsf{DP}}(\delta) \leq \sup_\alpha \frac{\varepsilon(\alpha)}{\alpha} + 2\sqrt{\left(\sup_\alpha \frac{\varepsilon(\alpha)}{\alpha}\right)\log(1/\delta)} = \Theta\left(\sqrt{\left(\sup_\alpha \frac{\varepsilon(\alpha)}{\alpha}\right)\log(1/\delta)}\right),$$

if $\frac{\varepsilon(\alpha)}{\alpha} = O(1)$. This suggests that $(\alpha, \varepsilon(\alpha))$-RDP implies $(\varepsilon_{\mathsf{DP}}(\delta), \delta)$-DP with $\varepsilon_{\mathsf{DP}}(\delta) = \Theta_\delta\left(\sqrt{\left(\sup_\alpha \frac{\varepsilon(\alpha)}{\alpha}\right)}\right)$.

D.2. Proof of Lemma 3.5

To see this, observe that $P_{X_1+X_2+\ldots+X_n} = P_{X_1+\ldots+X_{n-1}} \circ P_{X_n}$, where $\circ$ denotes the convolution operator. Since the convolution operator is linear, we have

$$P_{X_1+\ldots+X_{n-1}} \circ P_{X_n} = \lambda\left(P_{X_1+\ldots+X_{n-1}} \circ \mathsf{Ber}\left(\frac{1}{2}-\theta\right)\right) + (1-\lambda)\left(P_{X_1+\ldots+X_{n-1}} \circ \mathsf{Ber}\left(\frac{1}{2}+\theta\right)\right),$$

where $\lambda > 0$ is such that $p_n = \lambda\left(\frac{1}{2}-\theta\right) + (1-\lambda)\left(\frac{1}{2}+\theta\right)$. By the quasi-convexty of $D_\alpha(\cdot\|\cdot)$, it holds that $D_\alpha\left(P_{X_1+X_2+\ldots+X_n} \middle\| P_{X_1'+X_2+\ldots+X_n}\right)$ is upper bounded by

$$\max\left(D_\alpha\left(P_{X_1+X_2+\ldots+\mathsf{Ber}\left(\frac{1}{2}-\theta\right)} \middle\| P_{X_1'+X_2+\ldots+\mathsf{Ber}\left(\frac{1}{2}-\theta\right)}\right), D_\alpha\left(P_{X_1+X_2+\ldots+\mathsf{Ber}\left(\frac{1}{2}+\theta\right)} \middle\| P_{X_1'+X_2+\ldots+\mathsf{Ber}\left(\frac{1}{2}+\theta\right)}\right)\right).$$

By repetitively applying the quasi-convexity and the extrema argument for $n-1$ times, we arrive at

$$D_\alpha\left(P_{X_1+X_2+\ldots+X_n} \middle\| P_{X_1'+X_2+\ldots+X_n}\right) \leq \max_{k\in[n-1]} D_\alpha\left(P_{X_1+N_k} \middle\| P_{X_1'+N_k}\right),$$

where $N_k \sim \mathsf{Binom}(k, \frac{1}{2}-\theta) + \mathsf{Binom}(n-k-1, \frac{1}{2}+\theta)$.

In the last step, by making use of the *joint* quasi-convexity of Rényi divergence (i.e., $(P, Q) \mapsto D_\alpha(P\|Q)$ is quasi-convex), it suffices to show that

$$\begin{aligned}
&\left(P_{X_1+N_k}, P_{X_1'+N_k}\right) = \lambda_1^*\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+N_k}, P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+N_k}\right)\\
&+\lambda_2^*\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+N_k}, P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+N_k}\right) + \lambda_3^*\Big(P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+N_k},\\
&P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+N_k}\Big) + \lambda_4^*\left(P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+N_k}, P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+N_k}\right)
\end{aligned}$$

for some $\lambda_i^* \in [0,1]$ and $\sum_{i=1}^4 \lambda_i^* = 1$. To this end, since $X_i \sim \mathsf{Ber}(p_i)$ and $X_i' \sim \mathsf{Ber}(p_i')$ for some $p_i, p_i' \in \left[\frac{1}{2}-\theta, \frac{1}{2}+\theta\right]$, we must have

$$\begin{aligned}
P_{X_1+N_k} &= \lambda P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+N_k} + (1-\lambda)P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+N_k}, \text{ and}\\
P_{X_1'+N_k} &= \lambda' P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+N_k} + (1-\lambda')P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+N_k}
\end{aligned}$$

for some $\lambda, \lambda' \in [0,1]$. Therefore, by setting $\lambda_1^* = \lambda\lambda'$, $\lambda_2^* = \lambda(1-\lambda')$, $\lambda_3^* = (1-\lambda)\lambda'$, and $\lambda_4^* = (1-\lambda)(1-\lambda')$, we arrive at the desired result.

