Temporal Predictive Coding for Gradient Compression in Distributed Learning

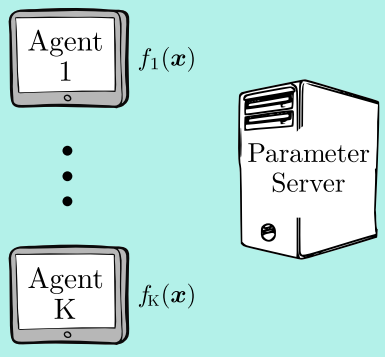
Adrian Edin[1], **Zheng Chen**[1], **Michel Kieffer**[2] **and Mikael Johansson**[3]
[1]Div. of Communication Systems, Dept. of Electrical Engineering (ISY), Linköping University (LiU), Sweden
[2]Université Paris-Saclay, CentraleSupelec, CNRS, Laboratoire des Signaux et Systèmes, Gif-sur-Yvette, France
[3]School of Electrical Engineering and Computer Science, KTH, Stockholm, Sweden

Distributed Learning

In the context of distributed learning we consider K distributed agents with local (possibly private) datasets. A global mapping $f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ is trained using local loss functions $f_k(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}, \forall k = 1, ..., K$, parameterized by $\boldsymbol{x} \in \mathbb{R}^d$.

We seek the solution to

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\, f(\boldsymbol{x}) = \underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \sum_{k=1}^{N} f_k(\boldsymbol{x}),$$



First order method: Distributed gradient decent

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \gamma \nabla f(\boldsymbol{x}^{(t)}) = \boldsymbol{x}^{(t)} - \gamma \boldsymbol{g}^{(t)}$$

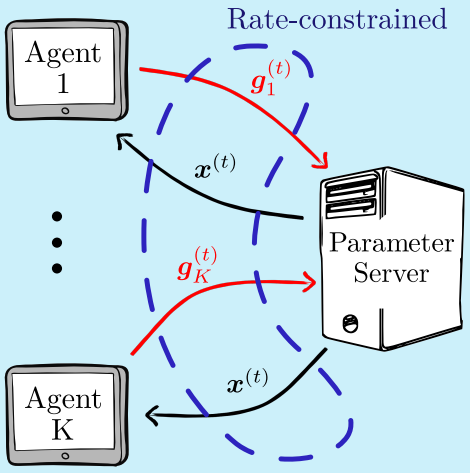
$$\boldsymbol{g}^{(t)} = \sum_{k=1}^{K} \boldsymbol{g}_k^{(t)} = \sum_{k=1}^{K} \nabla f_k(\boldsymbol{x}^{(t)})$$

Necessity of Compression

Four steps in each iteration:

1. PS broadcasts $\boldsymbol{x}^{(t)}$ to agents,
2. Agents compute local gradients $\boldsymbol{g}_k^{(t)}$,
3. Each agent transmits $\boldsymbol{g}_k^{(t)}$ to PS,
4. PS updates global model using aggregated $\boldsymbol{g}^{(t)}$.

Communication bottleneck problem arises when all agents simultaneously transmites their gradient to PS over **rate-constrained** digital links.



Common solutions:

- Reduce number of bits (Sparsification, quantization)
- Reduce number of particiapating agents (scheduling)
- Reduce transmission frequency (adaptive communication)

Emerging method:

Prediction-based approaches

Gradient Prediction

Motivation: Consecutive gradients might be correlated in time

$$\mathcal{M}_k^{(t)} = \left\{\widetilde{\boldsymbol{g}}_k^{(t-1)}, \widetilde{\boldsymbol{g}}_k^{(t-2)}, \dots\right\}$$

Agent k memory consisting of previous imperfect gradient estimates, stored by both agent k and the PS

$$\widehat{\boldsymbol{g}}_k^{(t)} = \sum_{i=1}^{s} a_{k,i}^{(t)} \widetilde{\boldsymbol{g}}_k^{(t-i)}$$

