The Distributed Discrete Gaussian Mechanism for Federated Learning with Secure Aggregation

Peter Kairouz, Ziyu Liu, Thomas Steinke
{kairouz, klz, steinke}@google.com
https://github.com/google-research/federated/tree/master/distributed_dp

Google Research

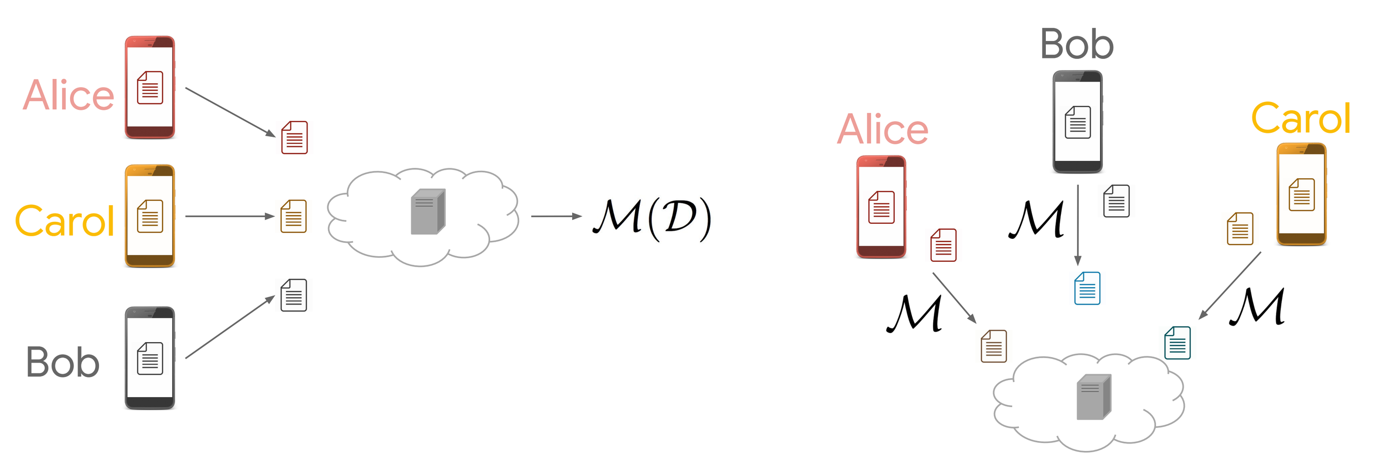
Background: Differentially Private FL

- While **Federated Learning (FL)** ensures raw data are kept decentralized, it **may not** provide **formal privacy guarantees**.
- **Differentially Private FL**: client updates (e.g. gradients) are **clipped** and **noised** appropriately to give **quantifiable, user-level** DP guarantees.

