Compression with Exact Error Distribution for Federated Learning

Mahmoud Hegazy[1] Rémi Leluc[1] Cheuck Ting Li[2] Aymeric Dieuleveut[1]

[1]CMAP, École Polytechnique [2]The Chinese University of Hong Kong

Context and Setting

- Compression schemes have been extensively used in federated learning to reduce the communication cost.
- Investigate compression and aggregation schemes that produce a specific error distribution (Gaussian or Laplace) on the sum of compression errors.

Aggregate AINQ schemes

Using shared randomness, n clients holding data $x_1, \dots, x_n$ and a server producing Y satisfies (AINQ) property if the quantization error follows a target distribution Q for any $\{x_i\}_{i=1}^n : Y - (\frac{1}{n}\sum_{i=1}^n x_i) \sim Q$.

We consider scalar mechanisms and then apply them coordinate-wise. The simplest such mechanism is subtractive dithering, which guarantees a uniformly distributed error.

Subtractive Dithering

For a given step size $w > 0$ and input X, subtractive dithering works by **sampling** $S \sim \mathcal{U}(-1/2, 1/2)$, **encoding** the message $M = \lceil X/w + S \rfloor$, **decoding** $Y = (M - S)\,w$. Then $(Y - X) \sim \mathcal{U}(-w/2; w/2)$, for any X.

Federated Learning Applications

1. FL and Differential Privacy

Gausian mechanism $\mathrm{G}(\mathcal{D}) = f(\mathcal{D}) + \mathcal{N}(0, \sigma^2 \mathrm{I})$ guarantees (ε, δ)-DP. Use AINQ mechanisms to directly obtain privacy guarantees with a reduced communication cost, e.g. setting the compression error to be a properly scaled Gaussian.

2. FL and Langevin Dynamics

For derivative function H of a potential, the stochastic Langevin dynamics is $\theta_{k+1} = \theta_k - \gamma H(\theta_k) + \sqrt{2\gamma} Z_{k+1}$ with $Z_k \sim \mathcal{N}_d(0, \mathrm{I}_d)$ and $\gamma > 0$. Reduce communication cost with $\mathscr{C}_\gamma$ such that $\mathscr{C}_\gamma(X) - X \sim \mathcal{N}_d(0, 2\mathrm{I}_d/\gamma)$ along with $\theta_{k+1} = \theta_k - \gamma \mathscr{C}_\gamma(H(\theta_k))$.