Linear predictor of the current gradient

$$\boldsymbol{a}_k^{*(t)} = \underset{\boldsymbol{a}_k^{(t)}\in\mathbb{R}^s}{\text{argmin}} \left\| \boldsymbol{g}_k^{(t)} - \widehat{\boldsymbol{g}}_k^{(t)} \right\|^2$$

Optimal prediction coefficient from the least-square (LS) estimate.

$$\boldsymbol{e}_k^{(t)} = \boldsymbol{g}_k^{(t)} - \widehat{\boldsymbol{g}}_k^{(t)}$$

Prediction residual

Properties from the LS-estimate:

- Orthogonality $\widehat{\boldsymbol{g}}_k^{(t)} \perp \boldsymbol{e}_k^{(t)}$
- Norm reduction: $\left\|\boldsymbol{e}_k^{(t)}\right\| \leq \left\|\boldsymbol{g}_k^{(t)}\right\|$

Event-triggered Prediction Residual Compression

To further reduce transmitted information, we compressed the residual using event-triggered transmission

$$\widetilde{\boldsymbol{e}}^{(t)} = \mathcal{C}\left(\boldsymbol{e}_k^{(t)}\right) = \begin{cases} 0, & \text{if } \left\|\boldsymbol{e}_k^{(t)}\right\| \leq e_{\text{th},k}^{(t)}, \\ \mathcal{Q}(\boldsymbol{e}_k^{(t)}), & \text{otherwise,} \end{cases}$$

$$\widetilde{\boldsymbol{g}}_k^{(t)} = \widehat{\boldsymbol{g}}_k^{(t)} + \widetilde{\boldsymbol{e}}_k^{(t)} \qquad \text{Imperfect gradient estimates}$$

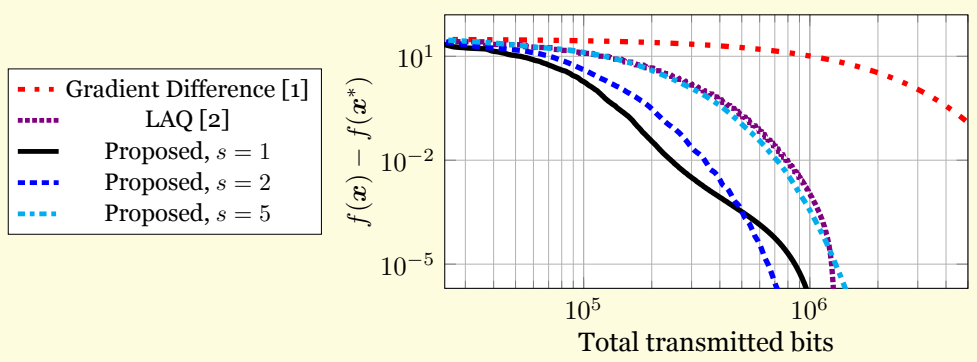
Trade-off: Aggregated gradient accuracy vs. Omitted residual transmissions.

We use the imperfect gradient estimate to update the parameters

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \gamma \widetilde{\boldsymbol{g}}^{(t)} = \boldsymbol{x}^{(t)} - \gamma \sum_{k=1}^{K} \widetilde{\boldsymbol{g}}_k^{(t)}$$

Simulation results

- $K = 10$ agents
- $T = 1000$ iterations
- Memory length s
- w8a dataset w. $d = 300$ parameters
- $e_{th,k}^{(t)} = c^{(t)} \left\|\boldsymbol{g}_k^{(t)}\right\|$, where $c^{(t)} = t/(K \cdot T)$
- $\mathcal{Q}(\cdot)$ is a stochastic quantizer.
- Quantized gradient/residual elements are further compressed by an entropy encoder (*e.g.* Huffman coding). Each residual transmission uses $R \times d$ bits, *i.e.*, $R = 6$ bits on average per residual element.



Comparison of communication cost until $f(\boldsymbol{g}^{(t)}) - f(\boldsymbol{x}^*) \leq 10^{-5}$.

