Privacy Amplification via Compression: Achieving the Optimal Privacy-Accuracy-Communication Trade-off in Distributed Mean Estimation

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

Abstract

Privacy and communication constraints are two major bottlenecks in federated learning (FL) and analytics (FA). We study the optimal accuracy of mean and frequency estimation for FL and FA respectively under joint communication and (ε, δ)-differential privacy (DP) constraints. We consider both the central and the multi-message shuffling DP models. We show that in order to achieve the optimal ℓ_2 error under (ε, δ)-DP, it is sufficient for each client to send $\Theta\left(n \min\left(\varepsilon, \varepsilon^2\right)\right)$ bits for FL and $\Theta\left(\log\left(n \min\left(\varepsilon, \varepsilon^2\right)\right)\right)$ bits for FA to the server, where n is the number of clients.

1. Introduction

In the basic setting of federated learning (FL) (McMahan et al., 2016; Konečný et al., 2016; Kairouz et al., 2021b) and analytics (FA), a server wants to execute a specific learning or analytics task on raw data that is kept on clients' devices. Consider, for example, model updates in FL or histogram estimation in FA, both of which can be modeled as a distributed mean estimation problem. Clients communicate targeted messages to the server and the privacy of the users' data is ensured (in terms of explicit differential privacy (DP) (Dwork et al., 2006) guarantees) by adding carefully calibrated noise to the computed mean at the server before releasing it to downstream modules (e.g., the server computes the average model update and corrupts it with the addition of noise). This is called the trusted server or central DP model, as it entrusts the central server with privatization and is one of the most common ways in which federated learning and analytics are implemented today [1].

In this paper, we ask the following question: given that the server needs to privatize the mean, can the clients communicate "less information" to the server? More precisely, can we leverage the fact that the server only needs to output a noisy (approximate) estimate of the mean to reduce the communication load without sacrificing accuracy? In recent years, there has been significant interest in the central DP model (Abadi et al., 2016) as well as communication efficiency and privacy for FL and FA under different models, including local DP (Warner, 1965; Kasiviswanathan et al., 2011; Kairouz et al., 2016; Ye & Barg, 2017; Barnes et al., 2019; Acharya et al., 2019c; Barnes et al., 2020a; Chen et al., 2020), shuffle (Erlingsson et al., 2019; Feldman et al., 2022a) and distributed DP (Agarwal et al., 2018; Kairouz et al., 2021a; Agarwal et al., 2021; Chen et al., 2022b;c); however, this basic question has remained open.

One natural way to reduce communication is to have clients communicate only partial information about their samples. For example, in the case of model updates, each client can update only a subset of the model coefficients. In histogram estimation, information about a client's sample can be "split" into multiple parts, and the client can communicate only a part. However, this results in less information at the server, or effectively fewer samples to estimate the target quantity, e.g., each model coefficient is now updated only by a subset of the clients. A quick calculation reveals that this increases the sensitivity of the estimate to each user's sample and therefore requires the addition of larger noise at the server to achieve the same privacy level. Hence reducing communication reduces accuracy for the same privacy guarantee.

We circumvent this challenge with a simple but insightful observation: when each client communicates only partial information about its sample, we can amplify privacy by randomly selecting the part contributed by each client. This random selection is hidden from a downstream module which has only access to the estimate revealed by the server, which leads to privacy amplification. Privacy amplification by subsampling has been studied in (Li et al., 2012; Balle et al., 2018) but usually refers to the selection of a random subset of the clients (from a larger pool of available clients). In our case, it is the "piece of information" that is randomly selected at each client.

This naturally leads to a follow-up question: can we leverage privacy amplification via compression and achieve the same three-way trade-off by using secure aggregation (Chen et al., 2022b) and shuffling (Erlingsson et al., 2019) type models which hide information from the server? For se-

[1]We assume a trusted service provider who applies the DP mechanism faithfully. This can be enforced by implementing the DP mechanism inside of a remotely attestable trusted execution environment (Allen et al., 2019).

cure aggregation the three-way trade-off has been studied in (Chen et al., 2022a) and the communication cost is significantly larger than the communication cost for central DP proved in this paper (see Table 1). For shuffling, the optimal communication cost has been posed as an open problem in (Chen et al., 2022a). We resolve this problem by showing that the optimal central DP trade-off can also be achieved with a multi-message shuffling scheme establishing the optimal communication cost. As before, our scheme leverages a privacy amplification gain.

Our contributions. We study distributed mean and frequency estimation for FL and FA[2], under both the central DP and the multi-message shuffling models. We characterize the order-optimal privacy-accuracy-communication trade-offs for mean estimation and provide an achievable scheme for frequency estimation (in Appendix C) under the central DP model. Our results reveal that privacy and communication efficiency can be achieved simultaneously with no additional penalty on accuracy. In particular, we show that $\tilde{O}\left(n\min\left(\varepsilon,\varepsilon^2\right)\right)$ and $\tilde{O}\left(\log\left(n\min\left(\varepsilon,\varepsilon^2\right)\right)\right)$ bits of (per-client) communication are sufficient to achieve the order-optimal error under (ε,δ)-privacy for mean and frequency estimation respectively, where n is the number of participating clients. Without compression, each client needs $O(d)$ bits and $\log d$ bits for the mean and frequency estimation problems respectively (where d is the number of trainable parameters in FL or the domain size in FA), which means that we can get significant savings in the regime $n\varepsilon^2 = o(d)$ (assuming $\varepsilon = O(1)$). We note that this is often the relevant regime not only for cross-silo but also for cross-device FL/FA. For instance, in practical FL, d usually ranges from 10^6–10^9, and n, the *per-epoch* sample size, is usually much smaller (e.g., of the order of 10^3–10^5). For mean estimation, we show that the central DP trade-off can also be achieved with a multi-message shuffling scheme (within a $\log d$ factor in communication cost).

We summarize the comparisons of our main results to local and distributed DP in Table 1.

2. Problem Formulation

Consider n clients each with local data $x_i \in \mathbb{R}^d$ that satisfies $\|x_i\|_2 \leq C$ for some constant $C > 0$ (one can think of x_i as a clipped local gradient). A server wants to learn an estimate $\hat{\mu}$ of the mean $\mu(x^n) \triangleq \frac{1}{n}\sum_i x_i$ from $x^n = (x_1, \ldots, x_n)$ after communicating with the n clients. Toward this end, each client locally compresses x_i into a b-bit message $Y_i = \mathsf{enc}_i\left(x_i\right) \in \mathcal{Y}$ through a local encoder $\mathsf{enc}_i : \mathcal{X} \mapsto \mathcal{Y}$ (where $|\mathcal{Y}| \leq 2^b$ and sends it to the central server, which upon receiving $Y^n = (Y_1, \ldots, Y_n)$ computes an estimate $\hat{\mu} = \mathsf{dec}\left(Y^n\right)$ that satisfies the following differential privacy:

[2]Due to the space constraint, we leave the analysis of our frequency estimation scheme into the appendix

Definition 2.1 (Differential Privacy)**.** The mechanism $\hat{\mu}$ is (ε,δ)-differentially private if for any neighboring datasets $x^n := (x_1, ..., x_i, ..., x_n)$, $x'^n := (x_1, ..., x'_i, ..., x_n)$, and measurable $\mathcal{S} \subseteq \mathcal{Y}$,

$$\Pr\left\{\hat{\mu} \in \mathcal{S} | x^n\right\} \leq e^{\varepsilon} \cdot \Pr\left\{\hat{\mu} \in \mathcal{S} | x'^n\right\} + \delta,$$

where the probability is taken over the randomness of $\hat{\mu}$.

Our goal is to minimize the ℓ_2^2 estimation error:

$$\min_{(\mathsf{enc}_i,\mathsf{dec})} \max_{x^n} \mathbb{E}\left[\left\|\hat{\mu}\left(\mathsf{enc}_1(x_1), ..., \mathsf{enc}_n(x_n)\right) - \mu(x^n)\right\|_2^2\right],$$

subject to b-bit communication and (ε,δ)-DP constraints.

3. Related Works

Distributed mean estimation. In this work, we consider the distributed mean estimation under a *central*-DP setting where the server is trusted, which is different from the local DP model (Kasiviswanathan et al., 2011; Duchi et al., 2013; Nguyên et al., 2016; Wang et al., 2019; Bhowmick et al., 2018; Chen et al., 2020) and the distributed DP model with secure aggregation (Bonawitz et al., 2016; Bell et al., 2020; Kairouz et al., 2021a; Agarwal et al., 2021; Chen et al., 2022b;c).

A key step in our mean estimation scheme is pre-processing the local data via Kashin's representation (Lyubarskii & Vershynin, 2010). While various compression schemes, based on quantization, sparsification, and dithering have been proposed in the recent literature, Kashin's representation has also been explored in a few works for communication efficiency (Fuchs, 2011; Studer et al., 2012; Caldas et al., 2018; Safaryan et al., 2020) and for LDP (Feldman et al., 2017) and is particularly powerful in the case of joint communication and privacy constraints as it helps spread the information in a vector evenly in every dimension.

Distributed frequency estimation. Distributed frequency estimation (a.k.a. histogram estimation) is another canonical task that has been heavily studied under a distributed setting with DP. Prior works either focus on 1) the local DP model with or without communication constraints, e.g., (Bassily & Smith, 2015; Bassily et al., 2017; Bun et al., 2018; 2019; Huang et al., 2022; Ye & Barg, 2017; Wang et al., 2019; Acharya et al., 2019c; Chen et al., 2020; Feldman & Talwar, 2021; Shah et al., 2022; Feldman et al., 2022b), or 2) the central DP model *without* communication constraints (Dwork et al., 2006; Ghosh et al., 2012; Korolova et al., 2009; Bun & Steinke, 2016; Balcer & Vadhan, 2017; Zhu et al., 2020; Cormode & Bharadwaj, 2022). In this work, we consider central DP but with explicit communication constraints.

	Communication (bits)	ℓ_2 error
Local DP (Chen et al., 2020; Feldman et al., 2017)	$\Theta\left(\lceil\varepsilon\rceil\right)$	$\Theta\left(\frac{d}{n\min(\varepsilon^2,\varepsilon)}\right)$
Distributed DP (with SecAgg) (Chen et al., 2022b)	$\tilde{O}\left(n^2\min\left(\varepsilon,\varepsilon^2\right)\right)$	$\Theta\left(\frac{d}{n^2\min(\varepsilon^2,\varepsilon)}\right)$
Central DP (Theorem 4.3)	$\tilde{O}\left(n\min\left(\varepsilon,\varepsilon^2\right)\right)$	$O\left(\frac{d\log d}{n^2\min(\varepsilon^2,\varepsilon)}\right)$
Shuffle DP (Theorem E.3)	$\tilde{O}\left(n\log(d)\min\left(\varepsilon,\varepsilon^2\right)\right)$	$O\left(\frac{d}{n^2\min(\varepsilon^2,\varepsilon)}\right)$

Table 1. Comparison of the communication costs of ℓ_2 mean estimation under local, distributed, central, and shuffle DP.

4. Main Results

In this section, we present mean estimation schemes that achieves the optimal $\tilde{O}_\delta\left(\frac{C^2d}{n^2\varepsilon^2}\right)$ error under (ε,δ)-DP while only using $\tilde{O}(n\varepsilon^2)$ bits of per-client communication.