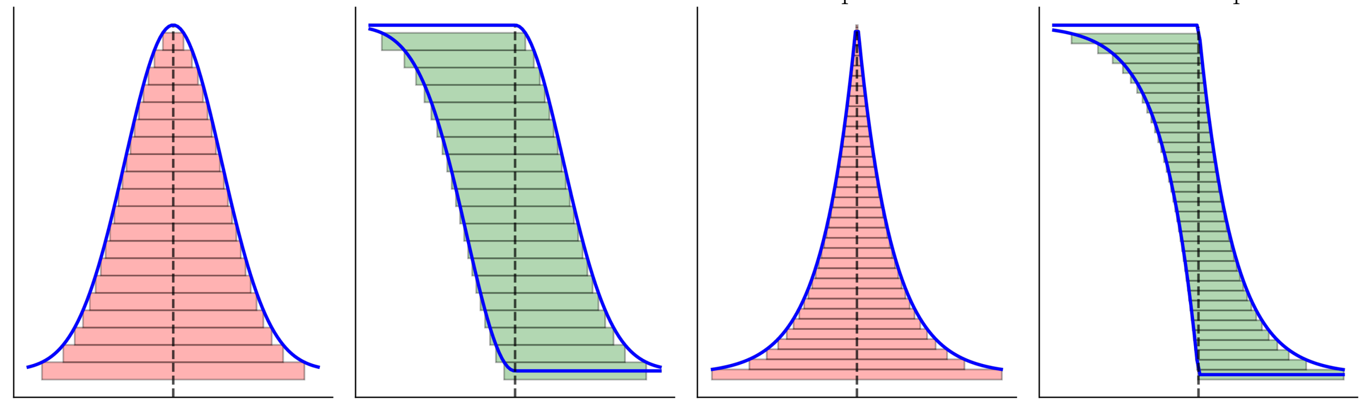
3. FL and Randomized Smoothing

$\min_{\theta \in \mathbb{R}^d} \{f(\theta) = \sum_{i=1}^n f_i(\theta)\}$ rely on *smoothed* $f_\sigma(\theta) = \mathbb{E}_\xi[f(\theta + \sigma\xi)]$ where $\xi \sim \mathcal{N}(0, \mathrm{I}_d)$ and $\sigma > 0$. Compress the model parameter θ with a Gaussian error $\mathscr{C}(\theta) = \theta + \sigma\xi$ and then evaluate the subgradients at compressed point as $g_i(\mathscr{C}(\theta))$ to recover the classical DRS algorithm.

Individual Mechanisms

Uniform error is easy and can be leveraged!

Multiple ways to use uniform distributions to generate (tile) other noises [1, 2]. The idea is to sample the quantization step size and a bias from a specific distribution.

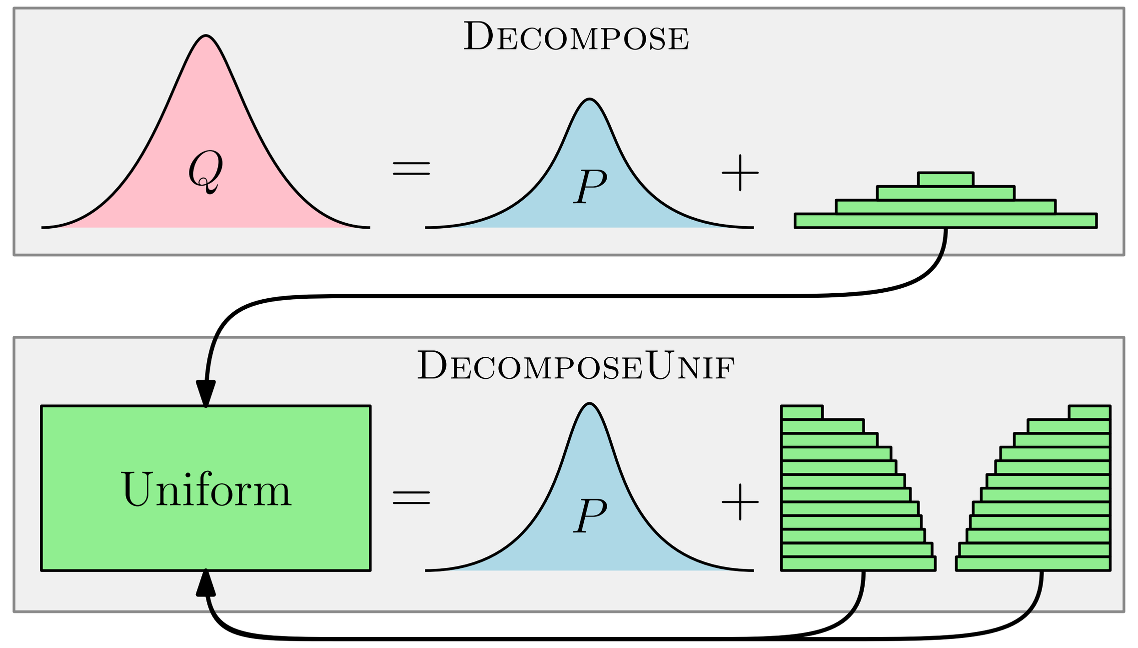


With multiple clients need to decompress before aggregation!

Aggregate Mechanisms

Irwin-Hall Mechanism

Let $S = (S_1, \dots, S_n) \overset{iid}{\sim} \mathcal{U}(-1/2, 1/2)$ and $T = 0$ to be degenerate. The encoding function is $M_i = \mathscr{E}(x_i, S_i) = \lceil x_i/w + S_i \rfloor$ where $w := 2\sigma\sqrt{3n}$, and the decoding function is $Y = w(\sum_i M_i - \sum_i S_i)$. The noise is a scaled Irwin-Hall distribution $\mathrm{IH}(n, 0, \sigma^2)$, where $\mathrm{IH}(n, \mu, \sigma^2)$ denotes the distribution of $n^{-1}\sum_{i=1}^n Z_i + \mu$ with $Z_1, \dots, Z_n \overset{iid}{\sim} \mathcal{U}(-\sigma\sqrt{3n}, \sigma\sqrt{3n})$.



