Differentially Private Distributed Mean Estimation with Malicious Security

Laasya Bangalore[1], Albert Cheu*[2], and Muthuramakrishnan Venkitasubramaniam[1]

[1]Department of Computer Science, Georgetown University
[2]Google

1 Introduction

Distributed mean estimation (DME) is a fundamental and important task as it serves as a sub-routine in convex optimization, aggregate statistics, and, more generally, federated learning. In many applications, clients are contributing input vectors that contain sensitive information. Thus, we should perform DME in a privacy-preserving manner. A surge of recent work has produced cryptographic protocols for secure aggregation, with varying privacy and security guarantees.[10, 4, 3, 2, 16].

Secure multiparty computation (MPC) is a viable tool for DME as it enables distributed computation of arbitrary tasks while simultaneously guaranteeing *privacy* where nothing beyond the output of the computation is revealed and *correctness* where the output is correct according to the specified function. Furthermore, MPC provides these guarantees even in the presence of an adversary that can control a subset of the parties and launch a coordinated attack on the protocol. While MPC shows *how to compute* in a distributed environment with the best possible security, there are two issues MPC fails to address: (1) it does not prevent adversaries from setting the inputs of corrupted parties arbitrarily thereby affecting the accuracy of the computation and (2) it does not specify *what to compute* and in many cases the underlying function (eg, sum of the inputs) itself can leak information of parties' inputs.

Towards mitigating the first issue, the Prio system by Corrigan-Gibbs and Boneh [10] designed a robust secure aggregation protocol. Prio is widely used by Apple, Google, Internet Services Research Group (ISRG), and Mozilla. For example, Mozilla uses a modified version of Prio to collect web usage statistics privately. In the Prio architecture, a set of clients holding private inputs delegate the task of aggregation to a set of servers. The Prio protocol preserves the privacy of an honest client's input even in the presence of a semi-honest (passive) adversary that corrupts an arbitrary subset of the servers. A key feature in their work is the *robustness* guarantee that protects the system from "faulty" inputs. More precisely, given some polynomial-time computable predicate P their system is able to weed out *bad* inputs, i.e. those that do not satisfy the predicate P via a "input certification" mechanism. Another attractive feature of their work is that the clients only need to send a single (i.e. non-interactive) message to all the servers. Such a feature allows the clients to participate over weak networks. The main drawback of Prio is that it only tolerates the semi-honest corruption of the servers and incurs large communication costs between the clients and servers[1].

A follow-up work by Talwar [16] improves the efficiency of Prio under the same threat model and architecture by making two relaxations. First, the robustness guarantee is relaxed to approximate robustness where invalid inputs could be accepted with a small (yet, non-negligible) probability. Second, the security guarantee allows differentially-private leakage of the honest clients' inputs in addition to the output of the computation.[2]

*Work done while a postdoctoral fellow at Georgetown University. Visiting researcher at Google employed via Magnit.

[1] The communication between every client and each server is proportional to the "circuit" size of the predicate P.

[2] In contrast, standard MPC security guarantees nothing beyond the output of the computation is revealed.

Another line of work, initiated by Bonawitz et al. [4] considers secure aggregation in the so-called star topology where the central (output) party is connected with all the clients and learns the result of the secure aggregation. The main feature of their construction is that the protocol execution goes to completion even when a subset (up to a threshold) of nodes drop out or behave maliciously. This is referred to as *guaranteed output delivery* in the MPC literature. A series of works [2, 6, 9] have refined this approach to additionally achieve robustness. One drawback that persists in this line of work is that they achieve security only against semi-honest corruption of the center node. [2] shows how to achieve security against malicious corruption of the center node but this comes with the cost of degrading the guaranteed output feature to *security with abort*.

Finally, all prior works only address the first of the two issues in MPC-based protocols discussed above and do not prevent leakage from the underlying function. Addressing this requires imposing a stronger guarantee of *differentially privacy* [11]. While prior works suggest using differentially private mechanisms, none of the works concretely design or analyze with this guarantee. In particular, no prior work shows how to combine robustness with differential privacy.