D.3. Proof of Lemma 3.6

$$\begin{aligned}
&P_{\mathsf{Binom}\left(1+k^*,\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*-1,\frac{1}{2}+\theta\right)} = P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*-1,\frac{1}{2}+\theta\right)} \circ P_{\mathsf{Binom}\left(k^*,\frac{1}{2}-\theta\right)}\\
&P_{\mathsf{Binom}\left(k^*,\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*,\frac{1}{2}+\theta\right)} = P_{\mathsf{Binom}\left(n-k^*,\frac{1}{2}+\theta\right)} \circ P_{\mathsf{Binom}\left(k^*,\frac{1}{2}-\theta\right)},
\end{aligned}$$

and by a data processing inequality, we have

$$\begin{aligned}
&D_\alpha\left(P_{\mathsf{Binom}\left(1+k^*,\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*-1,\frac{1}{2}+\theta\right)} \middle\| P_{\mathsf{Binom}\left(k^*,\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*,\frac{1}{2}+\theta\right)}\right)\\
&\leq D_\alpha\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*-1,\frac{1}{2}+\theta\right)} \middle\| P_{\mathsf{Binom}\left(n-k^*,\frac{1}{2}+\theta\right)}\right)\\
&\leq D_\alpha\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n',\frac{1}{2}+\theta\right)} \middle\| P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+\mathsf{Binom}\left(n',\frac{1}{2}+\theta\right)}\right),
\end{aligned}$$

where in the last inequality we use a data processing inequality and define $n' = \lceil \frac{n-1}{2} \rceil$. Similarly, when $k^* > \frac{n}{2}$, we decompose

$$\begin{aligned}
&P_{\mathsf{Binom}\left(1+k^*, \frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*-1, \frac{1}{2}+\theta\right)} = P_{\mathsf{Binom}\left(1+k^*, \frac{1}{2}-\theta\right)} \circ P_{\mathsf{Binom}\left(n-k^*-1, \frac{1}{2}+\theta\right)}\\
&P_{\mathsf{Binom}\left(k^*, \frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*, \frac{1}{2}+\theta\right)} = P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+\mathsf{Binom}\left(k^*, \frac{1}{2}-\theta\right)} \circ P_{\mathsf{Binom}\left(n-k^*-1, \frac{1}{2}+\theta\right)}.
\end{aligned}$$

By a data processing inequality, we obtain

$$\begin{aligned}
&D_\alpha\left(P_{\mathsf{Binom}\left(1+k^*, \frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*-1, \frac{1}{2}+\theta\right)} \middle\| P_{\mathsf{Binom}\left(k^*, \frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k^*, \frac{1}{2}+\theta\right)}\right)\\
&\leq D_\alpha\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}-\theta\right)} \middle\| P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}-\theta\right)}\right).
\end{aligned}$$

Therefore we conclude that (4) can be upper bounded by

$$\begin{aligned}
&\max_{k\in[n-1]} D_\alpha\left(P_{\mathsf{Binom}\left(1+k, \frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k-1, \frac{1}{2}+\theta\right)} \middle\| P_{\mathsf{Binom}\left(k, \frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n-k, \frac{1}{2}+\theta\right)}\right)\\
&\leq \max\Bigg(\underbrace{D_\alpha\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}+\theta\right)} \middle\| P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}+\theta\right)}\right)}_{\text{(a)}}\Bigg),\\
&\qquad\quad \underbrace{D_\alpha\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}-\theta\right)} \middle\| P_{\mathsf{Ber}\left(\frac{1}{2}+\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}-\theta\right)}\right)}_{\text{(b)}}. \qquad (7)
\end{aligned}$$

