CSER: Communication-efficient SGD with Error Reset

Cong Xie, Shuai Zheng, Oluwasanmi Koyejo, Indranil Gupta, Mu Li, Haibin Lin
Presenter: Cong Xie

November 16, 2020

Outline

1 Motivations

2 Preliminaries

3 CSER

4 Evaluation

5 Takeaways

Motivations

Motivations

- Reduce communication for distributed SGD

Motivations

- Reduce communication for distributed SGD
- Good convergence when communication is very low

Outline

1 Motivations

2 Preliminaries

3 CSER

4 Evaluation

5 Takeaways

Distributed SGD

Algorithm 1 Distributed SGD

```
Initialize $x_{i,0} = \hat{x}_0 \in \mathbb{R}^d, \forall i \in [n]$
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do
        $p_{i,t} \leftarrow -\eta \nabla f(x_{i,t-1}; z_{i,t})$
        $\bar{p}_t \leftarrow \frac{1}{n} \sum_{i \in [n]} p_{i,t}$
        $x_{i,t} \leftarrow x_{i,t-1} + \bar{p}_t$
    end for
end for
```

average, communication

Distributed EF-SGD

Algorithm 2 EF-SGD

```
**Input**: $\mathcal{C}_1$ - compressor
Initialize $x_{i,0} = \hat{x}_0 \in \mathbb{R}^d, e_{i,t} = \mathbf{0}, \forall i \in [n]$
**for all** iteration $t \in [T]$ **do**
    **for all** Workers $i \in [n]$ in parallel **do**
        EF: $p_{i,t} \leftarrow e_{i,t-1} - \eta \nabla f(x_{i,t-1}; z_{i,t})$,    $p'_{i,t} \leftarrow \mathcal{C}_1(p_{i,t})$,    $e_{i,t} \leftarrow p_{i,t} - p'_{i,t}$
        $\bar{p}'_t \leftarrow \frac{1}{n} \sum_{i\in[n]} p'_{i,t}$
        $x_{i,t} \leftarrow x_{i,t-1} + \bar{p}'_t$
    **end for**
**end for**
```

compressed message

residual error

QSparse-local-SGD

Algorithm 3 Qsparse-local-SGD

```
**Input**: $\mathcal{C}_1$ - compressor, $H > 0$ - synchronization interval
Initialize $x_{i,0} = \hat{x}_0 = \mathbf{0}, \forall i \in [n]$, $e_{i,0} = \mathbf{0}, \forall i \in [n]$
**for all** iteration $t \in [T]$ **do**
    **for all** Workers $i \in [n]$ in parallel **do**
        $x_{i,t-\frac{1}{2}} \leftarrow x_{i,t-1} - \eta \nabla f(x_{i,t-1}; z_{i,t})$
        **if** $\mod(t, H) \neq 0$ **then**
            Local update: $x_{i,t} \leftarrow x_{i,t-\frac{1}{2}}, \quad \hat{x}_t \leftarrow \hat{x}_{t-1}, \quad e_{i,t} \leftarrow e_{i,t-1}$
        **else**
            EF: $p_{i,t} \leftarrow e_{i,t-1} + x_{i,t-\frac{1}{2}} - \hat{x}_{t-1}, \quad p'_{i,t} \leftarrow \mathcal{C}_1(p_{i,t}), \quad e_{i,t} \leftarrow p_{i,t} - p'_{i,t}$
            $\bar{p}'_t \leftarrow \frac{1}{n} \sum_{i \in [n]} p'_{i,t}, \quad x_{i,t} \leftarrow \hat{x}_{t-1} + \bar{p}'_t, \quad \hat{x}_t \leftarrow \hat{x}_{t-1} + \bar{p}'_t$
        **end if**
    **end for**
**end for**
```

Outline

1 Motivations

2 Preliminaries

3 CSER

4 Evaluation

5 Takeaways

Error Feedback

Observations:

Error Feedback

Observations:

- With single worker, EF-SGD isn't reduced to vanilla SGD

Error Feedback

Observations:

- With single worker, EF-SGD isn't reduced to vanilla SGD
- Aggressive compressors (random sparsifier + high compression ratio) results in bad convergence

Error Feedback

Observations:

