Error Compensated Loopless SVRG for Distributed Optimization

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

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

The Problem

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

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

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

Algorithm 1: Error compensated Loopless SVRG (EC-LSVRG)

$x^0 = w^0 \in \mathbb{R}^d$; $e^0_\tau = 0 \in \mathbb{R}^d$; $u^0 = 1 \in \mathbb{R}$; params: stepsize $\eta > 0$; 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

    $g^k_\tau = \nabla f^{(\tau)}_{i_k^\tau}(x^k) - \nabla f^{(\tau)}_{i_k^\tau}(w^k)$, $\quad y^k_\tau = Q(\eta g^k_\tau + e^k_\tau)$,

    $e^{k+1}_\tau = e^k_\tau + \eta g^k_\tau - y^k_\tau$, $\quad u^{k+1}_\tau = 0$ 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^k_τ and u^{k+1}_τ to the other nodes. Send $\nabla f^{(\tau)}(w^k)$ to the other nodes if $u^k = 1$

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

  end

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

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

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

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 $x \in \mathbb{R}^d$,
$$\mathbb{E}\|x - Q(x)\|^2 \le (1-\delta)\|x\|^2. \tag{2}$$
- $\tilde{Q}$ is an *unbiased compressor* if there is $\omega \ge 0$ such that
$$\mathbb{E}[\tilde{Q}(x)] = x \quad \text{and} \quad \mathbb{E}\|\tilde{Q}(x)\|^2 \le (\omega+1)\|x\|^2 \tag{3}$$
for all $x \in \mathbb{R}^d$.
- $\frac{1}{\omega+1}\tilde{Q}$ is a contraction compressor with $\delta = \frac{1}{\omega+1}$.

Assumptions

Assumption 1: $\mathbb{E}[Q(x)] = \delta x$.
Assumption 2: For $x_\tau = \eta g^k_\tau + e^k_\tau \in \mathbb{R}^d$, $\tau = 1, ..., n$ and $k \ge 0$ in Algorithm 1, 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, $f^{(\tau)}$ is $\bar{L}$-smooth, f is L_f-smooth, and ψ is μ-strongly convex. $L_f \ge \mu$.
Assumption 4: $f_i^{(\tau)}$ is L-smooth, $f^{(\tau)}$ is $\bar{L}$-smooth, f is L_f-smooth and f is μ-strongly convex.

Composite Case

Convergence Result

Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3 holds. Let $w_k = \left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{-k}$, $W_k = \sum_{i=0}^k w_i$, and $\bar{x}^k = \frac{1}{W_k}\sum_{i=0}^k w_i x^i$. If $\eta \le \frac{\delta^2}{135(1-\delta)(L+L\delta)+53L_f\delta^2+53L\delta^2/n}$, then we have $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \le$

$$\frac{\frac{\mu}{2}\|x^0-x^*\|^2 + \frac{1}{2}(P(x^0)-P(x^*))}{1-(1-\min\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\})^{k+1}}\left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

In particular, if we choose

$$\eta = \frac{\delta^2}{135(1-\delta)(\bar{L}+L\delta)+53L_f\delta^2+53L\delta^2/n},$$

then $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \le \epsilon$, with $\epsilon \le \frac{\mu}{2}\|x^0 - x^*\|^2 + \frac{1}{2}(P(x^0) - P(x^*))$, as long as

$$k \ge O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \frac{(1-\delta)\bar{L}}{\delta^2\mu} + \frac{(1-\delta)L}{\delta\mu} + \frac{L_f}{\mu} + \frac{L}{n\mu}\right)\ln\frac{1}{\epsilon}\right).$$

Convergence Result

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

$$\eta \le \frac{\delta^2}{(1-\delta)(269L_f+1100L/n+1503L\delta/n)+53L_f\delta^2+53L\delta^2/n},$$

then we have $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \le$

$$\frac{\frac{\mu}{2}\|x^0-x^*\|^2 + \frac{1}{2}(P(x^0)-P(x^*))}{1-(1-\min\{\frac{\mu\eta}{3},\frac{\delta}{4},\frac{p}{2}\})^{k+1}}\left(1 - \min\left\{\frac{\mu\eta}{3}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

