The Problem

\[ \min_{x \in \mathbb{R}^d} P(x) := \frac{1}{2} \sum_{i=1}^{n} f_i(x) + \psi(x), \]

where \( f_i(x) := \frac{1}{2} \sum_j f_i^{(j)}(x) \) is an average of \( n \) smooth convex functions \( f_i^{(j)} \) distributed over \( n \) nodes, and \( \psi \) is a proper closed convex function. On each node, \( f_i^{(j)}(x) \) is an average of \( m \) smooth convex functions

\[ f_i^{(j)}(x) = \frac{1}{2} \sum_{k=1}^{m} f_i^{(j,k)}(x), \]

Algorithm

\begin{algorithm}
\textbf{Algorithm 1: Error compensated Loopless SVRG (EC-LSVRG)}

\texttt{x}^0 = w^0 \in \mathbb{R}^d, e_i^0 = 0 \in \mathbb{R}^d, \epsilon_i^0 = 1 \in \mathbb{R} ; \text{params:}
\texttt{stepsize} > 0, \texttt{probability} \texttt{p} \in (0, 1].
\texttt{for} \texttt{k} = 1, 2, \ldots , \texttt{do}
\texttt{for} \texttt{r} = 1, 2, \ldots , \texttt{do}
\texttt{Sample} \texttt{t} \texttt{uniformly and independently in} \texttt{[m]} \texttt{on each node}
\texttt{\( y_i^r = \nabla f_i^{(j)}(x_i^r) - \nabla f_i^{(j)}(w_{i}^r), y_i^r = Q(\epsilon_i^r + e_i^r), \)}
\texttt{\( e_i^{r+1} = \epsilon_i^r + y_i^r - y_i^r, \)}
\texttt{\( u_i^{r+1} = 0 \text{ with probability} \texttt{p} \text{ and probability} \texttt{1} - \texttt{p} \text{ for} \texttt{r} = 2, \ldots , \texttt{n} , \)}
\texttt{\( w_i^{r+1} = 1 \text{ with probability} \texttt{p} \) and \texttt{0} otherwise}
\texttt{Send \( y_i^r \) and \( u_i^{r+1} \) to the other nodes. Send \( \nabla f_i^{(j)}(u_i^r) \) to the other nodes if \( u_i^r = 1 \)}
\texttt{Receive \( y_i^r \) and \( u_i^{r+1} \) from the other nodes. Receive \( \nabla f_i^{(j)}(u_i^r) \) from the other nodes if \( u_i^r = 1 \)}
\texttt{end}
\texttt{end}
\texttt{end}
\texttt{end}
\end{algorithm}

Gradient Compression Methods

- \( Q : \mathbb{R}^d \rightarrow \mathbb{R}^d \) is a contraction compressor if there is a \( 0 < \delta < 1 \) such that for all \( x \in \mathbb{R}^d \),

\[ \left\| x - Q(x) \right\|^2 \leq (1 - \delta) \| x \|^2. \]

- \( Q \) is an unbiased compressor if there is a \( \omega \geq 0 \) such that

\[ \mathbb{E}\left[ Q(x) \right] = x \text{ and } \mathbb{E}\left[ \| Q(x) \|^2 \right] \leq (\omega + 1) \| x \|^2 \]

for all \( x \in \mathbb{R}^d \).

\[ \frac{1}{\sqrt{\omega + 1}} Q \text{ is a contraction compressor with } \delta = \frac{1}{\sqrt{\omega + 1}}. \]

Assumptions

- **Assumption 1:** \( E(Q(x)) = x \). \( E[Q(x)] = x \)
- **Assumption 2:** \( \text{For } x_i = \eta_i^r + e_i^r \in \mathbb{R}^d, r = 1, \ldots , n \)

and \( k \geq 0 \) in Algorithm 1, we have \( E[Q(x_i)] = Q(x_i) \), and

\[ \| \sum_{i=1}^{n} Q(x_i) - x_i \| \leq (1 - \delta) \left( \sum_{i=1}^{n} \| x_i \| \right). \]

**Assumption 3:** \( f_i^{(j)} \) is \( L \)-smooth, \( f_i^{(j,k)} \) is \( L_f \)-smooth, \( f \) is \( p \)-strongly convex.

