Error Compensated Distributed SGD can be Accelerated

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

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

The Problem

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

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 Katyusha (`ECLK`)

$x^0 = y^0 = z^0 = w^0 \in \mathbb{R}^d$; $e^0_\tau = 0 \in \mathbb{R}^d$; $u^0 = 1 \in \mathbb{R}$;
params: $\eta = \frac{1}{3\theta_1} > 0$, $\mathcal{L}_1 > 0$, $\sigma_1 = \frac{\mu_f}{2\mathcal{L}_1} \geq 0$,
$\theta_1, \theta_2 \in (0,1)$; probability $p \in (0,1]$

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

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

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

    $g_\tau^k = \nabla f_{i_k^\tau}^{(\tau)}(x^k) - \nabla f_{i_k^\tau}^{(\tau)}(w^k)$, $\tilde{g}_\tau^k = Q(\frac{\eta}{\mathcal{L}_1} g_\tau^k + e_\tau^k)$,

    $e_\tau^{k+1} = e_\tau^k + \frac{\eta}{\mathcal{L}_1} g_\tau^k - \tilde{g}_\tau^k$, $u_\tau^{k+1} = 0$ for $\tau = 2, \ldots, n$,

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

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

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

  end

  $\tilde{g}^k = \frac{1}{n}\sum_{\tau=1}^n \tilde{g}_\tau^k$, $u^{k+1} = \sum_{\tau=1}^n u_\tau^{k+1}$

  $z^{k+1} =$

  $\mathrm{prox}_{\frac{\eta}{(1+\eta\sigma_1)\mathcal{L}_1}\psi}\left(\frac{1}{1+\eta\sigma_1}\left(\eta\sigma_1 x^k + z^k - \tilde{g}^k - \frac{\eta}{\mathcal{L}_1}\nabla f(w^k)\right)\right)$

  $y^{k+1} = x^k + \theta_1(z^{k+1} - z^k)$,

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

  $x^{k+1} = \theta_1 z^{k+1} + \theta_2 w^{k+1} + (1-\theta_1-\theta_2) y^{k+1}$

end

Gradient Compression Methods

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

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

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

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

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 $x_\tau = \frac{\eta}{\mathcal{L}_1} g_\tau^k + e_\tau^k \in \mathbb{R}^d$, $\tau = 1, \ldots, n$ and $k \geq 0$ in Algorithm 1, we have $\mathbb{E}[Q(x_\tau)] = Q(x_\tau)$, and $\left\|\sum_{\tau=1}^n (Q(x_\tau) - x_\tau)\right\|^2 \leq (1-\delta)\left\|\sum_{\tau=1}^n x_\tau\right\|^2$.

Assumption 3: $f_i^{(\tau)}$ is L-smooth, $f^{(\tau)}$ is $\bar{L}$-smooth, f is L_f-smooth and μ_f-strongly convex. ψ is μ_ψ-strongly convex. $\mu_f \geq 0$, $\mu_\psi \geq 0$ and $\mu = \mu_f + \mu_\psi > 0$.

Some Notations

Let $e^k = \frac{1}{n}\sum_{\tau=1}^n e_\tau^k$ and $\tilde{z}^k = z^k - \frac{1}{1+\eta\sigma_1} e^k$. Define $\tilde{\mathcal{Z}}^k = \frac{\mathcal{L}_1 + \eta\mu/2}{2\eta}\|\tilde{z}^k - x^*\|^2$, $\mathcal{Y}^k = \frac{1}{\theta_1}(P(y^k) - P^*)$, and $\mathcal{W}^k = \frac{\theta_2}{pq\theta_1}(P(w^k) - P^*)$.

Convergence Result

Define

$$\Phi^k := \tilde{\mathcal{Z}}^k + \mathcal{Y}^k + \mathcal{W}^k + \frac{4\mathcal{L}_1}{\delta\eta}\cdot\frac{1}{n}\sum_{\tau=1}^n \|e_\tau^k\|^2.$$

Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3 holds. If $\mathcal{L}_1 \geq \max\{L_f, 3\mu\eta\}$, $\theta_1 + \theta_2 \leq 1$, and $\theta_2 \geq \frac{\mathcal{L}_2}{2\mathcal{L}_1}$, then we have $\mathbb{E}\left[\Phi^k\right] \leq$

$$\left(1 - \min\left(\frac{\mu}{\mu + 6\theta_1\mathcal{L}_1}, \theta_1 + \theta_2 - \frac{\theta_2}{q}, p(1-q), \frac{\delta}{6}\right)\right)^k \Phi^0.$$