We first consider a discrete setting with ℓ_∞ geometry: assume each client observes $x_i\in\{-c,c\}^d$ where $c>0$ is a constant, and a central server aims to estimate the mean $\mu\left(x^n\right):=\frac{1}{n}\sum_{i=1}^n x_i$ by minimizing the ℓ_2^2 error subject to the privacy and communication constraints. We argue later that solutions to the above ℓ_∞ problem can be used for ℓ_2 mean estimation by applying Kashin's representation.

To solve the aforementioned ℓ_∞ mean estimation problem, first observe that each client's local data can be expressed in d bits since each coordinate of x_i can only take values in $\{c,-c\}$. To reduce the communication load to $o(d)$ bits, each client adopts the following subsampling strategy: for each coordinate $j\in[d]$, client i chooses to send $x_i(j)$ to the server with probability γ. We assume that this subsampling step is performed with a seed shared by the client and the server[3], hence the server knows which coordinates are communicated by each client. Therefore upon receiving the client messages, it can compute the mean of each coordinate and privatize it by adding Gaussian noise. The key observation we leverage is that the randomness in the compression algorithm can be used to amplify privacy or equivalently reduce the magnitude of the Gaussian noise that is needed for privatization. Note that such randomness needs to be kept private from an adversary as the privacy guarantee of the scheme relies on it.

For the ℓ_2 mean estimation task formulated in Section 2, we pre-process local vectors by first computing their Kashin's representations and then performing randomized rounding (Kashin, 1977; Vershynin, 2018; Feldman et al., 2017; Chen et al., 2020). We leave the details to Appendix B.2. By combining Kashin's representation with the above sampling technique, we arrive at the following theorem:

Theorem 4.1 (ℓ_2 mean estimation)**.** *Let $x_1,...,x_n\in\mathcal{B}_2(C)$ (i.e., $\left\|x_i\right\|_2\le C$ for all $i\in[n]$).*

Then for any $\varepsilon,\delta>0$, Algorithm 1 combined with Kashin's representation and randomized rounding yields an (ε,δ)-DP unbiased estimator with ℓ_2^2 estimation error bounded by

$$O\left(\frac{dC^2}{nb}+\frac{C^2d^2\log(1/\delta)}{n^2b^2}+\frac{C^2d(\log(d/\delta)+\varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right).$$

Remark 4.2 (Unbiasedness). In mean estimation, we usually want the final mean estimator to be unbiased since standard convergence analyses of SGD (Ghadimi & Lan, 2013) require an unbiased estimate of the true gradient in each optimization round.

In Theorem 4.1, if we ignore the poly-logarithmic terms and assume $\varepsilon=O(1)$, the privatization error can be simplified to $\tilde{O}\left(\frac{dC^2}{n^2\varepsilon^2}\right)$, which dominates the total ℓ_2^2 error when $b=\tilde{\Omega}_\delta\left(\max\left(n\varepsilon^2,\sqrt{d}\varepsilon\right)\right)$.

4.1. Dimension-free communication cost

Next, we introduce a modification to the above scheme to remove the dependence on the dimension d in the communication cost $b=\tilde{\Omega}_\delta\left(\max\left(n\varepsilon^2,\sqrt{d}\varepsilon\right)\right)$ from the previous section, particularly in the *small-sample* regime $n\varepsilon^2=o(\sqrt{d}\varepsilon)$. We show that in this regime the performance of the scheme can be improved by a priori restricting the server's attention to a subset of the coordinates.

We make the following modification to the above scheme: before performing Algorithm 1, the server randomly selects $d'\approx O\left(\min(d,n^2\varepsilon^2)\right)$ coordinates and only requires clients to run Algorithm 1 on them. We present the modified scheme in Algorithm 2 in Appendix B.1 and summarize its performance in Theorem 4.3.

Theorem 4.3 (ℓ_2 mean estimation.)**.** *Let $x_1,...,x_n\in\mathcal{B}_2(C)$ (i.e., $\left\|x_i\right\|_2\le C$ for all $i\in[n]$), $d'=\min\left(d,nb,\frac{n^2\varepsilon^2}{(\log(1/\delta)+\varepsilon)\log(d/\delta)}\right)$.*

Then for any $\varepsilon,\delta>0$, Algorithm 2 is (ε,δ)-DP. In addition, the (average) per-client communication cost is $\gamma d=b$ bits,

[3] In practice, such randomness can be agreed by both sides ahead of time, or it can be generated by the server and communicated to each client.

and the ℓ_2^2 estimation error is at most

$$O\left(\max\left(\frac{C^2 d\log(d/\delta)}{nb}, \frac{C^2 d\log(d/\delta)(\log(1/\delta)+\varepsilon)}{n^2\varepsilon^2}\right)\right). \tag{1}$$

The above theorem implies that when $\varepsilon = O(1)$, $b = \tilde{\Omega}\left(n\varepsilon^2\right)$ bits per client are sufficient to achieve the order-optimal $\tilde{O}_\delta\left(\frac{c^2 d}{n^2\varepsilon^2}\right)$ error (even in the small sample regime $n \leq \sqrt{d}$), i.e. the communication cost of the scheme is independent of the dimension d.

4.2. Achieving the Optimal Trade-off via Shuffling

So far, we see that the communication cost can be reduced to ($\tilde{O}\left(n\varepsilon^2\right)$ for mean estimation while still achieving the order-wise optimal error, as long as the server is *trusted*. On the other hand, when the server is untrusted, (Chen et al., 2022b;a) show that optimal error under (ε, δ)-DP can be achieved with secure aggregation at a much higher communication cost ($\tilde{O}\left(n^2\varepsilon^2\right)$ bits per client). In this section, we show that the optimal communication-accuracy-privacy trade-off from the previous sections can be achieved if there exists a *secure* shuffler that randomly permutes clients' locally privatized messages and releases them to the server, even if the server is untrusted. We note that a similar result has been proven in a concurrent work (Girgis & Diggavi, 2023).

Our scheme makes use of a specific communication efficient LDP scheme SQKR (Chen et al., 2020) and amplifies the local DP via shuffling with the amplification lemma Feldman et al. (2022a). However, unlike in their result, we make use of multi-message shuffling lemma to achieve the optimal accuracy in *all* privacy regimes.

Privacy analysis. By making use of the amplification lemma (Feldman et al., 2022a) (see Appendix E for details), we design the local randomizers $\mathcal{M}_i$ that satisfy ε_0-LDP. Note that the amplification lemma is only tight when $\varepsilon_0 = O(1)$, thus restricting the (amplified) central $\varepsilon = O(1/\sqrt{n})$. To accommodate larger ε, users can send different portions of their messages to the server in separate shuffling rounds. Equivalently, we repeat the shuffled LDP mechanism for $T = O\left(\lceil n\varepsilon^2\rceil\right)$ rounds while ensuring that in each round clients communicate an independent piece of information about their sample to the server. More precisely, within each round, each client applies the local randomizers $\mathcal{M}_i$ with a per-round *local privacy budget* $\varepsilon_0 = O(1)$ and sends an independent message to the server. This results in (amplified) central $O(1/\sqrt{n})$-DP per round, which after composition over $T = O\left(\lceil n\varepsilon^2\rceil\right)$ rounds leads to ε-DP for the overall scheme as suggested by the composition theorem (Kairouz et al., 2016)). We detail the algorithm in Algorithm 4 in Appendix E.1.

We summarize the performance guarantee for the overall scheme in the following theorem.

Theorem 4.4 (ℓ_2 mean estimation)**.** *Let $x_1, ..., x_n \in \mathcal{B}_2(C)$ (i.e., $\|x_i\|_2 \leq C$ for all $i \in [n]$). For all $\varepsilon > 0, b > 0, n > 30$, and $\delta \in (\delta_{\min}, 1]$ where $\delta_{\min} = O\left(be^{-n}/\log d\right)$, There exsists a (ε, δ)-DP (given in Algorithm 4), uses no more than b bits of communication, and achieves*

$$\mathbb{E}\left[\left\|\mu\left(x^n\right) - \hat{\mu}\left(x^n\right)\right\|_2^2\right] = O\left(C^2 d\max\left(\frac{\log(d)}{nb}, \frac{\log(b/\delta)(\log(1/\delta)+\varepsilon)}{n^2\varepsilon^2}\right)\right).$$

4.3. Lower bounds

The estimation error in Theorem 4.3 and Theorem E.3 is optimal up to an $\log\left(d/\delta\right)$ factor. Specifically, Theorem 5.3 of (Chen et al., 2022a) shows that any b-bit *unbiased* compression scheme will incur $\Omega\left(\frac{C^2 d}{nb}\right)$ error for the ℓ_2 mean estimation problem (even when privacy is not required). This matches the first term in (1) up to a logarithmic factor.

On the other hand, the centralized Gaussian mechanism (under a central (ε, δ)-DP) achieves $O\left(\frac{C^2 d\log(1/\delta)}{n^2\varepsilon^2}\right)$ MSE (Balle & Wang, 2018) (which is order-optimal in most parameter regimes(Canonne et al., 2020)). Hence, the total communication received by the server has to be at least $\Omega(n^2\varepsilon^2)$ bits in order to achieve the same error. Therefore, the (average) per-client communication cost has to be at least $\Omega(n\varepsilon^2)$ bits.

4.4. Experiments

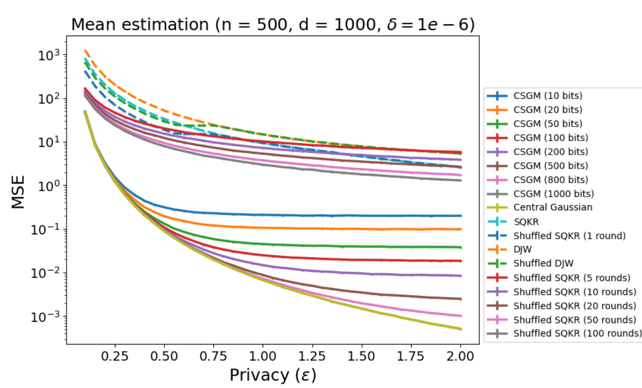


Figure 1. We compare the MSE of CSGM (Theorem 4.3) and shuffled SQKR (Theorem E.3) with other central and local DP schemes. Although both CSGM and shuffled SQKR are order-optimal, the pre-constants of CSGM are significantly lower. On the other hand, multi-round shuffling can improve the accuracy on the single-round ones. More experiments can be found in Appendix F.

5. Acknowledgements

This work was supported in part by NSF Award # CCF-2213223 and a Stanford Graduate Fellowship.

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.

Acharya, J. and Sun, Z. Communication complexity in locally private distribution estimation and heavy hitters. In *International Conference on Machine Learning*, pp. 51–60, 2019.

Acharya, J., Canonne, C. L., and Tyagi, H. Inference under information constraints: Lower bounds from chi-square contraction. In *Conference on Learning Theory*, pp. 3–17. PMLR, 2019a.

Acharya, J., Canonne, C. L., and Tyagi, H. Inference under information constraints ii: Communication constraints and shared randomness. *arXiv preprint arXiv:1905.08302*, 2019b.

Acharya, J., Sun, Z., and Zhang, H. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1120–1129, 2019c.

