Error Compensated Proximal SGD and RDA

Xun Qian[1] Hanze Dong [2] Peter Richtárik[1] Tong Zhang[2]

[1]KAUST [2]Hong Kong University of Science and Technology

The Problem

$$\min_{x\in\mathbb{R}^d} P(x) := \frac{1}{n}\sum_{\tau=1}^{n} f^{(\tau)}(x) + \psi(x), \quad (1)$$

where $f(x) := \frac{1}{n}\sum_\tau f^{(\tau)}(x)$ is an average of n smooth convex functions $f^{(\tau)}$ distributed over n nodes, and ψ is a proper closed convex function. On each node, $f^{(\tau)}(x)$ is an average of m smooth convex functions

$$f^{(\tau)}(x) = \frac{1}{m}\sum_{i=1}^{m} f_i^{(\tau)}(x).$$

Algorithm (ECSGD)

- $\mathrm{prox}_{\gamma\psi}(x) := \arg\min \left\{\frac{1}{2}\|x-y\|^2 + \gamma\psi(y)\right\}$

Algorithm 1: Error compensated proximal SGD (ECSGD)

$x^0 = w^0 \in \mathbb{R}^d$; $e_\tau^0 = 0 \in \mathbb{R}^d$; $u^0 = 1 \in \mathbb{R}$; params: stepsize $\gamma > 0$; probability $p \in (0,1]$.

for $k = 1, 2, \ldots$ **do**

  for $\tau = 1, ..., n$ **do**

    Sample i_k^τ uniformly and independently in $[m]$ on each node

    $g_\tau^k = \nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k), \quad y_\tau^k = Q(\gamma g_\tau^k + e_\tau^k),$

    $e_\tau^{k+1} = e_\tau^k + \gamma g_\tau^k - y_\tau^k, \quad u_\tau^{k+1} = 0$ for $\tau = 2, ..., n,$

    $u_1^{k+1} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$

    Send y_τ^k and u_τ^{k+1} to the other nodes. Send $\nabla f^{(\tau)}(w^k)$ to the other nodes if $u^k = 1$

    Receive y_τ^k and u_τ^{k+1} from the other nodes. Receive $\nabla f^{(\tau)}(w^k)$ from the other nodes if $u^k = 1$

  end

  $y^k = \frac{1}{n}\sum_{\tau=1}^n y_\tau^k, \quad u^{k+1} = \sum_{\tau=1}^n u_\tau^{k+1},$

  $x^{k+0.5} = x^k - (y^k + \gamma\nabla f(w^k)),$

  $x^{k+1} = \mathrm{prox}_{\gamma\psi}\left(x^{k+0.5}\right), \quad w^{k+1} = \begin{cases} x^k & \text{if } u^{k+1} = 1 \\ w^k & \text{otherwise} \end{cases}$

end

Gradient Compression Methods

- $Q : \mathbb{R}^d \to \mathbb{R}^d$ is a *contraction compressor* if there is a $0 < \delta \le 1$ such that for all $x \in \mathbb{R}^d$,

$$\mathbb{E}\|x - Q(x)\|^2 \le (1-\delta)\|x\|^2. \quad (2)$$

- $\tilde{Q}$ is an *unbiased compressor* if there is $\omega \ge 0$ such that

$$\mathbb{E}[\tilde{Q}(x)] = x \quad \text{and} \quad \mathbb{E}\|\tilde{Q}(x)\|^2 \le (\omega+1)\|x\|^2 \quad (3)$$

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

- $\frac{1}{\omega+1}\tilde{Q}$ is a contraction compressor with $\delta = \frac{1}{\omega+1}$.

