# GAL: Gradient Assisted Learning for Decentralized Multi-Organization Collaborations

Enmao Diao[1] Jie Ding[2] Vahid Tarokh[1]
[1]Duke University [2]University of Minnesota-Twin Cities

NEURAL INFORMATION PROCESSING SYSTEMS

## Overview

The main idea of Gradient Assisted Learning (GAL) is outlined below. In the training stage, the organization to be assisted, denoted by Alice, will calculate a set of "residuals" and broadcast these to other organizations. Subsequently, other organizations will fit the residuals using their local data, models, and objective functions and send the fitted values back to Alice. Next, Alice will line-search for the optimal gradient assisted learning rate along the calculated direction of learning. The above procedure is repeated until Alice accomplishes sufficient learning. In the inference stage, other organizations will send their locally predicted values to Alice, who will then assemble them to generate the final prediction.

- We propose a Gradient Assisted Learning (GAL) algorithm that is suitable for large-scale autonomous decentralized learning. Our method enables simultaneous collaboration among organizations without sharing data, models, and objective functions.
- We provide asymptotic convergence analysis and practical case studies of GAL. For the case of vertically distributed data, GAL generalizes the classical Gradient Boosting algorithm.
- Our proposed method can significantly outperform learning baselines and achieve near-oracle performance on various benchmark datasets. Compared with Assisted Learning (AL) [1,2], GAL does not need frequent synchronization of organizations. It also significantly reduces the computation and communication overhead.

## Motivation

***Question:*** *A large-scale autonomous decentralized learning method that can avoid data, models, and objective functions transparency?*

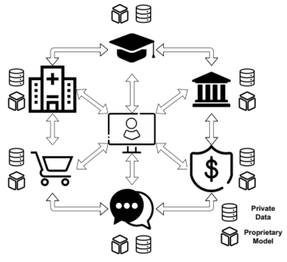


**Figure 1**. Decentralized organizations form a community of shared interest to provide better Machine Learning-as-a-Service.

The organizations can match the correspondence with common identifiers, such as user identification associated with the registration of different online platforms, timestamps associated with different clinics and health providers, and geo-locations.

**Examples:**

- A medical institute may be helped by multiple clinical laboratories and pharmaceutical entities to improve clinical treatment and facilitate scientific research.
- Financial organizations may collaborate with universities and insurance companies to predict loan default rate.
- Supply chain management can leverage decentralized manufacturing data to forecast future production and demand.

Paper

Code

## Method

### *Formulation*

- Suppose that there are $N$ data observations independently drawn from a joint distribution $p_{xy} = p_x p_{y|x}$, where $y \in Y$ and $x \in \mathbb{R}^d$ respectively represent the task label and feature variables, and $d$ is the number of features.
- Suppose that there are $M$ organizations. Each organization $m$ only holds $X_m$, a sub-vector of $X$. Alice, the organization to be assisted, has local data $x_1$ and task label $y_1$, while other $M-1$ organizations are collaborators which assist Alice and have local data $x_2 \dots x_M$.

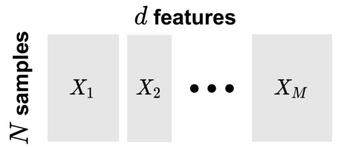


**Figure 2**. Illustration of organizations' vertically distributed data.

- Objectives (Alice, Oracle) :

$$F_{\text{Alone}} = \underset{F_1 \in \mathcal{F}_1}{\operatorname{argmin}} \mathbb{E}_N L_1(y_1, F_1(x_1)) \qquad F_{\text{Joint}} = \underset{F \in \mathcal{F}}{\operatorname{argmin}} \mathbb{E}_N L_1(y_1, F(x))$$

- GAL Objective:

$$f_m = \underset{f \in \mathcal{F}_m}{\operatorname{argmin}} \mathbb{E}_N \ell_m(r_1, f(x_m)) = \underset{f \in \mathcal{F}_m}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \ell_m(r_{i,1}, f(x_{i,m}))$$

### *The GAL algorithm*

- Local organizations iteratively fit the pseudo-residuals $r_1^t = -\left[\frac{\partial L_1(y_1, F^{t-1}(x))}{\partial F^{t-1}(x)}\right]$
- Optimize the gradient assistance weights

$$\hat{w}^t = \operatorname{argmin}_{w \in P_M} \mathbb{E}_N \ell_1\left(r_1^t, \textstyle\sum_{m=1}^{M} w_m f_m^t(x_m)\right)$$