Convergence Result

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

$$\Psi^k := \tilde{\mathcal{Z}}^k + \mathcal{Y}^k + \mathcal{W}^k + \frac{4\mathcal{L}_1}{\delta\eta}\|e^k\|^2 + \frac{28\mathcal{L}_1(1-\delta)}{\delta\eta n}\cdot\frac{1}{n}\sum_{\tau=1}^n \|e_\tau^k\|^2.$$

If $\mathcal{L}_1 \geq \max\{L_f, 3\mu\eta\}$, $\theta_1 + \theta_2 \leq 1$, and $\theta_2 \geq \frac{\mathcal{L}_3}{2\mathcal{L}_1}$, then we have $\mathbb{E}\left[\Psi^k\right] \leq$

$$\left(1 - \min\left(\frac{\mu}{\mu + 6\theta_1\mathcal{L}_1}, \theta_1 + \theta_2 - \frac{\theta_2}{q}, p(1-q), \frac{\delta}{6}\right)\right)^k \Psi^0.$$

Iteration Complexity

Assume the compressor Q in Algorithm 1 is a contraction compressor and Assumption 3 holds. Let $\mathcal{L}_1 = \max(\mathcal{L}_4, L_f, 3\mu\eta)$, $\theta_2 = \frac{\mathcal{L}_4}{2\max\{L_f, \mathcal{L}_4\}}$ and

$$\theta_1 = \begin{cases} \min\left(\sqrt{\frac{\mu}{\mathcal{L}_4 p}}\theta_2, \theta_2\right) & \text{if } L_f \leq \frac{\mathcal{L}_4}{p} \\ \min\left(\sqrt{\frac{\mu}{L_f}}, \frac{p}{2}\right) & \text{otherwise} \end{cases}.$$

- Let $\mathcal{L}_4 = \mathcal{L}_2 := \frac{4L}{n} + \frac{112(1-\delta)\bar{L}}{9\delta^2} + \frac{56(1-\delta)L}{9\delta}$. Then with some $q \in [\frac{2}{3}, 1)$, $\mathbb{E}[\Phi^k] \leq \epsilon\Phi^0$ for $k \geq$

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

- Let $\mathcal{L}_4 = \mathcal{L}_3 := \frac{4L}{n} + \frac{784(1-\delta)L_f}{9\delta^2} + \frac{56(1-\delta)L}{\delta n}$. If Assumption 1 or Assumption 2 holds, then for some $q \in [\frac{2}{3}, 1)$, we have $\mathbb{E}[\Psi^k] \leq \epsilon\Psi^0$ for $k \geq$

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

However, if $L_f = \bar{L} = L$, then the two above iteration complexities become

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

Optimal Choice of p

To minimize the total expected communication cost, the optimal choice of p is $O(r(Q))$.

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 $\Delta_1 r(Q) \geq O(1)$. Choose $p = O(r(Q))$. The total expected communication cost of the error compensated loopless Katyusha is

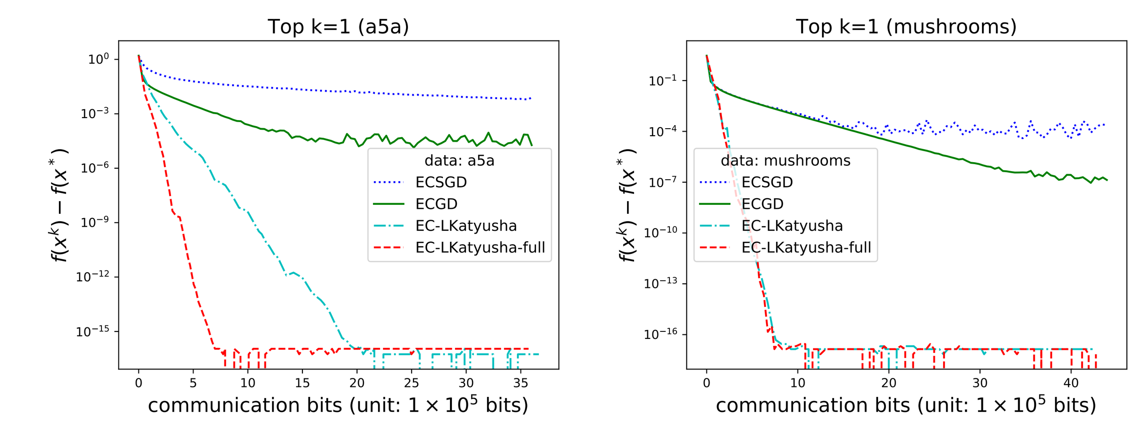
$$O\left(\Delta_1\left(\frac{r(Q)}{\delta} + \left(r(Q) + \frac{\sqrt{r(Q)}}{\sqrt{n}} + \frac{\sqrt{(1-\delta)r(Q)}}{\delta}\right)\sqrt{\frac{L}{\mu}}\right)\ln\frac{1}{\epsilon}\right).$$