Acharya, J., Canonne, C. L., and Tyagi, H. General lower bounds for interactive high-dimensional estimation under information constraints. *arXiv preprint arXiv:2010.06562*, 2020.

Acharya, J., Sun, Z., and Zhang, H. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pp. 48–78. PMLR, 2021.

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: 5052–5064, 2021.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In *Advances in Neural Information Processing Systems 30*, pp. 1709–1720, 2017.

Allen, J., Ding, B., Kulkarni, J., Nori, H., Ohrimenko, O., and Yekhanin, S. An algorithmic framework for differentially private data analysis on trusted processors. *Advances in Neural Information Processing Systems*, 32, 2019.

Balcer, V. and Cheu, A. Separating local & shuffled differential privacy via histograms. *arXiv preprint arXiv:1911.06879*, 2019.

Balcer, V. and Vadhan, S. Differential privacy on finite computers. *arXiv preprint arXiv:1709.05396*, 2017.

Balle, B. and Wang, Y.-X. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pp. 394–403. PMLR, 2018.

Balle, B., Barthe, G., and Gaboardi, M. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems*, 31, 2018.

Barnes, L. P., Han, Y., and Ozgur, A. Lower bounds for learning distributions under communication constraints via fisher information, 2019.

Barnes, L. P., Chen, W.-N., and Ozgur, A. Fisher information under local differential privacy. *arXiv preprint arXiv:2005.10783*, 2020a.

Barnes, L. P., Inan, H. A., Isik, B., and Ozgur, A. rtop-k: A statistical estimation approach to distributed sgd, 2020b.

Bassily, R. and Smith, A. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pp. 127–135, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746632. URL https://doi.org/10.1145/2746539.2746632.

Bassily, R., Nissim, K., Stemmer, U., and Thakurta, A. Practical locally private heavy hitters. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 2285–2293, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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.

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 federated learning on user-held data. *arXiv preprint arXiv:1611.04482*, 2016.

Braverman, M., Garg, A., Ma, T., Nguyen, H. L., and Woodruff, D. P. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1011–1020, 2016.

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

Bun, M., Nelson, J., and Stemmer, U. Heavy hitters and the structure of local privacy. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, SIGMOD/PODS '18, pp. 435–447, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450347068. doi: 10.1145/3196959.3196981. URL https://doi.org/10.1145/3196959.3196981.

Bun, M., Nelson, J., and Stemmer, U. Heavy hitters and the structure of local privacy. *ACM Transactions on Algorithms (TALG)*, 15(4):1–40, 2019.

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.

Chen, W.-N., Kairouz, P., and Ozgur, A. Breaking the communication-privacy-accuracy trilemma. *Advances in Neural Information Processing Systems*, 33, 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., Özgür, A., Cormode, G., and Baharadwaj, A. The communication cost of security and privacy in federated frequency estimation. *in submission*, 2022b.

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, 2022c.

Cheu, A., Smith, A., Ullman, J., Zeber, D., and Zhilyaev, M. Distributed differential privacy via shuffling. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 375–403. Springer, 2019.

Cormode, G. and Bharadwaj, A. Sample-and-threshold differential privacy: Histograms and applications. In *International Conference on Artificial Intelligence and Statistics*, pp. 1420–1431. PMLR, 2022.

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., 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.

Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2468–2479. SIAM, 2019.

Farokhi, F. Gradient sparsification can improve performance of differentially-private convex machine learning. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 1695–1700. IEEE, 2021.

Feldman, V. and Talwar, K. Lossless compression of efficient private local randomizers. *arXiv preprint arXiv:2102.12099*, 2021.

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.

Feldman, V., McMillan, A., and Talwar, K. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 954–964. IEEE, 2022a.

Feldman, V., Nelson, J., Nguyen, H., and Talwar, K. Private frequency estimation via projective geometry. In *International Conference on Machine Learning*, pp. 6418–6433. PMLR, 2022b.

Feldman, V., McMillan, A., and Talwar, K. Stronger privacy amplification by shuffling for rényi and approximate differential privacy. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4966–4981. SIAM, 2023.

Fuchs, J.-J. Spread representations. In *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pp. 814–817. IEEE, 2011.

Gandikota, V., Kane, D., Maity, R. K., and Mazumdar, A. vqsgd: Vector quantized stochastic gradient descent, 2019.

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

Ghazi, B., Golowich, N., Kumar, R., Pagh, R., and Velingker, A. On the power of multiple anonymous messages. *arXiv preprint arXiv:1908.11358*, 2019.

Ghazi, B., Kumar, R., Manurangsi, P., and Pagh, R. Private counting from anonymous messages: Near-optimal accuracy with vanishing communication overhead. In *International Conference on Machine Learning*, pp. 3505–3514. PMLR, 2020.

Ghosh, A., Roughgarden, T., and Sundararajan, M. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

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.

Girgis, A. M. and Diggavi, S. Multi-message shuffled privacy in federated learning. *arXiv preprint arXiv:2302.11152*, 2023.

Han, Y., Özgür, A., and Weissman, T. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory*, pp. 3163–3188. PMLR, 2018.

Hu, R., Gong, Y., and Guo, Y. Federated learning with sparsification-amplified privacy and adaptive optimization. *arXiv preprint arXiv:2008.01558*, 2020.

Huang, Z., Qiu, Y., Yi, K., and Cormode, G. Frequency estimation under multiparty differential privacy: One-shot and streaming. *Proc. VLDB Endow.*, 15 (10):2058–2070, jun 2022. doi: 10.14778/3547305.3547312. URL https://doi.org/10.14778/3547305.3547312.

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*, 2021a.

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. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021b.

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.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Korolova, A., Kenthapadi, K., Mishra, N., and Ntoulas, A. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, pp. 171–180, 2009.

Li, N., Qardaji, W., and Su, D. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pp. 32–33, 2012.

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.

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

Nguyên, T. T., Xiao, X., Yang, Y., Hui, S. C., Shin, H., and Shin, J. Collecting and analyzing data from smart device users with local differential privacy, 2016.

Safaryan, M., Shulgin, E., and Richtárik, P. Uncertainty principle for communication compression in distributed and federated learning and the search for an optimal compressor. *arXiv preprint arXiv:2002.08958*, 2020.

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.

Studer, C., Yin, W., and Baraniuk, R. G. Signal representations with minimum ℓ_∞-norm. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1270–1277. IEEE, 2012.

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.

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

Wang, T., Zhao, J., Yang, X., and Ren, X. Locally differentially private data collection and analysis. *arXiv preprint arXiv:1906.01777*, 2019.

Wangni, J., Wang, J., Liu, J., and Zhang, T. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 1299–1309, 2018.

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

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in neural information processing systems*, pp. 1509–1519, 2017.

Ye, M. and Barg, A. Optimal schemes for discrete distribution estimation under local differential privacy. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 759–763, June 2017. doi: 10.1109/ISIT.2017.8006630.

Zhu, W., Kairouz, P., McMahan, B., Sun, H., and Li, W. Federated heavy hitters discovery with differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 3837–3847. PMLR, 2020.

Zhu, Y. and Wang, Y.-X. Poission subsampled rényi differential privacy. In *International Conference on Machine Learning*, pp. 7634–7642. PMLR, 2019.

A. More Relevant Works

Federated learning and distributed mean estimation. Federated learning (Konečný et al., 2016; McMahan et al., 2016; Kairouz et al., 2019) emerges as a decentralized machine learning framework that provides data confidentiality by retaining clients' raw data on edge devices. In FL, communication between clients and the central server can quickly become a bottleneck (McMahan et al., 2016), so previous works have focused on compressing local model updates via gradient quantization (McMahan et al., 2016; Alistarh et al., 2017; Gandikota et al., 2019; Suresh et al., 2017; Wen et al., 2017; Wangni et al., 2018; Braverman et al., 2016), sparsification (Barnes et al., 2020b; Hu et al., 2020; Farokhi, 2021). To further enhance data security, FL is often combined with differential privacy (Dwork et al., 2006; Abadi et al., 2016; Agarwal et al., 2018). Among these works, (Hu et al., 2020) also employs gradient sparsification (or gradient subsampling) to reduce the problem dimensionality. However, the sparsification takes place *after* the aggregation of local gradients, so the randomness introduced during sparsification cannot be leveraged to amplify the differential privacy guarantee. As a result, this approach leads to a suboptimal trade-off between privacy and communication compared to our scheme.

Note that in this work, we consider FL (or more specifically, the distributed mean estimation) under a *central*-DP setting where the server is trusted, which is different from the local DP model (Kasiviswanathan et al., 2011; Duchi et al., 2013; Nguyên et al., 2016; Wang et al., 2019; Bhowmick et al., 2018; Chen et al., 2020) and the distributed DP model with secure aggregation (Bonawitz et al., 2016; Bell et al., 2020; Kairouz et al., 2021a; Agarwal et al., 2021; Chen et al., 2022b;c).

A key step in our mean estimation scheme is pre-processing the local data via Kashin's representation (Lyubarskii & Vershynin, 2010). While various compression schemes, based on quantization, sparsification, and dithering have been proposed in the recent literature, Kashin's representation has also been explored in a few works for communication efficiency (Fuchs, 2011; Studer et al., 2012; Caldas et al., 2018; Safaryan et al., 2020) and for LDP (Feldman et al., 2017) and is particularly powerful in the case of joint communication and privacy constraints as it helps spread the information in a vector evenly in every dimension.

Distributed frequency estimation and heavy hitters. Distributed frequency estimation (a.k.a. histogram estimation) is another canonical task that has been heavily studied under a distributed setting with DP. Prior works either focus on 1) the local DP model with or without communication constraints, e.g., (Bassily & Smith, 2015; Bassily et al., 2017; Bun et al., 2018; 2019; Huang et al., 2022) (under an ℓ_∞ loss for heavy hitter estimation) and (Kairouz et al., 2016; Ye & Barg, 2017; Wang et al., 2019; Acharya et al., 2019c; Acharya & Sun, 2019; Chen et al., 2020; Feldman & Talwar, 2021; Shah et al., 2022; Feldman et al., 2022b) (under an ℓ_1 or ℓ_2 loss), or 2) the central DP model *without* communication constraints (Dwork et al., 2006; Ghosh et al., 2012; Korolova et al., 2009; Bun & Steinke, 2016; Balcer & Vadhan, 2017; Zhu et al., 2020; Cormode & Bharadwaj, 2022). As suggested in (Duchi et al., 2013; Acharya et al., 2019a;b; 2020; Barnes et al., 2020a), compared to central DP, local DP models usually incur much larger estimation errors and can significantly decrease the utility. In this work, we consider central DP but with explicit communication constraints.

Local DP with shuffling. A recent line of works (Erlingsson et al., 2019; Cheu et al., 2019; Balcer & Cheu, 2019; Feldman et al., 2022a; Ghazi et al., 2019; 2020) considers *shuffle*-DP, showing that one can significantly boost the central DP guarantees by randomly shuffling local (privatized) messages. In this work, we show that the same shuffling technique can be used to achieve the optimal central DP error with nearly optimal communication cost. Therefore, we can obtain the same level of central DP with small communication costs while weakening the security assumption: achieving the optimal communication cost (under central DP) only requires a secure shuffler (as opposed to a fully trusted central server).