D.4. Proof of Lemma 3.7

Using the inequality $t \leq e^{t-1}$, we have

$$\begin{aligned}
D_\alpha\left(P \| Q\right) &= \frac{1}{\alpha-1}\log\left(\mathbb{E}_Q\left[\left(\frac{dP}{dQ}(X)\right)^\alpha\right]\right)\\
&\leq \frac{1}{\alpha-1}\log\left(\mathbb{E}_Q\left[e^{\alpha\left(\frac{dP}{dQ}(X)-1\right)}\right]\right)\\
&\leq \frac{1}{\alpha-1}\log\left(\mathbb{E}_Q\left[e^{\alpha\left(\frac{dP}{dQ}(X)-1\right)}\right]\right)\\
&\overset{\text{(a)}}{\leq} \frac{1}{\alpha-1}\log\left(e^{C_0\alpha^2\left\|\frac{dP}{dQ}(X)-1\right\|^2_{\psi_2}}\right)\\
&= C_0\frac{\alpha^2}{\alpha-1}\left\|\frac{dP}{dQ}(X)-1\right\|^2_{\psi_2},
\end{aligned}$$

where (a) holds for any sub-gaussian random variable (see, for instance, Vershynin (2018, Proposition 2.5.2), which states that for a zero-mean random variable Z with finite sub-gaussian norm, $\mathbb{E}\left[e^{\alpha Z}\right] \leq e^{C_0\alpha^2\|Z\|^2_{\psi_2}}$).

D.5. Proof of Lemma 3.8

Bounding (a) For notational simplicity, we denote the LR of term (a) in (7) as $R(i)$ for $i \in [n'+1]$ and bound it as follows.

$$\begin{aligned}
R(i) &\triangleq \frac{P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n',\frac{1}{2}+\theta\right)}(i)}{P_{\mathsf{Binom}\left(n'+1,\frac{1}{2}+\theta\right)}(i)}\\
&= \frac{\binom{n'}{i}\left(\frac{1}{2}+\theta\right)^{i+1}\left(\frac{1}{2}-\theta\right)^{n'-i}}{\binom{n'+1}{i}\left(\frac{1}{2}+\theta\right)^{i}\left(\frac{1}{2}-\theta\right)^{n'+1-i}} + \frac{\binom{n'}{i-1}\left(\frac{1}{2}+\theta\right)^{i}\left(\frac{1}{2}-\theta\right)^{n'-i+1}}{\binom{n'+1}{i}\left(\frac{1}{2}+\theta\right)^{i}\left(\frac{1}{2}-\theta\right)^{n'+1-i}}\\
&= \left(\frac{n'+1-i}{n'+1}\right)\left(\frac{1-2\theta}{1+2\theta}\right) + \left(\frac{i}{n'+1}\right)\left(\frac{1+2\theta}{1-2\theta}\right)\\
&= \left(\frac{1-2\theta}{1+2\theta}\right) + \frac{i}{n'+1}\left(\left(\frac{1+2\theta}{1-2\theta}\right) - \left(\frac{1-2\theta}{1+2\theta}\right)\right)\\
&= \left(\frac{1-2\theta}{1+2\theta}\right) + \frac{i}{n'+1}\left(\frac{8\theta}{1-4\theta^2}\right).
\end{aligned}$$

When $I \sim P_{\mathsf{Binom}\left(n'+1,\frac{1}{2}+\theta\right)}$, $I - \mathbb{E}[I]$ has a sub-gaussian norm $\|I-\mathbb{E}[I]\|_{\psi_2}^2 \leq \sigma_0^2(n'+1)$ for some universal constant σ_0 (since I is sum of $n'+1$ independent binary random variables). Also notice that $\mathbb{E}[R(I)] = 1$, so $R(I)-1$, which is a linear function of I, can be written as

$$R(I) - 1 = \frac{8\theta}{(n'+1)(1-4\theta^2)}\left(I - \mathbb{E}[I]\right).$$

Therefore, $R(I)-1$ has a sub-gaussian norm bounded by

$$\|R(I)-1\|_{\psi_2}^2 \leq C_1 \frac{\theta^2}{(1-4\theta^2)^2(n'+1)}.$$

By Lemma 3.7, we conclude that term (a) in (7) can be controlled by

$$\begin{aligned}
&D_\alpha\left(P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n',\frac{1}{2}+\theta\right)}\middle\| P_{\mathsf{Binom}\left(n'+1,\frac{1}{2}+\theta\right)}\right)\\
&\leq C_2\frac{\theta^2}{(1-4\theta^2)^2}\frac{\alpha^2}{\alpha-1}\frac{1}{n'+1},
\end{aligned} \tag{8}$$

for some constant $C_2 > 0$.