Algorithm (ECRDA)

Algorithm 2: Error compensated RDA (ECRDA)

$x^1 = w^1 = \arg\min_x h(x)$; $\bar{g}^0 = 0 \in \mathbb{R}^d$; $e_\tau^1 = 0 \in \mathbb{R}^d$; $u^1 = 1 \in \mathbb{R}$; params: an auxiliary function $h(x)$ that is strongly onvex on dom ψ and also satisfies

$$\arg\min_x h(x) \in \arg\min_x \psi(x);$$

a nonegative and nondecreasing sequence $\{\beta_k\}_{k\ge1}$.

for $k = 1, 2, \ldots$ **do**

  for $\tau = 1, ..., n$ **do**

    Sample i_k^τ uniformly and independently in $[m]$ on each node

    $g_\tau^k = \nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f^{(\tau)}(w^k), \quad y_\tau^k = Q(g_\tau^k + e_\tau^k),$

    $e_\tau^{k+1} = e_\tau^k + g_\tau^k - y_\tau^k, \quad u_\tau^{k+1} = 0$ for $\tau = 2, ..., n$,

    $u_1^{k+1} = \begin{cases} 1 & \text{with propobility } p \\ 0 & \text{with probability } 1-p \end{cases}$

    Send y_τ^k and u_τ^{k+1} to the other nodes. Send $\nabla f^{(\tau)}(w^k)$ to the other nodes if $u^k = 1$

    Receive y_τ^k and u_τ^{k+1} from the other nodes. Receive $\nabla f^{(\tau)}(w^k)$ from the other nodes if $u^k = 1$

  end

  $y^k = \frac{1}{n}\sum_{\tau=1}^n y_\tau^k, \quad u^{k+1} = \sum_{\tau=1}^n u_\tau^{k+1},$

  $\bar{g}^k = \frac{k-1}{k}\bar{g}^{k-1} + \frac{1}{k}(y^k + \nabla f(w^k))$

  $x^{k+1} = \arg\min_x\{\langle \bar{g}^k, x\rangle + \psi(x) + \frac{\beta_k}{k}h(x)\},$

  $w^{k+1} = \begin{cases} x^k & \text{if } u^{k+1} = 1 \\ w^k & \text{otherwise} \end{cases}$

end

Assumptions

Assumption 1: $\mathbb{E}[Q(x)] = \delta x$.

Assumption 2: For $x_\tau = \frac{\eta}{L}g_\tau^k + e_\tau^k \in \mathbb{R}^d$ ($x_\tau = g_\tau^k + e_\tau^k$), $\tau = 1, ..., n$ and $k \ge 0$ in Algorithm 1 (Algorithm 2), we have $\mathbb{E}[Q(x_\tau)] = Q(x_\tau)$, and

$$\left\|\sum_{\tau=1}^{n}(Q(x_\tau) - x_\tau)\right\|^2 \le (1-\delta)\left\|\sum_{\tau=1}^{n} x_\tau\right\|^2.$$

Assumption 3: $f_i^{(\tau)}$ is L-smooth for $1 \le i \le m$ and $1 \le \tau \le n$.

ECRDA

Assumption 4: $f_i^{(\tau)}$ is L-smooth. h is 1-strongly convex and $h(x^1) = \psi(x^1) = 0$.

Assumption 5: In Algorithm 2, $\|\nabla f_{i_k^\tau}^{(\tau)}(x^k)\|^2 \le G^2$, $\|\nabla f^{(\tau)}(w^k)\|^2 \le G^2$, and $\|\partial h(x^k)\|^2 \le H^2$ for $k \ge 1$. $h(x^*) \le D^2$.

Convergence Result ($\mathbb{E}[P(\bar{x}^k) - P(x^*)]$)

Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3 holds. Let $\bar{x}^k := \frac{1}{k}\sum_{j=1}^k x^j$.

$p = 0$: there exists constant stepsize $\gamma \le \frac{\delta^2}{48L}$ s.t.,

$$O\left(\frac{L\|x^0-x^*\|^2}{\delta^2 k} + \frac{\|x^0-x^*\|\sqrt{\sigma^2/\delta + L(P(w^0)-P(x^*))/\delta^2}}{\sqrt{k}}\right).$$

$p > 0$: there exists constant stepsize $\gamma \le \frac{\delta^2}{80L}$ s.t.,

$$O\left(\frac{1}{k}\left(\frac{L\|x^0-x^*\|^2}{\delta^2} + \frac{P(w^0)-P(x^*)}{p}\right) + \frac{\sigma\|x^0-x^*\|}{\sqrt{\delta k}}\right).$$

Under Assumption 1 or Assumption 2.

