Skellam Mixture Mechanism: A Novel Approach to Federated Learning with Differential Privacy

Ergute Bao, Yizheng Zhu, Xiaokui Xiao, Yin Yang, Beng Chin Ooi, Benjamin Tan and Khin Mi Mi Aung
Federated Learning with SGD

\[ \sum_{i=1}^{n} g_i \]

Aggregation

Server

Language model, Image classification, SVM, etc..
Privacy concerns

During the training process, gradients leak the training dataset [Zhu et al. ICML’ 18].
Privacy concerns

The server can be untrusted

Party $i$
Privacy concerns

Final model parameters remember the training dataset [Shokri et al. S&P’ 18, Carlini et al. Usenix’ 19].

Membership of the dataset (e.g. dataset of a rare disease)
Goal

A mechanism that protects individual privacy
• throughout the training process (Secure Aggregation)
• for the final model parameters (Differential Privacy)

Model accuracy should approach that in the centralized setting, which is seen as the lower bound for the distributed setting.
Differential Privacy [DMNS. TCC’ 06]

For any neighboring input databases $D_1$ and $D_2$, if mechanism $\mathcal{K}$’s output distributions are similar, then we say mechanism $\mathcal{K}$ is differentially private.

The similarity is quantified by $\epsilon$.

Credit: Cynthia Dwork [https://slideplayer.com/slide/6661339/]
Applying DP to FL with SGD

Party $i$:

- Injects Gaussian noise $z_i \sim \mathcal{N}(0,\sigma^2 \mathbf{I}_d)$ to original $g_i$. [Abadi et al. CCS’ 16]
- Scale of noise: $\sigma = \|g_i\|_2 / \epsilon$.

$$
\text{Aggregation} \\
\sum_{i=1}^{n} (g_i + z_i)
$$

Model update
Applying DP to FL with SGD

Participant $i$:

- Injects Gaussian noise $z_i \sim \mathcal{N}(0, \sigma^2 I_d)$ to original $g_i$. [Abadi et al. CCS’ 16]
- Scale of noise: $\sigma = \|g_i\|_2/\epsilon$.

Calibrating noise to the sensitivity of data [DMNS. TCC ’06]

To hide the private gradient, the noise must be as large as the gradient itself.
Trade-off between privacy and accuracy

• The amount of each individual noise determines the privacy level $\epsilon$.
• The amount of overall noise determines the model accuracy.

\[
\sum_{i=1}^{n} (g_i + z_i)
\]
Secure Aggregation

SecAgg [BIKMMPRSS, CCS’ 17] leverages MPC,
• Computing the sum of private inputs.
• Ensuring that the input is not revealed to any party (including the server).

\[ \sum_{i=1}^{n} g_i \]

Adversary only sees the sum.
Secure Aggregation

Think of SecAgg as a black-box function for securely computing the sum of inputs.

\[
\sum_{i=1}^{n} g_i
\]

Adversary only sees the sum.
Differential Privacy with Secure Aggregation

SecAgg amplifies privacy for individual participants:
- Assume that each participant adds a little i.i.d. Gaussian.
- Sum of Gaussian variates is a **larger** Gaussian (privacy amplification by $n$).

\[ \sum_{i=1}^{n} (g_i + z_i) = \sum_{i=1}^{n} g_i + \sum_{i=1}^{n} z_i \]

Adversary only sees the Noisy Sum.
Differential Privacy with Secure Aggregation

Challenge brought by SecAgg:
• Outputs of participants must be integers (required by MPC).
• We can not directly inject Gaussian noise to the real-valued gradients.
• Motivate new DP mechanisms:
  [Agarwal et al. NeurIPS’ 18], [KLS, ICML’ 21], [AKL, NeurIPS’ 22]
1. Pre-process the gradient $g_i \in \mathbb{R}^d$.
   For each real-valued parameter, say $a + b$ (e.g. $4.55 = 4 + 0.55$)
   • With probability $b$, round to $a + 1$
   • With probability $(1 - b)$, round to $a$
2. Inject integer-valued noise to processed gradient

Expectation of output is $(a + b)$
Existing Solutions

1. Pre-process the gradient \( g_i \in \mathbb{R}^d \).
   For each real-valued parameter, say \( a + b \) (e.g. \( 4.55 = 4 + 0.55 \))
   - With probability \( b \), round to \( a + 1 \) (cause sensitivity increase)
   - With probability \( 1 - b \), round to \( a \)

2. Inject integer-valued noise to processed gradient (of larger norm)
   - The noise is of scale \( (\|g_i\|_2 + \sqrt{d})/\epsilon \)

After rounding, gradients can be more different (requires more noise)
- 0.0001 could be round to 1, hence the rounded sensitivity is 1 instead of 0.0001.
Noise Overhead

- Common scenarios: $\|g_i\|_2 \ll \sqrt{d}$.
- Large DP noise drowns the signal of gradients.
- For integer representation using limited bits, large noise leads to overflow.
Our solution: intuition

We observe:
• Stochastic rounding is random.
• Differential privacy needs random noise.

We should leverage randomness in rounding for DP!!!
Building Block 1: Skellam Noise

• The difference of two independent Poisson variates.
• Looks like an ‘integer-valued’ Gaussian.
• Hence, it works like a Gaussian for DP (we improve existing analysis).
Building Block 2: Mixture of integer noises

\[ a : \text{integer part} \quad b : \text{fractional part} \]

Consider input \( a + b \)

omitted details…

1. Inject mixture of noises
   - With probability \( b \), sample \textit{integer} noise \textit{shifted} by \( a + 1 \)
   - With probability \( 1 - b \), sample \textit{integer} noise \textit{shifted} by \( a \)
Building Block 2: Mixture of integer noises

Consider input $a$ (integer part) + $b$ (fraction part).

1. Inject mixture of noises
   - With probability $b$, sample integer noise shifted by $a + 1$
   - With probability $(1 - b)$, sample integer noise shifted by $a$

No sensitivity overhead, which means tighter privacy guarantee!!

1. Pre-process the input $g_i$
   - With probability $b$, round to $a + 1$ (cause sensitivity increase)
   - With probability $(1 - b)$, round to $a$
2. Inject integer-valued noise to processed gradient (of larger norm)
   - The noise is of scale $(\|g_i\|_2 + \sqrt{d})/\epsilon$
Challenge: Privacy Analysis

Analyze the Rényi divergence of two mixtures of Skellam distributions (more details in our paper):

- Both mixtures consist of $n \cdot 2^d$ individual $d$—dimensional Skellam components.
  - Reduction to two 1—dimensional Skellam components.
- The mixtures & individuals of Skellam distributions are not well understood.
  - New tools for analyzing mixture of Skellams & individual Skellam.
Experiment on MNIST

lower bound    ours    existing solutions.

- • DPSGD
- □ SMM
- ★ Skellam
- △ DDG
- ○ cpSGD

$m = 6$  $m = 8$ bits per parameter  $m = 10$
• Existing solutions for Federated Learning with DP incur large sensitivity & noise overhead, causing utility degradation.

• We propose SMM that directly operates on real-valued input, and outputs an unbiased & integer-valued & private estimate.

• We develop new tools for analyzing mixture and individual Skellam noises for DP.