Aggregate Q Mechanism

Let $S_1, \dots, S_n \overset{iid}{\sim} \mathcal{U}(-1/2, 1/2)$ and $T = (A, B) \sim \pi_{A,B} \in \Pi_{A,B}(P, Q)$. The *aggregate Q mechanism* is defined by $w := 2\sigma\sqrt{3n}$ and

$$\mathscr{E}(x, s, a, b) := \lceil x/(aw) + s \rfloor,$$

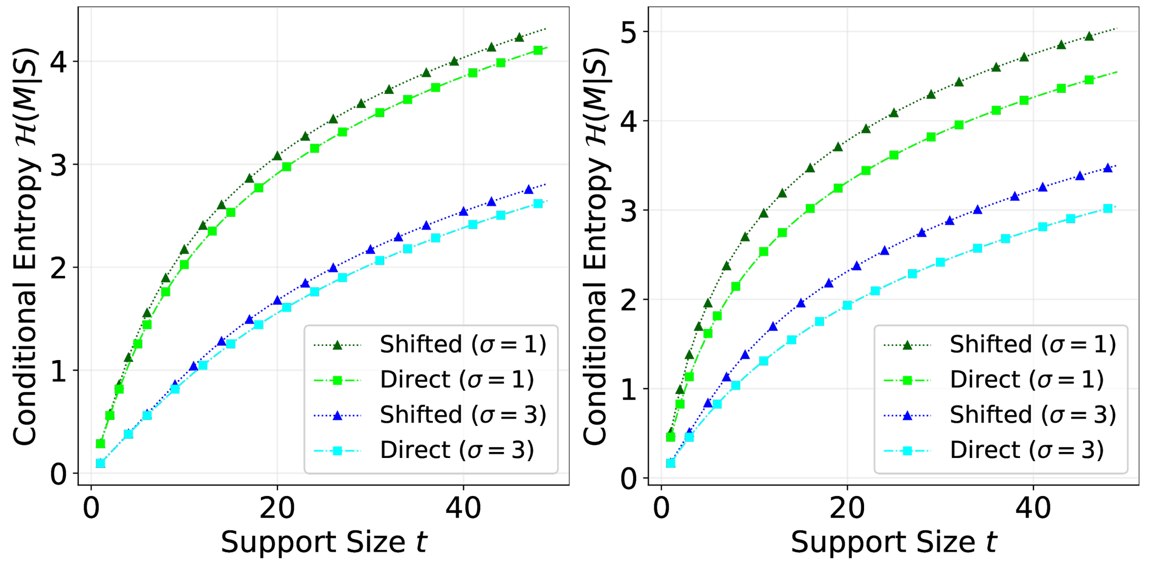
$$\overline{\mathscr{D}}((m_i)_i, (s_i)_i, a, b) := \frac{aw}{n}\left(\sum_{i=1}^n m_i - \sum_{i=1}^n s_i\right) + b.$$

Communication Complexity

Individual Mechanisms (Informal)

- (Optimality Gap) For an input $X \sim \mathcal{U}(0, t)$, the direct layered quantizer is optimal up to $1/t$ factor [1]. For a target unimodal symmetric noise distribution f_Z, the shifted layered quantizer uses at most 2 bits than the direct layered quantizer.
- (Minimal step size) Denote by η_Z the minimal step size of the shifted layer quantizer and assume X lies in a fixed interval of length t [2].

$$Z \sim \mathrm{Laplace}(0, \sigma/\sqrt{2}) \to \eta_Z = \sigma\sqrt{2}\ln 2$$
$$Z \sim \mathcal{N}(0, \sigma^2) \to \eta_Z = 2\sigma\sqrt{\ln 4}$$