$p = 0$: there exists constant stepsize $\gamma \le \frac{\delta^2}{(64+304/n)L}$ s.t.,

$$O\left(\frac{L\|x^0-x^*\|^2}{\delta^2 k} + \frac{\|x^0-x^*\|\sqrt{\sigma^2/(n\delta) + L(P(w^0)-P(x^*))/\delta^2}}{\sqrt{k}}\right).$$

$p > 0$: there exists constant stepsize $\gamma \le \frac{\delta^2}{(128+592/n)L}$ s.t.,

$$O\left(\frac{1}{k}\left(\frac{L\|x^0-x^*\|^2}{\delta^2} + \frac{P(w^0)-P(x^*)}{p}\right) + \frac{\sigma\|x^0-x^*\|}{\sqrt{n\delta k}}\right).$$

Convergence Result ($\mathbb{E}[P(\bar{x}^k) - P(x^*)]$)

Assume the compressor Q in Algorithm 2 is a contraction compressor and Assumptions 4, 5 hold. Let $\bar{x}^k := \frac{1}{k}\sum_{j=1}^k x^j$.

$p = 0$: for fixed $k \ge O(1/\delta)$, by choosing $\beta_j = 4\sqrt{\frac{k}{\delta}}\frac{\sqrt{G^2+L(P(w^1)-P(x^*))+\delta\sigma^2/4}}{D}$ for $j \ge 1$,

$$O\left(\frac{D\sqrt{G^2+L(P(w^1)-P(x^*))+\delta\sigma^2}}{\sqrt{\delta k}} + \left(\frac{DG}{\delta\sqrt{\delta k}} + H^2 + \frac{G^2}{\delta^2}\right)\frac{\ln k}{k}\right).$$

$p > 0$: for fixed $k \ge O(1/\delta^{\frac{5}{2}})$, by choosing $\beta_j = \frac{4\sqrt{k}}{\delta^{1/4}}\frac{\sqrt{\sigma^2+24G^2}}{D}$ for $j \ge 1$,

$$O\left(\frac{D\sqrt{\sigma^2+G^2}}{\delta^{1/4}\sqrt{k}} + \frac{LD(P(w^1)-P(x^*))}{k\sqrt{k}\delta^{5/4}p\sqrt{\sigma^2+G^2}} + \left(\frac{DG}{\sqrt{k}\delta^{7/4}} + \frac{H^2\delta^2+G^2}{\delta^2}\right)\frac{\ln k}{k}\right).$$

Under Assumption 1 or Assumption 2.

$p = 0$: for fixed $k \ge O(1/\delta)$, by choosing $\beta_j = 4\sqrt{\frac{k}{\delta}}\frac{\sqrt{G^2+(2+9/n)L(P(w^1)-P(x^*))+3\delta\sigma^2/n}}{D}$ for $j \ge 1$,

$$O\left(\frac{D\sqrt{G^2+L(P(w^1)-P(x^*))+\delta\sigma^2/n}}{\sqrt{\delta k}} + \left(\frac{DG}{\delta\sqrt{\delta k}} + H^2 + \frac{G^2}{\delta^2}\right)\frac{\ln k}{k}\right).$$

$p > 0$: for fixed $k \ge O(n^{\frac{3}{2}}/\delta^{\frac{5}{2}})$, by choosing $\beta_j = \frac{4\sqrt{k}}{(n\delta)^{1/4}}\frac{\sqrt{6\sigma^2+12G^2}}{D}$ for $j \ge 1$, ($A = \frac{D\sqrt{\sigma^2+G^2}}{(n\delta)^{1/4}\sqrt{k}}$)

$$O\left(A + \frac{n^{3/4}LD\big(P(w^1)-P(x^*)\big)}{k\sqrt{k}\delta^{5/4}p\sqrt{\sigma^2+G^2}} + \left(\frac{n^{1/4}DG}{\sqrt{k}\delta^{7/4}} + \frac{H^2\delta^2+G^2}{\delta^2}\right)\frac{\ln k}{k}\right).$$