Our Results. In this work, we address the two aforementioned issues by first constructing a robust secure aggregation protocol. Then, we design a differentially private mechanism for DME that uses the robust secure aggregation protocol developed in the first step as a black box.

Our secure aggregation protocol is robust with respect to predicate $P(\cdot)$, obtains privacy of honest clients' inputs in the presence of a malicious adversary and also achieves guaranteed output delivery. Further, our protocol only utilizes lightweight cryptography based on symmetric-key (i.e., collision-resistant hash functions) and is black-box in the underlying cryptographic primitives.

Theorem 1.1. *Let $h > 0$ and $P : \mathbb{F}^d \to \{0, 1\}$ be an arbitrary predicate. The protocol Σ_P involves $\mathsf{n_c}$ clients, $\mathsf{n_s}$ servers, and an output party $\mathcal{O}$ and securely computes the summation $\sum X_i \cdot P(X_i)$ where $X_i \in \mathbb{F}^d$ is the input vector of the i^{th} client. This protocol tolerates a malicious rushing adversary that can actively corrupt an arbitrary number of clients, up to t servers, the output party and drops out at most $\mathsf{d_s}$ servers when $\mathsf{n_s} > 3t + \mathsf{d_s} + 1$. Further, Σ_P achieves guaranteed output delivery.*

Moreover, the total communication complexity and between a client and each server is $\tilde{O}(d + \sqrt{\mathsf{n_s} \cdot d + |P|} + \mathsf{n_s} \cdot h/|\mathbb{F}|)$ field elements, and among the servers is $\tilde{O}(\mathsf{n_c} \cdot \mathsf{n_s}^2 \cdot d)$ field elements[3] *where $d >> \mathsf{n_s}$ and h is the output length of the hash function.*

In our DP DME protocol, clients execute a pre-processing algorithm PRE before communicating with Σ_P. Its output is differentially private due to the accumulation of noise introduced by PRE and resists malicious input due to the choice of P.

Theorem 1.2 (Upper bound for DP DME)**.** *Let (ε, δ) be target privacy parameters. Assume client data X_i $(i \in [\mathsf{n_c}])$ belongs to the Euclidean unit ball $\mathcal{B}^d$. There is a predicate P and pre-processing algorithm* PRE *such that, when all $\mathsf{n_c}$ clients evaluate* PRE *and communicate the results to Σ_P, the composition is (ε, δ)-differentially private. Moreover, the output party can post-process the aggregated value to obtain an unbiased estimate of the mean $\frac{1}{\mathsf{n_c}} \sum X_i$. Its variance is $O(\frac{d}{\varepsilon^2 \mathsf{n_c}^2} \log \frac{1}{\delta})$. If $t = O(\mathsf{n_c})$ clients are malicious, then we achieve $(O(\varepsilon), O(\delta))$-differential privacy*[4] *and the squared bias is $\tilde{O}(\frac{t^2}{\mathsf{n_c}^2} \cdot \frac{d}{\varepsilon^2 \mathsf{n_c}})$.*

In the fully honest case ($t = 0$), the above variance bound is asymptotically the same as the classic Gaussian mechanism. In the malicious case ($t > 0$), it is conceivable that smaller squared bias is achievable with other protocols that "wrap" alternative PRE, POST around Σ_P. We show this goal comes with a price for a natural class of PRE, POST algorithms.

[3] $\tilde{O}(\cdot)$ ignores polylogarithmic factors in $\mathsf{n_s}, \mathsf{n_c}, \mathbb{F}$.

[4] It is possible to re-parameterize the protocol to guarantee a target ε for, say, 2/3 corruptions instead of 0 corruptions. However, this comes at the price of increased error since each PRE must introduce slightly more noise for privacy.

