QUIC-FL: Quick Unblased Compression for Federated Learning

Ran Ben Basat, Shay Vargaftik, Amit Portnoy, Gil Einziger, Yaniv Ben-Itzhak, Michael Mitzenmacher

Gautam Singh

Indian Institute of Technology Hyderabad

November 22, 2023

The DME Problem

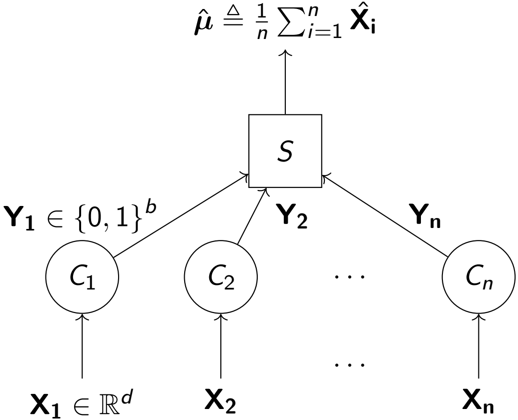

Figure 1: Illustration of the DME Problem. Here, $\hat{\mathbf{X}}_{\mathbf{i}}$ denotes the server estimate for $\mathbf{X_i}$.

vNMSE and NMSE

Definition (vNMSE)

The *vector Normalized Mean Square Error* of $\mathbf{x}$ is defined as

$$\mathrm{vNMSE}\left(\mathbf{x}\right) \triangleq \frac{\mathbb{E}\left[\left\|\hat{\mathbf{x}} - \mathbf{x}\right\|_2^2\right]}{\left\|\mathbf{x}\right\|_2^2}. \tag{1}$$

Definition (NMSE)

The *Normalized Mean Square Error* in the case of the DME problem is defined as

$$\mathrm{NMSE} \triangleq \frac{\mathbb{E}\left[\left\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right\|_2^2\right]}{\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x_i}\right\|_2^2} = \frac{\mathbb{E}\left[\left\|\hat{\boldsymbol{\mu}} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x_i}\right\|_2^2\right]}{\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x_i}\right\|_2^2}. \tag{2}$$

Randomized Hadamard Transform

Definition (Walsh-Hadamard Matrix)

The Walsh-Hadamard matrix $\mathbf{H}_{2^k}$ with $\mathbf{H}_1 = (1)$ is recursively defined as

$$\mathbf{H}_{2^k} = \begin{pmatrix} \mathbf{H}_{2^{k-1}} & \mathbf{H}_{2^{k-1}} \\ \mathbf{H}_{2^{k-1}} & \mathbf{H}_{2^{k-1}} \end{pmatrix}. \tag{3}$$

Definition (Randomized Hadamard Transform)

Let $\mathbf{H} \in \{-1, +1\}^{d \times d}$ be a Walsh-Hadamard Matrix and $\mathbf{D}$ be a diagonal matrix with uniform iid Rademacher entries *i.e.*, entries that are ± 1 with equal probability. Then, the *randomized Hadamard transform* (RHT) of $\mathbf{x} \in \mathbb{R}^d$ is given by

$$\mathcal{R}_{\mathbf{H}}(\mathbf{x}) \triangleq \mathcal{R}_{\mathbf{H}}(\mathbf{x}) = \left(\frac{1}{\sqrt{d}}\mathbf{HD}\right)\mathbf{x}. \tag{4}$$

Properties of RHT

Definition

Inverse RHT The *inverse randomized Hadamard transform* of **x** is given by

$$\mathcal{R}_{\mathbf{H}}^{-1}\left(x\right) \triangleq \mathcal{R}_{\mathbf{H}}^{-1}\mathbf{x} = \mathcal{R}_{\mathbf{H}}^{\top}\mathbf{x} = \left(\frac{1}{\sqrt{d}}\mathbf{DH}\right)\mathbf{x}. \tag{5}$$

Lemma

Define $F_{i,d}\left(x\right)$ to be the CDF of $\frac{1}{\sigma}\mathcal{R}_{\mathbf{H}}\left(\mathbf{x}\right)_i$ and $\Phi\left(x\right)$ the CDF of the standard normal distribution. Then, as $d \to \infty$, $\forall\ 1 \le j \le d$,

$$F_{i,d}\left(x_j\right) \xrightarrow{\mathrm{d}} \Phi\left(x_j\right). \tag{6}$$

Design Goals of QUIC-FL

1. Less computational complexity compared to other methods, at *both* client and server.
2. Lesser amount of shared global as well as private randomness.
3. Same (asymptotic) NMSE as other methods: $\mathcal{O}\left(\frac{1}{n}\right)$.

BSQ Algorithm

1. Uses a parameter $p \in (0,1]$, based on which a threshold $[-t_p, t_p]$ is chosen such that at most dp values lie outside this interval.
2. $\mathbf{x}$ separated into two parts: *large* values that lie outside this interval and *small* values that lie inside this interval.
3. To encode $\mathbf{x}$,
   1. *Large* values are sent exactly, along with their indices. Costs $\log \binom{d}{dp} \approx dp \log \left(\frac{1}{p}\right)$ bits. Can use delta encoding or no encoding at cost of $dp \lceil \log d \rceil$ bits (Authors assume $p \log d << 1$).
   2. *Small* values are stochastically quantized and sent using b bits per value.
4. To decode and compute $\hat{\mathbf{x}}$, decode the *large* values along with their indices. Then, estimate the quantized values.

Main Result of BSQ

Theorem

BSQ, without further assumptions, admits a worst-case NMSE of $\frac{1}{np\left(2^b-1\right)^2}$ *with* $\mathcal{O}\left(d\right)$ *and* $\mathcal{O}\left(nd\right)$ *complexity for encoding and decoding respectively.*

Arbitrarily Large NMSE

Notice that the constant in the NMSE can be made arbitrarily large. For example, $p = 2^{-10}, \ b = 1$.