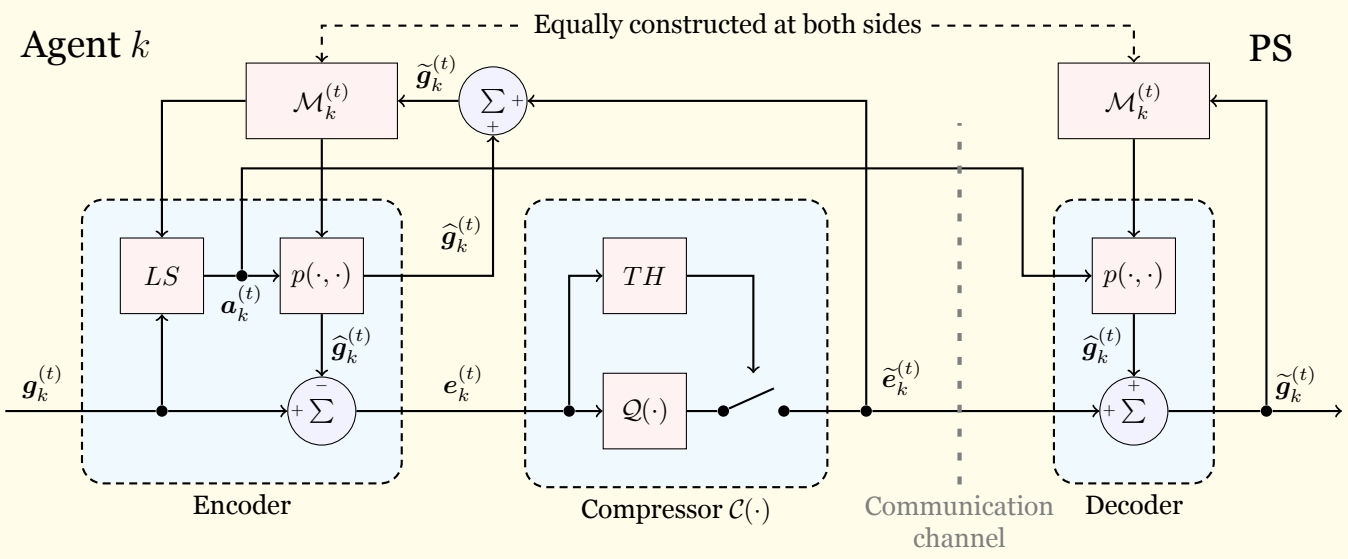
Method	Residual trans. frequency [%]	# iters.	# bits $[\times 10^5]$
Gradient Difference [1]	100	731	135.20
LAQ [2]	9.72	691	12.49
Proposed $s = 1$	4.75	726	8.60
Proposed $s = 2$	1.37	733	6.54
Proposed $s = 3$	1.01	733	8.40
Proposed $s = 5$	0.94	733	13.00

[1] K. Mishchenko, et al., "Distributed Learning with Compressed Gradient Differences," Dec. 28, 2023, arXiv: arXiv:1901.09269.
[2] J. Sun, et al., "Lazily Aggregated Quantized Gradient Innovation for Communication-Efficient Federated Learning," IEEE Trans. on Pattern Anal. and Mach. Intell., vol. 44, no. 4, pp. 2031–2044, Apr. 2022.

Conclusions

We proposed a prediction-based gradient compression with event-triggered communication for distributed learning. Remarks:

- The current gradient can be actively predicted using previous gradient information.
- Omitting transmission when prediction is accurate significantly reduce the amount of transmitted bits.



Block Diagram for Agent k

(Each agent is processed independently)

- Synchronized memory updates use imperfect gradient $\widetilde{\boldsymbol{g}}_k^{(t)}$
- Prediction coefficients $\boldsymbol{a}^{*(t)}$ transmitted with high bit count, *i.e.* negligible distortion.

$\Longrightarrow$ Both agent and PS knows $\widehat{\boldsymbol{g}}_k^{(t)}$

This work was presented at the 60th Allerton conference on Communication, Control, and Computing, Illinois, USA, 2024.

Link to the paper and more information: https://ostpopcorn.se/allerton24