Error Compensated Proximal SGD and RDA

Xun Qian1  Hanze Dong 2  Peter Richtárik1  Tong Zhang2
1KAUST  2Hong Kong University of Science and Technology

Algorithm (ECRDA)

\[ \min_{x \in E^d} P(x) := \frac{1}{\epsilon} f^{(i)}(x) + \psi(x), \]
where \( f^{(i)}(x) \) is an average of \( m \) smooth convex functions \( f_i^{(i)} \) distributed over \( n \) nodes, and \( \psi \) is a proper closed convex function. On each node, \( f^{(i)}(x) \) is an average of \( m \) smooth convex functions
\[ f^{(i)}(x) = \frac{1}{\epsilon} \sum_{i=1}^m f_i^{(i)}(x). \]

Algorithm (ECSGD)

\[ \text{proc}_{x_i}(x) := \arg \min \left\{ \| x - y \|_2^2 + \gamma \psi(y) \right\} \]

Example 1: Error compensated proximal SGD (ECSGD)

\[ x^0 = w^0 \in E^d; \quad c^0 = 0 \in E^d; \quad u^0 = 1 \in R; \quad \gamma \text{ params: stepsize} \]

for \( k = 1, 2, \ldots \) do

for \( \tau = 1, \ldots, n \) do

Sample \( i \) uniformly and independently in \([m] \) on each node

\[ g_i^\tau = \nabla f^{(i)}(x^\tau) - \nabla f^{(i)}(u^\tau); \quad y_i^\tau = Q(y_i^\tau + c_i^\tau); \quad c_i^\tau+1 = c_i^\tau + 2g_i^\tau - g_i^\tau+1 = 0 \text{ for } \tau = 2, \ldots, n; \quad u_i^\tau+1 = 1 \text{ with probability } p; \quad u_i^\tau = 0 \text{ with probability } 1 - p; \]

Send \( y_i^\tau \) and \( u_i^\tau+1 \) to the other nodes. Send \( \nabla f^{(i)}(u^\tau) \) to the other nodes if \( u^\tau = 1 \). Receive \( y_i^\tau \) and \( u_i^\tau+1 \) from the other nodes. Receive \( \nabla f^{(i)}(u^\tau) \) from the other nodes if \( u^\tau = 1 \).

end

end

Assumptions

Assumption 1: \( E(Q(x)) = 0 \).
Assumption 2: \( f^{(i)}(\frac{n}{2} x + \frac{1}{2} n c) \in E(x = \frac{n}{2} x + \frac{1}{2} n c, x \in E^d; \quad x \in E^d; \quad 1 \leq \tau \leq n \).
Assumption 3: \( f^{(i)} \) is \( L \)-smooth for \( 1 \leq i \leq m \) and \( 1 \leq \tau \leq n \).

ECRDA

Assumption 4: \( f^{(i)} \) is \( L \)-smooth. \( h \) is 1-strongly convex and \( h(x^\tau) = \psi(x^\tau) \).

Convergence Result \( (E(P(x^\tau) - P(x))]) \)

\[ \text{Assume the compressor } Q \text{ in Algorithm 1 is a contraction compressor and Assumption 3 holds. Let } \]
\[ x^\tau := \frac{1}{\epsilon} \sum_{\tau=1}^\tau x^\tau; \]
\[ p = 0; \quad \text{if there exists constant stepsize } \gamma \leq \frac{B^2}{\epsilon} \text{ s.t.,} \]
\[ O \left( \frac{1}{\epsilon} \sum_{\tau=1}^\tau (\nabla^2 f(x^\tau) + P(x^\tau - x^\tau)) \right) \]
\[ p > 0; \quad \text{if there exists constant stepsize } \gamma \leq \frac{B^2}{\epsilon} \text{ s.t.,} \]
\[ O \left( \frac{1}{\epsilon} \sum_{\tau=1}^\tau (\nabla^2 f(x^\tau) + P(x^\tau - x^\tau)) \right) \]

Under Assumption 1 or Assumption 2

\[ p = 0; \quad \text{if there exists constant stepsize } \gamma \leq \frac{B^2}{\epsilon} \text{ s.t.,} \]
\[ O \left( \frac{1}{\epsilon} \sum_{\tau=1}^\tau (\nabla^2 f(x^\tau) + P(x^\tau - x^\tau)) \right) \]

Communication Cost

\[ \text{Denote } \Delta_e \text{ as the communication cost of the uncompressed vector } x \in E^d \text{. Let} \]
\[ r(Q) := \sup_{x \in E^d} \left\{ E \left( \text{communication cost of } Q(x) \right) \right\}. \]

For efficiently small \( r \),

\[ \text{if } E(P(x^\tau) - P(x^\tau)) \leq \epsilon \text{ for ECSGD:} \]
\[ O \left( \Delta_e r(Q) + 1 \right); \]
\[ \text{if } E(P(x^\tau) - P(x^\tau)) \leq \epsilon \text{ for ECRDA:} \]
\[ O \left( \Delta_e r(Q) + 1 \right). \]

Numerical Results

1. Error Compensated and Full SGD/RDA

2. Comparison to Quantization and Randk-DIANA

References


The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication.


Distributed learning with compressed gradient differences.