Bounding (b) Similarly, let us denote the LR of term (b) in (7) as $R'(i) = \frac{1}{R(i)}$ for $i \in [n'+1]$. Let

$$I \sim P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n',\frac{1}{2}+\theta\right)}$$

and we control the sub-gaussian norm of $R'(I)-1$ as follows:

$$R'(I) - 1 = \frac{\mathbb{E}[R(I)] - R(I)}{R(I)} + \frac{1-\mathbb{E}[R(I)]}{R(I)}. \tag{9}$$

For the first term in the right hand side of (9), observe that

$$\begin{aligned}
\frac{\mathbb{E}[R(I)] - R(I)}{R(I)} &\leq \frac{|\mathbb{E}[R(I)] - R(I)|}{|R(I)|}\\
&\leq \left(\frac{1+2\theta}{1-2\theta}\right)|\mathbb{E}[R(I)] - R(I)|
\end{aligned} \tag{10}$$

where the last inequality holds since $R(I) \geq \left(\frac{1+2\theta}{1-2\theta}\right)$ almost surely. Therefore

$$\begin{aligned}
&\left\| \frac{\mathbb{E}[R(I)] - R(I)}{R(I)} \right\|_{\psi_2}^2 \\
&\leq \left(\frac{1+2\theta}{1-2\theta}\right)^2 \left\| R(I) - \mathbb{E}[R(I)] \right\|_{\psi_2}^2 \\
&\leq C_3 \left(\frac{1+2\theta}{1-2\theta}\right)^2 \frac{\theta^2}{(1-4\theta^2)^2} \frac{\alpha^2}{\alpha-1} \frac{1}{n'+1} \\
&= C_3 \frac{\theta^2}{(1-2\theta)^4} \frac{1}{n'+1}, \qquad (11)
\end{aligned}$$

where the second inequality is due to the fact that $I - \mathbb{E}[I]$ is sum of $n'+1$ zero-mean bounded random variables.

Next, the second term in (9) can be controlled by

$$\begin{aligned}
&\left| \frac{1 - \mathbb{E}[R(I)]}{R(I)} \right| \leq \left(\frac{1+2\theta}{1-2\theta}\right) \left| R(I) - 1 \right| \\
&= \left(\frac{1+2\theta}{1-2\theta}\right) \left| \left(\frac{1-2\theta}{1+2\theta}\right) + \frac{\mathbb{E}[I]}{n'+1} \left(\frac{8\theta}{1-4\theta^2}\right) - 1 \right| \\
&= \left(\frac{1+2\theta}{1-2\theta}\right) \left(\frac{8\theta^2}{(1-4\theta^2)(n'+1)} \right) \\
&= \frac{8\theta^2}{(1-2\theta)^2 (n'+1)}, \qquad (12)
\end{aligned}$$

where the second equality holds since

$$\mathbb{E}[I] = n' \left(1 + \frac{\theta}{2}\right) + \left(1 - \frac{\theta}{2}\right).$$

Combining (11) and (12), we obtain an upper bound on (9):

$$\begin{aligned}
&\left\| R'(I) - 1 \right\|_{\psi_2}^2 \\
&\leq 2 \left\| \frac{\mathbb{E}[R(I)] - R(I)}{R(I)} \right\|_{\psi_2}^2 + 2 \left\| \frac{8\theta^2}{(1-2\theta)^2 (n'+1)} \right\|_{\psi_2}^2 \\
&\leq 2C_3 \frac{\theta^2}{(1-2\theta)^4} \frac{1}{n'+1} + C_4 \left(\frac{8\theta^2}{(1-2\theta)^2 (n'+1)} \right)^2 \\
&\leq C_5 \frac{\theta^2}{(1-2\theta)^4} \frac{1}{n'+1},
\end{aligned}$$

for some $C_5 > 0$. Also notice that $\mathbb{E}[R'(I) - 1] = 0$ when $I \sim P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}+\theta\right)}$. Therefore, applying Lemma 3.7, we conclude that term (b) in (7) can be controlled by

$$\begin{aligned}
&D_\alpha \left(P_{\mathsf{Binom}\left(n'+1, \frac{1}{2}+\theta\right)} \middle\| P_{\mathsf{Ber}\left(\frac{1}{2}-\theta\right)+\mathsf{Binom}\left(n', \frac{1}{2}+\theta\right)} \right) \\
&\leq C_6 \frac{\theta^2}{(1-2\theta)^4} \frac{\alpha^2}{\alpha-1} \frac{1}{n'+1}, \qquad (13)
\end{aligned}$$

for some $C_6 > 0$.