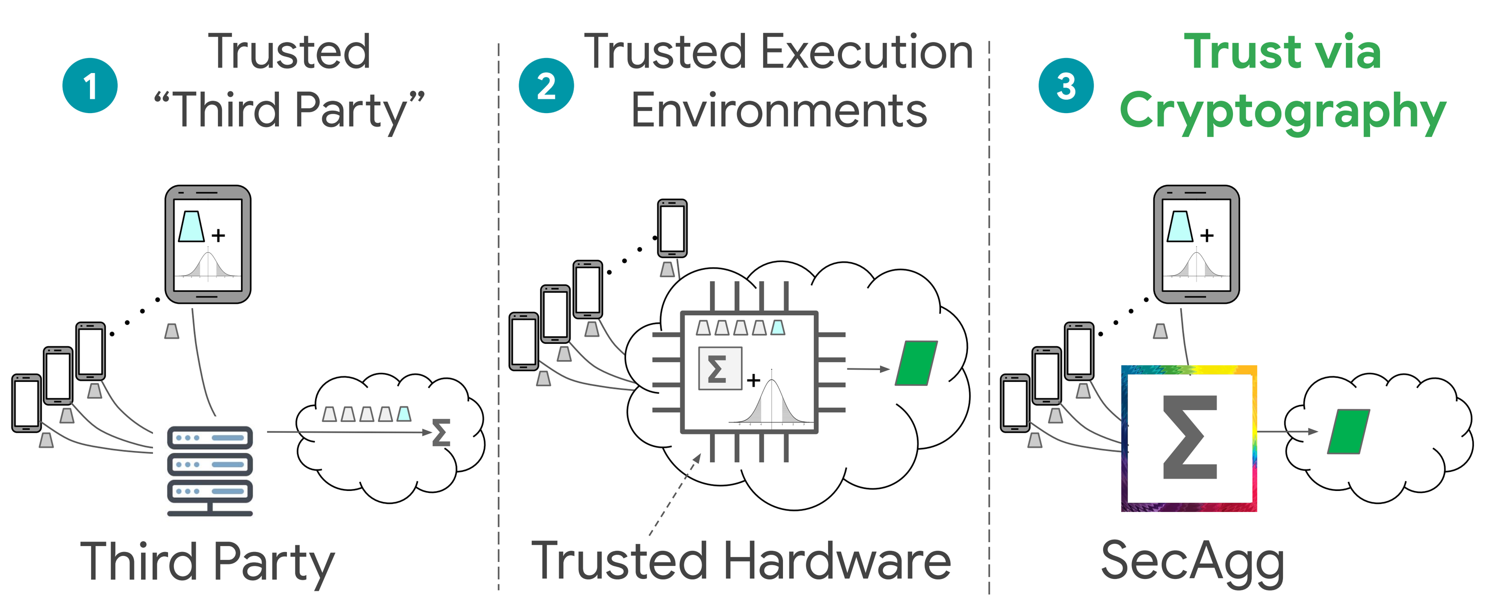
Privacy Models

Central DP: Noise@Server
- Full trust on server
- Better utility

Local DP: Noise@Clients
- No trust on server
- Poor utility



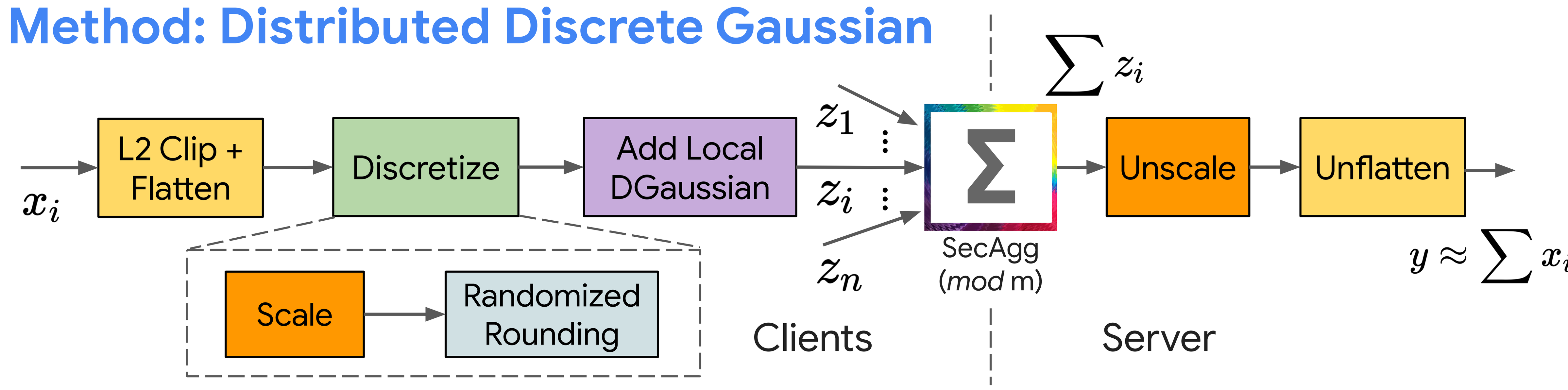
Distributed DP (this work)
Aims to achieve the utility of Central DP without fully trusting the server by "distributing" trust:



Some Challenges

- Secure Aggregation (SecAgg) operates on a **finite group** (integers with modular arithmetic)
  - Need **discrete** DP mechanisms
- **Sums** of Discrete Gaussians ≠ Discrete Gaussians
  - Need to carefully analyze the effects on DP
- **Communication** efficiency is vital for practical FL
  - Need to consider the trade-off against **privacy** and **utility** (both modular & quantization errors)

Method: Distributed Discrete Gaussian



Summary: An end-to-end system for differentially private FL combining **compression, SecAgg,** and **local noising** that matches the privacy / accuracy of **Central DP.**