B. Omitted Details of Distributed Mean Estimation

B.1. Algorithms

Algorithm 1 Coordinate Subsampled Gaussian Mechanism (CSGM)

Input: users' data $x_1, ..., x_n$, sampling parameters $\gamma \coloneqq b/d$, DP parameters (ε, δ).
Output: mean estimator $\hat{\mu}$.
for user $i \in [n]$ **do**
  for coordinate $j \in [d]$ **do**
    Draw $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\gamma)$.
    if $Z_{i,j} = 1$ **then**
      Send $x_i(j)$ to the server.
    end if
  end for
end for
for coordinate $j \in [d]$ **do**
  Server computes the average $\hat{\mu}_j \coloneqq \frac{1}{n\gamma} \sum_{i:Z_{ij}=1} x_i(j) + N(0, \sigma^2)$, where σ^2 is computed according to (2) in Theorem B.1.
end for
Return: $\hat{\mu} \coloneqq (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_d)$.

We summarize the scheme in Algorithm 1 and state its privacy and utility guarantees in the following theorem.

Theorem B.1 (ℓ_∞ mean estimation.). *Let $x_1, ..., x_n \in \{-c, c\}^d$ and let*

$$\sigma^2 = O\left(\frac{c^2 \log(1/\delta)}{n^2\gamma^2} + \frac{c^2 d(\log(d/\delta) + \varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right). \tag{2}$$

Then for any $\varepsilon, \delta > 0$, Algorithm 1 is (ε, δ)-DP and yields an unbiased estimator on μ. In addition, the (average) per-client communication cost is $\gamma \cdot d = b$ bits, and the ℓ_2^2 estimation error of $\hat{\mu}$ is at most

$$\begin{aligned}\mathbb{E}\left[\|\hat{\mu} - \mu\|_2^2\right] &\leq \frac{dc^2}{n\gamma} + d\sigma^2 \\ &= O\Big(\frac{d^2c^2}{nb} + \frac{d^3c^2\log(d/\delta)}{n^2b^2} \qquad (3) \\ &+ \frac{c^2d^2(\log(1/\delta) + \varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\Big). \qquad (4)\end{aligned}$$

B.2. ℓ_2 mean estimation via Kashin's representation (proof of Theorem 4.1)

If x_i has ℓ_2 norm bounded by C, then its Kashin's representation (with respect to a tight frame $K \in \mathbb{R}^{d\times D}$ where $D = \Theta(d)$) $\tilde{x}_i$ has bounded ℓ_∞ norm: $\|\tilde{x}_i\|_\infty \leq c = O\left(\frac{C}{\sqrt{d}}\right)$ and satisfies $x_i = K \cdot \tilde{x}_i$. This allows us to convert the ℓ_2 geometry to an ℓ_∞ geometry. Furthermore, by randomly rounding each coordinate of $\tilde{x}_i$ to $\{-c, c\}$ (see for example (Chen et al., 2020)), we can readily apply Algorithm 1 and obtain the desired results for ℓ_2 mean estimation. Theorem 4.1 is a direct consequence of combining Kashin's representation and Theorem B.1.

Algorithm 2 CSGM with Coordinate Pre-selection

Input: users' data $x_1, ..., x_n$, coordinate selection $d' \leq d$, sampling parameters $\gamma := b/d'$, DP parameters (ε, δ).
Output: mean estimator $\hat{\mu}$.
Randomly select d' coordinates $\mathcal{J} := \{j_1, ..., j_{d'}\} \subset [d]$.
for user $i \in [n]$ **do**
  Pre-processing x_i by restricting it on $\mathcal{J}$:
  $x_i(\mathcal{J}) := (x_i(j_1), ..., x_i(j_{|\mathcal{J}|}))$.
end for
Apply CSGM (Algorithm 1) on $x_i(\mathcal{J}), i \in [n]$:
$\hat{\mu}_{\mathcal{J}} \leftarrow \mathsf{CSGM}\left(x_i(\mathcal{J}), i \in [n]\right)$.
for $j \in [d]$ **do**
  if $j \in \mathcal{J}$ **then**
    $\hat{\mu}_j = \hat{\mu}_{\mathcal{J}}(j)$.
  else
    $\hat{\mu}_j = 0$.
  end if
end for
Return: $\hat{\mu} := \left(\frac{d}{d'}\hat{\mu}_1, \frac{d}{d'}\hat{\mu}_2, ..., \frac{d}{d'}\hat{\mu}_d\right)$.

B.3. Proof of Theorem B.1

It is trivial to see that the average communication cost is $d \cdot \gamma = b$ bits. To compute the ℓ_2^2 estimation error, observe that

$$
\begin{aligned}
&\mathbb{E}\left[\|\hat{\mu}_{x^n} - \mu_{x^n}\|_2^2\right] \\
&= \sum_{j=1}^{d} \mathbb{E}\left[\left(\frac{1}{n\gamma}\sum_i x_i(j) \cdot Z_{i,j} + N(0, \sigma^2) - \frac{1}{n}\sum_i x_i(j)\right)^2\right] \\
&= \sum_{j=1}^{d} \frac{1}{n^2}\mathbb{E}\left[\left(\frac{1}{\gamma}\sum_i x_i(j) \cdot Z_{i,j} - \sum_i x_i(j)\right)^2\right] + d\sigma^2 \\
&= \sum_{j=1}^{d} \frac{1}{n^2}\mathbb{E}\left[\left(\frac{1}{\gamma}\sum_i x_i(j) \cdot Z_{i,j}\right)^2\right] - \frac{1}{n^2}\left(\sum_i x_i(j)\right)^2 + d\sigma^2 \\
&= \sum_{j=1}^{d} \frac{1}{n^2}\mathbb{E}\left[\frac{1}{\gamma^2}\sum_i x_i^2(j) \cdot Z_{i,j}^2 + \frac{1}{\gamma^2}\sum_{i \neq i'} x_i(j)x_{i'}(j)Z_{i,j}Z_{i',j}\right] - \frac{1}{n^2}\left(\sum_i x_i(j)\right)^2 + d\sigma^2 \\
&= \sum_{j=1}^{d} \frac{1}{n^2}\left(\frac{1}{\gamma}\sum_i x_i^2(j) + \sum_{i \neq i'} x_i(j)x_{i'}(j)\right) - \frac{1}{n^2}\left(\sum_i x_i(j)\right)^2 + d\sigma^2 \\
&= \sum_{j=1}^{d} \frac{1}{n^2}\left(\frac{1}{\gamma} - 1\right)\left(\sum_i x_i^2(j)\right) + d\sigma^2 \\
&\leq \frac{dc^2}{n\gamma} + d\sigma^2,
\end{aligned}
$$

which yields the inequality of (3). Next, we analyze the privacy of Algorithm 1. We first the following two lemmas for subsampling and the Gaussian mechanism:

Lemma B.2 ((Li et al., 2012; Zhu & Wang, 2019))**.** *If $\mathcal{M}$ is (ε, δ)-DP, then $\mathcal{M}'$ that applies $\mathcal{M} \circ \mathsf{PoissonSample}$ satisfies (ε', δ')-DP with $\varepsilon' = \log\left(1 + \gamma\left(e^{\varepsilon} - 1\right)\right)$ and $\delta' = \gamma\delta$.*

Lemma B.3 ((Balle & Wang, 2018))**.** *For any $\varepsilon, \delta \in (0, 1)$, the Gaussian output perturbation mechanism with $\sigma^2 := \frac{\Delta^2 2\log(1.25/\delta)}{\varepsilon^2}$ satisfies (ε, δ)-DP, where Δ is the ℓ_2 sensitivity of the target function.*

Now, we use the above two lemmas to analyze the per-coordinate privacy leakage of Algorithm 1. For simplicity, we analyze the sum of $x_i(j)$'s instead (and normalized it in the last step). Let $S_j(x^n) := \sum_{i=1}^n (x_i(j))$, then clearly the sensitivity of $S_j(x^n)$ is c, so Lemma B.3 implies $S_j(x^n) + N(0, \sigma_1^2)$ satisfies $(\varepsilon_1, \delta_1)$-DP if we set $\sigma_1^2 = \frac{2c^2 \log(1.25/\delta_1)}{\varepsilon_1^2}$ (assuming $\varepsilon_1 < 1$). Next, if applying subsampling before computing the sum, i.e.,

$$S_j \circ \mathsf{PoissonSample}_\gamma(x^n) := \sum_{i=1}^n x_i(j) Z_{i,j},$$

where $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(1/\gamma)$ as defined in Algorithm 1, then by Lemma B.2,

$$S_j \circ \mathsf{PoissonSample}_\gamma(x^n) + N(0, \sigma_1^2)$$

satisfies $(\varepsilon_2, \delta_2)$-DP with $\varepsilon_2 := \log\left(1 + \gamma\left(e^{\varepsilon_1} - 1\right)\right) = C_1 \gamma \varepsilon_1$ (since we assume $\epsilon_1 < 1$) and $\delta_2 := \gamma \delta_1$. Equivalently, we have

$$\begin{cases} \varepsilon_1 = \tilde{C}_1 \frac{1}{\gamma} \varepsilon_2 \\ \delta_1 = \frac{1}{\gamma} \delta_2. \end{cases} \tag{5}$$

Now, since we have established the per-coordinate privacy leakage, we apply the following composition theorem to account for the total privacy budgets.

Theorem B.4. *For any $\varepsilon > 0$, $\delta \in [0,1]$ and $\tilde{\delta} \in (0,1]$, the class of (ε, δ)-DP mechanisms satisfies $(\tilde{\varepsilon}_{\tilde{\delta}}, d\delta + \tilde{\delta})$-DP under d-fold adaptive composition, for*

$$\tilde{\varepsilon}_{\tilde{\delta}} = d\varepsilon\left(e^{\varepsilon} - 1\right) + \varepsilon\sqrt{2d\log(1/\tilde{\delta})}.$$

According Theorem B.4, Algorithm 1 satisfies (ε, δ)-DP for

$$\varepsilon = d\varepsilon_2 (e^{\varepsilon_2} - 1) + \varepsilon_2 \sqrt{2d\log(1/\tilde{\delta})}, \tag{6}$$

and $\delta = d\delta_2 + \tilde{\delta}$ (where $\tilde{\delta}$ is a free parameter that we can optimize).

Consequently, for a pre-specified (total) privacy budget (ε, δ), we set parameters as follows. Let $\tilde{\delta} = \frac{\delta}{2}$ and $\delta_1 = \frac{1}{\gamma}\delta_2 = \frac{1}{2d\gamma}\delta$. Let $\varepsilon_2 \leq 1$ so that $e_2^{\varepsilon} - 1 \leq 2\varepsilon_2$ holds. Then (6) implies Algorithm 1 is

$$\varepsilon = 2d\varepsilon_2^2 + \varepsilon_2\sqrt{2d\log(1/\tilde{\delta})} \geq d\varepsilon_2(e^{\varepsilon_2} - 1) + \varepsilon_2\sqrt{2d\log(1/\tilde{\delta})}.$$

Solving the above quadratic (in-)equality for ε_2, it yields that

$$\varepsilon_2 = \min\left(1, \frac{-\sqrt{2d\log(2/\delta)} + \sqrt{2d\log(2/\delta) + 8\varepsilon d}}{4d}\right) = O\left(\min\left(1, \frac{\varepsilon}{\sqrt{d\left(\log(1/\delta) + \varepsilon\right)}}\right)\right).$$

Consequently, we set $\varepsilon_1 = \frac{\tilde{C}_1}{\gamma}\varepsilon_2 = O\left(\min\left(1, \frac{\varepsilon}{\gamma\sqrt{d(\log(1/\delta)+\varepsilon)}}\right)\right)$ (note that we require $\varepsilon_1 = O(1)$ so that (5) holds).

