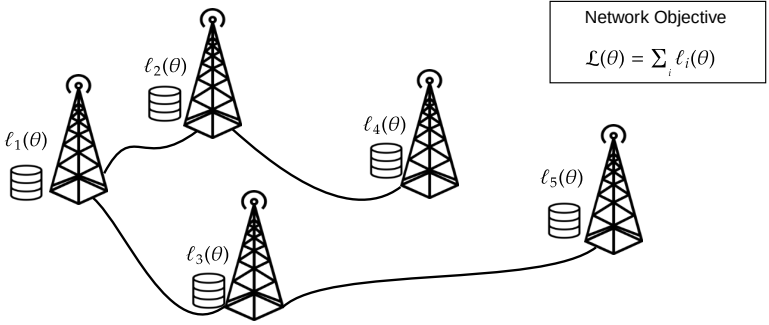
Communication-Efficient Distributionally Robust Decentralized Learning

Matteo Zecchin

SSIE 2021 - Student Workshop

Decentralized learning

Data-driven optimization methods deal with datasets distributed across multiple computing nodes. The costumary system objective is $\mathcal{L}(\cdot)$ the sum of loss terms at the network nodes $\{\ell_i(\cdot)\}_{i=1}^{m}$.

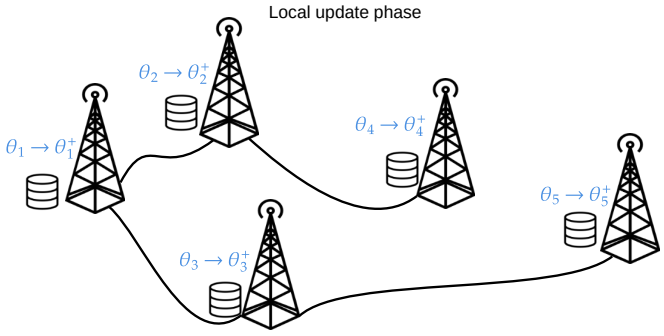


The goal is to find a configuration of the θ that minimizes the function $\mathcal{L}(\theta)$.

D-SGD

Decentralized (stochastic) gradient descent (D-SGD) is a versatile tool that harness data and computational resources in a privacy preserving and fault tolerant manner:

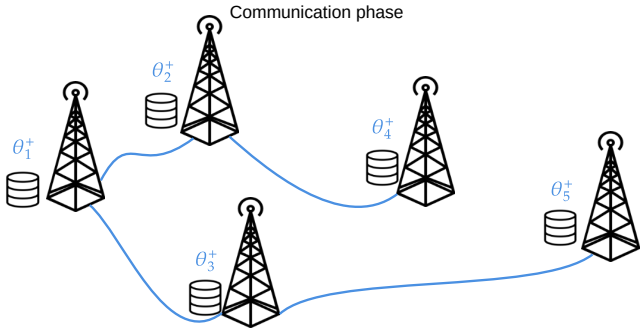
- Local data is never off-loaded and used only for local updates.
- Flexible w.r.t. communication topology.



D-SGD

Decentralized (stochastic) gradient descent (D-SGD) is a versatile tool that harness data and computational resources in a privacy preserving and fault tolerant manner:

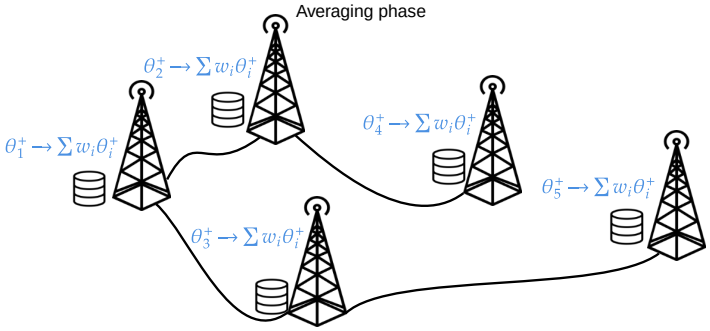
- Local data is never off-loaded and used only for local updates.
- Flexible w.r.t. communication topology .



D-SGD

Decentralized (stochastic) gradient descent (D-SGD) is a versatile tool that harness data and computational resources in a privacy preserving and fault tolerant manner:

- Local data is never off-loaded and used only for local updates.
- Flexible w.r.t. communication topology .

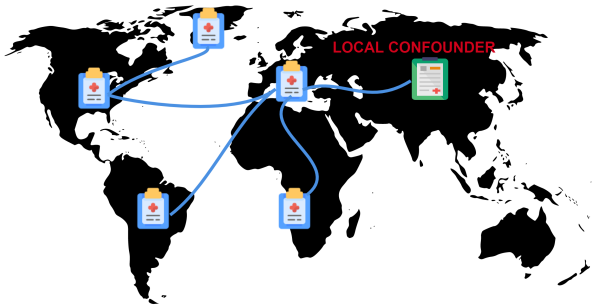


Fairness through distributional robustness

Data heterogeneity entail a fundamental challenge, deeply linked to the notion of **fairness.**

The vanilla formulation of loss function may neglect hard and under-represented sub-populations.

Example: Drug discovery with world-wide data and local extraneous factor.



The minimizer of the average loss formulation may perform badly and dangerously Asia.

Distributional Robust Formulation

Minimax formulation of the learning process as to focus on worst case performance.