- With single worker, EF-SGD isn't reduced to vanilla SGD
- Aggressive compressors (random sparsifier + high compression ratio) results in bad convergence
- Worse for QSparse-local-SGD

Error Reset

Error reset:

Error Reset

Error reset:

- Apply residual error to local model

Error Reset

Error reset:

- Apply residual error to local model
- Different workers maintain different local model

Error Reset

Error reset:

- Apply residual error to local model
- Different workers maintain different local model
- Use local models to compute gradients

Error Reset

Error reset:

- Apply residual error to local model
- Different workers maintain different local model
- Use local models to compute gradients
- Periodically reset/flush the error inside local models

Error Reset

Error Reset

Algorithm 5 Partial Synchronization (PSync)

```
function PSync($v_i \in \mathbb{R}^d$, $\mathcal{C}$ - compressor)
    On worker $i$:
    $v_i' = \mathcal{C}(v_i)$
    $r_i = v_i - v_i'$
    Partial synchronization:
        $\bar{v}' = \frac{1}{n} \sum_{i \in [n]} v_i'$
    $v_i' = \bar{v}' + r_i$
    return $v_i', r_i$
end function
```

Error Reset

Algorithm 4 Error Reset

```
Input: $\mathcal{C}_1$ - compressor
Initialize $x_{i,0} = \hat{x}_0 \in \mathbb{R}^d, e_{i,0} = \mathbf{0}, \forall i \in [n]$
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do
        $p_{i,t} \leftarrow e_{i,t-1} - \eta \nabla f(x_{i,t-1}; z_{i,t})$
        $p'_{i,t}, e_{i,t} \leftarrow PSync(p_{i,t}, \mathcal{C}_1)$
        $x_{i,t} \leftarrow x_{i,t-1}$
    end for
end for
```

Algorithm 5 Partial Synchronization (PSync)

```
function PSYNC($v_i \in \mathbb{R}^d$, $\mathcal{C}$ - compressor)
    On worker $i$:
    $v'_i = \mathcal{C}(v_i)$
    $r_i = v_i - v'_i$
    Partial synchronization:
        $\bar{v}' = \frac{1}{n} \sum_{i \in [n]} v'_i$
    $v'_i = \bar{v}' + r_i$
    return $v'_i, r_i$
end function
```

Error Reset

Algorithm 4 Error Reset

```
Input: $\mathcal{C}_1$ - compressor
Initialize $x_{i,0} = \hat{x}_0 \in \mathbb{R}^d, e_{i,0} = \mathbf{0}, \forall i \in [n]$
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do
        $p_{i,t} \leftarrow e_{i,t-1} - \eta \nabla f(x_{i,t-1}; z_{i,t})$
        $p'_{i,t}, e_{i,t} \leftarrow PSync(p_{i,t}, \mathcal{C}_1)$
        $x_{i,t} \leftarrow x_{i,t-1} - e_{i,t-1}$
    end for
end for
```

prev. residual error

Algorithm 5 Partial Synchronization (PSync)

```
function PSYNC($v_i \in \mathbb{R}^d$, $\mathcal{C}$ - compressor)
    On worker $i$:
    $v'_i = \mathcal{C}(v_i)$
    $r_i = v_i - v'_i$
    Partial synchronization:
        $\bar{v}' = \frac{1}{n} \sum_{i \in [n]} v'_i$
    $v'_i = \bar{v}' + r_i$
    return $v'_i, r_i$
end function
```

Error Reset

Algorithm 4 Error Reset

```
Input: $\mathcal{C}_1$ - compressor
Initialize $x_{i,0} = \hat{x}_0 \in \mathbb{R}^d, e_{i,0} = \mathbf{0}, \forall i \in [n]$
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do
        $p_{i,t} \leftarrow e_{i,t-1} - \eta \nabla f(x_{i,t-1}; z_{i,t})$
        $p'_{i,t}, e_{i,t} \leftarrow PSync(p_{i,t}, \mathcal{C}_1)$
        $x_{i,t} \leftarrow x_{i,t-1} - e_{i,t-1} + p'_{i,t}$
    end for
end for
```

partially synchronized

prev. residual error

Algorithm 5 Partial Synchronization (PSync)

```
function PSYNC($v_i \in \mathbb{R}^d$, $\mathcal{C}$ - compressor)
    On worker $i$:
    $v'_i = \mathcal{C}(v_i)$
    $r_i = v_i - v'_i$
    Partial synchronization:
        $\bar{v}' = \frac{1}{n} \sum_{i \in [n]} v'_i$
    $v'_i = \bar{v}' + r_i$
    return $v'_i, r_i$
end function
```

