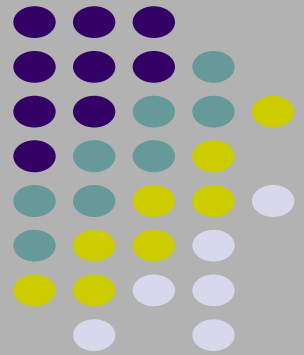
Detached Error Feedback for Distributed SGD with Random Sparsification

An Xu[1], Heng Huang[1]

[1]Electrical and Computer Engineering Department, University of Pittsburgh, PA, USA



Background

Communication compression for distributed sgd

- Computation overheads
  - Top-κ : $O(\kappa \log_2 d)$, inefficient on GPUs.
  - Random sparsification (RBGS): $O(1)$; ring-allreduce
  - large compression error, inferior performance.

Error Feedback

- Add current compression error to the next iteration
  - Assumptions 3.1 & 3.2 to bound the compression error

Assumption 3.1. (Bounded Variance) $\forall \theta \in \mathbb{R}^d$, the variance of the stochastic gradient satisfies $\mathbb{E}_{\xi \in \mathcal{S}} \|\nabla f(\theta; \xi) - \nabla F_{\mathcal{S}}(\theta)\|_2^2 \leq \sigma^2$.

Assumption 3.2. (Bounded Second Moment) $\forall \theta \in \mathbb{R}^d$, the second moment of the full gradient satisfies $\|\nabla F_{\mathcal{S}}(\theta)\|_2^2 \leq M^2$.

DEF

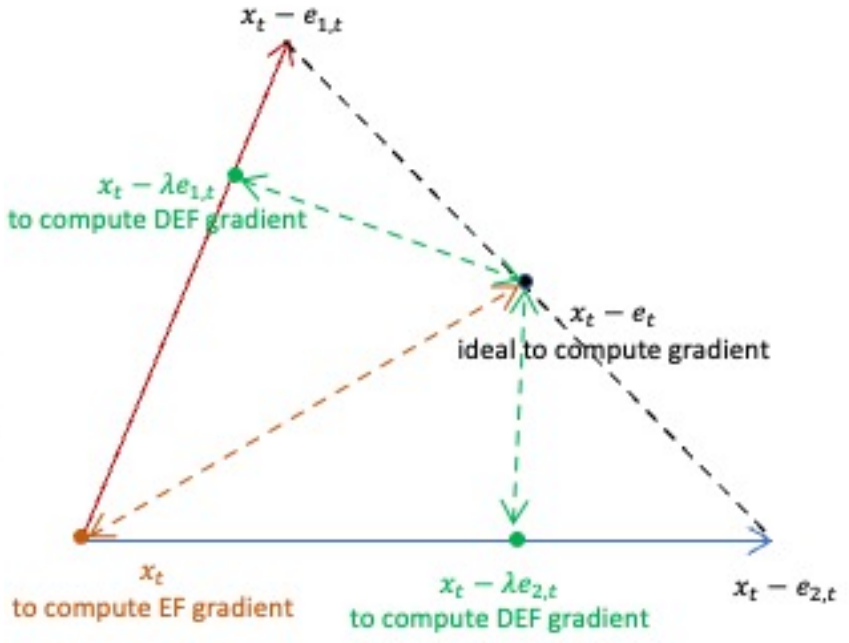
Detached Error Feedback

- Trade-off when second-moment cannot be bounded

$$(\alpha\sigma)^2 + ((1-\alpha)M)^2 \overset{\alpha=\frac{M^2}{\sigma^2+M^2}}{\geq} \frac{\sigma^2 M^2}{\sigma^2 + M^2} \overset{M\gg\sigma}{=} \sigma^2$$

- Compute gradient at a different point

$$\frac{1}{K}\sum_{k=1}^{K} \|x_t - \lambda e_{k,t} - y_t\|_2^2 = \frac{1}{K}\sum_{k=1}^{K} \|e_t - \lambda e_{k,t}\|_2^2 .$$



Theoretical Results

Trade-off for better convergence bounds without bounding second moment

- compression error $O(\frac{2\sigma^2 M^2}{\sigma^2+M^2})$

Trade-off for better excess risk

Theorem 4.6. *(Excess Risk Error of DEF(-A), Appendix B) Let Assumptions 3.1, 3.2, 3.3, 4.1 and 4.2 hold. Suppose* $\eta = \frac{c}{t+1}$, *where* $c > 0$ *is some constant.*

(1) The generalization error of DEF

$$\epsilon_{gen} = \mathcal{O}(T^{(1-\frac{K}{N})Lc/((1-\frac{K}{N})Lc+1)}) . \quad (22)$$

(2) Suppose $\eta \leq \frac{1}{4L}$. *The optimization error of DEF*

$$\epsilon_{opt} = \widetilde{\mathcal{O}}(T^{-\frac{\mu c}{2}} + T^{-1}) . \quad (23)$$

(3) For RBGS, the generalization error of DEF-A

$$\epsilon_{gen} = \mathcal{O}(T^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc/((1-\frac{K}{N})\delta^{\frac{1}{2}}Lc+1)}) . \quad (24)$$

(4) Suppose $\eta \leq \frac{1}{8L}$. *The optimization error of DEF-A*

$$\epsilon_{opt} = \widetilde{\mathcal{O}}(T^{-\frac{\mu\delta c}{2}} + T^{-1} + (1/\sqrt{1-\delta}-1)^{-2}) . \quad (25)$$

Extends to iterate averaging (IA)

- IA is a special case of DEF-A with compressor $C(\Delta) = \delta\Delta$
- Excess risk error analysis for IA
  - Trade-off for better excess risk of IA than SGD

Experiments

Table 2. The ImageNet test accuracy (%) comparison under various compression ratio settings with ResNet-50.

Ratio	SGD	EF	SAEF	PSync	DEF	DEF-A
1	76.04	—	—	—	—	—
16	—	75.29 (↓ 0.75)	75.83 (↓ 0.21)	75.63 (↓ 0.41)	75.98 (↓ 0.06)	**76.10** (↑ 0.06)
64	—	73.05 (↓ 2.99)	74.65 (↓ 1.39)	74.84 (↓ 1.20)	76.16 (↑ 0.12)	**76.37** (↑ 0.33)
128	—	63.80 (↓ 12.2)	74.26 (↓ 1.78)	74.12 (↓ 1.92)	**76.17** (↑ 0.13)	76.14 (↑ 0.10)
256	—	diverge	73.83 (↓ 2.21)	73.02 (↓ 3.02)	75.71 (↓ 0.33)	**76.00** (↓ 0.04)
512	—	diverge	73.00 (↓ 3.04)	72.60 (↓ 3.44)	75.52 (↓ 0.52)	**75.77** (↓ 0.27)
1024	—	diverge	71.89 (↓ 4.15)	71.82 (↓ 4.22)	**75.64** (↓ 0.40)	75.57 (↓ 0.47)

Take-aways

A new communication-efficient distributed training method DEF without bounding second moment.

First excessive risk analysis for communication-efficient distributed training.

Analysis can be extended to iterate averaging.

Empirical test accuracy improvement.

Poster Hall E #407