In particular, if we choose

$$\eta = \frac{\delta^2}{(1-\delta)(269L_f+1100L/n+1503L\delta/n)+53L_f\delta^2+53L\delta^2/n},$$

then $\mathbb{E}[P(\bar{x}^k) - P(x^*)] \le \epsilon$, with $\epsilon \le \frac{\mu}{2}\|x^0 - x^*\|^2 + \frac{1}{2}(P(x^0) - P(x^*))$, as long as

$$k \ge O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \frac{(1-\delta)L_f}{\delta^2\mu} + \frac{(1-\delta)L}{n\delta\mu} + \frac{L_f}{\mu} + \frac{L}{n\mu}\right)\ln\frac{1}{\epsilon}\right).$$

Smooth Case ($\psi \equiv 0$)

Convergence Result

Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 4 holds. Let $w_k = \left(1 - \min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^{-k}$, $W_k = \sum_{i=0}^k w_i$, and $\bar{x}^k = \frac{1}{W_k}\sum_{i=0}^k w_i x^i$. If

$$\eta \le \min\left\{\frac{1}{4L_f+24L/n}, \frac{\delta}{51\sqrt{(1-\delta)L_f\bar{L}}}, \frac{\sqrt{\delta}}{51\sqrt{(1-\delta)L_f L}}\right\},$$

then we have $\mathbb{E}[f(\bar{x}^k) - f(x^*)] \le$

$$\frac{9\mu\|x^0-x^*\|^2 + 9(f(x^0)-f(x^*))}{1-(1-\min\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\})^{k+1}}\left(1 - \min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

In particular, if we choose

$$\eta = \min\left\{\frac{1}{4L_f+24L/n}, \frac{\delta}{51\sqrt{(1-\delta)L_f\bar{L}}}, \frac{\sqrt{\delta}}{51\sqrt{(1-\delta)L_f L}}\right\},$$

then $\mathbb{E}[f(\bar{x}^k) - f(x^*)] \le \epsilon$ with $\epsilon \le 9\mu\|x^0 - x^*\|^2 + 9f(x^0) - f(x^*)$, as long as $k \ge$

$$O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \frac{\sqrt{(1-\delta)L_f\bar{L}}}{\mu\delta} + \frac{\sqrt{(1-\delta)L_f L}}{\mu\sqrt{\delta}} + \frac{L_f}{\mu} + \frac{L}{n\mu}\right)\ln\frac{1}{\epsilon}\right).$$

Convergence Result

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

$$\eta \le \min\left\{\frac{1}{4L_f+32L/n}, \frac{\delta}{84\sqrt{1-\delta}L_f}, \frac{\sqrt{n\delta}}{138\sqrt{(1-\delta)L_f L}}, \frac{\sqrt{n}\delta}{118\sqrt{(1-\delta)L_f L}}\right\}$$

then we have $\mathbb{E}[f(\bar{x}^k) - f(x^*)] \le$

$$\frac{12\mu\|x^0-x^*\|^2 + 12(f(x^0)-f(x^*))}{1-(1-\min\{\frac{\mu\eta}{2},\frac{\delta}{4},\frac{p}{2}\})^{k+1}}\left(1 - \min\left\{\frac{\mu\eta}{2}, \frac{\delta}{4}, \frac{p}{2}\right\}\right)^k.$$

In particular, if we choose

$$\eta = \min\left\{\frac{1}{4L_f+32L/n}, \frac{\delta}{84\sqrt{1-\delta}L_f}, \frac{\sqrt{n\delta}}{138\sqrt{(1-\delta)L_f L}}, \frac{\sqrt{n}\delta}{118\sqrt{(1-\delta)L_f L}}\right\}$$

then $\mathbb{E}[f(\bar{x}^k) - f(x^*)] \le \epsilon$ with $\epsilon \le 12\mu\|x^0 - x^*\|^2 + 12(f(x^0) - f(x^*))$ as long as

$$k \ge O\left(\left(\frac{1}{\delta} + \frac{1}{p} + \frac{\sqrt{(1-\delta)}L_f}{\mu\delta} + \frac{L_f}{\mu} + \frac{L}{n\mu}\right)\ln\frac{1}{\epsilon}\right).$$