Error Reset vs. Error Feedback

	Error feedback	Error reset

Error Reset vs. Error Feedback

	Error feedback	Error reset
Residual error	In $e_{i,t}$	In $e_{i,t}$ and $x_{i,t}$

Error Reset vs. Error Feedback

	Error feedback	Error reset
Residual error	In $e_{i,t}$	In $e_{i,t}$ and $x_{i,t}$
Synchronized variable	$x_{i,t}$	$x_{i,t} - e_{i,t}$

Error Reset vs. Error Feedback

	Error feedback	Error reset
Residual error	In $e_{i,t}$	In $e_{i,t}$ and $x_{i,t}$
Synchronized variable	$x_{i,t}$	$x_{i,t} - e_{i,t}$
Convergence error	$\frac{1}{n}\sum_{i\in[n]}\left\|e_{i,t}\right\|^2$	$\frac{1}{n}\sum_{i\in[n]}\left\|e_{i,t}\right\|^2 - \left\|\frac{1}{n}\sum_{i\in[n]} e_{i,t}\right\|^2$

CSER

Algorithm 6 CSER

```
Input: $\mathcal{C}_1, \mathcal{C}_2$ - compressors, $H > 0$ - error-reset interval
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do



        if $\mod (t, H) \neq 0$ then

        else


        end if
    end for
```

CSER

Algorithm 6 CSER

```
Input: $\mathcal{C}_1, \mathcal{C}_2$ - compressors, $H > 0$ - error-reset interval
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do
        $g_{i,t} \leftarrow \nabla f(x_{i,t-1}; z_{i,t})$, $z_{i,t} \sim \mathcal{D}_i$


        if $\mod(t, H) \neq 0$ then

        else


        end if
    end for
```

CSER

Algorithm 6 CSER

```
Input: $\mathcal{C}_1, \mathcal{C}_2$ - compressors, $H > 0$ - error-reset interval
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do
        $g_{i,t} \leftarrow \nabla f(x_{i,t-1}; z_{i,t}),\ z_{i,t} \sim \mathcal{D}_i$


        if $\mod (t, H) \neq 0$ then
            $x_{i,t} \leftarrow x_{i,t-\frac{1}{2}}, \quad e_{i,t} \leftarrow e_{i,t-\frac{1}{2}}$
        else
            $e'_{i,t-\frac{1}{2}}, e_{i,t} \leftarrow PSync(e_{i,t-\frac{1}{2}}, \mathcal{C}_1)$        ▷ model partial sync.
            $x_{i,t} \leftarrow x_{i,t-\frac{1}{2}} - e_{i,t-\frac{1}{2}} + e'_{i,t-\frac{1}{2}}$
        end if
    end for
```

CSER

Algorithm 6 CSER

```
Input: $\mathcal{C}_1, \mathcal{C}_2$ - compressors, $H > 0$ - error-reset interval
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do
        $g_{i,t} \leftarrow \nabla f(x_{i,t-1}; z_{i,t}),\ z_{i,t} \sim \mathcal{D}_i$
        $g'_{i,t}, r_{i,t} \leftarrow PSync(g_{i,t}, \mathcal{C}_2)$          ▷ gradient partial sync.

        if $\mod (t, H) \neq 0$ then
            $x_{i,t} \leftarrow x_{i,t-\frac{1}{2}}, \quad e_{i,t} \leftarrow e_{i,t-\frac{1}{2}}$
        else
            $e'_{i,t-\frac{1}{2}}, e_{i,t} \leftarrow PSync(e_{i,t-\frac{1}{2}}, \mathcal{C}_1)$          ▷ model partial sync.
            $x_{i,t} \leftarrow x_{i,t-\frac{1}{2}} - e_{i,t-\frac{1}{2}} + e'_{i,t-\frac{1}{2}}$
        end if
    end for
```