**Assumption 4:** \( f_i^{(j)} \) is \( L \)-smooth, \( f_i^{(j,k)} \) is \( L_f \)-smooth, \( f \) is \( p \)-strongly convex.

Smooth Case (\( \psi \equiv 0 \))

Convergence Result

Assume the compressor \( Q \) in Algorithm 1 is a contraction compressor and Assumption 4 holds. Let

\[ \| w_k \| = (1 - \min \left( \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right)) \| x_k \|, W_k = \sum_{i=0}^{k} w_i \text{ and } x_k = \frac{1}{\sum_{i=0}^{k} w_i} \sum_{i=0}^{k} w_i x_i. \]

If \( \eta \leq \frac{1}{\min \left( \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right)} \), then we have \( \| P(x_k) - P(x_{k+1}) \| \leq \frac{\epsilon_i^{r+1}}{\left( 1 - \min \left( \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right) \right)} \)

In particular, if we choose

\[ \eta = \min \left( \frac{1}{\min \left( \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right)}, \left( \frac{\left( \frac{\sqrt{p}}{\sqrt{2}L} + \frac{\sqrt{p}}{2} \right)}{1 + \frac{\left( \frac{\sqrt{p}}{\sqrt{2}L} + \frac{\sqrt{p}}{2} \right)}{\frac{\sqrt{p}}{2} + \frac{\sqrt{p}}{\sqrt{2}L}} \right) \right) \]

**Convergence Result**

Assume the compressor \( Q \) also satisfies Assumption 1 or Assumption 2. If

\[ \eta \leq \min \left( \frac{1}{\sqrt{\omega + 1} - \sqrt{1 - \delta}}, \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right) \]

then we have \( \| P(x_k) - P(x_{k+1}) \| \leq \frac{\epsilon_i^{r+1}}{\left( 1 - \min \left( \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right) \right)} \)

In particular, if we choose

\[ \eta = \min \left( \frac{1}{\sqrt{\omega + 1} - \sqrt{1 - \delta}}, \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right) \]

**Convergence Result**

Assume the compressor \( Q \) is a contraction compressor with \( \delta = \frac{1}{\sqrt{\omega + 1}} \),

\[ \| P(x_k) - P(x_{k+1}) \| \leq \frac{\epsilon_i^{r+1}}{\left( 1 - \min \left( \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right) \right)} \]

In particular, if we choose

\[ \eta = \min \left( \frac{1}{\sqrt{\omega + 1} - \sqrt{1 - \delta}}, \frac{\sqrt{p}}{\sqrt{2}L}, \frac{\sqrt{p}}{2} \right) \]

**Convergence Result**

Denote \( \Delta_i \) as the communication cost of the uncompressed vector \( x \in \mathbb{R}^d \). Let

\[ r(Q) := \sup_{x \in \mathbb{R}^d} \mathbb{E} \left[ \left( \text{communication cost of } Q(x) \right) \right]. \]

Assume \( L = \bar{L} = L \) and \( \Delta_i r(Q) \geq O(1) \). Choose \( p = O(r(Q)) \).

**Composite Case:**

\[ O \left( \frac{1}{\left( \frac{\sqrt{p}}{\sqrt{2}L} + \frac{\sqrt{p}}{2} \right)} + 1 + \frac{\left( \frac{\sqrt{p}}{\sqrt{2}L} + \frac{\sqrt{p}}{2} \right)}{\frac{\sqrt{p}}{2} + \frac{\sqrt{p}}{\sqrt{2}L}} \ln \frac{1}{\eta} \right). \]

**Smooth Case:**

\[ O \left( \frac{1}{\left( \frac{\sqrt{p}}{\sqrt{2}L} + \frac{\sqrt{p}}{2} \right)} + 1 + \frac{\left( \frac{\sqrt{p}}{\sqrt{2}L} + \frac{\sqrt{p}}{2} \right)}{\frac{\sqrt{p}}{2} + \frac{\sqrt{p}}{\sqrt{2}L}} \ln \frac{1}{\eta} \right). \]

References


Stochastic distributed learning with gradient quantization and variance reduction.


Log- and I-kahus with arbitrary sampling.