Aggregation and Transformation of Vector-Valued Messages in the Shuffle Model of Differential Privacy

Mary Scott, Graham Cormode and Carsten Maple (IEEE Transactions on Information Forensics and Security)

Differential privacy (DP) is a technique that provides a rigorous and provable privacy guarantee for aggregation and release. The Shuffle Model for DP has been introduced to balance the accuracy of local-DP algorithms with the privacy risks of central-DP. In this work we firstly provide a single message protocol for the summation of real vectors in the Shuffle Model. We then improve this bound through the implementation of a discrete Fourier Transform, greatly minimizing the perturbation error.

Our first contribution is a new protocol in the Shuffle Model for the private summation of vector-valued messages. This protocol extends an existing result from Balle et al. [3] by permitting the $n$ users to each submit a vector of real numbers rather than being restricted to submitting a scalar.

The local randomizer applies a generalized randomized response mechanism that:

- returns the true message $x$, with probability $1 - \gamma$,
- and a uniformly random message with probability $\gamma$.

It is necessary to find an appropriate $\gamma$ to optimize the proportion of random messages that are submitted, and also guarantee DP. The resulting estimator is unbiased and has normalized mean squared error (MSE) $O(\exp(-5\gamma^2))$, where $d$ is the dimension of each vector.

**Related Work**

The majority of research in the field of DP has focused on two contrasting models:

- **Centralized Model**: users submit their sensitive personal information directly to a trusted central data collector, who adds random noise to the raw data to provide DP before assembling and analyzing the aggregated results.

- **Local Model**: DP is guaranteed when each user applies a local randomizer to add random noise to their data before it is submitted. There is no need for a trusted party, but the level of noise required per user for the same privacy guarantee is much higher.

In recent years researchers have tried to create intermediate models that reap the benefits of both. In 2019, Cheu et al. [2] formalized the Shuffle Model to connect an additional shuffle step to the Local Model. This step is shown by the grey rectangle in Figure 1.

**Vector Sum in the Shuffle Model**

Our first contribution is a new protocol in the Shuffle Model for the private summation of vector-valued messages. This protocol extends an existing result from Balle et al. [3] by permitting the $n$ users to each submit a vector of real numbers rather than being restricted to submitting a scalar.

The local randomizer applies a generalized randomized response mechanism that:

- returns the true message $x$, with probability $1 - \gamma$,
- and a uniformly random message with probability $\gamma$.

It is necessary to find an appropriate $\gamma$ to optimize the proportion of random messages that are submitted, and also guarantee DP. The resulting estimator is unbiased and has normalized mean squared error (MSE) $O(\exp(-5\gamma^2))$, where $d$ is the dimension of each vector.

**Transforming Summation in the Shuffle Model**

An orthonormal transformation can be used to further tighten the bound we have obtained for private summation. The discrete Fourier transform (DFT) concentrates information about signals with a particular pattern more clearly, this graph has been plotted using a randomly generated synthetic dataset with a sinusoidal dependence on each coordinate.

Note that in Figure 2 the scale on the y-axis is logarithmic. Although the perturbation error initially looks much larger than its baseline counterpart, isolating these errors in Figure 3 clearly shows that the opposite is true for the ECG heartbeat categorization dataset. The same conclusion can be made for the synthetic dataset.

**Conclusion**

The experiments above confirm that picking $\gamma = 1$ and $k = 3$ serves to minimize the error. The lines of best fit, in Figure 3 for example, confirm the dependencies on the parameters $m, d, \epsilon$ and $\delta$.

We have seen via both theory and experiments that combining our new private summation protocol with a DFT reduces the MSE significantly, from a dependence on $d^{1/3}$ to $d^{1/6}$.

**References**


