Theoretical Limits of Pipeline Parallel Optimization and Application to Distributed Deep Learning

— Igor Colin, Ludovic Dos Santos, Kevin Scaman
Huawei Noah's Ark Research Lab, Paris

www.huawei.com

HUAWEI TECHNOLOGIES CO., LTD.

Introduction

Problem:

- Complex architectures, such as deep learning networks, require large amount of computational power
- Optimization has to be efficiently distributed, either
  - Data-wise or
  - Model-wise
- Pipeline distribution has shown remarkable results by parallelizing both
- Still lacks theoretical analysis, especially in extreme settings (e.g., few-shot learning)

Objective of the paper:

- Provide a general framework for the theoretical analysis of pipeline optimization
- Use randomized smoothing to better adapt to extreme settings

Proposed approach

Method:

- Combination of **pipeline distribution** and **randomized smoothing**:
  - Data flows through the computation graph as for pipeline distributed computing
  - Before sending the data to the next one, each node performs several evaluations using randomized smoothing
- Pipeline parallelization allows for minimum waiting time while randomized smoothing increases the overall convergence rate

Theoretical results:

1. On non-smooth optimization problem, the convergence rate is optimal up to a $d^{1/4}$ factor
2. Compared to pipeline optimization of depth Δ, the convergence rate on finite sums of size m is improved, from $O((RL/\varepsilon)^2(m+\Delta))$ to $O((RL/\varepsilon)^2 m + (RL/\varepsilon)\Delta d^{1/4})$

- This shows near optimality of pipeline optimization for large datasets and an accelerated way to solve problems where m is small (*e.g.*, few shot learning)

Experimental results

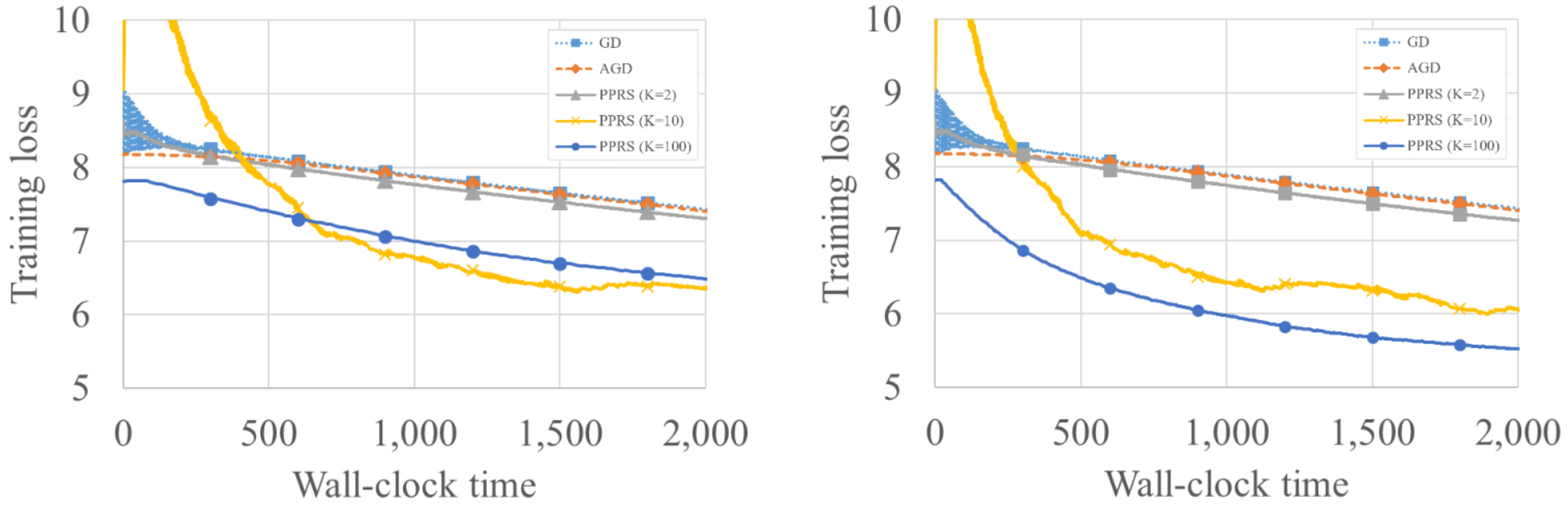

Figure 3: Comparison with GD and AGD. Increasing the number of samples increases the stability of PPRS and allows for faster convergence rates. Depth: (left) moderate, $\Delta = 20$. (right) high, $\Delta = 200$.

Conclusion

Wrap-up:

- We establish a new method for leveraging the downtimes in pipeline optimization
- We establish lower-bounds, showing optimality of our approach and its benefits in some settings
- Practical experiments on adversarial examples creation support the theoretical results

Future work:

- Adaptive hyperparameters selection is a challenging task that could lead to dramatic improvement of the convergence rates and the stability of the method in practical settings