Plug in $(\varepsilon_1, \delta_1)$ into σ_1^2, we have

$$\sigma_1^2 := \frac{2c^2\log(1.25/\delta_1)}{\varepsilon_1^2} = \Omega\left(\max\left(c^2\log(d/\delta), \frac{\gamma^2 c^2 d(\log(1/\delta) + \varepsilon)\log(d/\delta)}{\varepsilon^2}\right)\right).$$

Finally, as we are interested in estimating the (subsampled) mean instead of the sum, we will normalize the private sum by

$$\hat{\mu}_j(x^n) = \frac{1}{n\gamma}\left(S_j \circ \mathsf{PoissonSample}_\gamma(x^n) + N(0, \sigma_1^2)\right) = \frac{1}{n\gamma} S_j \circ \mathsf{PoissonSample}_\gamma(x^n) + N(0, \sigma^2),$$

where

$$\sigma^2 = O\left(\max\left(\frac{c^2\log(d/\delta)}{n^2\gamma^2}, \frac{c^2 d(\log(1/\delta) + \varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right)\right).$$

Plugging in σ^2 above and $\gamma = d/b$ yields the desired accuracy in Theorem B.1. □

Since we will reuse the above result, we summarize it into the following lemma:

Lemma B.5. *Let $f_i : \mathbb{R}^{d\times m} \mapsto \mathbb{R}^D$ for $i = 1, ..., B$ be n functions with sensitivity bounded by Δ (where the number of inputs m can be a random variable). Then*

$$\left(f_1 \circ \mathsf{PoissonSample}_\gamma(x^n) + N(0,\sigma^2), ..., f_B \circ \mathsf{PoissonSample}_\gamma(x^n) + N(0,\sigma^2)\right)$$

satisfies (ε, δ)-DP, if

$$\sigma^2 \geq O\left(\max\left(\Delta^2 \log(B/\delta), \frac{\gamma^2\Delta^2 B(\log(1/\delta)+\varepsilon)\log(B/\delta)}{\varepsilon^2}\right)\right).$$

B.4. Proof of Theorem 4.3

To prove Theorem 4.3, it suffices to prove the following ℓ_∞ version:

Theorem B.6. *Let $x_1, ..., x_n \in \{-c, c\}^d$, $d' = \min\left(nb, \frac{n^2\varepsilon^2}{(\log(1/\delta)+\varepsilon)\log(d/\delta)}\right)$, and*

$$\sigma^2 = O\left(\frac{c^2\log(1/\delta)}{n^2\gamma^2} + \frac{c^2 d'(\log(d'/\delta)+\varepsilon)\log(d'/\delta)}{n^2\varepsilon^2}\right). \tag{7}$$

Then Algorithm 2 is (ε, δ)-DP and yields an unbiased estimator on μ. In addition, the (average) per-client communication cost is $\gamma d' = b$ bits, and the ℓ_2^2 estimation error is at most

$$O\left(c^2 d^2 \log\left(\frac{d}{\delta}\right)\max\left(\frac{1}{nb}, \frac{(\log(1/\delta)+\varepsilon)}{n^2\varepsilon^2}\right)\right). \tag{8}$$

With a slight abuse of notation, we let $\mu_\mathcal{J} \in \mathbb{R}^d$ be such that

$$\mu_\mathcal{J}(j) = \begin{cases} 0, & \text{if} j \notin \mathcal{J} \\ \frac{d\mu_j}{d'}, & \text{else.}\end{cases}$$

Note that $\mu_\mathcal{J}$ is an unbiased estimate of μ if $\mathcal{J}$ is selected uniformly at random. Then the ℓ_2^2 error can be controlled by

$$\begin{aligned}
\mathbb{E}\left[\|\mu-\hat{\mu}\|_2^2\right] &\overset{\text{(a)}}{=} \mathbb{E}\left[\|\mu-\mu_\mathcal{J}\|_2^2\right] + \mathbb{E}\left[\|\mu_\mathcal{J}-\hat{\mu}\|_2^2\right] \\
&\overset{\text{(b)}}{\leq} \mathbb{E}\left[\|\mu-\mu_\mathcal{J}\|_2^2\right] + \frac{d^2}{d'^2} O\left(\max\left(\frac{d'^2c^2}{nb}, \frac{d'^3c^2\log(d/\delta)}{n^2b^2}, \frac{c^2d'^2(\log(1/\delta)+\varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right)\right) \\
&= \mathbb{E}\left[\|\mu-\mu_\mathcal{J}\|_2^2\right] + O\left(\max\left(\frac{d^2c^2}{nb}, \frac{d^2d'c^2\log(d/\delta)}{n^2b^2}, \frac{c^2d^2(\log(1/\delta)+\varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right)\right) \\
&\overset{\text{(c)}}{\leq} \frac{d^2c^2}{d'} + O\left(\max\left(\frac{d^2c^2}{nb}, \frac{d^2d'c^2\log(d/\delta)}{n^2b^2}, \frac{c^2d^2(\log(1/\delta)+\varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right)\right),
\end{aligned}$$

where (a) holds since $\mu_\mathcal{J}$ is an unbiased estimate of μ and conditioned on $\mathcal{J}$, $\hat{\mu}$ is an unbaised estimate of $\mu_\mathcal{J}$; (b) follows from Theorem B.1; (c) holds due to the following fact:

$$\mathbb{E}\left[\|\mu-\mu_\mathcal{J}\|_2^2\right] \leq \sum_{j\in\mathcal{J}} \mu_\mathcal{J}(j)^2 + \sum_{j\in[d]} \mu_j^2 \leq \frac{d^2c^2}{d'} + dc^2 \leq \frac{2d^2c^2}{d'}.$$

Therefore, by setting $d' = \min\left(nb, \frac{n^2\varepsilon^2}{(\log(1/\delta)+\varepsilon)\log(d/\delta)}\right)$ we ensure the first term in (c) is always smaller than the second term, and the second term can be simplified as follows:

$$\begin{aligned}
&O\left(c^2d^2\max\left(\frac{1}{nb}, \frac{d'\log(d/\delta)}{n^2b^2}, \frac{(\log(1/\delta)+\varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right)\right) \\
&\leq O\left(c^2d^2\max\left(\frac{1}{nb}, \frac{nb\log(d/\delta)}{n^2b^2}, \frac{(\log(1/\delta)+\varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right)\right) \\
&\leq O\left(c^2d^2\log(d/\delta)\max\left(\frac{1}{nb}, \frac{(\log(1/\delta)+\varepsilon)}{n^2\varepsilon^2}\right)\right).
\end{aligned}$$

Finally, applying the same trick of Kashin's representation, we can transform the ℓ_∞ geometry to ℓ_2 (similar to Proposition 4.1), hence proving Theorem 4.3. □

C. Distributed Frequency Estimation

In this section, we consider the frequency estimation problem for federated analytics. Recall that for the frequency estimation task, each client's private data $x_i \in \{0,1\}^d$ satisfies $\|x_i\|_0 = 1$, and the goal is to estimate $\pi := \frac{1}{n}\sum_i x_i$ by minimizing the ℓ_2 (or ℓ_1, ℓ_∞) error $\mathbb{E}\left[\|\pi - \hat{\pi}(Y^n)\|_2^2\right]$ subject to communication and (ε, δ)-DP constraints. When the context is clear, we sometimes use x_i to denote, by abuse of notation, the index of the item, i.e., $x_i \in [d]$.

To fully make use of the ℓ_0 structure of the problem, a standard technique is applying a Hadamard transform to convert the ℓ_0 geometry to an ℓ_∞ one and then leveraging the recursive structure of Hadamard matrices to efficiently compress local messages.

Specifically, for a given b-bit constraint, we partition each local item x_i into 2^{b-1} chunks $x_i^{(1)}, ..., x_i^{(2^b-1)} \in \{0,1\}^B$, where $B := d/2^{b-1}$ and $x_i^{(j)} = x_i[B\cdot(j-1) : B\cdot j - 1]$. Note that since x_i is one-hot, only one chunk of $x_i^{(j)}$ is non-zero. Then, client i performs the following Hadamard transform for each chunk: $y_i^{(\ell)} = H_B \cdot x_i^{(\ell)}$, where H_B is defined recursively as follows:

$$H_{2^n} = \frac{1}{\sqrt{2}}\begin{bmatrix} H_{2^{n-1}}, & H_{2^{n-1}} \\ H_{2^{n-1}}, & -H_{2^{n-1}} \end{bmatrix}, \text{ and } H_0 = \begin{bmatrix} 1 \end{bmatrix}.$$

Each client then generates a sampling vector $Z_{ij} \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}\left(\frac{1}{B}\right)$ via shared randomness that is also known by the server, and commits $(y_i^{(1)}(j), ..., y_i^{(2^{b-1})}(j))$ as its local report. Since $(y_i^{(1)}(j), ..., y_i^{(2^{b-1})}(j))$ only contains a single non-zero entry that can be $\frac{1}{\sqrt{B}}$ or $-\frac{1}{\sqrt{B}}$, the local report can be represented in b bits ($b-1$ bits for the location of the non-zero entry and 1 bit for its sign).

From the local reports, the server can compute an unbiased estimator by summing them together (with proper normalization) and performing an inverse Hadamard transform. Moreover, with an adequate injection of Gaussian noise, the frequency estimator satisfies (ε, δ)-DP.

The idea has been used in previous literature under local DP (Bassily et al., 2017; Acharya et al., 2019c;a; Chen et al., 2020), but in order to obtain the order-optimal trade-off under *central*-DP, one has to combine Hadamard transform with a random subsampling step and incorporate the privacy amplification due to random compression in the analysis. In Algorithm 3, we provide a summary of the resultant scheme which builds on the Recursive Hadamard Response (RHR) mechanism from (Chen et al., 2020), which was originally designed for communication-efficient frequency estimation under *local* DP.

In the following theorem, we control the ℓ_∞ error of Algorithm 3.