CSER

Algorithm 6 CSER

```
Input: $\mathcal{C}_1, \mathcal{C}_2$ - compressors, $H > 0$ - error-reset interval
for all iteration $t \in [T]$ do
    for all Workers $i \in [n]$ in parallel do
        $g_{i,t} \leftarrow \nabla f(x_{i,t-1}; z_{i,t}),\ z_{i,t} \sim \mathcal{D}_i$
        $g'_{i,t}, r_{i,t} \leftarrow PSync(g_{i,t}, \mathcal{C}_2)$                ▷ gradient partial sync.
        $x_{i,t-\frac{1}{2}} \leftarrow x_{i,t-1} - \eta g'_{i,t}, \quad e_{i,t-\frac{1}{2}} \leftarrow e_{i,t-1} - \eta r_{i,t}$
        if $\mod (t, H) \neq 0$ then
            $x_{i,t} \leftarrow x_{i,t-\frac{1}{2}}, \quad e_{i,t} \leftarrow e_{i,t-\frac{1}{2}}$
        else
            $e'_{i,t-\frac{1}{2}}, e_{i,t} \leftarrow PSync(e_{i,t-\frac{1}{2}}, \mathcal{C}_1)$       ▷ model partial sync.
            $x_{i,t} \leftarrow x_{i,t-\frac{1}{2}} - e_{i,t-\frac{1}{2}} + e'_{i,t-\frac{1}{2}}$
        end if
    end for
```

CSER

CSER:

CSER

CSER:

- Use error reset instead of error feedback

CSER

CSER:

- Use error reset instead of error feedback
- Add gradient partial sync. between model partial sync.

CSER

CSER:

- Use error reset instead of error feedback
- Add gradient partial sync. between model partial sync.
- Special cases:

CSER

CSER:

- Use error reset instead of error feedback
- Add gradient partial sync. between model partial sync.
- Special cases:

  $\mathcal{C}_2 = 0$: Partial-local-SGD/CSER-PL (replace EF by ER in QSparse-local-SGD)

CSER

CSER:

- Use error reset instead of error feedback
- Add gradient partial sync. between model partial sync.
- Special cases:

  $\mathcal{C}_2 = 0$: Partial-local-SGD/CSER-PL (replace EF by ER in QSparse-local-SGD)

  $\mathcal{C}_2 = 0, H = 1$: CSEA (replace EF by ER in EF-SGD)

CSEA vs. EF-SGD

Algorithm 7 CSEA

```
**Input**: $\mathcal{C}_1$ - compressor
Initialize $x_{i,0} = \hat{x}_0 \in \mathbb{R}^d, e_{i,t} = \mathbf{0}, \forall i \in [n]$
**for all** iteration $t \in [T]$ **do**
    **for all** Workers $i \in [n]$ in parallel **do**
        $p_{i,t} \leftarrow e_{i,t-1} - \eta \nabla f(x_{i,t-1}; z_{i,t})$
        $p'_{i,t} \leftarrow \mathcal{C}_1(p_{i,t})$
        $e_{i,t} \leftarrow p_{i,t} - p'_{i,t}$
        $\bar{p}'_t \leftarrow \frac{1}{n} \sum_{i \in [n]} p'_{i,t}$
        $x_{i,t} \leftarrow x_{i,t-1} - e_{i,t-1} + e_{i,t} + \bar{p}'_t$
    **end for**
**end for**
```

Algorithm 8 EF-SGD

```
**Input**: $\mathcal{C}_1$ - compressor
Initialize $x_{i,0} = \hat{x}_0 \in \mathbb{R}^d, e_{i,t} = \mathbf{0}, \forall i \in [n]$
**for all** iteration $t \in [T]$ **do**
    **for all** Workers $i \in [n]$ in parallel **do**
        $p_{i,t} \leftarrow e_{i,t-1} - \eta \nabla f(x_{i,t-1}; z_{i,t})$
        $p'_{i,t} \leftarrow \mathcal{C}_1(p_{i,t})$
        $e_{i,t} \leftarrow p_{i,t} - p'_{i,t}$
        $\bar{p}'_t \leftarrow \frac{1}{n} \sum_{i \in [n]} p'_{i,t}$
        $x_{i,t} \leftarrow x_{i,t-1} + \bar{p}'_t$
    **end for**
**end for**
```