Procedure

1. **L2 Clipping:** Initial bound on the client vector L2 sensitivity c
2. **Flattening**: Random unitary transform to spread values across vector dimensions
   - Controls the L-inf norm ⟶ Lower quantization errors / Less modular wrap-around
3. **Discretization:** Round input values to the discrete grid (rounding granularity γ)
   - Corresponds to scaling by $1/\gamma$ + rounding to integers
   - **Scaling**: Smaller γ ⟶ Less rounding errors, but larger values (more communication)
   - **Randomized rounding:** Unbiased discretization (e.g. 4.2 to 4/5 with 80%/20% prob)
   - **Norm inflation:** Rounding may increase norm ⟶ more DP noise for same privacy
   - **Conditional rounding:** we give a tighter probabilistic bound and retry rounding until the norm is smaller (less DP noise):

     β: rounding bias d: vector dimension

$$\Delta_2^2 := \min\left\{ \begin{array}{c} c^2 + \gamma^2 d/4 + \sqrt{2\log(1/\beta)} \cdot \gamma \cdot \left(c + \gamma\sqrt{d}/2\right), \\ (c+\gamma\sqrt{d})^2 \end{array} \right\}$$

4. **Local Noising**
   - Each client adds their own local discrete Gaussian noise
   - We give a **tight bound** on the sums of discrete Gaussians, which leads to **extremely close** privacy guarantees to **central DP** (central continuous/discrete Gaussian noise):

Theorem 11 (Convolution of two Discrete Gaussians). *Let $\sigma, \tau \geq \frac{1}{2}$. Let $X \leftarrow \mathcal{N}_{\mathbb{Z}}(0,\sigma^2)$ and $Y \leftarrow \mathcal{N}_{\mathbb{Z}}(0,\tau^2)$ be independent. Let $Z = X+Y$. Let $W \leftarrow \mathcal{N}_{\mathbb{Z}}(0,\sigma^2+\tau^2)$. Then*

$$\mathrm{D}_{\pm\infty}(Z\|W) = \sup_{z\in\mathbb{Z}} \left|\log\left(\frac{\mathbb{P}[Z=z]}{\mathbb{P}[W=z]}\right)\right| \leq 5 \cdot e^{-2\pi^2/(1/\sigma^2+1/\tau^2)}.$$

Privacy Guarantee
($\varepsilon^2/2$-concentrated DP, n clients):

$$\tau := 10 \cdot \sum_{k=1}^{n-1} e^{-2\pi^2 \frac{\sigma^2}{\gamma^2} \cdot \frac{k}{k+1}}$$

$$\varepsilon := \min\left\{ \begin{array}{c} \sqrt{\frac{\Delta_2^2}{n\sigma^2} + \frac{1}{2}\tau d} \\ \frac{\Delta_2}{\sqrt{n}\sigma} + \tau\sqrt{d} \end{array} \right\}$$

5. **SecAgg:** Securely sums locally clipped, scaled, rounded, and noised client vectors
   - SecAgg group size $m = 2^B$ determines the communication bit-width (for the sum)
   - Scaling $(1/\gamma)$ is chosen to keep modular wrapping infrequent (often <0.05% prob)
6. **Server Post-Processing:** Unscale and undo the flattening transform
   - Extension: may optionally collect metrics to update c and γ for the next iteration

Empirical Results

Stack Overflow Next Word Prediction
>10^8 training question/answer sentences grouped by >340k Stack Overflow users