Theorem C.1. *Let $\hat{\pi}(x^n)$ be the output of Algorithm 3. Then it holds that for all $j \in [d]$,*

$$\mathbb{E}\left[|\pi(j) - \hat{\pi}(j)|\right] \leq \sqrt{\frac{\sum_i \mathbb{1}_{\{x_i \in [B\cdot(j-1):B\cdot j-1]\}}}{n^2} + \frac{\sigma^2}{B}}, \tag{9}$$

and the ℓ_2^2 and ℓ_1 errors are bounded by

$$\mathbb{E}\left[\|\pi - \hat{\pi}\|_2^2\right] \leq \frac{B}{n} + \frac{d\sigma^2}{B}, \text{ and} \tag{10}$$

$$\mathbb{E}\left[\|\pi - \hat{\pi}\|_1\right] \leq \sqrt{\frac{dB}{n} + \frac{d^2\sigma^2}{B}}. \tag{11}$$

Theorem C.2. *For any $\varepsilon, \delta > 0$, Algorithm 3 is (ε, δ)-DP, if*

$$\sigma^2 \geq O\left(\frac{B^2 \log(B/\delta)}{n^2} + \frac{B(\log(1/\delta) + \varepsilon)\log(B/\delta)}{n^2\varepsilon^2}\right).$$

By combining Theorem C.1 and Theorem C.2, we conclude that Algorithm 3 achieves (ε, δ)-DP with ℓ_2^2 error

$$O\left(\frac{B}{n} + \frac{dB\log(B/\delta)}{n^2} + \frac{d(\log(1/\delta)+\varepsilon)\log(B/\delta)}{n^2\varepsilon^2}\right)$$
$$= O\left(\frac{d}{n2^b} + \frac{d^2\log(d/\delta)}{n^2 2^b} + \frac{d(\log(1/\delta)+\varepsilon)\log(d/\delta)}{n^2\varepsilon^2}\right).$$

Algorithm 3 Subsampled Recursive Hadamard Response

Input: user data $x_1, ..., x_n \in \{0,1\}^d$ (where d is a power of two), DP parameters (ε, δ), communication budget b.
Output: frequency estimate $\hat{\pi}$
Set $B := d/2^{b-1}$ and partition each one-hot vector x_i into 2^{b-1} chunks: $x_i^{(1)}, ..., x_i^{(2^b-1)} \in \{0,1\}^B$.
for user $i \in [n]$ **do**
  Compute the Hadamard transform of each chunk: $y_i^{(\ell)} = H_B \cdot x_i^{(\ell)}$.
  for coordinate $j \in [B]$ **do**
    Draw $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}\left(\frac{1}{B}\right)$
    if $Z_{i,j} = 1$ **then**
      Send $(y_i^{(1)}(j), ..., y_i^{(2^{b-1})}(j))$ to the server.
    end if
  end for
end for
Server computes the average: $\forall \ell \in [2^{b-1}], j \in [B]$,

$$\hat{y}^{(\ell)}(j) := \frac{B}{n} \sum_{i: Z_{ij}=1} y_i^{(\ell)}(j) + N(0, \sigma^2),$$

where σ^2 is computed according to Theorem C.2.
Server performs the inverse Hadamard transform $\hat{\pi}^{(\ell)} = H_B \cdot \hat{y}^{(\ell)}$, for $\ell = 1, ..., B$.
Return: $\hat{\pi} = \left(\left(\hat{\pi}^{(1)}\right)^{\intercal}, ..., \left(\hat{\pi}^{(2^{b-1})}\right)^{\intercal} \right)$.

Notice that when $n = \tilde{\Omega}(d)$, the error can be simplified to

$$O\left(\frac{d}{n 2^b} + \frac{d(\log(1/\delta) + \varepsilon)\log(d/\delta)}{n^2 \varepsilon^2} \right),$$

which matches the order-optimal estimation error (up to a $\log d$ factor) subject to a b-bit constraint (Han et al., 2018; Acharya et al., 2019a;b) and (ε, δ)-DP constraint (Balle & Wang, 2018; Acharya et al., 2021).

C.1. Proof of Theorem C.1

Let $\pi := \frac{1}{n}\sum_i x_i$ and $\pi^{(\ell)}$ be defined in the same way as $x_i^{(\ell)}$ for $\ell \in [B]$. Then our goal is to bound $\left|\pi^{(\ell)}(j) - \hat{\pi}^{(\ell)}(j)\right|$, for all $\ell \in [2^{b-1}]$ and $j \in [B]$.

To this end, let $y^{(\ell)} := H_B \cdot \pi^{(\ell)}$ (so it holds that $\pi^{(\ell)} = \frac{1}{B} H_B \cdot y^{(\ell)}$). Then we have

$$\begin{aligned} \mathbb{E}\left[\left|\pi^{(\ell)}(j) - \hat{\pi}^{(\ell)}(j)\right|\right] &\overset{\text{(a)}}{\leq} \sqrt{\mathbb{E}\left[\left(\pi^{(\ell)}(j) - \hat{\pi}^{(\ell)}(j)\right)^2\right]} \\ &= \sqrt{\mathbb{E}\left[\left(\frac{1}{B} H_B \cdot \left(y^{(\ell)} - \hat{y}^{(\ell)}\right)(j)\right)^2\right]}. \end{aligned} \tag{12}$$

Next, observe that due to the subsampling step, for all $\ell \in [2^{b-1}]$ and $j \in [B]$,

$$\hat{y}^{(\ell)}(j) = \frac{B}{n} \sum_{i=1}^{n} \langle (H_B)_j, x_i^{(\ell)} \rangle \cdot Z_{ij} + N(0, \sigma^2),$$

where recall that $Z_{ij} \overset{\text{i.i.d.}}{\sim} \mathsf{Ber}(1/B)$. Therefore, $\hat{y}^{(\ell)}(j)$ is an unbiased estimator of $y^{(\ell)}(j)$. In addition, since we choose Z_{ij}

independently in Algorithm 3, $\hat{y}^{(\ell)}(j)$'s are independent for different j's, so we have

$$\begin{aligned}\mathbb{E}\left[\left(\hat{y}^{(\ell)}(j) - y^{(\ell)}(j)\right)^2\right] &= \mathsf{Var}\left(\hat{y}^{(\ell)}(j)\right) \\ &= \sigma^2 + \frac{B^2}{n^2}\sum_{i=1}^{n}\langle (H_B)_j, x_i^{(\ell)}\rangle^2 \mathsf{Var}\left(Z_{ij}\right) \\ &\leq \sigma^2 + \frac{B}{n^2}\sum_{i=1}^{n}\langle (H_B)_j, x_i^{(\ell)}\rangle^2 \\ &= \sigma^2 + \frac{B}{n^2}\underbrace{\sum_{i=1}^{n}\mathbb{1}_{\{x_i \in \ell\text{-th chunk}\}}}_{:=C_\ell}, \end{aligned} \tag{13}$$

and for all $j \neq j'$

$$\mathbb{E}\left[\left(\hat{y}^{(\ell)}(j) - y^{(\ell)}(j)\right)\cdot\left(\hat{y}^{(\ell)}(j') - y^{(\ell)}(j')\right)\right] = 0. \tag{14}$$

Therefore, we continue bounding (12) as follows:

$$\begin{aligned}\sqrt{\mathbb{E}\left[\left(\frac{1}{B}H_B\cdot\left(y^{(\ell)} - \hat{y}^{(\ell)}\right)(j)\right)^2\right]} &= \sqrt{\frac{1}{B^2}\mathbb{E}\left[\langle (H_B)_j, \left(\hat{y}^{(\ell)} - y^{(\ell)}\right)\rangle^2\right]} \\ &= \sqrt{\frac{1}{B^2}\mathbb{E}\left[\left(\sum_{k=1}^{B}(H_B)_{jk}\cdot\left(\hat{y}^{(\ell)}(k) - y^{(\ell)}(k)\right)\right)^2\right]} \\ &\overset{\text{(a)}}{=} \sqrt{\frac{1}{B^2}\mathbb{E}\left[\sum_{k=1}^{B}\left(\hat{y}^{(\ell)}(k) - y^{(\ell)}(k)\right)^2\right]} \\ &\overset{\text{(b)}}{=} \sqrt{\frac{C_\ell}{n^2} + \frac{\sigma^2}{B}} \\ &\overset{\text{(c)}}{\leq} \sqrt{\frac{1}{n} + \frac{\sigma^2}{B}}, \end{aligned}$$

where (a) holds since each entry of H_B takes value in $\{-1, 1\}$ and by (14), (b) holds due to (13), and (c) holds because $C_\ell \leq n$ for all ℓ.

Finally, to bound the ℓ_2^2 error, observe that the above analysis ensures that

$$\mathbb{E}\left[\left(\pi^{(\ell)}(j) - \hat{\pi}^{(\ell)}(j)\right)^2\right] \leq \frac{C_{\ell(j)}}{n^2} + \frac{\sigma^2}{B},$$

where $\ell(j) \in [2^{b-1}]$ is the index of the chuck containing j. Therefore, summing over $j \in [d]$, we must have

$$\mathbb{E}\left[\left\|\pi^{(\ell)} - \hat{\pi}^{(\ell)}\right\|_2^2\right] \leq \sum_{j=1}^{d}\frac{C_{\ell(j)}}{n^2} + \frac{d\sigma^2}{B} = \frac{B}{n} + \frac{d\sigma^2}{B},$$

since

$$\sum_j C_{\ell(j)} = \sum_{\ell=1}^{2^{b-1}}\sum_{j'\in\ell\text{-th chunk}}\sum_{i=1}^{n}\mathbb{1}_{\{i\in\ell-\text{th chunk}\}} = B\sum_{\ell=1}^{2^{b-1}}\sum_{i=1}^{n}\mathbb{1}_{\{i\in\ell-\text{th chunk}\}} = B\cdot n.$$

□

D. Proof of Theorem C.2

Let $f_j(x^n) \coloneqq (\pi^{(1)}(j), ..., \pi^{(2^{b-1})}(j))$, for $j = 1, ..., B$. Then the ℓ_2 sensitivity of f_j is $\Delta = \frac{B}{n}$. Set the sampling rate $\gamma = \frac{1}{B}$ and the proof is complete by Lemma B.5. □

E. Additional Details for Shuffle-DP

In this section, we present a mean estimation scheme that combines a local-DP mechanism with privacy amplification via shuffling by building on the following recent result (Erlingsson et al., 2019; Feldman et al., 2022a):

Lemma E.1 ((Feldman et al., 2022a))**.** *Let $\mathcal{M}_i$ be an independent $(\varepsilon_0, 0)$-LDP mechanism for each $i \in [n]$ with $\varepsilon_0 \leq 1$ and π be a random permutation of $[n]$. Then for any $\delta \in [0, 1]$ such that $\varepsilon_0 \leq \log\left(\frac{n}{16\log(2/\delta)}\right)$, the mechanism $\mathcal{S}$: $(x_1, \ldots, x_n) \mapsto \left(\mathcal{M}_1\left(x_{\pi(1)}\right), \ldots, \mathcal{M}_n\left(x_{\pi(n)}\right)\right)$ is (ε, δ)-DP for some ε such that $\varepsilon = O\left(\varepsilon_0 \frac{\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)$.*

Privacy analysis. With the above amplification lemma, we only need to design the local randomizers $\mathcal{M}_i$ that satisfy ε_0-LDP. Note that the above lemma is only tight when $\varepsilon_0 = O(1)$, thus restricting the (amplified) central $\varepsilon = O(1/\sqrt{n})$, i.e. to be very small. To accommodate larger ε, users can send different portions of their messages to the server in separate shuffling rounds. Equivalently, we repeat the shuffled LDP mechanism for $T = O\left(\lceil n\varepsilon^2\rceil\right)$ rounds while ensuring that in each round clients communicate an independent piece of information about their sample to the server. More precisely, within each round, each client applies the local randomizers $\mathcal{M}_i$ with a per-round *local privacy budget* $\varepsilon_0 = O(1)$ and sends an independent message to the server. This results in (amplified) central $O(1/\sqrt{n})$-DP per round, which after composition over $T = O\left(\lceil n\varepsilon^2\rceil\right)$ rounds leads to ε-DP for the overall scheme as suggested by the composition theorem (Kairouz et al., 2016)). We detail the algorithm in Algorithm 4 in Appendix E.1.