For uncompressed L-Katyusha, by choosing $p = 1$, the total expected communication cost is

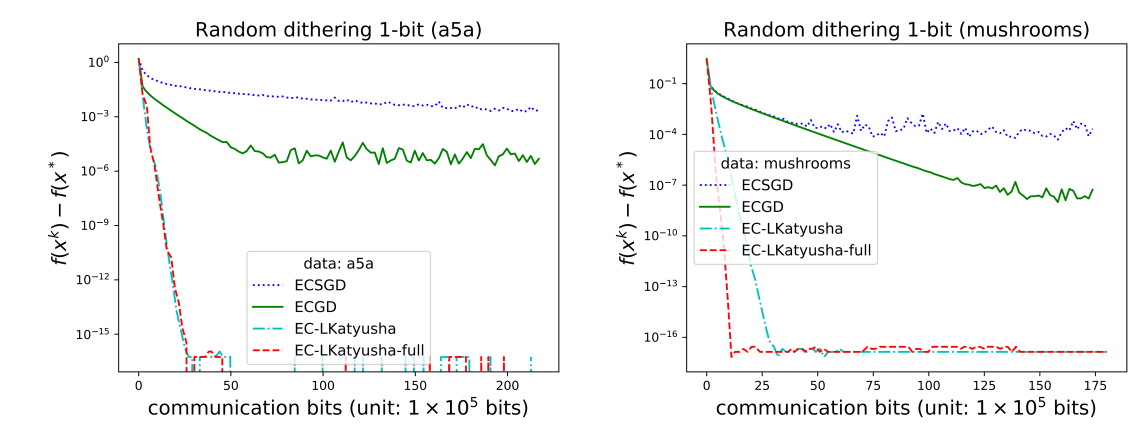
$$O\left(\Delta_1\sqrt{\frac{L}{\mu}}\ln\frac{1}{\epsilon}\right).$$

Numerical Results

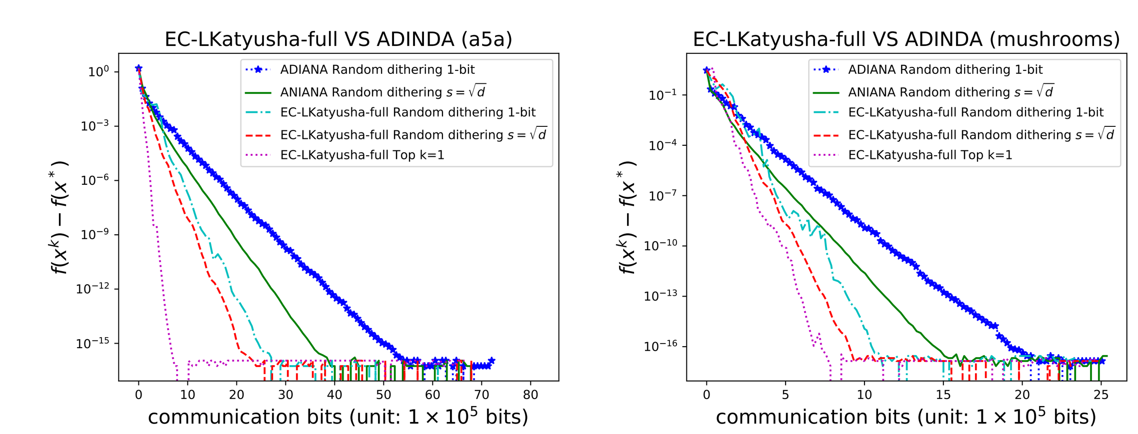
1. ECSGD vs ECGD vs EC-LKatyusha vs EC-LKatyusha-full for Top k=1 compressor



2. ECSGD vs ECGD vs EC-LKatyusha vs EC-LKatyusha-full for Random dithering 1-bit compressor



3. EC-LKatyusha-full vs ADIANA



References

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

[2] Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.