Optimal Choice of p

Denote the iteration complexity as $O\left(\left(\frac{1}{p} + a\right)\ln\frac{1}{\epsilon}\right)$, where a is independent of p. To minimize the total expected communication cost, the optimal choice of p is

$$O\left(\min\left\{r(Q), \frac{1}{a}\right\}\right) \le p \le O\left(\max\left\{r(Q), \frac{1}{a}\right\}\right).$$

Communication Cost

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

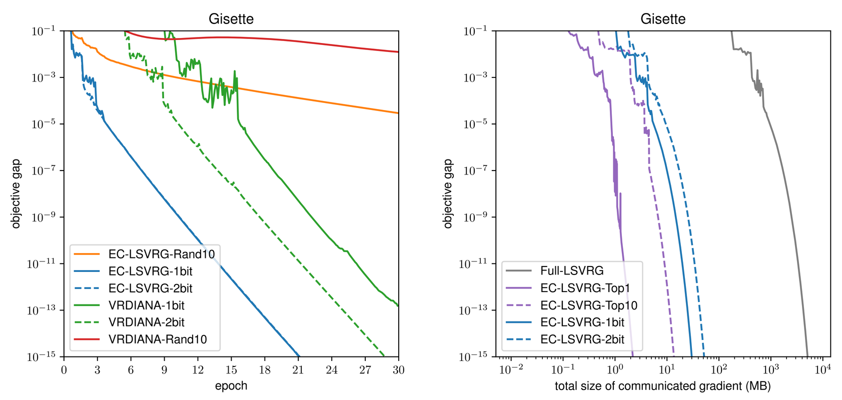
$$r(Q) := \sup_{x\in\mathbb{R}^d}\left\{\mathbb{E}\left[\frac{\text{communication cost of } Q(x)}{\Delta_1}\right]\right\}.$$

Assume $L_f = \bar{L} = L$ and $\Delta_1 r(Q) \ge O(1)$. Choose $p = O(r(Q))$.

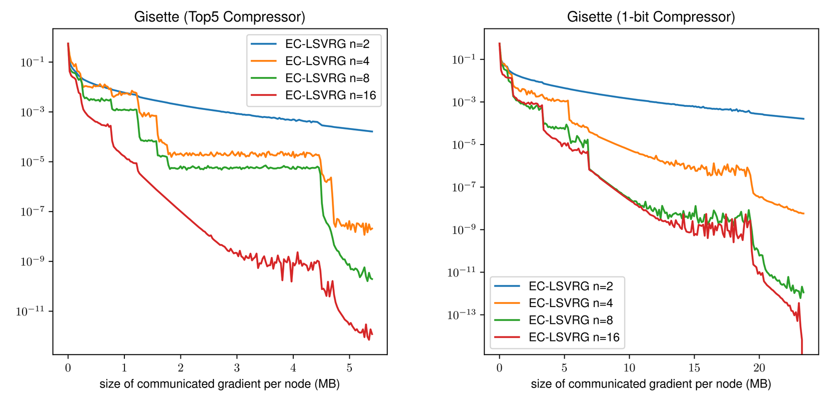
- Composite case:
$$O\left(\Delta_1\left(\frac{r(Q)}{\delta} + 1 + \left(r(Q) + \frac{(1-\delta)r(Q)}{\delta^2}\right)\frac{L}{\mu}\right)\ln\frac{1}{\epsilon}\right).$$
- Smooth case:
$$O\left(\Delta_1\left(\frac{r(Q)}{\delta} + 1 + \left(r(Q) + \frac{\sqrt{(1-\delta)}r(Q)}{\delta}\right)\frac{L}{\mu}\right)\ln\frac{1}{\epsilon}\right).$$

Numerical Results

1. Compare to compressed algorithm ($p = \frac{1}{mn}$)



2. Distributed Experiment ($p = 10^{-4}$)



References

[1] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik.
Stochastic distributed learning with gradient quantization and variance reduction.
arXiv: 1904.05115, 2019.

[2] Xun Qian, Zheng Qu, and Peter Richtárik.
L-svrg and l-katyusha with arbitrary sampling.
arXiv preprint arXiv:1906.01481, 2019.