- Line-search for the gradient assisted learning rate

$$\hat{\eta}^t = \operatorname{argmin}_{\eta \in \mathbb{R}} \mathbb{E}_N L_1\left(y_1, F^{t-1}(x) + \eta \textstyle\sum_{m=1}^{M} \hat{w}_m^t f_m^t(x_m)\right)$$

- Alice perform a functional gradient descent step in the form of

$$F^1 \leftarrow F^0 - \eta \cdot \frac{\partial}{\partial F} \mathbb{E}_{p_{x,y}} L_1(y, F(x))\,|_{F=F^0} = F^0 - \eta \cdot \mathbb{E}_{p_{x,y}} \frac{\partial}{\partial F} L_1(y, F(x))\,|_{F=F^0}$$

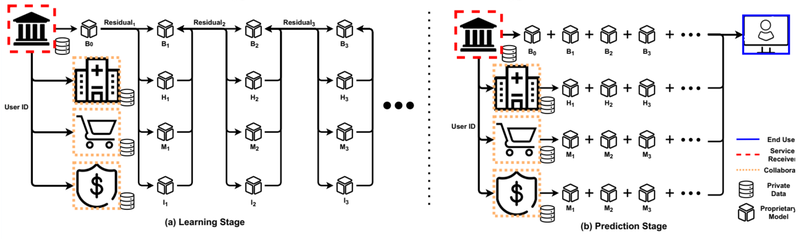


**Figure 3**. Learning and Prediction Stages for Gradient Assisted Learning (GAL).

### *Theoretical analysis*

We also provide an asymptotic convergence analysis for the GAL algorithm, where the goal is to minimize a loss $f \mapsto \mathcal{L}(f)$ over a function class through step-wise function aggregations. Because of the greedy nature of GAL, we consider the function class to be the linear span of organization-specific $\mathcal{F}_m$.

$$\mathcal{F}_M = \left\{ f : x \mapsto \sum_{m=1}^{M} w_m f_m(x_m), \forall f_m \in \mathcal{F}_m, x \in \mathbb{R}^d, w \in P_M \right\}$$

**Theorem 1** *Assume that the loss (functional) $f \mapsto \mathcal{L}(f)$ is convex and differentiable on $\mathcal{F}$, the function $u \mapsto \mathcal{L}(f + ug)$ has an upper-bounded second-order derivative $\partial^2 \mathcal{L}(f+ug)/\partial u^2$ for all $f \in span(\mathcal{F}_1, \dots, \mathcal{F}_M)$ and $g \in \cup_{m=1}^{M} \mathcal{F}_m$, and the ranges of learning rates $\{a_t\}_{t=1,2,\dots}$ satisfy $\sum_{t=1}^{\infty} a_t = \infty$, $\sum_{t=1}^{\infty} a_t^2 < \infty$. Then, the GAL algorithm satisfies $\mathcal{L}(F^t) \to \inf_{f \in span(\mathcal{F}_1, \dots, \mathcal{F}_M)} \mathcal{L}(f)$ as $t \to \infty$, with a convergence rate at the order of $O(\sum_{\tau=1}^{t} (a_{1:\tau}/a_{1:t}) a_\tau^2)$.*

## Experiments

- **Datasets:** UCI, MNIST, CIFAR10, ModelNet40, ShapeNet55, and MIMIC3

### *Model Autonomy*

**Table 1**. Results of the UCI datasets ($M = 8$) with Linear, GB, SVM and GB-SVM models. an organization with little informative data and free choice of its local model (model autonomy) can leverage others' local data and models and even achieve near-oracle performance.

| Dataset | Model | Diabetes(↓) | BostonHousing(↓) | Blob(↑) | Wine(↑) | BreastCancer(↑) | QSAR(↑) |
|---|---|---|---|---|---|---|---|
| Late | Linear | 136.2(0.1) | 8.0(0.0) | 100.0(0.0) | 100.0(0.0) | 96.9(0.4) | 76.9(0.8) |
| Joint | Linear | 43.4(0.3) | 3.0(0.0) | 100.0(0.0) | 100.0(0.0) | 98.9(0.4) | 84.0(0.2) |
| Alone | Linear | 59.7(9.2) | 5.8(0.9) | 41.3(10.8) | 63.9(15.6) | 92.5(3.4) | 68.8(3.4) |
| AL | Linear | 51.5(4.6) | 4.7(0.6) | 97.5(2.5) | 95.1(3.6) | 97.7(1.1) | 70.6(5.2) |
| GAL | Linear | 42.7(0.6) | 3.2(0.2) | 100.0(0.0) | 96.5(3.0) | 98.5(0.7) | 82.5(0.8) |
| GAL | GB | 56.5(2.8) | 3.8(0.5) | 96.3(2.2) | 95.8(1.4) | 96.1(1.0) | 84.8(0.9) |
| GAL | SVM | 46.6(1.4) | 2.9(0.2) | 96.3(4.1) | 96.5(1.2) | 99.1(1.1) | 85.5(0.7) |
| GAL | GB-SVM | 49.8(2.6) | 3.4(0.8) | 70.0(7.9) | 95.8(1.4) | 93.2(1.6) | 82.9(1.5) |