Theorem 1.3 (A Lower bound for Wrapped DP DME)**.** *For any pair of pre- and post-processing algorithms that preserves $(\varepsilon, \delta = o(1/\mathsf{n_c}))$-differential privacy where additionally the post-processing algorithm is affine, the resulting system* either *produces biased estimates when there are no malicious clients* or *there is an explicit attack by t malicious clients that results in $\frac{t^2}{\mathsf{n_c}^2} \cdot \frac{d}{\varepsilon^2 \mathsf{n_c}}$ expected squared error.*

Remark 1.4 (Benefits of Modularity)**.** *Our upper and lower bounds for DP are agnostic to the implementation of secure aggregation. We could wrap our* PRE *and* POST *algorithms around any other MPC protocol that performs input-certified secure aggregation—for example, one with improved communication efficiency—and obtain the same error guarantee as in Theorem 1.2. By the same token, replacing our secure aggregation protocol will not circumvent the constraint stated in Theorem 1.3*

2 Our Robust Secure Aggregation Protocol

We are interested in scenarios wherein large corporations, such as Apple, Google, or Meta, deploy powerful servers to securely compute on their users' data.Typically, secure aggregation systems involve a single center node (or server) that efficiently manages a large number of clients by undertaking a bulk of the workload. Nonetheless, many existing systems encounter the following obstacles: (i) *Malicious adversaries:* Effectively managing malicious adversaries poses a significant challenge, as they can arbitrarily deviate from the protocol. In particular, a malicious client could deliberately provide malformed inputs to manipulate the output. Additionally, a malicious center node or server represents a single point of failure and cannot be entirely relied upon to handle sensitive data from a large number of clients. (ii) *Unreliable clients (i.e., client dropouts):* The clients primarily consist of mobile devices, which are inherently deemed unreliable due to their limited capabilities. As a result, it is unrealistic to expect them to engage in multiple rounds of interaction with the servers or handle computationally intensive tasks. (iii) *Handling a large number of clients (Scalability):* The servers are tasked with managing a significant number of clients, potentially numbering in the thousands. Therefore, it is crucial for the server-side costs to scale well in terms of both communication and computation. Specifically, any of the fixes to handle malicious adversaries or dropouts should incur minimal costs per client. For instance, certifying the well-formedness of the inputs involves checking if the input X satisfies some predicate $P(\cdot)$ i.e., $P(X) = 1$. Prior works achieve this by requiring server communication proportional to the size of the predicate[5] per client, which do not scale well for large predicates.

We now provide a high-level overview of our protocol and discuss how we address each of the aforementioned obstacles. As a starting point, we consider a simple protocol based on a semi-honest variant of BGW using the packed secret sharing scheme [12]. The clients share their inputs among the servers who then perform the computation. Note that our protocol leverages the multi-server setting with an honest majority to achieve guaranteed output delivery[6].

To achieve robustness, we enhance this simple protocol with an input certification mechanism using zero-knowledge (ZK) proofs, similar to [10], RoFL [6], and EiFFEL [9]. Specifically, each client generates succinct ZK proofs to prove to each server that the inputs are well-formed (i.e., satisfy a predicate P) and that the input share given to the server is consistent with this well-formed input. The client transmits her input share and (additionally) the associated ZK proof to each server. Our main technical contribution lies in the design of a lightweight protocol for input certification where the communication complexity from the client to the server is proportional to $\sqrt{|P|}$. Also, the communication overhead incurred by the servers after receiving the inputs shares and proofs (from the clients) is independent of the size of the certification predicate $|P|$. This resolves the first and the third obstacles mentioned above.

[5] The size of the predicate $P(\cdot)$ is defined as the number of gates in the circuit representing P.

[6] Our guarantees differ from prior works [5, 3, 6, 2, 9] in the single server setting because their security guarantees do not hold when the server is maliciously corrupt.

To address the second obstacle, we incorporate the following three techniques. First, each client needs to just speak only once i.e., interact with the servers over just one round of communication. Second, clients employ a simple sharing scheme (rather than communication-intensive verifiable secret sharing (VSS)) to share clients' inputs which minimize the clients' communication costs. Third, the clients use lightweight and succinct zero-knowledge proofs to certify well-formedness of their inputs, thereby incurring lower computation costs.