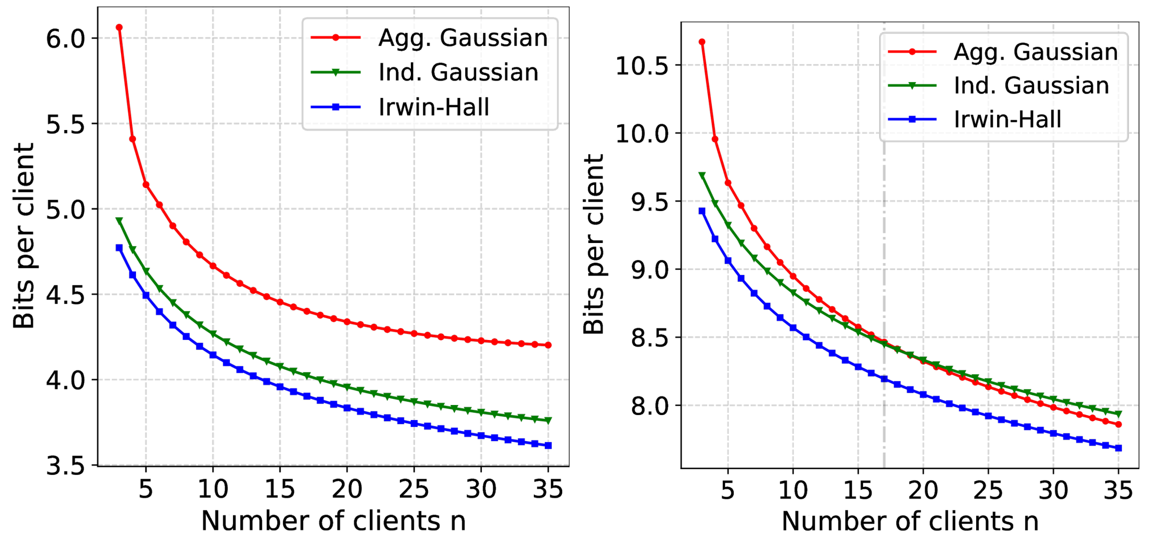
Aggregate Mechanism

- (complexity): Let $P = \mathrm{IH}(n, 0, \sigma^2)$ and assume $|x_i| \le t/2$. There exists an aggregate AINQ mechanism for simulating Q, with an expected amount of communication per client upper-bounded by

$$-h_{\mathrm{M}}(Q\|P) + \log\frac{t}{2\sigma\sqrt{3n}} + \frac{6\sigma\sqrt{3n}\log e}{t}\cdot\frac{\mathbb{E}_{Z\sim Q}[|Z|]}{\mathbb{E}_{Z\sim P}[|Z|]} + 1.$$

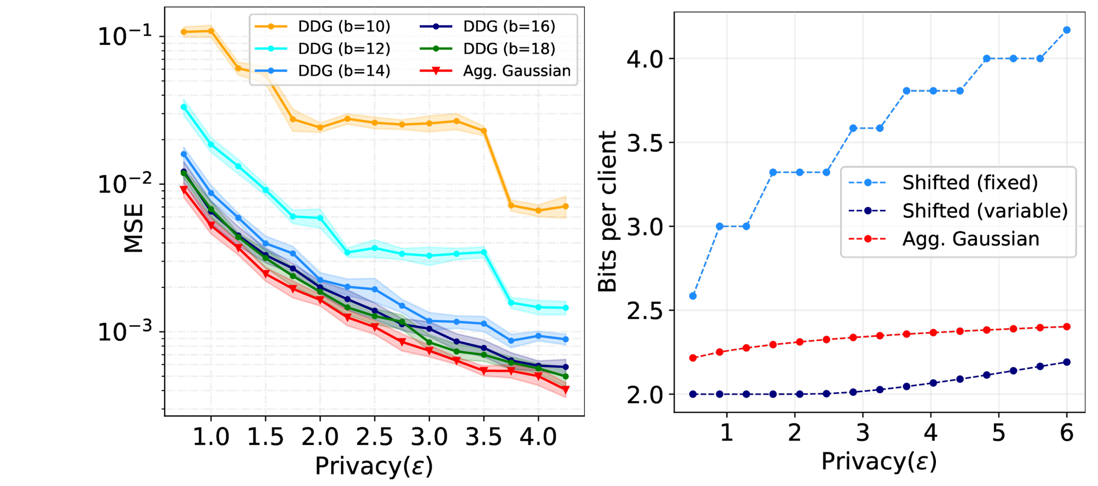
- (lower bound) For P, Q with pdfs f, g (unimodal,symmetric) with $L := 2\sup\{x : f(x) > 0\} < \infty$ and $\lambda := \inf_{x>0} \mathrm{d}g(x)/\mathrm{d}f(x)$, we have