### *Deep Model Sharing*

- Because local deep learning models such as CNN can consume extensive computation space, we propose Deep Model Sharing (DMS) to allow sharing feature extractors of deep models across all iterations to save memory.

**Table 2**. Results of the MNIST and CIFAR10 ($M = 8$) datasets with CNN model.

| Dataset | MNIST(↑) | CIFAR10(↑) |
|---|---|---|
| Interm | 98.8(0.1) | 78.2(0.2) |
| Late | 98.0(0.1) | 74.4(0.3) |
| Joint | 99.4(0.0) | 80.1(0.2) |
| Alone | 24.2(0.1) | 46.3(0.3) |
| AL | 34.3(0.1) | 51.1(0.2) |
| GAL | 96.3(0.6) | **74.3(0.2)** |
| $\text{GAL}_{\text{DMS}}$ | **96.3(0.5)** | 67.0(0.3) |

### *Comparison with AL*

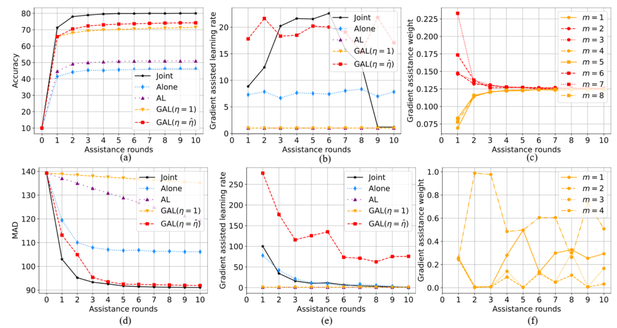


**Figure 4**. Results of the CIFAR10 (a-c) ($M = 8$) and MIMICL (d-f) ($M = 4$) datasets. GAL significantly outperforms 'Alone' and 'AL'. Our method also performs close to the centralized baselines. The gradient assisted learning rate diminishes to zero as the overarching loss converges. A constant gradient assisted learning rate ($\eta = 1$) converges much slower. The gradient assistance weights exhibits interpretability of the importance of organizations as the weights of the central image patches ($m = \{2,3,6,7\}$) of CIFAR10 dataset are larger than the boundary patches ($m = \{1,4,5,8\}$).

### *Case Studies*

**Table 3**. Results of case studies of 3D object recognition and medical time series forecasting.

| Dataset | ModelNet40(↑) | ShapeNet55(↑) | MIMICL(↓) | MIMICM(↑) |
|---|---|---|---|---|
| Interm | 75.3(18.2) | 88.6(0.1) | 64.6(0.9) | 0.90(0.0) |
| Late | 86.6(0.2) | 88.4(0.1) | 71.4(0.2) | 0.91(0.0) |
| Joint | 46.3(1.4) | 16.3(0.0) | 91.1(0.7) | 0.82(0.0) |
| Alone | 76.4(1.1) | 81.3(0.6) | 106.1(0.3) | 0.78(0.0) |
| AL | 77.3(2.8) | 83.8(0.0) | 119.3(0.3) | 0.86(0.0) |
| GAL | 83.0(0.2) | 84.1(0.6) | **91.9(2.3)** | **0.88(0.0)** |
| $\text{GAL}_{\text{DMS}}$ | **83.2(0.3)** | **85.3(0.2)** | 97.7(2.9) | 0.81(0.0) |

### *Noisy training with gradient assistance weights*

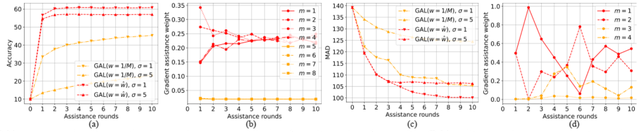


**Figure 5**. GAL equipped with gradient assistance weight significantly outperforms the GAL with direct average under noise injections.

## References

[1] Xian, Xun, et al. "Assisted learning: A framework for multi-organization learning." *Advances in Neural Information Processing Systems* 33 (2020): 14580-14591.

[2] Resources related to Assisted Learning (AL) can be found at http://www.assisted-learning.org.