Overall, our robust secure aggregation protocol is concretely efficient and achieves guaranteed output delivery and robustness, and privacy against malicious adversaries.

3 Achieving Differential Privacy

Upper Bound. We now sketch how to perform DP DME by building around Σ_P, a protocol that only adds input that accepted by predicate P. Assume for now that Σ_P accepts infinite-precision inputs. Clients can add independent Gaussian noise $\mathbf{N}(0, (\sigma^2/\mathsf{n_c}) \cdot I)$ to their data vectors, where I is the identity matrix. If all $\mathsf{n_c}$ clients are honest, the aggregated vector has noise distributed as $\mathbf{N}(0, \sigma^2 I)$. It is well-established that this ensures a target parameter ε of differential privacy when $\sigma^2 \approx 1/\varepsilon^2$. If $t > 0$ clients are malicious, the aggregate has less variance than intended. As a result, the effective privacy parameter will not be the same as the target parameter ε but it will follow a smooth function of t, roughly $\varepsilon \cdot \sqrt{\mathsf{n_c}/(\mathsf{n_c} - t)}$.

In reality, we will only be able to operate with finite-precision values. So instead of sampling from $\mathbf{N}(0, \sigma^2/\mathsf{n_c})$ for each coordinate, each client samples from $\mathbf{Bin}(h/\mathsf{n_c}, 1/2)$. Clients simply need to discretize their data in order to match the discrete nature of binomial noise.

Cheu, Joseph, Mao, and Peng show that binomial noise suffices for DP DME [8]. But, due to composition, their error bound has polylogarithmic factors not present in the Gaussian mechanism's bound. We instead quantify how well the binomial approximates the distribution formed by rounding a Gaussian to the nearest integer. The De Moivre-Laplace theorem is a classic asymptotic result: as $h \to \infty$, $\mathbf{Bin}(h, 1/2) - h/2$ approaches $\mathsf{round}(\mathbf{N}(0, \sigma^2 = h/4))$. We derive a variant that allows us to choose h for a target level of approximation: if $h \approx \frac{1}{\varepsilon^2} \log \frac{1}{\delta}$, then the binomial is (ε, δ)-close to the rounded Gaussian, where "(ε, δ)-close" is the condition that appears in approximate differential privacy. The upshot is that greater fidelity of approximation requires a greater magnitude of binomial noise. Clients re-scale their data to match that scale of noise, preventing their signal from being drowned out.

The output party obtains an estimate of the mean by undoing the scaling that the clients performed. When there are no malicious clients, there is no bias because the noise introduced is spherical around 0. The presence of t malicious clients naturally introduces bias but this is bounded by way of input certification: P rejects vectors whose Euclidean norms exceed a threshold computed from a tail bound on the idealized spherical Gaussian.

Comparison with Prior work. DP DME protocols built atop secure aggregation already exist in the literature [13, 1, 7]. The primary distinction of our work is bounding error when input certification is available. In the full version of our work, we also construct a predicate P for the protocol by Chen, Özgür, and Kairouz and repeat the analysis steps [7]. The prior protocols have Rényi and concentrated DP guarantees but we argue our protocol remains competitive with respect to composition: we use Gaussian DP to budget across multiple rounds and only resort to approximate DP to quantify simulation fidelity.

Lower Bound. We finally give intuition for our lower bound (Theorem 1.3). Any DP solution for mean estimation in the ℓ_2 ball must have expected squared ℓ_2 error $\approx d/\varepsilon^2 \mathsf{n_c}$ [14, 15]. If the estimate is unbiased, this is a lower bound on the variance. Now, for a wrapped protocol $\Pi = (\mathsf{PRE}, P, \mathsf{POST})$ where POST is affine, this implies a lower bound on the variance of $\mathsf{PRE}(X_i)$ for any datum X_i. So an "extreme" value must lie inside the support of PRE. This is what a malicious client can send to skew the result.