$$h_{\mathrm{M}}(Q\|P) \ge -(1-\lambda)\left(Lf(0) + \log\frac{eL(g(0) - \lambda f(0))}{2(1-\lambda)}\right).$$



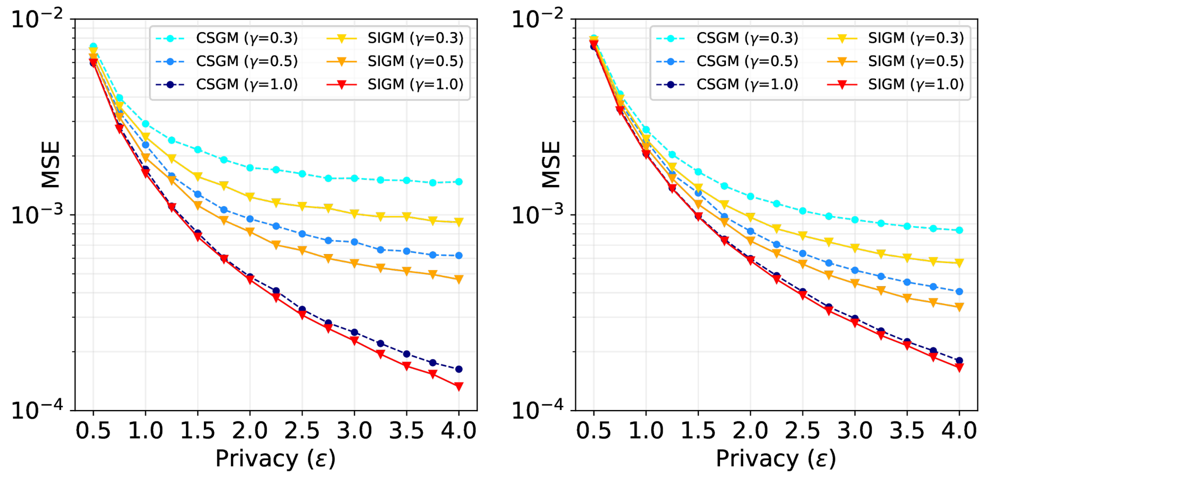
Experiments

Comparison against DDG mechanism:



MSE (left) and bits per client (right) against ε. The DDG mechanism can require up to $b = 18$ bits to match the privacy-utility tradeoff of aggregate Gaussian, where the latter only requires ≤ 2.5 bits on average. We also plot the bits per client for the shifted layered quantizer (using a fixed or variable-length code) on the right figure for comparison (shifted layered quantizer is incompatible with SecAgg).

Improving Privacy Amplification by Subsampling:



Comparison of the Subsampled Individual Gaussian Mechanism (SIGM) and the CSGM scheme of [3]. CSGM leverages privacy amplification through sub-sampling to reduce the amount of noise added by the Gaussian mechanism. We show that even in this setting with lower noise magnitude, it is possible to improve the accuracy-communication tradeoff with our methods.

References

[1] Mahmoud Hegazy and Cheuk Ting Li. Randomized quantization with exact error distribution. In *2022 IEEE Information Theory Workshop (ITW)*, pages 350–355. IEEE, 2022.

[2] David Bruce Wilson. Layered multishift coupling for use in perfect sampling algorithms (with a primer on cftp). *Monte Carlo Methods*, 26:141–176, 2000.

[3] Wei-Ning Chen, Dan Song, Ayfer Ozgur, and Peter Kairouz. Privacy amplification via compression: Achieving the optimal privacy-accuracy-communication trade-off in distributed mean estimation. *Advances in Neural Information Processing Systems*, 36, 2024.