Communication costs. The communication cost of the above T-round scheme can be computed as follows. As shown in (Chen et al., 2020), the optimal communication cost of an ε_0-LDP mean estimation is $O\left(\lceil\varepsilon_0\rceil\right)$ bits. In addition, the (private-coin) SQKR scheme proposed in (Chen et al., 2020) uses $O\left(\lceil\varepsilon_0\rceil \log d\right)$ bits of communication (we state the formal performance guarantee for this scheme in Lemma E.2), where compression is done by subsampling coordinates and privatization is performed with Randomized Response. Therefore, since the per-round $\varepsilon_0 = O(1)$, the total per-client communication cost is $O\left(n\varepsilon^2 \log d\right)$, matching the optimal communication bounds in Section 4 within a $\log d$ factor.

Lemma E.2 (SQKR (Chen et al., 2020))**.** *For all $\varepsilon_0 > 0, b_0 > 0$, there exists a $(\varepsilon_0, 0)$-LDP mechanism using $b_0 \log(d)$ bits such that $\hat{\mu}$ is unbiased and satisfies $\mathbb{E}\left[\left\|\mu\left(x^n\right) - \hat{\mu}\left(x^n\right)\right\|_2^2\right] = O\left(\frac{c^2 d}{n \min\left(\varepsilon_0^2, \varepsilon_0, b_0\right)}\right)$.*

Finally, we summarize the performance guarantee for the overall scheme (Algorithm 4) in the following theorem.

Theorem E.3 (ℓ_2 mean estimation)**.** *Let $x_1, ..., x_n \in \mathcal{B}_2(C)$ (i.e., $\left\|x_i\right\|_2 \leq C$ for all $i \in [n]$). For all $\varepsilon > 0, b > 0, n > 30$, and $\delta \in (\delta_{\min}, 1]$ where $\delta_{\min} = O\left(\frac{be^{-n}}{\log(d)}\right)$, Algorithm 4 combined with Kashin's representation and randomized rounding is (ε, δ)-DP, uses no more than b bits of communication, and achieves*

$$\mathbb{E}\left[\left\|\mu\left(x^n\right) - \hat{\mu}\left(x^n\right)\right\|_2^2\right] = O\left(C^2 d \max\left(\frac{\log(d)}{nb}, \frac{\log(b/\delta)(\log(1/\delta) + \varepsilon)}{n^2\varepsilon^2}\right)\right).$$

Remark E.4. As opposed to previous schemes Algorithm 1-3, the shuffled SQKR requires some condition on δ, i.e., $\delta \in [\delta_{\min}, 1]$ due to the specific shuffling lemma we used. In practice, however, $\delta_{\min}$ is small due to the exponential dependence on n. The order-wise optimal error of $O\left(\frac{C^2 d}{n^2 \min(\varepsilon^2, \varepsilon)}\right)$ is achieved, up to logarithmic factors, when $b = \Omega_\delta\left(n \log(d) \min\left(\varepsilon^2, \varepsilon\right)\right)$.

Remark E.5. We note that similar ideas of private mean estimation based on shuffling have been studied before, see, for instance, (Girgis et al., 2021). However, these papers do not use the above privacy budget splitting trick over multiple rounds, so their result is only optimal when ε is very small. The above scheme can be viewed as a multi-message shuffling scheme (Cheu et al., 2019; Ghazi et al., 2020), and in particular, can be regarded as a generalization of the scalar mean estimation scheme (Cheu et al., 2019) to d-dim mean estimation.

E.1. Algorithm of Shuffled SQKR

Algorithm 4 Shuffled SQKR

Input: users' data $x_1, \ldots, x_n$, local-DP parameter ε_0, communication parameters b_0, T
Output: mean estimator $\hat{\mu}$
for round $k \in [T]$ **do**
  for user $i \in [n]$ **do**
    Sample $s(i,1), \ldots, s(i,b_0) \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}[d]$
    Sample $Z \sim \mathsf{Bern}\left(\frac{e^{\varepsilon_0}}{e^{\varepsilon_0}+2^{b_0}-1}\right)$
    if Z=1 **then**
      Set $Y(i,1), \ldots, Y(i,b_0) \leftarrow x_i(s(i,1)), \ldots, x_i(s(i,b_0))$
    else
      Sample $Y(i,1), \ldots, Y(i,b_0) \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}\{-c, c\}$
    end if
    Send $Y(i,1), \ldots, Y(i,b_0)$ and $s(i,1), \ldots, s(i,b_0)$ to shuffler
  end for
  Shuffler samples a permutation $\pi \sim \mathsf{Unif}\{f : [n] \to [n] \text{ bijective}\}$
  for $j \in [b_0]$ **do**
    Shuffler sends $Y(\pi(1), j), \ldots, Y(\pi(n), j)$ and $s(\pi(1), j), \ldots, s(\pi(n), j)$ to server
  end for
  $\hat{\mu}^{(k)} \leftarrow \frac{d}{nb_0} \frac{e^{\varepsilon_0}+2^{b_0}-1}{e^{\varepsilon_0}-1} \sum_{i=1}^{n} \sum_{j=1}^{b_0} Y(\pi(i), j) e_{s(\pi(i),j)}$
end for
Return $\hat{\mu} := \frac{1}{T} \sum_{k=1}^{T} \hat{\mu}^{(k)}$

E.2. Proof of Theorem E.3

Each round $x^n \mapsto \hat{\mu}^{(k)}$ of Algorithm 4 implements the private-coin SQKR scheme of (Chen et al., 2020), achieving the communication cost and error as stated in Lemma E.2.

Lemma E.6 (SQKR (Chen et al., 2020))**.** *For all $\varepsilon_0 > 0, b_0 > 0$, the random mapping $x_i \mapsto y(i,1), \ldots, y(i,b_0), s(i,1), \ldots, s(i,b_0)$ in Algorithm 4 is $(\varepsilon_0, 0)$-LDP and has output that can be communicated with $b_0 \log(d)$ bits, and the $\hat{\mu}^{(k)}$ computed from $y(i,1), \ldots, y(i,b_0), s(i,1), \ldots, s(i,b_0)$ is an unbiased estimator satisfying*

$$\max_{x^n} \mathbb{E}\left[\left\|\mu\left(x^n\right) - \hat{\mu}^{(k)}\left(x^n\right)\right\|_2^2\right] = O\left(\frac{c^2 d}{n \min\left(\varepsilon_0^2, \varepsilon_0, b_0\right)}\right). \tag{15}$$

We now characterize the error performance of Algorithm 4 for general choices of parameters that satisfy privacy and communication constraints.

Proposition E.7. *For all $\varepsilon > 0, b > 0, n > 0$, with any arbitrary choice of*

$$\delta_1 \in \left(e^{-n}, 1\right] \tag{16}$$
$$\delta_2 \in (0, 1], \tag{17}$$

there exists a choice of parameters ε_0, b_0, T such that Algorithm 4 is $(\varepsilon, T\delta_1 + \delta_2)$-DP, uses no more than b bits of communication, and

$$\max_{x^n} \mathbb{E}\left[\|\mu - \hat{\mu}\|_2^2\right] = O\left(\max\left(\frac{c^2 d \log(d) b_0}{nb}, \frac{c^2 d \log(1/\delta_1)\left(\log(1/\delta_2) + \varepsilon\right)}{n^2 \varepsilon^2}\right)\right). \tag{18}$$

Proof. For arbitrary choice of

$$b_0 < \log\left(\frac{n}{16 \log(2)}\right), \tag{19}$$

it suffices to choose

$$T = \left\lfloor \frac{b}{(\log_2(d)+1)b_0} \right\rfloor \tag{20}$$

$$\varepsilon_0 = O\left(\min\left(1, \frac{\varepsilon\sqrt{n}}{\sqrt{T\log(1/\delta_1)\left(\log(1/\delta_2)+\varepsilon\right)}}\right)\right). \tag{21}$$

Since it takes b_0 bits to send $y(i,1),\dots,y(i,b_0)$ and $\log_2(d)$ bits to send each of $s(i,1),\dots,s(i,b_0)$, and this is done T times, Algorithm 4 using less than b bits is immediate from the choice of T.

Applying Lemma E.6, by construction the mapping from each x_i to $y(i,1),\dots,y(i,b_0)$ is $(\varepsilon_0, 0)$-LDP. By assumption

$$\delta_1 > e^{-n/16e} > e^{-n}, \tag{22}$$

the inequality

$$1 < \log\left(\frac{n}{16\log(2/\delta_1)}\right) \tag{23}$$

is satisfied. Then the choice of

$$\varepsilon_0 \leq 1 \tag{24}$$

also satisfies $\varepsilon_0 \leq \log\left(\frac{n}{16\log(2/\delta)}\right)$, so by Lemma E.1 the mapping $x^n \mapsto \hat{\mu}^{(k)}$ is $(\varepsilon_1, \delta_1)$-DP. where

$$\varepsilon_1 = O\left(\frac{\varepsilon_0\sqrt{\log(1/\delta_1)}}{\sqrt{n}}\right). \tag{25}$$

Since the output of Algorithm 4 is a function of $\left(\hat{\mu}^{(1)},\dots,\hat{\mu}^{(T)}\right)$, by B.4 it suffices to have

$$\varepsilon_1 = O\left(\min\left(1, \frac{\varepsilon}{\sqrt{T(\log(1/\delta_2)+\varepsilon)}}\right)\right) \tag{26}$$

for Algorithm 4 to be $(\varepsilon, T\delta_1+\delta_2)$-DP. The first inequality follows from the assumption of $\delta_1 > e^{-n}$ and choice of $\varepsilon_0 = O(1)$, and the second from choice of

$$\varepsilon_0 = O\left(\frac{\varepsilon\sqrt{n}}{\sqrt{T\log(1/\delta_1)\left(\log(1/\delta_2)+\varepsilon\right)}}\right). \tag{27}$$

Since $\varepsilon_0 \leq 1 \leq b$, we have $\min(\varepsilon_0^2, \varepsilon_0, b) = \varepsilon_0^2$. Applying Lemma E.6,

$$\max_{x^n} \mathbb{E}\left[\|\mu - \hat{\mu}\|_2^2\right] = \frac{1}{T}\max_{x^n}\mathbb{E}\left[\left\|\mu - \hat{\mu}^{(1)}\right\|_2^2\right] \tag{28}$$

$$= O\left(\frac{d}{Tn\varepsilon_0^2}\right) \tag{29}$$

$$= O\left(\max\left(\frac{d}{Tn}, \frac{d\log(1/\delta_1)\left(\log(1/\delta_2)+\varepsilon\right)}{n^2\varepsilon^2}\right)\right). \tag{30}$$

Substituting the choice of T gives the desired result. □

To show Theorem E.3, it suffices to choose

$$b_0 = 1 \tag{31}$$

$$\delta_1 = \frac{\delta}{2T} \tag{32}$$

$$\delta_2 = \frac{\delta}{2}, \tag{33}$$

which requires $n > 16e\log(2) \approx 30.14$ due to (19), and apply the previous proposition.

E.3. Rényi-DP for Shuffled SQKR

We can use the following result for Rényi-DP (RDP) guarantees for Algorithm 4.

