3PC: Three Point Compressors for Communication-Efficient Distributed Training and a Better Theory for Lazy Aggregation

Peter Richtárik¹ Igor Sokolov² Illya Fattahkulin² Elmur Gasanov¹ Zhize Li¹ Eduard Gorbunov¹

1KAUST 2ETH Zurich 3MIPT

The problem
Nonconvex distributed optimization problem:
\[ \min_{x} \{ f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \} , \]

- \( n \): number of clients
- \( f(x) \): smooth local loss function, i.e., \( \| \nabla f_i(x) - \nabla f_j(x) \| \leq L \| x - y \| \) for all \( x, y \in \mathbb{R}^d \)

Goal: find \( x \) such that \( \mathbb{E} \| \nabla f(x) \| \leq \epsilon \)

Compressed learning
Contractive compressor: a (possibly randomized) map \( C: \mathbb{R}^d \rightarrow \mathbb{R}^d \) is called a contractive compressor if there exists a constant \( 0 < \alpha \leq 1 \)
\[ \mathbb{E} \| C(x) - C(y) \| \leq (1 - \alpha) \| x - y \| , \forall x, y \in \mathbb{R}^d \]

Top-\( \ell \) (greedy) quantization operator is defined via
\[ C(x) = \frac{1}{\sum_{i=1}^{\ell} \mathbb{1}_{[p_i, p_{i+1})}(x)} \sum_{i=1}^{\ell} \mathbb{1}_{[p_i, p_{i+1})}(x) x \]

Error feedback with contractive compressor
- \( \mathbb{Q} \): Motivation for error feedback: the type of method
- \( \alpha > 1 \)
- \( \alpha \): optimal complexity \( \mathcal{O}(\sqrt{\epsilon}) \)
- \( \alpha \): practical method

Laziness aggregation
- \( \mathbb{Q} \): Motivation for laziness [3]: reduce communication by sending gradients only when they change significantly
- \( \gamma \): optimal complexity \( \mathcal{O}(1/\gamma) \)
- \( \gamma \): practical method

2. Distributed compressed GD with 3PC

1. Three point compressor (3PC)

3PC: we say that \( f \in \mathbb{R}^d \) if \( f(x) \)
\[ \mathbb{E} \| C_i(x) - C_j(x) \| \leq (1 - \alpha) \| x - y \| , \forall x, y \in \mathbb{R}^d \]

The vectors \( x \in \mathbb{R}^d \) and \( y \in \mathbb{R}^d \) are parameters defining the compressor.

3. Special cases

GD: if we do not employ any compression, i.e., we set \( \alpha = 1 \), then Algorithm 1 reduces to vanilla GD and (1) holds with \( B = 1 \) and \( A = 0 \).

EF21 [2]: let \( C \) be a contractive compressor

then Algorithm 2 reduces to EF21 and (1) holds with \( A = 1 - (1 - \alpha/(1 + \alpha)) \) and \( B = (1 - (1 - \alpha)/(1 + \alpha)) \), where \( \alpha \) satisfies \( \alpha < 1 \) and \( (1 - \alpha)/(1 + \alpha) < 1 \).

LAG [3]: and CLAG, let \( C \) be a contractive compressor. Choose a trigger \( \zeta > 0 \), and define

\[ C_i(x) = \left\{ \begin{array}{ll} A \mid C - x \mid & \text{if } \| C - x \| > \zeta \| x \| \end{array} \right. \]

then Algorithm 1 reduces to CLAG and (1) holds with \( A = 1 - (1 - \alpha/(1 + \alpha)) \) and \( B = (1 - (1 - \alpha)/(1 + \alpha)) \), where \( \alpha \) satisfies \( \alpha < 1 \) and \( (1 - \alpha)/(1 + \alpha) < 1 \).

If \( C \notin \mathbb{R}^d \) (i.e., we recover LAG).

In Table 1, we summarize several further 3PC compressors and the new algorithms they lead to (e.g., 3PCv1-3PCv5).

4. Main result

Assumption 1. \( f(x) \) is \( \mathbb{R}^d \)-valued. Moreover, \( f(x) \in \mathbb{R}^d \) for all \( x \in \mathbb{R}^d \).

Assumption 2. \( f(x) \) is \( \mathbb{R}^d \)-valued. Moreover, \( f(x) \in \mathbb{R}^d \) for all \( x \in \mathbb{R}^d \).

Let \( \mathbb{I} \) be a constant \( \mathbb{I} \subseteq \mathbb{R} \), such that \( \sum_{i=1}^{n} f_i(x) =: \mathbb{I} \). Then, for any \( T \geq 0 \) we have

\[ \mathbb{E} \| \nabla \mathbb{I}(x) \| \leq \mathbb{I} \| x \| \]

where \( \mathbb{I} \) is sampled uniformly at random from the points \( \{ x, x', \ldots, x^{k-1} \} \) produced by 3PC, \( \mathbb{I} = f(x) = f(x') = f(x^{k-1}) \), and \( C \) is a contractive compressor.

Let the assumptions of Theorem 1 hold and choose the stepsize \( \gamma \) such that

\[ \frac{1}{\gamma} \mathbb{E} \| \nabla \mathbb{I}(x) \| \leq \mathbb{I} \| x \| \]

then, to achieve \( \mathbb{E} \| \nabla \mathbb{I}(x) \| \leq \mathbb{I} \| x \| \) for some \( x > 0 \), the 3PC method requires

\[ T = \mathbb{E} \{ \mathbb{I}(x) \} / \gamma \] (denominations = communication rounds)

5. Comparison of methods with lazy aggregation

Table 2: Comparison of existing and proposed theoretically-supported methods exploiting lazy aggregation. In the notation for our methods, \( M_1 = L + \mathbb{L}(x) \) and \( M_0 = 0 + \mathbb{L}(x) \). For all methods, \( M_1 = 0 + \mathbb{L}(x) \) and \( M_0 = 0 + \mathbb{L}(x) \). For all methods, \( M_1 = 0 + \mathbb{L}(x) \) and \( M_0 = 0 + \mathbb{L}(x) \).

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>3PCv1</th>
<th>3PCv2</th>
<th>3PCv3</th>
<th>3PCv4</th>
<th>3PCv5</th>
</tr>
</thead>
<tbody>
<tr>
<td>3PCv1</td>
<td>CLAG</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3PCv2</td>
<td>CLAG</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3PCv3</td>
<td>CLAG</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3PCv4</td>
<td>CLAG</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3PCv5</td>
<td>CLAG</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

\[ \min_{x} \{ f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \} \]

where \( n \), the flattened representations of images with \( d = 784 \), \( D \) and \( E \) are homogenous parameters.

Figures

Figure 1: Number of clients \( n = 100 \), compression level \( K = 25 \).

Figure 2: Number of clients \( n = 20 \). The red-centered oval indicates the experiment with the smallest communication cost.

Figure 3: Comparison of existing and proposed theoretically-supported methods exploiting lazy aggregation. In the notation for our methods, \( M_1 = L + \mathbb{L}(x) \) and \( M_0 = \mathbb{L}(x) \).

References