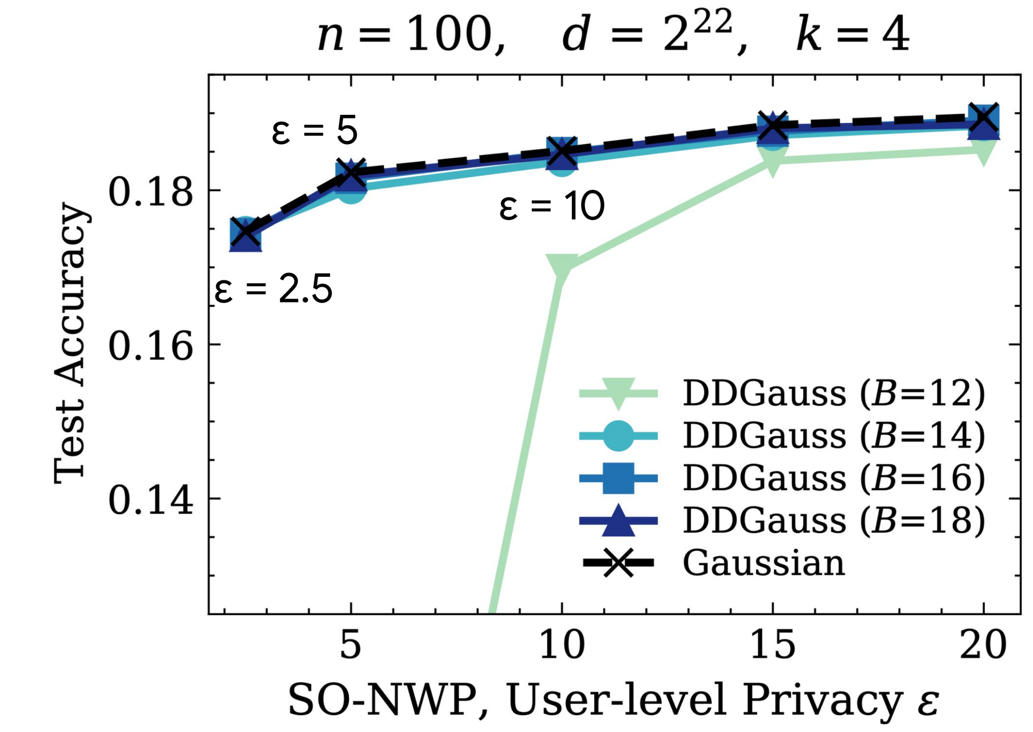


Fig. 1. Our method matches the **central continuous Gaussian** if bit-width B is sufficient (≥ 14). $\delta = 10^{-6}$.

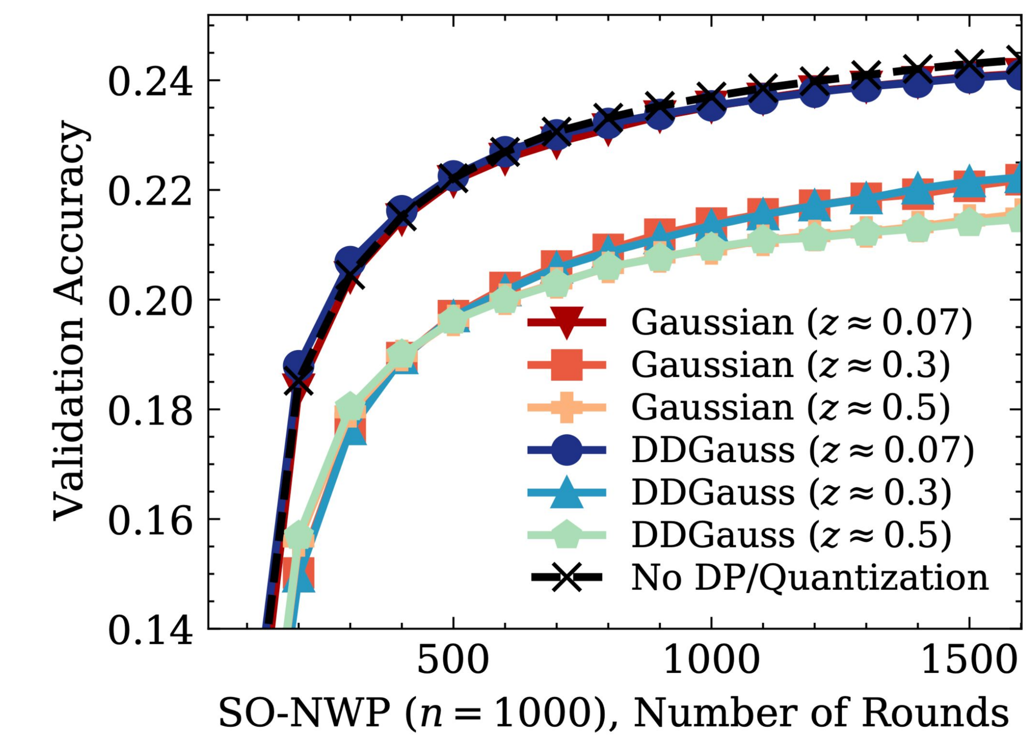


Fig. 2. DDG works in production-scale (1000 clients) and low-noise (utility-first) settings. z: noise multiplier.

See full version (arXiv:2102.06387) for more!

Conclusion & Future Directions

- Distributed DP achieves accuracy similar to central DP with only 16 bits per value
- **Next steps:** (a) Discrete Fourier Transform instead of Walsh-Hadamard Transform for better compression efficiency, (b) lower bound on communication, privacy, and accuracy trade-offs, (c) exploring the role of sparsity under distributed DP.