Lemma E.8 ((Feldman et al., 2023) Corollary 4.3)**.** *Let $\mathcal{M}_i$ be an independent $(\varepsilon_0, 0)$-LDP mechanism for each $i \in [n]$ with $\varepsilon_0 \leq 1$ and π be a random permutation of $[n]$. Then for any $\alpha < \frac{n}{16\varepsilon_0 \exp(\varepsilon_0)}$, the mechanism*

$$\mathcal{S} : (x_1, \ldots, x_n) \mapsto \left(\mathcal{M}_1\left(x_{\pi(1)}\right), \ldots, \mathcal{M}_n\left(x_{\pi(n)}\right)\right)$$

is $(\varepsilon(\alpha), \delta)$-RDP where

$$\varepsilon(\alpha) = O\left(\alpha\left(1 - e^{-\varepsilon_0}\right)^2 \frac{e^{\varepsilon_0}}{n}\right). \tag{34}$$

Applying Lemma E.6, by construction the mapping from each x_i to $y(i, 1), \ldots, y(i, b_0)$ is $(\varepsilon_0, 0)$-LDP. By Lemma E.8, the mapping $x^n \mapsto \hat{\mu}^{(k)}$ is (ε_1, α)-RDP where

$$\varepsilon_1 = O\left(\alpha\left(1 - e^{-\varepsilon_0}\right)^2 \frac{e^{\varepsilon_0}}{n}\right) \tag{35}$$

By composition, Algorithm 4 is $(T\varepsilon_1, \alpha)$-RDP.

F. Additional Experiments

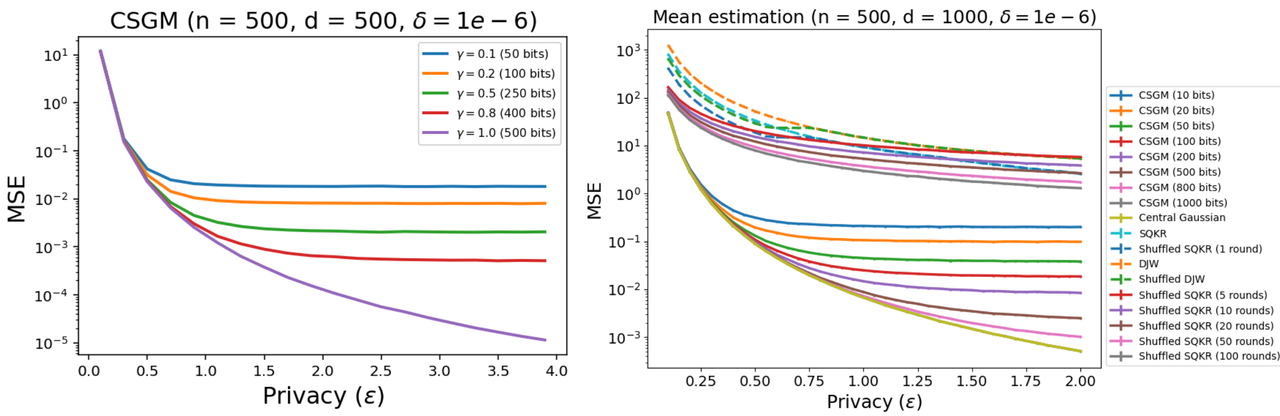
In this section, we empirically evaluate our mean estimation scheme (CSGM) from Section 4, examine its privacy-accuracy-communication trade-off, and compare it with other DP mechanisms (including the shuffling-based mechanism introduced in Section 4.2).

Setup. For a given dimension d, and number of samples n, we generate local vectors $X_i \in \mathbb{R}^d$ as follows: let $X_i(j) \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{d}}\left(2 \cdot \mathsf{Ber}(0.8) - 1\right)$ where $\mathsf{Ber}(0.8)$ is a Bernoulli random variable with bias $p = 0.8$. This ensures $\left\|X_i\right\|_\infty \leq 1/\sqrt{d}$ and $\left\|X_i\right\|_2 \leq 1$, and in addition, the empirical mean $\mu\left(X^n\right) \coloneqq \frac{1}{n}\sum_i X_i$ does not converge to 0. Note that as our goal is to construct an unbiased estimator, we did not project our final estimator back to the ℓ_∞ or ℓ_2 space as the projection step may introduce bias. Therefore, the ℓ_2 estimation error can be greater than 1. We account for the privacy budget with Rényi DP (Mironov, 2017) and the privacy-amplification by subsampling lemma in (Zhu & Wang, 2019) and convert Rényi DP to (ε, δ)-DP via (Canonne et al., 2020).

Privacy-accuracy-communication trade-off of CSGM. In the first experiment (left of Figure 2), we apply Algorithm 1 with different sampling rates γ, which leads to different communication budgets ($b = \gamma d$). Note that when $\gamma = 1$, the scheme reduces to the central Gaussian mechanism without compression. In Figure2, we see that with a fixed communication budget, CSGM approximates the central (uncompressed) Gaussian mechanism in the high privacy regime (small ε) and starts deviating from it when ε exceeds a certain value. In addition, that value of ε depends only on sample size n and the communication budget b and not the dimension d as predicted by our theory: recall that the compression error dominates the total error, and hence the performance starts to deviate from the (uncompressed) Gaussian mechanism when $b = o(n\varepsilon^2)$, a condition that is independent of d. Observe, for example, that when $b = 50$ bits, the Gaussian mechanism starts outperforming CSGM at $\varepsilon \geq 0.5$ for both $d = 500$ and $d = 5000$. Hence, for $\varepsilon \approx 0.5$ CSGM is able to provide 10X compression when $d = 500$, but 100X compression when $d = 5000$ without impacting MSE.

Comparison with local and shuffle DP. Next, we compare the CSGM with local and shuffled DP for $d = 10^3$ and $n = 500$. For local DP, we consider the private-coin SQKR scheme introduced in Section 4.2 which uses $\lceil \log d \rceil + 1) T = 11T$ bits for T shuffling rounds and DJW (Duchi et al., 2013) which is known to be order-optimal when $\varepsilon = O(1)$ (but is not communication-efficient). For shuffle-DP, we apply the amplification lemma in (Feldman et al., 2022a) to find the corresponding local ε_0 (see Section 4.2 for more details) and simulate both SQKR and DJW as the local randomizers

The MSEs of all mechanisms are reported in the right of Figure 2. Our results suggest that for a fixed communication budget (say, 10 bits), the practical performance of CSGM significantly outperforms shuffled-DP mechanisms, including the shuffled SQKR and DJW, eventhough they have the same order-wise guarantees theoretically. In addition, the amplification gain of single-round shuffling diminishes fast as ε increases. Indeed, when $\varepsilon \geq 0.8$, we observe no amplification gain compared to the pure local DP.



(a) (b)

Figure 2. MSEs of CSGM (Algorithm 1) and shuffle LDP schemes.

Benefits of multi-message shuffling. Figure 3 illustrates separation between Algorithm 4 and LDP schemes. Algorithm 4 achieves error decreasing quadratically with n as guaranteed by Theorem E.3. With only one round of shuffling, there is separation from the LDP scheme only when n is sufficiently large, and thus order-optimal error performance only occurs for large n (or equivalently small ε). This problem is avoided with multiple rounds of shuffling.

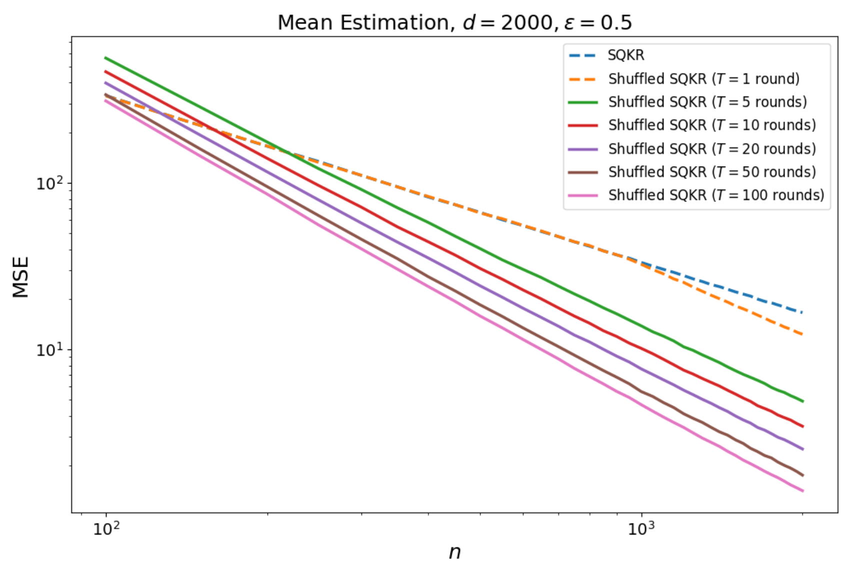


Figure 3. Comparison of MSE vs. number of clients n for LDP scheme (SQKR) and shuffled SQKR. For shuffled SQKR, we set $b_0 = 1$ and choose ε_0 using results in Section E.3. Communication cost is $\lceil(\log_2(2000) + 1)\rceil = 12$ bits per round.

Benefits of coordinate pre-selection. Figure 4 compares the performance of CSGM with and without coordinate pre-selection. In this regime coordinate pre-selection improves performance for all b. As predicted by Corollary 4.1 and Theorem 4.3, the MSE decreases with b but is effectively constant for sufficiently high b where the privacy term dominates. We can determine the communication cost needed for order-optimal central DP error performance to be the b at which the MSE is within some fixed constant factor away from the limiting value. We see that the communication cost increases with dimension d with the vanilla CSGM scheme, but a dimension-free communication cost is achieved with coordinate pre-selection.

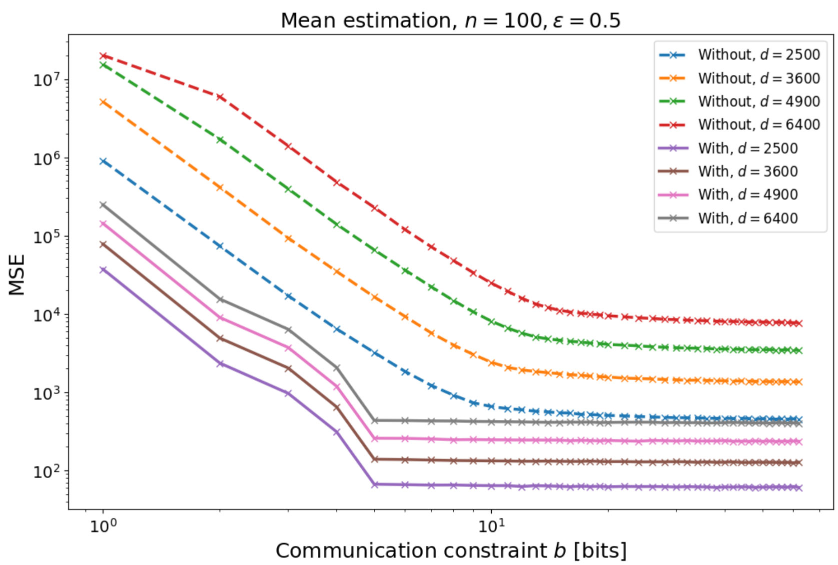


Figure 4. CSGM with and without coordinate pre-selection using $d' = 833$.