Outline

1 Motivations

2 Preliminaries

3 CSER

4 Evaluation

5 Takeaways

Evaluation

Experiments:

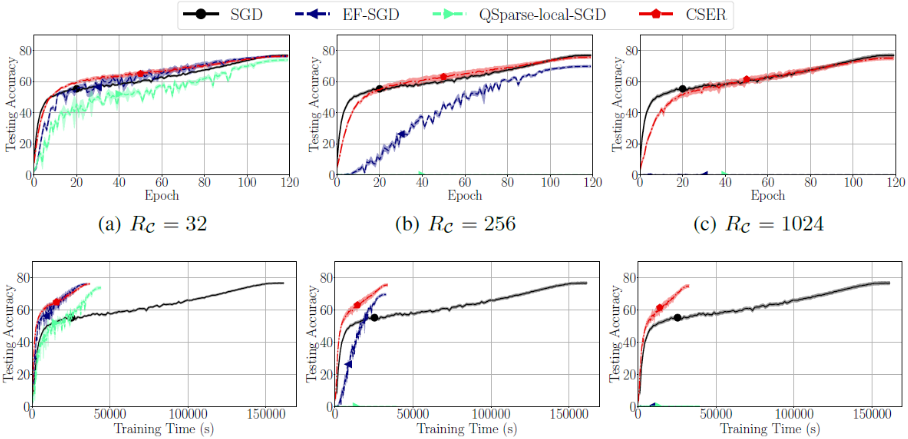
- Baseline: SGD, EF-SGD, QSparse-local-SGD
- Datasets (model): CIFAR-100 (WideResNet-40-8), ImageNet (ResNet-50)
- Compressor: random sparsifier
- 8 P3.2xlarge instances: V100 GPU, 10Gbps bandwidth
- AllReduce for average

Evaluation

Testing accuracy (%) on CIFAR-100 with different overall compression ratios ($R_{\mathcal{C}}$).

	Baseline			Proposed algorithm		
Optimizer/ $R_{\mathcal{C}}$	SGD	EF-SGD	QSparse-local -SGD	CSEA	CSER	CSER-PL
1	**87.01±0.11**	None	None	None	None	None
2	None	87.20±0.10	87.16±0.03	87.17±0.21	**87.47±0.03**	None
4	None	86.97±0.08	87.08±0.22	87.25±0.23	87.22±0.03	**87.33±0.05**
8	None	86.61±0.23	87.15±0.10	87.14±0.05	87.09±0.05	**87.27±0.04**
16	None	85.69±0.31	87.02±0.13	87.15±0.09	**87.28±0.04**	86.72±0.05
32	None	85.17±0.12	86.70±0.04	86.83±0.20	86.90±0.15	**86.92±0.26**
64	None	84.65±0.07	80.64±0.47	86.63±0.16	86.78±0.11	**86.91±0.15**
128	None	83.50±0.87	70.27±2.37	86.30±0.15	**86.81±0.17**	86.36±0.21
256	None	83.92±0.55	diverge	86.34±0.20	**86.68±0.07**	86.27±0.02
512	None	76.05±0.56	diverge	85.75±0.34	**86.20±0.09**	85.68±0.12
1024	None	diverge	diverge	85.13±0.13	**85.66±0.07**	84.94±0.37

Evaluation



(a) $R_{\mathcal{C}} = 32$ (b) $R_{\mathcal{C}} = 256$ (c) $R_{\mathcal{C}} = 1024$

(d) $R_{\mathcal{C}} = 32$ (e) $R_{\mathcal{C}} = 256$ (f) $R_{\mathcal{C}} = 1024$

Figure 2: Testing accuracy with different algorithms, for ResNet-50 on ImageNet.

Acceleration

$256\times$ compression ratio:

- $10\times$ for CIFAR-100
- $4.5\times$ for ImageNet

Outline

1 Motivations

2 Preliminaries

3 CSER

4 Evaluation

5 Takeaways

Takeaways

- Error reset improves convergence for aggressive compressors
- A better alternative of error feedback
- Combination of gradient and model compression with tuned hyperparameters improves convergence

Q & A