References

[1] Agarwal, N., Kairouz, P., Liu, Z.: The skellam mechanism for differentially private federated learning. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 5052–5064 (2021), https://proceedings.neurips.cc/paper/2021/hash/285baacbdf8fda1de94b19282acd23e2-Abstract.html

[2] Bell, J., Gascón, A., Lepoint, T., Li, B., Meiklejohn, S., Raykova, M., Yun, C.: ACORN: input validation for secure aggregation. IACR Cryptol. ePrint Arch. p. 1461 (2022), https://eprint.iacr.org/2022/1461

[3] Bell, J.H., Bonawitz, K.A., Gascón, A., Lepoint, T., Raykova, M.: Secure single-server aggregation with (poly)logarithmic overhead. In: Ligatti, J., Ou, X., Katz, J., Vigna, G. (eds.) CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020. pp. 1253–1269. ACM (2020). https://doi.org/10.1145/3372297.3417885, https://doi.org/10.1145/3372297.3417885

[4] Bonawitz, K.A., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: Thuraisingham, B.M., Evans, D., Malkin, T., Xu, D. (eds.) Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017. pp. 1175–1191. ACM (2017). https://doi.org/10.1145/3133956.3133982, https://doi.org/10.1145/3133956.3133982

[5] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: Thuraisingham, B.M., Evans, D., Malkin, T., Xu, D. (eds.) Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017. pp. 1175–1191. ACM (2017). https://doi.org/10.1145/3133956.3133982, https://doi.org/10.1145/3133956.3133982

[6] Burkhalter, L., Lycklama, H., Viand, A., Küchler, N., Hithnawi, A.: Rofl: Attestable robustness for secure federated learning. CoRR **abs/2107.03311** (2021), https://arxiv.org/abs/2107.03311

[7] Chen, W., Özgür, A., Kairouz, P.: The poisson binomial mechanism for unbiased federated learning with secure aggregation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 3490–3506. PMLR (2022), https://proceedings.mlr.press/v162/chen22s.html

[8] Cheu, A., Joseph, M., Mao, J., Peng, B.: Shuffle private stochastic convex optimization. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), https://openreview.net/forum?id=DrZXuTGg2A-

[9] Chowdhury, A.R., Guo, C., Jha, S., van der Maaten, L.: Eiffel: Ensuring integrity for federated learning. In: Yin, H., Stavrou, A., Cremers, C., Shi, E. (eds.) Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022. pp. 2535–2549. ACM (2022). https://doi.org/10.1145/3548606.3560611, https://doi.org/10.1145/3548606.3560611

[10] Corrigan-Gibbs, H., Boneh, D.: Prio: Private, robust, and scalable computation of aggregate statistics. In: Akella, A., Howell, J. (eds.) 14th USENIX Symposium on Networked Systems Design and

Implementation, NSDI 2017, Boston, MA, USA, March 27-29, 2017. pp. 259–282. USENIX Association (2017), https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/corrigan-gibbs

[11] Dwork, C., McSherry, F., Nissim, K., Smith, A.D.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings. Lecture Notes in Computer Science, vol. 3876, pp. 265–284. Springer (2006). https://doi.org/10.1007/11681878_14

[12] Franklin, M., Yung, M.: Communication complexity of secure computation (extended abstract). In: Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing. p. 699–710. STOC '92, Association for Computing Machinery, New York, NY, USA (1992). https://doi.org/10.1145/129712.129780, https://doi.org/10.1145/129712.129780

[13] Kairouz, P., Liu, Z., Steinke, T.: The distributed discrete gaussian mechanism for federated learning with secure aggregation. CoRR **abs/2102.06387** (2021), https://arxiv.org/abs/2102.06387

[14] Kamath, G., Ullman, J.R.: A primer on private statistics. CoRR **abs/2005.00010** (2020), https://arxiv.org/abs/2005.00010

[15] Steinke, T.: Upper and Lower Bounds for Privacy and Adaptivity in Algorithmic Data Analysis. Ph.D. thesis, Harvard University, Cambridge, MA (2016)

[16] Talwar, K.: Differential secrecy for distributed data and applications to robust differentially secure vector summation. CoRR **abs/2202.10618** (2022), https://arxiv.org/abs/2202.10618