Given a bunch of dataset from differently distributed dataset $\mathcal{D}_i \sim P_i^{\otimes n_i}$ for $i = 1, \ldots, m$ define the convex hull

$$\mathcal{P} := \left\{ \sum_i \lambda_i P_i \middle| \lambda_i \geq 0, \ \sum_i \lambda_i = 1 \right\}$$

and the distributionally robust objective

$$\max_{P \in \mathcal{P}} \min_{\theta \in \mathbb{R}^d} \mathbb{E}_{z \sim P} \left[\ell(\theta, z) \right]$$

Interpretation: two player game, player one tunes θ to minimize the loss w.r.t. to P while player two chooses the most challenging distribution P for the current model θ.

Fairness and distributional robust have been previously considered in federated learning [1].

[1] Mohri, Mehryar, Gary Sivek, and Ananda Theertha Suresh. "Agnostic federated learning." International Conference on Machine Learning. 2019.

Agnostic Decentralized Gradient Descent Ascent (AD-GDA)

We propose a AD-GDA, a distributionally robust learning algorithm for the **fully decentralized** learning.

AD-GDA objective:

$$\max_{\lambda \in \Lambda} \min_{\theta \in \mathbb{R}^d} \underbrace{\sum_{i=1}^{m} \lambda_i \mathbb{E}_{z \sim P_i} \left[\ell(\theta, z)\right]}_{\text{adversarial loss mixture}} + \mu r(\vec{\lambda})$$

regularizer to prevent degenerate adversary config.

Key features of AD-GDA

- **Single loop** ascent-descent: computing nodes query once the local gradient oracle and perform a **parallel** gradient descent step in θ and an ascent one in λ.
- **Communication efficient**: nodes exploit compressed gossip to share reduce the network communication load.

Guarantees

AD-GDA is a **provably convergent** algorithm, we provide convergence guarantees in both convex and non-convex settings

- In the **convex** scenario, AD-GDA approaches a solution with optimality gap $\mathcal{O}(\frac{1}{\sqrt{T}})$
- In the **non-convex** scenario, AD-GDA returns a approximate stationary solution with gradient vanishing as $\mathcal{O}(\frac{1}{\sqrt{T}})$

Key proof ingredient: the step-size for θ must be smaller that the one used to update λ.

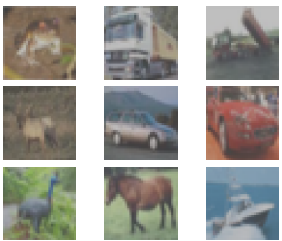
$$\Downarrow$$

The optimization problem w.r.t. θ changes slowly enough to optimize w.r.t. λ.

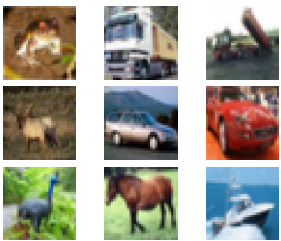
Experiment

Set-up:

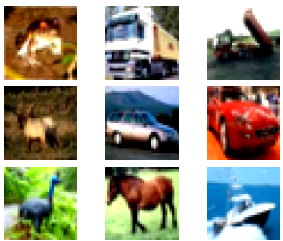
- Classification task over CIFAR-10 dataset scattering data over a 10 nodes connected according to a ring topology.
- One node in the system store images with lower contrast and one with higher contrast.



(a) low contrast



(b) original contrast

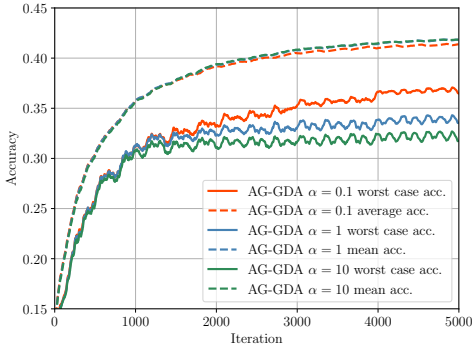


(c) high contrast

Experiment Results

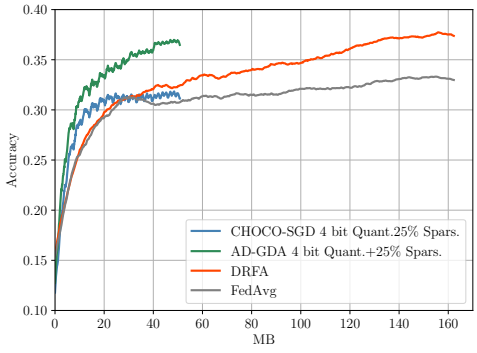
Effect of regularization:

- AD-GDA increases the worst case accuracy with a modest reduction in the avearage accuracy.
- Regularization param. allows to trade the two.



Communication Efficiency:

- AD-GDA employs compression to reduce the comm. load.
- AD-GDA beats distributionally robust federated learning and CHOCO-SGD.



Recap

Problem: decentralized learning is becoming increasingly popular but in case of heterogeneous local data sources can produce **unfair** predictors.

Starting from the **distributionally robust** formulation of the learning task, we provide a AD-GDA which:

- effectively trades average accuracy for worst case performance (fair predictors).
- is based on a single loop procedure (simple and fast).
- employs message compression (communication efficient).
- is provably convergent (sound algorithm).

Recap

Problem: decentralized learning is becoming increasingly popular but in case of heterogeneous local data sources can produce **unfair** predictors.

Starting from the **distributionally robust** formulation of the learning task, we provide a AD-GDA which:

- effectively trades average accuracy for worst case performance (fair predictors).
- is based on a single loop procedure (simple and fast).
- employs message compression (communication efficient).
- is provably convergent (sound algorithm).

Thank you for your attention :)