Communication Cost

Denote Δ_1 as the communication cost of the uncompressed vector $x \in \mathbb{R}^d$. Let

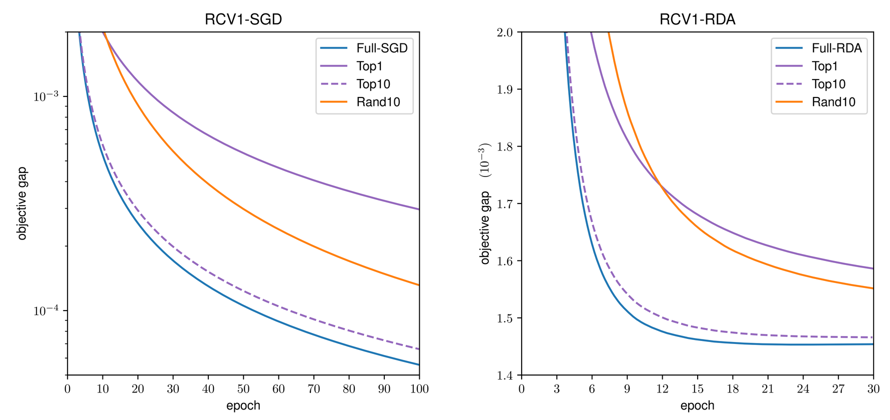
$$r(Q) := \sup_{x\in\mathbb{R}^d}\left\{\mathbb{E}\left[\frac{\text{communication cost of } Q(x)}{\Delta_1}\right]\right\}.$$

For effieiently small ϵ,

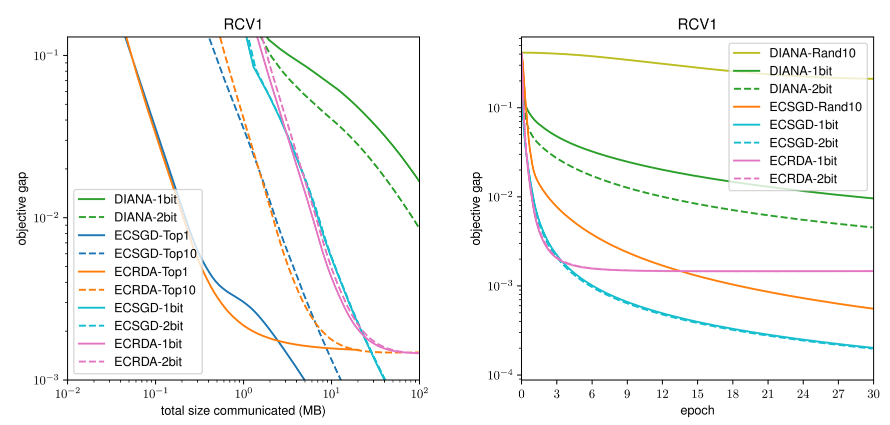
- $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \le \epsilon$ for ECSGD:
$$O((\Delta_1 r(Q) + 1)\tfrac{1}{\delta\epsilon^2});$$
- $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \le \epsilon$ for ECRDA:
$$O((\Delta_1 r(Q) + 1)\tfrac{1}{\sqrt{\delta}\epsilon^2}).$$

Numerical Results

1. Error Compensated and Full SGD/RDA



2. Comparison to Quantization and RandK-DIANA



References

[1] S. U. Stich and S. P. Karimireddy.
The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication.
arXiv: 1909.05350, 2019.

[2] K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik.
Distributed learning with compressed gradient differences.
arXiv: 1901.09269, 2019.